Universidad de Valparaíso
CHILE
Facultad de Ciencias
Instituto de Estadística

MODELOS LINEALES GENERALIZADOS PARA VARIABLES DE RESPUESTA BINARIA BAJO PLANES DE MUESTREO COMPLEJO

Trabajo de titulación presentado por:
Alexis David Silva Barraza

para optar al grado de:
Licenciado en Estadística

y al título profesional de:
Ingeniero en Estadística

Profesor guía:
Carlos Felipe Henríquez Roldán, PhD

Valparaíso, Chile
2014

Agradecimientos

A mi madre, Marisol Barraza, por toda la ayuda que me brindó para poder estudiar, empezando por la educación básica y media, obligándome a hacer las tareas, ayudándome a estudiar una y otra vez sin quejarse, a pesar de todas las tareas propias que tenía. Además de siempre creer en mí, lo cual me ayudó mucho para llegar a terminar este trabajo, ya que muchas veces uno piensa que no lo podrá lograr y ese apoyo es lo que te hace seguir adelante, ¡gracias mamita!

A mi padre, Mario Silva, por todo el esfuerzo que realizó trabajando día y noche sin descanso, para costear mis estudios, además de todas sus enseñanzas para ser una mejor persona.

A mi hermano, Felipe Silva, que quiero dejar bien claro que es un perro, por esperar fielmente a que terminara mis estudios para salir a hacer deporte y jugar conmigo. Gracias por tu compañía y apoyo.

A mi hermana, Alejandra Silva, que a pesar de todas las discusiones, siempre estuvo ahí para apoyarme en lo que pudiese.

A mi profesor guía Carlos Henríquez quien me apoyo con toda su sabiduría para la elaboración de este proyecto teniendo una tremenda paciencia, haciendo así posible el desarrollo de este trabajo. Además de brindarme herramientas importantes para mi desarrollo como profesional.

A todos mis amigos de la universidad, por esas largas noches de estudio llenas de conocimiento compartido y muchas tallas.

A Ester Barrios, quien a pesar de llegar a mi vida en el último año de universidad, siempre me ha da su apoyo y me ha motivado a dar y ser más, sacando lo mejor de mí en varios aspectos de mí vida.

Resumen

El principal objetivo de esta tesis es estudiar la regresión logística bajo planes de muestreo complejos, pero para llegar a esto se debe realizar una descripción de la regresión logística: fórmulas, usos y otras cosas. Para después mostrar cómo se debe realizar un análisis de regresión logística cuando el muestro utilizado es un muestreo aleatorio simple, acá se verán los pasos que se deben seguir en el análisis. Además, se buscará un método vía simulación para estimar el tamaño muestral necesario para un correcto análisis de regresión logístico, cuando el muestreo a utilizar es el muestreo aleatorio simple. Por ejemplo para una población de 10.000 individuos se deben encuestar a 4500 individuos con un error estándar de estimación del 5 % y un nivel de confianza del 95 %.

Al llegar al estudio de muestras complejas se analizan algunas variables de la encuesta CASEN en donde se puede ver la diferencia que existe al analizar una muestra compleja de forma errónea sin incorporar el plan de muestreo a la forma correcta incorporando el plan de muestreo. Al realizar estos análisis se pudo apreciar la gran diferencia que existe en los resultaos al incorporar el plan de muestreo utilizado para la captura de los datos, comparando con no incorporarlo. Por otro lado para ver el tamaño muestral necesario cuando el plan de muestreo a utilizar es complejo, se simularon variables a partir de una base de datos real la que contenía el número de manzanas censales por comuna y el número de viviendas, después se obtuvo una muestra fija de 10 individuos por estrato y se calcularon los coeficientes de regresión logística y el efecto de diseño $DEFF$ el que indica de cuanto más debe ser el tamaño de la muestra, ya que, al multiplicar el $DEFF$ por el tamaño muestral se sabrá cuál es el tamaño de muestra necesario para realizar un correcto análisis de regresión logístico.

Introducción

El tema surge en una conversación con el profesor Carlos Henríquez, ante la consulta, ¿se debe analizar de la misma forma una muestra que ha sido obtenida con un diseño muestral complejo a otra obtenida con un muestreo aleatoria simple? En eso el profesor responde: excelente pregunta, claro que se debe analizar de diferente forma, ya que al utilizar un muestreo complejo las ponderaciones cambian por etapa, y se deben incorporar los factores de expansión para el análisis, lo que hace variar los resultados a hacer de cuenta de que los datos provienen de una muestra aleatoria simple; este es un típico error que cometen casi todas los que realizan análisis de datos que fueron capturados a través de un plan de muestreo complejo, ya que, el software analiza los datos con la programación por defecto, es decir como una muestra aleatoria simple, ahí nace el principal interés por estudiar las muestras complejas, por otro lado el tema que siempre había encontrado interesante era la regresión logística ya que a partir de respuestas binarias se pueden calcular las probabilidades que tiene un evento en ocurrir, además en el análisis se pueden incluir variables discretas y continuas lo que lo hace aún más interesante. Ahí nació el interés y motivación para estudiar la regresión logística en muestras complejas.

Al revisar diferentes libros y artículos dedicados a la regresión logística no se encontró material sobre tamaño muestral necesario para realizar un correcto análisis de regresión logístico, al revisar se encontró una fórmula 10*(k+1) donde k representa el número de variables regresoras, sin tomar en cuenta la variabilidad, la población total, el error estándar con el que se desea trabajar y el nivel de confianza; lo que hace pensar que la fórmula es errónea. Por lo tanto quedaba una interrogante ¿cuál es el tamaño muestral necesario para realizar un correcto análisis de regresión logístico? Por otro lado en los cursos de muestreo, siempre se habló de encontrar el tamaño muestral vía simulación con el software Stata, lo que motivo a buscar un método vía simulación que indique cual es el tamaño muestral necesario para un correcto análisis de regresión logístico, en muestreo aleatorio simple y complejo.

El interés principal recae en ver si existe diferencia al analizar una muestra compleja como aleatoria simple sin incorporar el plan de muestro y como compleja incorporando el

plan de muestreo, ya que si no hubiese diferencia daría lo mismo incorporar o no el plan de muestreo; para probar la diferencia que existe en el análisis, se utilizara la encuesta CASEN en la que los datos son obtenidos con un marco muestral complejo; la que será analizada sin incorporar el plan de muestreo e incorporando el plan de muestreo, para posteriormente comparar los resultados.

Abreviaturas y Símbolos

A continuación se presentarán todos los símbolos y abreviaturas utilizados en la investigación.

mlg	Modelo lineal generalizado.
rg	Regresión logística.
vm	Verosimilitud máxima.
mcp	Mínimos cuadrados ponderados.
smv	Seudo-máxima verosimilitud.
fda	Función de distribución acumulada
cll	Complemento log-log
fe	Factor de expansión
mas	Muestreo aleatorio simple.
mc	Muestreo complejo.
va	Variable aleatoria.

Objetivos

Objetivo General

- Estudiar la regresión logística bajo planes de muestreo complejos.

Objetivos Específicos

- Realizar una descripción de la regresión logística.
- Estudiar la regresión logística tradicional bajo muestreo aleatorio simple.
- Estudiar la regresión logística bajo panes de muestreo complejos.
- Determinar tamaños de muestras para regresión logística bajo planes de muestreo complejo.
- Aplicar resultados en datos reales y simulados incorporando plan de muestreo.

Índice general

Capítulo 1

Regresión logística y tipos de muestreo

1.1 Introducción

El objetivo de este capítulo es dar una noción general sobre los modelos logísticos como el modelo *logit*, *probit* y complemento $log - log$, además de mostrar los tipos de muestreos probabilísticos básicos más utilizados.

Los modelos lineales son de gran utilidad para modelar variables continuas, pero ¿qué ocurre cuando la variable respuesta no es continua si no binaria? Si este fuese el caso y se utilizara erroneamente un modelo lineal para modelar variables binarias, se obtendrían estimaciones absurdas que sobrepasarían los rangos permisibles. Una solución a esto, son los modelos lineales generalizados; de los cuales existe una gran variedad, sin embargo en este trabajo de titulación se enfocará en los modelos lineales generalizados para variables de respuesta binaria; siendo el más conocido el modelo de regresión logística. Este modelo explica funciones de las probabilidades de éxito en función de covariables. Su aplicación fluye en diversas áreas del saber, salud, economía y las ciencias sociales.

1.2 Modelos lineales generalizados

Para explicar que son los modelos lineales generalizados, se debe partir por la base de los modelos lineales más conocidos en la literatura estadística como: regresión y ANOVA; los que se basan en supuestos como: (1) los errores se distribuyen normalmente, (2) la varianza es constante, y (3) la variable respuesta se relaciona linealmente con la o las variables regresoras. Pero en muchas ocasiones, uno o varios de estos supuestos no se cumplen. Por ejemplo, encuestas donde la variable respuesta solamente tenga dos posibles resultados. En éste caso, no se podría aplicar un modelo lineal, dada la no existencia de linealidad entre la variable respuesta y las variables regresoras. Por esta razón, una alternativa a los modelos lineales (ml) son los modelos lineales generalizados (mlg) que permiten utilizar distribuciones no normales como: distribución binomial, Poisson y muchas otras.

Ciertos tipos de variables respuesta, sufren inevitablemente la violación de los supuestos en los modelos normales y los mlg ofrecen una buena alternativa para tratarlos. Específicamente, se puede considerar utilizar mlg cuando la variable respuesta es: un conteo de casos (por ejemplo, en estudios epidemiológicos); acotada en dos posibles resultados (ej. ciencias de la salud, tiene o no una enfermedad).

En este trabajo de titulación se estudiarán los mlg donde la variable respuesta tiene solo dos posibles resultados, más conocidos como variable binaria. Los mlg para variables de respuesta binarias y los ml para variables de respuesta continua tienen en común que la estimación de una ecuación de regresión que relaciona la variable respuesta y con una o más variables regresoras x. En ml el valor esperado de y es la media condicional de y dado un vector de covariables x, y se estima por una ecuación que es lineal en los parámetros de regresión:

$$Y = E(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad , \tag{1.1}$$

pero cuando Y es una variable binaria la $E(y|x) = \pi(x)$ que es la probabilidad condicional que $y = 1$ dado el vector de covariable x. Un enfoque inicial para modelar $\pi(x)$ es como función lineal de x, pero hay varios problemas con este enfoque tales como: (1) la variable respuesta sigue una distribución binomial, causando una grave violación en la normalidad y homogeniedad de varianza, supuestos necesarios para la eficiente estimación de los parámetros a través del método de mínimos cuadrados; (2) un modelo de regresión lineal para $\pi(x)$ no refleja con precisión la relación entre x e y como se puede ver en la figura 1.1, incluso podría producir valores para $\pi(x)$ fuera de los rangos permisibles de 0 a 1. Una alternativa para solucionar esto es identificar una función no lineal para $\pi(x)$, por ejemplo $g(\pi(x))$, esto produciría un modelo de regresión ajustado el que es lineal en los parámetro del predictor. Por otra parte, la función estimada, $g(\pi(x))$, debe ser elegida de modo que cuando se trasformen nuevamente los valores de $\pi(x)$, estén en el rango de 0 a 1. La función $g(\pi(x))$ es denominada función de enlace, donde se utiliza comúnmente las funciones *logit*, *probit* y *complemento log − log(cll)*, Heeringa West & Berglund [2010].

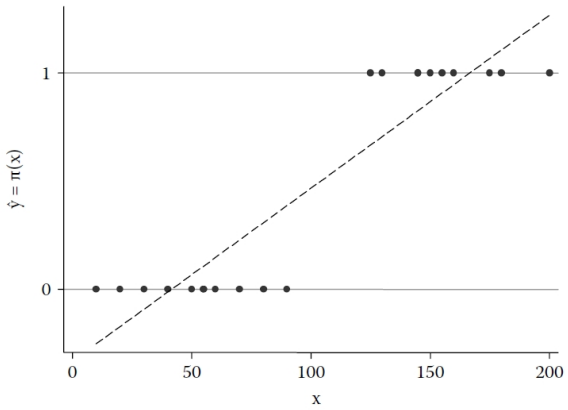

Figura 1.1: Modelo ingenuo Heeringa West & Berglund [2010]

La Figura 1.1 muestra el uso ingenuo de regresión lineal para variable con respuesta binaria.

1.2.1 Modelo de regresión *logit*

Para un modelo de regresión logístico (mrl), la función de enlace es el *logit* y estaría dada de la siguiente forma:

$$g(\pi(x)) = \text{logit}(\pi(\boldsymbol{x})) = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad , \tag{1.2}$$

donde *logit* es no lineal en $\pi(\boldsymbol{x})$, pero se supone linealidad en los parámetros, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$. Basado en el modelo de regresión ajustado para el *logit*, el valor estimado de $\pi(\boldsymbol{x})$ se puede recuperar con la función inversa de *logit*:

$$\widehat{\pi}(\boldsymbol{x}) = g^{-1}(\widehat{g}(\pi(\boldsymbol{x}))) = \frac{exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_p x_p)}{1 + exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_p x_p)}. \tag{1.3}$$

La función inversa $\widehat{g}^{-1}(\cdot)$ es la función de distribución acumulada (fda) de la distribución de probabilidad logística. Bajo el modelo logístico en la ecuación (1.2) y $\widehat{\pi}(\boldsymbol{x})$ es la fda logística evaluados en el *logit* estimado con el vector de covariables x.

Para los modelos *logit* no existe una solución directa para estimar los parámetros, como lo es en los ml con el método de mínimos cuadrados. Para estimar los coeficientes del regresor en el modelo logit debe usarse métodos iterativos, tales como el método de Newton-Raphson o Fisher scoring algorithm Agresti [2002] los que se utilizan para determinar los valores de los coeficientes estimados que maximizan la siguiente función de seudo-

probabilidad ponderada.

$$PL(\boldsymbol{\beta}|X) = \prod_{i=1}^{n} \left\{ \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \right\}^{w_i} \tag{1.4}$$

Es importante tener en cuenta que la función de enlace no es una transformación explícita de la variable y, sino más bien una transformación de $E(y|x) = \pi(\boldsymbol{x})$. También que la evaluación de seudo-riesgo en la ecuación (1.4) requiere las observaciones originales, y_i, los valores modelados de $\pi(x_i)$, y en el caso de los datos de encuestas complejas, los pesos del muestreo w_i, como se verá en el capítulo 3.

Según Agresti [2002] cada ciclo en la estimación iterativa del modelo de regresión logística requiere cuatro operaciones:

1 El valor del *logit* en la ecuación (1.2) se calcula para cada encuestado basado en los valores de iteración actual de los parámetros estimados: $Z_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p$

2 Para cada caso el *logit* se transforma a la probabilidad de escala, mediante la evaluación de la fda logística en los valores de z_i como se ilustra en la ecuación (1.3).

3 El valor de la probabilidad de la ecuación (1.4) se evalúa a continuación, en $i = 1, ..., n$ los valores de $\widehat{\pi}(x)$ y observaciones y_i.

4 El algoritmo a continuación ajusta el valor del parámetro individual de $\widehat{\beta}_j$ para maximizar la función de probabilidad, y vuelve a repetir el ciclo.

El algoritmo iterativo se detiene cuando el cambio en las estimaciones del vector de valores de β ya no aumenta el valor de la función de probabilidad.

