Universidad Nacional Autónoma de México

Facultad de Ciencias

AJUSTE DE UN MODELO DE REGRESIÓN LOGÍSTICA A LA POBLACIÓN CENSADA DE LA ENCUESTA INTERCENSAL 2015

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A:

JUAN ANDRÉS CERVANTES SANDOVAL

TUTORA

DRA. GUILLERMINA ESLAVA GÓMEZ

CIUDAD DE MÉXICO, 2021

Agradecimientos

Porque de él, y por él, y para él, son todas las cosas.
A él sea la gloria por los siglos. Amén
Romanos 11:36

A mis papás Raúl y Mariana y a mis hermanos Daniela y Joaquín por su vida, comprensión y esfuerzo. A mis abuelos Juan y Alicia por su constante amor, sostén y confianza. A mis tíos Yoyo, Lemuel, Ismael y Miguel por su aprecio y hacer mi vida más divertida.

A Pedro, Fernanda y Diana por sus pláticas, enseñanzas y risas, lo difícil de la facultad valía la pena solo con los divagues que creabamos. Sobre todo a Diana, por su tiempo, amor, presencia y ánimo, por escucharme cada día y aconsejarme en todo, sin ti no lo hubiera logrado.

A la doctora Guillermina por su confianza, enseñanza y paciencia, por todo el tiempo y trabajo invertido en mi. A mis sinodales por todas las molestias tomadas en revisar el trabajo y por sus precisas observaciones.

A mi abuelito Jorge, por su vida y sabiduría, por ser gran parte de quien soy. Solo desearía que me hubieras visto lograrlo.

Resumen

En este trabajo se ejemplificó el uso de modelos de regresión logística a una gran base de datos, tomando ventaja de la información recabada en los municipios que fueron completamente censados en la Encuesta Intercensal 2015 del INEGI. La variable respuesta en el modelo es dicotómica y corresponde a si hay respuesta ($Y = 1$) o no hay respuesta ($Y = 0$) en la pregunta sobre el ingreso recibido en el último mes; las variables explicativas son Edad, Sexo, Nivel Educativo, Situación Laboral, hablar Lengua indígena, Situación Conyugal y Entidad de residencia. La población de interés se compone de aquellos elementos que son sujetos a responder la pregunta sobre ingreso dentro de la población censada y consta de 686,243 individuos u observaciones.

Primero. Utilizando la población de interés y auxiliándose de métodos automatizados de selección de modelos se ajustó un modelo que incluye a las siete variables explicativas y catorce interacciones de primer y segundo orden. El ajuste de este modelo supone que la población censal proviene de una superpoblación.

Segundo. Se ajustó un modelo a una muestra aleatoria simple, *SI*, del 20 % de la población de interés, ésta se compone de 137,243 observaciones. El modelo seleccionado fue aquel que incluyó las siete variables explicativas, y las cinco interacciones: Sexo con Situación Laboral, Sexo con Nivel Educativo, Lengua indígena con Situación Laboral, Lengua indígena con Nivel Educativo y Lengua indígena con Situación Conyugal.

Tercero. Se seleccionaron dos muestras aleatorias con diseño estratificado y por conglomerados, *STSIC*, en la primera se seleccionaron tres conglomerados por estrato y en la segunda solamente dos. Se ajustó un modelo de regresión logística a cada muestra. Considerando que es difícil sostener el supuesto de que las observaciones en la variable respuesta $\{y_1, \cdots, y_n\}$ son independientes debido a que fueron obtenidas con un diseño muestral que usa estratificación y conglomeramiento de la población, se usó la estimación con pseudo-máxima verosimilitud para dar una mejor estimación de la varianza, para esto se utilizó el paquete *Survey* del software *R*. Los dos modelos seleccionados fueron aquellos con las siete variables explicativas, cuatro interacciones en la muestra de mayor tamaño (Sexo con Situación Laboral, Sexo con Nivel Educativo, Lengua indígena con Situación Laboral, Lengua indígena con Nivel Educativo) y con tres interacciones en la de menor tamaño (Sexo con Situación Laboral, Sexo con Nivel Educativo, Lengua indígena con Situación Laboral).

Finalmente, se tomó el modelo seleccionado en la muestra con diseño muestral *STSIC*, el que contiene a los siete efectos principales y cuatro interacciones de primer orden, y se ajustó a la población y a las tres muestras para comparar los valores estimados de los coeficientes y de sus varianzas. Se observó un claro aumento en la magnitud de las varianzas estimadas en las muestras complejas, donde se usa el método que considera el diseño, en particular el conglomeramiento de las observaciones.

Índice General

Índice de Tablas

Introducción

La Encuesta Intercensal 2015 (EIC2015) levantada durante el mes de marzo del 2015, provee la información recabada en alrededor de 5.9 millones de viviendas mediante el cuestionario ampliado. Se trata de una encuesta de cobertura temática amplia que actualiza la información sobre el volumen, composición y distribución de la población y de las viviendas particulares habitadas del territorio nacional. Atendiendo la solicitud de los usuarios y con el propósito de generar información que posibilite dar seguimiento a los grupos vulnerables, se determinó captar la información de todas las viviendas de 814 municipios del país.

Dado que estos municipios fueron completamente censados, las más de tres millones de observaciones en esta población fueron consideradas en una base de datos por separado, ignorando desde luego el diseño muestral de la encuesta. En el *Capítulo 2* se detalla la construcción de esta base de datos, donde además se trabajó la selección de observaciones que cumplen criterios para ser ajustados en la regresión logística, según las variables respuesta Ingreso *(Y)* y explicativas seleccionadas: Sexo, Edad, Situación Laboral, Nivel Educativo, Lengua indígena, Situación Conyugal y Entidad. Esta base, nombrada población de interés, consiste en 686,243 elementos.

En el *Capítulo 3* se describe el ajuste un modelo de regresión logística a la población de interés, donde se consideraron los supuestos estadísticos que el conjunto de observaciones en la variable respuesta, $\{y_1, \cdots, y_{686,243}\}$ constituye una muestra aleatoria de observaciones independientes e idénticamente distribuidas. Posteriormente se realizó una muestra aleatoria simple del 20 % de la población de interés que redujo la carga computacional en los algoritmos de selección del modelo de regresión logística. Esta muestra consistió de $\{y_1, \cdots, y_{137,243}\}$ observaciones que toman en cuenta los mismos supuestos de independencia que en la población.

Con el fin de ejemplificar el ajuste de un modelo de regresión logística cuando las observaciones provienen de un esquema muestral complejo, en el *Capítulo 4* se describe la selección de dos muestras siguiendo el diseño similar al utilizado por el INEGI en la EIC2015 considerando el identificador de Estrato y UPM de pertenencia registrado en cada observación. El ajuste del modelo de regresión logística, la estimación de los coeficientes y de su desviación estándar consideran el supuesto de independencia de las observaciones. Este supuesto es difícil de sostener cuando las observaciones provienen de un esquema complejo de selección, por esto la estimación de los parámetros se realizó utilizando software especializado, el paquete *Survey* del software *R*.

En el *Capítulo 5*, se tomó el modelo seleccionado en la muestra con diseño complejo de mayor tamaño y se estimaron los parámetros en la población y tres muestras. Se analizaron los intervalos de confianza para los coeficientes estimados y se reagruparon dos categorías en la variable Entidad. Al final se analizaron las estimaciones a través de la razón de momios.

Capítulo 1. Conceptos previos

1.1 Modelos lineales generalizados y regresión logística

Los modelos lineales generalizados (*GLM* por sus siglas en inglés *General Lineal Models*) son una extensión de los modelos de regresión lineal, estos abarcan distribuciones con respuesta no-normal y modelan funciones de la media. Los *GLM* tienen tres componentes:

- *Componente aleatorio*: la variable respuesta y es aleatoria y tiene una función de distribución perteneciente a la familia exponencial.
- *Predictor lineal*: $\beta_0 + \sum_{i=1}^{p} \beta_1 x_1 = \boldsymbol{x\beta}$
- *Función liga* g: es una función monótona que relaciona $\mu = E(y|x)$ con las variables explicativas (ver Agresti (2015), p.2):

$$g(E(y|\boldsymbol{x})) = \boldsymbol{x\beta}.$$

Un *GLM* es un modelo lineal para la media de la variable respuesta condicionada en $x_1, ..., x_n$ variables que tiene distribución en la familia exponencial.

Los principales componentes de los *GLM* para respuesta continua y discreta son:

Componente aleatorio	**Función liga**	**Modelo**
Normal	Identidad	Regresión Análisis de varianza
Familia Exponencial	Cualquiera	Modelo lineal generalizado
Binomial	Logit	Regresión logística
Multinomial	Logit	Respuesta multinomial
Poisson	Logaritmo	Loglineal

Agresti (2015), p.5 (tabla 1.1)

Un modelo de regresión logística es aquel que asume una variable $Y \sim$ *Bernoulli* $(\pi(x))$, donde $\pi(x)$ es la probabilidad de éxito, $Y : \{0, 1\}$.

Si se busca estimar el parámetro asociado a la probabilidad de éxito de la variable respuesta $\pi(x)$, con x una variable explicativa, un modelo no lineal, llamado regresión logística es el siguiente (Agresti (2013), p.119)

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}, \quad \text{donde} \quad \pi(x) = P(y = 1|x).$$

Sea $\boldsymbol{x} = (x_1, ..., x_p)$ un vector de p variables explicativas y sea $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ un vector de p parámetros, entonces un modelo no lineal para $\pi(\boldsymbol{x})$ es (Agresti (2013), p.182)

$$P(y=1|\boldsymbol{x}) = \pi(\boldsymbol{x}) = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)} \tag{1.1}$$

y su forma lineal

$$logit[\pi(\boldsymbol{x})] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{1.2}$$

donde la función logit para $\pi(\boldsymbol{x})$ está definida como

$$logit[\pi(\boldsymbol{x})] := log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = log\left(\frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})}\right).$$

De esta manera, $E[Y] = \pi(\boldsymbol{x})$. El valor observado de la variable respuesta Y_i, con i el indicador para cada observación en la muestra, tiene como estimador

$$\hat{P}(y_i = 1|\boldsymbol{x}) = \widehat{E[Y_i]} = \hat{\pi}_i(\boldsymbol{x}_i)$$

con $\hat{\pi}_i(\boldsymbol{x}_i) \in (0,1)$ para toda i. De forma general, una observación con características $\boldsymbol{x}_i$ se clasifica en un primer grupo si se cumple que

$$\hat{\pi}_i(\boldsymbol{x}_i) = \hat{P}(Y_i = 1|X = \boldsymbol{x}_i) > \tau$$

con τ un punto de corte predeterminado que puede reflejar el costo del error de clasificación (comúnmente $\tau = 0.5$, pero no necesariamente). Se clasifica para un segundo grupo en el caso de no cumplirse.

1.1.1 Ecuaciones de verosimilitud de la regresión logística

Considerando que $\pi(\boldsymbol{x})$ es la probabilidad condicional de que Y sea igual a 1 dado el vector $\boldsymbol{x}$, entonces $1 - \pi(\boldsymbol{x})$ es la probabilidad condicional de que Y sea igual a 0. Suponiendo una base de datos que tiene n observaciones independientes $(\boldsymbol{x}_i, y_i)$ con $i = (1, 2, ..., n)$, una forma conveniente de expresar la función de probabilidad para cada observación i es:

$$P(y_i|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)^{y_i}[1 - \pi(\boldsymbol{x}_i)]^{1-y_i}, \quad y_i = \{0, 1\} \tag{1.3}$$

con

$$\pi(\boldsymbol{x}_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}.$$

Buscando obtener los valores estimados de los parámetros $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$, la función de verosimilitud es el producto de los términos dados en la ecuación (1.3)

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}[1 - \pi(\boldsymbol{x}_i)]^{1-y_i}.$$

Para encontrar los valores de $\boldsymbol{\beta}$ que maximicen $l(\boldsymbol{\beta})$ se utiliza la función de logverosimilitud

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(\boldsymbol{x}_i)] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x}_i)]\}.$$

Con el fin de obtener los valores estimados se deriva la logverosimilitud respecto a los $p + 1$ coeficientes y se iguala a cero, lo que resulta en $p + 1$ ecuaciones, estas son (Hosmer et al. (2013), p. 37):

$$\begin{aligned}
\sum_{i=1}^{n} [y_i - \pi(\boldsymbol{x}_i)] &= 0 \\
\sum_{i=1}^{n} x_{i1}[y_i - \pi(\boldsymbol{x}_i)] &= 0 \\
\sum_{i=1}^{n} x_{i2}[y_i - \pi(\boldsymbol{x}_i)] &= 0 \\
&\vdots \\
\sum_{i=1}^{n} x_{ip}[y_i - \pi(\boldsymbol{x}_i)] &= 0.
\end{aligned}$$

Este sistema de ecuaciones no lineales se resuelve usando métodos numéricos y despejando cada $\hat{\beta}_j$. Se denota $\hat{\boldsymbol{\beta}}$ la solución de este sistema de ecuaciones. Los valores ajustados para el modelo de regresión logística son $\hat{\pi}(\boldsymbol{x}_i)$, el valor de la expresión (1.1) calculada usando $\hat{\boldsymbol{\beta}}$ y $\boldsymbol{x}_i$.

Es importante observar que los estimadores $\hat{\boldsymbol{\beta}}$ tienen distribución normal asintótica. Esto permite hacer inferencias sobre los efectos que tienen las variables explicativas.

Prueba de hipótesis para efectos de variables

El vector de parámetros $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ puede ser explicado como el efecto que tiene cada valor del vector $\boldsymbol{x}$ sobre la variable respuesta y_i. Por lo que si el valor de alguna $\beta_j = 0$ entonces la variable x_j no tiene efecto estadísticamente significativo sobre y_i.

Para medir este efecto estadísticamente significativo se usan pruebas de hipótesis que plantean

$$H_0 : \beta_j = \beta_0 \quad vs \quad H_1 : \beta_j \neq \beta_0$$

con $\beta_0 \in \mathbb{R}$.

Una prueba es la *Prueba de Wald*, donde se define el estadístico de prueba:

$$z = \frac{\hat{\beta}_j - \beta}{\sqrt{\widehat{var}(\hat{\beta}_j)}}.$$

La hipótesis nula se rechaza al nivel de significancia α si y solo si $z \leq -z_{1-\frac{\alpha}{2}}$ ó $z \geq z_{1-\frac{\alpha}{2}}$, donde $z_{1-\frac{\alpha}{2}}$ es el cuantil de una distribución normal estándar. Equivalentemente, la hipótesis nula se rechaza si y solo si el valor p es menor que el nivel de significancia α.

También se puede considerar la estadística z^2, la cual tiene una distribución asintótica jí-cuadrada (χ^2) con un grado de libertad. De este modo, el criterio de rechazo para la hipótesis nula con nivel de significancia α es rechazar H_0 si y solo si $z^2 \geq z_{1-\alpha}$, siendo $z_{1-\alpha}$ el cuantil de una distribución χ^2 con un grado de libertad.

Un intervalo de confianza para β_j se obtiene de encontrar dos valores para los cuales

$$-z_{1-\frac{\alpha}{2}} \leq z \leq z_{1-\frac{\alpha}{2}}.$$

Usando la aproximación a la distribución normal y recordando la definición del estadístico z de la prueba de Wald:

$$\hat{\beta}_j - z_{1-\frac{\alpha}{2}}\hat{se}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + z_{1-\frac{\alpha}{2}}\hat{se}(\hat{\beta}_j).$$

Otra prueba es la *Razón de verosimilitudes*, donde usa un estadístico

$$\Lambda = \frac{L_0}{L_1},$$

con L_0 como los estimadores máximo verosímiles obtenidos con la función de logverosimilitud bajo la hipótesis nula, y L_1 en todo el espacio paramétrico. Donde, de forma asintótica a la familia exponencial, $-2log\Lambda$ tiene una distribución a χ^2 con un grado de libertad. La hipótesis nula se rechaza a un nivel α si y sólo si $-2log\Lambda \geq z_{1-\alpha}$ con $z_{1-\alpha}$ siendo el cuantil de una χ^2 y un nivel α de significancia dado.

Interacciones entre variables explicativas en el modelo

Una interacción en un modelo de regresión logística ocurre si la relación entre una variable explicativa x_k y la variable respuesta Y_i dependen del valor de otra variable explicativa x_h. Se dice que x_h es el modificador o magnificador de la variable x_k. Una variable de interacción representa un efecto proveniente de la multiplicación de las respuestas x_k y x_h, por lo que $x_k \times x_h$ es un efecto no aditivo que está sobre el efecto lineal que tienen las variables por su cuenta. El coeficiente de regresión que tiene este producto indica qué tan importante es

la relación para el modelo de estas dos variables. Por lo que, tomando en cuenta todas las interacciones:

$$logit[\pi(\boldsymbol{x})] = \beta_0+\beta_1 x_1+...+\beta_p x_p+\beta_{p+1} x_1 x_2+...+\beta_{p+\binom{p}{2}} x_{(p-1)} x_p+...+\beta_{\binom{p}{1}+\binom{p}{2}+...+\binom{p}{p}} x_1 x_2 ... x_p. \tag{1.4}$$

Ejemplo 1. Modelo nulo y saturado (considerando interacciones de variables explicativas)

Considerando el ejemplo dado en Fitzmaurice et al. (1997), donde se usa una muestra del *Censo Británico de 1991*. En un sector de la población, la variable respuesta de forma dicotómica es: $Y = 1$ si pertenece a una clase profesional o gerencial, $Y = 0$ si no pertenece a esta clase. Se consideran tres variables explicativas, Sexo (G=0 si es hombre y G=1 si es mujer), educación (E) y Edad (A). Educación se trata como una variable dicotómica que considera si se obtuvieron estudios de postsecundaria (E=1) o no (E=0). La variable edad de las observaciones está categorizada de 25 a 34 años (A=0), de 35 a 44 años (A=1), de 45 a 54 años (A=2) y de 55 a 65 años (A=3).

Un *modelo nulo* se define como

$$logit[\pi(\boldsymbol{x})] = log\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \beta_0$$

en el cual el estimador del parámetro de la probabilidad $\pi(\boldsymbol{x})$ es independiente de cualquier variable explicativa y es constante para todos los valores de las variables explicativas.

El modelo de la ecuación *(1.4)* es el más complejo que puede explicar la asociación de los datos. En este ejemplo incluye las interacciones de primer orden (las variables explicativas dos a dos) y las de segundo orden (las tres variables). Cuando los valores estimados son tantos como todas las variables e interacciones posibles de ellas se llama *modelo saturado*:

$$\begin{aligned} logit[\pi(\boldsymbol{x})] =& \beta_0 + \beta_1 G + \beta_2(A=1) + \beta_3(A=2) + \beta_4(A=3) + \beta_5 E \\ &+ \beta_6 G \times (A=1) + \beta_7 G \times (A=2) + \beta_8 G \times (A=3) \\ &+ \beta_9 G \times E + \beta_{10} E \times (A=1) + \beta_{11} E \times (A=2) \\ &+ \beta_{12} E \times (A=3) + \beta_{13} G \times E \times (A=1) \\ &+ \beta_{14} G \times E \times (A=2) + \beta_{15} G \times E \times (A=3). \end{aligned}$$

Si el número de variables explicativas es grande este modelo es poco útil en la práctica, ya que se buscan dos metas: que el modelo sea lo bastante complejo para ajustar a los datos y ser lo bastante sencillo para que se pueda ajustar con la menor incertidumbre posible.

1.1.2 Bondad de ajuste del modelo

Se refiere a verificar que tan acertadas son las estimaciones del modelo ajustado en una base de datos. Se buscan entonces métricas que digan cuantitativamente qué tan cercanas son las estimaciones con los valores observados. Estas medidas darán un resultado estadístico con el cual poder tomar decisiones en la selección de modelos. La bondad de ajuste de un

modelo de regresión logística generalmente se mide a través de la devianza.

Análisis de devianza

Compara el modelo saturado con un modelo seleccionado, toma los estimadores máximo verosímiles del modelo saturado $\boldsymbol{\mu} = (\mu_0, ..., \mu_n)$ $(L(\boldsymbol{\mu}))$ y los estimadores máximo verosímiles del modelo seleccionado $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)$ $(L(\hat{\boldsymbol{\beta}}))$. La expresión de la devianza es:

$$dev = -2[L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\mu})].$$

Si los estimadores máximo verosímiles del modelo ajustado con p variables explicativas hace que la expresión anterior sea cero, se tendría que es el mismo comportamiento que el modelo saturado. Por lo que se busca el valor más pequeño de la devianza. Cuando se toma la verosimilitud evaluada en los valores estimados para el modelo saturado y es el caso de una regresión logística con $Y_i \sim Bernoulli(\pi_i)$, se cumple que

$$L(modelo \quad saturado) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i} = 1,$$

esto porque al ser el modelo saturado, los valores estimados $\hat{y}_i = y_i$, entonces la devianza es:

$$dev = -2(L(\hat{\beta}_0, ..., \hat{\beta}_p)).$$

La devianza permite la comparación de modelos anidados, esto es cuando el conjunto de parámetros de un modelo 1 con q variables explicativas es subconjunto del conjunto de parámetros de otro modelo 2 con p variables explicativas, es decir, los parámetros del modelo 1 son: $\beta_{i_1}, ..., \beta_{i_q}$ para $i_1, ..., i_q \in 1, ..., p$. Dado que, en regresión logística, $dev_1 - dev_2$ se distribuye asintóticamente con una χ^2 con $p - q$ grados de libertad, es posible realizar la prueba de hipótesis

$$H_0 : dev_1 - dev_2 = 0 \quad vs \quad H_1 : dev_1 - dev_2 \neq 0.$$

La hipótesis nula se rechaza si y solo si el modelo 2 presenta mejor comportamiento que el modelo 1, es decir, que se rechaza sólo si hay significancia estadística que el modelo con más variables presenta un mejor comportamiento que el que tiene menos.

AIC

El Criterio de Información de Akaike, evalúa un modelo por la cercanía con la que los valores se aproximan a la media del modelo saturado. Además de que busca penalizar la sobrecarga de variables que no son relevantes. Si $\hat{\boldsymbol{\beta}}$ es el vector de estimadores con máxima verosimilitud para el modelo ajustado M, el AIC se define como:

$$AIC = 2k - 2L(\hat{\boldsymbol{\beta}}_M),$$

donde k es el número de parámetros del modelo M. Se tiene que, un menor valor de AIC muestra un mejor comportamiento de un modelo comparado con otro.

BIC

El Criterio de Información Bayesiano también evalúa un modelo por la cercanía con la que los valores se aproximan a la media del modelo saturado y penaliza la sobrecarga de variables que no son relevantes. Este reemplaza el 2 del AIC con $\ln(n)$, n siendo el número de observaciones (Agresti (2013), p.212). Su fórmula está dada por:

$$BIC = ln(n)k - 2L(\hat{\boldsymbol{\beta}}_M).$$

Este criterio penaliza con mayor severidad el número de parámetros en el modelo cuando n es grande, así el modelo seleccionado con este criterio no es más complejo que el elegido con AIC si n es grande. El criterio BIC es más propenso de elegir un mejor modelo que los otros criterios mientras la n converge a infinito.

El AIC *y* BIC son medidas meramente comparativas para selección de modelos, de forma independiente y calculados en el ajuste bajo la misma base de datos.

Significancia en el modelo

La forma para evaluar la significancia de variables de forma global del modelo se realiza con la prueba de hipótesis:

$$H_0 : (\beta_1, ..., \beta_p) = \mathbf{0} \quad vs \quad H_1 : \beta_j \neq 0, \quad \text{para alguna} \quad j \in (1, .., p),$$

lo que da una forma de decidir si existe efecto entre la variable media y cualquiera de las variables explicativas. Cuando se rechaza la hipótesis nula se decide que al menos una de las variables explicativas tiene un efecto significativo en la respuesta, es decir, proporciona un mejor modelo que el que no tiene esa variable explicativa.

La devianza del *modelo nulo* ($logit[\pi(x)] = \beta_0$) es llamada *devianza nula*. Por lo que la prueba de hipótesis prueba que la resta de las devianzas del modelo nulo y el seleccionado sea igual cero o no. Esto equivale a la prueba de hipótesis del modelo M y el Modelo nulo:

$$H_0 : dev_M - dev_{Null} = 0 \quad vs \quad H_1 : dev_M - dev_{Null} \neq 0.$$

El criterio de rechazo para la hipótesis nula al nivel α se hará si y solo si el valor p es menor que el nivel de significancia α.

Procesos de selección de modelo

Cuando se quiere comprobar la significancia estadística de variables o interacciones entre ellas, en especial cuando éstas son numerosas, un método algorítmico es de gran utilidad en seleccionar un modelo comprobando la significancia del ajuste de las variables e interacciones.

Uno de estos procedimientos de selección es el *step*, actúan agregando o retirando secuencialmente variables e interacciones en el modelo y efectuando pruebas estadísticas usando un criterio predeterminado, este puede ser *AIC*, *BIC* o la devianza. Este proceso se puede dividir en *forward, backward y backward-forward*. A continuación, se explica el algoritmo para la selección *forward* y se describe de forma general para *backward* y *bacward-forward*. Los algoritmos están descritos de forma detallada e ilustrados en Hosmer et al. (2013), p. 212.

Forward selection

El algoritmo comienza suponiendo que se tiene un total de p posibles variables que explican una variable respuesta Y. Se ajusta el modelo nulo y se evalúan los estimadores máximo verosímiles, posteriormente, se agregan una a una las p variables. Con cada variable x_j se evalúa la logverosimilitud, se realiza el cálculo del criterio de selección elegido (*AIC*, *BIC*, devianza) y se revisa el valor p obtenido de la prueba de hipótesis respectiva. Se realiza con cada variable y la variable que haya presentado el valor p más pequeño es seleccionada y añadida al modelo. Cada variable es ahora añadida, evaluando su logverosimilitud y comparando con el modelo anterior. El proceso continúa con cada variable y posteriormente con cada interacción entre variables, terminando cuando agregar cualquiera no mejora el modelo de forma estadísticamente significativa según el criterio seleccionado.

Backward selection

El algoritmo en este caso es inverso al anterior, se empieza con todas las variables e interacciones explicativas en el modelo, el modelo saturado. Posteriormente, de forma iterativa se va probando la significancia estadística de cada una según el criterio seleccionado, así retirando las que no lo cumplan. La selección se detiene cuando eliminar cualquier variable o interacción no mejora el comportamiento del modelo.

Backward-forward selection

Este algoritmo es más elaborado, ya que combina los dos anteriores. En este caso se inicia sin variables ni interacciones explicativas, el modelo nulo, después de que una variable haya sido agregada, se revisa el modelo y elimina cualquiera que ahora no cumpla con el criterio. El algoritmo se detiene cuando agregar o retirar alguna variable o interacción no mejora el desempeño.

Datos agrupados y no agrupados

En el caso de que todas las variables explicativas sean categóricas, las observaciones se pueden presentar en una tabla de contingencia, tomando los datos agrupados y asignando la frecuencia en cada celda.

Ejemplo 2. Tabla de contingencia para variables categóricas

Tomando el *Ejemplo 1*, sean $N = 491,403$ observaciones, considerando las tres variables explicativas y la variable respuesta, se genera una base de datos de 4 columnas y 491,403 filas. Se considera que las observaciones por cada categoría de las variables explicativas están distribuidas de la siguiente forma:

- Sexo (G). Hombre (G=0) con 314,498 observaciones y Mujer (G=1) con 176,905.
- Educación (E). Sin estudios de postsecundaria (E=0) 383,294 y Con estudios de postsecundaria (E=1) 108,109 observaciones.
- Edad (A). De 25-34 años (A=0) 93,367, de 35-44 años (A=1) 127,765, de 45-54 (A=2) años 140,248 y de 55-65 años (A=3) 130,023 observaciones.
- Variable respuesta, puesto profesional o gerencial. No (Y=0) 289,928 y Si (Y=1) 201,475 observaciones.

Una tabla de contingencia toma las combinaciones de cada una de las variables y asigna la frecuencia de observaciones que cumplen con esos casos. Por lo que en el ejemplo la tabla de contingencia es de 32 filas originadas por las categorías de las variables [Sexo (2) × Educación (2) × Edad (4) × respuesta a puesto profesional o gerencial (2)= 32]; tiene 5 columnas de las 4 variables y la asignación de frecuencia.

Tabla 1: Tabla de contingencia considerando variables Sexo (G), Edad (A), Educación (E) y respuesta (Y) en el Ejemplo 2

G	E	A	Y	Frec
0	0	0	0	27,380
0	0	0	1	19,249
0	0	1	0	37,788
0	0	1	1	26,191
0	0	2	0	41,221
0	0	2	1	28,678
0	0	3	0	38,397
0	0	3	1	26,495
0	1	0	0	7,767
0	1	0	1	5,296
0	1	1	0	10,570
0	1	1	1	7,360
0	1	2	0	11,682
0	1	2	1	8,115
0	1	3	0	10,738
0	1	3	1	7,571

G	E	A	Y	Frec
1	0	0	0	15,506
1	0	0	1	10,691
1	0	1	0	21,245
1	0	1	1	14,568
1	0	2	0	23,113
1	0	2	1	16,346
1	0	3	0	21,446
1	0	3	1	14,980
1	1	0	0	4,442
1	1	0	1	3,036
1	1	1	0	5,938
1	1	1	1	4,105
1	1	2	0	6,556
1	1	2	1	4,537
1	1	3	0	6,139
1	1	3	1	4,257

La tabla de contingencia permite manejar la información de una forma computacionalmente más eficiente, además al hacer esto ocurre:

- El ajuste de un modelo de regresión logística usando los datos agrupados o no agrupados produce los mismos valores ajustados de las estimaciones porque son los mismos datos, $\hat{P}(Y = 1|X = \boldsymbol{x})$.
- El valor de la devianza de cada ajuste es diferente.
- El valor de la devianza para datos agrupados sirve para probar la bondad del modelo al compararlo con otros modelos ajustados a los datos arupados, no así con los datos no agrupados.

1.1.3 Razón de momios e interpretación del modelo de regresión logística

El valor del estimador de los parámetros $\hat{\boldsymbol{\beta}}$ corresponde al efecto de las variables explicativas sobre la variable respuesta, sin embargo, las interpretaciones no son directas y deben darse en términos de $\hat{\pi}_i(\boldsymbol{x})$. La interpretación del modelo puede darse a través de las definiciones de momios y razón de momios.

