UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
ACATLAN

EXPLICACION DE LA PARTICIPACION POLITICA EN MEXICO MEDIANTE REGRESION LOGISTICA A PARTIR DE UNA ENCUESTA PREELECTORAL REALIZADA EN EL AÑO 2000

T E S I N A

QUE PARA OBTENER EL TITULO DE

LICENCIADO EN ACTUARIA

P R E S E N T A

JORGE ARMANDO BARRERA CEBALLOS

ASESOR ACT. MAHIL HERRERA MALDONADO

DICIEMBRE 2004


U.N.A.M. FES ACATLÁN
DIC 13 2004
DEPTO. DE TÍTULOS
PROFESIONALES Y CERTIFICACIÓN
UNIDAD DE ADMINISTRACIÓN ESCOLAR

AGRADECIMIENTOS

A mis padres. Porque este trabajo representa la recompensa a todo el esfuerzo realizado para brindarme una educación de calidad. Por todos los principios de responsabilidad y esfuerzo que me enseñaron siempre. Por esto y por todo lo que me han brindado, les agradezco de todo corazón. Los quiero más que a nadie en el mundo.

A mi hermano. Por apoyarme y estar conmigo siempre, en las buenas y en las malas. Eres una gran persona. También te quiero más que a nadie en el mundo.

A mis amigos. Gracias porque cuento con ustedes siempre. Igualmente saben que pueden contar conmigo siempre. Confíen en que a su lado siempre habrá una persona que les extenderá la mano incondicionalmente.

A Mahil Herrera. Porque sin tus conocimientos y tu asesoría estadística esta tesis no sería sin duda la misma. Además por el respaldo que me brindaste en la realización de esta tesina.

A Iván Castro y a Alicia de la Macorra. Porque a ustedes agradezco mi crecimiento profesional y sobretodo les agradezco el aprendizaje adquirido en lo que a la estadística se refiere. Su apoyo y guía han sido cruciales para mí. Además les agradezco ser personas en las que puedo confiar totalmente.

A Selene González. Por tus innumerables asesorías en el ámbito político del fenómeno del turnout y por la experiencia de vida.

A BIMSA. Muchas gracias por haber hecho de mi vida profesional algo tan interesante y divertido. En BIMSA surgió la idea de hacer esta tesina y creo que nunca hubiera tenido esta forma de no haber estado nunca ahí.

A Patricia de Zúñiga. Por la revisión de estilo y redacción tan profunda que me brindaste.

INDICE

Síntesis

La participación electoral en México en el año 2000 se explica por: el nivel de educación, el tipo de localidad (Urbano, Rural/Mixto), el nivel de identificación con algún partido político y la edad, conclusión derivada de la aplicación de la regresión logística. Existen tres líneas de investigación para abordar el tema de la participación electoral: el enfoque racional, sociológico, y el psicológico, las cuales fueron puestas a prueba resultando la combinación de las dos últimas la más adecuada para explicar porqué la gente votó en la elección del 2000. Los datos aquí utilizados corresponden a la encuesta preelectoral (Febrero) de la elección presidencial del año 2000 realizada por el grupo de investigación del periódico Reforma.

Introducción

El esfuerzo de los académicos en darle una explicación consistente al fenómeno que dota de sentido a la democracia en la actualidad, es un ejercicio que ha generado un debate, que como todos, no está acabado... El fenómeno de la participación electoral ha sido abordado desde diferentes perspectivas y enfoques de análisis. Por un lado, tenemos el argumento sociológico, por otro lado el psicológico y finalmente el racional (Harrop and Miller, 1987). La literatura más reciente del comportamiento electoral en Estados Unidos hace referencia a estos enfoques y son tomados como la base teórica de los estudios que hoy día se realizan no sólo en Estados Unidos, sino en diversos países, entre ellos México. Las tres líneas de investigación son amplias y a su vez complejas por lo que los hallazgos y las contribuciones metodológicas y teóricas han influido en el estudio de los procesos sociales y políticos de otros países.

Con la ayuda de la teoría estadística es posible modelar las explicaciones psicológica, sociológica y racional a la participación electoral, que la Ciencia Política ha aportado. En este trabajo se pretende probar el poder explicativo de cada una de las teorías planteadas y sus posibles combinaciones utilizando el método estadístico de regresión logística. Esta manera de explicar la participación electoral resulta adecuada dado que la teoría estadística señala que para la aplicación de la regresión logística es necesario que la variable dependiente sea dicotómica, es decir, tenga solamente dos categorías. En este sentido, el votante solamente tiene dos opciones en una elección: acudir a emitir su voto o no acudir. El trabajo no desarollará los modelos de cada escuela por separado debido a que por sí mismos no fueron poderosos en términos explicativos. En cambio, las combinaciones de los modelos sí arrojaron resultados importantes. Se mostrarán las combinaciones de los modelos que logran una mejor aproximación al objetivo del trabajo: la combinación de los modelos racional-sociológico y psicológico-sociológico.

La elección presidencial del año 2000 es importante debido a que es uno de los momentos políticos conyunturales más importantes de nuestro país en las últimas décadas. Por esta razón, en este trabajo se analizarán datos provenientes de una encuesta preelectoral realizada a nivel nacional en febrero del 2000, cinco meses antes de la elección, por el periódico Reforma.

Una vez aplicada la regresión logística a los modelos planteados se encuentra que la combinación de los argumentos psicológicos y sociológicos resulta la más adecuada para la explicación de la participación en México en las elecciones presidenciales del año 2000. El argumento racional combinado con el argumento sociológico también es modelado, sin embargo, en este caso los resultados no presentan suficiente evidencia estadística para explicar adecuadamente la participación electoral.

1. Explicaciones al fenómeno del voto

Para la escuela de enfoque sociológico, el elector es un producto social y su decisión de voto responde a los clivages[1] sociales de tal manera que los cambios en los sistemas políticos se explican en gran medida por los cambios en las estructuras políticas y económicas. Así, el acto de ir a votar se explica por variables estructurales o variables determinadas por el contexto social como son el ingreso o la educación, la edad, el contexto político, entre otras cosas.

En contraste con este punto de vista en el que el voto se explica por el contexto social, las escuelas psicológica y de elección racional basan su explicación de la participación electoral en las decisiones individuales. Para el enfoque psicológico, el voto se explica por las identificaciones de los electores con los partidos políticos principalmente, de tal forma que la identificación partidista moldea las decisiones políticas de un individuo. Las investigaciones que esta escuela ha desarrollado sugieren que las preferencias partidistas se transmiten de generación en generación desde las etapas tempranas de formación. Esto tiene sentido al tomar en cuenta que las personas van definiendo su manera de ver el mundo con base en las enseñanzas de los padres. Una implicación importante de este modo de entender el comportamiento humano es que conforme pasa el tiempo los individuos intensifican las posturas que han aprendido. En suma, para esta escuela las decisiones y evaluaciones políticas son una expresión de la psicología humana.

Ahora, con base en este último enfoque, ¿cómo se explica el acto político más fundamental? La respuesta tiene que ver con el concepto de compromiso político, el cual nos muestra que la participación política depende de los niveles de political engagement (compromiso político) de un individuo. Así, es más probable que una persona más interesada en política participe en las urnas que una que no está interesada en absoluto. La intuición detrás de esto es que un individuo identificado con alguna de las fuerzas políticas implicadas en los procesos electorales, siente el compromiso de apoyar las ideas que su partido de preferencia sustenta, por lo cual es muy posible que acuda a votar. No así alguien que no siente ningún tipo de compromiso político con alguna fuerza electoral.

[1] Conflictos principalmente ideológicos de una sociedad. Para comprender el concepto de clivage considérese el ejemplo de clivage más típico que es el de derecha-izquierda, en el cual las clases bajas apoyan a la izquierda y las medias-altas a la derecha.

Hay quienes argumentan que los procesos de cambio en las sociedades no se dan de una elección a otra sino que éstos toman tiempo. Dicho de otro modo, los pobres no se vuelven ricos de una elección a otra ni viceversa como para asegurar que este cambio en la estructura social es causa de las alteraciones en las tasas de participación. Con base en esta observación, ¿cómo podrían los partidarios de la escuela sociológica explicar que las tasas de participación varían tanto de una elección a otra? Parece que algo falta para comprender exhaustivamente la participación electoral desde este enfoque. Con respecto a la escuela psicológica, nos encontramos ante otro dilema que ha sido planteado en el debate, a saber, que un alto número de personas sin identificación partidista participa políticamente en las elecciones, lo cual hace pensar que no es esta la variable fundamental para darle una explicación consistente al fenómeno que estamos estudiando. Por estas y otras razones, resulta necesario acudir a otro tipo de argumentos.

Con base en la escuela racional, el acto de ir a votar es el cálculo que hacen los electores de los beneficios y costos de esta acción. John Aldrich ha argumentado extensamente que una elección racional es aquella en la que se toman en cuenta las utilidades asociadas con ciertos resultados. Así, si la utilidad de ir a votar es mayor a los costos, entonces, si el individuo toma sus decisiones racionalmente, acudirá a las urnas y no lo hará en caso contrario.

Riker y Ordeshook, partidarios de la escuela racional, plantean una ecuación retomada por Aldrich para calcular la utilidad de votar para un individuo (Riker y Ordeshook, 1968):

$$R= P\ B(a-b) - C +D$$

En donde R es la utilidad de participar en una elección, P es la probabilidad de que el voto sea decisivo en la elección, B(a-b) el diferencial en beneficio entre el candidato a y b, C el costo de votar y D el civismo o elemento expresivo. De acuerdo con esta ecuación un individuo votará siempre y cuando R sea positivo.

Supongamos una elección en la que compiten dos candidatos. Es inmediato suponer también que hay tres posibles resultados, a saber, que gane el candidato a, que pierda o que empate con el candidato b. El individuo racional deberá hacer el cálculo de la utilidad que cada resultado le reporta y entonces decidir votar o no votar. Si por alguna razón, el elector apoya al candidato a, el resultado que le generará más utilidad será que a sea el ganador; posteriormente preferirá que empaten y finalmente que a pierda. Siguiendo el argumento racional, el individuo hará el cálculo de su utilidad y si ésta es mayor a cero, entonces votará; si es negativo no lo hará (Aldrich, 1993).