1.2.2 Modelo de regresión *probit*

El modelo *probit* es una alternativa al modelo de regresión *logit* para modelar variables de respuesta binaria Agresti [2002]. Los modelos de regresión *probit* también son modelos lineales generalizados, y los procedimientos para la estimación son paralelos a los del modelos *logit*. Con la diferencia de que los modelos probit utilizan la distribución normal inversa:

$$g(\pi(x)) = \Phi^{-1}(\pi(x)) = z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1.5}$$

En contraste con la regresión logística, donde se asume que Z sigue una distribución de probabilidad logística, el probit asume una distribución normal estándar. Una vez más, como en la regresión logística, la transformación posterior de la escala probit a las cantidades de interés $\pi(x)$ requiere una evaluación de la fda normal en el valor estimado de probit, $Z_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}$

$$\widehat{\pi}(x_i)_{probit} = \Phi(z_i) = Prob(Z \leq z_i) = \int_{-\infty}^{\widehat{\beta}_o + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_{pi}} \frac{1}{2\pi} exp Z^2 dZ \tag{1.6}$$

Los pasos en la estimación del modelo de regresión probit son idénticos al descrito anteriormente para la regresión *logit*. Por lo tanto en general, las inferencias derivadas de la regresiones *logit* y *probit* no difieren significativamente Heeringa West & Berglund [2010].

1.2.3 Modelo complementario *log-log*

Se utiliza comúnmente en el análisis de regresión cuando la variable respuesta es binaria. La función de enlace *cll* está relacionada con la distribución de Gompertz y el enlace puede ser denominado como el "*gompit*." La distinción primaria para el enlace de *cll* es que a diferencia de la *logit* o probit la fda no se requiere que sea simétrica respecto al punto medio $\pi(x) = 0,5$. Su aplicación más común es en situaciones donde $\pi(x)$ es o muy cercana a 0 o de 1 (en uno o en el otro extremo). Una gran parte de software que en sus funciones permite el análisis *logit* y *probit* también permite la opción *cll*. Para mayor información puede ver Allison [1999].

1.2.4 Relación entre *logit*, *probit* y complemento *log-log*

En las secciones 1.2.1 y 1.2.2 se mostró que los modelo *logit* y *probit* estiman los parámetros y errores estándares de forma parecida. La diferencia se produce debido que la estimación de los parámetros de interes para el modelo *logit* se realiza con la función *logarítmica* mientras que en el modelo *probit* se utiliza la distribución normal. Con la función complemento *log-log* ocurre algo semejante, ya que,este modelo es una mezcla de los dos modelos anteriormente mencionados. Por esta razón al estimar los parámetros para diferentes tamaños muestrales los resultados de los tres modelos serían parecidos, haciendo innecesario el estudio de cada modelo por separado. Es por ello que este trabajo de titulación se centrara a estudiar el modelo *logit*.

1.3 Muestreo

Lohr [2000] señala que el muestreo es una herramienta de investigación científica, cuya función es determinar que parte de la población debe ser examinada, con la finalidad de hacer inferencias sobre ésta.

Para seleccionar una muestra (subconjunto de una población) a analizar se debe tener en cuenta: la unidad de observación (objeto sobre el cual se desea realizar el estudio); población objetivo (el total que podría extraerse); población muestreada (es la población de donde se extrae la muestra); unidad de muestreo (es la unidad de estudio que fisicamente se muestrea); marco de muestreo (es la lista de unidades de muestreo, ejemplo números telefónicos, direcciones, entre otros).

La muestra debe lograr una representación adecuada de la población, para que ésta sea representativa y por tanto útil. Deben de reflejarse las similitudes y diferencias encontradas en la población, es decir ejemplificar las características de ésta.

Existen diferentes criterios de clasificación de los diferentes tipos de muestreo, aunque en general pueden dividirse en dos grupos: muestreo probabilístico y no probabilístico. Los que se plantean en este trabajo de titulación son los muestreos probabilísticos, ya que permiten realizar una representación de la población; porque al extraer la muestra todos los individuos tienen la misma probabilidad de ser seleccionados. Los tipos de muestreo probabilísticos que existen son: aleatorio simple, estratificado, conglomerado, sistemático y complejo.

1.3.1 Muestreo aleatorio simple

El muestreo aleatorio simple es la forma más sencilla de muestreo probabilístico y proporciona la base teórica de muestreos más elaborados. Existen dos formas de extraer una muestra aleatoria simple: con remplazo, donde la unidad puede repetirse y sin remplazo, donde todas las unidades son distintas.

Una muestra aleatoria simple con remplazo de tamaño n, obtenida a partir de una población de tamaño N, en donde una unidad se extrae de una población al azar, para ser la primera unidad muestreada con probabilidad $1/N$. Luego, la unidad muestreada vuelve a la población y se repite éste proceso hasta que la muestra cuente con n elementos, donde puede haber elementos repetidos.

Sin embargo, cuando la población es finita resulta inadecuado tener elementos repetidos, ya que, no proporciona mayor información, por esta razón es conveniente utilizar el muestreo aleatorio simple sin remplazo donde cada individuo tiene la misma probabilidad de ser seleccionado. Siendo éste sin remplazo, existirán $\binom{N}{n}$ muestras posibles, donde cada una es igualmente probable, por lo tanto la probabilidad de elegir cualquier muestra

individual t de n es:

$$P(t) \;=\; \frac{1}{\binom{N}{n}} \;=\; \frac{n!(N-n)!}{N!} \tag{1.7}$$

Comúnmente para seleccionar los individuos a encuestar, no se utilizan métodos totalmente aleatorios sino seudo-aleatorios, ya que los datos se encuentran en el computador, éste genera un algoritmo para seleccionar cada elemento que será incluido en la muestra Lohr [2000].

1.3.2 Muestreo estratificado

Cuando se quiere muestrear una población, con frecuencia se dispone de información adicional sobre ésta, que ayudará a escoger el plan de muestreo más eficiente. Por ejemplo, antes de realizar una encuesta se sabe que los hombres comen más que las mujeres, que las verduras son más caras en el norte de Chile que en el sur, la cantidad de personas que viven en cada región y así muchas otras características se conocen a priori del levantamiento de los datos. El disponer de esta información adicional ayuda a diseñar un plan de muestreo.

Si la variable de interés asume distintos valores promedio en diferentes subpoblaciones, se obtendrían estimaciones más precisas de la población, al tomar una **muestra aleatoria estratificada**, los estratos en conjunto conforman la población completa, de modo que cada unidad de muestreo pertenece exactamente a un estrato. Para el análisis se extrae una muestra independiente de cada estrato y, posteriormente, se reune la información para obtener las estimaciones globales de las poblaciones.

Un ejemplo de porque es necesario utilizar el muestreo estratificado, es suponer una población de 2000 estudiantes donde 1000 estudiantes son hombres y 1000 estudiantes son mujeres, de la cual se desea obtener una muestra de 100 personas. Ésto desde un punto de vista teórico al aplicar un muestreo aleatorio simple es posible obtener una muestra solo de hombres o mujeres, por lo cual un gran número de personas diría que la muestra no es representativa, para evitar que suceda se podría extraer una muestra aleatoria simple de 50 hombres y 50 mujeres, así se garantiza que la proporción fuese la misma que la población, Lohr [2000].

1.3.3 Muestreo por conglomerado

Es un método en el cual la unidad de muestreo es un grupo de unidades a muestrear. Es decir, que cada grupo o conglomerado es lo que se debe muestrear. Cada conglomerado es considerado como una unidad de muestreo siendo éstas de diferentes tamaños.

En un muestreo por conglomerados se tienen 2 tipos de unidades:

1 Unidades de muestreo

2 Conglomerados

Puesto que en muestreo lo más caro es llegar a la unidad muestreada, entonces ¿sería posible llegar a un lugar y encuestar a todos en ese mismo lugar? Esto es lo que se hace en el muestreo por conglomerados; es decir, se incrementa el tamaño de la muestra a bajo costo.

La principal razón para usar el muestreo por conglomerados, es que encuestar a muchas personas ubicadas en lugares geográficos distintos puede ser difícil, caro incluso en algunas circunstancias imposible de obtener. Dado que es difícil llegar a todos los elementos, pero es más accesible tener todos los elementos de un lugar pequeño. Además de abaratar los costos de realizar una encuesta no obstante, se introduce un problema de dependencia entre las unidades que pertenecen a un mismo conglomerado Lohr [2000].

1.3.4 Muestreo sistemático

El muestreo sistemático es una forma ordenada de obtener una muestra, se utiliza cuando el universo o población es de gran tamaño. Primero se deben identificar las unidades y determinar fechas (cuando se deba). Luego hay que calcular una constante, que se denomina coeficiente de elevación $K = N/n$; donde N es el tamaño de la población (este caso es para poblaciones finitas cuando es infinita K puede ser dado) y n el tamaño de la muestra. Posteriormente se debe elegir al azar un número entre 1 y K; de ahí en adelante se registra la observación cada K intervalos.

Esto quiere decir que si se tiene un determinado número de personas que es la población (N) y se quiere escoger de esa población un número más pequeño el cual es la muestra (n), se divide el número de la población por el número de la muestra n que se quiere tomar y el resultado de esta operación será el intervalo, entonces se escoge un número al azar desde uno hasta el número del intervalo, y a partir de este número se escogen los demás siguiendo el orden pre obtenido Lohr [2000].

1.3.5 Muestreo complejo

El muestreo complejo es la combinación de los muestreos anteriormente mencionados, también conocidos como muestreo por etapas. Al utilizarse dos diferentes tipos de muestreo para la extracción de una muestra, se puede denominarlo como una muestra compleja. Esto es muy utilizado en la práctica, pero se puede ver con mayor frecuencia en encuestas nacionales como la encuesta CASEN, ya que al ser una población de millones de personas se buscan diferentes formas de obtener una muestra representativa a nivel país.

Capítulo 2

Regresión logística en muestreo aleatorio simple

2.1 Introducción

El software Stata 12 cuenta con varios comandos para el análisis de regresión logística, los que por defecto vienen implementados para realizar un análisis para datos que fueron capturados con un muestreo aleatorio simple, es por ello que en el presente capítulo se mostrarán los pasos a seguir para su modelación y los comandos que se utilizan.

Cuando se desea realizar un análisis de regresión logística, lo que menos se toma en cuenta es cuál es el tamaño de muestra necesario para realizar un correcto análisis. Ya que, normalmente los análisis se realizan con una muestra y ésta es utilizada para inferir sobre una población. Es por ello que en el presente trabajo de titulación se plantea realizar simulaciones para identificar cual será el tamaño de muestra necesario.

El objetivo de este capítulo es mostrar cómo se realiza un análisis de regresión logística, cuando el muestreo utilizado para la captura de los datos, es un muestreo aleatorio simple y además estimar cual es el tamaño de muestra necesario para realizar un correcto análisis de regresión logística considerando que el muestreo a utilizar es el muestreo aleatorio simple.

2.2 Construcción de un modelo de regresión logística

La construcción de un modelo de regresión logística se realiza de la misma forma que un modelo lineal y consta de las siguientes etapas:

1 La especificación del modelo.

2 La estimación de los parámetros y errores estándares.

3 La evaluación del modelo.

4 Realizar una interpretación de los resultados e inferencias basados en el modelo final.

Al igual que en todos los procesos de construcción de modelos estadísticos, la etapa 1-3 definen un proceso iterativo diseñado para refinar y probar el modelo. Varios ciclos de especificación del modelo, estimación y secuencias de evaluación suelen ser necesarios antes de que se identifique un modelo definitivo y se pueda hacer inferencias de la población a partir de los resultados obtenidos, Heeringa West & Berglund [2010].

2.3 Especificación del modelo

De todas las variables que pueda tener un estudio, ¿qué variables deben introducirse en el modelo? El modelo debe ser lo más reducido y que explique lo más que pueda (principio de parsimonia), y que además sea congruente e interpretable. Por otro lado hay que tener en cuenta que un mayor número de variables en el modelo implicará mayores errores estándares de los estimadores.

El modelo de regresión logístico, es una herramienta flexible para modelar, que permite la incorporacion de variables del tipo cualitativas y cuantitativas, e interacciones, además de la variable respuesta. Como en todos los modelos de regresión, la identificación de un mejor modelo de regresión logística para datos de encuestas debe seguir un proceso sistemático, científico mediante el que los predictores son identificados y evaluados. Se debe seguir un proceso gradual especificando el modelo, para el perfeccionamiento del conjunto de predictores, los pasos a seguir serían los siguientes:

I Realizar un análisis de la variable respuesta con cada variable regresora (se le conoce como análisis bivariado).

II Seleccionar las variables regresoras que están asociadas a la variable respuesta con significación $p \leq 0,25$ como candidato para los efectos principales en el modelo de regresión logística multivariante.

III Evaluar la contribución de cada variable regresora en el modelo multivariado.

IV Comprobar si existe linealidad entre las variables regresoras continuas.

En el paso I se analizará la variable respuesta con cada variable regresora, en el software Stata el que entrega un detalle de esta relación.

En el paso II se habla de significación de $p \leq 0,25$ y no de $p \leq 0,05$, ya que este último sería un criterio excesivamente restrictivo, además es recomendable utilizar $p \leq 0,25$ para incluir covariables con una débil asociación a la variable respuesta, porque asociada a las otra covariables podría demostrar ser un fuerte predictor.

En el paso III la comprobación de significancia estadística de cada uno de los coeficientes de regresión en el modelo. Para esto se pueden usar tres métodos: prueba de Wald, la prueba G de razón de verosimilitud y la prueba de Score.

En el paso IV, se requiere comprobar si existe linealidad entre las variables regresoras continuas (también conocida como multicolinealidad) se realiza verificando la correlación entre las variables regresoras. Si esta correlación fuese grande o relativamente confiable, su efecto sería el incremento de los errores estándar, y en ocasiones, del valor estimado para los coeficientes de regresión, lo que hace que las estimaciones sean poco creíbles.

2.4 Estimación de parámetro y errores estándar

Dado un modelo de regresión logístico de la forma $logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \ldots + \beta p x_p$, se calcularán las estimaciones de los parámetros junto con sus errores estándares. Para datos donde su captura fue realizada con un muestreo aleatorio simple. Los parámetros del mrl y los errores estándares pueden ser estimados usando el método de vm. La función de vm para un modelo de regresión logística, se basa en la distribución Bernoulli:

$$L(\beta|X) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \quad , \tag{2.1}$$

donde $\pi(x_i)$ está vinculado a los coeficientes del modelo de regresión a través de la función de distribución logística:

$$\pi(x_i) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)} \quad . \tag{2.2}$$

Para el cálculo de estimación por el método de vm se recurre a métodos iterativos, como el de Newton-Raphson. Ya que el cálculo es complejo, normalmente hay que recurrir al uso de rutinas de programación o paquetes estadísticos. De este modo no solo obtenemos la estimación de los parámetros sino también la de los errores estándares.

2.5 Evaluación y diagnóstico

Para probar la significancia de los parámetros, los efectos individuales y evaluar la bondad de ajuste global de un modelo de regresión logística, es necesario corroborarlo a través de pruebas estadísticas. Estas pruebas estadísticas pueden realizarse de forma individual; es decir, probar parámetro por parámetro o de forma conjunta con algunas variables de interés o el modelo completo. En Stata 12 se puede ver al utilizar el comando *logit* o de forma aparte con el comando *test*.

2.5.1 Pruebas para coeficientes individuales

Cuando se quiere probar si una variable es significativa para el modelo, se puede usar una prueba estadística. Una de las más conocidas es la prueba de Wald la cual es utilizada para probar si la contribución de los parámetros en un modelo de regresión logística es significativa como lo muesta Long & Freese [2006].