El cociente

$$\Omega = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} \qquad con \quad 0 < \pi(\boldsymbol{x}) < 1$$

da una interpretación de la probabilidad de que ocurra un evento comparado con la probabilidad de que no ocurra, este cociente es llamado *Momio*. Cuando $0 < \Omega < 1$ es más probable la ausencia del evento que su presencia. Si $\Omega > 1$, se tiene que es más probable que ocurra el evento a que no ocurra.

Analizando dos eventos distintos, se considera el cociente:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1(\boldsymbol{x})/(1 - \pi_1(\boldsymbol{x}))}{\pi_2(\boldsymbol{x})/(1 - \pi_2(\boldsymbol{x}))}$$

que se conoce como *razón de momios*. El cual, si $\theta = 1$ implica que $\pi_i = \pi_j$, cuando $\theta > 1$, se tiene $\pi_i > \pi_j$.

De esta manera, la transformación *logit* representa el logaritmo del momio y este compara la probabilidad de éxito contra la de fracaso de cada uno de los ensayos Bernoulli de la variable respuesta. Tomando la ecuación (1.2) y aplicándole la función exponencial:

$$\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j).$$

La interpretación para la estimación de cada coeficiente de regresión (Agresti (2013), p.164) es que el momio se multiplica por la exponencial de las betas por cada unidad incrementada

de una variable explicativa x_k siempre que las otras permanezcan constantes. Lo que quiere decir:

$$\frac{Momio \quad evaluado \quad en \quad x_k + 1}{Momio \quad evaluado \quad en \quad x_k} = \frac{\exp[(\beta_0 + \sum_{j=1, j\neq k}^{p} \beta_j x_j) + \beta_k(x_k + 1)]}{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)} = \exp(\beta_k).$$

Entonces:

$$Momio \quad evaluado \quad en \quad x_k + 1 = exp(\beta_k) * Momio \quad evaluado \quad en \quad x_k$$

lo que significa que por cada unidad aumentada de la variable x_k, el momio se multiplica por $\exp(\beta_k)$. Además, es importante notar que el signo que tendrá la estimación de β_k definirá si la probabilidad aumenta o disminuye.

Un intervalo de confianza para la razón de momios es aplicando la función exponencial en el intervalo de Wald, dando un intervalo del $(1 - \alpha) \times 100\,\%$ de confianza para $exp(\beta_k)$:

$$\exp[\hat{\beta}_k - z_{1-\frac{\alpha}{2}} \times \widehat{se}(\hat{\beta}_k)] < \exp(\beta_k) < \exp[\hat{\beta}_k + z_{1-\frac{\alpha}{2}} \times \widehat{se}(\hat{\beta}_k)].$$

1.2 Diseño muestral

Sea una población de N elementos $U = \{1, ..., k, ..., N\}$, es posible obtener bajo un esquema probabilístico de selección un subconjunto s de tamaño n llamado *muestra*. Con un esquema probabilístico dado, la *probabilidad de seleccionar la muestra* s se denota con $p(s)$.

Como se especifica en Särndal et al. (1992), p. 28, para un diseño de muestra dado $p(.)$, es posible considerar cualquier muestra s como el resultado de una variable aleatoria establecida S, cuya distribución de probabilidad está especificada por la función $p(.)$. Sea ℓ el conjunto de todas las muestras s, entonces se tiene que $P(S = s) = p(s)$ para cualquier $s \in \ell$. Porque $p(s)$ es la distribución de probabilidad de ℓ y se tiene que

$$\begin{aligned} i) \quad & p(s) \geq 0, \quad \text{para toda} \quad s \in \ell, \\ ii) \quad & \sum_{s \in \ell} p(s) = 1. \end{aligned}$$

Suponiendo un diseño muestral dado, la inclusión en la muestra de un elemento fijo k es un evento aleatorio indicado por la variable aleatoria I_k, definida como:

$$I_k = \begin{cases} 1, & \text{si k} \in s \\ 0, & \text{si no.} \end{cases}$$

La *probabilidad de inclusión* π_k, es la probabilidad que la unidad k sea seleccionada en muestra, definida como (Särndal et al. (1992), p. 31, (2.4.2)):

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{k \in s} p(s).$$

Observe que la notación π_k es similar a la de probabilidad de respuesta en la regresión logística, pero son expresiones distintas.

El correspondiente *factor de expansión* de la unidad poblacional k es $w_k = \frac{1}{\pi_k}$. La unidad k es auto representada si $w_k = 1$.

Muestreo Aleatorio Simple (SI)

Bajo un diseño aleatorio simple cada muestra s de tamaño fijo n recibe la misma probabilidad de ser seleccionada, esto es:

$$p(s) = \begin{cases} 1/\binom{N}{n} & \text{si s es de tamaño n.} \\ 0, & \text{e.o.c.} \end{cases}$$

las probabilidades de inclusión para cada unidad k y kl son (Särndal et al. (1992), p. 66):

$$\pi_k = \frac{n}{N} = f, \quad \text{k=1,..,N}$$

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)} = f, \quad \text{k} \neq l = 1, .., N$$

Muestreo estratificado

En este diseño (descrito en Särndal et al. (1992), p. 100) la población finita $U = \{1, ..., k, ..., N\}$ es particionada en H subpoblaciones disjuntas llamadas *estratos*, y denotadas por $U_1, ..., U_h, ...U_H$. Con muestreo estratificado se refiere a que una muestra s_h es seleccionada de un estrato U_h de acuerdo a un diseño $p_h(.)$ y que la selección en un estrato es independiente de la selección de otro estrato. El total de la muestra, denotado s, estará compuesto de

$$s = s_1 \cup s_2 \cup .. \cup s_H$$

Y por la independencia en la selección de cada estrato

$$p(s) = p_1(s_1)p_2(s_2)...p_H(s_H).$$

El número de elementos en el estrato h, llamado tamaño del estrato se denota N_h, que se asume conocido. Como los estratos son una partición de U, se tiene:

$$N = \sum_{h=1}^{H} N_h.$$

Si el diseño probabilístico para todos los estratos es un aleatorio simple, la probabilidad de selección para la unidad k en el estrato h está dada por

$$\pi_{h_k} = n_h/N_h$$

y se denota a este esquema como STSI.

Muestreo unietápico por conglomerados

En este diseño muestral, la población $U = 1, .., k, ...N$ es agrupada en N_I subpoblaciones llamadas conglomerados (o Unidades Primarias de Muestreo) y denotados por $U_1, ..., U_i, ..., U_{N_I}$, el conjunto de conglomerados es representado como:

$$U_I = \{1, ..., i, ..., N_I\}.$$

La notación I representa la etapa de conglomeramiento de las observaciones, en el caso unietápico solo existe un primer conglomeramiento. El número de elementos en el i-ésimo conglomerado U_i es denotado por N_i. La partición de U es expresada por las ecuaciones:

$$U = \bigcup_{i \in U_I} U_i \qquad y \qquad N = \sum_{i \in U_I} N_i.$$

Una muestra s_I de conglomerados de U_I tiene probabilidad de ser seleccionada de acuerdo al diseño dado $p_I(.)$. Cada elemento en el conglomerado seleccionado es observado. Entonces la muestra s está definida por

$$s = \bigcup_{i \in S_I} U_i$$

y el tamaño de s es

$$n_s = \sum_{s_I} N_i.$$

Es importante notar que incluso cuando $p_I(.)$ tenga un tamaño en el diseño fijo, el número de los elementos observados n_s no lo será, porque los tamaños de los conglomerados N_i pueden variar. Las probabilidades de inclusión de primer y segundo orden bajo el diseño $p_I(.)$ son (Särndal et al., (1992), p. 127)

$$\pi_{I_i} = \sum_{i \in S_I} p_I(S_I) \quad y$$
$$\pi_{I_{ij}} = \sum_{i \& j \in S_I} p_I(S_I).$$

Se explican los dos diseños anteriores para entender el funcionamiento del diseño estratificado por conglomerados, en una sola etapa aleatoria simple *(STSIC)*. Sin embargo, este diseño se describe con detalle en el *Capítulo 4* dado que se desarrolla en esa sección con una notación distinta a la que se ha trabajado.

Capítulo 2. Información sobre la base de datos

Esta sección presenta el conjunto de datos considerados para el ajuste de los modelos de regresión logística que estiman la probabilidad dar respuesta en la variable Ingreso (Y=1) considerando siete variables explicativas. Es importante resaltar que para el ajuste de los modelos en el *Capítulo 3* solo se tomaron los datos censales y no se consideró el diseño muestral, su descripción es para entender como está construida la población censada, además de que el diseño se considera en el *Capítulo 4*.

Se trabajó con distintas etapas de selección de población que son definidas en su sección correspondiente, pero para considerar una nomenclatura general se presenta la siguiente *Tabla 2*.

Tabla 2: Etapas de selección de conjuntos poblacionales

Nombre	Observaciones de	Total	
Población encuestada 21 Estados	Estados que tienen municipios que fueron censados	20,017,705	*Tabla 4*
Población censal (pc)	Localidades donde se aplicó censo	3,737,259	*Sección 2.1*
Población de interés(*PI*)	Localidades de *Población censal* conforme a cumplimiento de variables definidas	686,243	*Sección 2.2*
Muestra SI del 20 % (*SI 0.2*)	Muestra *SI* del 20 % de la población de interés	137,249	*Sección 3.2*
Muestra *STSIC* $m_k = 3$ (*STSIC*3)	Muestra *STSIC* seleccionando $m_k = 3$ UPMs por estrato de la población de interés	110,453	*Sección 4.1*
Muestra *STSIC* $m_k = 2$ (*STSIC*2)	Muestra *STSIC* seleccionando $m_k = 2$ UPMs por estrato de la población de interés	77,980	*Sección 4.1*

Encuesta Intercensal 2015 (INEGI)

El esquema de muestreo de la Encuesta Intercensal 2015 es probabilístico, estratificado, por conglomerados y unietápico. En la muestra se seleccionaron áreas geográficas completas (UPM) utilizando muestreo aleatorio simple, y en su interior, se visitaron todas las viviendas para captar sus características y las de sus residentes. En las localidades menores a 15 mil se seleccionaron al menos dos UPM por tamaño de localidad y estrato socioeconómico al interior del municipio. En tanto que en las localidades de 15 mil a 49,999 habitantes se estratificaron las AGEB por estrato socioeconómico, mediante una selección de al menos dos UPM por estrato al interior de cada localidad. Para las localidades de 50 mil o más habitantes se estratificaron las AGEB por estrato socioeconómico, haciendo una distribución proporcional al número de viviendas del estrato y asignando al menos una UPM por AGEB. Los municipios y localidades que a petición de diferentes instituciones fueron censados, se consideraron en estratos independientes.

2.1 Población censal

Se determinó captar la información de las viviendas de 814 municipios (aunque solo se logró en 810), divididos en 21 entidades de la República, donde su población era de menos de 1,300 viviendas habitadas o cumplieran algún criterio de vulnerabilidad. Los municipios considerados estaban en alguna de las siguientes categorías, tomadas de *Encuesta Intercensal 2015. Síntesis metodológica y conceptual*, (p.69, 79-123) donde puede encontrase la relación completa de los municipios:

Tabla 3: Características de los municipios con cobertura completa

Característica	**Municipios**
Con menos de 1,300 viviendas según censo 2010	713
Forman parte de los 100 municipios con menor Índice de Desarrollo Humano (IDH)	88
Grado de rezago social muy alto	113
Alta concentración de localidades con población afrodescendiente	8
Forman parte de los 100 municipios con mayor porcentaje de personas en pobreza alimentaria	80
Mayor porcentaje de personas en pobreza extrema	100
Forman parte de los 100 municipios con mayor porcentaje de personas con carencia por acceso a la alimentación	75

Un mismo municipio puede cumplir varias características

Al ser localidades de las que se pedía especial atención, todas fueron tomadas como estratos, y todas las unidades primarias de muestreo en estos estratos fueron seleccionadas. Aunque no se encuentran en la población censada considerada en este trabajo también entraron con certeza a la muestra localidades que tuvieran: i) Un porcentaje estimado de 2 % de población afromexicana, ii) hablantes de lenguas indígenas en peligro de extinción o iii) hablantes de lengua indígena de interés para la CDI (Comisión Nacional para el Desarrollo de los Pueblos Indígenas). Esto se detalla en *Encuesta Intercensal 2015. Síntesis metodológica y conceptual*, (p.69).

Tabla 4: Estratificación de la población en estados con localidades censadas

		Municipios en			Estratos en			UPMs de		
cod	**Estado**	**muestra**	**censo**	**Total**	**muestra**	**censo**	**Total**	**muestra**	**censo**	**Total**
20	Oaxaca	129	**441**	*568	809	**1,403**	2,212	13,223	**15,125**	28,348
21	Puebla	152	**65**	217	1,066	**242**	1,308	15,241	**3,147**	18,388
30	Veracruz	182	**30**	212	1,522	**127**	1,649	22,672	**1,768**	24,440
14	Jalisco	109	**16**	125	971	**77**	1,048	14,951	**1,294**	16,245
15	México	120	**5**	125	1,303	**25**	1,328	14,737	**307**	15,044
7	Chiapas	86	**32**	118	867	**212**	1,079	11,796	**5,373**	17,169
16	Michoacán	110	**3**	113	924	**14**	938	14,183	**202**	14,385
31	Yucatán	51	**55**	106	288	**157**	445	5,500	**3,022**	8,522
13	Hidalgo	81	**3**	84	619	**13**	632	9,460	**272**	9,732
12	Guerrero	55	**26**	81	544	**172**	716	8,957	**4,677**	13,634
26	Sonora	34	**38**	72	317	**112**	429	6,465	**1,782**	8,247
8	Chihuahua	42	**25**	67	302	**104**	406	7,835	**3,782**	11,617
29	Tlaxcala	47	**13**	60	255	**37**	292	4,355	**666**	5,021
24	S.L. Potosí	53	**5**	58	421	**24**	445	7,757	**654**	8,411
32	Zacatecas	45	**13**	58	374	**62**	436	6,679	**761**	7,440
19	Nuevo León	37	**14**	51	300	**43**	343	6,372	**870**	7,242
11	Guanajuato	45	**1**	46	650	**5**	655	8,056	**83**	8,139
28	Tamaulipas	36	**7**	43	313	**30**	343	6,673	**500**	7,173
10	Durango	31	**8**	39	236	**35**	271	5,372	**1,085**	6,457
5	Coahuila	29	**9**	38	275	**25**	300	5,578	**502**	6,080
18	Nayarit	19	**1**	20	167	**7**	174	3,067	**467**	3,534
	Total	1,493	**810**	2,303	12,523	**2,926**	15,449	198,929	**46,339**	245,268

*Solo en Oaxaca hay 2 municipios mixtos (censados y muestreados).

La *Tabla 4* detalla el número de municipios que fueron censados con respecto a los muestreados y el total por estado, de igual manera, se aprovechó la división geográfica en estratos y UPMs para dar información de cuantos se censaron con respecto al total. Este censo planeaba abarcar cada municipio por completo, sin embargo, hay dos municipios en Oaxaca donde solo se censo una parte. Se observa que este estado contiene una clara mayoría de sus municipios en la parte censada con respecto al total.

Tabla 5: Distribución de la población censal por estados

Entidad		Población				Hogares				Municipios	
Cod	Nombre	$\sum$ Pesos muestrales	%	Total de individuos	%	$\sum$ Pesos muestrales	%	Total	%	Total	%
20	Oaxaca	1,187,027	30.9	1,178,043	31.5	303,404	32.9	301,220	33.5	441	54.4
7	Chiapas	809,888	21.1	760,120	20.3	161,033	17.5	151,883	16.9	32	4.0
12	Guerrero	462,887	12.0	449,802	12.0	105,216	11.4	102,354	11.4	26	3.2
21	Puebla	255,431	6.6	255,077	6.8	64,788	7.0	64,709	7.2	65	8.0
30	Veracruz	217,589	5.7	214,368	5.7	51,529	5.6	50,767	5.6	30	3.7
8	Chihuahua	200,548	5.2	189,873	5.1	52,113	5.7	49,522	5.5	25	3.1
31	Yucatán	188,237	4.9	184,655	4.9	49,792	5.4	48,834	5.4	55	6.8
14	Jalisco	69,239	1.8	65,276	1.7	18,461	2.0	17,551	2.0	16	2.0
10	Durango	64,344	1.7	62,523	1.7	13,972	1.5	13,529	1.5	8	1.0
29	Tlaxcala	62,942	1.6	61,369	1.6	14,760	1.6	14,427	1.6	13	1.6
26	Sonora	57,143	1.5	55,297	1.5	17,973	2.0	17,448	1.9	38	4.7
24	S. L. Potosí	45,367	1.2	44,258	1.2	11,317	1.2	11,044	1.2	5	0.6
18	Nayarit	42,514	1.1	42,232	1.1	7,882	0.9	7,827	0.9	1	0.1
32	Zacatecas	37,824	1.0	37,355	1.0	10,769	1.2	10,640	1.2	13	1.6
15	México	34,145	0.9	34,118	0.9	8,172	0.9	8,167	0.9	5	0.6
19	Nuevo León	28,928	0.8	27,111	0.7	9,073	1.0	8,541	1.0	14	1.7
13	Hidalgo	25,527	0.7	23,750	0.6	6,222	0.7	5,821	0.6	3	0.4
5	Coahuila	18,085	0.5	17,872	0.5	5,332	0.6	5,276	0.6	9	1.1
28	Tamaulipas	18,458	0.5	17,503	0.5	5,290	0.6	4,996	0.6	7	0.9
16	Michoacán	11,450	0.3	11,396	0.3	3,078	0.3	3,066	0.3	3	0.4
11	Guanajuato	5,261	0.1	5,261	0.1	1,340	0.1	1,340	0.1	1	0.1
	Total	3,842,834	100.0	3,737,259	100.0	921,516	100.0	898,962	100.0	*810	100.0

*No se logró todo el censo en 1 municipio de Oaxaca y 3 de Chihuahua

A pesar de que estas localidades fueron censadas y en teoría su factor de expansión debería ser uno (están autorepresentadas), al estar contenidas en una base de datos de la muestra intercensal, se consideran pesos muestrales distintos, ya que se aplicó de forma general una corrección en el factor de expansión en cada estrato por la no respuesta atribuida al informante. En este trabajo se aisló esta población censada de la muestra nacional, por lo que la información recabada se analizó sin los pesos muestrales.

2.2 Población de interés para el modelo de regresión logística

El modelo de regresión logística ajustado a la población solo consideró al conjunto de observaciones que están sujetas a la pregunta sobre el ingreso recibido según el cuestionario del INEGI, así como a siete variables con las que se ajustó el modelo.

Consideración sobre la respuesta en la variable Ingreso

La variable binaria (Y=$\{0,1\}$) distingue a los individuos que no dieron respuesta a la pregunta *¿Cuánto gana (NOMBRE) por ese trabajo?* de los que si dieron información sobre su ingreso en el último mes (Los valores que aparecen codificados *NA* para la variable

INGTRMEN en la base de datos son los que no respondieron).

Como primer paso, se seleccionaron aquellas observaciones que están sujetas a contestar la pregunta en esa instancia, caracterizada según el cuestionario del encuestador como: *población ocupada de 12 y más años que obtiene o recibe del(los) trabajo(s) que desempeñó en la semana de referencia.* Por lo tanto, se definió a las personas que reciben ingreso a aquellas observaciones con las siguientes características en la base de datos:

- *Edad.* Mayor a 12 años cumplidos y menor a 90 años cumplidos.
- *Situación en el trabajo.* Aquellos que sean Empleado(a) u obrero(a), Jornalero(a) o peón(a), Ayudante con pago, Patrón(a) o empleador(a) o Trabajador(a) por cuenta propia.
- *Condición de actividad.* La semana pasada Trabajó, Hizo o vendió algún producto, Ayudó en algún negocio, Crio animales o cultivó algo, Ofreció algún servicio por un pago, Atendió su propio negocio, Tenía trabajo, pero no trabajó.

Variables explicativas seleccionadas para el modelo de regresión logística

En la Encuesta Intercensal 2015 se tienen un total de 86 variables en el cuestionario de personas, se encuentran organizadas en los subtemas de Población Total y Estructura, Fecundidad y Mortalidad, Migración, Movilidad Cotidiana, Etnicidad, Servicios de Salud, Educación, Características Económicas y Trabajo no Remunerado.

Se consideraron siete variables como variables explicativas para el modelo. Algunas variables se recategorizaron para reducir el número de sus categorías, también algunas se crearon a partir de otras variables originales con el fin de dar categorías más complejas, se eliminaron casos donde no hubiera respuesta a alguna de estas variables.

Las variables y su descripción están presentadas en las siguiente *Tabla 6*:

Tabla 6: Variables seleccionadas para el modelo de regresión logística

Nombre	Tipo	Descripción	Niveles	Notación
Ingreso	Creada de *Ingresos por trabajo*	Dio o no respuesta al ingreso percibido en el último mes	Si, No	*INGTRMEN_RESP*
Sexo	Original	Condición biológica de nacimiento	Hombre*, Mujer	*SEXO*
Edad	Original, recategorizada	Años cumplidos de la persona	12-19*, 20-29, 30-59, 60+	*EDAD_r*
Nivel educativo	Creada de *Escolaridad* y *Nivel de Escolaridad*	Grado máximo de estudios obtenido al momento	Ninguno o Primaria Incompleta*, Primaria completa, Secundaria Completa y Bachillerato Completo o más	*NIVACAD_r*
Situación Laboral	Original, recategorizada	Relación de propiedad con el negocio, empresa o establecimiento en el que trabaja	Trabajador por cuenta propia*, Empleado, obrero o asistente, Jornalero o Peón y Patrón o Empleador	*SITUACION_TRAB_r*
Lengua Indígena	Original	Hablar o no alguna lengua indígena	Sí*, No	*HLENGUA*
Situación Conyugal	Original, recategorizada	Personas unidas o no en matrimonio	No casado*, Si casado	*SITUA_CONYUGAL_r*
Entidad	Original	Entidad geográfica donde reside	Oaxaca∗, Chiapas, Guerrero, Puebla, Veracruz, Chihuahua, Yucatán, Jalisco, Durango, Tlaxcala, Sonora, San Luis Potosí, Nayarit, Zacatecas, México, Nuevo León, Hidalgo, Coahuila, Tamaulipas, Michoacán y Guanajuato	*ENT*

*Categoría de referencia

Casos especiales eliminados

Para eliminar posibles errores de captura se analizaron casos en los cuales el ingreso reportado no corresponde con su situación laboral establecida, eliminando 23 casos en los cuales el ingreso de un Ayudante, jornalero o peón era mayor al cuantil 0.99 ($38,571) de ingreso en toda la población.

En segunda instancia se eliminaron 92 observaciones donde el ingreso era mayor a $50,000 y la edad menor a 20 años o el nivel educativo era menor a primaria completa; si su situación

laboral era empleado, trabajador por cuenta propia, empleador o patrón, entonces que su educación fuera de bachillerato completo o haya establecido su ocupación como funcionario gubernamental. Esto después de analizar las observaciones.

Tabla 7: Número de observaciones en cada etapa de selección de la población de interés

Etapa	Total
Municipios Cobertura Completa	3,737,259
Edad entre 12 y 90 años, situación laboral y condición de actividad con ingreso definido	688,669
Casos Completos en las 7 variables de interés	686,358
Eliminando casos atípicos donde el ingreso no corresponde con su situación laboral, Escolar u Ocupación	686,243

La *Tabla 7* contiene el número de observaciones existentes en cada conjunto descrito y como se va reduciendo a medida que se define a mayor detalle la población de interés. El total final de esta población (686,243 observaciones) corresponde al 18.36 % de toda la población que fue censada en primera instancia.

Tabla 8: Distribución de la población de interés por estados

	Observaciones		UPMs		Estratos		Municipios	
Entidad	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**
Oaxaca	207,994	30.3	14,279	32.9	1,361	47.4	441	54.4
Chiapas	97,993	14.3	4,811	11.1	210	7.3	32	4.0
Guerrero	69,133	10.1	4,296	9.9	170	5.9	26	3.2
Yucatán	60,480	8.8	2,985	6.9	156	5.4	55	6.8
Puebla	55,480	8.1	3,101	7.1	239	8.3	65	8.0
Veracruz	43,359	6.3	1,751	4.0	127	4.4	30	3.7
Chihuahua	36,170	5.3	3,440	7.9	103	3.6	25	3.1
Tlaxcala	20,477	3.0	659	1.5	37	1.3	13	1.6
Sonora	17,804	2.6	1,728	4.0	111	3.9	38	4.7
Jalisco	15,241	2.2	1,181	2.7	77	2.6	16	2.0
S. Luis Potosí	9,590	1.4	643	1.5	24	0.8	5	0.6
México	9,391	1.4	307	0.7	25	0.9	5	0.6
Nuevo León	7,953	1.2	828	1.9	43	1.5	14	1.7
Zacatecas	7,575	1.1	718	1.7	61	2.1	13	1.6
Durango	6,859	1.0	902	2.1	35	1.2	8	1.0
Coahuila	5,623	0.8	493	1.1	25	0.9	9	1.1
Hidalgo	5,092	0.7	271	0.6	13	0.5	3	0.4
Tamaulipas	3,501	0.5	440	1.0	30	1.0	7	0.9
Nayarit	2,726	0.4	352	0.8	7	0.2	1	0.1
Michoacán	2,538	0.4	192	0.4	14	0.5	3	0.4
Guanajuato	1,264	0.2	82	0.2	5	0.2	1	0.1
Total	686,243	100.0	43,459	100.0	2,873	100.0	810	100

La *Tabla 8* muestra que la distribución por estados es muy desigual, por ejemplo, Oaxaca tiene casi un tercio del total las observaciones (30.3 %), además de que esta entidad cuenta con casi el 50 % de los estratos. Los 6 estados con menos observaciones representan menos del 1 % cada uno en la población.

2.3 Análisis exploratorio de las variables en la población de interés

Definida la población de interés (686,243 observaciones) se analizó el ingreso reportado, este análisis exploratorio da una perspectiva de la distribución del ingreso y su respuesta en función de las variables explicativas, esto permitió un mejor proceso de selección del modelo de regresión logística.