No obstante de la lógica planteada en el ejemplo anterior, hay quienes han argüido que es problemático tratar de explicar la participación a través de un argumento racional pues la fórmula contiene elementos inconsistentes (Barry, 1970). Si observamos el elemento P en la ecuación podemos ver que dado que ésta tiende a cero, sobre todo en electorados muy grandes, el individuo tendría muy pocos incentivos de participar pues es muy improbable que su voto influya de manera decisiva en la elección. Bajo la lógica racional, el individuo no debería participar políticamente pues sabe que un resultado no se altera con un solo voto. Sin embargo, los electores acuden a votar... De esta manera nos encontramos con la llamada "paradoja del voto".

A este respecto, autores como Downs, Riker y Ordeshook han planteado que no es que la participación electoral sea un acto irracional sino que más bien, el individuo recibe beneficios al sentir que de alguna manera está contribuyendo al proceso democrático y que está cumpliendo con sus deberes ciudadanos. A partir de este argumento, se incluyó el término D en la ecuación dándole validez al cálculo de la utilidad de votar. Por este elemento es posible pensar que el saldo será positivo a pesar de que la probabilidad de incidir en la elección sea muy pequeña.

Otro de los debates en torno al enfoque racional tiene que ver con el argumento que afirma que el costo de participar es muy alto en tanto que para hacer una decisión política, como es la participación, se necesita información, reflexión y tiempo para obtener el registro como ciudadano con derecho al voto. Nuevamente esto resulta problemático pues si esto fuera cierto, tendríamos además de una P muy chica, una C muy grande por lo que sería menos probable obtener una utilidad positiva.

Además se ha observado que los partidos políticos disminuyen los costos de información a través de las campañas políticas. El individuo ya no tiene que ir a buscar la información suficiente para tomar una decisión ya que cada vez es más frecuente que los partidos envíen correspondencia explicando sus plataformas o hagan llamadas telefónicas, entre otros medios de movilización del voto. Asimismo, las instituciones electorales se han esforzado por agilizar los trámites para obtener el registro en el padrón electoral, con lo cual también se disminuyen los costos de asistir a las votaciones. Para Aldrich el argumento racional sigue teniendo sentido al tomar en cuenta la sensibilidad del individuo al valor de la C.

El elemento de la ecuación B(a-b) se refiere al beneficio que obtiene un elector de que gane el candidato preferido (a); (a-b) es la diferencia que existe entre el candidato a y b en cuanto a propuestas, ideología o características personales. De aquí que mientras mayor sea la distancia entre un candidato y otro, mayor será el beneficio para el elector. Esto porque se asume que si la distancia entre un candidato y otro es muy grande, sería peor para el individuo que ganara el candidato contrario ya que este se encontraría muy lejos de sus opiniones políticas. Por tanto, si resultara electo el preferido, le reportaría más beneficios.

Con respecto a los inconvenientes teóricos del beneficio del diferencial B(a-b), el teorema de Hotelling (Gravelle and Reef, 1981) indica que las plataformas de un candidato y otro son muy similares por lo que (a-b) sería muy cercano a cero y de nueva cuenta, esto obligaría a que la utilidad de votar se redujera. No obstante, podemos argumentar que la elección del año 2000 fue particularmente distinta a las demás pues había muchos elementos para diferenciar a los candidatos. La posibilidad del cambio que Vicente Fox representaba y argumentaba en campaña lo hacía muy distinto a Labastida y a Cárdenas. Esto hace pensar que el inconveniente del diferencial puede no influir en el modelo de elección racional en este caso particular.

Como se ha visto, las diferentes escuelas que abordan nuestro tema tienen problemas teóricos que son debatibles en muchos sentidos. Sin embargo, han aportando conceptos que sí han esclarecido, si bien no de manera definitiva, sí de manera parcial el fenómeno de la participación electoral. Por este motivo, los académicos toman variables de todas las escuelas para realizar sus modelos de participación.

2. Teoría de la Regresión Logística

2.1. Introducción al Modelo de Regresión Logística

Los métodos de regresión se han convertido en un componente integral de cualquier análisis de datos que trata de explicar la relación entre una variable dependiente y una o más variables explicativas, resultando común el caso en el que la variable dependiente es discreta, teniendo dos o más categorías. Así, el modelo de regresión logística ha representado, en muchas disciplinas, el método estándar de análisis en esta situación.

Resulta importante señalar que la finalidad del análisis de regresión logística es la misma de cualquier otro método de modelaje utilizado en estadística : encontrar el mejor y más sencillo modelo para describir la relación entre una variable dependiente y un conjunto de variables explicativas. Dichas variables explicativas son comúnmente llamadas covariables. El ejemplo más común de un modelo de regresión es el de regresión lineal en el cual se asume que la variable dependiente es continua. Este es un aspecto que distingue a un modelo de regresión logística de un modelo de regresión lineal ya que la variable dependiente, en el primer caso, es dicotómica (dos categorías) o politómica (más de dos categorías).

La regresión logística requiere de algunos de los supuestos de la regresión lineal como son:

1. El modelo sea especificado de manera correcta, es decir, que ninguna variable explicativa importante sea omitida y que las variables explicativas sean medidas sin error.

2. Las observaciones sean independientes.

3. Ninguna de las variables explicativas sea función lineal del resto. Multicolinealidad perfecta (perfecto grado de correlación lineal entre las variables explicativas) impide la estimación de los coeficientes del modelo; fuerte multicolinealidad ocasiona estimaciones imprecisas.

Casos influyentes, también ocasionan problemas en la estimación de los parámetros al modelo de regresión logística al igual que al modelo de regresión lineal.

Si dichas condiciones se cumplen, los estimadores de máxima verosimilitud de los coeficientes del modelo deben, teóricamente, tener las propiedades deseables de insesgamiento, eficiencia, y normalidad (en muestras lo suficientemente grandes). Existen ciertas reglas prácticas para determinar si la muestra es suficiente (por ejemplo, para el caso del modelo de regresión logística si n- k excede 100, siendo k el número de parámetros). El tamaño de muestra representa una condición necesaria pero no suficiente ya que las propiedades estadísticas de los estimadores bajo este modelo dependen también del número de casos dentro de una combinación de valores fijos de X y Y. Distribuciones sesgadas de Y son particularmente problemáticas. Por ejemplo, una muestra de 200 casos, pero solamente con 5 casos con Y = 1, provee poca información acerca de los efectos parciales de las variables explicativas.

Otra diferencia importante entre regresión lineal y logística consiste en la naturaleza de la relación entre la variable dependiente y las variables explicativas. En cualquier problema de regresión una cantidad clave es el valor medio de la variable dependiente dado el valor de las distintas variables independientes incluidas en el modelo. Esta cantidad es conocida como la **media condicional** y esta expresada mediante:

$$E[Y \mid (x_1, x_2, ..., x_k)]$$

donde:

Y = variable dependiente

$(x_1, x_2, ..., x_k)$ = valor de cada una de las variables independientes

k = número de variables independientes incluidas en el modelo

Por simplicidad, $(x_1, x_2, ..., x_k)$ será representado por **X**. Dicha cantidad se lee como "el valor esperado de Y, dado el valor de **X**".

En regresión lineal se asume que este valor medio puede ser expresado como una ecuación lineal en **X** (o alguna tranformación de **X** o de *Y*), tal como :

$$E[Y \mid (x_1, x_2, ..., x_k)] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

Esta expresión implica que es posible para $E[Y \mid (x_1, x_2, ..., x_k)]$ tomar cualquier valor según como X tome valores comprendidos entre $-\infty$ a $+\infty$.

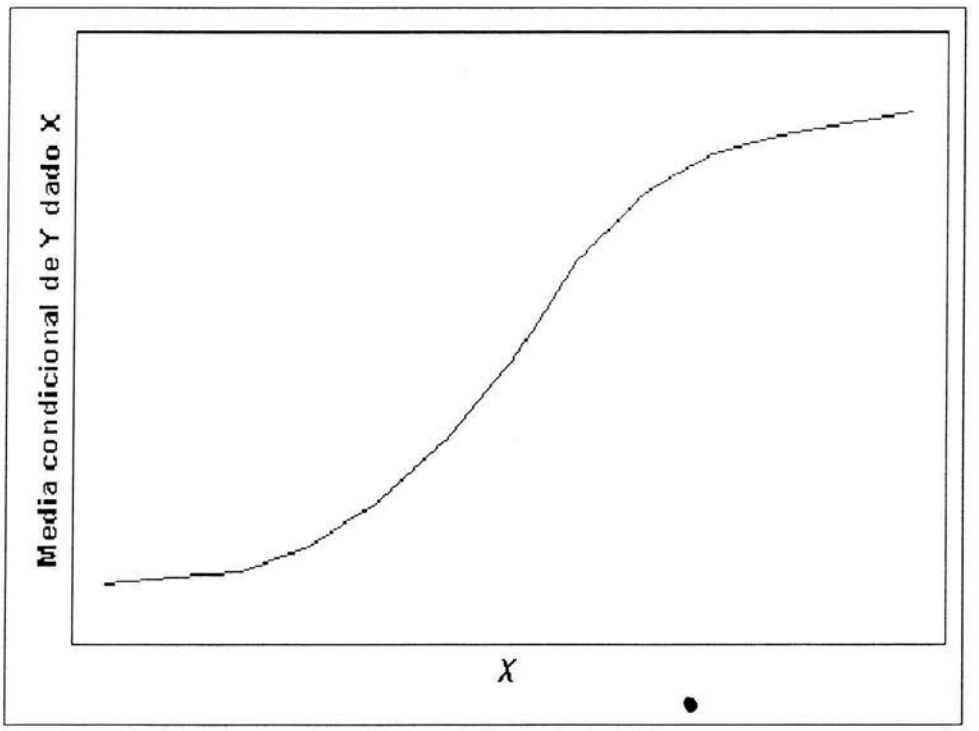
Con el fin de simplificar la notación, se utilizará la cantidad

$$\pi(X) = E[Y \mid (x_1, x_2, ..., x_k)]$$

Para representar la media condicional de Y dado **X** cuando la distribución logística es utilizada. La forma específica del modelo de regresión logística que se utilizará es la siguiente:

$$\pi(X) = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_k X_k}}$$

el cual tiene la siguiente forma:



Gráfica 2.1. $\pi(X)$ vs X

Como puede observarse a partir de la gráfica de $\pi(\mathbf{X})$, cuando la variable dependiente es dicotómica la media condicional de Y debe ser mayor o igual que cero y menor o igual que uno [*i.e.*, $0 \leq E(Y \mid \mathbf{X}) \leq 1$]. Además, la gráfica muestra que la media condicional de Y se aproxima a cero y a uno de manera gradual.