El estadístico Wald, para determinar el valor-p usa de referencia la distribución χ^2_1, que es una distribución con un grado de libertad. Además, en *Stata* se puede utilizar el comando *test* con la variable de interés (ejemplo *test* x_1), la cual mostrara el valor de χ^2_1 y el valor p. Por otro lado, el valor de χ^2_1 es el cuadrado de la prueba-z, que vemos en *Stata* al utilizar el comando *logit*.

Otro comando utilizado para medir la contribución de un coeficiente es *lrtest* para poder ilustrar como se utiliza, se presentara el siguiente ejemplo.

Dado un modelo de regresión logística con las variables y_1 x_1 x_2 x_3 x_4 x_5 en *Stata* se define de la siguiente forma:
logit y_1 x_1 x_2 x_3 x_4 x_5, *nolog*
luego se guarda el modelo con el comando
estimates store fmodel
Nuevamente definimos el modelo sin la variable x_3 para medir su contribución
logit y_1 x_1 x_2 x_4 x_5, *nolog*
Volvemos a guardar el modelo ahora con el nombre de nmodel
estimates store nmodel
Ahora realizamos la prueba *lrtest* con respecto a los modelos
Lrtest fmodel nmodel

La interpretación de *lrtest* con respecto a la variable x_3 se realiza utilizando el valor-p, sí el valor-p $\leq$ 0,01 se puede decir que su contribución es buena; en caso contrario es mejor quitar la variable (dependiendo con que nivel de significancia se desee trabajar normalmente se utiliza $\alpha = 0,05$ *o* $0,01$).

2.5.2 Pruebas para coeficientes múltiples

Para probar la contribución de dos o más parámetros se pueden utilizar los mismos comandos utilizados en pruebas para un solo coeficiente. El cambio que se produce es en los grados de libertad utilizados para la prueba de $Wald$, ya que estos cambian dependiendo del número de variables que se desea probar. Si son dos variables tendrá dos grados de libertad, si son tres tendrá tres grados de libertad y así sucesivamente. La interpretación del resultado se realiza de la misma forma que para un parámetro tanto para el comando *test* como para el *lrtest*

2.5.3 Análisis de residuos

Una forma importante de evaluar el ajuste del modelo, es con un análisis de residuos. Los residuos son la diferencia entre el modelo ajustado y las observaciones de la muestra,las diferencias pueden ser de distintos tamaño, pero las que son de interés son los casos más alejados, que son conocidos como datos atípicos (o outliers en inglés). Una observación atípica puede llegar a tener un gran efecto en la estimación de los parámetros es por esto que es de interés realizar el análisis residual, Long & Freese [2006].

2.5.3.1 Residuos

Se define la probabilidad predicha para un determinado conjunto de variables independientes como

$$\pi_i = Pr(y_i = 1|x_i) \quad ,$$

entonces la desviación $y_i - \pi_i$ es heterocedástica (cuando la varianza de las perturbaciones no es constante a lo largo de las observaciones), con $Var(y_i - \pi_i|x_i) = \pi_i(1-\pi_i)$ esto implica que la varianza sea mayor cuando $\pi_i = 0,5$ y menor cuando π_i se encuentre más cerca de 0 o 1. Por ejemplo $\pi = 0,5$ entonces la $Var = 0,5(1-0,5) = 0,25$ y si $\pi = 0,01$ entonces la $Var = 0,01(1-0,01) = 0,0099$. En otras palabras la heterocedasticidad depende de las probabilidad positivas. Lo que sugiere el uso de los residuos de Pearson, que dividen el residuo $y - \hat{\pi}$ por la desviación estándar.

$$r_i = \frac{y_i - \hat{\pi}}{\sqrt{\hat{\pi}(1-\hat{\pi})}} \quad .$$

El valor de r es la diferencia entre el valor ajustado y el observado. *Pregibon* [1981] muestra que la varianza de r no es 1, como $Var(y_i - \pi_i|x_i) \neq \pi_i(1-\pi_i)$, y propone los residuos de Pearson estandarizados

$$r_i^{std} = \frac{r_i}{\sqrt{1-h_{ii}}} \quad ,$$

donde

$$h_{ii} = \hat{\pi}_i(1 - \pi_i)x_i\hat{Var}(\hat{\beta})x_i^\top . \tag{2.3}$$

Aunque es preferible r^{std} a r dado a su varianza constante, en la práctica se puede encontrar que ambos modelos residuales son muy parecidos.

2.6 Interpretación de resultados

Del modelo de regresión logística, se puede hacer inferencias sobre el significado y la importancia de las variables predictoras de varias maneras. Sobre la hipótesis nula de que un único coeficiente es igual a cero, $H_0 : \beta_j = 0$, o más complejas sobre múltiples parámetros en el modelo ajustado. Los intervalos de confianza (IC) de los coeficientes del modelo individual también se pueden usar para sacar conclusiones acerca de la importancia de los predictores y proporcionar información sobre la magnitud potencial y la incertidumbre asociada con los efectos estimados de las variables predictoras individuales. Un intervalo de confianza basado en el diseño para el parámetro de regresión logística el que se calcula de la siguiente forma:

$$CI_{1-\alpha}(\beta_j) = \hat{\beta}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{\beta}_j) \quad , \tag{2.4}$$

por lo general, $\alpha = 0,05$ se utiliza (junto con los grados de libertad), y el resultado es un nivel de confianza del 95 % para el parámetro. En teoría, la inferencia correcta es que en un muestreo repetido, se espere que 95 de 100 muestras se encuentre en el intervalo de confianza calculado de esta manera se espera incluya el valor real de la población de β_j. Si el IC estimado incluye ln $(1) = 0$, los analistas pueden elegir para inferir que $H_0 : \beta_j = 0$ se acepta con una tasa de error tipo I, $\alpha = 0,05$. Inferencia sobre el significado e importancia de predictores puede llevase a cabo directamente por los β_j (en la escala log-odds). Sin embargo, para cuantificar la magnitud del efecto de un predictor individual, es más útil transformar la inferencia a una escala que es interpretada fácilmente por científicos y público en general. En un modelo de regresión logística con un solo predictor, x_1, una estimación de los odds-ratio correspondientes a un aumento de una unidad en el valor de x_1 se puede obtener exponencialmente con el coeficiente de regresión logística:

$$\hat{\psi} = exp(\hat{\beta}_1) \quad . \tag{2.5}$$

Si el modelo contiene sólo un único predictor, el resultado es una estimación de la odds ratio ajustada. Si el modelo de regresión logística equipada incluye múltiples predictores, es decir,

$$logit(\hat{pi}(x)) = ln\left[\frac{\hat{\pi}(x)}{1 - \hat{\pi}(x)}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p \quad . \tag{2.6}$$

El resultado de $\hat{\psi}_j|\hat{\beta}_{k\neq j} = exp(\hat{\beta}_j)$ es un odds ratio ajustado. En general, el odds ratio ajustado representa el efecto multiplicativo de un aumento de una unidad en la variable x_j, predictor de las probabilidades de la variable de resultado es igual a 1, manteniendo

constantes todas las demás variables predictoras. Los límites de confianza también se pueden calcular de la siguiente forma:

$$CI(\psi_j) = exp(\hat{\beta}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{\beta}_j)) \quad . \tag{2.7}$$

Los procedimientos de software para el análisis de regresión logística, ofrecen al analista la posibilidad de estimar los parámetros y errores estándares en la escala log-odds ($\beta_j s$ los originales) o estimaciones transformadas de los correspondientes odds ratios ajustados y los intervalos de confianza. Los odds ratios ajustados y los intervalos de confianza pueden ser estimados y reportados para cualquier tipo de variable predictora, incluidas las variables categóricas, ordinales, y continuas, Heeringa West & Berglund [2010].

2.6.1 Odds

Un odds corresponde a la razón entre la probabilidad de experimentar un evento en relación con la probabilidad de no experimentar; es decir, un cuociente de dos probabilidades. El cálculo de un odds (o chance) es muy utilizado en estudios de prevalencia (estudios a través del tiempo).

En un estudio de casos y controles, el cálculo de odds es de gran utilidad. De hecho, se puede calcular más de un odds. Pero, ¿cómo se ve en un modelo de regresión logística?

Un modelo de regresión logística está dado por

$$G(E(y)) = \beta_0 + \beta_1 x_1 = \beta^\top X \quad , \tag{2.8}$$

donde G( ) es una función que en este caso será ln y $E(y) = \frac{\pi}{1-\pi}$ que es lo mismo que decir la probabilidad de que ocurra un evento dado que no ocurra, por lo tanto se tendra.

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta^\top X \quad , \tag{2.9}$$

ahora despejando $\frac{\pi}{1-\pi}$ queda

$$\frac{\pi}{1-\pi} = exp(\beta^\top X) \quad , \tag{2.10}$$

donde

$$\pi = \frac{exp\beta^\top X}{1 + exp\beta^\top X} \quad . \tag{2.11}$$

Para poder entender mejor como funcionan los Odds, se verá un ejemplo: cual es la probabilidad de que una persona tenga cáncer pulmonar dado que fuma. Las variables en este problema serían Y e X. Donde Y es la variable respuesta con posibles valores 0 para las personas sin cáncer pulmonar y 1 para los que tienen cáncer pulmonar; X como una variable regresora que tendrá como posibles respuesta 0 para las personas que no fuman frecuentemente y 1 para las que si lo hacen.

x \ y	0	1	total
0	π_{00}	π_{01}	$\pi_{0\cdot}$
1	π_{10}	π_{11}	$\pi_{1\cdot}$
total	$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	$\pi_{\cdot\cdot}$

Cuadro 2.1: Relación entre los valores esperados de la variable tener cáncer y con la variable fumar x.

No obstante la relación entre la variable de respuesta y y la covariable x descrita en esta tabla, sigue trabajando con la notación $\pi \equiv \pi_{\cdot 1}$ así, se tiene $ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x$ entonces:

$$ln(\tfrac{\pi}{1-\pi}) = \begin{cases} \beta_0 & \text{si} \quad \text{x=0} \\ \beta_0 + \beta_1 & \text{si} \quad \text{x=1} \end{cases}$$

Es ahí el interés de saber si $\beta_1 = 0$ lo que se probará a través de test estadísticos donde la hipótesis nula $H_0 : \beta_1 = 0$. Una vez comprobado que el parámetro es distinto de 0, se tendría que:

$$\frac{\pi}{1-\pi} = \frac{\text{Probabilidad de tener cáncer}}{\text{Probabilidad de no tener cáncer}} = exp(\beta + \beta_1 x_1) \tag{2.12}$$

Se tendrá que $\begin{cases} e^{\beta_0} & x = 0 \quad \text{Odds o chance de tener cáncer dado que la persona no fuma} \\ e^{\beta_0 \beta_1} & x = 1 \quad \text{Odds o chance de tener cáncer dado que la persona fuma} \end{cases}$

Por lo tanto, $\pi = P(\text{tener cáncer}) = P(C)$ y $Odds(C) = odds = \frac{\pi}{1-\pi}$

donde C= probabilidad de tener cáncer.

Algunos ejemplos de su uso serian:

Dado un $\pi = 0,5$ el Odds $= \frac{0,5}{1-0,5} = 1:1$ quiere decir que, las personas que fuman y las que no fuma tiene la misma probabilidad de tener cáncer.

Por el contrario un ejemplo irreal es si $\pi = 0,1$ el odds $= \frac{0,1}{1-0,1} = \frac{0,1}{0,9} = 1:9$ lo que quiere decir que, las personas que fuman tienen una menor probabilidad de tener cáncer pulmonar; ya que, de 10 personas hay 1 posibilidad a favor de tener cáncer y 9 de no tener cáncer.

Un Odds es interpretable pero la información que proporciona no es suficiente, es ahí donde surge la necesidad de trabajar con los odds ratio.

2.6.2 Odds ratio

Los odds ratio no son muy utilizados por los estadísticos. Éstos son utilizados principalmente en el área de la salud. El odds ratio se puede interpretar como una proporción o una razón. La mayor utilidad de este indicador está expresada en el cálculo de un cuociente entre dos odds.

El odds ratio, tal y como está construido (cociente entre dos odds), siempre será mayor o igual a 0. Su campo de variación va de 0 hasta ∞, y su interpretación se realizara en función de que el valor sea igual, menor o mayor a 1. Si toma el valor 1 significa que no hay asociación entre las variables analizadas. Si este cociente es menor a 1 el factor de riesgo podría pasar a ser un factor de protección; si el cociente es mayor a 1 el factor de riesgo estaría influenciando. Para ilustrar mejor el uso de los odds ratios se utilizará el ejemplo anterior.

$$\frac{odds_f}{odds_{f^c}} = \frac{\text{odds de las personas que fuman}}{\text{odds de las personas que no fuman}} = \frac{\dfrac{P(c/F)}{P(c^c/F)}}{\dfrac{P(c/F^c)}{P(c^c/F^c)}}$$

$$= \frac{\dfrac{\pi_{11}}{\pi_{\cdot 1}}\dfrac{\pi_{\cdot 0}}{\pi_{10}}}{\dfrac{\pi_{01}}{\pi_{\cdot 1}}\dfrac{\pi_{\cdot 0}}{\pi_{00}}} = \frac{\pi_{11}}{\pi_{\cdot 1}}\frac{\pi_{\cdot 0}}{\pi_{10}}\frac{\pi_{\cdot 1}}{\pi_{01}}\frac{\pi_{00}}{\pi_{\cdot 0}} = \frac{\pi_{11}}{\pi_{10}}\frac{\pi_{00}}{\pi_{01}} = [0, \infty[$$

Donde $P(c/F)$ es la probabilidad de tener cáncer dado que la persona fuma; $P(c^c/F)$ es la probabilidad de no tener cáncer dado que la persona fuma; $P(c/F^c)$ es la probabilidad de no tener cáncer dado que la persona no fuma y $P(c^c/F^c)$ es la probabilidad de no tener cáncer dado que la persona no fuma.

Así los posibles resultados son:

$$\text{Odds ratio} = \begin{cases} < 1 & \text{factor de protección} \\ = 1 & \text{factor nulo} \\ > 1 & \text{factor de riesgo} \end{cases}$$

En Stata al utilizar el comando *logistic* entrega automáticamente los odds ratio ajustados.

Algunos ejemplos e interpretación de los resultados serían:

Con un odds ratio irreal de 0,5 puede decir que fumar es un factor de protección ante la posibilidad de tener cáncer lo que esta lejos de ser real; es decir, el fumar ayuda a

prevenir el cáncer.

Con un odds ratio de 1 se dice que fumar no influye en tener cáncer pulmonar, es decir una persona que fuma tiene la misma probabilidad de tener cáncer que una persona que no fuma.

Con un odds ratio de 10 se puede decir que una persona que fuma tiene un riesgo 10 veces mayor a tener cáncer pulmonar respecto a las personas que no fuman.

2.7 Tamaños muestrales

Cuando se desea realizar un estudio, en el cual los análisis se realizan con regresión logística se debe tener en cuenta de que es prácticamente imposible estudiar a toda la población, además de innecesario, ya que, se podrían obtener los mismos resultados con una muestra, pero ¿cuál es el tamaño de muestra necesario para un correcto análisis utilizando regresión logística? Al revisar la literatura como estimar el tamaño de muestra Ortega y Cayuela [2002] y Molinero [2001] indica que este debiese calcularse con la siguiente fórmula "10*(k+1)" donde k es el número de variables regresoras, lo que no parece ser del todo correcto; ya que, no toma en cuenta los niveles de confianza, tamaño total de la población y la variabilidad de los datos que es imprescindible para su cálculo a través de fórmula. Es por ello que en este trabajo de titulación se propone que el tamaño de la muestra se calcule vía simulación en el software *Stata*.