Tabla 9: Proporción observada de la respuesta en Ingreso según variables explicativas en población de interés N=689,243

Variable	Categoría	No respondió	%	Respondió	%	Total	%
Sexo	Hombre	107,847	20.8	410,109	79.2	517,952	75.5
	Mujer	17,102	10.2	151,185	89.8	168,287	24.5
Edad	12-19	10,268	18.9	43,918	81.1	54,186	7.9
	20-29	27,872	15.4	153,063	84.6	180,935	26.4
	30-59	66,179	17.4	314,253	82.6	380,432	55.4
	60+	20,630	29.2	50,060	70.8	70,690	10.3
Nivel Educativo	Nin/Prim Incomp	49,851	24.6	152,451	75.4	202,302	29.5
	Prim Comp	35,188	19.7	143,174	80.3	178,362	26.0
	Sec Comp	26,104	15.2	145,740	84.8	171,844	25.0
	Bach Comp y más	13,806	10.3	119,929	89.7	133,735	19.5
Situación Laboral	Cuenta Propia	93,962	34.4	179,357	65.6	273,319	39.8
	Empleado/Asist	19,812	6.4	291,965	93.6	311,777	45.4
	Jornalero/Peón	8,738	10.2	76,934	89.8	85,672	12.5
	Patrón/Empleador	2,437	15.7	13,038	84.3	15,475	2.3
Situación Conyugal	Casado	57,020	19.1	241,390	80.9	298,410	43.5
	No Casado	67,929	17.5	319,904	82.5	387,833	56.5
Lengua indígena	Si	73,437	23.4	240,526	76.6	313,963	45.8
	No	51,512	13.8	320,768	86.2	372,280	54.2
Entidad	Oaxaca	46,773	22.5	161,221	77.5	207,994	30.3
	Chiapas	24,659	25.2	73,334	74.8	97,993	14.3
	Guerrero	15,335	22.2	53,798	77.8	69,133	10.1
	Yucatán	5,662	9.4	54,818	90.6	60,480	8.8
	Puebla	8,052	14.5	47,428	85.5	55,480	8.1
	Veracruz	5,003	11.5	38,356	88.5	43,359	6.3
	Chihuahua	5,648	15.6	30,522	84.4	36,170	5.3
	Tlaxcala	2,282	11.1	18,195	88.9	20,477	3.0
	Sonora	1,347	7.6	16,457	92.4	17,804	2.6
	Jalisco	1,801	11.8	13,440	88.2	15,241	2.2
	San Luis Potosí	921	9.6	8,669	90.4	9,590	1.4
	México	907	9.7	8,484	90.3	9,391	1.4
	Nuevo León	803	10.1	7,150	89.9	7,953	1.2
	Zacatecas	1,018	13.4	6,557	86.6	7,575	1.1
	Durango	1,384	20.2	5,475	79.8	6,859	1.0
	Coahuila	437	7.8	5,186	92.2	5,623	0.8
	Hidalgo	1,523	29.9	3,569	70.1	5,092	0.7
	Tamaulipas	477	13.6	3,024	86.4	3,501	0.5
	Nayarit	352	12.9	2,374	87.1	2,726	0.4
	Michoacán	464	18.3	2,074	81.7	2,538	0.4
	Guanajuato	101	8.0	1,163	92.0	1,264	0.2
Total		**124,949**	**18.2**	**561,294**	**81.8**	**686,243**	**100.0**

Gráfico 1. Proporción observada en respuesta Ingreso en Entidad

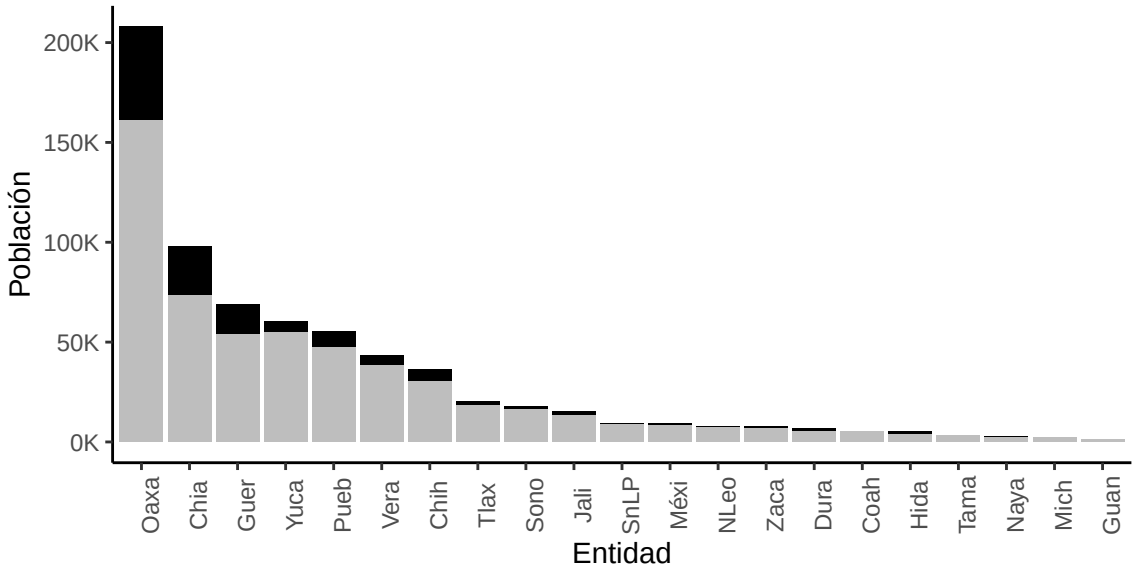


Gráfico 2. Proporción observada respuesta Ingreso en Sexo

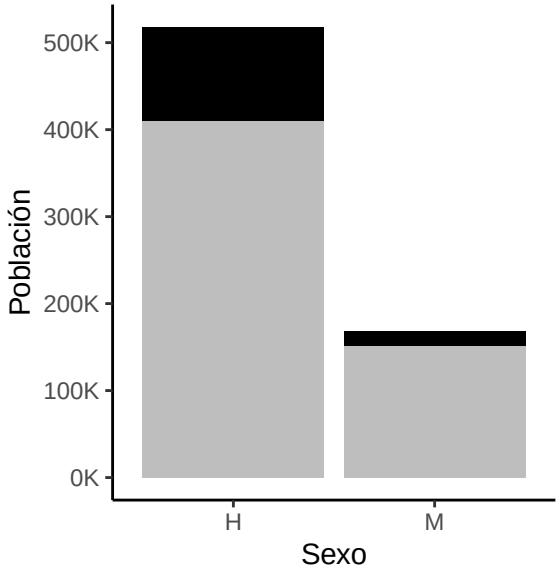
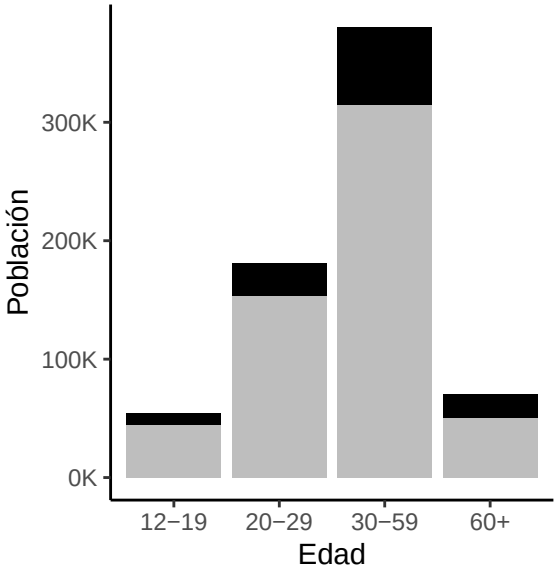
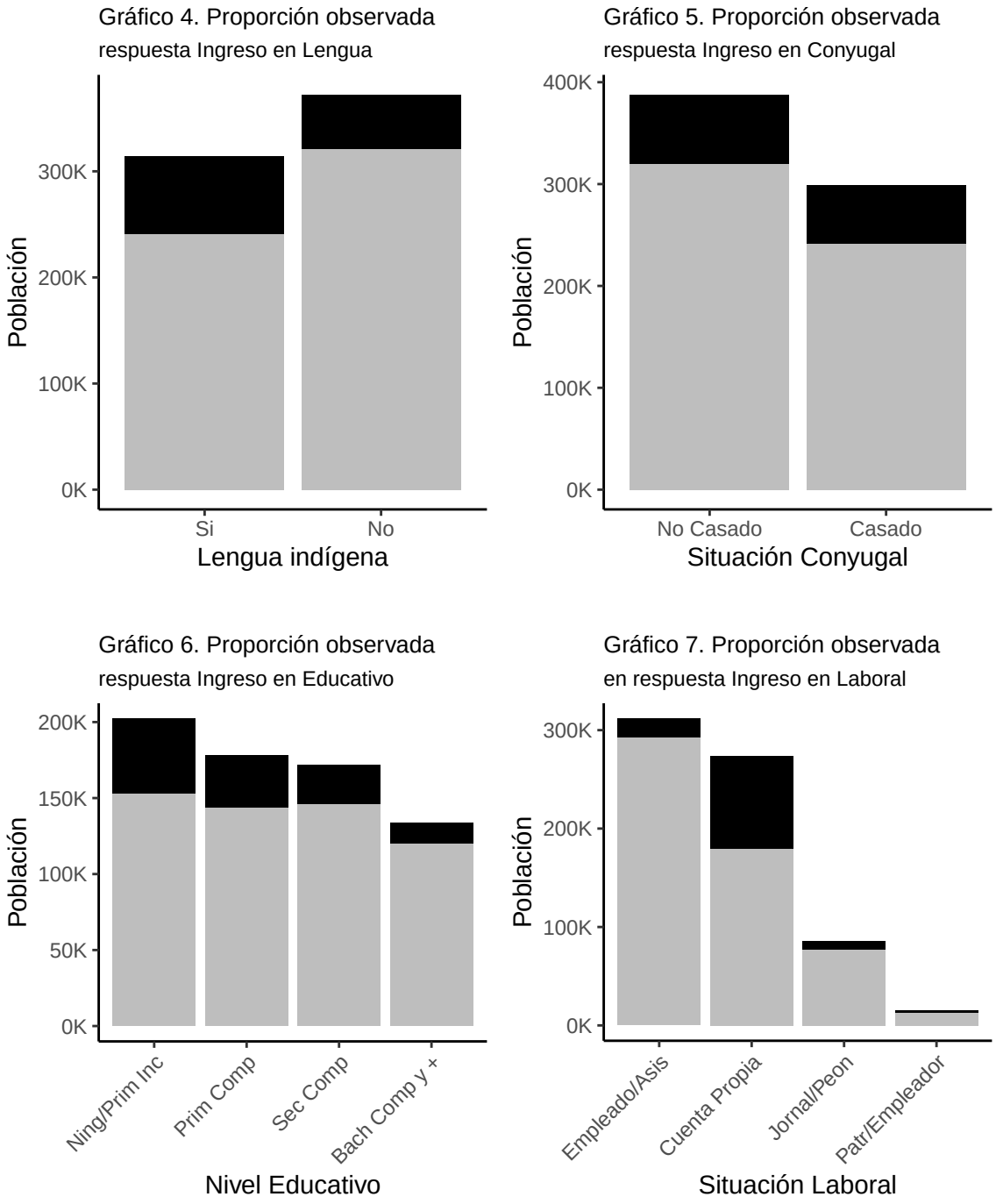


Gráfico 3. Proporción observada respuesta Ingreso en Edad





La *Tabla 9* muestra la proporción del total por cada categoría, así como la proporción de cada categoría con la sí respuesta (Y=1) y no respuesta (Y=0) en la variable Ingreso. Con la ayuda visual del *Gráfico 2* se observa una distribución desigual en la variable Sexo dado que más del 75 % corresponde a Hombre; de igual forma en la variable Edad, donde existen 4 categorías, más de 55 % de las observaciones corresponden 30-59 y 12-19 representa menos del 8 %. La variable Situación Laboral tiene 2.3 % de las observaciones en Patrón/Empleador, un número pequeño a comparación de las otras categorías. Hay desproporción dentro de las 21 categorías de la variable Entidad como se observa en el

Gráfico 1, Oaxaca tiene una gran mayoría con más del 30 % de las observaciones. En las restantes tres variables (Nivel Educativo, Lengua indígena y Situación Conyugal) se observa mayor igualdad en la distribución de las observaciones por categorías, como se observa en los *Gráficos* 4, 5 y 6.

Del total de observaciones, 561,294 (81.8 %) corresponden a si respuesta en la variable Ingreso y 124,949 (18.2 %) a no respuesta. Esta tendencia se sigue en cada categoría, con Cuenta Propia de la variable Situación Laboral como aquella que tiene el valor más bajo en si respuesta (65.6 %).

Considerando a las observaciones que si respondieron en la variable Ingreso (561,294) se analiza el ingreso reportado.

Tabla 10: Información de la cantidad de ingreso ($)

Mín.	1er cuartil	Mediana	Media	3er cuartil	Max.
0	1,500	2,571	3,498	4,286	999,998

Tabla 11: Distribución de las observaciones por ingreso

ingreso ($)	**Total**	**%**
$[0 - 1,500]$	148,138	26.39
$(1,500 - 3,000]$	175,786	31.32
$(3,000 - 4,500]$	111,447	19.86
$(4,500 - 10,000]$	109,831	19.57
$(10,000 - 50,000]$	15,940	2.84
$(50,000 - 999,999)$	152	0.03
Total	**561,294**	**100.0**

Gráfico 8. Distribución de ingreso [0−50,000] (cuantil 0−0.97)

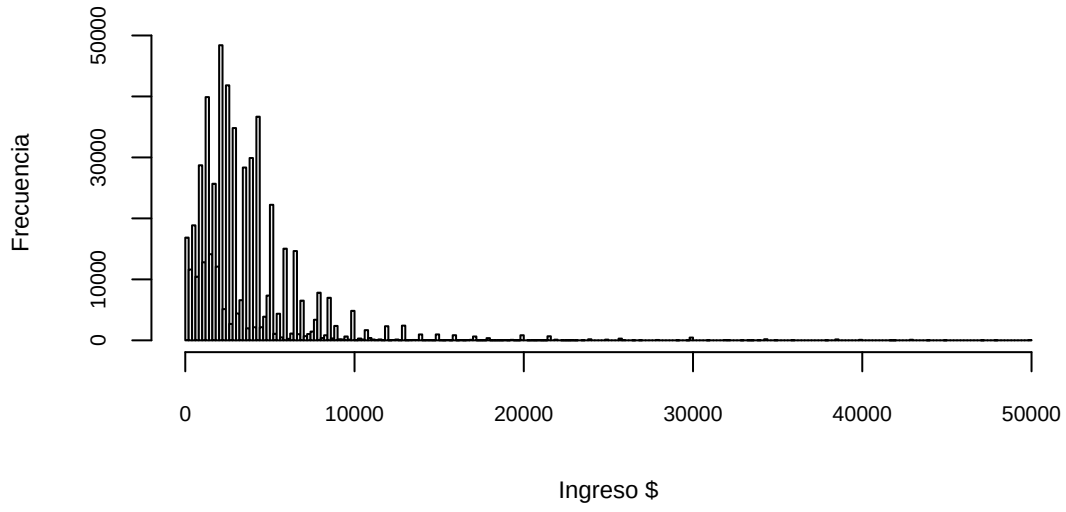


Gráfico 9. Distribución de ingreso [50,000−999,998] (cuantil 0.97−1)

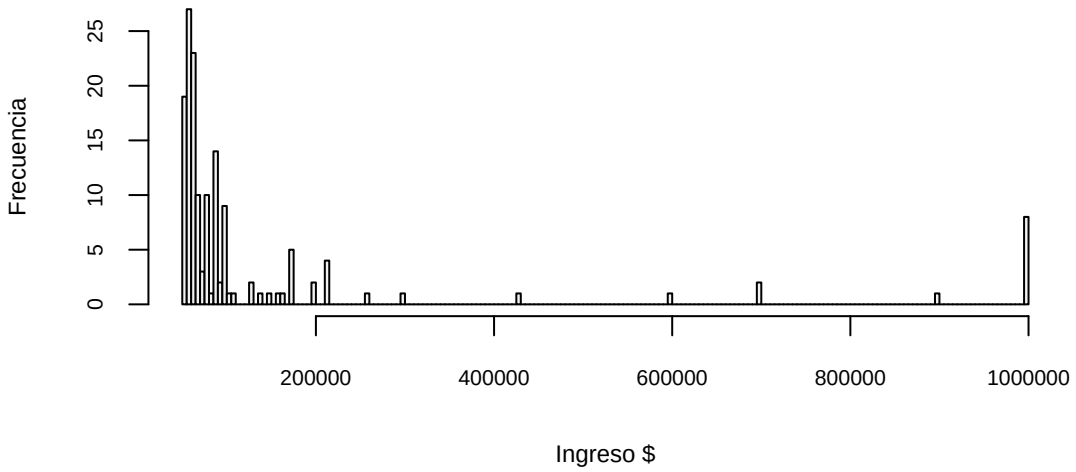
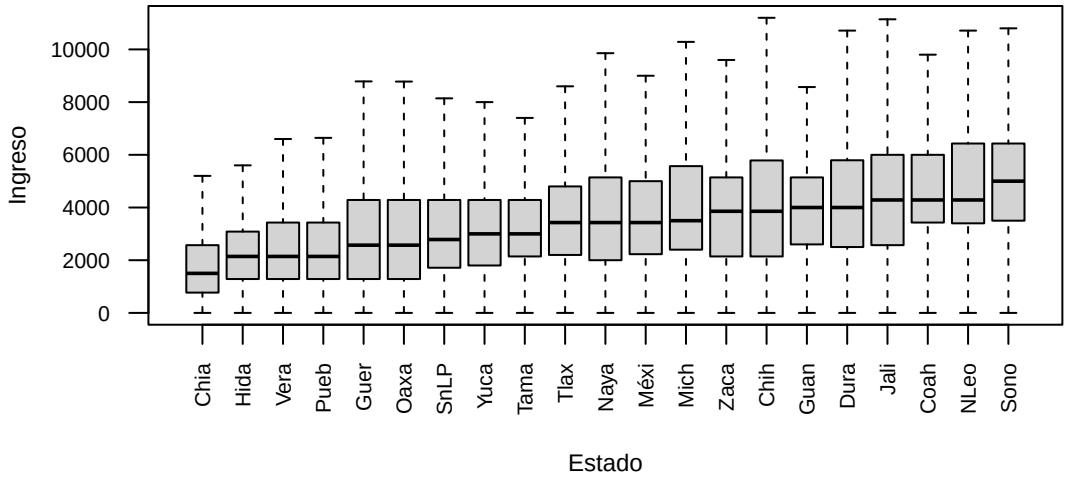


Gráfico 10. Ingreso Promedio por Estado



El ingreso promedio es de $3,498$ pesos al mes. En la *Tabla 11* se observa que el $97.19\,\%$ de la población reporta recibir menos de $10,000$ pesos al mes. Existen 152 observaciones con valores extremos que reportan recibir más de $50,000$ y hasta $999,998$ pesos al mes, aún después de haber eliminado casos atípicos en la *Sección 2.2*, estas observaciones se tomaron en cuenta dado que existen municipios que fueron censados por tener menos de 1,300 viviendas y no necesariamente por ser vulnerables. En el *Gráfico 10* se observa que hay diferencias en los ingresos por entidad geográfica con Chiapas, Veracruz, Hidalgo y Puebla en los que menos se recibió con un promedio de $2,556$ pesos mensuales y Sonora

donde más con un ingreso promedio de $5,948$ pesos al mes.

Una vez definida la población de interés y descrito las variables explicativas y la variable respuesta se realizaron procesos de selección de modelo de regresión logística para predecir la probabilidad de la sí respuesta $Y = 1$ en la variable Ingreso y la relación con las siete variables explicativas, además de analizar las implicaciones que tienen las 686,243 observaciones.

Capítulo 3. Ajuste de un modelo de regresión logística a la población de interés

En la primera parte de este capítulo se presenta la aplicación de métodos automatizados de selección de modelos en la base de datos de la población de interés. Todos los modelos seleccionados en este capítulo supusieron que las observaciones de la variable respuesta son independientes y que la población de interés es una base de datos de una muestra, esto para que tenga sentido los ajustes de la estimación de varianza.

De las variables explicativas, la única variable continua en la encuesta era Edad, que al ser categorizada en grupos se convirtió en discreta. Esto permitió trasladar los datos a una tabla de contingencia, esto es, considerar los datos de forma agrupada, con el fin de facilitar computacionalmente el proceso de selección del modelo. Los valores de las estimaciones de los coeficientes de la regresión son los mismos que con los datos no agrupados porque es la misma información.

El número de registros del archivo de datos original se redujo de 686,243 filas a 21,504 en el archivo correspondiente a la tabla de contingencia. Esta tabla se compone con los registros derivados del producto [Sexo (2) × Edad (4) × Nivel Educativo (4) × Situación Laboral (4) × Situación Conyugal (2) × Lengua Indígena (2) × Estado (21) × respuesta Ingreso (2) = 21, 504]. La dimensión de la tabla de contingencia es de 21,504 filas y 9 columnas de las 8 variables y la asignación de frecuencias. Para comprender la estructura de la tabla de contingencia, se muestran las primeras y últimas celdas en el *Cuadro 1.*

Cuadro 1. Estructura de la tabla de contingencia de la población de interés (BaseDT)

> head(BaseDT)

	SEXO	EDAD_r	SITUACION_TRAB_r	NIVACAD_r	HLENGUA	SITUA_CONYUGAL_r	ENT	INGTRMEN_RESP	Freq
1	H	12-19	Cuenta Propia	Ning/Prim Inc	Si	No Casado	20	0	159
2	M	12-19	Cuenta Propia	Ning/Prim Inc	Si	No Casado	20	0	21
3	H	20-29	Cuenta Propia	Ning/Prim Inc	Si	No Casado	20	0	418
4	M	20-29	Cuenta Propia	Ning/Prim Inc	Si	No Casado	20	0	47
5	H	30-59	Cuenta Propia	Ning/Prim Inc	Si	No Casado	20	0	1694

> tail(BaseDT)

	SEXO	EDAD_r	SITUACION_TRAB_r	NIVACAD_r	HLENGUA	SITUA_CONYUGAL_r	ENT	INGTRMEN_RESP	Freq
21500	M	20-29	Patr/Empleador	zBach Comp y +	No	Casado	32	1	3
21501	H	30-59	Patr/Empleador	zBach Comp y +	No	Casado	32	1	37
21502	M	30-59	Patr/Empleador	zBach Comp y +	No	Casado	32	1	10
21503	H	60+	Patr/Empleador	zBach Comp y +	No	Casado	32	1	3
21504	M	60+	Patr/Empleador	zBach Comp y +	No	Casado	32	1	0

Tabla 12: Número de celdas en la tabla de contingencia de la población de interés

Frecuencia	Total	%
Diferente de cero	10,474	48.7
Cero	11,030	51.3
Total	21,504	100.0

La tabla de contingencia resultó con 11,030 celdas con frecuencia cero como se observa en la *Tabla 12*; un ejemplo en particular es la última fila del *Cuadro 1* donde la frecuencia es cero. Esto es debido a lo especifico en los cruces de cada celda y la naturaleza de las variables, considerando que algunas categorías tienen un bajo porcentaje de observaciones (ver *Tabla 9*), por ejemplo, la variable Estado al ser de carácter geográfico y tener 21 categorías, algunas tienen muy pocas observaciones del total (menos del 1 %) como es el caso de Guanajuato, esto hace que los cruces tengan frecuencia cero.

A continuación, se describe el proceso de selección del modelo de regresión logística a la población de interés (`BaseDT`). Se utilizaron las funciones para modelos lineales generalizados (`glm()`) en el software *R*. Se empezó con el modelo aditivo simple, todas las variables explicativas seleccionadas de forma aditiva y sin considerar interacciones entre ellas.

Cuadro 2. Ajuste del modelo aditivo simple (m1) a la población de interés

```
> m1<- glm(INGTRMEN_RESP ~ SEXO+ EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA
+ SITUA_CONYUGAL_r + ENT, family=binomial,weight=Freq ,data=BaseDT)

> formula(m1)

 INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r
 + NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r

> drop1(m1, test="Chisq", k=log(686243))

                Df Deviance    BIC   LRT  Pr(>Chi)
<none>               549778 550222
ENT             20   555581 555756  5803 < 2.2e-16 ***
SEXO             1   559078 559508  9299 < 2.2e-16 ***
EDAD_r           3   550724 551127   946 < 2.2e-16 ***
SITUACION_TRAB_r 3   608979 609382 59201 < 2.2e-16 ***
NIVACAD_r        3   550585 550988   807 < 2.2e-16 ***
HLENGUA          1   551211 551641  1432 < 2.2e-16 ***
SITUA_CONYUGAL_r 1   549834 550264    55 1.003e-13 ***
```

Todas las variables son estadísticamente significativas bajo la prueba de hipótesis jí-cuadrada χ^2 con α=0.05 como se observa en el *Cuadro 2*. Por lo tanto, se concluyó que todas las

variables son significativas para el ajuste de la respuesta en la variable Ingreso.

Posteriormente se buscó un modelo más complejo que el aditivo, se utilizó la función `step()` en R, este cuenta con tres opciones para realizar la revisión de la significancia de las variables y sus interacciones: *Backward, Forward y Backward-Forward*. Se aplicaron los tres algoritmos en la base de datos y el tercero (*Backward-forward procedure*) mostró ser el más efectivo, ya que permite una revisión de la significancia añadiendo y quitando las variables en varias etapas.

La bondad de ajuste en la función `step()` se consideró y cálculo con dos criterios: *AIC* y *BIC*. Se optó por darle importancia a este último, ya que el *BIC* penaliza más el uso de variables e interacciones que tienen poca relevancia en el modelo considerando un número de observaciones grande.

En conclusión, se revisaron las combinaciones de obtención de modelo, con los tres algoritmos de *step* y los dos criterios, pero solo se presentan los modelos obtenidos con el proceso *Backward-forward* y el criterio *BIC* por las razones mencionadas anteriormente.

Ejemplo 3. Tabla de contingencia creada de interacciones entre variables

Cuando se habla de variables categóricas, las interacciones pueden ser expresadas como una tabla de contingencia que aumentará de tamaño conforme se agreguen interacciones.

Considerando las variables Sexo (S) con categorías Hombre(0) y Mujer (1), la variable Situación Conyugal (C) con categorías No Casado (0) y Casado (1) y Lengua indígena (L) con categorías Si(0) y No(1) en la población censal, la tabla de contingencia generada por las interacciones entre ellas se encuentra en la *Tabla 13.*

Tabla 13: Frecuencia de las celdas en la tabla de contingencia de las interacciones para el Ejemplo 3

Grado Interacción	Variable	Categorías	Frecuencia	Total
Sin interacción	S	S(0)	517,956	686,243
		S(1)	168,287	
	L	L(0)	313,963	686,243
		L(1)	372,280	
	C	C(1)	298,410	686,243
		C(0)	387,833	
1er orden	S×L	S(0)×L(0)	246,472	686,243
		S(0)×L(1)	271,484	
		S(1)×L(0)	67,491	
		S(1)×L(1)	100,796	
	S×C	S(0)×C(0)	273,738	686,243
		S(0)×C(1)	244,218	
		S(1)×C(0)	114,095	
		S(1)×C(1)	54,192	
	L×C	L(0)×C(0)	182,988	686,243
		L(0)×C(1)	130,975	
		L(1)×C(0)	204,845	
		L(1)×C(1)	167,435	
2do orden	S×L×C	S(0)×L(0)×C(0)	135,739	686,243
		S(0)×L(0)×C(1)	110,733	
		S(0)×L(1))×C(0)	137,999	
		S(0)×L(1)×C(1)	133,485	
		S(1)×L(0)×C(0)	47,249	
		S(1)×L(0)×C(1)	20,242	
		S(1)×L(1)×C(0)	66,846	
		S(1)×L(1)×C(1)	33,950	

Como se observa, la frecuencia de las celdas disminuye cuando se agregan interacciones. La función `step()` en *R* considera cada una de las celdas con interacciones hasta el orden que se le indique. Cuando el número de variables, categorías e interacciones aumenta es evidente que se aumenta la posibilidad de celdas con frecuencia cero.

Es importante notar que las interacciones de variables crean una nueva tabla de contingencia dentro del algoritmo (ver *Ejemplo 3*), en ésta se agregan las combinaciones de las variables explicativas. Por lo tanto, es evidente que las interacciones de la variable Entidad con otras variables causa muchas celdas frecuencia cero, debido al número de categorías con pocas observaciones. Los procesos *step* tomando en cuenta las interacciones de la variable Entidad con las demás variables fueron efectuados, pero dada la poca significancia y el aumento de tiempo computacional, se decidió tomarla solo como efecto aditivo.

Con estas especificaciones, se seleccionaron las interacciones de primer orden (variables que involucran la relación dos a dos de las variables explicativas).

Cuadro 3. Algoritmo de selección y ajuste de modelo (m2BIC) a la población de interés

> m2BIC<- step(glm(INGTRMEN_RESP~ ENT + (SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA+ SITUA_CONYUGAL_r)^2, family=binomial,data=BaseDT, weight=Freq), direction="both", k=log(686243))

> formula(m2BIC)

```
INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA +
SITUA_CONYUGAL_r + SEXO:EDAD_r + SEXO:SITUACION_TRAB_r + SEXO:NIVACAD_r +
SEXO:HLENGUA + EDAD_r:SITUACION_TRAB_r + EDAD_r:NIVACAD_r + EDAD_r:SITUA_CONYUGAL_r
SITUACION_TRAB_r:NIVACAD_r + SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA +
HLENGUA:SITUA_CONYUGAL_r
```

> drop1(m2BIC,test="Chisq",k=log(686243))

	Df	Deviance	BIC	LRT	Pr(>Chi)	
<none>		543573	544648			
ENT	20	549345	550152	5772.1	< 2.2e-16	***
SEXO:EDAD_r	3	543640	544675	67.0	1.913e-14	***
SEXO:SITUACION_TRAB_r	3	545950	546985	2376.7	< 2.2e-16	***
SEXO:NIVACAD_r	3	543697	544732	123.9	< 2.2e-16	***
SEXO:HLENGUA	1	543589	544651	15.7	7.400e-05	***
EDAD_r:SITUACION_TRAB_r	9	543975	544929	401.4	< 2.2e-16	***
EDAD_r:NIVACAD_r	9	543790	544744	216.4	< 2.2e-16	***
EDAD_r:SITUA_CONYUGAL_r	3	543719	544754	146.1	< 2.2e-16	***
SITUACION_TRAB_r:NIVACAD_r	9	544009	544963	435.6	< 2.2e-16	***
SITUACION_TRAB_r:HLENGUA	3	544197	545232	623.6	< 2.2e-16	***
NIVACAD_r:HLENGUA	3	543750	544785	176.9	< 2.2e-16	***
HLENGUA:SITUA_CONYUGAL_r	1	543627	544688	53.4	2.654e-13	***

El resultado mostrado en el *Cuadro 3* son las interacciones seleccionadas y su significancia estadística. Es un modelo que mejora con respecto al modelo aditivo simple *m1* bajo el criterio *BIC*.

Del mismo modo se procedió a seleccionar las interacciones de segundo orden (relaciones tres a tres y menores de las variables explicativas).

Cuadro 4. Algoritmo de selección y ajuste de modelo (m3BIC) a la población de interés

> m3BIC <- step(glm(INGTRMEN_RESP~ ENT + (SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r)^3, family=binomial, data=BaseDT, weight=Freq), direction="both", k=log(686243)

> formula(m3BIC)

```
  INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r +
  HLENGUA + SITUA_CONYUGAL_r + SEXO:EDAD_r + SEXO:SITUACION_TRAB_r +
  SEXO:NIVACAD_r + SEXO:HLENGUA + SEXO:SITUA_CONYUGAL_r + EDAD_r:SITUACION_TRAB_r +
  EDAD_r:NIVACAD_r + EDAD_r:SITUA_CONYUGAL_r + SITUACION_TRAB_r:NIVACAD_r +
  SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA + HLENGUA:SITUA_CONYUGAL_r +
  SEXO:NIVACAD_r:HLENGUA + SEXO:HLENGUA:SITUA_CONYUGAL_r
```

> drop1(m3BIC,test="Chisq", k=log(686243))

```
                              Df Deviance    BIC    LRT  Pr(>Chi)
<none>                             543347 544489
ENT                           20   549123 549996 5775.9 < 2.2e-16 ***
SEXO:EDAD_r                    3   543418 544520   71.0 2.638e-15 ***
SEXO:SITUACION_TRAB_r          3   545720 546822 2372.8 < 2.2e-16 ***
EDAD_r:SITUACION_TRAB_r        9   543746 544768  399.5 < 2.2e-16 ***
EDAD_r:NIVACAD_r               9   543568 544590  221.4 < 2.2e-16 ***
EDAD_r:SITUA_CONYUGAL_r        3   543497 544599  150.6 < 2.2e-16 ***
SITUACION_TRAB_r:NIVACAD_r     9   543793 544814  446.0 < 2.2e-16 ***
SITUACION_TRAB_r:HLENGUA       3   543924 545026  577.1 < 2.2e-16 ***
SEXO:NIVACAD_r:HLENGUA         3   543456 544558  109.1 < 2.2e-16 ***
SEXO:HLENGUA:SITUA_CONYUGAL_r  1   543471 544600  124.1 < 2.2e-16 ***
```

El *Cuadro 4* muestra las interacciones seleccionadas y su significancia estadística. Este modelo mejoró con respecto a los modelos *m1* y *m2BIC*, bajo el criterio *BIC*.

Debido a limitaciones computacionales, no se evaluaron interacciones de mayor grado ni fue posible analizar el modelo saturado para comparar sus efectos con los otros modelos encontrados.

Los modelos están anidados, es decir, los modelos más complejos contienen variables e interacciones del anterior. Se aplicó una prueba de hipótesis para ver la significancia del modelo con más variables e interacciones con respecto a los que tienen menos.

$$H_0 : dev_1 - dev_2 = 0 \quad vs \quad H_1 : dev_1 - dev_2 \neq 0.$$

Cuadro 5. Prueba de hipótesis para modelos anidados, ajustados a la población de interés

```
#m1. Modelo Aditivo Simple
#m2BIC. Modelo interacciones de dos, seleccionado en step() con BIC
#m3BIC. Modelo interacciones de tres, seleccionado en step() con BIC
```

> anova(m1,m2BIC, test="Chisq")

```
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     10441     549778
2     10394     543573 47     6205 < 2.2e-16 ***
```

```
> anova(m2BIC,m3BIC, test="Chisq")

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     10394     543573
2     10389     543347  5   226.42 < 2.2e-16 ***
```

Como se observa en el *Cuadro 5*, en este caso al comparar el modelo *m1* con el *m2BIC* el valor reportado de p fue menor a 0.05, de igual forma al comparar el modelo *m2BIC* con el modelo *m3BIC*. Por lo tanto, se consideró que el desempeño del modelo con más interacciones es estadísticamente mejor que el que tiene menos. La comparación de *AIC*, *BIC* y devianza se encuentra en el *anexo a*.