El cambio en $\pi(\mathbf{X})$ por unidad de cambio en $\mathbf{X}$ se convierte progresivamente más pequeño conforme la media condicional se acerca a cero o a uno. Se dice que la curva tiene la forma de una letra S.

La siguiente transfomación de $\pi(\mathbf{X})$ es conocida como el logit, esta trasformación esta definida en términos de $\pi(\mathbf{X})$:

$$g(\mathbf{X}) = \ln\left[\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right]$$

$$= \ln\left[\frac{\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}}{1-\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}}\right]$$

$$= \ln\left[\frac{\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}}{\frac{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}-e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}}\right]$$

$$= \ln\left[\frac{\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}}{\frac{1}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}}\right]$$

$$= \ln\left[\frac{\left(e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}\right)*\left(1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}\right)}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}\right]$$

$$= \ln[(e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k})]$$

$$= \beta_0 + \beta_1X_1 + \ldots + \beta_kX_k$$

La importancia del logit consiste en que tiene algunas propiedades del modelo de regresión lineal. Por ejemplo, $g(\mathbf{X})$ es lineal en sus parámetros, puede ser continua y puede variar de $-\infty$ a $+\infty$. Esta última propiedad de $g(\mathbf{X})$ asegura que $\pi(\mathbf{X})$ se encuentre entre 0 y 1 como ya se comentó anteriormente.

Una diferencia muy importante entre el modelo de regresión lineal y el modelo de regresión logística radica en la distribución condicional de la variable dependiente. En el modelo de regresión lineal se asume que una observación de la variable dependiente puede ser expresada mediante:

$$y = E(Y \mid \mathbf{X}) + \varepsilon$$

La cantidad ε es conocida como error de predicción, el cual representa la diferencia entre el verdadero valor de Y_j en la población (un valor posiblemente diferente al valor observado en la muestra) y el valor estimado de Y_j .

En regresión lineal, el residual es comúnmente denotado por e siendo $e_j = Y_j - \widehat{Y}_j$, siendo este la diferencia entre el valor estimado y el valor observado de Y para la $j-ésima$ observación.

Así, en regresión lineal, se desprende que la distribución condicional de la variable dependiente dado X será normal con media $E(Y \mid \mathbf{X})$ y varianza constante. Sin embargo, este no es el caso cuando la variable dependiente es dicotómica. En esta situación se puede expresar el valor de la variable dependiente dado $\mathbf{X}$ como:

$$y = \pi(\mathbf{X}) + \varepsilon$$

Aquí, ε puede tomar dos distintos valores. Si $y = 1$ entonces $\varepsilon = 1 - \pi(\mathbf{X})$ con probabilidad $\pi(\mathbf{X})$ y si $y = 0$ entonces $\varepsilon = -\pi(\mathbf{X})$ con probabilidad $1 - \pi(\mathbf{X})$. Por lo tanto ε sigue una distribución con media:

$$\begin{aligned} E(\varepsilon) &= (1 - \pi(\mathbf{X})) * \pi(\mathbf{X}) + (-\pi(\mathbf{X})) * (1 - \pi(\mathbf{X})) \\ &= \pi(\mathbf{X}) - (\pi(\mathbf{X}))^2 - \pi(\mathbf{X}) + (\pi(\mathbf{X}))^2 = 0 \end{aligned}$$

y varianza:

$$Var(\varepsilon) = (1 - \pi(\mathbf{X}))^2 * \pi(\mathbf{X}) + (-\pi(\mathbf{X}))^2 * (1 - \pi(\mathbf{X}))$$

$$= \pi(\mathbf{X}) - 2(\pi(\mathbf{X}))^2 + (\pi(\mathbf{X}))^3 + (\pi(\mathbf{X}))^2 - (\pi(\mathbf{X}))^3$$

$$= \pi(\mathbf{X}) - (\pi(\mathbf{X}))^2$$

$$= \pi(\mathbf{X})[1 - \pi(\mathbf{X})]$$

De esta forma, la distribución condicional de la variable dependiente dado $\mathbf{X}$ sigue una distribución binomial con probabilidad dada por la media condicional $\pi(\mathbf{X})$.

En resumen, cuando la variable dependiente es dicotómica se tiene:

1) La media condicional de la ecuación de regresión debe ser formulada para estar acotada entre 0 y 1. El modelo logístico $\pi(\mathbf{X})$ utlizado en este caso cumple con este requisito.

2) La distribución binomial describe la distribución de los errores.

2.2. El Modelo de Regresión Logística Múltiple

Considérese un conjunto de p variables independientes representadas por el vector:

$$\mathbf{X}' = (X_1, X_2, ..., X_p)$$

Sea la probabilidad condicional de que ocurra el evento representada de la siguiente manera:

$$P(Y = 1 \mid \mathbf{X}) = \pi(\mathbf{X})$$

Entonces el logit del modelo de regresión logística esta dado por la ecuación:

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

en cuyo caso:

$$\pi(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_k X_k}} \quad ... \ (1)$$

Si algunas de las variables son nominales u ordinales, sería inapropiado incluirlas en el modelo como si fueran continuas. Esto dado que los números usados para representar los distintos niveles de dichas variables son meros identificadores y no tienen ningún significado numérico. Bajo esta situación el método apropiado consiste en utilizar un conjunto de variables de diseño (variables dicotómicas).

Así, de manera general si una variable nominal u ordinal tiene k posibles valores, entonces $k-1$ variables de diseño será necesario inlcuir en el modelo. La notación es la siguiente: supóngase que la $j-ésima$ variable independiente, X_j tiene k_j posibles valores. Las $k_j - 1$ variables de diseño serán representadas como D_{ju} y los coeficientes para estas variables de diseño serán representados mediante β_{ju}, $u = 1, 2, ..., k_j - 1$. Por lo tanto el logit para un modelo con p variables indpendientes siendo la $j-ésima$ variable nominal u ordinal será:

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + ... + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + \beta_p X_p$$

En general, cuando se desarrollen los modelos de regresión logística expuestos en este trabajo, se eliminarán los índices en el símbolo de suma y los dobles subíndices para indicar que se están utilizando variables de diseño.

2.3. Ajuste del modelo de Regresión Logística

Supóngase que se tiene una muestra de n observaciones independientes del par $(\mathbf{x}_i, y_i)$, $i = 1,2,...,n$. El ajuste del modelo requiere la obtención de estimadores del vector $\beta' = (\beta_0, \beta_1, ..., \beta_p)$. El método de estimación utilizado será el de máxima verosimilitud, teniendo como función de verosimilitud la siguiente:

$$l(\beta) = \prod_{i=1}^{n} \zeta(\mathbf{x}_i) \quad ... \quad (2)$$

Para aquellos pares $(\mathbf{x}_i, y_i)$ donde $y_i = 1$, la contribución a la función de verosimilitud es $\pi(\mathbf{x}_i)$ y para aquellos pares para los cuales $y_i = 0$ la contribución a la función de verosimilitud es $1 - \pi(\mathbf{x}_i)$. Una manera conveniente de expresar la contribución a la función de verosimilitud para el par $(\mathbf{x}_i, y_i)$ es a través del término:

$$\zeta(\mathbf{x}_i) = \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i}$$

El principio de máxima verosimilitud establece que se deberá utilizar como estimador de β el valor que maximice la expresión de la ecuación (2). Sin embargo, resulta más sencillo el trabajar con el logaritmo de la ecuación (2). Esta expresión, está definida por:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]\} ... \quad (3)$$

y es conocida como el logaritmo de la verosimilitud (Log Likelihood).

De esta manera existirán $p + 1$ ecuaciones de verosimilitud, las cuales son obtenidas por diferenciación del logaritmo de la función de verosimilitud respecto a los $p + 1$ coeficientes. Las funciones de verosimlitud resultantes pueden ser expresadas mediante:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0$$

para el caso de la diferenciación de la ecuación (3) respecto a β_0

y

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0$$

para el caso de la diferenciación de la ecuación (3) respecto a cada uno de

los j parámetros β_j.

donde $j = 1,2,...,p$.

La solución a las ecuaciones anteriores se representa mediante $\hat{\boldsymbol{\beta}}$. De esta forma, los valores estimados del modelo de regresión logística son $\hat{\pi}(\mathbf{x}_i)$, el valor calculado de la expresión (1) utilizando $\hat{\boldsymbol{\beta}}$ y $\mathbf{x}_i$.

2.4. Estimación de errores estándar

El método de estimación de las varianzas y covarianzas de los coeficientes estimados proviene de la teoría estimación basada en máxima verosimilitud. Esta teoría establece que los estimadores son obtenidos mediante la matriz de derivadas parciales de segundo orden del logaritmo de la función de verosimilitud. Las derivadas parciales tienen la siguiente forma general:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)] \quad \ldots (4)$$

y

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu} \pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)] \quad \ldots (5)$$

para $j, u = 0, 1, 2, \ldots, p$.

Sea $I(\beta)$ la matriz que contiene el negativo de los términos dados en las ecuaciones (4) y (5). Dicha matriz es conocida como la matriz de información. Las varianzas y las covarianzas de los coeficientes estimados son obtenidos a partir de la inversa de la matriz de información, la cual será representada mediante la expresión:

$$\sum(\beta) = I^{-1}(\beta)$$

Se utilizará $\sigma^2(\beta_j)$ para representar el $j-$ *ésimo* elemento de la diagonal de esta matriz, el cual es la varianza de β_j, y $\sigma(\beta_j, \beta_u)$ para representar algún elemento fuera de la diagonal de la matriz, el cual es la covarianza entre β_j y β_u. Los estimadores de las varianzas y covarianzas serán representados por $\hat{\sum}\left(\hat{\beta}\right)$ y obtenidos mediante la evaluación de $\sum(\beta)$ en $\hat{\beta}$. Se utilizará $\hat{\sigma}^2(\hat{\beta}_j)$ y $\hat{\sigma}(\hat{\beta}_j, \hat{\beta}_u)$, $j = 0, 1, 2, \ldots, p$ para representar los valores en esta matriz. La expresión para representar el estimador del error estándar del $j-$ *ésimo* coeficiente estimado será la siguiente:

$$\hat{SE}(\hat{\beta}_j) = \left[\hat{\sigma}^2(\hat{\beta}_j)\right]^{1/2}$$

para $j = 0, 1, 2, \ldots, p$.