Antes de realizar la simulación se comenzará por plantear un problema, el que ayudará a entender mejor la dinámica de la simulación. Por ejemplo se desea estudiar si fumar frecuentemente es un factor de riesgo ante la posibilidad de tener cáncer pulmonar, donde las variables serán y, x_1, x_2 y x_3 donde: la variable respuesta y tendrá como posibles resultados: 0 para indicar que el paciente no tiene cáncer pulmonar; 1 para decir que tiene cáncer pulmonar. La variable regresora x_1 donde la posible respuesta será: 0 no fuma frecuentemente; 1 fuma frecuentemente, estas variables son del tipo dicotómica. La variable regresora x_2 será la edad, la que corresponde a una variable del tipo discreta y la variable x_3 niveles de arsénico en la sangre que corresponde a una variable continua.

Para realizar los estudios de tamaños muestrales se realizarán estudios de dos tipos, con la o las variables regresoras: binaria; binaria y discreta conjuntamente; binaria y continua conjuntamente .

2.7.1 Tamaño muestral con una variable regresora binaria

Es bien sabido en estudios de tamaños muestrales que el tamaño de muestra más grande, se da cuando la variabilidad es alta, es decir la varianza es grande. En regresión logística la varianza es máxima cuando $\pi = 0,5$ es por ello que en la presente simulación se buscara que $\pi = 0,5$ para así tener el tamaño muestral máximo, dejando cubierta poblaciones con menor variabilidad.

Para comenzar la simulación se supondrá una población, donde los datos serán los siguientes: De los presentes datos se calculan los parámetros β_0 y β_1, para posteriormente

x \ y	0	1	Total
0	2208	261	2469
1	2212	5319	5731
Total	4420	5580	10000

Cuadro 2.2: 10.000 observaciones generadas vía simulación en el software Stata 12 por la variable x según la variable y.

compararlo con el de las muestras. Después se comenzará a tomar muestras de distintos tamaños, para ver cuál es la más parecido a las de la población. En las *mas* se calcularán los $\hat{\beta}$, esto para comparar con los parámetros poblacionales, si algunos de los tamaños muestrales tuviese estimaciones en promedio parecidas al de la población, sería el tamaño adecuado. Para saber si realmente se debe utilizar ese tamaño muestral, se trabajará con un error estándar y nivel de confianza como se verá más adelante, si el error es pequeño y el nivel de confianza es alto será el tamaño adecuado, sino busca otros tamaño de muestra y se repite el proceso anterior sucesivamente hasta encontrar el tamaño de muestra adecuado. Ese es el proceso que se realizará en la simulación.

Al aplicar el modelo logístico

$$ln\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 \quad , \tag{2.13}$$

que es equivalente a *logit* y x_1 en *Stata* se obtienen los siguientes resultados:

variable	Coeficiente	Error estándar	Valor-p	Intervalo de	Confianza del 95 %
x1	3,0127	0,0701	0,0000	2,8751	3,1502
Constante	-2,1353	0,0654	0,0000	-2,2636	-2,0070

Cuadro 2.3: Modelo logístico de la ecuación 2.13 realizado en Stata 12 por variable según valores

Ya contando con el valor de los coeficientes de $\beta_1 = 3,0127$ y $\beta_0 = -2,1353$ los que corresponden al parámetro de x_1 y la constante, respectivamente, se procederá a obtener

distintos tamaños de muestra.

Se obtienen mil muestras de cada tamaño muestral, en este caso se obtuvieron mil muestras de tamaño quinientos, otras mil de tamaño mil así de quinientos en quinientos hasta llegar a diez mil. Una vez ya obtenidas las muestras el siguiente paso es calcular las estimaciones de los parámetros, los que se calculan en cada muestra, es decir, los parámetros son calculados 1000 veces para cada tamaño muestral. Realizado ésto, se calcula la media y la desviación estándar de los parámetros por cada tamaño muestral, en consecuencia, para el tamaño 500 de las 1000 muestras obtenidas se tendrá el promedio de los $\hat{\beta}_0$ y $\hat{\beta}_1$ con su desviación estándar; lo que se realiza para tener una visión general de los datos simulados. Esto mismo se realiza para cada tamaño muestral. Posteriormente para estimar el tamaño muestral se analizará la variabilidad, error estándar y nivel de confianza para distintos tamaños muestrales.

En esta simulación los resultados que se obtuvieron son:

n	Promedio.(b0)	Des. estandar.(b0)	Promedio.(b1)	Des. estandar.(b1)
500	-2,1801	0,3063	3,0620	0,3295
1000	-2,1594	0,2044	3,0365	0,2175
1500	-2,1384	0,1554	3,0134	0,1676
2000	-2,1418	0,1342	3,0186	0,1457
2500	-2,1392	0,1160	3,0167	0,1248
3000	-2,1398	0,1006	3,0175	0,1083
3500	-2,1347	0,0896	3,0123	0,0955
4000	-2,1399	0,0831	3,0172	0,0882
4500	-2,1398	0,0718	3,0176	0,0780
5000	-2,1382	0,0649	3,0171	0,0686
5500	-2,1397	0,0593	3,0171	0,0630
6000	-2,1346	0,0547	3,0120	0,0584
6500	-2,1353	0,0481	3,0126	0,0516
7000	-2,1364	0,0429	3,0144	0,0463
7500	-2,1366	0,0375	3,0141	0,0403
8000	-2,1352	0,0315	3,0127	0,0336
8500	-2,1357	0,0266	3,0129	0,0283
9000	-2,1352	0,0221	3,0129	0,0232
9500	-2,1352	0,0148	3,0125	0,0160
10000	-2,1353	0	3,0127	0

Cuadro 2.4: Resumen de los resultado de las simulaciones, donde se muestra el promedio y la desviación estandar de cada tamaño muestral

La variabilidad de los datos no se puede apreciar solo al ver estos resultados, por lo tanto para tener una mejor apreciación de los datos se realizaran histogramas y un gráfico de dispersión, los que se muestran a continuación.

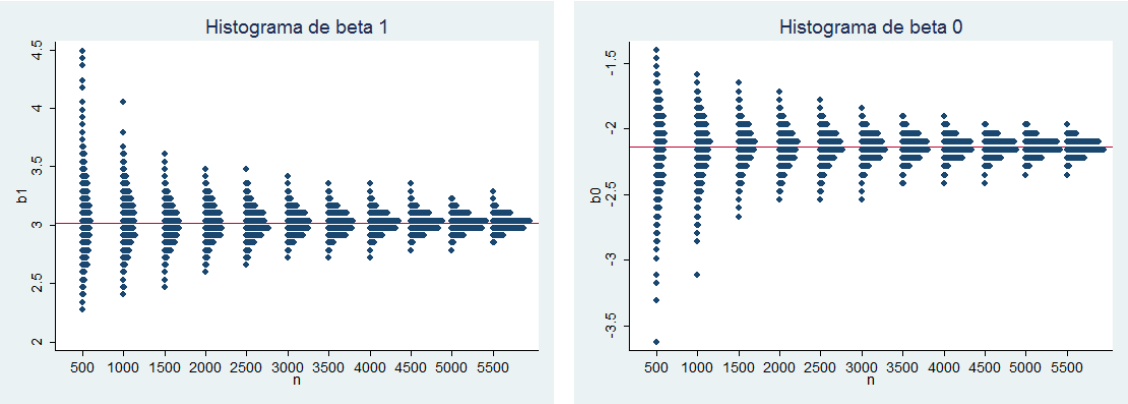

Figura 2.1: Histograma de las 1.000 estimaciones de β para los diferentes tamaños muestrales n

Los gráficos de la Figura 2.1 no se encuentran en la misma escala ya que estos miden cosas distintas. Estos gráficos son estimaciones de β_1 y β_0, donde en cada tamaño muestral existen 1000 estimaciones. En estos gráficos se puede notar, que a medida que el tamaño de la muestra aumenta los datos se tienden a concentrar alrededor del β_1 y β_0 poblacional, el que corresponde a la línea roja que cruza el gráfico, por lo tanto sí se desea un mayor nivel de confianza lo lógico es aumentar el tamaño muestral.

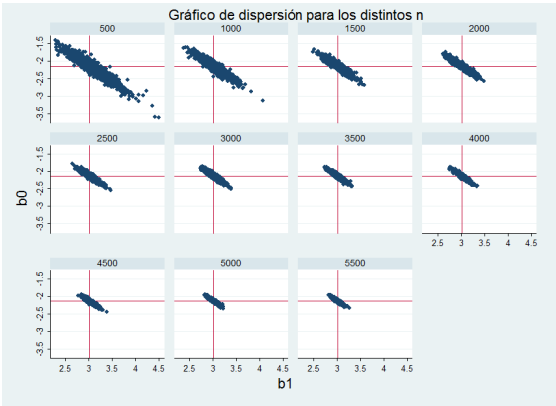

Figura 2.2: Gráfico de dispersión entre β_1 y β_0 para los diferentes tamaño muestral (n)

En la Figura 2.2 se muestra la dispersión de las estimaciones de β_1 y β_0, para distintos tamaños de muestra n, se puede ver como a medida que el tamaño muestral aumenta las observaciones se aglomeran alrededor de los valores poblaciones que corresponden a las líneas rojas en el gráfico.

Ahora, para encontrar el tamaño de muestra adecuado, se realizarán cálculos para los diferentes n con variados errores de estimación. Estos cálculos se realizarán considerando

la siguiente expresión para los esquemas de simulación

$$P(|\beta_1 - \hat{\beta}_{1,n}| < e\beta_1) \geq 1 - \alpha \quad , \tag{2.14}$$

donde β_1 se supone conocido (es el parámetro poblacional del modelo logístico dado la ecuación (2.13)) y, a partir de éste se genera una población de 10.000 unidades con una covariable binaria x_1; $\hat{\beta}_{1,n}$ es el parámetro calculado para x_1 en la muestra de tamaño n, se seleccionan 1000 *mas* para cada tamaño muestral; el e es el error estándar con el que se permitirá trabajar y $1 - \alpha$ el nivel de confianza. Con esta fórmula se determina la proporción empírica de los estimadores que no están alejados del parámetro más allá del error de estimación (e). Esta proporción se considera como, el nivel de significancia empírico.

n \ e	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
500	8,0 %	15,4 %	22,1 %	30,3 %	36,7 %	43,2 %	49,1 %	54,6 %	59,2 %	63,6 %
1000	11,0 %	21,2 %	30,9 %	42,2 %	52,0 %	61,1 %	68,2 %	73,5 %	80,2 %	84,3 %
1500	13,2 %	28,0 %	42,0 %	54,1 %	64,5 %	72,6 %	80,2 %	85,6 %	90,1 %	92,7 %
2000	15,7 %	30,7 %	46,2 %	59,5 %	69,6 %	77,5 %	84,8 %	89,3 %	93,3 %	96,6 %
2500	18,9 %	37,5 %	54,1 %	68,9 %	78,6 %	86,0 %	90,8 %	93,9 %	96,6 %	97,8 %
3000	20,0 %	41,0 %	58,3 %	72,4 %	84,2 %	91,3 %	95,2 %	97,3 %	99,0 %	99,3 %
3500	22,9 %	45,9 %	65,2 %	78,7 %	89,4 %	95,0 %	97,5 %	98,6 %	99,3 %	99,8 %
4000	28,6 %	53,2 %	71,5 %	82,9 %	91,1 %	95,6 %	97,8 %	98,9 %	99,7 %	99,8 %
4500	30,5 %	56,3 %	75,5 %	88,6 %	95,1 %	98,3 %	99,0 %	99,4 %	99,8 %	99,9 %
5000	32,7 %	62,1 %	81,8 %	92,5 %	96,7 %	98,9 %	100 %	100 %	100 %	100 %
5500	37,8 %	64,7 %	83,7 %	94,7 %	98,9 %	99,5 %	100 %	100 %	100 %	100 %
6000	38,0 %	68,6 %	88,2 %	96,6 %	98,9 %	100 %	100 %	100 %	100 %	100 %
6500	43,4 %	75,2 %	92,2 %	97,6 %	99,7 %	100 %	100 %	100 %	100 %	100 %
7000	49,5 %	80,8 %	94,8 %	99,3 %	100 %	100 %	100 %	100 %	100 %	100 %
7500	54,0 %	86,4 %	97,3 %	99,5 %	100 %	100 %	100 %	100 %	100 %	100 %
8000	64,8 %	91,9 %	99,3 %	99,9 %	100 %	100 %	100 %	100 %	100 %	100 %
8500	69,8 %	96,7 %	99,9 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
9000	80,9 %	99,1 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
9500	94,5 %	99,9 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
10000	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Cuadro 2.5: Nivel de confianza empírico basado en simulaciones de 1000 muestras realizado con Stata 12 por error de estimación (e) según tamaño muestral (n)

El procedimiento en el software es el siguiente: cuando se cumple la condición de la ecuación (2.14) se da el valor de 1 y cuando no se cumple será 0; por lo cual, al calcular el promedio se tendrá la proporción de muestras que cumplen con la condición. Se espera que este nivel de significancia empírico este próximo al nivel de confianza $(1 - \alpha)$. Este valor será interpretable al ser expresado en porcentaje.

Los valores que se muestran en el Cuadro 2.5 son para β_1; ya que, es por lo general el parámetro más relevante en el estudio.

En el encabezado de las columnas del Cuadro 2.5 se muestran los errores estándares de $\hat{\beta}_1$ a través de simulaciones, estos porcentajes corresponden a los errores típicos utilizados en estadística para realizar análisis de muestras (por ejemplo cálculo de media, parámetros y otros); en el encabezado de las filas, se representan los distintos tamaños muestrales considerados en las simulaciones (que van de n=500(500)10.000). Con la intersección de éstas, se obtiene como resultado un nivel de confianza, en particular si se desea trabajar con un error estándar del 5 %, se debe buscar el tamaño muestral donde el nivel de confianza sea lo más próximo al 95 %, en este caso el tamaño de muestra necesario con el que se cumplen estas condiciones es de 4500. Lo que quiere decir que de 1000 muestras que se tomen de tamaño 4500, 951 estarán dentro del intervalo conformado por el parámetro más menos el 5 %. Adicionalmente, si se desea un tamaño muestral más depurado, se debe ver la tabla que va de 500 en 500 y el error estándar del 5 % (u otro a elección), si se encuentra un nivel de confianza de 94 % para un tamaño (ejemplo 4000) y 96 % para otro (4500), se puede realizar un proceso a parte de simulación, pero ahora de 10 en 10 (n=4000(10)4500) y donde se vea que se obtiene un nivel de confianza del 95 % estará el tamaño muestral exacto.

Para seleccionar el error estándar se debe tener en cuenta que la estimación del parámetro será el valor más menos el error estándar. Por ejemplo, se desea saber el porcentaje de votos que tendrá el candidato presidencial Marco Enriquez Ominami, para ésto se analiza a una parte de la población (muestra) en el que se decide trabajar con un error estándar del 3 %. Sí el resultado estimado fuese 15 % de los votos, se dice que tendrá el 15 % de los votos de la población más menos el 3 %; es decir, el resultado esperado estará entre 12 % y 18 % lo que es aceptable. Pero si se decide trabajar con un error estándar del 10 % los resultados estarían entre 5 % y 25 %, en este caso el rango es demasiado amplio y deja mucho en incertidumbre, por lo tanto ¿con qué error estándar trabajaría? La respuesta a esta pregunta varía dependiendo de la precisión que se desee para el estudio y el dinero con el que se dispone para realizarlo.