Se seleccionó el modelo *m3BIC* ya que con la prueba de hipótesis de la devianza y bajo el *BIC* tiene el mejor ajuste. Sin embargo, al tratar con una base de datos de 686,243 entradas se presentaron problemas, como que un mejor análisis quedó rezagado debido a la carga y tiempo computacional.

Además, como se explica en Fitzmaurice et al. (1997), una gran cantidad de datos puede ocasionar sobredispersión, esto se refiere a cuando los datos muestran una mayor variabilidad que la predicha por la relación varianza-media al hacer inferencia estadística sobre ellos. Cuando esto ocurre, las estrategias de selección de modelos basadas en pruebas de criterio de información (como el *AIC* o *BIC*) se ejecutarán pobremente. Si se ignora esto, un modelo con muchos parámetros puede ser seleccionado ya que se tiene un mal cálculo de la varianza de los estimadores.

Con grandes cantidades de datos, el uso directo de métodos estadísticos estándar tiende a producir estimaciones de aparente muy alta precisión, esencialmente debido a fuertes suposiciones de dependencia que son débiles bajo estos métodos (Véase más en Cox (2015)). Igualmente, si la cantidad de datos es de cientos de miles o más, los procesos de selección de modelo son tardados y es difícil analizar toda la base, como se comenta en Hastie, T. (2018), realizar el análisis en una muestra aleatoria simple de las observaciones es aconsejable para calcular los valores estimados.

3.1 Selección de una muestra aleatoria simple de la población de interés

Esta parte continúa con la selección de una muestra aleatoria simple del 20 % de la población de interés y la aplicación de métodos automatizados de selección de modelo de regresión logística en ella. De igual forma, los modelos seleccionados supusieron que las observaciones de la variable respuesta son independientes y no se consideró el diseño muestral de la encuesta.

Debido a que es un experimento computacional fue posible tomar una muestra aleatoria simple de los datos, esto permitió trabajar con una muestra probabilística con diseño sencillo. Se realizaron muestras de distintos tamaños y se seleccionó del 20 % de la población de interés con un total de 137, 249 observaciones.

En las *Tablas 14 y 15* se compara esta muestra con respecto a la población de interés para observar los cambios que hayan podido tener los porcentajes de distribución geográfica y en las variables explicativas con respecto a la respuesta Ingreso.

Tabla 14: Distribución de la población de interés y muestra *SI 0.2* por estado

	Observaciones				UPMs				Estratos				Municipios			
	población		muestra		población		muestra		población		muestra		población		muestra	
Entidad	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**
Oaxaca	207,994	30.3	41,465	30.2	14,279	32.9	11,436	32.6	1,361	47.4	1,246	46.2	441	54.4	439	54.4
Chiapas	97,993	14.3	19,572	14.3	4,811	11.1	3,882	11.1	210	7.3	203	7.5	32	4.0	32	4.0
Guerrero	69,133	10.1	13,937	10.1	4,296	9.9	3,476	9.9	170	5.9	165	6.1	26	3.2	26	3.2
Yucatán	60,480	8.8	11,879	8.7	2,985	6.9	2,701	7.7	156	5.4	145	5.4	55	6.8	55	6.8
Puebla	55,480	8.1	11,174	8.1	3,101	7.1	2,731	7.8	239	8.3	225	8.3	65	8.0	65	8.0
Veracruz	43,359	6.3	8,750	6.4	1,751	4.0	1,585	4.5	127	4.4	118	4.4	30	3.7	30	3.7
Chihuahua	36,170	5.3	7,193	5.2	3,440	7.9	2,526	7.2	103	3.6	103	3.8	25	3.1	25	3.1
Tlaxcala	20,477	3.0	4,180	3.0	659	1.5	637	1.8	37	1.3	37	1.4	13	1.6	13	1.6
Sonora	17,804	2.6	3,490	2.5	1,728	4.0	1,232	3.5	111	3.9	107	4.0	38	4.7	38	4.7
Jalisco	15,241	2.2	3,030	2.2	1,181	2.7	953	2.7	77	2.6	75	2.8	16	2.0	16	2.0
SL Potosí	9,590	1.4	1,946	1.4	643	1.5	531	1.5	24	0.8	24	0.9	5	0.6	5	0.6
México	9,391	1.4	1,914	1.4	307	0.7	293	0.8	25	0.9	25	0.9	5	0.6	5	0.6
Nuevo León	7,953	1.2	1,583	1.2	828	1.9	600	1.7	43	1.5	41	1.5	14	1.7	14	1.7
Zacatecas	7,575	1.1	1,517	1.1	718	1.7	573	1.6	61	2.1	59	2.2	13	1.6	13	1.6
Durango	6,859	1.0	1,451	1.1	902	2.1	589	1.7	35	1.2	34	1.3	8	1.0	8	1.0
Coahuila	5,623	0.8	1,192	0.9	493	1.1	398	1.1	25	0.9	25	0.9	9	1.1	9	1.1
Hidalgo	5,092	0.7	963	0.7	271	0.6	240	0.7	13	0.5	12	0.4	3	0.4	3	0.4
Tamaulipas	3,501	0.5	710	0.5	440	1.0	295	0.8	30	1.0	29	1.1	7	0.9	7	0.9
Nayarit	2,726	0.4	545	0.4	352	0.8	219	0.6	7	0.2	7	0.3	1	0.1	1	0.1
Michoacán	2,538	0.4	510	0.4	192	0.4	164	0.5	14	0.5	14	0.5	3	0.4	3	0.4
Guanajuato	1,264	0.2	248	0.2	82	0.2	62	0.2	5	0.2	5	0.2	1	0.1	1	0.1
Total	686,243	100.0	137,249	100.0	43,459	100.0	35,123	100.0	2,873	100.0	2,699	100.0	810	100	808	100.0

población: población de interés *PI*
muestra: muestra aleatoria simple del 20 % de población de interés SI 0.2

De forma general se mantuvo la proporción de población, estratos y UPMs por estados. La razón por la que se perdieron estratos y UPMs son las pocas observaciones que contenían algunos (más del 6 % de los estratos y más del 21 % de las UPMs tenían menos de 5

observaciones). El número de observaciones en estos grupos geográficos es debido a las pocas viviendas que existen en algunos de ellos, aunado a la selección de observaciones que se realizó para solo tomar aquellas sujetas a responder la pregunta sobre ingreso. Mostrar la distribución en UPM y Estrato es meramente ilustrativo.

Tabla 15: Proporción observada de la respuesta en Ingreso por variables explicativas, en la población de interés *N=686,243* **y muestra** ***SI 0.2*** *n=137,249*

Variable	Categoría	población		muestra		población		muestra		población		muestra	
		N resp	%	N resp	%	Resp	%	Resp	%	Total	%	Total	%
Sexo	Hombre	107,847	20.8	21,479	20.7	410,109	79.2	82,101	79.3	517,952	75.5	103,580	75.5
	Mujer	17,102	10.2	3,479	10.3	151,185	89.8	30,190	89.7	168,287	24.5	33,669	24.5
Edad	12-19	10,268	18.9	2,043	18.8	43,918	81.1	8,815	81.2	54,186	7.9	10,858	7.9
	20-29	27,872	15.4	5,571	15.3	153,063	84.6	30,902	84.7	180,935	26.4	36,473	26.6
	30-59	66,179	17.4	13,198	17.4	314,253	82.6	62,519	82.6	380,432	55.4	75,717	55.2
	60+	20,630	29.2	4,146	29.2	50,060	70.8	10,055	70.8	70,690	10.3	14,201	10.3
Nivel Educativo	Nin/Prim Incomp	49,851	24.6	9,935	24.6	152,451	75.4	30,483	75.4	202,302	29.5	40,418	29.4
	Prim Comp	35,188	19.7	7,005	19.6	143,174	80.3	28,769	80.4	178,362	26.0	35,774	26.1
	Sec Comp	26,104	15.2	5,200	15.1	145,740	84.8	29,171	84.9	171,844	25.0	34,371	25.0
	Bach Comp y más	13,806	10.3	2,818	10.6	119,929	89.7	23,868	89.4	133,735	19.5	26,686	19.5
Situación Laboral	Cuenta Propia	93,962	34.4	18,708	34.2	179,357	65.6	35,962	65.8	273,319	39.8	54,670	39.8
	Empleado/Asist	19,812	6.4	4,065	6.5	291,965	93.6	58,118	93.5	311,777	45.4	62,183	45.3
	Jornalero/Peón	8,738	10.2	1,696	9.8	76,934	89.8	15,547	90.2	85,672	12.5	17,243	12.6
	Patrón/Empleador	2,437	15.7	489	15.5	13,038	84.3	2,664	84.5	15,475	2.3	3,153	2.3
Situación Conyugal	Casado	57,020	19.1	11,471	19.2	241,390	80.9	48,145	80.8	298,410	43.5	59,616	43.4
	No Casado	67,929	17.5	13,487	17.4	319,904	82.5	64,146	82.6	387,833	56.5	77,633	56.6
Lengua indígena	Si	73,437	23.4	14,709	23.5	240,526	76.6	47,885	76.5	313,963	45.8	62,594	45.6
	No	51,512	13.8	10,249	13.7	320,768	86.2	64,406	86.3	372,280	54.2	74,655	54.4
Entidad	Oaxaca	46,773	22.5	9,276	22.4	161,221	77.5	32,189	77.6	207,994	30.3	41,465	30.2
	Chiapas	24,659	25.2	4,944	25.3	73,334	74.8	14,628	74.7	97,993	14.3	19,572	14.3
	Guerrero	15,335	22.2	3,096	22.2	53,798	77.8	10,841	77.8	69,133	10.1	13,937	10.2
	Yucatán	5,662	9.4	1,124	9.5	54,818	90.6	10,755	90.5	60,480	8.8	11,879	8.7
	Puebla	8,052	14.5	1,642	14.7	47,428	85.5	9,532	85.3	55,480	8.1	11,174	8.1
	Veracruz	5,003	11.5	1,012	11.6	38,356	88.5	7,738	88.4	43,359	6.3	8,750	6.4
	Chihuahua	5,648	15.6	1,071	14.9	30,522	84.4	6,122	85.1	36,170	5.3	7,193	5.2
	Tlaxcala	2,282	11.1	468	11.2	18,195	88.9	3,712	88.8	20,477	3.0	4,180	3.0
	Sonora	1,347	7.6	277	7.9	16,457	92.4	3,213	92.1	17,804	2.6	3,490	2.5
	Jalisco	1,801	11.8	347	11.5	13,440	88.2	2,683	88.5	15,241	2.2	3,030	2.2
	San Luis Potosí	921	9.6	173	8.9	8,669	90.4	1,773	91.1	9,590	1.4	1,946	1.4
	México	907	9.7	177	9.2	8,484	90.3	1,737	90.8	9,391	1.4	1,914	1.4
	Nuevo León	803	10.1	178	11.2	7,150	89.9	1,405	88.8	7,953	1.2	1,583	1.2
	Zacatecas	1,018	13.4	222	14.6	6,557	86.6	1,295	85.4	7,575	1.1	1,517	1.1
	Durango	1,384	20.2	273	18.8	5,475	79.8	1,178	81.2	6,859	1.0	1,451	1.1
	Coahuila	437	7.8	98	8.2	5,186	92.2	1,094	91.8	5,623	0.8	1,192	0.9
	Hidalgo	1,523	29.9	291	30.1	3,569	70.1	672	69.8	5,092	0.7	963	0.7
	Tamaulipas	477	13.6	96	13.5	3,024	86.4	614	86.5	3,501	0.5	710	0.5
	Nayarit	352	12.9	78	14.3	2,374	87.1	467	85.7	2,726	0.4	545	0.4
	Michoacán	464	18.3	89	17.5	2,074	81.7	421	82.5	2,538	0.4	510	0.4
	Guanajuato	101	8.0	26	10.5	1,163	92.0	222	89.5	1,264	0.2	248	0.2
Total		**124,949**	**18.2**	**24,958**	**18.2**	**561,294**	**81.8**	**112,291**	**81.8**	**686,243**	**100.0**	**137,249**	**100.0**

población: población de interés *PI*

muestra: muestra aleatoria simple del 20 % de población de interés SI 0.2

En el caso de las proporciones observadas de respuesta en Ingreso según las variables explicativas, estas se mantuvieron muy parecidas, lo que dio un indicio de que las categorías

en cada variable mantuvieron su representatividad porcentual.

3.2 Ajuste de un modelo de regresión logística a la muestra *SI* seleccionada

La muestra permitió realizar el proceso de selección de modelo de regresión logística de forma más eficiente computacionalmente, sin embargo, la base de datos sigue teniendo un gran tamaño (137,249 filas) por lo que una tabla de contingencia disminuyó el trabajo computacional. Tomando en cuenta las mismas variables que en la población de interés se obtuvo una base con $21,504$ filas [Sexo (2) $\times$ Edad (4) $\times$ Nivel Educativo (4) $\times$ Situación Laboral (4) $\times$ Situación Conyugal (2) $\times$ Lengua Indígena (2) $\times$ Estado (21) $\times$ respuesta Ingreso (2) $= 21,504$].

Tabla 16: Número de celdas en la tabla de contingencia de la muestra *SI 0.2*

Frecuencia	**Total**	**%**
Diferente de cero	7,229	33.6
Cero	14,275	66.4
Total	21,504	100.0

Se observa en la *Tabla 16* que 14,275 celdas contienen frecuencia cero, principalmente ocasionado por las pocas observaciones en algunas categorías de las variables. El aumento de las celdas vacías a comparación de la población de interés (15.1 % más) es debido al tamaño de observaciones que muchas celdas tenían (alrededor del 20 % de las que tenían frecuencia diferente de cero era con menos de 5 observaciones), dado que la muestra fue al 20 % existía una alta probabilidad que estos casos no fueran representados en la muestra.

La selección comenzó con el ajuste del modelo aditivo simple. Al analizar las variables en el modelo presentado en el *Cuadro 6*, se notó que fueron significativas según la prueba de hipótesis con jí-cuadrada χ^2.

```
Cuadro 6. Ajuste del modelo aditivo simple (m1.20) a la muestra SI 0.2
```

```
> m1.20 <- glm(INGTRMEN_RESP ~ SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r +
HLENGUA + SITUA_CONYUGAL_r + ENT, family=binomial, weight=Freq, data=BaseDT.20)
```

```
> formula(m1.20)
```

```
INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r
+ NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r
```

```
> drop1(m1.20,test="chisq", k=log(137249))
```

```
         Df Deviance    BIC     LRT Pr(>Chi)
<none>                110314 110705
ENT              20   111402 111555  1087.2  < 2e-16 ***
SEXO              1   112100 112479  1786.0  < 2e-16 ***
EDAD_r            3   110487 110841   172.2  < 2e-16 ***
SITUACION_TRAB_r  3   121863 122217 11548.2  < 2e-16 ***
NIVACAD_r         3   110454 110809   139.3  < 2e-16 ***
HLENGUA           1   110641 111019   326.3  < 2e-16 ***
SITUA_CONYUGAL_r  1   110319 110698     5.0  0.02477 *
```

De manera similar al ajuste efectuado a la población de interés, buscando un modelo con interacciones, se realizaron los tres algoritmos del método *step* y se tomaron ambos criterios *AIC y BIC*. Se determinó que los modelos a ser comparados serían aquellos resultantes del proceso *Backward-forward* y dándole importancia al criterio *BIC*. Así mismo, la variable Entidad solo se incluyó a los modelos como efecto aditivo, ya que las interacciones con las otras variables causan muchos casos con frecuencia cero. Con esto definido, se procedió a buscar modelos más complejos.

Se inicio buscando interacciones de hasta primer orden y se llegó a un modelo. Posteriormente se realizó un proceso de selección de modelo que involucró interacciones de segundo orden y se llegó al mismo modelo que el obtenido en el paso anterior, por lo que ninguna interacción de variables se añadió. Finalmente, al realizarse el proceso buscando interacciones de mayor orden, éstas ya no fueron significativas y se llegó al mismo modelo, este fue nombrado *m220BIC*.

Cuadro 7. Ajuste del modelo considerando las interacciones de mayor orden (m220BIC) a la muestra SI 0.2

```
> m220BIC <- step(glm(INGTRMEN_RESP~ ENT + (SEXO + EDAD_r + SITUACION_TRAB_r +
NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r)^., family=binomial, data=BaseDT.20, weight=Freq),
direction="both", k=log(137249))

> formula (m220BIC)

INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r +
HLENGUA + SITUA_CONYUGAL_r + SEXO:SITUACION_TRAB_r + SEXO:NIVACAD_r +
SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA + HLENGUA:SITUA_CONYUGAL_r

> drop1(m220BIC,test="Chisq", k=log(137249))

                        Df Deviance    BIC     LRT  Pr(>Chi)
<none>                       109235 109779
ENT                      20   110335 110643 1100.57 < 2.2e-16 ***
EDAD_r                    3   109392 109901  157.68 < 2.2e-16 ***
SEXO:SITUACION_TRAB_r     3   109804 110313  569.47 < 2.2e-16 ***
SEXO:NIVACAD_r            3   109274 109783   39.31 1.490e-08 ***
SITUACION_TRAB_r:HLENGUA  3   109416 109924  180.93 < 2.2e-16 ***
NIVACAD_r:HLENGUA         3   109280 109789   45.53 7.138e-10 ***
HLENGUA:SITUA_CONYUGAL_r  1   109259 109792   24.60 7.063e-07 ***
```

Se evaluaron también los modelos ajustados en la población de interés aplicados en la muestra para revisar la significancia de las interacciones y comparar su efectividad con la prueba de hipótesis de la devianza.

Cuadro 8. Ajuste de los modelos m2BIC y m3BIC a la muestra SI 0.2

> m2BIC.20 <- glm(formula=formula(m2BIC), family=binomial, data=Base.20)

> drop1(m2BIC.20,test="Chisq", k=log(137249))

	Df	Deviance	BIC	LRT	Pr(>Chi)	
<none>		108992	109939			
ENT	20	110073	110783	1080.62	< 2.2e-16	***
SEXO:EDAD_r	3	109016	109927	23.54	3.116e-05	***
SEXO:SITUACION_TRAB_r	3	109531	110442	538.74	< 2.2e-16	***
SEXO:NIVACAD_r	3	109031	109942	38.33	2.411e-08	***
SEXO:HLENGUA	1	108997	109932	5.05	0.02467	*
EDAD_r:SITUACION_TRAB_r	9	109090	109930	98.09	< 2.2e-16	***
EDAD_r:NIVACAD_r	9	109032	109872	39.73	8.511e-06	***
EDAD_r:SITUA_CONYUGAL_r	3	109020	109931	27.31	5.069e-06	***
SITUACION_TRAB_r:NIVACAD_r	9	109062	109902	69.71	1.739e-11	***
SITUACION_TRAB_r:HLENGUA	3	109124	110035	131.26	< 2.2e-16	***
NIVACAD_r:HLENGUA	3	109028	109939	35.59	9.143e-08	***
HLENGUA:SITUA_CONYUGAL_r	1	109020	109955	28.03	1.197e-07	***

> m3BIC.20 <- glm(formula=formula(m3BIC), family=binomial, data=Base.20)

> drop1(m3BIC.20,test="Chisq", k=log(137249))

	Df	Deviance	BIC	LRT	Pr(>Chi)	
<none>		108951	109956			
ENT	20	110032	110801	1081.37	< 2.2e-16	***
SEXO:EDAD_r	3	108975	109945	23.94	2.575e-05	***
SEXO:SITUACION_TRAB_r	3	109493	110463	541.78	< 2.2e-16	***
EDAD_r:SITUACION_TRAB_r	9	109048	109947	97.24	< 2.2e-16	***
EDAD_r:NIVACAD_r	9	108992	109891	40.55	6.043e-06	***
EDAD_r:SITUA_CONYUGAL_r	3	108979	109949	28.14	3.396e-06	***
SITUACION_TRAB_r:NIVACAD_r	9	109022	109921	70.68	1.121e-11	***
SITUACION_TRAB_r:HLENGUA	3	109071	110041	120.44	< 2.2e-16	***
SEXO:NIVACAD_r:HLENGUA	3	108981	109951	29.54	1.721e-06	***
SEXO:HLENGUA:SITUA_CONYUGAL_r	1	108965	109959	13.87	0.000196	***

Nuevamente los modelos están anidados, los más complejos contienen las variables e interacciones de los otros *(m1 ⊂ m220BIC ⊂ m2BIC ⊂ m3BIC)*. Se realizó la prueba de hipótesis con los modelos dos a dos:

$$H_0 : dev_1 - dev_2 = 0 \quad vs \quad H_1 : dev_1 - dev_2 \neq 0.$$

Cuadro 9. Prueba de hipótesis para modelos anidados, ajustados a la muestra SI 0.2

#m1. Modelo Aditivo Simple
#m2BIC. Modelo buscando interacciones de dos, seleccionado en step con BIC en PI
#m3BIC. Modelo buscando interacciones de tres, seleccionado en step con BIC en PI
#m220BIC. Modelo buscando todas las interacciones, seleccionado en Step con BIC en SI 0.2

```
> anova(m1.20, m220BIC.20, test="Chisq")

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    137216     110314
2    137203     109235 13  1079.62 < 2.2e-16 ***

> anova(m220BIC.20, m2BIC.20, test="Chisq")

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    137203     109235
2    137169     108992 34   242.35 < 2.2e-16 ***

> anova(m2BIC.20, m3BIC.20, test="Chisq")

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    137169     108992
2    137164     108951  5    41.40 7.781e-08 ***
```

La prueba de hipótesis mostrada en el *Cuadro 9* señaló que el mejor modelo es el que tiene más interacciones, es decir, se rechaza la hipótesis nula en cada prueba. Sin embargo, para la muestra se seleccionó el modelo *m220BIC*, el cual tiene 9 interacciones de variables menos que el modelo *m3BIC* y facilita su presentación. Además, este modelo tiene comparativamente el *BIC* más bajo al ajustar los datos de la muestra, por lo que fue seleccionado con el algoritmo, además, se supone que su ajuste tiene menos error por ser menos observaciones. La comparación de *AIC*, *BIC* y devianza se encuentra en el *anexo b*.

En el siguiente capítulo se consideró una muestra obtenida con un diseño complejo y se ajustó un modelo de regresión logística considerando esto.

Capítulo 4. Ajuste de un modelo de regresión logística a las muestras *STSIC*

En esta parte se ajustó un modelo de regresión logística a una muestra compleja, se consideró que los datos provienen de una encuesta, que las muestras se obtienen con un diseño muestral *STSIC* y se aplicaron los métodos algorítmicos sugeridos por autores como Hosmer et al. (2013) y Heeringa et al. (2017), que consideran el diseño para una mejor estimación de la varianza.

4.1 Selección de muestras con diseño *STSIC* de la población de interés

Para obtener la muestra se siguió el esquema que usó el INEGI para la población muestreada de la Encuesta Intercensal 2015. El esquema es estratificado, por conglomerados, aleatorio simple en una sola etapa de selección, este diseño está descrito a continuación:

Muestreo estratificado por conglomerados, unietápico y aleatorio simple (*STSIC*)

Para detallar este esquema se utilizan los diseños expuestos en la *Sección 1.2* pero con la notación empleada en Hosmer (2013), p. 233. Tomando a la población U se particiona en $k = 1, 2, ..., K$ estratos, posteriormente cada k estrato se divide $j = 1, 2, ..., M_k$ conglomerados (Unidades Primarias de Muestreo) disjuntos con $i = 1, 2, ..., N_{kj}$ elementos cada uno. Por lo tanto, el total de elementos de la población U es:

$$N = \sum_{k=1}^{K} \sum_{j=1}^{M_k} N_{kj}.$$

Considerando el esquema probabilístico aleatorio simple en todos los estratos, la probabilidad de inclusión para cada j-ésimo conglomerado es:

$$\pi_{kj} = \frac{m_k}{M_k}, \quad \text{para cada} \quad j = 1, 2, ..., M_k,$$

donde m_k es el número de conglomerados seleccionados para la muestra en cada estrato k. Su factor de expansión es:

$$w_{kj} = \frac{1}{\pi_{kj}} = \frac{M_k}{m_k}.$$

Dado que es unietápico, todos los elementos del j-ésimo conglomerado seleccionado serán observados, por lo tanto, la probabilidad de inclusión y factor de expansión para cada elemento i será igual al del conglomerado al que pertenece:

$$\pi_{kji} = \frac{m_k}{M_k},$$
$$w_{kji} = \frac{M_k}{m_k}.$$

Considerando la población de interés se imitó el diseño *STSIC* aplicado por el INEGI, sin embargo, el diseño exacto es desconocido dado que en ninguna publicación se especifica el número de UPMs seleccionadas (m_k) para cada estrato. Por lo que se seleccionaron dos muestras *STSIC* diferentes, las cuales toman un número fijo de UPMs para todos los estratos: la primera selecciona $m_k = 3$ y la segunda selecciona $m_k = 2$. La población de interés tiene 2,873 estratos, de los cuales 412 solo tienen una UPM y 268 tienen dos UPM, por lo que en aquellos donde existen menos de m_k conglomerados se seleccionó con certeza los existentes y se dio una probabilidad de inclusión igual a uno. Las muestras seleccionadas son descritas en las siguientes *Tablas 17, 18 y 19*.

Tabla 17: Tamaño de las muestras *STSIC* seleccionadas

Muestra	**Número de UPM en muestra**	**% ***	**Número de obs en muestra**	**% ***	$\boldsymbol{w_{kji}}$
$m_k = 3$ UPMs por estrato	7,527	17.31	110,453	16.09	$\frac{M_k}{3}$, si $M_K \geq 3$ 1 si $M_k = (1, 2)$
$m_k = 2$ UPMs por estrato	5,334	12.27	77,980	11.36	$\frac{M_k}{2}$, si $M_K \geq 2$ 1 si $M_k = 1$

*Porcentaje del total de la población de interés

Tabla 18: Proporción observada de la respuesta en Ingreso por variables explicativas, en población de interés *N=686,243* y muestra *STSIC* $m_k = 3$ *n=110,453*

Variable	Categoría	población		muestra		población		muestra		población		muestra	
		N resp	%	N resp	%	Resp	%	Resp	%	Total	%	Total	%
Sexo	Hombre	107,847	20.8	19,541	22.8	410,109	79.2	66,230	77.2	517,952	75.5	85,771	77.2
	Mujer	17,102	10.2	2,720	11.0	151,185	89.8	21,962	89.0	168,287	24.5	24,682	22.3
Edad	12-19	10,268	18.9	1,879	20.2	43,918	81.1	7,437	79.8	54,186	7.9	9,316	8.4
	20-29	27,872	15.4	4,923	16.8	153,063	84.6	24,305	83.2	180,935	26.4	29,228	26.5
	30-59	66,179	17.4	11,609	19.3	314,253	82.6	48,402	80.7	380,432	55.4	60,011	54.3
	60+	20,630	29.2	3,850	32.4	50,060	70.8	8,048	67.6	70,690	10.3	11,898	10.8
Nivel Educativo	Nin/Prim Incomp	49,851	24.6	9,122	26.6	152,451	75.4	25,178	73.4	202,302	29.5	34,300	31.1
	Prim Comp	35,188	19.7	6,536	21.3	143,174	80.3	24,191	78.7	178,362	26.0	30,727	27.8
	Sec Comp	26,104	15.2	4,603	16.6	145,740	84.8	23,114	83.4	171,844	25.0	27,717	25.1
	Bach Comp y más	13,806	10.3	2,000	11.3	119,929	89.7	15,709	88.7	133,735	19.5	17,709	16.0
Situación Laboral	Cuenta Propia	93,962	34.4	17,107	37.8	179,357	65.6	28,120	62.2	273,319	39.8	45,227	40.9
	Empleado/Asist	19,812	6.4	2,915	6.2	291,965	93.6	44,191	93.8	311,777	45.4	47,106	42.6
	Jornalero/Peón	8,738	10.2	1,737	11.0	76,934	89.8	14,079	89.0	85,672	12.5	15,816	14.3
	Patrón/Empleador	2,437	15.7	502	21.8	13,038	84.3	1,802	78.2	15,475	2.3	2,304	2.1
Situación Conyugal	Casado	57,020	19.1	10,148	21.5	241,390	80.9	37,081	78.5	298,410	43.5	47,229	42.8
	No Casado	67,929	17.5	12,113	19.2	319,904	82.5	51,111	80.8	387,833	56.5	63,224	57.2
Lengua indígena	Si	73,437	23.4	12,080	25.1	240,526	76.6	36,035	74.9	313,963	45.8	48,115	43.6
	No	51,512	13.8	10,181	16.3	320,768	86.2	52,157	83.7	372,280	54.2	62,338	56.4
Entidad	Oaxaca	46,773	22.5	11,192	24.9	161,221	77.5	33,794	75.1	207,994	30.3	44,986	40.7
	Chiapas	24,659	25.2	3,050	24.5	73,334	74.8	9,382	75.5	97,993	14.3	12,432	11.3
	Guerrero	15,335	22.2	2,070	27.6	53,798	77.8	5,422	72.4	69,133	10.1	7,492	6.8
	Yucatán	5,662	9.4	780	11.9	54,818	90.6	5,770	88.1	60,480	8.8	6,550	5.9
	Puebla	8,052	14.5	1,523	15.1	47,428	85.5	8,581	84.9	55,480	8.1	10,104	9.1
	Veracruz	5,003	11.5	1,093	12.8	38,356	88.5	7,437	87.2	43,359	6.3	8,530	7.7
	Chihuahua	5,648	15.6	454	14.4	30,522	84.4	2,689	85.6	36,170	5.3	3,143	2.8
	Tlaxcala	2,282	11.1	198	8.8	18,195	88.9	2,063	91.2	20,477	3.0	2,261	2.0
	Sonora	1,347	7.6	236	8.2	16,457	92.4	2,628	91.8	17,804	2.6	2,864	2.6
	Jalisco	1,801	11.8	343	13.0	13,440	88.2	2,292	87.0	15,241	2.2	2,635	2.4
	San Luis Potosí	921	9.6	112	10.6	8,669	90.4	945	89.4	9,590	1.4	1,057	1.0
	México	907	9.7	173	9.8	8,484	90.3	1,596	90.2	9,391	1.4	1,769	1.6
	Nuevo León	803	10.1	124	11.2	7,150	89.9	985	88.8	7,953	1.2	1,109	1.0
	Zacatecas	1,018	13.4	278	17.9	6,557	86.6	1,276	82.1	7,575	1.1	1,554	1.4
	Durango	1,384	20.2	234	24.8	5,475	79.8	709	75.2	6,859	1.0	943	0.9
	Coahuila	437	7.8	49	6.1	5,186	92.2	756	93.9	5,623	0.8	805	0.7
	Hidalgo	1,523	29.9	114	20.8	3,569	70.1	434	79.2	5,092	0.7	548	0.5
	Tamaulipas	477	13.6	98	14.7	3,024	86.4	569	85.3	3,501	0.5	667	0.6
	Nayarit	352	12.9	19	12.7	2,374	87.1	131	87.3	2,726	0.4	150	0.1
	Michoacán	464	18.3	109	18.5	2,074	81.7	481	81.5	2,538	0.4	590	0.5
	Guanajuato	101	8.0	12	4.5	1,163	92.0	252	95.5	1,264	0.2	264	0.2
Total		**124,949**	**18.2**	**22,261**	**20.2**	**561,294**	**81.8**	**88,192**	**79.8**	**686,243**	**100.0**	**110,453**	**100.0**

población: población de interés *PI*

muestra: muestra *STSIC* con $m_k = 3$ *STSIC3*

Tabla 19: Proporción observada de la respuesta en Ingreso por variables explicativas, en población de interés *N=686,243* y muestra *STSIC* $m_k = 2$ *n=77,980*