La representación de la matriz de información será útil cuando se exponga la evaluación del ajuste del modelo, así $\hat{I}(\hat{\beta}) = \mathbf{X}'\mathbf{V}\mathbf{X}$ donde $\mathbf{X}$ es una matriz n por $p+1$ elementos que contiene la información para cada uno de los integrantes de la muestra. $\mathbf{V}$ es una matriz diagonal de n por n elementos con elemento general $\hat{\pi}_i\ (1-\hat{\pi}_i)$. De esta forma, la matriz $\mathbf{X}$ es:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ & & \dots & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$$

y la matriz $\mathbf{V}$ es:

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1\ (1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2\ (1-\hat{\pi}_2) & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \hat{\pi}_n\ (1-\hat{\pi}_n) \end{bmatrix}$$

2.5. Medidas de Bondad de Ajuste del Modelo de Regresión Logística

2.5.1 El estadístico Log-Likelihood

El estadístico Log-Likelihood provee una medida de las desviaciones entre los valores observados y los valores estimados por el modelo. Este estadístico generalmente se denota como -2 veces el log-likelihood ($-2LL$), y sigue aproximadamente una distribución χ^2. Una indicación del ajuste del modelo puede ser obtenida mediante la comparación del valor de $-2LL$ para el modelo que contiene las k variables explicativas con el valor $-2LL$ para el modelo nulo. Este estadístico se denotará como $-2LL_{diff}$ y está representado mediante la siguiente ecuación:

$$-2LL_{diff} = (-2LL_0) - (-2LL_1)$$

$-2LL_0$ representa la medida de las desviaciones para el modelo nulo con logit $g(\mathbf{X}) = \beta_0$

$-2LL_1$ representa la medida de las desviaciones para el modelo con logit $g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$

El cambio en $-2LL$ representa el efecto que tienen las variables explicativas en las desviaciones del modelo, un efecto que puede ser evaluado utilizando la distribución χ^2 con tantos grados de libertad como la diferencia en el número de términos entre los modelos bajo comparación.

Este estadístico $-2LL$ tiene una relación cercana con el estadístico de bondad de ajuste utilizado en regresión lineal. De hecho, la suma de residuales al cuadrado, la cual es una medida de las desviaciones en regresión lineal, puede ser vista como análoga a $-2LL$ la cual es una medida de las desviaciones utilizada en regresión logística. De manera similar, el estadístico F utilizado en regresión lineal puede ser considerado como análogo al estadístico χ^2 utilizado en regresión logística.

	Medida de desviaciones	Distribución de referencia
Regresión lineal	Suma de residuales al cuadrado	F
Regresión logística	Log-likelihood	χ^2

Tabla 2.1. Estadísticos de Bondad de Ajuste para regresión lineal y logística

2.5.2 El estadístico de Wald

El estadístico de Wald prueba la hipótesis de que el coficiente de regresión para la $k-esima$ variable explicativa es cero, es decir, que dicha variable explicativa no tiene efecto sobre la variable dependiente. Esto es:

$$H_0 : \beta_K = 0 \quad \text{vs} \quad H_1 : \beta_K <> 0$$

El coeficiente β_K representa la relación que existe entre la variable dependiente y la k –ésima variable explicativa en cuestión. Un valor igual a cero significa que no existe relación alguna, por lo cual para concluir que la variable explicativa es importante en cuanto a su aportación al modelo buscaremos rechazar H_0.

El estadístico de Wald se calcula mediante la siguiente ecuación:

$$\text{Estadístico de Wald} = \frac{\beta_K^2}{SE_{\beta_K}}$$

donde:

β_K = coeficiente de regresión para la k –ésima variable explicativa

SE_{β_K} = error estándar de β_K

El estadístico de Wald sigue una distribución aproximadamente $Normal(0,1)$ y su significancia se prueba utilizando el criterio basado en el valor-p[1]. Un estadístico de Wald significativo sugiere que la variable explicativa tiene efecto sobre la variable dependiente.

[1] El valor - p de una prueba de hipótesis es igual a la probabilidad observada de cometer un error tipo I (rechazar la hipótesis Ho cuando esta es cierta). Por lo tanto, el criterio basado en el valor-p se refiere a que deben buscarse valores-p pequeños (generalmente menores a 0.1) para rechazar Ho.

Falta página

N° 20

donde:

$n_{Y=0}$ =número de casos observados con $Y = 0$

$n_{Y=1}$ =número de casos observados con $Y = 1$

Por otro lado, el número de errores con el modelo se calcula mediante la siguiente expresión:

$$p_e = b + c$$

2.6. Diagnósticos de Regresión Logística

Para la explicación de los diagnósticos de regresión logística será necesario introducir el término "patrón de covariable", el cual es una combinación específica de los valores de las distintas variables independientes incluidas en el modelo de regresión logística estimado. Por ejemplo, supóngase que en un modelo se tienen dos variables independientes codificadas de la siguiente manera:

Género:

1. Mujer

2. Hombre

Nivel socioeconómico:

1. Bajo

2. Medio

3. Alto

Así, se tendría para este caso 6 patrones de covariables en total:

Género	Nivel Socioeconómico
Hombre	Bajo
Hombre	Medio
Hombre	Alto
Mujer	Bajo
Mujer	Medio
Mujer	Alto

Tabla 2.3. Ejemplo de patrones de covariables

En esta explicación se asumirá que el modelo ajustado contiene p variables independientes y que ellas forman J patrones de covariables indexados mediante $j = 1, 2, ..., J$.

Las cantidades clave para los diagnósticos de regresión logística son las sumas de residuales al cuadrado. Además, en regresión logística se tienen errores que siguen la distribución binomial, por lo cual, como resultado se tiene que la varianza del error es una función de la media condicional:

$$var(Y_j \mid \mathbf{x}_j) = m_j E(Y_j \mid \mathbf{x}_j) * [1 - E(Y_j \mid \mathbf{x}_j)] = m_j \pi(\mathbf{x}_j)[1 - \pi(\mathbf{x}_j)]$$

donde:

m_j = número de casos con patrón de covariable X_j.

En el caso de regresión logística se tendrán residuales definidos mediante la siguiente ecuación:

$$res(y_j, \hat{\pi}_j) = \frac{\left(y_j - m_j \hat{\pi}_j\right)}{\left(m_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right)\right)^{1/2}}$$

donde y_j = número de respuestas positivas, $y = 1$, entre los m_j casos con patrón de covariable j.

Dado que cada residual es dividido por una estimación de su error estándar, se esperaría que estás cantidades tuvieran media aproximadamente igual a cero y varianza aproximadamente igual a uno. Estos residuales son conocidos como residuales de Pearson y servirán más adelante para definir el estadístico χ^2 basado en dichos residuales.

El diagnóstico del modelo será realizado mediante las siguientes medidas:

1) Las medidas globales de distancia entre Y^T y $\hat{Y}^T$

2) La contribución de cada par (Y_i , $\hat{Y}_i$), $i = 1, ..., n$ a dichas medidas globales de distancia

Como una primera medida global de distancia entre Y^T y $\hat{Y}^T$ será utilizada la prueba de hipótesis de Hosmer y Lemeshow, la cual está basada en agrupaciones de las probabilidades estimadas a partir del modelo y tiene como hipótesis nula aquella que indica que el modelo tiene buen ajuste.

Para la prueba de Hosmer y Lemeshow inicialmente se debe construir una tabla de contingencia de valores observados que tenga como columnas las probabilidades estimadas para cada uno de los J patrones de covariables observados en la muestra y como renglones los dos posibles valores de la variable dependiente observada $(0, 1)$, así la tabla resultante será de tamaño $2X\ J$. El siguiente paso consiste en colapsar el número de columnas de la tabla de contingencia basándose en percentiles de las probabilidades estimadas, para el caso de este trabajo se formarán diez agrupaciones de probabilidades estimadas basadas en los deciles de dicha distribución. Se denotará el número de agrupaciones formadas con la letra g.

Para la estimación de los valores esperados de cada una de las celdas se lleva a cabo el siguiente procedimiento:

1) Para el renglón $Y = 1$, se suman las probabilidades estimadas para todos los individuos en el grupo correspondiente

2) Para el renglón $Y = 0$, el valor esperado estimado se obtiene sumando para todos los individuos en el grupo correspondiente, uno menos la probabilidad estimada

La expresión para el cálculo del estadístico χ^2 correspondiente a dicha prueba es la siguiente:

$$\chi^2 = \sum_{j=1}^{C} \frac{(o_j - e_j)^2}{e_j}$$

donde:

C = número de celdas = $2 * g$

o_j = valores observados para la j-ésima celda

e_j = valores esperados para la j-ésima celda

Dicho estadístico sigue una distribución χ^2 con $(g - 2)$ grados de libertad.

Otra medida global de distancia entre Y^T y $\hat{Y}^T$ es la prueba de hipótesis basada en los residuales de Pearson, la cual tiene como hipótesis nula que el modelo ajusta de manera adecuada. Para el cálculo de estos residuales se debe tomar en cuenta que en regresión logística los valores estimados para la variable dependiente son calculados para cada uno de los patrones de covariables presentes en la muestra y dependen de la probabilidad estimada para dicho patrón de covariable, así los valores estimados son los siguientes:

$$m_j \hat{\pi}_j = m_j \left(\frac{\exp\left(\hat{g}(X_j)\right)}{1+\exp\left(\hat{g}(X_j)\right)} \right)$$

donde:

X_j = patrón de covariable j

m_j = número de individuos con el patrón de covariable X_j

$\hat{\pi}_j$ = media condicional estimada de Y dado X_j

$\hat{g}(X_j)$ = representa el logit estimado

$j = 1,...,J$

El estadístico χ^2 de Pearson se define de la siguiente manera:

$$\chi^2 = \sum_{j=1}^{J} res(y_j, \hat{\pi}_j)^2$$

El cual sigue una distribución χ^2 con $J - (p + 1)$ grados de libertad.