En esta simulación se puede ver que el tamaño muestral más indicado para realizar un correcto análisis de regresión logística es de 4500, ya que se trabajaría con un error estándar de estimación del 5 % y un nivel de confianza del 95,1 %.

2.7.2 Tamaño muestral para dos variables regresoras una binaria y otra discreta

Para saber cuál es el tamaño muestral necesario, cuando las variables regresoras son binaria y discreta, se realizará el mismo proceso de antes, con la diferencia de que se agregará una variable discreta x_2 dentro de las variables regresoras.

Al aplicar el modelo logístico

$$ln\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad , \tag{2.15}$$

que es equivalente a *logit* y x_1 x_2 en *Stata* se obtienen los siguientes resultados:

Variable	Coeficiente	Error estándar	Valor-p	Intervalo de	confianza del 95 %
x1	3,003169	0,0802977	0,000	2,845789	3,16055
x2	0,219887	0,0054279	0,000	0,209248	0,23052
cons	-8,582272	0,1855886	0,000	-8,946019	-8,21852

Cuadro 2.6: Salida del modelo ajustado en Stata para la población de 10.000 individuos con las variables x_1 y x_2

Ya contando con el valor de los coeficientes de $\beta_1 = 3,003169$, $\beta_2 = 0,219887$ y $\beta_0 = -8,582272$ los que corresponden al parámetro de x_1, x_2 y la constante, respectivamente, se procederá a obtener los distintos tamaños de muestra. A continuación, se muestra el promedio y desviación estándar para los diferentes tamaños muestrales (n).

Para tener una mejor apreciación de la dispersión de los datos se mostraran histogramas para cada tamaño muestral en la siguiente figura.

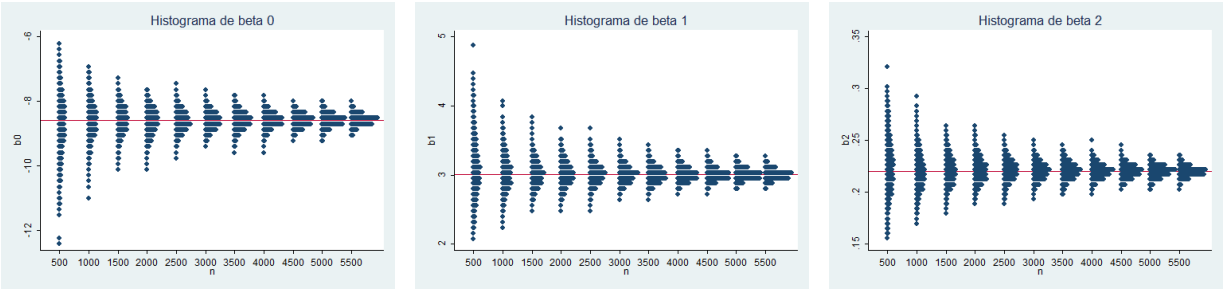

Figura 2.3: Histograma de las 1.000 estimaciones de β para los diferentes tamaños muestrales (n)

Los gráficos de la Figura 2.3 no se encuentran todos en la misma escala, la razón es que cada uno mide cosas distintas, éstos gráficos muestran la variabilidad de las 1.000

n \ est.	Prom.(b0)	Des. est.(b0)	Prom.(b2)	Des. est.(2)	Prom.(b1)	Des. est.(b1)
500	-8.7764	.8917	.2246	.0254	3.0684	.3976
1000	-8.6243	.6111	.2205	.0178	3.0276	.2649
1500	-8.6184	.4752	.2208	.0139	3.0114	.1996
2000	-8.6203	.4020	.2208	.0117	3.0153	.1727
2500	-8.6187	.3550	.2208	.0103	3.0120	.1529
3000	-8.6069	.2943	.2205	.0087	3.0087	.1289
3500	-8.5999	.2708	.2205	.0079	3.0028	.1167
4000	-8.5960	.2411	.2201	.0069	3.0082	.1048
4500	-8.5950	.2175	.2201	.0063	3.0083	.0936
5000	-8.5904	.1963	.2200	.0057	3.0097	.0806
5500	-8.5872	.1751	.2198	.0051	3.0076	.0755
6000	-8.5808	.1596	.2198	.0046	3.0023	.0694
6500	-8.5890	.1439	.2201	.0043	3.0033	.0610
7000	-8.5792	.1285	.2197	.0037	3.0039	.0548
7500	-8.5836	.1146	.2198	.0033	3.0045	.0485

Cuadro 2.7: Promedio y desviación estándar de las estimaciones de los parámetros para 1000 muestras en cada tamaño muestral (n)

n \ est.	Prom.(b0)	Des. est.(b0)	Prom.(b2)	Des. est.(2)	Prom.(b1)	Des. est.(b1)
8000	-8.5802	.0998	.2198	.0029	3.0029	.0401
8500	-8.5838	.0869	.2199	.0025	3.0039	.0348
9000	-8.5809	.0657	.2198	.0019	3.0038	.0269
9500	-8.5836	.0467	.2199	.0013	3.0031	.0192
10000	-8.5822	0	.2198	0	3.0031	0

Cuadro 2.8: Promedio y desviación estándar de las estimaciones de los parámetros para 1000 muestras en cada tamaño muestral (n)

estimaciones de los parámetros para cada tamaños muestrales, estas son representadas para β_0, β_1 y β_2, respectivamente, donde se puede notar que a medida que aumenta el tamaño muestral disminuye la variabilidad.

n \ e	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
500	05,9 %	12,5 %	19,1 %	24,6 %	31,0 %	37,8 %	43,1 %	48,0 %	52,9 %	57,5 %
1000	07,3 %	15,1 %	25,8 %	34,7 %	44,5 %	51,9 %	57,5 %	64,7 %	70,3 %	75,0 %
1500	12,7 %	24,1 %	36,0 %	46,4 %	55,7 %	63,6 %	70,9 %	77,7 %	82,5 %	87,3 %
2000	13,1 %	27,6 %	40,2 %	51,4 %	60,5 %	69,7 %	77,7 %	81,8 %	88,2 %	92,1 %
2500	17,2 %	31,4 %	46,2 %	59,4 %	69,6 %	77,3 %	84,0 %	88,7 %	92,2 %	94,8 %
3000	19,2 %	36,7 %	51,1 %	63,1 %	75,4 %	83,7 %	90,3 %	94,4 %	96,4 %	97,8 %
3500	16,4 %	34,3 %	53,7 %	69,0 %	79,0 %	87,9 %	93,4 %	96,6 %	98,8 %	99,2 %

n \ e	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
4000	21,4 %	42,8 %	61,6 %	75,1 %	84,4 %	92,0 %	95,6 %	97,6 %	98,9 %	99,3 %
4500	25,3 %	48,7 %	67,5 %	81,4 %	89,0 %	94,2 %	96,8 %	98,8 %	99,4 %	99,6 %
5000	28,9 %	52,9 %	74,5 %	87,0 %	93,9 %	97,2 %	99,1 %	99,8 %	99,9 %	100 %
5500	32,0 %	58,3 %	76,1 %	88,2 %	95,3 %	98,3 %	99,5 %	99,9 %	99,9 %	100 %
6000	32,2 %	60,6 %	80,2 %	90,5 %	97,0 %	99,3 %	99,8 %	99,9 %	100 %	100 %
6500	36,6 %	67,8 %	84,8 %	94,8 %	98,8 %	99,5 %	100 %	100 %	100 %	100 %
7000	39,8 %	72,1 %	88,8 %	97,1 %	99,4 %	99,9 %	100 %	100 %	100 %	100 %
7500	45,1 %	77,5 %	92,8 %	98,9 %	99,7 %	100 %	100 %	100 %	100 %	100 %
8000	52,6 %	86,3 %	96,9 %	99,7 %	100 %	100 %	100 %	100 %	100 %	100 %
8500	58,8 %	91,4 %	98,9 %	99,9 %	100 %	100 %	100 %	100 %	100 %	100 %
9000	69,3 %	96,9 %	99,8 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
9500	83,6 %	99,7 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
10000	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Cuadro 2.9: Nivel de confianza empírico para β_1 basado en simulaciones de 1000 muestras realizado con Stata 12 por error de estimación (e) según tamaño muestral (n)

n \ e	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
500	05,4 %	12,4 %	18,6 %	27,9 %	34,9 %	39,9 %	46,1 %	52,2 %	57,3 %	63,0 %
1000	09,7 %	20,1 %	28,5 %	39,0 %	47,7 %	55,0 %	61,3 %	68,6 %	73,2 %	78,1 %
1500	11,0 %	22,7 %	33,9 %	44,2 %	54,4 %	65,7 %	73,8 %	79,3 %	85,3 %	88,8 %
2000	13,0 %	25,4 %	40,0 %	53,2 %	64,3 %	73,6 %	80,2 %	86,5 %	92,0 %	94,1 %
2500	15,5 %	31,3 %	46,8 %	59,5 %	71,5 %	79,5 %	86,7 %	91,4 %	94,1 %	97,0 %
3000	19,3 %	37,6 %	53,5 %	69,1 %	78,5 %	86,8 %	92,1 %	96,1 %	97,9 %	98,4 %
3500	22,3 %	41,2 %	60,1 %	73,5 %	83,9 %	90,1 %	94,2 %	96,6 %	98,4 %	99,2 %
4000	25,6 %	47,9 %	66,4 %	80,2 %	89,0 %	93,4 %	97,3 %	98,9 %	99,5 %	99,9 %
4500	25,7 %	50,0 %	68,8 %	83,5 %	92,0 %	96,6 %	98,8 %	99,5 %	99,8 %	99,9 %
5000	30,9 %	54,2 %	75,0 %	88,2 %	95,1 %	97,5 %	99,3 %	99,7 %	100 %	100 %
5500	34,5 %	63,3 %	81,7 %	91,3 %	96,3 %	98,5 %	99,4 %	99,9 %	100 %	100 %
6000	35,7 %	64,2 %	83,5 %	94,4 %	98,8 %	99,9 %	100 %	100 %	100 %	100 %
6500	37,6 %	69,3 %	86,1 %	96,1 %	99,1 %	99,8 %	100 %	100 %	100 %	100 %
7000	43,3 %	75,7 %	92,0 %	98,0 %	99,8 %	100 %	100 %	100 %	100 %	100 %
7500	48,1 %	82,0 %	95,5 %	99,2 %	99,9 %	100 %	100 %	100 %	100 %	100 %
8000	54,5 %	87,9 %	97,0 %	99,1 %	100 %	100 %	100 %	100 %	100 %	100 %
8500	61,1 %	90,7 %	99,0 %	99,9 %	100 %	100 %	100 %	100 %	100 %	100 %
9000	74,4 %	98,0 %	99,9 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
9500	90,3 %	99,8 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
10000	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Cuadro 2.10: Nivel de confianza empírico para β_2 basado en simulaciones de 1000 muestras realizado con Stata 12 por error de estimación (e) según tamaño muestral (n)

Para determinar el tamaño de muestra necesario con el que se realizaría un correcto análisis de regresión logística cuando la variable regresora es discreta y binaria; se utiliza la ecuación (2.14) de la sección 2.7.2 y se realiza el mismo proceso tanto para β_1 y β_2.

Al realizar el proceso en el software los resultados que se obtuvieron se muestran en los cuadros 2.9 y 2.10.

Como síntesis final de las simulación realizada para variables del tipo binaria y discreta se puede decir que el tamaño muestral de una población de 10.000 observaciones, debe ser de 5000 ya que se trabajaría con un error estándar del 5 % y un nivel de confianza del 93,9 %, para llegar a esta conclusión se observaron las dos tablas y se utilizó el nivel de confianza más pequeño para que así fuese representativo para los dos tipos de variables.

2.7.3 Tamaño muestral para variable regresora del tipo binaria y continuas

Cuando se desea estudiar cual es el tamaño de muestra necesario para realizar un correcto análisis de regresión logística donde la variable regresora es continua; el cálculo del tamaño de la muestra es mucho más complejo, por lo que se escapa de los objetivos de este trabajo de titulación. Sí se desea más información sobre este tema puede revisar el libro Long & Freese [2006] el que indica que se debe observar la curva de regresión logística para diferentes intervalos donde la pendiente varia (por ejemplo donde la curva tiene menor pendiente, pendiente media y una pendiente mayor), en el caso de que la pendiente no se tomara en cuenta y se realizase la simulación como en las sub-secciones anteriores (tamaño muestral con una variable regresora binaria y tamaño muestral para dos variables regresoras una binaria y otra discreta), los tamaños muestrales se incrementarían, dejando como única alternativa trabajar con el total de la población. La razón por la que ocurre esto es; como los datos son continuos las posibles muestras son infinitas y difiere mucho una de otra, por lo cual también el cálculo del parámetro de regresión logístico.

Capítulo 3

Regresión logística en muestreo complejo

3.1 Introducción

Las muestras complejas son de gran utilidad cuando se desea analizar una población y se cuenta con información previa de esta; como por ejemplo la cantidad de regiones, provincias y comunas con la que cuenta un país; la cantidad de manzanas censales que tiene cada comuna; el número de viviendas que tiene cada manzana censal y otras variables propias de un marco muestral, las que se pueden saber antes de seleccionar una muestra.

Stata 12 cuenta con un conjunto de comandos que permiten realizar análisis de muestras complejas, los que permiten incorporar tanto los factores de expansión como características del plan de muestreo complejo en el análisis y así realizar las estimaciones de los parámetros y de sus respectivos errores estándares, cumpliendo todos los supuestos que conlleva el análisis de muestras complejas.

El objetivo de este capítulo es mostrar la diferencia que existe en los resultados al incorporar el plan de muestreo que se utilizó en la captura de los datos, antes de utilizar el análisis por defecto que realiza el software Stata. Además se mostrará un método para estimar el tamaño de muestra necesario para un correcto análisis de regresión logística cuando el muestreo a utilizar es complejo.

3.2 Estimación de parámetros y errores estándar en muestras complejas

Cuando los datos son capturados con un muestreo complejo, ya no es posible una aplicación directa del método de *vm*, por dos razones. En primer lugar, las probabilidades de selección para el individuo $i = 1, ..., n$ observaciones de la muestra ya no son (en general) iguales. Por lo tanto se requieren ponderaciones muestrales en la estimación de parámetros de un modelo de regresión logístico. En segundo lugar, la estratificación y la agrupación de las observaciones de muestras complejas viola el supuesto de independencia en las observaciones, que es crucial para el enfoque de *vm*.

Los autores mencionados a continuación son citados por Heeringa West & Berglund [2010]. Dos enfoques se han desarrollado para estimar los parámetros y errores estándar del modelo de regresión logística en un marco muestral complejo. Grizzle, Starmer Koch (1969) formularon por primera vez un enfoque basado en los mínimos cuadrados ponderados (*mcp*). El método de estimación de *mcp* fue programado originalmente para la regresión logística en el paquete de software GENCAT Landis (1976) y todavía sigue estando disponible como una opción en programas como SAS, PROC y CATMOD. Más tarde, Binder (1981) Binder (1983) presentó un segundo marco general para la regresión logística de montaje y otros modelos lineales generalizados para datos de encuestas complejas. Posteriormente se propuso el método de seudo-máxima verosimilitud (*smv*) como una técnica para la estimación de los parámetros del modelo. El enfoque de *smv* para la estimación de parámetros se combinó con un estimador linealizado de la matriz de varianza y covarianza de las estimaciones de los parámetros, teniendo características del diseño muestral en cuenta.