Variable	Categoría	población		muestra		población		muestra		población		muestra	
		N resp	%	N resp	%	Resp	%	Resp	%	Total	%	Total	%
Sexo	Hombre	107,847	20.8	14,164	23.3	410,109	79.2	46,636	76.7	517,952	75.5	60,800	78.0
	Mujer	17,102	10.2	1,911	11.1	151,185	89.8	15,269	88.9	168,287	24.5	17,180	22.0
Edad	12-19	10,268	18.9	1,351	21.1	43,918	81.1	5,038	78.9	54,186	7.9	6,389	8.2
	20-29	27,872	15.4	3,558	17.3	153,063	84.6	17,005	82.7	180,935	26.4	20,563	26.4
	30-59	66,179	17.4	8,407	19.8	314,253	82.6	34,096	80.2	380,432	55.4	42,503	54.5
	60+	20,630	29.2	2,759	32.4	50,060	70.8	5,766	67.6	70,690	10.3	8,525	10.9
Nivel Educativo	Nin/Prim Incomp	49,851	24.6	6,591	26.7	152,451	75.4	18,080	73.3	202,302	29.5	24,671	31.6
	Prim Comp	35,188	19.7	1,408	11.4	143,174	80.3	10,996	88.6	178,362	26.0	12,404	15.9
	Sec Comp	26,104	15.2	4,751	21.8	145,740	84.8	17,025	78.2	171,844	25.0	21,776	27.9
	Bach Comp y más	13,806	10.3	3,325	17.4	119,929	89.7	15,804	82.6	133,735	19.5	19,129	24.5
Situación Laboral	Cuenta Propia	93,962	34.4	12,470	38.6	179,357	65.6	19,875	61.4	273,319	39.8	32,345	41.4
	Empleado/Asist	19,812	6.4	2,154	6.5	291,965	93.6	30,875	93.5	311,777	45.4	33,029	42.4
	Jornalero/Peón	8,738	10.2	1,136	10.4	76,934	89.8	9,775	89.6	85,672	12.5	10,911	14.0
	Patrón/Empleador	2,437	15.7	315	18.6	13,038	84.3	1,380	81.4	15,475	2.3	1,695	2.2
Situación Conyugal	Casado	57,020	19.1	7,355	22.0	241,390	80.9	26,049	78.0	298,410	43.5	33,404	42.8
	No Casado	67,929	17.5	8,720	19.6	319,904	82.5	35,856	80.4	387,833	56.5	44,576	57.2
Lengua indígena	Si	73,437	23.4	8,769	25.8	240,526	76.6	25,264	74.2	313,963	45.8	34,033	43.6
	No	51,512	13.8	7,306	16.6	320,768	86.2	36,641	83.4	372,280	54.2	43,947	56.4
Entidad	Oaxaca	46,773	22.5	8,351	25.4	161,221	77.5	24,463	74.6	207,994	30.3	32,814	42.1
	Chiapas	24,659	25.2	2,288	28.3	73,334	74.8	5,809	71.7	97,993	14.3	8,097	10.4
	Guerrero	15,335	22.2	1,200	26.3	53,798	77.8	3,361	73.7	69,133	10.1	4,561	5.8
	Yucatán	5,662	9.4	557	11.9	54,818	90.6	4,108	88.1	60,480	8.8	4,665	6.0
	Puebla	8,052	14.5	1,119	15.1	47,428	85.5	6,303	84.9	55,480	8.1	7,422	9.5
	Veracruz	5,003	11.5	791	12.8	38,356	88.5	5,407	87.2	43,359	6.3	6,198	7.9
	Chihuahua	5,648	15.6	306	14.1	30,522	84.4	1,859	85.9	36,170	5.3	2,165	2.8
	Tlaxcala	2,282	11.1	160	9.4	18,195	88.9	1,541	90.6	20,477	3.0	1,701	2.2
	Sonora	1,347	7.6	180	9.3	16,457	92.4	1,763	90.7	17,804	2.6	1,943	2.5
	Jalisco	1,801	11.8	245	13.1	13,440	88.2	1,629	86.9	15,241	2.2	1,874	2.4
	San Luis Potosí	921	9.6	91	14.0	8,669	90.4	560	86.0	9,590	1.4	651	0.8
	México	907	9.7	121	8.6	8,484	90.3	1,284	91.4	9,391	1.4	1,405	1.8
	Nuevo León	803	10.1	91	11.1	7,150	89.9	728	88.9	7,953	1.2	819	1.1
	Zacatecas	1,018	13.4	557	11.9	6,557	86.6	4,108	88.1	7,575	1.1	4,665	6.0
	Durango	1,384	20.2	142	25.5	5,475	79.8	415	74.5	6,859	1.0	557	0.7
	Coahuila	437	7.8	45	8.8	5,186	92.2	467	91.2	5,623	0.8	512	0.7
	Hidalgo	1,523	29.9	60	15.5	3,569	70.1	327	84.5	5,092	0.7	387	0.5
	Tamaulipas	477	13.6	54	12.5	3,024	86.4	379	87.5	3,501	0.5	433	0.6
	Nayarit	352	12.9	10	9.9	2,374	87.1	91	90.1	2,726	0.4	101	0.1
	Michoacán	464	18.3	55	15.4	2,074	81.7	301	84.6	2,538	0.4	356	0.5
	Guanajuato	101	8.0	15	12.9	1,163	92.0	101	87.1	1,264	0.2	116	0.1
Total		**124,949**	**18.2**	**16,075**	**20.6**	**561,294**	**81.8**	**61,905**	**79.4**	**686,243**	**100.0**	**77,980**	**100.0**

población: población de interés *PI*
muestra: muestra *STSIC* con $m_k = 2$ *STSIC2*

Se observa en las *Tablas 18 y 19* que en ambas muestras, de forma general, las categorías mantuvieron un porcentaje similar (no cambian más del 5 %) en el total por variable y en respuesta a Ingreso. Sin embargo, en las categorías de Entidad existieron cambios importantes, un resultado esperado, ya que como se observa en la *Tabla 8* la distribución de los estratos era desigual y al obtener la muestra con un esquema estratificado causó un sesgo a tener más población en aquellas entidades con más estratos. Por ejemplo, Oaxaca tenía el 47.4 % de los estratos en la población, lo que ocasionó que su porcentaje de observaciones

pasará de 30.3 % a 40.7 % en la muestra *STSIC* $m_k = 3$ y 42.1 % en la muestra *STSIC* $m_k = 2$.

Considerando estas dos muestras se procedió a seleccionar y ajustar modelos de regresión logística.

4.2 Ajuste de modelos de regresión logística a las muestras con diseño *STSIC*

Como se apunta en Heeringa et al. (2017), p.263, en muestras obtenidas con un diseño aleatorio simple (*SI*), los coeficientes y errores estándar de la regresión logística pueden ser estimados usando el método para máxima verosimilitud dada en la *sección 1.1.1.* Sin embargo, cuando la muestra es obtenida bajo un diseño muestral complejo, la aplicación directa de este método de máxima verosimilitud no es correcta, por las siguientes razones:

- Primero, las probabilidades de selección y respuesta (y_i) para las $i = 1, 2, ..., n$ observaciones en la muestra ya no son las mismas. Entonces, los factores de expansión w_{kji} son necesarios para una buena estimación.
- Segundo, la estratificación y conglomeración de la muestra viola la independencia que se asume de las respuestas (y_i) en las observaciones y que es crucial para el cálculo de desviaciones estándar bajo el método usado hasta el momento.

Estimación con pseudo-máxima verosimilitud

El método nombrado estimación con pseudo-máxima verosimilitud *(pseudo-maximum likelihood estimation, PLME)* es una propuesta para la estimación de los parámetros que considera los puntos anteriores. Este enfoque se hizo usando un estimador lineal de la matriz de varianza-covarianza examinando el diseño de la muestra, usa la solución del siguiente vector de ecuaciones (Hosmer et al. (2013), p. 233, (6.8)):

$$\sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}}[w_{kji} \times y_{kji}] \times \ln[\pi(\boldsymbol{x}_{kji})] + [w_{kji} \times (1 - y_{kji})] \times \ln[1 - \pi(\boldsymbol{x}_{kji})].$$

Donde la notación es la dada en la sección anterior, con $k = 1, 2, ..., K$ los estratos, $j = 1, 2, ..., m_k$ las UPMs seleccionadas en la muestra para cada estrato, $i = 1, 2, .., n_{kj}$ el número de observaciones por conglomerado y w_{kji} es el factor de expansión de la kji-ésima observación. Así mismo, y_{kji} es el resultado dicotómico de la variable respuesta, $\boldsymbol{x}_{kji}$ es el vector de variables explicativas para la kji-ésima observación. En consecuencia $\pi(\boldsymbol{x}_{kji})$ se define como:

$$\pi(\boldsymbol{x}_{kji}) = \frac{exp(\beta_0 + \beta_1 x_{kji1} + ... + \beta_p x_{kjip})}{1 + exp(\beta_0 + \beta_1 x_{kji1} + ... + \beta_p x_{kjip})}.$$

Derivar la ecuación respecto a los p parámetros desconocidos da un vector de $p+1$ ecuaciones e igualando cero es:

$$\boldsymbol{X'W}(\boldsymbol{y}-\boldsymbol{\pi})=\boldsymbol{0}$$

donde $\boldsymbol{X}$ es $n\times(p+1)$ la matriz de la covarianza de valores, $\boldsymbol{W}$ es la matriz diagonal $n\times n$ que contiene los factores de expansión, $\boldsymbol{y}$ es el vector $n\times 1$ de las respuestas observadas y $\boldsymbol{\pi}=[\pi(x_{111}),...,\pi(x_{Km_Knkj})]'$ el vector $n\times 1$ de las log-probabilidades para la variable respuesta. Resolver este sistema de ecuaciones da como resultado las estimaciones de los coeficientes.

La correcta estimación de la varianza está dada por (Hosmer et al. (2013), p. 234, (6.10)):

$$\widehat{Var}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X'DX})^{-1}\boldsymbol{S}(\boldsymbol{X'DX})^{-1}$$

con $\boldsymbol{D}=\boldsymbol{WV}$ una matriz diagonal $n\times n$ con el elemento general $w_{kji}\times\hat{\pi}(x_{kji})[1-\hat{\pi}(x_{kji})]$ y $\boldsymbol{S}$ un estimador agrupado dentro del estrato de la matriz de covarianza de $\boldsymbol{X'W}(\boldsymbol{y}-\boldsymbol{\pi})$.

La estimación de los coeficientes y varianzas requiere de métodos numéricos, una herramienta computacional que resuelve estos cálculos es el paquete *Survey* en el software *R*.

Paquete *Survey*.

Survey es un paquete que se utiliza para estimar parámetros a partir de datos muestrales. Para esto, se requiere además de la base de datos, especificar el diseño muestral y proveer la variable con los pesos muestrales. Algunas de sus funciones son:

`svydesign()` : *survey design* es una función que permite especificar el esquema de muestreo. Se compone de unos argumentos como el *id* el cual identifica a las *UPM*, *strata* que considera la estratificación y *weights* que permiten añadir el vector de factores de expansión para cada observación.

`svyglm()` : *survey-weighted generalized linear models* es una función que ajusta modelos de regresión a partir de datos que provienen de muestras complejas, usando los factores de expansión y la estimación con pseudo-máxima verosimilitud.

Se realizó el proceso de selección de un modelo de regresión logística a las dos muestras con diseño *STSIC* usando la función `svyglm()`, el proceso se realizó con el método *step* dando las mismas especificaciones que en la selección del modelo en la población de interés y la muestra aleatoria simple del 20 %, estas son: proceso *Backward-forward*, dándole importancia al criterio *BIC* y tomando la variable Entidad solo como efecto aditivo.

El *Cuadro 10* muestra el diseño dado a la función `svydesign()` y la función `step()` para la muestra *STSIC* $m_k = 3$.

Cuadro 10. Selección y ajuste de un modelo considerando diseño muestral a la muestra STSIC mk=3

```
> ST_3svy<- svydesign(id = ~UPM ,strata = ~ESTRATO,
data = Base_3.ST, weights = ~W)

> options(survey.lonely.psu="adjust")

  #Se asigna a las UPMs con solo una observación la varianza
  promedio de las UPMs con más de 1 observación.

> ST_3MBIC <- step(svyglm(INGTRMEN_RESP~ ENT +
(SEXO + EDAD_r + SITUACION_TRAB_r +
NIVACAD_r+ HLENGUA+ SITUA_CONYUGAL_r)^2,
family=binomial,design=ST_3svy),
k=log(nrow(Base_3.ST)),trace=F)

> formula(ST_3MBIC)

INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r +
NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r + SEXO:SITUACION_TRAB_r +
SEXO:NIVACAD_r + SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA

> Anova(ST_3MBIC,type="II",test="Chisq")

                        Df     Chisq Pr(>Chisq)
ENT                     20  162.1653  < 2.2e-16 ***
SEXO                     1  118.5271  < 2.2e-16 ***
EDAD_r                   3   47.4317  2.813e-10 ***
SITUACION_TRAB_r         3 1076.5270  < 2.2e-16 ***
NIVACAD_r                3   59.9870  5.916e-13 ***
HLENGUA                  1   25.5018  4.420e-07 ***
SITUA_CONYUGAL_r         1    4.5009   0.033877 *
SEXO:SITUACION_TRAB_r    3  207.2237  < 2.2e-16 ***
SEXO:NIVACAD_r           3   23.9870  2.514e-05 ***
SITUACION_TRAB_r:HLENGUA 3    5.4506   0.141624
NIVACAD_r:HLENGUA        3   15.0952   0.001737 **
```

El modelo seleccionado para la muestra *STSIC* $m_k = 3$ contiene a las siete variables explicativas y cuatro interacciones entre ellas. En esta muestra todas las variables explicativas son significativas, la interacción de Situación Laboral con Lengua indígena fue seleccionada bajo el criterio *BIC* pero la prueba de hipótesis Jí-cuadrada muestra que no es estadísticamente significativa.

El *Cuadro 11* muestra el diseño dado a la función `svydesign()` y la función `step()` para la muestra *STSIC* $m_k = 2$.

Cuadro 11. Selección y ajuste de un modelo considerando diseño muestral a la muestra STSIC mk=2

```
> ST_2svy<- svydesign(id = ~UPM ,strata = ~ESTRATO,
data = Base_2.ST, weights = ~W)

> options(survey.lonely.psu="adjust")

  #Se asigna a las UPMs con solo una observación la varianza
  promedio de las UPMs con más de 1 observación.

> ST_2MBIC <- step(svyglm(INGTRMEN_RESP~ ENT +
(SEXO + EDAD_r + SITUACION_TRAB_r +
NIVACAD_r+ HLENGUA+ SITUA_CONYUGAL_r)^2,
family=binomial,design=ST_2svy),
k=log(nrow(Base_2.ST)),trace=F)

> formula(ST_2MBIC)

INGTRMEN_RESP ~ ENT + SEXO + EDAD_r + SITUACION_TRAB_r +
NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r + SEXO:SITUACION_TRAB_r +
SEXO:NIVACAD_r + SITUACION_TRAB_r:HLENGUA

> Anova(ST_2MBIC,type="II",test="Chisq")

                          Df    Chisq Pr(>Chisq)
ENT                       20 117.7168  7.537e-16 ***
SEXO                       1 133.4463  < 2.2e-16 ***
EDAD_r                     3  57.5393  1.971e-12 ***
SITUACION_TRAB_r           3 708.6194  < 2.2e-16 ***
NIVACAD_r                  3  42.7038  2.844e-09 ***
HLENGUA                    1   1.9073  0.1672641
SITUA_CONYUGAL_r           1   3.7816  0.0518182 .
SEXO:SITUACION_TRAB_r      3 116.8088  < 2.2e-16 ***
SEXO:NIVACAD_r             3  14.3330  0.0024852 **
SITUACION_TRAB_r:HLENGUA   3  19.2009  0.0002485 ***
```

El modelo seleccionado para la muestra *STSIC* $m_k = 2$ contiene a las siete variables explicativas y tres interacciones entre ellas. La prueba de hipótesis mostró que la variable Lengua indígena no es significativa, sin embargo, su interacción con Situación Laboral lo es.

En el *Capítulo 5* se seleccionó un modelo para comparar la estimación de los coeficientes y desviaciones estándar en la población y muestras bajo los mismos parámetros.

Capítulo 5. Ajuste de un modelo de regresión logística común a la población de interés y las muestras

Los modelos candidatos estaban entre los obtenidos en la población de interés y tres muestras, de este modo se comparó la estimación de los coeficientes y errores estándar obtenidos en cada caso bajo las mismas variables y categorías. La siguiente *Tabla 20* describe las poblaciones y modelos seleccionados, candidatos para usar en común.

Tabla 20: Descripción de población de interés, muestras y los modelos seleccionados

Conjunto	Descripción	Fórmula de modelo seleccionado	Int	Nombre
Población de interés (*PI*, `Base`)	Base de datos con localidades censadas, 686,243 observaciones	INGTRMEN_RESP ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r + SEXO:EDAD_r + SEXO:SITUACION_TRAB_r + SEXO:NIVACAD_r + SEXO:HLENGUA + SEXO:SITUA_CONYUGAL_r + EDAD_r:SITUACION_TRAB_r + EDAD_r:NIVACAD_r + EDAD_r:SITUA_CONYUGAL_r + SITUACION_TRAB_r:NIVACAD_r + SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA + HLENGUA:SITUA_CONYUGAL_r + SEXO:NIVACAD_r:HLENGUA + SEXO:HLENGUA:SITUA_CONYUGAL_r	14	*m3BIC*
Muestra del 20 % *(SI 0.2, `Base.20`)*	Muestra aleatoria simple del 20 % de la población de interés, 137,249 observaciones.	INGTRMEN_RESP ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r + SEXO:SITUACION_TRAB_r + SEXO:NIVACAD_r + SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA + HLENGUA:SITUA_CONYUGAL_r	5	*m220BIC*
Muestra *STSIC* $m_k = 3$ (*STSIC3*, `Base_3.ST`)	Muestra *STSIC* seleccionando $m_k = 3$ UPMs por estrato de la población de interés, 110,453 observaciones.	INGTRMEN_RESP ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r + SEXO:SITUACION_TRAB_r + SEXO:NIVACAD_r + SITUACION_TRAB_r:HLENGUA + NIVACAD_r:HLENGUA	4	*ST_3MBIC*
Muestra *STSIC* $m_k = 2$ (*STSIC2*, `Base_2.ST`)	Muestra *STSIC* seleccionando $m_k = 2$ UPMs por estrato de la población de interés, 77,980 observaciones.	INGTRMEN_RESP ENT + SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA + SITUA_CONYUGAL_r + SEXO:SITUACION_TRAB_r + SEXO:NIVACAD_r + SITUACION_TRAB_r:HLENGUA	3	*ST_2MBIC*

Se observa que los modelos están anidados $ST\_2MBIC \subset ST\_3MBIC \subset m220BIC \subset m3BIC$.

El modelo seleccionado fue el que contiene 4 interacciones *(ST_3MBIC)*, toma interacciones que son significativas en casi todos los conjuntos, lo que permite una comparación de la estimación de los coeficientes con valores que se esperan estadísticamente significativos, como se observa en el *Cuadro 12*. Por lo tanto, se ajustó en la población y muestras.

Cuadro 12. Ajuste del modelo ST_3MBIC a la población de interés y muestras

#Ajuste en la población de interés

> pi.ST_3MBIC <- glm(formula=formula(ST_3MBIC), family=binomial, data=Base)

> drop1(pi.ST_3BIC,test="Chisq", k=log(nrow(Base)))

```
                      Df Deviance    BIC    LRT  Pr(>Chi)
<none>                      544818 545423
ENT                   20   550764 551100 5946.0 < 2.2e-16 ***
EDAD_r                 3   545676 546241  858.1 < 2.2e-16 ***
SITUA_CONYUGAL_r       1   544873 545464   54.9 1.257e-13 ***
SEXO:SITUACION_TRAB_r  3   547413 547977 2594.7 < 2.2e-16 ***
SEXO:NIVACAD_r         3   544999 545563  180.5 < 2.2e-16 ***
SITUACION_TRAB_r:HLENGUA 3 545801 546365  982.6 < 2.2e-16 ***
NIVACAD_r:HLENGUA      3   545023 545587  204.5 < 2.2e-16 ***
```

#Ajuste en la muestra SI del 20%

> SI.ST_3MBIC <- glm(formula=formula(ST_3MBIC), family=binomial, data=Base.20)

> drop1(SI.ST_3BIC,test="Chisq", k=log(nrow(Base.20)))

```
                        Df Deviance    BIC     LRT  Pr(>Chi)
<none>                      109259 109792
ENT                    20   110365 110661 1105.63 < 2.2e-16 ***
EDAD_r                  3   109416 109913  156.57 < 2.2e-16 ***
SITUA_CONYUGAL_r        1   109264 109784    4.55   0.03299 *
SEXO:SITUACION_TRAB_r   3   109833 110330  573.56 < 2.2e-16 ***
SEXO:NIVACAD_r          3   109299 109796   39.66 1.259e-08 ***
SITUACION_TRAB_r:HLENGUA 3  109455 109952  196.00 < 2.2e-16 ***
NIVACAD_r:HLENGUA       3   109300 109797   41.02 6.462e-09 ***
```

#Ajuste en la muestra STSIC mk=3

> STSIC3.ST_3MBIC <- svyglm(formula=formula(ST_3MBIC), family=binomial, design=ST_3svy)

> Anova(STSIC3.ST_3MBIC,type="II")

```
                      Df     Chisq Pr(>Chisq)
ENT                   20  162.1653  < 2.2e-16 ***
SEXO                   1  118.5271  < 2.2e-16 ***
EDAD_r                 3   47.4317  2.813e-10 ***
```

```
SITUACION_TRAB_r          3 1076.5270  < 2.2e-16 ***
NIVACAD_r                 3   59.9870  5.916e-13 ***
HLENGUA                   1   25.5018  4.420e-07 ***
SITUA_CONYUGAL_r          1    4.5009   0.033877 *
SEXO:SITUACION_TRAB_r     3  207.2237  < 2.2e-16 ***
SEXO:NIVACAD_r            3   23.9870  2.514e-05 ***
SITUACION_TRAB_r:HLENGUA  3    5.4506   0.141624
NIVACAD_r:HLENGUA         3   15.0952   0.001737 **
```

```
#Ajuste en la muestra STSIC mk=2
```

> STSIC2.ST_3MBIC <- svyglm(formula=formula(ST_3MBIC), family=binomial, design=ST_2svy)

> Anova(STSIC2.ST_3MBIC,type="II")

```
                          Df    Chisq Pr(>Chisq)
ENT                       20 117.1842  9.451e-16 ***
SEXO                       1 133.0524  < 2.2e-16 ***
EDAD_r                     3  54.7892  7.615e-12 ***
SITUACION_TRAB_r           3 704.4395  < 2.2e-16 ***
NIVACAD_r                  3  50.9583  4.993e-11 ***
HLENGUA                    1   2.5990   0.106933
SITUA_CONYUGAL_r           1   3.6080   0.057503 .
SEXO:SITUACION_TRAB_r      3 116.0602  < 2.2e-16 ***
SEXO:NIVACAD_r             3  16.2477   0.001009 **
SITUACION_TRAB_r:HLENGUA   3  23.2803  3.530e-05 ***
NIVACAD_r:HLENGUA          3   7.1842   0.066253 .
```

La estimación de los coeficientes y la desviación estándar se encuentran en el *anexo d* y en el *anexo e* se encuentra el gráfico de los intervalos de confianza del 95 % para cada coeficiente. Se observó que en la estimación para la categoría Tamaulipas $E(28)$, en todos los conjuntos, el intervalo de confianza contenía al cero, lo que implica que no es estadísticamente significativo y su valor es igual al de la categoría de referencia, se revisó con un proceso anova. Por lo tanto, estas categorías se reagruparon.

Para comprobar que el reagrupamiento realizado no muestra un cambio estadísticamente significativo en la estimación de los coeficientes se realizaron pruebas de hipótesis:

$$H_0 : dev_1 - dev_2 = 0 \quad vs \quad H_1 : dev_1 - dev_2 \neq 0$$

Los resultados se encuentran en el *anexo f*, en todos los casos no se rechazó la hipótesis nula: los parámetros adicionales son cero, con un nivel de significancia de 0.05. Por lo que se seleccionan los ajustes del modelo con menos categorías.

La variable Entidad entonces quedó con 20 categorías, en adelante se muestran los resultados considerando este agrupado. Con el fin de comparar las estimaciones del ajuste del modelo

ST_3MBIC aplicado a cada conjunto, en la *Tabla 21* se muestran las categorías y notación simplificada.

Tabla 21: Correspondencia de notación para las variables explicativas y sus categorías

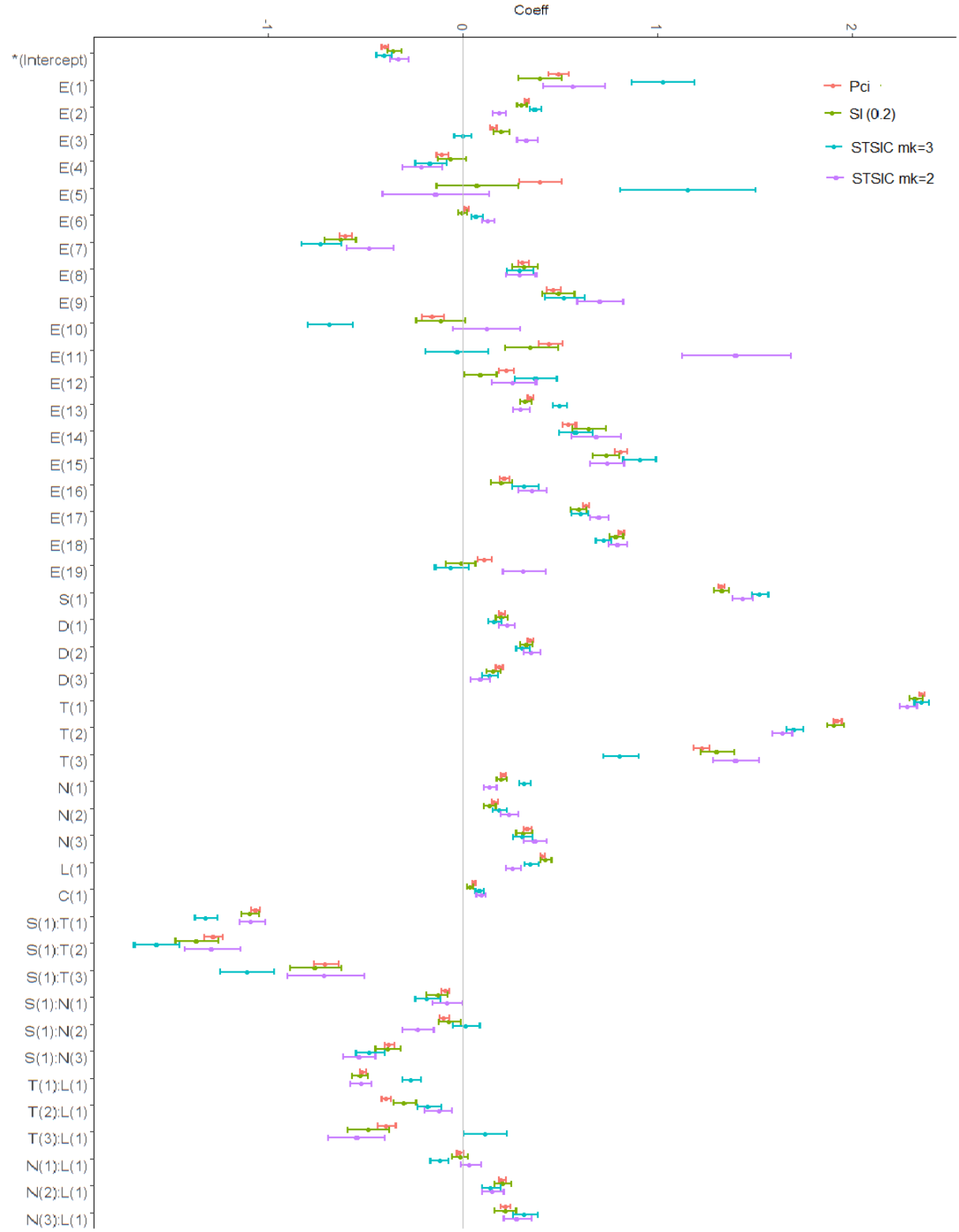
Descriptiva		Base Computacional		Simplificada	
Variable	**Categoría**	**Variable**	**Categoría**	**Var**	**Cat**
Sexo	**Hombre**	*SEXO*	**H**	*S*	0
	Mujer		M		1
Edad	**12 a 19**	*EDAD_r*	**12-19**	*D*	0
	20 a 29		20-29		1
	30 a 59		30-59		2
	60 y más		60+		3
Nivel Educativo	**Ninguno Primaria Incompleta**	NIVACAD_r	**Ning/Prim Inc**	N	0
	Primaria Completa		Prim Comp		1
	Secundaria Completa		Sec Comp		2
	Bachillerato Completo y más		zBach Comp y +		3
Situación Laboral	**Cuenta Propia**	SITUACION_TRAB_r	**Cuenta Propia**	T	0
	Empleado o Asistente		Empleado/Asis		1
	Jornalero o Peón		Jornal/Peon		2
	Patrón o Empleador		Patr/Empleador		3
Situación Conyugal	**No Casado**	SITUA_CONYUGAL_r	**No Casado**	C	0
	Casado		Casado		1
Lengua indígena	**Habla lengua indígena**	HLENGUA	**Si**	L	0
	No habla lengua indígena		No		1
Entidad	**Oaxaca, Tamaulipas**	ENT	**20,28**	E	0
	Coahuila		5		1
	Chiapas		7		2
	Chihuahua		8		3
	Durango		10		4
	Guanajuato		11		5
	Guerrero		12		6
	Hidalgo		13		7
	Jalisco		14		8
	México		15		9
	Michoacán		16		10
	Nayarit		18		11
	Nuevo León		19		12
	Puebla		21		13
	S.L. Potosí		24		14
	Sonora		26		15
	Tlaxcala		29		16
	Veracruz		30		17
	Yucatán		31		18
	Zacatecas		32		19

Categoría de referencia

El siguiente *Gráfico 11* muestra los intervalos de confianza para cada coeficiente estimado en la población de interés y tres muestras usando la notación simplificada de la *Tabla 21*. El

valor de los coeficientes puede encontrarse en el *anexo g*.