Por otro lado, como medidas de diagnóstico se revisarán el nivelaje, el $\Delta \hat{\beta}_j$ y el $\Delta\chi_j^2$.

El nivelaje representa una medida de distancia de cada una de las observaciones en la muestra respecto al valor promedio de la probabilidad estimada $\hat{\pi}(\mathbf{X})$. Este concepto de distancia respecto a la media de dicha variable resulta importante, dado que es más probable que aquellos valores que se encuentran localizados lejos del valor promedio de la probabilidad estimada, tengan mayor influencia sobre los valores estimados de los parámetros. Entre más se aleja la probabilidad estimada de su valor medio (0.5) el nivelaje se incrementa hasta que dicha probabilidad estimada se vuelve menor que 0.1 ó mayor que 0.9, regiones en las cuales el nivelaje debe ser el más pequeño. La siguiente tabla muestra el comportamiento del nivelaje que se esperaría ocurriera si el modelo tiene un ajuste adecuado y si no existen observaciones con fuerte influencia sobre los valores estimados de los parámetros.

Probabilidad estimada	0-0.1	0.1-0.3	0.3-0.7	0.7-0.9	0.9-1
Nivelaje	Pequeño	Grande	Moderado a pequeño	Grande	Pequeño

Tabla 2.4. Comportamiento adecuado del nivelaje dentro de las cinco regiones definidas por el valor de la probabilidad estimada

En regresión lineal, además de los residuales para cada uno de los patrones de covariables, otras cantidades importantes para la formación e interpretación de los diagnósticos son la matriz de "sombrero"y los valores de nivelaje derivados a partir de ella. Sea $\mathbf{X}$ la matriz de tamaño J por $(p+1)$ que contiene los valores de todos los patrones de covariables formados a partir de los valores observados de las p variables independientes, con la primera columna igual a 1 para reflejar la presencia de una constante en el modelo. La matriz $\mathbf{X}$ es conocida comúnmente como la matriz de diseño. En regresión lineal la matriz de "sombrero"es $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Pregibon (1981) derivó una expresión para la matriz de "sombrero"en el caso de regresión logística. La matriz es:

$$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$$

donde $\mathbf{V}$ es una matriz diagonal de tamaño JXJ con elemento general $v_i = m_i \hat{\pi}(\mathbf{x}_i)\left[1 - \hat{\pi}(\mathbf{x}_i)\right]$.

Sea h_j el nivelaje para el $j-ésimo$ patrón de covariable y además sea h_j la cantidad que representa al $j-ésimo$ elemento en la diagonal de la matriz $\mathbf{H}$. La expresión para el cálculo de h_j es la siguiente:

$$h_j = m_j \hat{\pi}(\mathbf{x}_j)\left[1 - \hat{\pi}(\mathbf{x}_j)\right](1, \mathbf{x}_j')(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}(1, \mathbf{x}_j')' = v_j * b_j$$

donde:

$$b_j = (1, \mathbf{x}_j')(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}(1, \mathbf{x}_j')'$$

con

$$\sum_{j=1}^{J} h_j = (p+1)$$

el número de parámetros en el modelo.

Una medida útil de diagnóstico es la que mide el cambio en el valor de los coeficientes estimados debido a la eliminación de los elementos en la muestra con patrón de covariable j. Esta medida se obtiene mediante la siguiente expresión:

$$\Delta \hat{\beta}_j = \frac{r_j^2 h_j}{(1-h_j)^2}$$

Los rangos aceptables para dicho diagnóstico son los que se muestran a continuación:

Probabilidad estimada	0-0.1	0.1-0.3	0.3-0.7	0.7-0.9	0.9-1
$\Delta \hat{\beta}$	Pequeño	Grande	Moderado	Grande	Pequeño

Tabla 2.5. Comportamiento adecuado de $\Delta \hat{\beta}$ dentro de las cinco regiones definidas por el valor de la probabilidad estimada

Otra medida de diagnóstico es aquella que mide la reducción en el valor del estadístico χ^2 de Pearson debida a la eliminación de los elementos en la muestra con patrón de covariable j. La expresión para el cálculo de dicha medida es la siguiente:

$$\Delta\chi_j^2 = \frac{r_j^2}{(1-h_j)}$$

Los rangos aceptables para dicho diagnóstico son los que se muestran a continuación:

Prob. estimada	0-0.1	0.1-0.3	0.3-0.7	0.7-0.9	0.9-1
$\Delta\chi^2$	Grande o Pequeño	Moderado	Moderado a pequeño	Moderado	Grande o Pequeño

Tabla 2.6. Comportamiento adecuado de $\Delta\chi^2$ dentro de las cinco regiones definidas por el valor de la probabilidad estimada

2.7 Interpretación de coeficientes

Se iniciará la interpretación de coeficientes de regresión logística bajo la situación en la que la $j-ésima$ variable independiente es dicotómica. Para este fin, supóngase que x_j está codificada como 0 ó 1. Los posibles valores de $\pi(X)$ se muestran en la siguiente tabla:

	$x_j=1$	$x_j=0$
Y=1	$\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_j+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_j+\cdots+\beta_kX_k}}$	$\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}$
Y=0	$1-\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)=\frac{1}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_j+\cdots+\beta_kX_k}}$	$1-\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)=\frac{1}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}$
Total	1	1

Tabla 2.7. Valores del modelo de regresión logística cuando la $j-ésima$ variable independiente es dicotómica

Los momios resultantes de observar la característica de la variable dependiente entre los individuos con $x_j = 1$ están definidos por :

$$\frac{\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)}{1-\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)}$$

De manera similar, los momios resultantes de observar la característica de la variable dependiente entre los individuos con $x_j = 0$ están definidos por :

$$\frac{\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)}{1-\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)}$$

El logit queda definido de la siguiente manera:

$$g(X_1 = x_1,...,X_j = 1,...,X_k = x_k) = \ln\left[\frac{\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)}{1-\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)}\right]$$

y

$$g(X_1 = x_1,...,X_j = 0,...,X_k = x_k) = \ln\left[\frac{\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)}{1-\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)}\right]$$

De esta forma, la razón de momios, denotada por ψ, se define como el cociente de los momios para $x_j = 1$ respecto a los momios para $x_j = 0$, y esta dado por la ecuación:

$$\psi = \frac{\frac{\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)}{1-\pi(X_1=x_1,...,X_j=1,...,X_k=x_k)}}{\frac{\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)}{1-\pi(X_1=x_1,...,X_j=0,...,X_k=x_k)}}$$

El logaritmo de la razón de momios es:

$$ln(\psi) = \ln\left[\frac{\frac{\pi(X_1=x_1,\ldots,X_j=1,\ldots,X_k=x_k)}{1-\pi(X_1=x_1,\ldots,X_j=1,\ldots,X_k=x_k)}}{\frac{\pi(X_1=x_1,\ldots,X_j=0,\ldots,X_k=x_k)}{1-\pi(X_1=x_1,\ldots,X_j=0,\ldots,X_k=x_k)}}\right] =$$

$$g(X_1 = x_1,\ldots,X_j = 1,\ldots,X_k = x_k) - g(X_1 = x_1,\ldots,X_j = 0,\ldots,X_k = x_k)$$

Utilizando las expresiones de la tabla 2.7, la razón de momios es:

$$\psi = \frac{\left[\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_j+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_j+\ldots+\beta_kX_k}}\right]\left[\frac{1}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}\right]}{\left[\frac{1}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_j+\ldots+\beta_kX_k}}\right]\left[\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}\right]}$$

$$= \left[\frac{e^{\beta_0+\beta_1X_1+\ldots+\beta_j+\ldots+\beta_kX_k}}{e^{\beta_0+\beta_1X_1+\ldots+\beta_kX_k}}\right]$$

$$= e^{\beta_j}$$

Así, para el modelo de regresión logística, la razón de momios para la j – *ésima* variable independiente dicotómica es:

$$\psi = e^{\beta_j}$$

y

$$ln(\psi) = \ln(e^{\beta_j}) = \beta_j$$

La razón de momios es una medida de asociación que indica cuanto más probable o menos probable es encontrar la característica de la variable dependiente entre aquellos con $x_j = 1$ que entre aquellos con $x_j = 0$.

El estimador de ψ es $\hat{\psi} = e^{\hat{\beta}_j}$.

El intervalo al $100(1-\alpha)\%$ de confianza estimado de la razón de momios puede ser obtenido calculando primero los puntos extremos del intervalo de confianza para $\hat{\beta}_j$, después exponenciando estos valores se obtiene la siguiente expresión:

$$\exp(\hat{\beta}_j \pm Z_{1-\alpha/2} SE(\hat{\beta}_j))$$

donde $SE(\hat{\beta}_j)$ representa el error estándar de $\hat{\beta}_j$.

Para el caso en el que alguna de las variables independientes es nominal u ordinal con más de dos categorías se tiene otro tratamiento. Cabe recordar que en la sección 2.2 se revisó que el logit para un modelo con p variables independientes siendo la $j-$ *ésima* variable nominal u ordinal con k categorías es el siguiente:

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \sum_{u=1}^{k-1} \beta_{ju} D_{ju} + \beta_p X_p$$

teniendo a la $k-$ *ésima* categoría como la categoría de referencia.