Un mayor desarrollo y evaluación del enfoque de *smv* se presenta en Roberts, Rao Kumar (1897) y Skinner, Holt Smith (1989). El enfoque de la *smv* es el método estándar para el modelo de regresión logística en todos los principales sistemas de software que apoyan el análisis de datos de encuestas complejas.

La estimación de los parámetros con el método de *smv* en poblaciones finitas se obtiene mediante la maximización de la siguiente estimación de la probabilidad, que es una función ponderada de los datos de la muestra observados y los valores de $\pi(x_i)$:

$$PL(\beta|x) = \prod_{i=1}^{N} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i w_i} \quad , \tag{3.1}$$

donde $\pi(x_i) = exp(x_i\beta)/[1 + exp(x_i\beta)]$
y w_i Son los pesos muestrales

Al igual que el procedimiento de *vm*, esta función puede ser maximizada usando el método de Newton-Raphson, o algoritmos iterativos relacionados, como se mostro en el capítulo anterior. Nuestro interés es la estimación a través de software así que solo se

verá como realizarlo en *Stata*, más adelante se realizará un ejemplo para mostrar cómo funciona.

El siguiente obstáculo en el análisis de modelos de regresión logística para datos de encuestas complejas es estimar las varianzas y covarianzas de muestreo en las estimaciones de los parámetros. Carpeta (1983) propone una solución a este problema que aplica una versión multivariante de series de Taylor linealizada (*stl*) la linealización es un procedimiento que permite aproximar un modelo no lineal, por otro que si lo es y que cumple las propiedades de los sistemas lineales, en particular el principio de superposición. El resultado es un estimador de varianza de tipo sándwich de la forma

$$var(\hat{\beta}) = (J^{-1})var[S(\hat{\beta})](J^{-1}) \tag{3.2}$$

donde J es la matriz de segunda derivadas con respecto a β_j de la seudo-log de probabilidad para los datos (derivados mediante la aplicación de la función logaritmo natural de la probabilidad se define en (3.1)) y la $var(\hat{\beta})$ es la matriz de varianza-covarianza de la muestra suma de los valores ponderados función de puntuación para las observaciones individuales que se utilizan para ajustar el modelo.

3.3 Análisis de regresión logística bajo planes de muestreo complejos

El objetivo de este trabajo de titulación es el estudio de la regresión logística en muestras complejas, para poder realizalo se analizarán algunas variables de la encuesta CASEN, ya que, la captura de los datos de está fue realizado con un muestreo complejo. Por lo cual, cuenta con los factores de expansión por etapa; lo que permite incluirlos para el análisis de regresión logística como se verá a continuación.

Antes de aplicar un modelo de regresión logística se debe ver qué variables serán incluidas para el análisis, por lo cual, se revisarán todas las variables que conforma la encuesta CASEN, luego se escogerán las variables que puedan ser transformadas a 0 y 1 para poder realizar el análisis.

Después de haber revisado las variables, se decidió estudiar si el saber leer y escribir influya en que una persona trabaje o no trabaje. Para realizar este análisis se dejará la variable trabajo como 0 si la persona no ha trabajado durante la última semana y como 1 si la persona ha trabajado; después como 0 si la persona no sabe leer y escribir y como 1 si sabe leer y escribir. Este análisis será aplicado para las personas que estén en edad de trabajar es decir hombres entre 18 y 65 años y mujeres entre 18 y 60 años.

3.3.1 Análisis logístico incorporando el plan de muestreo

Para el análisis de muestras que han sido obtenidas a partir de un plan de muestreo complejo su usarán los comandos *svy* en el software Stata los que permiten incorporar el plan de muestreo para el análisis. Además permite trabajar con sub-poblaciones, como es en este caso.

De una muestra de 246.925 se tomaron en cuenta 172.893 observaciones para el análisis, ya que estos cumplen con los requisitos de edad. Para efectos del análisis no se eliminó ningún dato, solo se tomó como subpoblación con el comando *subpop* dentro de los *svy* en Stata, tomando en cuenta 602 estratos y haciendo que la muestra sea representativa a una población de 11.669.778 personas, con lo que se obtuvieron los siguientes resultados.

variable	Coeficiente	Error estándar	Valor-p	Intervalo de confianza del 95 %
sabe leer y escribir	0 ,8843	0,0440	0,000	0,7979 ; 0,9706
constante	-0,7758	0,0430	0,000	-0,8601 ;-0,6915

Cuadro 3.1: Salida en Stata 12 de un modelo logístico incorporando plan de muestreo, con la variable respuesta la persona trabajo la última semana si (1) no (0) y la variable regresora sabe leer y escribir si (1) no (0)

Posterior al análisis se calculó el efecto de diseño de la muestra conocido como DEFF el que se obtiene en Stata con el comando $estat\ effects$ y nos entrega los siguientes resultados.

variable	Coeficiente	Error estándar	DEFF
sabe leer y escribir	0,8843	0,0440	1,8394
constante	-0,7758	0,0430	1,7938

Cuadro 3.2: Efecto de diseño del modelo logístico mostrado en el cuadro 3.1

Lo que nos indica que la muestra debe ser 1,8349 veces más grande para que se pueda utilizar este plan de muestreo, es decir si nuestra muestra es de 172.893 debiese multiplicarse por 1,8349 y nos daría 317.241 que corresponde a la cantidad de individuos que debiesen encuestarse para que la muestra sea representativa a nivel país.

Si se observa el cuadro 3.1 donde se aplicó el modelo *logit* se obtiene un coeficiente de 0,8843 que indica que si la persona sabe leer y escribir el logit estimado aumenta en promedio 0.8843 unidades lo que sugiere una relación positiva en el estar trabajando y saber leer y escribir. Una interpretación con más sentido sería con el antilogaritmo o odds ratio como se verá en el siguiente cuadro.

variable	Odds ratio	Error estándar	Valor-p	Intervalo de confianza del 95 %
sabe leer y escribir	2,4213	0,1066	0,000	2,2210 ; 2,6396
constante	0,4603	0,0197	0,000	0,4231 ; 0,5007

Cuadro 3.3: Salida en Stata 12 de un modelo logístico incorporando plan de muestreo, para sus odds ratio con la variable respuesta la persona trabajo la última semana si (1) no (0) y la variable regresora sabe leer y escribir si (1) no (0)

En este caso el odds ratio calculado para la variable saber leer y escribir es de 2,4213 que indica que si una persona sabe leer y escribir tiene una probabilidad 2 veces superior de encontrar trabajo a una persona que no sabe leer y escribir.

3.3.2 Análisis logístico sin incorporar el plan de muestreo

Si no se incorporase el plan de muestreo utilizado para la captura de los datos y se analizase como si fuese una muestra aleatoria simple los resultados que se obtendrían serían los siguientes:

variable	Coeficiente	Error estándar	Valor-p	Intervalo de confianza del 95 %
sabe leer y escribir	1,0828	0,0219	0,000	1,0398 ; 1,1259
constante	-1,2221	0,0214	0,000	-1,2641 ;-1,1800

Cuadro 3.4: Salida en Stata 12 de un modelo logístico sin incorporar plan de muestre, para la variable respuesta la persona trabajo la última semana si (1) no (0) y la variable regresora sabe leer y escribir si (1) no (0)

Y al calcular los odds ratios se obtendrias:

variable	Coeficiente	Error estándar	Valor-p	Intervalo de confianza del 95 %
sabe leer y escribir	2,9532	0,0648	0,000	2,8287 ; 3,0831
constante	0,2946	0,0063	0,000	0,2824 ; 0,3072

Cuadro 3.5: Salida en Stata 12 de un modelo logístico sin incorporar plan de muestreo, para los odds ratio con la variable respuesta la persona trabajo la última semana si (1) no (0) y la variable regresora sabe leer y escribir si (1) no (0)

En este caso el odds ratio calculado para la variable saber leer y escribir es de 2.9532 que indica que si una persona sabe leer y escribir tiene una probabilidad 3 veces superior de encontrar trabajo a una persona que no sabe leer y escribir.

3.3.3 Comparación de coeficientes

Suponga que desea trabajar con un nivel de confianza del 95 % y un error estándar del 5 % y se analiza la muestra como si esta fuese una muestra aleatoria simple, es decir, sin tomar en cueta el plan de muestreo. Los resultados de la estimación de los parámetros serían de 1,0828 con un intervalo de confianza del 95 % el valor estimado estaría entre 1,0398 y 1,1259. En cambio si se incorpora el plan de muestreo en la estimación de parámetros, el resultado de este sería de 0,8843 con un intervalo de confianza del 95 % este valor estaría entre 0 ,7979 y 0,9706. Por lo tanto los resultados de las estimaciones de los parámetros al no incluir el plan de muestreo y al incluirlo no se asemejan y difieren mucho más que un 5 % por lo cual se debe tener muy claro cómo fueron capturado los datos antes de realizar el análisis, ya que, no tomar en cuenta el plan de muestreo puede llevar a grandes errores. Este análisis se realizó con los coeficientes pero ocurre lo mismo en el caso de los odds ratios, como se vera a continuación.

Muestreo	Coeficiente	Intervalo de confianza del 95 %
mas	1,0828	1,0398 ; 1,1259
mc	0,8843	0,7979 ; 0,9706

Cuadro 3.6: Comparación de coeficientes sin incorporar el plan de muestreo e incorporándolo

3.3.4 Comparación de odds ratio

Para los odds ratio, al igual que el caso anterior se supone que se desea trabajar con un nivel de confianza del 95 % y un error estándar del 5 % y se analiza la muestra como si esta fuese una muestra aleatoria simple, es decir, sin tomar encueta el plan de muestreo. Los resultados de la estimación de los parámetros serían de 2,9532 con un intervalo de confianza del 95 % el valor estimado estarían entre 2,8287 y 3,0831. En cambio sí se incorpora el plan de muestreo en la estimación de parámetros, el resultado de este sería de 2,4213 con un intervalo de confianza del 95 % este valor estaría entre 2,2210 y 2,6396. Por lo tanto los resultados de las estimaciones de los parámetros al no incluir el plan de muestreo y al incluirlo no se asemeja y difieren mucho más que un 5 % por lo cual se debe tener muy claro cómo fueron capturado los datos antes de realizar el análisis, ya que, no tomar en cuenta el plan de muestreo puede llevar a grandes errores.

Muestreo	Odds ratio	Intervalo de confianza del 95 %
mas	2,9532	2,8287 ; 3,0831
mc	2,4213	2,2210 ; 2,6396

Cuadro 3.7: Comparación de odds ratio sin incorporar el plan de muestreo e incorporándolo

3.4 Estudio de tamaño muestral

Para realizar un estudio de tamaño muestral se supondrá una interrogante, la que trata de buscar si existe alguna relación entre el estar casado y tener hijos, así la unidad de análisis serán las viviendas. Donde en cada vivienda se registrara si la pareja está casada y si tiene hijos; las variables se dejarán de la siguiente forma: si en la vivienda hay una pareja y esta tiene hijos se registrara como 1 si no tiene hijos como 0, esta será la variable respuesta; por otro lado la variable regresora será el registro de si la pareja está casada, si esta está casada se anotará como 1 y si no, como 0.

Todo el diseño muestral complejo y las variables involucradas en el análisis serán generadas utilizando el software Stata, a partir de una base de datos que contiene información real de las manzanas censales de la región de Valparaíso por comuna. Contando con toda esta información se realizará un plan de muestreo por etapas, el que permitirá saber a cuantos individuos se deben encuestar para realizar un correcto análisis de regresión logística.

3.4.1 Plan de muestreo

El plan de muestreo pensado para la región de Valparaíso consta de dos etapas; en la primera etapa se usará un muestreo por estrato y en la segunda un muestreo por conglomerado.

1 Muestreo por estrato: con la información que se cuenta se dividirá la región de Valparaíso en 35 estratos, donde cada estrato es una comuna de la región. La región de Valparaíso tiene un total de 37 comunas, por lo cual en este plan de muestreo no se incorporan las comunas de Isla de Pascua y Juan Fernández.

2 Muestreo por conglomerado: los conglomerados serán las manzanas censales seleccionadas al azar, en cada estrato.

Para realizar los análisis se necesita los factores de expansión; ya que, si no se contara con estos no se podría realizar el análisis de regresión logística como muestra compleja.

3.4.2 Factores de expansión

Los factores de expansión (fe) se deben calcular en cada etapa del muestreo, es decir, en este caso deben ser calculados por estrato y por manzana censal. El fe es calculado con la probabilidad de selección de cada individuo por lo cual queda de la siguiente forma:

$$fe = \frac{1}{\text{probabilidad de selección}} \quad , \tag{3.3}$$

esto representa la cantidad de individuos que representa el dato seleccionado. Para mayor información puede ver Lohr [2000].

3.4.3 Tamaño muestral

El tamaño muestral n, se determina primero para un mas. En este caso se calculará a través de la fórmula estándar descrita en la mayoría de los textos de muestro por ejemplo (Lohr 2000). Esta fórmula está dada por.

$$n_{mas} = \frac{N \cdot Z_\alpha^2 \cdot (1-p)}{e^2 \cdot (N-1) + Z_\alpha^2 \cdot p \cdot (1-p)} \quad ; \tag{3.4}$$

donde, p es la proporción de casos favorables, 100(1-α) es el nivel de confianza y Z_α es el percentil α-ésimo superior de la distribución normal.

Para determinar el tamaño muestral n en la población generada a partir de la base de datos de la región de Valparaíso, que cuenta con todas las manzanas censales, se realizará la estimación con un 95 % de confianza y un error de estimación del 5 %. La población total que consta de 20.393 manzanas censales (N) con una cantidad variable de hogares en cada una de estas. Se supone variabilidad máxima, que se obtiene cuando p=0,5, el tamaño muestral total es de 377 manzanas censales.

Se estudia el tamaño muestral para un mas, ya que, Lohr [2000] describe una relación entre un tamaño muestral para muestreo aleatorio simple (n_{mas}) y el correspondiente tamaño muestral bajo el plan de muestreo complejo (n_{pmc}) a través de la siguiente expresión.

$$n_{pmc} = DEFF * n_{mas} \quad . \tag{3.5}$$

Esta expresión ayuda a determinar el tamaño muestral para una población cuando el muestreo a utilizar es un muestreo complejo. Nótese que sí el efecto de diseño es igual a 1, el n determinado para el mas es el mismo que el n determinado para el plan de muestreo complejo. En este caso se desconoce el efecto del diseño ($DEFF$), por lo cual es necesario obtener el tamaño muestral para el plan de muestreo complejo a través de simulaciones.

Para la población simulada se estima el tamaño muestral de la siguiente forma:

1 La población cuenta con 35 estratos, de los cuales se extraerán 10 manzanas censales de cada estrato haciendo un total de 350 manzanas censales de una población total de 20.393.

2 Para cada manzana censal se registraran 10 viviendas dando un total de 3.500 viviendas de un total de 437.120 viviendas.

3 Se realizan las estimaciones con el software Stata.

Al realizar estos pasos los resultados que se obtienen son los siguientes.

Parámetro	Coeficiente	Error estándar	Valor-p	Intervalo de confianza del 95 %
$\beta_1(casado)$	2,2006	0,1752	0,000	1,8572 ; 2,5440
$\beta_0(constante)$	1,0379	0,0804	0,000	0,8803 ; 1,1957

Cuadro 3.8: Ajuste de un modelo logístico al incorporar el plan de muestre complejo, para la variable respuesta: tienen hijos (si o no) y la covariable están casados (si o no)

El coeficiente 2,2 indica que si las personas están casadas el *logit* estimado aumenta en promedio 2,2 unidades lo que sugiere una relación positiva en que las personas estén casadas y estás tengan hijos.