Gráfico 11. Intervalos al 95% confianza de coeficientes, ST_3MBIC (4 int)



La *Tabla 22* muestra los coeficientes estimados con la notación correspondiente a cada beta β numerada y las desviaciones estándar ajustados con el modelo *ST_3MBIC*, también usando la notación simplificada de la *Tabla 21.*

Tabla 22: Parámetros estimados del ajuste con el modelo *ST_3MBIC* en la población y las muestras

Var explicativa	*PI*		*SI 0.2*		STSIC3		STSIC2	
	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{\beta}$	$\hat{se}(\hat{\beta})$
β_0, intercept (**E**=0)	-0.403	0.0163	-0.355	0.0365	-0.407	0.1116	-0.330	0.1191
β_1, (**E**=1)	0.489	0.0520	0.392	0.1103	1.029	0.2983	0.569	0.2357
β_2, (**E**=2)	0.327	0.0105	0.303	0.0235	0.369	0.1500	0.186	0.1755
β_3, (**E**=3)	0.155	0.0169	0.198	0.0385	-0.002	0.1560	0.328	0.1564
β_4, (**E**=4)	-0.111	0.0332	-0.062	0.0739	-0.166	0.1657	-0.211	0.1730
β_5, (**E**=5)	0.395	0.1063	0.072	0.2137	1.156	0.4289	-0.142	0.6689
β_6, (**E**=6)	0.015	0.0115	-0.006	0.0257	0.070	0.1151	0.128	0.1203
β_7, (**E**=7)	-0.608	0.0348	-0.629	0.0798	-0.731	0.4424	-0.480	0.5583
β_8, (**E**=8)	0.309	0.0273	0.317	0.0619	0.293	0.1443	0.294	0.1872
β_9, (**E**=9)	0.463	0.0371	0.488	0.0836	0.520	0.1791	0.703	0.2821
β_{10},(**E**=10)	-0.159	0.0555	-0.117	0.1256	-0.681	0.2704	0.122	0.1378
β_{11},(**E**=11)	0.445	0.0621	0.347	0.1335	-0.032	0.4465	1.404	0.4777
β_{12},(**E**=12)	0.223	0.0395	0.088	0.0848	0.374	0.1594	0.261	0.1728
β_{13},(**E**=13)	0.349	0.0142	0.323	0.0315	0.493	0.0965	0.300	0.1090
β_{14},(**E**=14)	0.543	0.0370	0.646	0.0846	0.576	0.1659	0.686	0.1932
β_{15},(**E**=15)	0.814	0.0302	0.737	0.0668	0.906	0.1301	0.742	0.1181
β_{16},(**E**=16)	0.213	0.0245	0.200	0.0541	0.318	0.1489	0.357	0.2127
β_{17},(**E**=17)	0.632	0.0171	0.592	0.0381	0.602	0.1142	0.699	0.2108
β_{18},(**E**=18)	0.812	0.0161	0.788	0.0360	0.722	0.1089	0.796	0.1100
β_{19},(**E**=19) (**S**=0)	0.110	0.0363	-0.014	0.0784	-0.059	0.1302	0.312	0.1599
β_{20}, (**S**=1) (**D**=0)	1.328	0.0162	1.331	0.0363	1.525	0.0895	1.436	0.1141
β_{21}, (**D**=1)	0.197	0.0142	0.198	0.0318	0.165	0.0569	0.227	0.0584
β_{22}, (**D**=2)	0.347	0.0138	0.326	0.0310	0.310	0.0602	0.354	0.0701
β_{23}, (**D**=3) (**T**=0)	0.186	0.0169	0.157	0.0377	0.136	0.0699	0.087	0.0854
β_{24}, (**T**=1)	2.357	0.0146	2.329	0.0325	2.359	0.1182	2.286	0.1117
β_{25}, (**T**=2)	1.923	0.0174	1.911	0.0389	1.704	0.1544	1.641	0.1696
β_{26}, (**T**=3) (**N**=0)	1.229	0.0392	1.307	0.0901	0.808	0.2450	1.402	0.1914
β_{27}, (**N**=1)	0.206	0.0121	0.198	0.0271	0.319	0.0780	0.138	0.0684
β_{28}, (**N**=2)	0.162	0.0143	0.136	0.0320	0.188	0.0642	0.238	0.0745
β_{29}, (**N**=3) (**L**=0)	0.333	0.0191	0.314	0.0422	0.309	0.1083	0.370	0.1070
β_{30}, (**L**=1) (**C**=0)	0.405	0.0132	0.421	0.0295	0.350	0.0916	0.258	0.1025
β_{31}, (**C**=1)	0.056	0.0076	0.036	0.0170	0.085	0.0399	0.092	0.0486
β_{32}, (**S**=1)×(**T**=1)	-1.067	0.0222	-1.098	0.0489	-1.323	0.0982	-1.089	0.1089
β_{33}, (**S**=1)×(**T**=2)	-1.283	0.0491	-1.371	0.1095	-1.576	0.1817	-1.291	0.1870
β_{34}, (**S**=1)×(**T**=3)	-0.707	0.0611	-0.762	0.1314	-1.110	0.2368	-0.709	0.2455

β_{35}, (**S**=1)×(**N**=1)	-0.094	0.0243	-0.133	0.0539	-0.182	0.0924	-0.079	0.1113
β_{36}, (**S**=1)×(**N**=2)	-0.100	0.0261	-0.072	0.0581	0.015	0.0977	-0.232	0.1239
β_{37}, (**S**=1)×(**N**=3)	-0.380	0.0285	-0.386	0.0630	-0.481	0.1101	-0.537	0.1420
β_{38}, (**T**=1)×(**L**=1)	-0.520	0.0183	-0.533	0.0406	-0.265	0.1233	-0.525	0.1227
β_{39}, (**T**=2)×(**L**=1)	-0.398	0.0245	-0.299	0.0555	-0.175	0.1737	-0.126	0.1932
β_{40}, (**T**=3)×(**L**=1)	-0.394	0.0477	-0.488	0.1074	0.111	0.2759	-0.547	0.2349
β_{41}, (**N**=1)×(**L**=1)	-0.020	0.0177	-0.015	0.0396	-0.123	0.0928	0.036	0.0912
β_{42}, (**N**=2)×(**L**=1)	0.200	0.0192	0.204	0.0430	0.143	0.0793	0.151	0.0957
β_{43}, (**N**=3)×(**L**=1)	0.218	0.0241	0.218	0.0534	0.318	0.1225	0.280	0.1197

PI: población de interés
SI 0.2: muestra del 20 %
*STSIC*3: muestra *STSIC* con $m_k = 3$
*STSIC*2: muestra *STSIC* con $m_k = 2$

El modelo de regresión logística con la notación de la *Tabla 22* es:

$$\begin{aligned} logit[\pi(\boldsymbol{x})] =& \beta_0 + \beta_1 E(1) + \beta_2 E(2) + \beta_3 E(3) + \beta_4 E(4) + \beta_5 E(5) + \beta_6 E(6) \\ &+ \beta_7 E(7) + \beta_8 E(8) + \beta_9 E(9) + \beta_{10} E(10) + \beta_{11} E(11) \\ &+ \beta_{12} E(12) + \beta_{13} E(13) + \beta_{14} E(14) + \beta_{15} E(15) + \beta_{16} E(16) \\ &+ \beta_{17} E(17) + \beta_{18} E(18) + \beta_{19} E(19) + \beta_{20} S(1) + \beta_{21} D(1) \\ &+ \beta_{22} D(2)) + \beta_{23} D(3) + \beta_{24} T(1) + \beta_{25} T(2) + \beta_{26} T(3) \\ &+ \beta_{27} N(1) + \beta_{28} N(2) + \beta_{29} N(3)) + \beta_{30} L(1) + \beta_{31} C(1) \\ &+ \beta_{32} S(1) \times T(1) + \beta_{33} S(1) \times T(2) + \beta_{34} S(1) \times T(3) \\ &+ \beta_{35} S(1) \times N(1) + \beta_{36} S(1) \times N(2) + \beta_{37} S(1) \times N(3) \\ &+ \beta_{38} T(1) \times L(1) + \beta_{39} T(2) \times L(1) + \beta_{40} T(3) \times L(1) \\ &+ \beta_{41} N(1) \times L(1) + \beta_{42} N(2) \times L(1) + \beta_{43} N(3) \times L(1) \end{aligned}$$

La interpretación de los coeficientes no es directa pero usando la razón de momios podemos revisar algunos resultados, de forma similar a como se realiza en Fitzmaurice et al. (1997), p.423. Se presentan los siguientes como ejemplo:

Considerando a un individuo con la respuesta a las variables explicativas en su categoría de referencia, es decir de Oaxaca, 12-19 años, ninguna o primaria incompleta, trabajador por cuenta propia, no casado y hablante de lengua indígena; si es mujer multiplica en *PI*, 3.77 ($\exp\hat{\beta}_{20} = \exp(1.328)$) veces, en *SI 0.2*, 3.78 ($\exp\hat{\beta}_{20} = \exp(1.331)$) veces, en *STSIC*3, 4.60 ($\exp\hat{\beta}_{20} = \exp(1.525)$) veces y en *STSIC*2, 4.20 ($\exp\hat{\beta}_{20} = \exp(1.436)$) veces el momio de si responder a la pregunta ingreso que siendo hombre, con un intervalo de confianza ajustado del 95 % de (3.67, 3.88), (3.57, 4.02), (3.97, 5.32) y (3.48, 5.07) respectivamente.

De igual forma, considerando a un individuo con las respuestas en las categorías de referencia y solo cambiando los casos de la variable Situación Laboral, ser empleado o asistente multiplica en *PI*, 10.56 ($\exp\hat{\beta}_{24} = \exp(2.357)$) veces, en *SI 0.2*, 10.27

($\exp\hat{\beta}_{24} = \exp(2.329)$) veces, en *STSIC*3, 10.58 ($\exp\hat{\beta}_{24} = \exp(2.359)$) veces y en *STSIC*2, 9.84 ($\exp\hat{\beta}_{24} = \exp(2.286)$) veces el momio de si responder a la pregunta ingreso que siendo trabajador por cuenta propia, que es la de referencia (con un intervalo de confianza al 95 % de (10.31, 10.82), (9.73, 10.84), (8.71, 12.85) y (8.18, 11.82) respectivamente). Ahora, ser jornalero o peón multiplica en *PI*, 6.48 ($\exp\hat{\beta}_{25} = \exp(1.923)$) veces, en *SI 0.2*, 6.76 ($\exp\hat{\beta}_{25} = \exp(1.911)$) veces, en *STSIC*3, 5.50 ($\exp\hat{\beta}_{25} = \exp(1.704)$) veces y en *STSIC*2, 5.16 ($\exp\hat{\beta}_{25} = \exp(1.641)$) veces el momio de si responder a la pregunta ingreso que siendo trabajador por cuenta propia (con un intervalo del 95 % de (6.65, 7.04), (6.34, 7.21), (4.26, 7.08) y (3.90, 6.82) respectivamente). Finalmente, ser patrón o empleador multiplica en *PI*, 3.42 ($\exp\hat{\beta}_{26} = \exp(1.229)$) veces, en *SI 0.2*, 3.70 ($\exp\hat{\beta}_{26} = \exp(1.307)$) veces en *STSIC*3, 2.24 ($\exp\hat{\beta}_{26} = \exp(1.808)$) veces y en *STSIC*2, 4.06 ($\exp\hat{\beta}_{26} = \exp(1.402)$) veces el momio de si responder a la pregunta ingreso que siendo trabajador por cuenta propia (con un intervalo de confianza del 95 % de (3.20, 3.65), (3.19, 4.29), (1.50, 3.36) y (2.97, 5.57) respectivamente).

Por lo tanto, podemos observar que la interacción de Sexo×Situación Laboral implica que, considerando a un individuo con respuesta de las variables explicativas en la categoría de referencia, revisando los casos de la variable Sexo y Situación Laboral se tiene que ser mujer y empleado o asistente multiplica en *PI*, 13.71 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{24} + \hat{\beta}_{32}) = \exp(2.62)$) veces, en *SI 0.2*, 12.96 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{24} + \hat{\beta}_{32}) = \exp(2.56)$) veces, en *STSIC*3, 12.95 ($\exp(\hat{\beta}_{20}+\hat{\beta}_{24}+\hat{\beta}_{32}) = \exp(2.56)$) veces y en *STSIC*2, 13.92 ($\exp(\hat{\beta}_{20}+\hat{\beta}_{24}+\hat{\beta}_{32}) = \exp(2.13)$) veces el momio de si responder a la pregunta ingreso que siendo hombre y trabajador por cuenta propia, que son las categorías de referencia de estas variables (con un intervalo de confianza del 95 % de (12.56, 14.96), (10.68, 15.73), (7.83, 21.42) y (8.02, 24.13) respectivamente). De igual forma, ser mujer y jornalero o peón multiplica en *PI*, 7.16 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{25} + \hat{\beta}_{33}) = \exp(1.97)$) veces, en *SI 0.2*, 6.49 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{25} + \hat{\beta}_{33}) = \exp(1.87)$) veces, en *STSIC*3, 5.22 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{25} + \hat{\beta}_{33}) = \exp(1.65)$) veces y en *STSIC*2, 5.97 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{25} + \hat{\beta}_{33}) = \exp(1.79)$) veces el momio de si responder a la pregunta ingreso que siendo hombre y trabajador por cuenta propia 8con un intervalo de confianza del 95 % de (6.25, 8.20), (4.79, 8.80), (2.59, 10.52) y (2.75, 12.94) respectivamente). Por último y de forma similar, ser mujer y patrón o empleador multiplica en *PI*, 6.36 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{26} + \hat{\beta}_{34}) = \exp(1.85)$) veces, en *SI 0.2*, 6.53 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{26} + \hat{\beta}_{34}) = \exp(1.88)$) veces, en *STSIC*3, 3.40 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{26} + \hat{\beta}_{34}) = \exp(1.22)$) veces y en *STSIC*2, 8.41 ($\exp(\hat{\beta}_{20} + \hat{\beta}_{26} + \hat{\beta}_{34}) = \exp(2.13)$) veces el momio de si responder a la pregunta ingreso que siendo hombre y trabajador por cuenta propia (con un intervalo de confianza del 95 % de (5.25, 7.70), (4.27, 9.97), (1.33, 8.69) y (3.40, 20.81) respectivamente).

Como se puede observar, las estimaciones de los coeficientes en cada población y muestras suelen ser parecidas, sin embargo, los intervalos de confianza crecen considerablemente en las muestras *STSIC* dada su estimación considerando la muestra compleja.

Conclusión

Se ejemplificó el uso de modelos de regresión logística ajustándolos a una base de datos de gran tamaño. Esto tuvo complicaciones, ya que como se explica en Cox et al. (2018), los procedimientos de inferencia estadística usualmente relacionan la varianza de los estimadores con el número de observaciones en la muestra, lo que causa desviaciones estándar pequeñas cuando este número es muy grande. También existen problemas al preparar los datos computacionalmente, dado que se requiere mayor procesamiento de las observaciones y algún problema en la información puede ser pasado por alto e influir en el análisis y estimación. Existe *software* que permite la preparación de grandes bases de datos con menor costo computacional (como *H20* en *R*), que no se utilizó en este trabajo. Para disminuir ambos problemas, obtener una muestra aleatoria simple de las observaciones permitió utilizar los algoritmos automatizados de selección considerando modelos con interacciones de mayor grado.

Es importante notar que los métodos estadísticos para el ajuste del modelo de regresión aplicados a la población censada, de no haber sido tratada como una base de datos obtenida de una muestra de una superpoblación, se tendrían que haber cambiado por métodos de estimación considerando la información como una población finita (ya que es un censo), se encuentra más de esto en Heeringa et al. (2017). A pesar de que en el *Capítulo 3* no se consideró el diseño de la encuesta y se tomó solo como una base de datos de gran tamaño, se sospechó que las observaciones $\{y1, \ldots, y686, 243\}$ no eran independientes, ya que como se indica en Fitzmaurice et al. (1997): en información obtenida de encuestas sociales a gran escala donde la conglomeración ocurre naturalmente (familias, viviendas, vecindarios, etc.) la sobredispersión puede existir, debido a que las respuestas están relacionadas, lo que implica una mala estimación de la varianza, sobre este aspecto se encuentra más información en Fleiss et al. (2003), en el capítulo llamado *Analysis of Correlated Binary Data*, donde se señala que para una estimación más atinada podrían utilizarse diversos métodos, como por ejemplo los *generalized estimation equations* o los modelos mixtos. Así mismo, otros métodos de estimación de los coeficientes y varianza para la muestra pudieron ser a través del remuestreo (bootstrap, jacknife) o de forma similar a la que se realiza en Fitzmaurice et al. (1997).

En el *Capítulo 4* se aprovechó que existía registro de la estratificación y conglomeración usada para la población que se muestreo en la EIC2015. Fue una gran ventaja, como se apunta en Heeringa et al. (2017), p.38, para un análisis correcto de la información se requiere una correcta estratificación y conglomeramiento de la población. Esto permitió ejemplificar los métodos de ajuste de un modelo de regresión logística bajo la perspectiva de muestreo, dado que, en muestras obtenidas de esquemas probabilísticos complejos, suele existir un aumento en los valores de las desviaciones estándar porque se violan supuestos de independencia que otros métodos consideran.

En el trabajo no se tiene como objetivo buscar o seleccionar modelos predictivos, sino más bien ilustrar el ajuste de modelos descriptivos. Esta distinción entre modelos predictivos y no

predictivos la presenta por ejemplo Efron (2020). Sin embargo, es de interés observar el poder predictivo de los modelos ajustados a través de las tablas de clasificación. El análisis usando el modelo *ST_3MBIC* aplicado a la población y a las tres muestras se encuentra en el *anexo h*. De forma general, los modelos ajustados carecieron de poder predictivo: los individuos son pronosticados en su mayoría a la categoría de responder *sí* a la pregunta sobre el Ingreso ($Y = 1$).

Anexos

a) Modelos candidatos para la población de interés

Tabla 23: Comparación de los modelos candidatos en la población de interés

Nombre	Descripción	Int	*AIC*	*BIC*	dev
m1	Aditivo	0	549,844	550,222	549,778
m2BIC	Interacciones de dos (ENT aditivo) en *step* con *BIC*	11	543,733	544,648	543,573
m3BIC	Interacciones de tres (ENT aditivo) en *step* con *BIC*	14	543,517	**544,489**	543,347

b) Modelos candidatos para la muestra aleatoria simple del 20 % de la población de interés

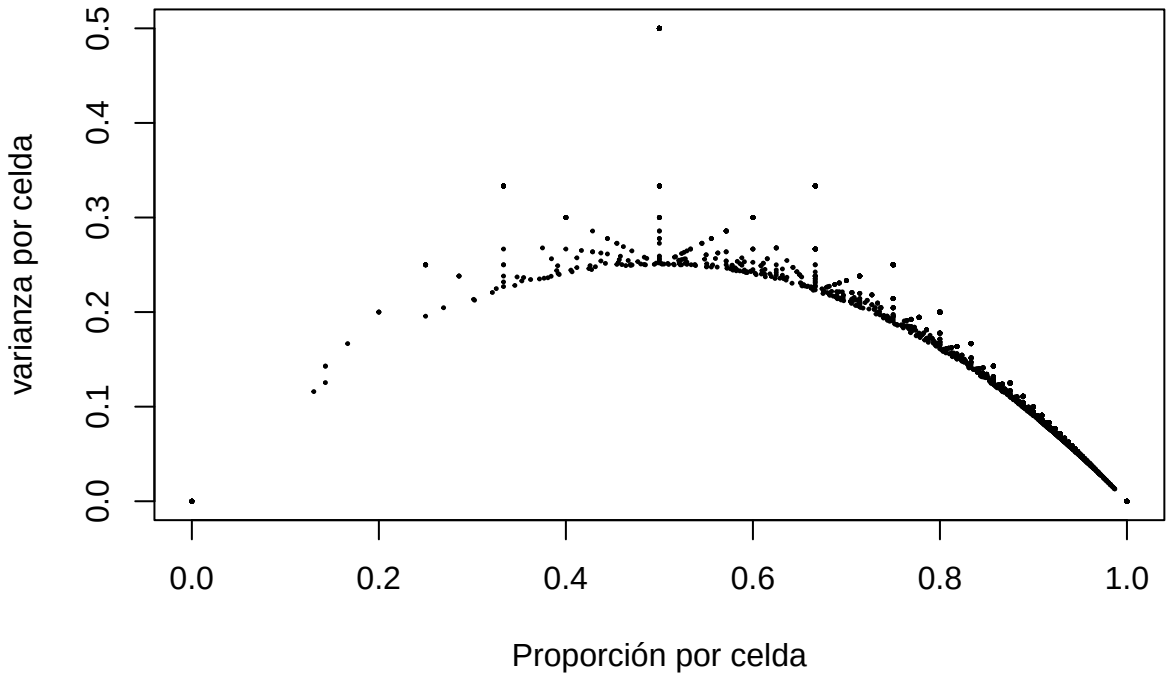
Tabla 24: Comparación de los modelos candidatos en la muestra *SI 0.2*

Nombre	Obtención	Int	*AIC*	*BIC*	dev
m1	Aditivo	0	110,380	110,705	110,314
m2BIC	Interacciones de dos con *BIC* en la población de interés	11	109,152	109,939	108,992
m3BIC	Interacciones de dos con *BIC* en la población de interés	14	109,121	109,956	108,951
m220BIC	Interacciones con *BIC* en muestra SI del 20 % de población de interés	5	109,327	**109,779**	109,235

c) Proporción de respuesta a Ingreso por celda en la tabla de frecuencia para la muestra *SI 0.2*

Considerando las siete variables y las categorías de cada una se tiene una tabla de frecuencia de 10,752 casos (celdas). En la muestra del 20 % de la población de interés (137,249 observaciones) se tiene que 5,950 de ellas están vacías, es decir, no tienen frecuencia. La proporción de respuesta a la variable Ingreso entonces solo puede ser calculada en 4,802 celdas, la varianza necesita por lo menos dos observaciones por celda, por lo que el siguiente gráfico compara la proporción con su varianza en las 3,770 celdas que tienen más de una observación.

Gráfico 12. Distribución por celda de la tabla de contingencia



Se observa que tiende a haber más respuestas positivas a Ingreso en general por celda.

d) Coeficientes y desviaciones estándar estimadas bajo el modelo *ST_3MBIC* en la población y muestras

	PI	SI	STSIC3	STSIC2
(Intercept)	-0.4036	-0.3549	-0.4061	-0.3305
SE	0.0163	0.0365	0.0409	0.0486
z	-24.70	-9.73	-9.92	-6.79
ENT5	0.4898	0.3924	1.0248	0.5706
SE	0.0521	0.1104	0.1620	0.1595
z	9.41	3.56	6.33	3.58
ENT7	0.3277	0.3030	0.3679	0.1866
SE	0.0105	0.0235	0.0262	0.0315
z	31.11	12.89	14.02	5.92
ENT8	0.15531	0.19862	-0.00536	0.32962
SE	0.01694	0.03851	0.04264	0.05219
z	9.17	5.16	-0.13	6.32
ENT10	-0.1103	-0.0611	-0.1698	-0.2098
SE	0.0332	0.0739	0.0794	0.0988
z	-3.32	-0.83	-2.14	-2.12

ENT11	0.3956	0.0726	1.1501	-0.1400
SE	0.1063	0.2137	0.3460	0.2746
z	3.72	0.34	3.32	-0.51
ENT12	0.01516	-0.00511	0.06769	0.12932
SE	0.01155	0.02575	0.02881	0.03451
z	1.31	-0.20	2.35	3.75
ENT13	-0.6074	-0.6289	-0.7340	-0.4793
SE	0.0348	0.0798	0.0985	0.1227
z	-17.45	-7.88	-7.45	-3.91
ENT14	0.3093	0.3181	0.2883	0.2957
SE	0.0273	0.0619	0.0693	0.0779
z	11.33	5.14	4.16	3.80
ENT15	0.4635	0.4892	0.5155	0.7045
SE	0.0371	0.0837	0.1021	0.1211
z	12.49	5.85	5.05	5.82
ENT16	-0.1579	-0.1168	-0.6854	0.1230
SE	0.0555	0.1256	0.1159	0.1715
z	-2.85	-0.93	-5.91	0.72
ENT18	0.4453	0.3474	-0.0347	1.4045
SE	0.0621	0.1335	0.1599	0.2831
z	7.17	2.60	-0.22	4.96
ENT19	0.2239	0.0884	0.3686	0.2623
SE	0.0395	0.0849	0.1068	0.1133
z	5.67	1.04	3.45	2.31
ENT21	0.3497	0.3238	0.4902	0.3006
SE	0.0142	0.0315	0.0376	0.0415
z	24.65	10.28	13.03	7.24
ENT24	0.5434	0.6469	0.5732	0.6869
SE	0.0370	0.0847	0.0900	0.1260
z	14.70	7.64	6.36	5.45
ENT26	0.8147	0.7375	0.9009	0.7435
SE	0.0302	0.0668	0.0802	0.0866
z	26.98	11.04	11.24	8.59
ENT28	0.0323	0.0351	-0.1904	0.0882
SE	0.0522	0.1166	0.1218	0.1728
z	0.62	0.30	-1.56	0.51
ENT29	0.2131	0.2003	0.3138	0.3579
SE	0.0245	0.0541	0.0658	0.0726
z	8.70	3.70	4.77	4.93
ENT30	0.6327	0.5929	0.5987	0.6996
SE	0.0171	0.0381	0.0430	0.0475
z	36.94	15.55	13.93	14.71
ENT31	0.8127	0.7884	0.7196	0.7964
SE	0.0161	0.0361	0.0397	0.0474
z	50.54	21.86	18.13	16.81

ENT32	0.1109	-0.0133	-0.0640	0.3140
SE	0.0363	0.0785	0.0862	0.1083
z	3.05	-0.17	-0.74	2.90
SEXOM	1.3276	1.3308	1.5246	1.4358
SE	0.0162	0.0363	0.0418	0.0497
z	81.98	36.62	36.51	28.88
EDAD_r20-29	0.1972	0.1980	0.1646	0.2268
SE	0.0142	0.0318	0.0359	0.0430
z	13.87	6.23	4.59	5.28
EDAD_r30-59	0.3468	0.3261	0.3104	0.3534
SE	0.0138	0.0310	0.0350	0.0418
z	25.04	10.53	8.88	8.46
EDAD_r60+	0.1862	0.1566	0.1366	0.0865
SE	0.0169	0.0377	0.0426	0.0505
z	11.01	4.15	3.20	1.71
SITUACION_TRAB_rEmpleado/Asis	2.3570	2.3292	2.3594	2.2861
SE	0.0146	0.0325	0.0368	0.0438
z	161.22	71.72	64.08	52.21
SITUACION_TRAB_rJornal/Peon	1.9226	1.9111	1.7045	1.6412
SE	0.0174	0.0389	0.0415	0.0514
z	110.36	49.15	41.03	31.95
SITUACION_TRAB_rPatr/Empleador	1.2291	1.3069	0.8078	1.4023
SE	0.0392	0.0901	0.0896	0.1132
z	31.37	14.51	9.02	12.39
NIVACAD_rPrim Comp	0.2062	0.1984	0.3195	0.1377
SE	0.0121	0.0271	0.0302	0.0362
z	17.04	7.34	10.58	3.80
NIVACAD_rSec Comp	0.1619	0.1359	0.1877	0.2383
SE	0.0143	0.0320	0.0353	0.0431
z	11.33	4.24	5.32	5.53
NIVACAD_rzBach Comp y +	0.3330	0.3139	0.3090	0.3698
SE	0.0191	0.0422	0.0480	0.0576
z	17.42	7.43	6.43	6.42
HLENGUANo	0.4048	0.4210	0.3517	0.2570
SE	0.0132	0.0295	0.0333	0.0387
z	30.66	14.25	10.57	6.64
SITUA_CONYUGAL_rCasado	0.0563	0.0362	0.0847	0.0924
SE	0.0076	0.0170	0.0191	0.0224
z	7.41	2.13	4.43	4.13
SEXOM:SITUACION_TRAB_rEmpleado/Asis	-1.0671	-1.0982	-1.3232	-1.0894
SE	0.0222	0.0489	0.0570	0.0662
z	-48.05	-22.46	-23.22	-16.47
SEXOM:SITUACION_TRAB_rJornal/Peon	-1.2830	-1.3703	-1.5781	-1.2902
SE	0.0491	0.1095	0.1155	0.1390

z	-26.15	-12.52	-13.66	-9.28
SEXOM:SITUACION_TRAB_rPatr/Empleador	-0.7066	-0.7620	-1.1104	-0.7092
SE	0.0611	0.1314	0.1401	0.1954
z	-11.57	-5.80	-7.93	-3.63
SEXOM:NIVACAD_rPrim Comp	-0.0939	-0.1335	-0.1820	-0.0795
SE	0.0243	0.0539	0.0629	0.0747
z	-3.86	-2.48	-2.89	-1.06
SEXOM:NIVACAD_rSec Comp	-0.0998	-0.0725	0.0151	-0.2318
SE	0.0261	0.0581	0.0674	0.0803
z	-3.83	-1.25	0.22	-2.89
SEXOM:NIVACAD_rzBach Comp y +	-0.3801	-0.3863	-0.4806	-0.5373
SE	0.0285	0.0630	0.0721	0.0849
z	-13.33	-6.13	-6.66	-6.33
SITUACION_TRAB_rEmpleado/Asis:HLENGUANo	-0.5200	-0.5328	-0.2638	-0.5257
SE	0.0183	0.0406	0.0463	0.0540
z	-28.45	-13.12	-5.69	-9.73
SITUACION_TRAB_rJornal/Peon:HLENGUANo	-0.3985	-0.3002	-0.1691	-0.1282
SE	0.0245	0.0556	0.0595	0.0717
z	-16.24	-5.40	-2.84	-1.79
SITUACION_TRAB_rPatr/Empleador:HLENGUANo	-0.3945	-0.4890	0.1146	-0.5480
SE	0.0477	0.1074	0.1115	0.1420
z	-8.27	-4.55	1.03	-3.86
NIVACAD_rPrim Comp:HLENGUANo	-0.0201	-0.0149	-0.1236	0.0358
SE	0.0177	0.0396	0.0446	0.0524
z	-1.14	-0.38	-2.77	0.68
NIVACAD_rSec Comp:HLENGUANo	0.2001	0.2043	0.1437	0.1508
SE	0.0192	0.0430	0.0480	0.0570
z	10.42	4.75	2.99	2.65
NIVACAD_rzBach Comp y +:HLENGUANo	0.2184	0.2185	0.3173	0.2808
SE	0.0241	0.0534	0.0607	0.0718
z	9.08	4.09	5.23	3.91