La siguiente tabla muestra las variables de diseño correspondientes:

	D_{j1}	D_{j2}	...	D_{jk-1}
*Categoría*1	1	0	0	0
*Categoría*2	0	1	0	0
...				
Categoríak − 1	0	0	0	1
Categoríak	0	0	0	0

Tabla 2.8. Variables de diseño para el caso en el que la $j-$ *ésima* variable independiente es nominal u ordinal con k categorías

De esta forma se puede calcular la razón de momios para la categoría $k-1$ respecto a la $k-$ *ésima* categoría (la de referencia) de la siguiente manera:

$$g(X_1 = x_1, \ldots, D_{j1} = 0, D_{j2} = 0, \ldots, D_{jk-1} = 1, \ldots, X_p = x_p)$$

$$- g(X_1 = x_1, \ldots, D_{j1} = 0, D_{j2} = 0, \ldots, D_{jk-1} = 0, \ldots, X_k = x_k)$$

$$= (\beta_0 + \beta_1 X_1 + ... + \beta_{j1}(0) + \beta_{j2}(0) + ... + \beta_{jk-1}(1) + \beta_p X_p) -$$

$$(\beta_0 + \beta_1 X_1 + ... + \beta_{j1}(0) + \beta_{j2}(0) + ... + \beta_{jk-1}(0) + \beta_p X_p)$$

$$= \beta_{jk-1}$$

Dado que la diferencia en el logit es $ln(\psi)$, se tiene que $ln(\psi) = \beta_{jk-1}$, por lo tanto:

$$\psi = e^{\beta_{jk-1}}$$

El estimador de ψ es $\hat{\psi} = e^{\hat{\beta}_{jk-1}}$.

La razón de momios es una medida de asociación que indica cuanto más probable o menos probable es encontrar la característica de la variable dependiente entre aquellos con respuesta a alguna de las categorías comprendidas entre 1 y $k = 1$ que entre aquellos con respuesta a la k – *ésima* categoría de la variable independiente.

El intervalo al $100(1 - \alpha)\%$ de confianza estimado de la razón de momios puede ser obtenido calculando primero los puntos extremos del intervalo de confianza para $\hat{\beta}_{jk-1}$, después exponenciando estos valores se obtiene la siguiente expresión:

$$\exp(\hat{\beta}_{jk-1} \pm Z_{1-\alpha/2} \hat{SE}(\hat{\beta}_{jk-1}))$$

donde $\hat{SE}(\hat{\beta}_{jk-1})$ representa el error estándar de $\hat{\beta}_{jk-1}$.

En lo que respecta a la interpretación de la razón de momios de variables continuas incluidas en el modelo, se deben tomar en cuenta algunas consideraciones.

La primera consideración consiste en que ha sido probada la linealidad de dichas variables continuas. Además, supóngase que el logit es el siguiente:

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + ... + \beta_j X_j + ... + \beta_p X_p$$

siendo la j – *ésima* variable continua.

De esta ecuación se desprende que el coeficiente β_j corresponde al cambio en el logit para un incremento de "1"unidad en la variable conitnua X_j, esto es:

$$\beta_j = g(X_1 = x_1, ..., X_j = x_j + 1, ..., X_p = x_p) - g(X_1 = x_1, ..., X_j = x_j, ..., X_p = x_p)$$

para cualquier valor de la variable X_j.

Sin embargo, incrementos de una sola unidad en la variable X_j pueden resultar de poco interés. Para proveer una útil interpretación de dicha variable se presentará a continuación un método de estimación puntual y de intervalo para cambios arbitrarios de "c"unidades en variables continuas.

El cambio en el logit debido al cambio en "c_1"unidades en X_j se obtiene a partir de la diferencia:

$$c_1\beta_j = g(X_1 = x_1, ..., X_j = x_j + c_1, ..., X_p = x_p) - g(X_1 = x_1, ..., X_j = x_j, ..., X_p = x_p)$$

y la razón de momios asociada se obtiene exponenciando dicha diferencia:

$$\Psi = \exp(c_1\beta_j)$$

Un estimador puede ser obtenido reemplazando el coeficiente β_j por su respectivo estimador de máxima verosimilitud $\hat{\beta}_j$.

Con el fin de obtener la estimación por intervalo de confianza para la variable X_j es necesario obtener un estimador del error estándar de $\hat{\beta}_j$. Una vez que este ha sido calculado, se multiplica el error estándar de $\hat{\beta}_j$ por c_1, así los extremos del intervalo estimado al $100(1-\alpha)\%$ de confianza son los siguientes:

$$\exp\left[c_1 \hat{\beta}_j \pm Z_{1-\alpha/2} * c_1 * \hat{SE}(\hat{\beta}_j) \right]$$

3. Aplicación de la Regresión Logística en las Teorías del Voto

3.1 Aplicación de las Teorías del Votante Racional y Sociológica

De acuerdo con los fundamentos relacionados a la teoría del votante racional, se tiene la siguiente ecuación:

$$Ri = PiBi - Ci + Di...(1)$$

En donde:

Pi = Probabilidad de que el voto del votante i sea decisivo en la elección

Bi = Beneficio que obtendrá el votante i al votar por su candidato preferido

Ci = Costo para el votante i de emitir su voto

Di = Obligaciones cívicas o elemento expresivo del votante i

$i = 1,...,n$

n = número de entrevistados

En la ecuación (1) si $Ri > 0$, resulta razonable para el votante i emitir su voto, en caso de que $Ri <= 0$, para el votante i resulta un acto no razonable el acudir a votar por lo cual no votará el día de la elección.

Con el objeto de probar las teorías racional y sociológica, se relacionarán las respuestas individuales a la pregunta referente a la participación del entrevistado el día de la elección, con un conjunto de variables las cuales pueden ser agrupadas bajo los siguientes encabezados:

1. **Características sociodemográficas del entrevistado (Teoría Sociológica)**

2. **Variables de reducción de costo (Teoría del Votante Racional)**

3. Valor estratégico del acto de votar (Teoría del Votante Racional)

4. Obligaciones cívicas (Teoría del Votante Racional)

Teniéndose las siguientes relaciones entre dichos conjuntos de variables y los componentes de la ecuación (1):

Conjunto de variables	Componente de la ecuación de la Teoría del Votante Racional
1	–
2	C_i
3	P_i, B_i
4	D_i

Tabla 3.1. Conjuntos de variables

La pregunta sobre participación el día de la elección es la siguiente:

*P*12. En una escala del 0 al 10, donde 0 significa que usted definitivamente no va a votar en las elecciones para Presidente y 10 que definitivamente usted si va a votar.

¿ Por favor dígame qué tan probable es que usted vote en las elecciones presidenciales de este año ?

No va a votar — Sí va a votar

0	1	2	3	4	5	6	7	8	9	10

Recodificada de la siguiente manera:

No va a votar — Sí va a votar

0 a 9 = 0	10 = 1

El modelo estadístico a través del cual se pretenden probar las hipótesis establecidas en la ecuación (1) consiste en un modelo de regresión logística teniendo como variable dependiente la correspondiente a la intención de acudir a votar el día de la elección:

$Yi = Ri = 1$ si el entrevistado está seguro de que acudirá a votar el día de la elección

$Yi = Ri = 0$ si el entrevistado no está seguro de que acudirá a votar el día de la elección

Las variables explicativas son las siguientes:

1. Características sociodemográficas del entrevistado

1.1 Nivel de educación (*Ei*)

1 = No estudió

2 = Primaria

3 = Secundaria

4 = Preparatoria

5 = Universidad o más

1.2 Nivel socioeconómico del hogar del entrevistado, aproximado por el número de focos en el hogar (variable continua, *FOCOSi*)

El histograma de la variable FOCOS se observa de la siguiente manera:

Gráfica 3.1. Histograma de focos en el hogar

Dicho histograma se observa sesgado hacia un número de focos menor a 30, con algunos casos extremos con un número mayor a 50 focos en el hogar.

Además si se cruza esta última variable con el nivel socioeconómico AMAI[1] se observa que efectivamente la proporción de focos en un hogar determinado crece conforme es más alto el nivel socioeconómico del hogar, por lo cual la variable número de focos es una buena representación continua del nivel socioeconómico al cual pertenece el entrevistado.

Nivel Socioeconómico de la Vivienda	Número de Focos Promedio
A/B	13
C+	11
C	8
D+	7
DE	4

Tabla 3.2. Nivel Socioeconómico vs Número de Focos Promedio

[1] Nivel socioeconómico del hogar determinado por la Asociación Mexicana de Agencias de Investigación. Revisar Anexo A.

1.3 Tipo de Localidad en la cual se encuentra el hogar del entrevistado (*LOCi*)

0 = Urbana

1 = Rural / Mixta

1.4 Edad del entrevistado (Mayores de 18 años, *EDADi*)

2. Variables de reducción de costo

2.1 Información obtenida a partir de algún partido político (representación de *Ci*)

0 = no ha tenido ningún tipo de acercamiento directo con algún partido político

1 = ha tenido algún tipo de contacto con algún partido político ya sea

a través de asistir a un mitin político o bien por haber sido contactado

por algún partido político mediante un representante o vía propagandas

o cartas

3. Valor estratégico del acto de votar

3.1 Percepción de lo "cerrado" que va a estar la elección (representación de *Pi*)

0 = está seguro de que su candidato favorito va a ganar la elección

1 = no está seguro de que su candidato favorito va a ganar la elección

De esta forma queda representado el término *Pi*, dado que si el entrevistado no está seguro de que su candidato favorito gane la elección, entonces será más probable que acuda a votar para apoyarlo con su voto.

3.2 Decisión definitiva sobre cual candidato va a votar el entrevistado (representación de *Bi*)

0 = no está seguro que va a votar por el candidato mencionado como favorito

1 = está seguro que va a votar por el candidato mencionado como favorito

Así queda representado el término *Bi*, dado que si el entrevistado está seguro de que va a votar por el candidato mencionado como favorito siginfica que le representa mayor utilidad el votar por él.

Según la teoría del votante racional, la interacción de dichas variables (*PiBi*) es la variable de interés para representar adecuadamente el valor estratégico del acto de votar, quedando esta categorizada de la siguiente manera:

1 = está seguro que va a votar por el candidato mencionado como favorito, sin embargo, no está seguro de que su candidato favorito va a ganar la elección La categoría 0 se divide en tres casos:

Caso 1:

Está seguro que va a votar por el candidato mencionado como favorito y está seguro de que su candidato favorito va a ganar la elección

Caso 2:

Está seguro de que el candidato mencionado como favorito va a ganar la elección pero su intención de voto no es definitiva

Caso 3:

No está seguro de que el candidato mencionado como favorito va a ganar la elección y su intención de voto no es definitiva

4. Obligaciones cívicas

4.1 Percepción de México como un país demócrata (representación de *Di*)

0 = opina que México no es una democracia

1 = opina que México es una democracia

En este caso se tiene la variable discreta Nivel de educación (*Ei*), la cual sería inapropiado incluir en el modelo como si fuera una variable medida en una escala continua dado que los diferentes niveles de dicha variable son meros identificadores y no tienen significado numérico. En una situación como esta el método adecuado consiste en incluir en el modelo una colección de variables de diseño (variables dicotómicas) que representen las diferentes categorías de la variable en cuestión.