Otro forma de interpretar los resultados es con los odds ratio; el cual en Stata se obtiene con el comando $svy: \ logit\ tiene\_hijo\ casado, or$ lo que entrega los siguientes resultados.

Parámetro	Odds ratio	Error estándar	Valor-p	Intervalo de confianza del 95 %
$e^{\beta_1}(casado)$	9,0308	1,5820	0,000	6,4061 ; 12,7309
$e^{\beta_0}(constante)$	2,8235	0,2271	0,000	2,4116 ; 3,3057

Cuadro 3.9: Ajuste de un modelo logístico para los odds ratio incorporando el plan de muestre complejo, para la variable respuesta: tienen hijos (si o no) y la covariable están casados (si o no)

Esto implica que una pareja casada tiene un chance 9 veces más alta de tener hijos respecto a personas que no están casada.

Para saber si el tamaño muestral es adecuado se calcula el efecto de diseño de la muestra conocido como $DEFF$ el que es obtenido en Stata 12 con el comando $estat\ effects$ el que entrega los siguientes resultados.

Parámetro	Coeficiente	Error estándar	DEFF
$\beta_1(casado)$	2,2006	0,1752	4,3633
$\beta_0(constante)$	1,0379	0,0804	3,9717

Cuadro 3.10: Efecto de diseño del modelo logístico mostrado en el cuadro 3.8

Los resultados muestran un $DEFF$ mayor a 1 lo que indica que el tamaño muestral debe ser recalculado con la expresión (3.5), por lo cual el total de manzanas censales a encuestar en la región de Valparaíso debe ser de 1.645 dando así un aproximado de 16.450 viviendas.

Capítulo 4

Conclusión

Como síntesis final de este trabajo de titulación se puede decir que, se tienen dos métodos para calcular el tamaño muestral para cuando se desea ajustar un modelo de regresión logístico. Por un lado, se dispone de la fórmula clásica para proporciones que no considera ningún tipo de covariables bajo un muestreo aleatorio. Por otro lado, el método vía simulación que sirve tanto para un muestreo aleatorio simple como para un muestreo complejo. El tamaño muestral es de suma importancia, ya que, los valores de los coeficientes pueden variar significativamente con una mala muestra, lo que llevaría a interpretaciones que no reflejan la realidad.

Por otro lado se demostró la hipótesis de este trabajo, la que consistió en ver si existe diferencia al analizar una muestra capturada con un muestreo complejo como: una muestra aleatoria simple. Es decir, ¿qué ocurre si se excluye el plan de muestreo y se procesan los datos como si fuesen obtenidos a través de un plan de muestreo aleatorio simple? Para esto se utilizó como referencia la encuesta CASEN del 2009, donde los resultados demostraron una diferencia significativa, por lo que es importante revisar el proceso de captura de datos, para así realizar los análisis de la forma correcta tomando en cuenta todos los supuestos que conlleva el modelo de regresión logístico; permitiendo así, realizar inferencias que se aproximan más a la realidad.

Bibliografía

Agresti, A. (2002). Categorical Data Analysis. John Wiley & Sons, New York.

Allison, P. (1999). Logistic Regression Using the SAS R System: Theory and Application. Cary, NC.

Binder, D. (1981). On the variances of asymptotically normal estimators from complex surveys. Survey Methodology, 7, 157-170.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. In: Federation Proceeding; 21. p. 58-61.

Cochran, G. (1977). Sampling techniques. John Wiley & Sons, Toronto .

Day, N. & Kerridge, D. (1967). A general maximum likelihood discriminant. Biometrics; 23: 313-323.

Grizzle, J., Starmer, C., & Koch, G. (1969). Analysis of Categorical Data by Linear Models, Biometrics, 25, 489-504. .

Heeringa, S., West, B. & Berglund, P. (2010). Applied Survey Data Analysis. Chapman and Hall, Boca Raton, FL.

Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression, 2d ed.. John Wiley & Sons, New York.

Long, S. & Freese, J. (2006). Regression Models for Categorical Dependent Variables Using Stata. 2d ad.. Stat Press, College Station, Texas .

Lohr, S. L.(2000). Muestreo diseño y análisis. Thomson, Arizona.

McCullagh, P. & Nelder, J. (1989). Generalized Linear Models, 2d ed.. Chapman and Hall, London.

McCulloch, C. & Searle, S. (2001). Generalized, Linear and Mixed Models. John Wiley & Sons, New York.

Molinero, L. (2001). Fuente http://www.seh-lelha.org/rlogis2.htm. Recuperado el 1 sep 2013.

Ortega M. y Cayuela A. (2002). Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. Revista Española de salud publica, vol. 76, núm. 2. Fuente http://www.redalyc.org/articulo.oa?id=17076202.

Pregibon, D. (1981). Logistic regression diagnostics. Annals of Statistics: 705-724.

Roberts, G., Rao, J. & Kumar, S. (1987). Logistic regression analysis of sample survey data. Biometrika, 74, 1-12.

Skinner, C., Holt, D. & Smith, T. (1989). Analysis of Complex Surveys. John Wiley & Sons, New York.

Capítulo 5

Anexo códigos

Nota: estos códigos son de propiedad del autor de este trabajo de titulación, si desea utilizarlo solicite permiso por escrito a Alexis Silva correo electronico alex.sil.barr@gmail.com.

5.1 Código para simulación muestreo aleatorio simple

```
clear all
version 12
set more off
cd "C:\Users\Alexis\Documents\CARRERA\5 año\tesis\tesis CFHR\codigos"
/***********************************/
/* Tamaño de muestra via simulacion */
/***********************************/
set seed 3102013
set obs 10000
/*********************************************/
/**************generar muestra***************/
/*********************************************/
gen x = runiform()
gen p = runiform()
gen n = rpoisson(33)
gen n2 = rpoisson(25)
gen y = 1 if x>0.45
gen id = _n
replace y =0 if x<=0.45
```

```
table y
/****************************/
/*****manipulando pi*********/
/**para maxima variabilidad**/
/****************************/
gen x1 = 1 if p>0.5
replace x1 =0 if p<=0.5
replace x1 =1 if x>0.5

gen x2 = n if x>0.4
replace x2 = n2 if x<=0.4
replace x2 = n if x>0.8

/****************************/
/**revisando los parametros**/
/****************************/

keep y x1 x2
logit y x2
logit y x1 x2


/**********************/
/**guadando la muestra*/
/**********************/
save logit.dta, replace

/*********************************************************/
/*********Programa para la extraccion de muestras*******/
/**********solo con variables dicotomicas***************/
/*********************************************************/
capture program drop varias_ma_prop
program varias_ma_prop
  clear all
  set seed ‘1’
  postfile mis_prop_sim n id b0 b1 using mis_p_sim‘2’, replace
   forvalues i = 1(1)1000 {
    use logit.dta, clear
    quietly {
      sample ‘2’, count
  logit y x1
  matrix b=e(b)
```

```
}
    post mis_prop_sim (e(N)) (‘i’) (b[1,2]) (b[1,1])
  }
  postclose mis_prop_sim
end

/*************************************/
/**usar el programa para seleccionar***/
/*******las muestras aleatorias********/
/*************************************/


forvalues n=500(500)10000{
  varias_ma_prop ‘n’ ‘n’
}


/***********************************/
/**juntar las muestras en una sola**/
/***********************************/
use mis_p_sim500, clear
forvalues n=1000(500)10000{
  append using mis_p_sim‘n’
}



/***********************************/
/**estadisticas descriptiva de la muestra**/
/**por tamaño muestral**/
/***********************************/
table n, c(mean b0 sd b0 mean b1 sd b1)

/*****************************/
/**Calculo errores estandar**/
/**********para beta 1*******/
/*****************************/
forvalues e=1(1)9 {
  generate i_b1_0_0‘e’ = 1 if abs(b1-3.012711) < 0.0‘e’*3.012711
  replace  i_b1_0_0‘e’ = 0 if i_b1_0_0‘e’ == .
  }
```

```
 forvalues p=10(1)20 {
 generate i_b1_0_‘p’ = 1 if abs(b1-3.012711) < 0.‘p’*3.012711
 replace  i_b1_0_‘p’ = 0 if i_b1_0_‘p’ == .
  }




table n, c(mean i_b1_0_01 mean i_b1_0_02 mean i_b1_0_03 mean
 i_b1_0_04 mean i_b1_0_05)
table n, c(mean i_b1_0_06 mean i_b1_0_07 mean i_b1_0_08 mean
i_b1_0_09  mean i_b1_0_10)

table n, c(mean i_b1_0_11 mean i_b1_0_12 mean i_b1_0_13 mean
i_b1_0_14 mean i_b1_0_15)
table n, c(mean i_b1_0_16 mean i_b1_0_17 mean i_b1_0_18 mean
i_b1_0_19  mean i_b1_0_20 )




/*******************************/
/****para hacer los graficos****/
/*******************************/
forvalues o=1(1)9 {
preserve
contract n i_b1_0_0‘o’ if i_b1_0_0‘o’ == 1
generate algo = 1
rename _freq frec_b‘o’
g id_n =_n
save frec‘o’.dta, replace
restore
}

preserve
contract n i_b1_0_10 if i_b1_0_10 == 1
generate algo = 1
rename _freq frec_b10
g id_n =_n
save frec10.dta, replace
restore
```

```
forvalues n=1(1)10{
  append using frec‘n’
}


#delimit ;
  twoway (connected frec_b1 frec_b2 frec_b3 frec_b4
          frec_b5 frec_b6 frec_b7 frec_b8  frec_b9 frec_b10
  n, msymbol(x x x x x x x x x)
          ytitle("Muestras", size(small))
  xtitle("n")
  xlabel(500 1000 1500 2000 2500 3000 3500 4000 4500
  5000 5500 6000 6500 7000
  7500 8000 8500 9000 9500 10000,alternate labsize(small))
  legend( size(small) rows(2)
  order(1 "0,01" 2 "0,02" 3 "0,03" 4 "0,04" 5 "0,05"
  6 "0,06" 7 "0,07" 8 "0,08" 9 "0,09" 10 "0,10")
         )

  )
 ;
#delimit cr




forvalues o=10(1)20 {
preserve
contract n i_b1_0_‘o’ if i_b1_0_‘o’ == 1
generate algo = 1
rename _freq frec_b‘o’
g id_n =_n
save frec‘o’.dta, replace
restore
}

forvalues n=10(1)20{
  append using frec‘n’
}
```

```
#delimit ;
  twoway (connected frec_b10 frec_b11 frec_b12 frec_b13 frec_b14
          frec_b15  n ,
  connect(L)
          ytitle("Muestras")
  msymbol(x x x x x x x x x)
  xtitle(" " "n")
  clcolor(blue)

  xlabel(500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500)
  legend(size(tiny) rows(2)
  order(1 "0,10" 2 "0,11" 3 "0,12" 4 "0,13" 5 "0,14" 6 "0,15" )
  )
  )

  ;
#delimit cr

#delimit ;
  twoway (connected frec_b16 frec_b17 frec_b18  frec_b19  frec_b20  n ,
  connect(L)
          ytitle("Muestras")
  msymbol(x x x x x x x x x)
  xtitle(" " "n")
  clcolor(blue)

  xlabel(500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500)
  legend(size(tiny) rows(2)
  order(1 "0,16" 2 "0,17" 3 "0,18" 4 "0,19" 5 "0,20")
  )
  )

  ;
#delimit cr


twoway (mean i_b1_0_01, by(n)

dotplot b1, over(n) yline(3.012711) title(Histograma de beta 1)
dotplot b0, over(n) yline(-2.135322) title(Histograma de beta 0)
```

```
scatter b0 b1, by(n) yline(-2.135322) xline(3.012711)


/*******************************************************/
/*********Programa para la extraccion de muestras*******/
/******* variables discreta y binaria ************/
/*******************************************************/

use logit.dta,replace
logit y x1 x2
capture program drop varias_ma_prop
program varias_ma_prop
  clear all
  set seed ‘1’
  postfile mis_prop_sim n id b0 b2 b1 using mis_p_sim‘2’, replace
   forvalues i = 1(1)1000 {
    use logit.dta, clear
    quietly {
      sample ‘2’, count
  logit y x1 x2
  matrix b=e(b)
}
    post mis_prop_sim (e(N)) (‘i’) (b[1,3]) (b[1,2]) (b[1,1])
  }
  postclose mis_prop_sim
end

forvalues n=500(500)10000{
  varias_ma_prop ‘n’ ‘n’
}

use mis_p_sim500, clear
forvalues n=1000(500)5500{
  append using mis_p_sim‘n’
}

table n, c(mean b0 sd b0 mean b2 sd b2 )
table n, c(mean b1 sd b1)


// realizar gráficos
dotplot b0, over(n) yline(-8.582272) title(Histograma de beta 0)
```

```
dotplot b1, over(n) yline(3.003169) title(Histograma de beta 1)
dotplot b2, over(n) yline(0.2198871)  title(Histograma de beta 2)

scatter b0 b1, by(n) yline(-8.582272) xline(3.003169)
scatter b0 b2, by(n) yline(-8.582272) xline(.219887)
scatter b1 b2, by(n) yline(3.003169)  xline(.219887)
// termina gráficos

//para el beta 1
forvalues e=1(1)9 {
  generate i_b1_0_0‘e’ = 1 if abs(b1-3.012711) < 0.0‘e’*3.003169
  replace  i_b1_0_0‘e’ = 0 if i_b1_0_0‘e’ == .
  }

 forvalues p=10(1)20 {
 generate i_b1_0_‘p’ = 1 if abs(b1-3.012711) < 0.‘p’*3.003169
 replace  i_b1_0_‘p’ = 0 if i_b1_0_‘p’ == .
  }

 //para el beta 2
forvalues e=1(1)9 {
  generate i_b2_0_0‘e’ = 1 if abs(b2-.219887) < 0.0‘e’*.219887
  replace  i_b2_0_0‘e’ = 0 if i_b2_0_0‘e’ == .
  }

 forvalues p=10(1)20 {
 generate i_b2_0_‘p’ = 1 if abs(b2-0.219887 ) < 0.‘p’*0.219887
 replace  i_b2_0_‘p’ = 0 if i_b2_0_‘p’ == .
  }



table n, c(mean i_b1_0_01 mean i_b1_0_02
mean i_b1_0_03 mean i_b1_0_04  mean i_b1_0_05)
table n, c(mean i_b1_0_06 mean i_b1_0_07 mean
i_b1_0_08 mean i_b1_0_09  mean i_b1_0_10)
table n, c(mean i_b1_0_11 mean i_b1_0_12
mean i_b1_0_13 mean i_b1_0_14  mean i_b1_0_15)
table n, c(mean i_b1_0_16 mean i_b1_0_17
mean i_b1_0_18 mean i_b1_0_19  mean i_b1_0_20 )
```

```
table n, c(mean i_b2_0_01 mean i_b2_0_02
 mean i_b2_0_03 mean i_b2_0_04  mean i_b2_0_05)
table n, c(mean i_b2_0_06 mean i_b2_0_07
 mean i_b2_0_08 mean i_b2_0_09  mean i_b2_0_10)
table n, c(mean i_b2_0_11 mean i_b2_0_12
 mean i_b2_0_13 mean i_b2_0_14  mean i_b2_0_15)
table n, c(mean i_b2_0_16 mean i_b2_0_17
 mean i_b2_0_18 mean i_b2_0_19  mean i_b2_0_20)


forvalues o=1(1)9 {
preserve
contract n i_b1_0_0‘o’ if i_b1_0_0‘o’ == 1
generate algo = 1
rename _freq2 frec2_b‘o’
g id_n =_n
save frec2‘o’.dta, replace
restore
}

preserve
contract n i_b1_0_10 if i_b1_0_10 == 1
generate algo = 1
rename _freq2 frec2_b10
g id_n =_n
save frec210.dta, replace
restore

forvalues n=1(1)10{
  append using frec2‘n’
}


#delimit ;
  twoway (connected frec2_b1 frec2_b2 frec2_b3 frec2_b4
          frec2_b5 frec2_b6 frec2_b7 frec2_b8  frec2_b9 frec2_b10
  n, msymbol(x x x x x x x x x x)
          ytitle("Muestras", size(small))
  xtitle("n")
  xlabel(500 1000 1500 2000 2500 3000
  3500 4000 4500 5000 5500 6000 6500 7000
  7500 8000 8500 9000 9500 10000,alternate labsize(small))
```

```
  legend( size(small) rows(2)
  order(1 "0,01" 2 "0,02" 3 "0,03" 4 "0,04" 5 "0,05"
  6 "0,06" 7 "0,07" 8 "0,08" 9 "0,09" 10 "0,10")
         )

  )
 ;
#delimit cr
```