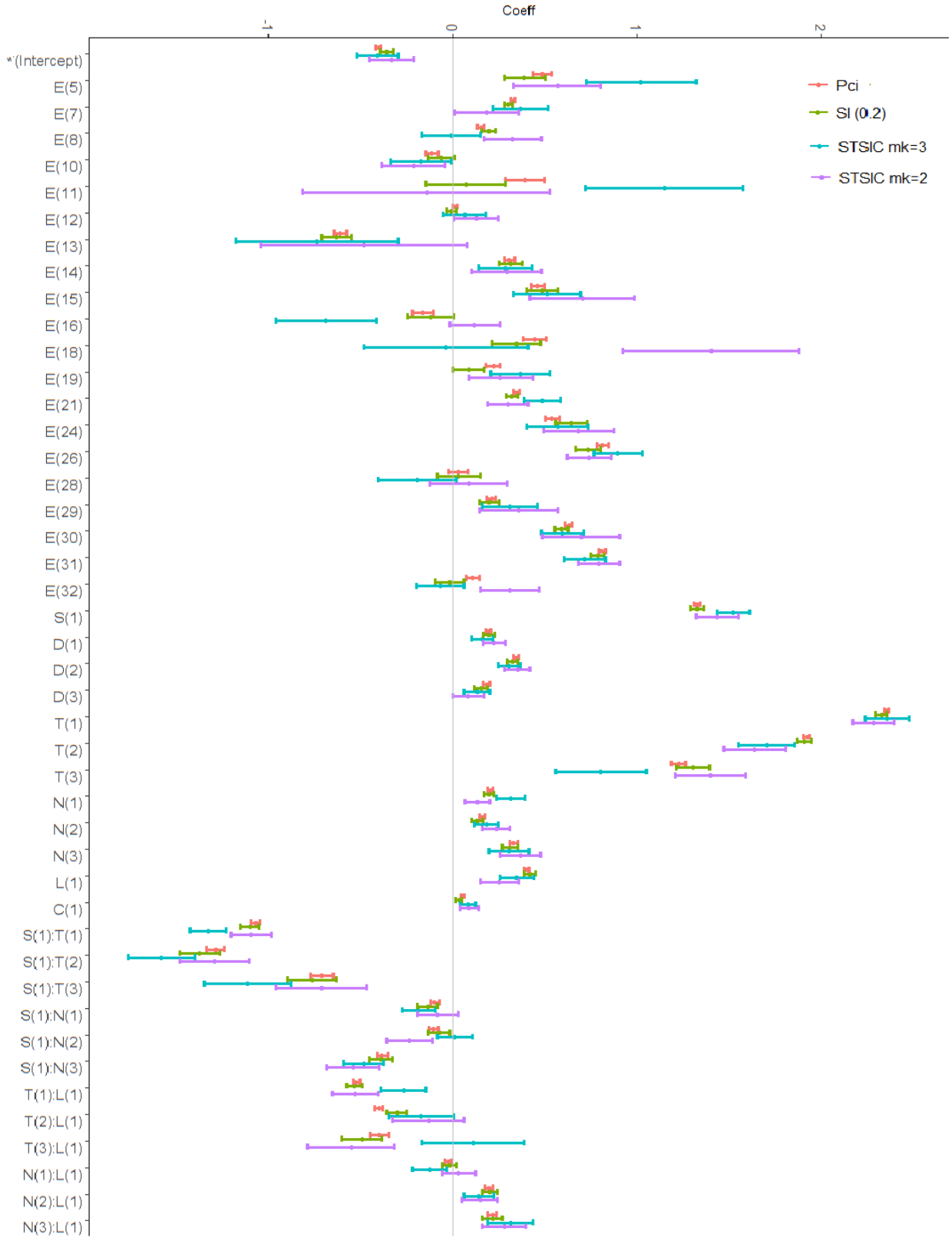
e) Intervalos de confianza para las estimaciones de los coeficientes del modelo *ST_3MBIC* ajustado a la población de interés y tres muestras

El siguiente *Gráfico 13* muestra el valor estimado de cada coeficiente y el respectivo intervalo de confianza, con la variable Entidad en 21 categorías, en cada grupo poblacional. Con el fin de facilitar la lectura de este gráfico la notación de las variables se cambió a la siguiente:

- Sexo, Hombre S(0), Mujer S(1).
- Edad, 12-19 D(0), 20-29 D(1), 30-59 D(2), 60+ D(3).
- Situación Laboral, Cuenta Propia T(0), Empleado Asistente T(1), Jornalero Peón T(2), Patrón Empleador T(3).
- Nivel Educativo, Ning/Prima Incom N(0), Prim Comp N(1), Sec Comp N(2), Bach comp y más N(3).

- Situación Conyugal, No casado C(0), Casado C(1).
- Lengua indígena, Si(0), No (1).
- Entidad, Oaxaca E(20), Coahuila E(5), Chiapas E(7), Chihuahua E(8), Durango E(10), Guanajuato E(11), Guerrero E(12), Hidalgo E(13), Jalisco E(14), México E(15), Michoacán E(16), Nayarit E(18), Nuevo León E(19), Puebla E(21), S.L. Potosí E(24), Sonora E(26), Tamaulipas E(28), Tlaxcala E(29), Veracruz E(30), Yucatán E(31), Zacatecas E(32).

Gráfico 13. Intervalos al 95% confianza de coeficientes, ST_3MBIC (4 int)
ENT no recategorizada



f) Pruebas de hipótesis en recategorización de variable Entidad

$$H_0 : dev_1 - dev_2 = 0 \quad vs \quad H_1 : dev_1 - dev_2 \neq 0.$$

```
# pi.ST_3MBIC: Ajuste de ST_3MBIC en población de interés sin recategorizar ENT
# pi.ST_3MBIC.2: Ajuste de ST_3MBIC en población de interés con ENT recategorizada
```

> anova(pi.ST_3MBIC,pi.ST_3MBIC.2,test="Chisq")

```
    Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1    686198     544818
2    686199     544818 -1 -0.38364   0.5357
```

```
# SI.ST_3MBIC: Ajuste de ST_3MBIC en muestra SI del 20\% sin recategorizar ENT
# SI.ST_3MBIC.2: Ajuste de ST_3MBIC en muestra SI del 20\% con ENT recategorizada
```

> anova(SI.ST_3MBIC,SI.ST_3MBIC.2,test="Chisq")

```
    Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1    137204     109259
2    137205     109259 -1 -0.091322   0.7625
```

```
# STSIC3.ST_3MBIC:  Ajuste de ST_3MBIC en muestra SITIC $m_k=3$ sin recategorizar ENT
# STSIC3.ST_3MBIC.2:  Ajuste de ST_3MBIC en muestra SITIC $m_k=3$ con ENT recategorizada
```

> anova(STSIC3.ST_3MBIC,STSIC3.ST_3MBIC.2,test="Chisq")

```
     Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
1     4610      86279
2     4611      86281     -1     -2.4    0.132
```

```
# STSIC2.ST_3MBIC: Ajuste de ST_3MBIC en muestra SITIC $m_k=2$ sin recategorizar ENT
# STSIC2.ST_3MBIC.2: Ajuste de ST_3MBIC en muestra SITIC $m_k=2$ con ENT recategorizada
```

> anova(STSIC2.ST_3MBIC,STSIC2.ST_3MBIC.2,test="Chisq")

```
     Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2417      61849
2      2418      61849    -1   -0.3   0.6129
```

---

g) Coeficientes y desviaciones estándar estimadas bajo el modelo *ST_3MBIC* en la población y muestras, con Entidad recategorizada

	PI	SI	STSIC3	STSIC2
(Intercept)	-0.4035	-0.3547	-0.4071	-0.3303
SE	0.0163	0.0365	0.0409	0.0486
z	-24.69	-9.72	-9.94	-6.79
ENT5	0.489	0.392	1.029	0.569
SE	0.052	0.110	0.162	0.160
z	9.40	3.55	6.36	3.57
ENT7	0.3274	0.3028	0.3695	0.1861
SE	0.0105	0.0235	0.0262	0.0315
z	31.11	12.89	14.09	5.91
ENT8	0.15478	0.19802	-0.00177	0.32847
SE	0.01692	0.03846	0.04257	0.05215
z	9.15	5.15	-0.04	6.30
ENT10	-0.1109	-0.0618	-0.1660	-0.2111
SE	0.0332	0.0739	0.0794	0.0988
z	-3.34	-0.84	-2.09	-2.14
ENT11	0.3947	0.0716	1.1562	-0.1419
SE	0.1063	0.2137	0.3460	0.2746
z	3.71	0.34	3.34	-0.52
ENT12	0.01475	-0.00557	0.07027	0.12843
SE	0.01153	0.02571	0.02876	0.03447
z	1.28	-0.22	2.44	3.73
ENT13	-0.6078	-0.6293	-0.7311	-0.4803
SE	0.0348	0.0798	0.0985	0.1227
z	-17.46	-7.89	-7.42	-3.91
ENT14	0.3086	0.3174	0.2928	0.2943
SE	0.0273	0.0619	0.0693	0.0778
z	11.31	5.13	4.23	3.78
ENT15	0.4628	0.4885	0.5199	0.7029
SE	0.0371	0.0836	0.1020	0.1211
z	12.47	5.84	5.10	5.81
ENT16	-0.1585	-0.1175	-0.6810	0.1216
SE	0.0555	0.1256	0.1159	0.1715
z	-2.86	-0.94	-5.88	0.71
ENT18	0.4451	0.3471	-0.0321	1.4037
SE	0.0621	0.1335	0.1598	0.2831
z	7.17	2.60	-0.20	4.96
ENT19	0.2231	0.0876	0.3735	0.2607
SE	0.0395	0.0848	0.1067	0.1133
z	5.65	1.03	3.50	2.30
ENT21	0.3492	0.3233	0.4934	0.2996
SE	0.0142	0.0315	0.0375	0.0415

z	24.65	10.27	13.14	7.22
ENT24	0.5429	0.6464	0.5764	0.6857
SE	0.0370	0.0846	0.0900	0.1259
z	14.69	7.64	6.40	5.44
ENT26	0.8140	0.7367	0.9055	0.7419
SE	0.0302	0.0668	0.0801	0.0865
z	26.97	11.03	11.30	8.58
ENT29	0.2125	0.1996	0.3177	0.3566
SE	0.0245	0.0541	0.0658	0.0725
z	8.68	3.69	4.83	4.92
ENT30	0.6322	0.5925	0.6016	0.6988
SE	0.0171	0.0381	0.0429	0.0475
z	36.94	15.55	14.02	14.70
ENT31	0.8123	0.7880	0.7219	0.7956
SE	0.0161	0.0360	0.0397	0.0474
z	50.55	21.86	18.20	16.80
ENT32	0.1102	-0.0141	-0.0592	0.3125
SE	0.0363	0.0784	0.0861	0.1083
z	3.04	-0.18	-0.69	2.89
SEXOM	1.3276	1.3308	1.5249	1.4357
SE	0.0162	0.0363	0.0418	0.0497
z	81.98	36.62	36.52	28.88
EDAD_r20-29	0.1972	0.1980	0.1647	0.2268
SE	0.0142	0.0318	0.0359	0.0430
z	13.87	6.23	4.59	5.28
EDAD_r30-59	0.3468	0.3262	0.3100	0.3536
SE	0.0138	0.0310	0.0350	0.0418
z	25.04	10.54	8.87	8.46
EDAD_r60+	0.1863	0.1568	0.1361	0.0869
SE	0.0169	0.0377	0.0426	0.0505
z	11.02	4.16	3.19	1.72
SITUACION_TRAB_rEmpleado/Asis	2.3570	2.3292	2.3592	2.2862
SE	0.0146	0.0325	0.0368	0.0438
z	161.22	71.73	64.08	52.22
SITUACION_TRAB_rJornal/Peon	1.9227	1.9111	1.7037	1.6414
SE	0.0174	0.0389	0.0415	0.0514
z	110.37	49.15	41.02	31.95
SITUACION_TRAB_rPatr/Empleador	1.2291	1.3069	0.8079	1.4023
SE	0.0392	0.0901	0.0896	0.1132
z	31.37	14.51	9.02	12.39
NIVACAD_rPrim Comp	0.2062	0.1984	0.3194	0.1377
SE	0.0121	0.0271	0.0302	0.0362
z	17.05	7.34	10.58	3.80
NIVACAD_rSec Comp	0.1619	0.1359	0.1876	0.2384

SE	0.0143	0.0320	0.0353	0.0431
z	11.33	4.25	5.32	5.53
NIVACAD_rzBach Comp y +	0.3331	0.3139	0.3086	0.3700
SE	0.0191	0.0422	0.0480	0.0576
z	17.42	7.44	6.43	6.43
HLENGUANo	0.4051	0.4214	0.3501	0.2576
SE	0.0132	0.0295	0.0333	0.0387
z	30.70	14.28	10.52	6.66
SITUA_CONYUGAL_rCasado	0.0563	0.0362	0.0846	0.0924
SE	0.0076	0.0170	0.0191	0.0224
z	7.41	2.13	4.42	4.13
SEXOM:SITUACION_TRAB_rEmpleado/Asis	-1.0671	-1.0981	-1.3233	-1.0894
SE	0.0222	0.0489	0.0570	0.0662
z	-48.05	-22.46	-23.22	-16.47
SEXOM:SITUACION_TRAB_rJornal/Peon	-1.2833	-1.3706	-1.5759	-1.2908
SE	0.0491	0.1095	0.1155	0.1390
z	-26.15	-12.52	-13.64	-9.29
SEXOM:SITUACION_TRAB_rPatr/Empleador	-0.7067	-0.7621	-1.1098	-0.7091
SE	0.0611	0.1314	0.1401	0.1954
z	-11.57	-5.80	-7.92	-3.63
SEXOM:NIVACAD_rPrim Comp	-0.0939	-0.1335	-0.1822	-0.0795
SE	0.0243	0.0539	0.0629	0.0747
z	-3.86	-2.48	-2.90	-1.06
SEXOM:NIVACAD_rSec Comp	-0.0998	-0.0724	0.0149	-0.2318
SE	0.0261	0.0581	0.0674	0.0803
z	-3.83	-1.25	0.22	-2.89
SEXOM:NIVACAD_rzBach Comp y +	-0.3801	-0.3863	-0.4806	-0.5372
SE	0.0285	0.0630	0.0721	0.0849
z	-13.33	-6.13	-6.66	-6.33
SITUACION_TRAB_rEmpleado/Asis:HLENGUANo	-0.5198	-0.5326	-0.2650	-0.5253
SE	0.0183	0.0406	0.0463	0.0540
z	-28.45	-13.12	-5.72	-9.72
SITUACION_TRAB_rJornal/Peon:HLENGUANo	-0.3977	-0.2993	-0.1752	-0.1258
SE	0.0245	0.0555	0.0593	0.0716
z	-16.23	-5.39	-2.95	-1.76
SITUACION_TRAB_rPatr/Empleador:HLENGUANo	-0.3941	-0.4884	0.1109	-0.5471
SE	0.0477	0.1074	0.1115	0.1420
z	-8.27	-4.55	0.99	-3.85
NIVACAD_rPrim Comp:HLENGUANo	-0.0202	-0.0150	-0.1233	0.0357
SE	0.0177	0.0396	0.0446	0.0524
z	-1.14	-0.38	-2.76	0.68
NIVACAD_rSec Comp:HLENGUANo	0.2002	0.2043	0.1435	0.1510
SE	0.0192	0.0430	0.0480	0.0570
z	10.43	4.75	2.99	2.65

NIVACAD_rzBach Comp y +:HLENGUANo	0.2183	0.2184	0.3179	0.2805
SE	0.0241	0.0534	0.0607	0.0718
z	9.08	4.09	5.24	3.91

h) Error predictivo del modelo *ST_3MBIC* aplicado a la población de interés y tres muestras

El poder predictivo de un modelo habla de la capacidad de acierto de la predicción de la observación i dada la variable respuesta estimada con el modelo $\hat{y}_i$ y comparándola con la respuesta real observada. A partir de una tabla cruzada de clasificación, se observa los datos mal clasificados en el modelo por categoría.

Tabla 25: Tabla de clasificación teórica de error predictivo

	Valores ajustados Y	
Valores observados Y	*0*	*1*
0	x_{11}	x_{12}
1	x_{21}	x_{22}

Los elementos que estén en la diagonal x_{21} y x_{12} serán las observaciones clasificadas de manera incorrecta por el modelo. Por lo que su porcentaje de error será el cociente de $x_{12} + x_{21}$ entre $x_{11} + x_{12} + x_{21} + x_{22}$

Se comparó el poder predictivo del modelo *(ST_3MBIC)* aplicado a la población de interés y tres muestras, se analizó la capacidad de predicción con el error usando tablas de clasificación.

Tabla 26: Error predictivo del modelo *ST_3MBIC*, en la población y las muestras

Valores observados en *PI*	**Valores ajustados** *No Respuesta*	*Respuesta*	**%Error**
No Respuesta	17,434	107,515	86.04
Respuesta	13,725	547,569	2.44
			17.70
PI: población de interés			
Valores observados en *SI 0.2*	**Valores ajustados** *No Respuesta*	*Respuesta*	**%Error**
No Respuesta	2,618	22,340	89.51
Respuesta	1,961	110,330	1.74
			17.66
SI 0.2: muestra del 20 %			
Valores observados en *STSIC3*	**Valores ajustados** *No Respuesta*	*Respuesta*	**%Error**
No Respuesta	3,626	18,635	83.71
Respuesta	2,589	85,603	2.93
			19.21
*STSIC*3: muestra *STSIC* con $m_k = 3$			
Valores observados en *STSIC2*	**Valores ajustados** *No Respuesta*	*Respuesta*	**%Error**
No Respuesta	1,508	14,567	90.61
Respuesta	1,181	60,724	1.90
			20.19
*STSIC*2: muestra *STSIC* con $m_k = 2$			

Aunque el error total en la población y las muestras es en promedio de 19 %, un error

predictivo bajo, el desempeño predictivo del modelo es muy pobre, esto es considerando que por categorías hay un error promedio en los valores observados con No respuesta del 87 %. Es decir, el modelo asigna a la mayoría de los casos como Respuesta cuando eran No respuesta. Para mejorar esta predicción podrían usarse métodos para optimizar los hiperparámetros como el punto de corte, sin embargo, no se realizó porque no es el fin del trabajo.

i) Código de R

Base de datos

#Librerías

```
library(readr)
library(dplyr)
library(anchors)
library(tidyr)
library(car)
library(gmodels)
library(plyr)
library(survey)
library(plotrix)
library(xtable)
library(doBy)
library(MASS)
library(sampling)
library(data.table)
```

##Cargar y filtrar cada base de datos por entidad y agruparlas

```
Estado <- read_csv("TR_PERSONA05.CSV",
          col_types = cols(COBERTURA = col_character(),
          ID_PERSONA = col_character(),
          ID_VIV = col_character(),
          UPM = col_character ()),
          locale = locale(encoding = "ISO-8859-1"))
Estado05 <- filter(Estado,COBERTURA==1)
```

#Se hace en cada estado y agrupa

```
BasePO1 <- rbind(Estado05,Estado07,Estado08,Estado10,Estado11,Estado12,Estado13,Estado14,
             Estado15,Estado16,Estado18,Estado19,Estado20,Estado21,Estado24,Estado26,
             Estado28,Estado29,Estado30,Estado31,Estado32)
BasePO1 <- dplyr::select(BasePO1,
         ENT , MUN, NOM_MUN, ESTRATO, UPM, ID_VIV, ID_PERSONA, FACTOR,
         TAMLOC, COBERTURA, SEXO, EDAD, ESCOACUM, CONACT,
         NIVACAD, ALFABET, PERTE_INDIGENA, HLENGUA, HESPANOL, ELENGUA,
         SITUACION_TRAB, OCUPACION_C,SITUA_CONYUGAL, AGUINALDO, VACACIONES,  SERVICIO_MEDICO,
         INGTRMEN, SAR_AFORE)
G1 <- c("EDAD","OCUPACION_C")
G2 <- c("ESCOACUM", "CONACT","NIVACAD")
G3 <- c("PERTE_INDIGENA", "HLENGUA", "HESPANOL", "ELENGUA", "ALFABET", "SAR_AFORE",
    "AGUINALDO", "VACACIONES", "SERVICIO_MEDICO","SITUACION_TRAB","SITUA_CONYUGAL")
G4 <- c("INGTRMEN")
BasePO1<-replace.value(BasePO1, G1, from = c(999),to = NA)
BasePO1<-replace.value(BasePO1, G2, from = c(99),to = NA)
BasePO1<-replace.value(BasePO1, G3, from = c(9),to = NA)
BasePO1<-replace.value(BasePO1, G4, from = c(999999),to = NA)
BasePO1<-filter(BasePO1, EDAD >= 12 & EDAD <=90 & SITUACION_TRAB <= 5 &
              (CONACT >= 10 & CONACT <= 16))
```

##Etiquetado y recategorización de variables

```
 BasePO1$TAMLOC <- factor(BasePO1$TAMLOC, levels = c(1:5), labels = c("< 2,500 habs",
                                                 "[2,500-15,000 habs)",
                                                 "[15,000-50,000 habs)",
                                                 "[50,000-100,000 habs)",
                                                 "[100,000 + habs"))
```

```
BasePO1$COBERTURA <- factor(BasePO1$COBERTURA, levels = c(1:3), labels = c("Censado",
                                                    "Muestreado", "Con Muestra insuficiente"))

BasePO1$PERTE_INDIGENA<-factor(BasePO1$PERTE_INDIGENA, levels=c(1,2,3,8), labels = c("Si",
                                                          "Si, en parte","No", "No sabe"))

BasePO1$HESPANOL <- factor(BasePO1$HESPANOL, levels = c(5,7), labels = c("Si","No"))

BasePO1$ELENGUA <- factor(BasePO1$ELENGUA, levels=c(1,3), labels = c("Si","No"))

BasePO1$ALFABET <- factor(BasePO1$ALFABET, levels=c(5,7), labels = c("Si","No"))

BasePO1$CONACT <- factor(BasePO1$CONACT, levels=c(10:16), labels=c("Trabajo",
                           "Hizo o vendio algún producto", "Ayudo en algun negocio",
                 "Crio animales o cultivo algo", "Ofrecio algun servicio por un pago",
                        "Atendio su propio negocio", "Tenia trabajo, pero no trabajo"))

BasePO1$AGUINALDO <- factor(BasePO1$AGUINALDO, levels = c(1:2),labels = c("Si","No"))

BasePO1$VACACIONES <- factor(BasePO1$VACACIONES, levels = c(3:4), labels = c("Si","No"))

BasePO1$SERVICIO_MEDICO <- factor(BasePO1$SERVICIO_MEDICO, levels = c(5:6), labels =c("Si","No"))

BasePO1$SAR_AFORE <- factor(BasePO1$SAR_AFORE, levels = c(3:4), labels =c("Si","No"))

BasePO1<-unite(BasePO1, MUN_r,c(1:2),  sep = "", remove = F)

BasePO1$SEXO <- factor(BasePO1$SEXO, levels = c(1,3), labels = c("H","M"))

BasePO1$HLENGUA <- factor(BasePO1$HLENGUA, levels = c(1,3), labels = c("Si","No"))

BasePO1$SITUA_CONYUGAL_r <- BasePO1$SITUA_CONYUGAL
BasePO1$SITUA_CONYUGAL_r[which(BasePO1$SITUA_CONYUGAL!=5)]<-1
BasePO1$SITUA_CONYUGAL_r <- factor(BasePO1$SITUA_CONYUGAL_r,levels=c(1,5),labels=c("No Casado",
                                                                          "Casado"))
BasePO1$SITUA_CONYUGAL <- factor(BasePO1$SITUA_CONYUGAL, levels=c(1:6),
                             labels = c("Union","Separado", "Divorcio",
                                        "Viudo","Casado","Soltero"))

BasePO1$SITUACION_TRAB_r <- BasePO1$SITUACION_TRAB
BasePO1$SITUACION_TRAB_r[which(BasePO1$SITUACION_TRAB==1 | BasePO1$SITUACION_TRAB==3)]<-0
BasePO1$SITUACION_TRAB_r[which(BasePO1$SITUACION_TRAB==5)]<-1
BasePO1$SITUACION_TRAB_r[which(BasePO1$SITUACION_TRAB==2)]<-2
BasePO1$SITUACION_TRAB_r[which(BasePO1$SITUACION_TRAB==4)]<-3
BasePO1$SITUACION_TRAB_r<- factor(BasePO1$SITUACION_TRAB_r,levels=c(0:3),
                              labels=c("Empleado/Asis","Cuenta Propia","Jornal/Peon",
                                       "Patr/Empleador"))

BasePO1$SITUACION_TRAB <- factor(BasePO1$SITUACION_TRAB, levels = c(1:5),
  labels = c("Empl/Obrero", "Jorna/Peon", "Ayudante", "Patr/Empleador", "Cuenta Propia"))


BasePO1$NIVACAD_r<-NA
BasePO1$NIVACAD_r[which( BasePO1$NIVACAD <= 2 &
                               (BasePO1$ESCOACUM < 6 |
                                  is.na(BasePO1$ESCOACUM) == TRUE ))] <- 1
BasePO1$NIVACAD_r[which( (BasePO1$NIVACAD == 2 |
                                BasePO1$NIVACAD == 6) &
                               BasePO1$ESCOACUM >= 6 &
                               is.na(BasePO1$ESCOACUM) == FALSE)] <- 2
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 3 &
                              BasePO1$ESCOACUM < 9 |
                              is.na(BasePO1$ESCOACUM) == TRUE)] <- 2
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 3 &
                              BasePO1$ESCOACUM == 9 |
                              is.na(BasePO1$ESCOACUM) == TRUE)] <- 3
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 4 &
                              BasePO1$ESCOACUM < 12 |
```

```
                              is.na(BasePO1$ESCOACUM) == TRUE)] <- 3
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 5 &
                              BasePO1$ESCOACUM < 12 |
                              is.na(BasePO1$ESCOACUM)==TRUE)] <- 3
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 7 &
                              BasePO1$ESCOACUM < 12 |
                              is.na(BasePO1$ESCOACUM) == TRUE)] <- 3
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 9 &
                              BasePO1$ESCOACUM < 12 |
                              is.na(BasePO1$ESCOACUM) == TRUE)] <- 3
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 4 &
                              BasePO1$ESCOACUM >= 12 &
                              is.na(BasePO1$ESCOACUM)== FALSE)] <- 4
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 5 &
                              BasePO1$ESCOACUM >= 12 &
                              is.na(BasePO1$ESCOACUM) == FALSE)] <- 4
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 7 &
                              BasePO1$ESCOACUM >= 12 &
                              is.na(BasePO1$ESCOACUM) == FALSE)] <- 4
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 8 &
                              BasePO1$ESCOACUM >= 12 &
                              is.na(BasePO1$ESCOACUM) == FALSE)] <- 4
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD == 9 &
                              BasePO1$ESCOACUM >= 12 &
                              is.na(BasePO1$ESCOACUM) == FALSE)] <- 4
BasePO1$NIVACAD_r[which(BasePO1$NIVACAD >= 10 &
                              BasePO1$ESCOACUM >=12 |
                              is.na(BasePO1$ESCOACUM) == TRUE) ] <- 4
BasePO1$NIVACAD_r[which(is.na(BasePO1$NIVACAD)== TRUE)] <- NA
BasePO1$NIVACAD_r<-factor(BasePO1$NIVACAD_r, levels  = c(1:4),
                              labels = c("Ning/Prim Inc",
                                         "Prim Comp","Sec Comp",
                                        "zBach Comp y +") )

BasePO1$NIVACAD<-factor(BasePO1$NIVACAD, levels = c(0:14),
  label = c("Ninguno", "Preescolar o kinder", "Primaria", "Secundaria",
        "Preparatoria o bachillerato general","Bachillerato tecnologico",
        "Estudios tecnicos o comerciales con primaria terminada",
        "Estudios tecnicos o comerciales con secundaria terminada",
        "Estudios tecnicos o comerciales con preparatoria terminada",
        "Normal con primaria o secundaria terminada", "Normal de licenciatura",
        "Licenciatura", "Especialidad", "Maestria", "Doctorado"))

BasePO1$EDAD_r <- cut(
BasePO1$EDAD, breaks =c(12,20,30,60,Inf),
 labels = c("12-19","20-29","30-59","60+"),
 include.highest = FALSE, right = FALSE)

BasePO1$EDAD_r1 <- cut(
BasePO1$EDAD, breaks =c(12,20,40,60,Inf),
  labels = c("12-19","20-39","40-59","60+"),
   include.highest = FALSE, right = FALSE)

BasePO1$EDAD_r2 <- cut(
BasePO1$EDAD, breaks =c(12,25,40,60,Inf),
  labels = c("12-24","24-39","40-59","60+"),
  include.highest = FALSE, right = FALSE)

BasePO1$INGTRMEN_RESP <- NA
BasePO1$INGTRMEN_RESP[which(is.na(BasePO1$INGTRMEN) == TRUE)] <- 0
BasePO1$INGTRMEN_RESP[which(is.na(BasePO1$INGTRMEN) == FALSE)] <- 1

BasePO1<-droplevels(BasePO1)


 # Eliminación de casos sin respuesta completa de 6 variables


Indice <-which(complete.cases(BasePO1[c("SEXO", "EDAD_r",
```

```
                                    "SITUACION_TRAB_r", "NIVACAD_r",
                                    "HLENGUA","SITUA_CONYUGAL_r")]))
BasePO1 <- BasePO1[Indice,]
```