La variable Nivel de educación tiene cinco posibles valores, por lo cual será necesario incluir en el modelo cuatro variables de diseño, lo anterior se debe a que una de las categorías de la variable original es considerada como categoría de referencia. De esta forma la variable *EDUi* quedará representada por las siguientes cuatro variables de diseño, tomando como categoría de referencia No estudió. Cabe aclarar que SPSS, el software utilizado a lo largo de este trabajo, permite la especificación de este tipo de restricciones, es decir, al utilizar SPSS no es necesario crear las cuatro variables de diseño ya que lo hace de manera automática una vez que se ha especificado la variable en cuestión como categórica y la categoría que servirá como categoría de referencia, en este caso No estudió.

Nivel de Educación	E_4	E_5	E_6	E_7
No estudió	0	0	0	0
Primaria	1	0	0	0
Secundaria	0	1	0	0
Preparatoria	0	0	1	0
Universidad o más	0	0	0	1

Tabla 3.3. Variables de diseño para la variable Educación

Así, el logit del modelo de regresión logística múltiple que se pretende ajustar es el siguiente:

$$g_1(\mathbf{X}) = \beta_0 + \beta_1 PB + \beta_2 D + \beta_3 C + \sum \beta jEj + \beta_8 FOCOS + \beta_9 LOC + +\beta_{10} EDAD. (2)$$

En donde:

$j = 4, 5, 6, 7$

$\mathbf{X} = (PB, D, C, E_4, E_5, E_6, E_7, FOCOS, LOC, EDAD)$

En cuyo caso:

$$\pi(\mathbf{X}) = \frac{e^{g_1(\mathbf{X})}}{1+e^{g_1(\mathbf{X})}} = E(Y \mid \mathbf{X})$$

En la siguiente tabla se puede observar la prueba basada en el estadístico Log-Likelihood (revisar capítulo 2):

$-2LL_0$	806.226
$-2LL_1$	737.271
$-2LL_{diff}$	68.955
d.f.	10
sig.	0.000

Tabla 3.4. Aplicación del estadístico Log-Likelihood al modelo con Logit g_1

Esta tabla sugiere que de manera conjunta, las variables independientes son importantes para explicar la variable dependiente Participación.

Los coeficientes estimados son los siguientes:

								95% CI	for β
Variable	B	S.E.	Wald	d.f.	Sig.	R	Exp(β)	Lower	Upper
PB	0.166	0.181	0.841	1	0.359	0	1.181	0.828	1.685
D	0.253	0.188	1.819	1	0.178	0	1.288	0.892	1.861
C	-0.043	0.258	0.028	1	0.868	0	0.958	0.578	1.589
*E			18.232	4	0.001	0.113			
*E4	1.395	0.443	9.922	1	0.002	0.099	4.033	1.694	9.605
*E5	1.074	0.457	5.515	1	0.019	0.066	2.927	1.194	7.174
*E6	1.366	0.494	7.663	1	0.006	0.084	3.92	1.49	10.311
*E7	1.923	0.506	14.461	1	0	0.124	6.842	2.539	18.433
FOCOS	0.029	0.016	3.26	1	0.071	0.04	1.029	0.998	1.062
*LOC	0.91	0.244	13.883	1	0	0.121	2.483	1.539	4.007
*EDAD	0.046	0.008	31.807	1	0	0.192	1.047	1.03	1.064
CONSTANT	-2.586	0.65	15.818	1	0				

*Variables estadisticamente significativas al 95% de confianza

Tabla 3.5. Coeficientes estimados para el modelo Racional-Sociológico

Como podemos observar en tabla de coeficientes estimados del modelo racional-sociológico, los coeficientes correspondientes a los términos PB, D y C de la ecuación (2) no son significativamente distintos de cero, por lo cual concluimos que dichas variables no son importantes a la hora de explicar participación. Sin embargo, también se obtiene como resultado que todas las variables sociodemográficas, excepto *FOCOS*, si están relacionadas de manera importante con nuestra variable dependiente. Por lo anterior, se estimarán los coeficientes del siguiente modelo:

$$g_2(\mathbf{X}) = \beta_0 + \sum \beta j Ej + \beta_8 LOC + \beta_9 EDAD ...(3)$$

En donde:

$j = 4,5,6,7$

$\mathbf{X} = (E_4, E_5, E_6, E_7, LOC, EDAD)$

En cuyo caso:

$$\pi(\mathbf{X}) = \frac{e^{g_2(\mathbf{X})}}{1+e^{g_2(\mathbf{X})}} = E(Y \mid \mathbf{X})$$

Los resultados de la prueba basada en el estadístico Log-Likelihood para este caso son los siguientes:

$-2LL_0$	806.226
$-2LL_1$	743.059
$-2LL_{diff}$	63.167
d.f.	6
sig.	0.000

Tabla 3.6. Aplicación del estadístico Log-Likelihood al modelo con Logit g2

La tabla anterior sugiere que el efecto que tienen las variables explicativas sobre la reducción de las desviaciones del modelo es significativo al 95% de confianza.

La tabla de coeficientes estimados es la siguiente:

								95% CI	for β
Variable	B	S.E.	Wald	d.f.	Sig.	R	Exp(β)	Lower	Upper
*E			19.217	4	0.0007	0.118			
*E4	1.312	0.443	8.768	1	0.0031	0.091	3.713	1.558	8.848
*E5	0.980	0.457	4.594	1	0.0321	0.056	2.664	1.087	6.529
*E6	1.286	0.492	6.828	1	0.0090	0.077	3.618	1.379	9.495
*E7	1.899	0.501	14.369	1	0.0002	0.123	6.679	2.502	17.831
*LOC	0.854	0.238	12.849	1	0.0003	0.116	2.348	1.472	3.744
*EDAD	0.044	0.008	29.950	1	0.0000	0.186	1.044	1.028	1.061
CONSTANT	-2.023	0.597	11.458	1	0.0007				

*Variables estadisticamente significativas al 95% de confianza

Tabla 3.7. Coeficientes estimados para el modelo Sociológico

De esta forma se puede concluir que después de aplicar el modelo Racional-Sociológico, las únicas variables estadísticamente significativas para explicar la participación son las del modelo Sociológico.

Resulta muy importante mencionar que a partir de este resultado no se pretende de ninguna manera concluir que la Teoría del Votante Racional no funcione para el caso de encuestas preelectorales realizadas en México, dado que este hecho puede deberse más bien a que las variables seleccionadas para la representación de los términos correspondientes a la ecuación ...(1) no miden de una manera totalmente adecuada los conceptos implícitos en cada uno de los términos en los cuales esta sustentada la Teoría del Votante Racional. Sin embargo, para la selección de dichas variables se revisaron los cuestionarios correspondientes a distintas encuestas preelectorales realizadas en el año 2000 por empresas encuestadoras, resultando el cuestionario de Reforma el que contenía las variables más cercanas a la representación de los conceptos de la ecuación...(1). Por esta razón se recomienda a las empresas encuestadoras dedicadas a la realización de estudios preelectorales, la inclusión de variables que realmente representen de una manera adecuada los conceptos de la ecuación...(1), tomando la experiencia de estudios realizados en Estados Unidos sobre este tema (Riker y Ordeshook 1968) con el objeto de contar con elementos más sólidos de prueba de dicha teoría.

3.2 Aplicación de las Teorías Sociológica y Psicológica

Debido a que se ha demostrado que las variables relacionadas con la Teoría Racional no son las más adecuadas para explicar la participación en la elección que estamos estudiando, es pertinente investigar si las variables que aluden a las teorías Psicológica y Sociológica explican en mayor medida el objeto fundamental de este estudio.

En primer lugar, recordemos que en la parte teórica de este trabajo se expuso que la variable más importante para la escuela psicológica es la identificación partidista, por esta razón se introdujo en el siguiente modelo dicha variable. Como es de esperarse, aquellas personas más fuertemente identificadas con alguna de las fuerzas políticas tendrán una mayor probabilidad de participar en la elección que aquellas menos identificadas políticamente. El tratamiento que se le dio a esta variable es el siguiente:

Identificación partidista (IDi)

0 = algo o poco identificados con algún partido político o bien independientes

1 = muy identificados con algún partido político

En segundo lugar, también, de la experiencia con el modelo Racional-Sociológico obtuvimos como conclusión que las variables sociodemográficas, comúnmente utilizadas por la escuela Sociológica, son aquellas que resultan importantes para explicar la participación. De aquí que se decidió incluirlas también en este nuevo modelo.

Cabe mencionar que el manejo de la variable dependiente representante de participación es igual que en el modelo anterior.

Así, el logit de este modelo queda de la siguiente manera:

$$g_3(\mathbf{X}) = \beta_0 + \sum \beta jEj + \beta_9 LOC + \beta_{10} EDAD + \beta_{11} ID ...(4)$$

En donde:

$j = 4,5,6,7$

$\mathbf{X} = (E_4, E_5, E_6, E_7, LOC, EDAD, ID)$

En cuyo caso:

$$\pi(\mathbf{X}) = \frac{e^{g_3(\mathbf{X})}}{1+e^{g_3(\mathbf{X})}} = E(Y \mid \mathbf{X})$$

Los resultados después de aplicar el estadístico Log-Likelihood son los siguientes:

$-2LL_0$	806.226
$-2LL_1$	722.672
$-2LL_{diff}$	83.554
d.f.	7
sig.	0.000

Tabla 3.8. Aplicación del estadistico Log-Likelihood al modelo con Logit g3

Así, se puede concluir que el efecto que tienen las variables explicativas sobre la reducción de las desviaciones del modelo es significativo al 95% de confianza. Los coeficientes estimados bajo este modelo son los siguientes:

								95% CI	for β
Variable	B	S.E.	Wald	d.f.	Sig.	R	Exp(β)	Lower	Upper
*ID	0.981	0.228	18.450	1	0.000	0.143	2.666	1.704	4.171
*E			19.740	4	0.001	0.121			
*E4	1.469	0.452	10.578	1	0.001	0.103	4.343	1.792	10.521
*E5	1.279	0.470	7.409	1	0.007	0.082	3.592	1.430	9.021
*E6	1.530	0.506	9.157	1	0.003	0.094	4.618	1.714	12.441
*E7	2.127	0.514	17.137	1	0.000	0.137	8.391	3.065	22.972
*LOC	0.727	0.243	8.948	1	0.003	0.093	2.068	1.285	3.330
*EDAD	0.044	0.008	29.518	1	0.000	0.185	1.045	1.029	1.062
CONSTANT	-2.486	0.620	16.071	1	0.000				

*Variables estadisticamente significativas al 95% de confianza

Tabla 3.9. Coeficientes estimados para el modelo Psicológico- Sociológico

Por otro lado, si solamente manejamos un modelo con las variables utilizadas por la escuela sociológica, es decir, eliminando de este último análisis la variable de Identificación partidista y dejando solamente las variables sociodemográficas, se tiene el mismo logit manejado en la ecuación (3). Los coeficientes estimados para este modelo son los mostrados en la Tabla 3.7.