5.2 Código para comparar resultados de encuesta CASEN

```
cd "C:\Users\Alexis\Documents\CARRERA\5 año\
tesis\tesis CFHR\CASEN"
set more off

svyset id [pweight=expr],strata(estrato)
 vce(linearized) singleunit(missing) || id
svy: proportion o1

table o1, row col
gen __o1 = 1 if o1 == 1
replace __o1=0 if o1 == 2
table __o1
/**********************************
*****condiciones********************
table edad
drop if edad <  18
drop if edad > 65
table sexo edad
drop if sexo == 2  & edad >60
table edad sexo
count // cantidad de datos 146664
e1 sabe leer y escribir
e7t   tipo de estudios
e7 curso
e7t tipo de curso
table e7t
t19c  ¿conoce ud...? información de
derechos ciudadanos en los servicios público
*/
//drop subpob
generate subpob = o1 if 18<edad<64
```

```
table subpob, row
replace subpob = . if sexo ==2 & edad>60
table subpob,row

gen __e1 = 1 if e1 == 1
replace __e1=0 if e1 ==2


****************************************
***** ################################
****************************************

svy: logit __o1 __e1

svy, subpop(subpob): logit __o1 __e1
estat effects


svy, subpop(subpob): logit __o1 __e1, or
estat effects

logit __o1 __e1
logit __o1 __e1, or
estat class
mfx
predict prob, p
list prob

count
table e1
```

5.3 Código para estimar tamaño de muestra complejo

```
clear all
version 12
set more off
cd "C:\Users\Alexis\Documents\CARRERA\5 año\tesis
\tesis CFHR\codigos simulacion compleja
\nueva compleja\mi codigo"

insheet using "selecfin.csv", delimiter(";")  /* Regiones: V y RM */
```

```
generate lambda = persona/vivienda

set seed 0504
table comuna
drop if comuna == "San Pedro"
drop if comuna == "Valle Hermoso"
drop if comuna == "Valdivia de Pain"
drop if comuna == "El Colorado"
drop if comuna == "El Maitén"
drop if comuna == "Lampa"
drop if comuna == "Estación Colina"
drop if comuna == "Batuco"
drop if comuna == "Lo Herrera"
keep cun manzent vivienda hogar persona ide_encavi comuna lambda
format manzent %16.0g
generate cant_vivienda = vivienda
generate cun2= string(cun)
generate c_region = real(substr(cun2,1,1)) if length(cun2)==4
replace  c_region = real(substr(cun2,1,2)) if length(cun2)==5
generate c_provincia = real(substr(cun2,2,1)) if length(cun2)==4
replace  c_provincia = real(substr(cun2,3,1)) if length(cun2)==5
generate c_comuna = real(substr(cun2,3,2)) if length(cun2)==4
replace  c_comuna = real(substr(cun2,4,2)) if length(cun2)==5
summ cant_vivienda if cun2 == "5101"
tabulate cant_vivienda if cun2 == "5101"
keep if c_region == 5

gen id = _n
keep  cun manzent comuna vivienda hogar persona
ide_encavi id lambda lambda
replace lambda = 0 if lambda == .
replace comu = "San Felipe" if comu == "Curimón"
replace comu = "El tabo" if comu == "Las Cruces"
replace comu = "Puchuncaví" if comu == "Las Ventanas"
replace comu = "Valparaíso"  if comu == "Placilla"
replace comu = "Quilpué" if comu == "Villa Almendros"
replace comu = "El tabo" if comu == "El Tabo"
table comu

save manzanas.dta, replace
```

```
************************************
** Seleccion de manzanas censales
************************************
clear all
capture drop program selecciona_mc
program selecciona_mc
     use manzanas.dta, clear
     bsample ‘1’ if comu == "‘3’" & vivienda != 0 & vivienda != .
 save m_c‘2’, replace
end


selecciona_mc 10 1 "Algarrobo"
selecciona_mc 10 2 "Cabildo"
selecciona_mc 10 3 "Calera"
selecciona_mc 10 4 "Calle Larga"
selecciona_mc 10 5 "Cartagena"
selecciona_mc 10 6 "Casa Blanca"
selecciona_mc 10 7 "Catemu"
selecciona_mc 10 8 "Concón"
selecciona_mc 10 9 "El tabo"
selecciona_mc 10 10 "Hijuelas"
selecciona_mc 10 11 "La Cruz"
selecciona_mc 10 12 "La Ligua"
selecciona_mc 10 13 "Limache"
selecciona_mc 10 14 "Llaillay"
selecciona_mc 10 15 "Los Andes"
selecciona_mc 10 16 "Nogales"
selecciona_mc 10 17 "Olmué"
selecciona_mc 10 18 "Panquehue"
selecciona_mc 10 19 "Papudo"
selecciona_mc 10 20 "Petorca"
selecciona_mc 10 21 "Placilla Penuela"
selecciona_mc 10 22 "Puchuncaví"
selecciona_mc 10 23 "Putaendo"
selecciona_mc 10 24 "Quillota"
selecciona_mc 10 25 "Quilpué"
selecciona_mc 10 26 "Quintero"
selecciona_mc 10 27 "Rinconada"
selecciona_mc 10 28 "San Antonio"
selecciona_mc 10 29 "San Esteban"
selecciona_mc 10 30 "San Felipe"
```

```
selecciona_mc 10 31 "Santa María"
selecciona_mc 10 32 "Santo Domingo"
selecciona_mc 10 33 "Valparaíso"
selecciona_mc 10 34 "Villa Alemana"
selecciona_mc 10 35 "Viña del Mar"



*******************************
*** juntar muestras ************
*******************************
use m_c1, clear
  forvalues i=2(1)35 {
    append using m_c`i'
  }


table comu
generate FE_comu =37.5 if comu == "Algarrobo"
replace FE_comu = 19.6  if comu == "Cabildo"
replace FE_comu = 50.9  if comu == "Calera"
replace FE_comu = 5     if comu == "Calle Larga"
replace FE_comu = 65.3  if comu == "Cartagena"
replace FE_comu = 15.9  if comu == "Casa Blanca"
replace FE_comu = 11    if comu == "Catemu"
replace FE_comu = 47.8  if comu == "Concón"
replace FE_comu = 84.6  if comu == "El tabo"
replace FE_comu = 63.4  if comu == "Hijuelas"
replace FE_comu = 7.8   if comu == "La Cruz"
replace FE_comu = 13.6  if comu == "La Ligua"
replace FE_comu = 34.1  if comu == "Limache"
replace FE_comu = 36.4  if comu == "Llaillay"
replace FE_comu = 21.7  if comu == "Los Andes"
replace FE_comu = 80.7  if comu == "Nogales"
replace FE_comu = 12.2  if comu == "Olmué"
replace FE_comu = 17.5  if comu == "Panquehue"
replace FE_comu = 3.8   if comu == "Papudo"
replace FE_comu = 16.1  if comu == "Petorca"
replace FE_comu = 11.4  if comu == "Placilla Penuela"
replace FE_comu = 32.1  if comu == "Puchuncaví"
replace FE_comu = 26.9  if comu == "Putaendo"
replace FE_comu = 13.1  if comu == "Quillota"
```

```
replace FE_comu = 73.6  if comu == "Quilpué"
replace FE_comu = 179.5 if comu == "Quintero"
replace FE_comu = 56.7  if comu == "Rinconada"
replace FE_comu = 6.1   if comu == "San Antonio"
replace FE_comu = 131.5 if comu == "San Esteban"
replace FE_comu = 6.2   if comu == "San Felipe"
replace FE_comu = 68.1  if comu == "Santa María"
replace FE_comu = 5.3   if comu == "Santo Domingo"
replace FE_comu = 25.4  if comu == "Valparaíso"
replace FE_comu = 329.7 if comu == "Villa Alemana"
replace FE_comu = 428.8 if comu == "Viña del Mar"


generate strata = 1 if comu == "Algarrobo"
replace strata = 2 if comu == "Cabildo"
replace strata = 3 if comu == "Calera"
replace strata = 4 if comu == "Calle Larga"
replace strata = 5 if comu == "Cartagena"
replace strata = 6 if comu == "Casa Blanca"
replace strata = 7 if comu == "Catemu"
replace strata = 8 if comu == "Concón"
replace strata = 9 if comu == "El tabo"
replace strata = 10 if comu == "Hijuelas"
replace strata = 11 if comu == "La Cruz"
replace strata = 12 if comu == "La Ligua"
replace strata = 13 if comu == "Limache"
replace strata = 14 if comu == "Llaillay"
replace strata = 15 if comu == "Los Andes"
replace strata = 16 if comu == "Nogales"
replace strata = 17 if comu == "Olmué"
replace strata = 18 if comu == "Panquehue"
replace strata = 19 if comu == "Papudo"
replace strata = 20 if comu == "Petorca"
replace strata = 21 if comu == "Placilla Penuela"
replace strata = 22 if comu == "Puchuncaví"
replace strata = 23 if comu == "Putaendo"
replace strata = 24 if comu == "Quillota"
replace strata = 25 if comu == "Quilpué"
replace strata = 26 if comu == "Quintero"
replace strata = 27 if comu == "Rinconada"
replace strata = 28 if comu == "San Antonio"
replace strata = 29 if comu == "San Esteban"
```

```
replace strata = 30 if comu == "San Felipe"
replace strata = 31 if comu == "Santa María"
replace strata = 32 if comu == "Santo Domingo"
replace strata = 33 if comu == "Valparaíso"
replace strata = 34 if comu == "Villa Alemana"
replace strata = 35 if comu == "Viña del Mar"

drop id
generate id_muestra = _n

save muestra_pob_05, replace
*************************************
** expancion de la muestra*************
*************************************
use muestra_pob_05,clear
table comuna vivienda


set more off
forvalues i=1(1)350 {
  use muestra_pob_05, clear
  keep if id_muestra ==‘i’
  local id = id_muestra[1]
  local c  = cun[1]
  local m = manzent[1]
  local co = comuna[1]
  local v = vivienda[1]
  local h = hogar[1]
  local p = persona[1]
  local ie = ide_encavi
  local l = lambda
  local f = FE_comu[1]
  local s = strata[1]
  drop cun manzent comuna vivienda hogar persona
   ide_encavi id_muestra lambda FE_comu   strata
  set obs ‘v’
  generate id_muestra = ‘id’
  generate cun = ‘c’
  generate manzent = ‘m’
  generate comu = "‘co’"
  generate vivienda   = ‘v’
  generate hogar = ‘h’
```

```
  generate persona = ‘p’
  generate id_encavi = ‘ie’
  generate hijos = rpoisson(‘l’)
  generate FE_comu = ‘f’
  generate strata = ‘s’
  save p‘i’, replace
}


use p1, clear
  forvalues i=2(1)350 {
    append using p‘i’
  }

save muestra_exp_hogares, replace
count
*******************************************
*******Creacion de variables y**************
******* Analisis de la poblacion************
*******************************************

use muestra_exp_hogares.dta,clear
set seed 25122013


table hijos
replace hijos = rpoisson(1) if hijos >= 16
table hijos


hist hijos
generate tiene_hijo = 1 if hijos>=1
replace tiene_hijo= 0 if tiene==.

table tiene_hijo, row



gen u  = runiform()
gen u2 = runiform()

gen casado = 1 if u>=0.5 &  tiene_hijo == 1
```

```
replace casado= 0 if u < 0.8 & casado == .
replace casado =1 if u <0.2
replace casado = 0 if casado == .
table casado, row

gen id = _n
drop vivienda hogar persona
drop u u2
drop id_encavi
drop cun
format manzent %16.0g

logit tiene_hijo casado
logit tiene_hijo casado,or


**********************************
**generar FE por manzana censal ****
**********************************

table comu
generate FE_manzana = 55.8 if comu == "Algarrobo"
replace  FE_manzana = 16.376   if comu == "Cabildo"
replace  FE_manzana = 43.955   if comu == "Calera"
replace  FE_manzana = 4.089    if comu == "Calle Larga"
replace  FE_manzana = 50.154   if comu == "Cartagena"
replace  FE_manzana = 15.584   if comu == "Casa Blanca"
replace  FE_manzana = 24.642   if comu == "Catemu"
replace  FE_manzana = 37.118   if comu == "Concón"
replace  FE_manzana = 28.478   if comu == "El tabo"
replace  FE_manzana = 9.561     if comu == "Hijuelas"
replace  FE_manzana = 6.935     if comu == "La Cruz"
replace  FE_manzana = 44.431    if comu == "La Ligua"
replace  FE_manzana = 25.869    if comu == "Limache"
replace  FE_manzana = 27.383    if comu == "Llaillay"
replace  FE_manzana = 75.434    if comu == "Los Andes"
replace  FE_manzana = 12.360    if comu == "Nogales"
replace  FE_manzana = 22.349    if comu == "Olmué"
replace  FE_manzana = 3.151     if comu == "Panquehue"
replace  FE_manzana = 9.888     if comu == "Papudo"
replace  FE_manzana = 11.583    if comu == "Petorca"
replace  FE_manzana = 35.031    if comu == "Placilla Penuela"
```

```
replace  FE_manzana = 43.170    if comu == "Puchuncaví"
replace  FE_manzana = 13.637    if comu == "Putaendo"
replace  FE_manzana = 54.238    if comu == "Quillota"
replace  FE_manzana = 206.698   if comu == "Quilpué"
replace  FE_manzana = 51.932    if comu == "Quintero"
replace  FE_manzana = 2.658     if comu == "Rinconada"
replace  FE_manzana = 93.585    if comu == "San Antonio"
replace  FE_manzana = 7.134     if comu == "San Esteban"
replace  FE_manzana = 98.356    if comu == "San Felipe"
replace  FE_manzana = 5.771     if comu == "Santa María"
replace  FE_manzana = 16.699    if comu == "Santo Domingo"
replace  FE_manzana = 367.304   if comu == "Valparaíso"
replace  FE_manzana = 102.547   if comu == "Villa Alemana"
replace  FE_manzana = 429.638   if comu == "Viña del Mar"

save muestra_poblacion_valpo.dta, replace
***********************************
**analisis como muestra compleja****
***********************************
use muestra_poblacion_valpo.dta, clear
table comu
svyset id [pweight=FE_manzana],strata(strata) vce(linearized)
singleunit(missing) || id

svy: logit tiene_hijo casado
estat effects
logit tiene_hijo casado



svy: logit tiene_hijo casado, or
estat effects
di 350* 3.53206
logit tiene_hijo casado, or
```