#Exclusión de outlies de Ingreso y Situación Laboral

```
BasePO1%>%do(data.frame(t(quantile(.$INGTRMEN, probs = c(0.999),na.rm = T ))))
```

#quantile 99.9 en Ingreso de toda la población 38571

#23 se excluyen

```
BasePO1 <-BasePO1[-which(
(BasePO1$SITUACION_TRAB=="Ayudante" |
 BasePO1$SITUACION_TRAB=="Jorna/Peon") &
BasePO1$INGTRMEN>38571),]
```

#Se eliminan 95 más al analizar los que tienen ingreso mayor a 50,000

```
BasePO1<-filter(BasePO1, (is.na(INGTRMEN)==T) |
                (INGTRMEN<50000) |
                (INGTRMEN>=50000 &
                   EDAD_r!="12-19" &
                  NIVACAD_r!= "Ning/Prim Inc" &
                   ((SITUACION_TRAB_r=="Patr/Empleador" |
                       SITUACION_TRAB_r =="Cuenta Propia" ) |
                      (SITUACION_TRAB_r== "Empleado/Asis"
                       & AGUINALDO== "Si"
                       & SERVICIO_MEDICO == "Si"
                       & (NIVACAD_r=="Bach Comp y +" | OCUPACION_C==111)
                      ))))
```

#Variables factor

```
BasePO1$MUN_r <- as.factor(BasePO1$MUN_r)
BasePO1$ESTRATO <- as.factor(BasePO1$ESTRATO)
BasePO1$UPM <- as.factor(BasePO1$UPM)
BasePO1$SEXO<-as.factor(BasePO1$SEXO)
BasePO1$EDAD_r<-as.factor(BasePO1$EDAD_r)
BasePO1$EDAD_r1<-as.factor(BasePO1$EDAD_r1)
BasePO1$EDAD_r2<-as.factor(BasePO1$EDAD_r2)
BasePO1$ENT<-as.factor(BasePO1$ENT)
BasePO1$SITUACION_TRAB_r<-as.factor(BasePO1$SITUACION_TRAB_r)
BasePO1$SITUA_CONYUGAL_r<-as.factor(BasePO1$SITUA_CONYUGAL_r)
BasePO1$HLENGUA<-as.factor(BasePO1$HLENGUA)
BasePO1$NIVACAD_r<-as.factor(BasePO1$NIVACAD_r)
BasePO1$INGTRMEN_RESP<-as.factor(BasePO1$INGTRMEN_RESP)
BasePO1$ID_VIV<-as.character(BasePO1$ID_VIV)
BasePO1$ID_PERSONA<-as.character(BasePO1$ID_PERSONA)
```

#Orden para guardar variables

```
orden <- c(
 "ENT","MUN","MUN_r", "NOM_MUN", "ESTRATO", "UPM", "ID_VIV", "ID_PERSONA", "FACTOR",
 "TAMLOC", "COBERTURA", "SEXO", "EDAD", "EDAD_r", "EDAD_r1","EDAD_r2","SITUA_CONYUGAL","SITUA_CONYUGAL_r",
 "ESCOACUM", "NIVACAD", "NIVACAD_r","ALFABET", "HLENGUA", "HESPANOL", "ELENGUA","PERTE_INDIGENA",
 "OCUPACION_C", "CONACT", "SITUACION_TRAB", "SITUACION_TRAB_r", "AGUINALDO", "VACACIONES",
 "SERVICIO_MEDICO","SAR_AFORE","INGTRMEN","INGTRMEN_RESP")

Base<-BasePO1[,orden]
```

Análisis exploratorio

#Variables y respuesta a ingreso

```
CrossTable(Base$ENT,Base$SEXO,digits=1,prop.t=F,prop.chisq=F,format="SPSS")
CrossTable(Base$ENT,Base$EDAD_r,digits=1,prop.t=F,prop.chisq=F,format="SPSS")
CrossTable(Base$ENT,Base$NIVACAD_r,digits=1,prop.t=F,prop.chisq=F,format="SPSS")
CrossTable(Base$ENT,Base$SITUACION_TRAB_r,digits=1,prop.t=F,prop.chisq=F,format="SPSS")
CrossTable(Base$ENT,Base$SITUA_CONYUGAL_r,digits=1,prop.t=F,prop.chisq=F,format="SPSS")
CrossTable(Base$ENT,Base$HLENGUA,digits=1,prop.t=F,prop.chisq=F,format="SPSS")
CrossTable(Base$ENT,Base$INGTRMEN_RESP,digits=1,prop.t=F,prop.chisq=F,format="SPSS")

length(unique(Base$UPM))
length(unique(Base$ESTRATO))
```

#Análisis Ingreso

```
use=c("SEXO","INGTRMEN_RESP")
Base2<-Base[,use]
BaseT=as.table(ftable(Base2))
BaseDT.2=as.data.frame(BaseT)
rm(use,BaseT,Base2)
S<- ggplot(data=BaseDT.2,aes(x=SEXO,y=Freq, fill= INGTRMEN_RESP)) +
geom_bar(stat="identity") + labs(title=""Gráfico 2. Proporción observada ",
subtitle="respuesta Ingreso en Sexo ", x="Sexo", y="Población", fill="") +
scale_fill_manual(labels=c("No Respuesta","Respuesta"),values=c("black",
"grey"))+scale_y_continuous(labels=label_number(suffix="K",
scale=1e-3))+theme_classic()
S<-S+theme(plot.title = element_text(size=10))+theme(plot.subtitle =
element_text(size=9))+theme(legend.position="bottom")+theme(legend.position="bot
tom",legend.key.size=unit(.25,"cm"),legend.text=element_text(size=8))

use=c("EDAD_r","INGTRMEN_RESP")
 Base2<-Base[,use]
 BaseT=as.table(ftable(Base2))
 BaseDT.2=as.data.frame(BaseT)
 rm(use,BaseT,Base2)
 E<- ggplot(data=BaseDT.2,aes(x=EDAD_r,y=Freq, fill= INGTRMEN_RESP)) +
 geom_bar(stat="identity") + labs(title=""Gráfico 3. Proporción observada ",
 subtitle="respuesta Ingreso en Edad ", x="Edad", y="Población", fill="") +
 scale_fill_manual(labels=c("No Respuesta","Respuesta"),values=c("black",
 "grey"))+scale_y_continuous(labels=label_number(suffix="K",
 scale=1e-3))+theme_classic()
E<-E+theme(plot.title = element_text(size=10))+theme(plot.subtitle =
element_text(size=9))+theme(legend.position="bottom")+theme(legend.position="bot
tom",legend.key.size=unit(.25,"cm"),legend.text=element_text(size=8))


#Lo mismo con cada variable
```

Métodos automatizados de selección del modelo en población de interés

Establecer categoría de referencia

```
Base$NIVACAD_r <- relevel(Base$NIVACAD_r,ref="Ning/Prim Inc")
Base$HLENGUA <- relevel (Base$HLENGUA,ref="Si")
Base$SITUA_CONYUGAL_r <- relevel (Base$SITUA_CONYUGAL_r,ref="No Casado")
Base$ENT <- relevel(Base$ENT,ref="20")
```

```
Base$SITUACION_TRAB_r <- relevel (Base$SITUACION_TRAB_r,ref="Cuenta Propia")
Base$EDAD_r <- relevel(Base$EDAD_r,ref="12-19")
Base$SEXO <- relevel(Base$SEXO,ref="H")
```

Tabla de contingencia

```
use=c("SEXO", "EDAD_r",
  "SITUACION_TRAB_r", "NIVACAD_r",
  "HLENGUA","SITUA_CONYUGAL_r","ENT","INGTRMEN_RESP")
Base2<-Base[,use]
BaseT=as.table(ftable(Base2))
BaseDT=as.data.frame(BaseT)
rm(use,BaseT,Base2)
```

Modelo aditivo simple

```
m1<-glm(INGTRMEN_RESP ~ SEXO+ EDAD_r + SITUACION_TRAB_r + NIVACAD_r +  HLENGUA
    + SITUA_CONYUGAL_r + ENT, family=binomial,weight=Freq ,data=BaseDT)
```

Interacciones de segundo orden

```
m2BIC=step(glm(INGTRMEN_RESP~ ENT  +
             (SEXO + EDAD_r + SITUACION_TRAB_r
              + NIVACAD_r+ HLENGUA+ SITUA_CONYUGAL_r)^2,
           family=binomial,data=BaseDT,weight=Freq), k=log(sum(BaseDT$Freq)))
```

Interacciones de tercer orden

```
m3BIC=step(glm(INGTRMEN_RESP~ ENT +
             (SEXO + EDAD_r + SITUACION_TRAB_r
              + NIVACAD_r+ HLENGUA + SITUA_CONYUGAL_r )^3,
           family=binomial,data=BaseDT,weight=Freq), k=log(sum(BaseDT$Freq)))

anova(m1L, test="Chisq")
anova(m2BIC, test="Chisq")
anova(m3BIC, test="Chisq")
```

Selección y ajuste a la muestra SI de la población de interés

---

Selección de muestra

```
 veinte<-sample(1:nrow( Base ), round(nrow(Base)*.2,0))
 Base.20 <- Base[ veinte, ]
 rm(,veinte)

 m1L.20=glm(INGTRMEN_RESP ~ SEXO+ EDAD_r + SITUACION_TRAB_r + NIVACAD_r +  HLENGUA
       + SITUA_CONYUGAL_r + ENT, family=binomial, data=Base.20)
```

Proceso automatizado de selección de modelo de regresión logística a muestra del 20 % de la población de interés, todas las interacciones posibles

```
m220BIC <- step(glm(INGTRMEN_RESP~ ENT + (SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r + HLENGUA +
SITUA_CONYUGAL_r)^6, family=binomial, data=BaseDT.20, weight=Freq), direction="both", k=log(137249))
```

Modelos anteriores aplicado a muestra del 20 % de la población de interés

```
m2BICL.20=glm(formula=formula(m2BIC),family=binomial,data=Base.20)
m3BICL.20=glm(formula=formula(m3BIC),family=binomial,data=Base.20)

drop1(m1L.20, test="Chisq")
drop1(m2BICL.20, test="Chisq")
drop1(m3BICL.20, test="Chisq")
drop1(m220BIC, test="Chisq")
```

Selección de muestras *STSIC* y ajuste de modelo considerando su diseño

---

```
# Muestra STSIC m_k=3


n__=3
Base_3.a<- Base %>%
group_by(ESTRATO) %>%
mutate(Nh = n_distinct(UPM)) %>%
select(ESTRATO,UPM,Nh) %>%
filter(Nh==1) %>%
distinct()  %>%
mutate(Probs = as.numeric(ifelse(n_distinct(UPM)>=n__, n__/n_distinct(UPM), 1))) %>%
arrange(ESTRATO)
Base_3.a1<- Base %>%
group_by(ESTRATO) %>%
mutate(Nh = n_distinct(UPM)) %>%
select(ESTRATO,UPM,Nh) %>%
filter(Nh==2) %>%
distinct()  %>%
mutate(Probs = as.numeric(ifelse(n_distinct(UPM)>=n__, n__/n_distinct(UPM), 1))) %>%
arrange(ESTRATO)
Base_3.b<- Base %>%
group_by(ESTRATO) %>%
mutate(Nh = n_distinct(UPM)) %>%
select(ESTRATO,UPM,Nh) %>%
filter(Nh>=n__) %>%
distinct()  %>%
mutate(Probs = as.numeric(ifelse(n_distinct(UPM)>=n__, n__/n_distinct(UPM), 1))) %>%
arrange(ESTRATO)
ST_3.a <- sampling:::strata(Base_3.a, stratanames="ESTRATO",
                      size=c(rep(1,length(unique(Base_3.a$ESTRATO)))),
                      method=c("srswor"),description=F)
ST_3.a<-as.data.frame(getdata(Base_3.a, ST_3.a))

ST_3.a1 <- sampling:::strata(Base_3.a1, stratanames="ESTRATO",
                      size=c(rep(2,length(unique(Base_3.a$ESTRATO)))),
                      method=c("srswor"),description=F)
ST_3.a1<-as.data.frame(getdata(Base_3.a1, ST_3.a1))
set.seed(2323)
ST_3.b <- sampling:::strata(Base_3.b, stratanames="ESTRATO",
                      size=c(rep(n__,length(unique(Base_3.b$ESTRATO)))),
                      method=c("srswor"),description=F)
ST_3.b<-as.data.frame(getdata(Base_3.b, ST_3.b))
ST_3=rbind(ST_3.a,ST_3.a1,ST_3.b)%>% mutate(W = 1/Prob)
rm(Base_3.a,Base_3.b,Base_3.a1,ST_3.a,ST_3.b, ST_3.a1)
Base_3.ST<- subset(Base, UPM %in% ST_3$UPM) %>%  group_by(UPM) %>%
merge(y=select(ST_3,UPM,Prob,W), by.x = "UPM")

ST_3svy<- svydesign(
 id = ~UPM ,
  strata = ~ESTRATO ,
  data = Base_3.ST ,
  weights = ~W)


# Muestra STSIC m_k=2
```

```
n_=2
Base.a<- Base %>%
 group_by(ESTRATO) %>%
 mutate(Nh = n_distinct(UPM)) %>%
select(ESTRATO,UPM,Nh) %>%
filter(Nh<n_) %>%
distinct()  %>%
mutate(Probs = as.numeric(1)) %>%
arrange(ESTRATO)
Base.b<- Base %>%
group_by(ESTRATO) %>%
mutate(Nh = n_distinct(UPM)) %>%
select(ESTRATO,UPM,Nh) %>%
filter(Nh>=n_) %>%
distinct()  %>%
mutate(Probs = as.numeric(n_/n_distinct(UPM))) %>%
arrange(ESTRATO)

ST.a <- sampling:::strata(Base.a, stratanames="ESTRATO",
                     size=c(rep(1,length(unique(Base.a$ESTRATO)))),
                     method=c("srswor"),description=F)
ST.a<-as.data.frame(getdata(Base.a, ST.a))
set.seed(2323)
ST.b <- sampling:::strata(Base.b, stratanames="ESTRATO",
                      size=c(rep(n_,length(unique(Base.b$ESTRATO)))),
                      method=c("srswor"),description=F)
ST.b<-as.data.frame(getdata(Base.b, ST.b))
ST=rbind(ST.a,ST.b)%>% mutate(W = 1/Prob)
rm(Base.a,Base.b,ST.a,ST.b)

Base_2.ST<- subset(Base, UPM %in% ST$UPM) %>%  group_by(UPM) %>%
 merge(y=select(ST,UPM,Prob,W), by.x = "UPM")

ST_2svy<- svydesign(
 id = ~UPM ,
 strata = ~ESTRATO ,
 data = Base_2.ST ,
 weights = ~W)
```

Selección de modelo de regresión logística en Muestra STSIC m_k=3

```
options(survey.lonely.psu="adjust")
ST_3MBIC <- step(svyglm(INGTRMEN_RESP~ ENT +
            (SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r+ HLENGUA+
            SITUA_CONYUGAL_r)^3,
            family=binomial,design=ST_3svy), k=log(nrow(Base_3.ST)),trace=F)

 Anova(ST_3MBIC,type="II",test="Chisq")
```

#Selección de modelo de regresión logística en Muestra STSIC m_k=2

```
options(survey.lonely.psu="adjust")
ST_2MBIC <- step(svyglm(INGTRMEN_RESP~ ENT +
            (SEXO + EDAD_r + SITUACION_TRAB_r + NIVACAD_r+ HLENGUA+
             SITUA_CONYUGAL_r)^3,
            family=binomial,design=ST_2svy), k=log(nrow(Base_2.ST)),trace=F)

Anova(ST_2MBIC,type="II",test="Chisq")
```

Análisis de modelo ajustado en la población y muestras

Ajuste del modelo ST_3MBIC en población y muestras

```
pi.ST_3MBIC <- glm(formula=formula(ST_3MBIC), family=binomial, data=Base)
drop1(pi.ST_3BIC,test="Chisq", k=log(nrow(Base)))

SI.ST_3MBIC <- glm(formula=formula(ST_3MBIC), family=binomial, data=Base.20)
drop1(SI.ST_3BIC,test="Chisq", k=log(nrow(Base.20)))

STSIC3.ST_3MBIC <- svyglm(formula=formula(ST_3MBIC), family=binomial, design=ST_3svy)
Anova(STSIC3.ST_3MBIC,type="II")

STSIC2.ST_3MBIC <- svyglm(formula=formula(ST_3MBIC), family=binomial, design=ST_2svy)
Anova(STSIC2.ST_3MBIC,type="II")


 compareCoefs(pi.ST_3MBIC,SI.ST_3MBIC,STSIC3.ST_3MBIC,STSIC2.ST_3MBIC, se = T,
 zvals=T,print = T, digits = 3)
```

Recategorización de variable ENT

```
BaseNC <- Base
table(BaseNC$ENT)
levels(BaseNC$ENT)[c(1,17)]="20,28"
table(BaseNC$ENT)
BaseNC$ENT <- relevel(BaseNC$ENT,ref="20,12")
pi.ST_3MBIC.2<-glm(formula=formula(ST_3MBIC),family="binomial",data=BaseNC)
anova(pi.ST_3MBIC,pi.ST_3MBIC.2,test="Chisq")

BaseNC <- Base.20
table(BaseNC$ENT)
levels(BaseNC$ENT)[c(1,17)]="20,12"
table(BaseNC$ENT)
BaseNC$ENT <- relevel(BaseNC$ENT,ref="20,28")
SI.ST_3MBIC.2<-glm(formula=formula(ST_3MBIC),family="binomial",data=BaseNC)
anova(SI.ST_3MBIC,SI.ST_3MBIC.2,test="Chisq")

BaseNC <- Base_3.ST
table(BaseNC$ENT)
levels(BaseNC$ENT)[c(1,17)]="20,28"
table(Base.NC$ENT)
BaseNC$ENT <- relevel(BaseNC$ENT,ref="20,28")
STsvySS<- svydesign(
id = ~UPM ,
strata = ~ESTRATO ,
data = BaseNC ,
weights = ~W)
STSIC3.ST_3MBIC.2<-svyglm(formula(ST_3MBIC),family="binomial",design=STsvySS)
anova(STSIC3.ST_3MBIC,STSIC3.ST_3MBIC.2,test="Chisq")

BaseNC <- Base_2.ST
table(BaseNC$ENT)
levels(BaseNC$ENT)[c(1,17)]="20,28"
table(Base.NC$ENT)
BaseNC$ENT <- relevel(BaseNC$ENT,ref="20,28")
STsvyS<- svydesign(
id = ~UPM ,
strata = ~ESTRATO ,
data = BaseNC ,
weights = ~W)
STSIC2.ST_3MBIC.2<-svyglm(formula(ST_3MBIC),family="binomial",design=STsvyS)
anova(STSIC2.ST_3MBIC,STSIC2.ST_3MBIC.2,test="Chisq")

compareCoefs(pi.ST_3MBIC.2,SI.ST_3MBIC.2,STSIC3.ST_3MBIC.2,STSIC2.ST_3MBIC.2,
se = T, zvals=T,print = T, digits = 3)
```

#Gráfico 11. Coeficientes e intervalos de confianza del modelo ST_3MBIC en población y muestras, ENT recategorizada

```
sm=summary(pi.ST_3MBIC.2)
```

```
sc <- as.data.frame(sm$coefficients)$'Estimate'
sd <- as.data.frame(sm$coefficients)$'Std. Error'
Sest3 <- data.frame(sc,sd)

sma=summary(SI.ST_3MBIC.2)
sc <- as.data.frame(sma$coefficients)$'Estimate'
sd <- as.data.frame(sma$coefficients)$'Std. Error'
Sesta3 <- data.frame(sc,sd)

tsm=summary(STSIC3.ST_3MBIC.2)
sc <- as.data.frame(tsm$coefficients)$'Estimate'
sd <- as.data.frame(tsm$coefficients)$'Std. Error'
tSest3 <- data.frame(sc,sd)

stm=summary(STSIC2.ST_3MBIC.2)
sc <- as.data.frame(stm$coefficients)$'Estimate'
sd <- as.data.frame(stm$coefficients)$'Std. Error'
Sestt3 <- data.frame(sc,sd)


Vare=c("*(Intercept)", "E(1)", "E(2)", "E(3)","E(4)", "E(5)", "E(6)", "E(7)",
"E(8)", "E(9)", "E(10)", "E(11)",  "E(12)",  "E(13)",  "E(14)", "E(15)",  "E(16)",
"E(17)", "E(18)", "E(19)", "S(1)","D(1)",  "D(2)",   "D(3)",  "T(1)", "T(2)",
"T(3)", "N(1)",  "N(2)",  "N(3)", "L(1)",   "C(1)", "S(1):T(1)",  "S(1):T(2)",
"S(1):T(3)",  "S(1):N(1)",  "S(1):N(2)",  "S(1):N(3)",  "T(1):L(1)", "T(2):L(1)",
"T(3):L(1)", "N(1):L(1)", "N(2):L(1)",  "N(3):L(1)")

Sest3<-Sest3 %>%   mutate(Var=Vare, Pob=rep("PI",44))
Sesta3<-Sesta3 %>% mutate(Var=Vare, Pob=rep("SI (0.2)",44))
tSest3<-tSest3 %>% mutate(Var=Vare,Pob=rep("STSI mk=3",44))
Sestt3<-Sestt3 %>% mutate(Var=Vare,Pob=rep("STSI mk=2",44))
Pob <- c("PI","SI (0.2)","STSI mk=3","STSI mk=2")
cols <- c("sc","sd","Var","Pob")
colnames(Sesta3) <- cols
colnames(Sestt3) <- cols
colnames(tSest3) <- cols
S_3<-rbind(Sest3,Sesta3,tSest3,Sestt3)
S_3$Var <- factor(S_3$Var, levels = Vare)
S_3$Pob <- factor(S_3$Pob, levels = Pob)

(GrafST_3MBIC.2<-ggplot(S_3,aes(x=Var,y=sc,color=Pob)) +geom_point(position =
position_dodge(width = 0.6),size=1) +theme_classic() +
geom_errorbar(aes(ymin = sc+sd, ymax = (sc-sd)), width = 0.5,position =
position_dodge(width = 0.6)) +
labs(title="Gráfico 11. Intervalos al 95% confianza de coeficientes, ST_3MBIC
ENT recategorizado", x="", y="Coeff", color="")+
theme(plot.title = element_text(size=12),axis.text.x=element_text(angle=90,hjust=1)) +
geom_hline(yintercept=0,color="gray") + theme(legend.position = "top")
```

Anexos

#"Gráfico de proporción y varianza por celda

```
Base.20R <- dplyr::select(Base.20, INGTRMEN_RESP, ENT, SEXO, EDAD_r,
SITUACION_TRAB_r, NIVACAD_r,HLENGUA, SITUA_CONYUGAL_r)
Base.20R<-as.data.frame(Base.20R)
Base.20R$INGTRMEN_RESP <-
as.numeric(levels(Base.20R$INGTRMEN_RESP)[Base.20R$INGTRMEN_RESP])
Means=tapply(Base.20R[,1], Base.20R[,2:8],mean)
Vars=tapply(Base.20R[,1], Base.20R[,2:8],var)
SD=sqrt(Vars)
plot(Means,Vars, xlab="proporción por celda", ylab="varianza por celda", pch=19,
cex=.2)
plot(Means,SD, xlab="proporción por celda", ylab="des est por celda", pch=19, cex=.2)
```

#Coeficientes e intervalos de confianza del modelo ST_3MBIC en población y muestras

```
sm=summary(pi.ST_3MBIC)
sc <- as.data.frame(sm$coefficients)$'Estimate'
sd <- as.data.frame(sm$coefficients)$'Std. Error'
Sest3 <- data.frame(sc,sd)

sm=summary(SI.ST_3MBIC)
sc <- as.data.frame(sma$coefficients)$'Estimate'
sd <- as.data.frame(sma$coefficients)$'Std. Error'
Sesta3 <- data.frame(sc,sd)

summary(STSIC3.ST_3MBIC)
sc <- as.data.frame(tsm$coefficients)$'Estimate'
sd <- as.data.frame(tsm$coefficients)$'Std. Error'
tSest3 <- data.frame(sc,sd)

summary(STSIC2.ST_3MBIC)
sc <- as.data.frame(stm$coefficients)$'Estimate'
sd <- as.data.frame(stm$coefficients)$'Std. Error'
Sestt3 <- data.frame(sc,sd)


Vare=c("*(Intercept)", "E(5)", "E(7)", "E(8)","E(10)", "E(11)", "E(12)", "E(13)",
"E(14)", "E(15)", "E(16)", "E(18)",  "E(19)",  "E(21)",  "E(24)", "E(26)",  "E(28)",
"E(29)", "E(30)", "E(31)", "E(32)", "S(1)","D(1)",  "D(2)",   "D(3)",  "T(1)", "T(2)",
"T(3)", "N(1)",  "N(2)",  "N(3)", "L(1)",   "C(1)", "S(1):T(1)",  "S(1):T(2)",
"S(1):T(3)",  "S(1):N(1)",  "S(1):N(2)",  "S(1):N(3)",  "T(1):L(1)", "T(2):L(1)",
"T(3):L(1)", "N(1):L(1)", "N(2):L(1)",  "N(3):L(1)")

Sest3<-Sest3 %>%   mutate(Var=Vare, Pob=rep("PI",45))
Sesta3<-Sesta3 %>% mutate(Var=Vare, Pob=rep("SI (0.2)",45))
tSest3<-tSest3 %>% mutate(Var=Vare,Pob=rep("STSI mk=3",45))
Sestt3<-Sestt3 %>% mutate(Var=Vare,Pob=rep("STSI mk=2",45))
Pob <- c("PI","SI (0.2)","STSI mk=3","STSI mk=2")
cols <- c("sc","sd","Var","Pob")
colnames(Sesta3) <- cols
colnames(Sestt3) <- cols
colnames(tSest3) <- cols
S_3<-rbind(Sest3,Sesta3,tSest3,Sestt3)
S_3$Var <- factor(S_3$Var, levels = Vare)
S_3$Pob <- factor(S_3$Pob, levels = Pob)

(GrafST_3MBIC<-ggplot(S_3,aes(x=Var,y=sc,color=Pob)) +geom_point(position =
position_dodge(width = 0.6),size=1) +theme_classic() +
geom_errorbar(aes(ymin = sc+sd, ymax = (sc-sd)), width = 0.5,position =
position_dodge(width = 0.6)) +
labs(title="Intervalos al 95% confianza de coeficientes, ST_3MBIC
(4 int)", x="", y="Coeff", color="")+
theme(plot.title = element_text(size=12),axis.text.x=element_text(angle=90,hjust=1)) +
geom_hline(yintercept=0,color="gray") + theme(legend.position = "top")
```

#Tabla de predicción del modelo ST_3MBIC aplicado en población y muestras, con ENT recategorizada

```
table(Base$INGTRMEN_RESP,predict(pi.ST_3MBIC.2)>0)
table(Base.20$INGTRMEN_RESP,predict(SI.ST_3MBIC.2)>0)
table(Base_3.ST$INGTRMEN_RESP,predict(STSIC3.ST_3MBIC.2)>0)
table(Base_2.ST$INGTRMEN_RESP,predict(STSIC2.ST_3MBIC.2)>0)
```

Bibliografía

[1] Agresti, A. (2013) *Categorical data analysis*, tercera edición, New Jersey: John Wiley & Sons Inc.

[2] Agresti, A. (2015) *Foundations of linear and generalized linear models*, primer edición, Hoboken, New Jersey: John Wiley & Sons Inc.

[3] Särndal, C., Swensson, B. y Wretman, J. (1992) *Model Assited Survey Sampling*, primera edición, Nueva York: Springer.

[4] INEGI. (2015). Encuesta Intercensal 2015. *Síntesis metodológica y conceptual.* http://internet.contenidos.inegi.org.mx/contenidos/productos//prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825078836.pdf

[5] Fitzmaurice G.M., Heath A.F. y Cox, D.R. (1997) *Detecting Overdispersion in Large Scale Surveys: Application to a Study of Education and Social Class in Britain*, Appl. Statist. 4, 415-432.

[6] Hosmer, D.W., Lemeshow, S. y Sturdivant, R.X. (2013) *Applied Logistic Regression*, tercera edición, John Wiley & Sons Inc.

[7] Cox, D.R. (2015) *Big data and precision*, Biometrika 102, 712–716.

[8] Fleiss, J.L., Levin, B. y Paik, C.M. (2003) *Statistical Methods for Rates and Proportions*, tercera edición, Hoboken, New Jersey: John Wiley & Sons Inc.

[9] Cox, D.R., Kartsonaki, C. y Keogh, R. (2018) *Big data: Some Statistical issues*, Statistics and Probability Letters 136, 111–115.

[10] Heeringa, S.G., West, B.T. y Berglund, P.A. (2017) *Applied Survey Data Analysis*, segunda edición, SW, Florida: Taylor & Francis Group.

[11] Efron, B., (2020) *Prediction, Estimation, and Attribution*, Journal of the American Statistical Association, 115:530, 636-655.

[12] Hastie, T. [University of Bristol] (2018, noviembre 14) *Statistical learning with big data. A talk by Trevor Hastie* [Video]. Youtube. www.youtube.com/watch?v=0EWJZIC4JxA&ab_channel=UniversityofBristol