Como resultado de comparar ambos modelos se observa que tanto el modelo que combina las escuelas sociológica y piscológica así como el modelo que sólo utiliza la escuela sociológica funcionan muy bien para el propósito fundamental de este trabajo. Con el fin de determinar si es importante el efecto que la variable *ID* aporta al modelo de la escuela sociológica, se calculará el estadístico $-2LL_{diff}$, el cual se obtiene de la siguiente manera:

$$-2LL_{diff} = (-2LL_0) - (-2LL_1)$$

En este caso, $-2LL_0$ corresponde al modelo con logit representado por g_2 (escuela sociológica) y $-2LL_1$ corresponde al modelo cuyo logit es g_3 (combinación de escuela sociológica y psicológica).

De esta forma, se tienen los siguientes resultados:

$-2LL_0$	743.059
$-2LL_1$	722.672
$-2LL_{diff}$	20.387
d.f.	1
sig.	6.32581E-06

Tabla 3.10. Aplicación del estadístico Log-Likelihood

A partir del cuadro anterior, dado que el valor-p es menor a 0.05 se puede concluir al 95% de confianza que la variable Identificación Partidista tiene efecto en la reducción de las desviaciones obtenidas por el modelo de la escuela sociológica.

Una vez que han sido identificadas las variables de las escuelas sociológica y psicológica como significativas, se procederá a investigar si la variable continua *EDAD* debe ser incluida en el modelo de manera lineal, como se ha hecho hasta este momento. Para tal efecto, se aplicará la transformación de Guerrero y Jonhson (Hosmer and Lemeshow, 1989) . Esta transformación añade el término $xln(x)$ al modelo para cada una de las variables continuas. Si el coeficiente para esta variable es significativo, se tiene evidencia de no linealidad en el logit.

Después de añadir al modelo el término $EDADT = EDAD * LN(EDAD)$ se obtiene el siguiente cuadro resumen:

								95% CI	for β
Variable	B	S.E.	Wald	d.f.	Sig.	R	Exp(β)	Lower	Upper
EDADT	-0.023	0.036	0.424	1	0.514	.000	0.976	0.908	1.049

Tabla 3.11. Resultados de la transformación de Guerrero y Jonhson

Dado que el coeficiente de dicho término no resulta estadísticamente significativo, se puede suponer linealidad de la variable *EDAD*.

Como se mencionó en el capítulo 2, uno de los supuestos bajo el cual las estimaciones de máxima verosimilitud de los coeficientes del modelo deben, teóricamente, tener las propiedades deseables de insesgamiento, eficiencia, y normalidad es la no presencia de casos graves de multicolinealidad. La presencia de fuerte multicolinealidad tiende a producir coeficientes estimados de las variables explicativas que parecen grandes de manera poco razonable: como regla práctica si algunos de los coeficientes estimados para dichas variables son mayores que 2 deberán ser examinados para determinar si existe fuerte multicolinealidad (Menard, 2000). Para el modelo que combina las escuelas sociológica y psicológica se observa que los coeficientes estimados de las variables *ID*, *LOC* y *EDAD* son menores que 1, en el caso de las variables de diseño correspondientes a la variable nivel de educación se observan coeficientes estimados mayores que 1 (en el caso de la categoría Universidad se presenta un coeficiente mayor que 2), este hecho se debe a que de entrada las variables de diseño están relacionadas entre sí, dado que la pertenencia a una de las categorías indica la no presencia en el resto de las categorías. Sin embargo, se puede concluir que no existen problemas de multicolinealidad dado que el resto de los coeficientes son menores que 1.

3.3 Diagnóstico del modelo

A continuación se realizarán las pruebas correspondientes a la bondad de ajuste, es decir, se aplicarán un conjunto de pruebas de hipótesis para saber que tan efectivo es el modelo ajustado en describir la variable dependiente participación de voto.

Para tal fin, sean $Y^T = (Y_1,...,Y_n)$ los valores muestrales observados para la variable dependiente Participación de voto y sean $\hat{Y}^T = (\hat{Y}_1,..., \hat{Y}_n)$ los valores estimados a partir del modelo para la misma variable, en donde $n = 678$ entrevistados. Así, concluiremos que el modelo ajusta de una manera adecuada si se cumplen las siguientes dos condiciones:

1) Las medidas globales de distancia entre Y^T y $\hat{Y}^T$ son estadísticamente pequeñas

2) La contribución de cada par $(Y_i, \hat{Y}_i)$, $i = 1,...,n$ a dichas medidas globales de distancia ocurre de manera no sistemática, es decir, la magnitud de dichas medidas no depende del comportamiento de algún patrón de covariable en particular

De esta forma una completa descripción del ajuste del modelo involucra tanto el cálculo de medidas globales de distancia entre Y^T y $\hat{Y}^T$, así como una inspección de los componentes individuales de dichas medidas globales.

Como una primera medida global de distancia entre Y^T y $\hat{Y}^T$ será utilizada la prueba de Hosmer y Lemeshow. La tabla de contingencia de valores observados queda de la siguiente manera:

	0.19-0.48	0.49-0.56	0.57-0.63	0.64-0.67	0.68-0.72	0.73-0.77	0.78-0.82	0.83-0.87	0.88-0.91	0.92-0.98
0	38	31	23	29	19	16	14	12	7	2
1	30	37	46	39	48	52	54	57	61	63
Total	68	68	69	68	67	68	68	69	68	65

Tabla 3.12. Frecuencias observadas para los deciles de las probabilidades estimadas

La tabla de valores esperados estimados resultante es la siguiente:

	0.19-0.48	0.49-0.56	0.57-0.63	0.64-0.67	0.68-0.72	0.73-0.77	0.78-0.82	0.83-0.87	0.88-0.91	0.92-0.98
0	37.20	32.40	26.94	23.37	19.86	16.87	13.35	10.27	7.07	3.67
1	30.80	35.60	42.06	44.63	47.14	51.13	54.65	58.73	60.93	61.33
Total	68	68	69	68	67	68	68	69	68	65

Tabla 3.13. Frecuencias esperadas estimadas para los deciles de las probabilidades estimadas

Los valores resultantes al aplicar dicha prueba son los siguientes:

χ^2	Grados de libertad	valor-p
4.47	8	0.812

Tabla 3.14. Resultados de la prueba de Hosmer y Lemeshow

La tabla anterior sugiere que no se puede rechazar la hipótesis nula, es decir, no se puede rechazar que el modelo ajuste de manera adecuada.

Otra medida global de distancia entre Y^T y $\hat{Y}^T$ es la prueba de hipótesis basada en los residuales de Pearson (Hosmer and Lemeshow, 1989). Aplicando este estadístico al modelo aquí estimado se tiene:

χ^2	Grados de libertad	valor-p
367.06	375	0.61

Tabla 3.15. Resultados de la prueba χ^2 basada en los residuales de Pearson

De esta forma no se puede rechazar que el modelo tiene buen ajuste.

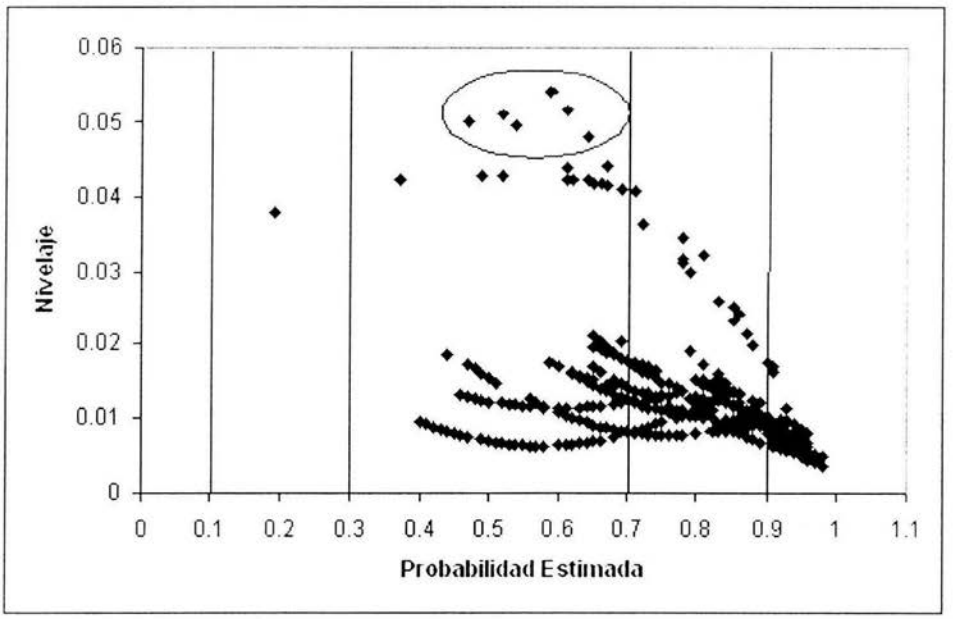
Las pruebas de hipótesis de Hosmer y Lemeshow y la prueba basada en el estadístico χ^2 cuadrada de Pearson proveen un número único que resume el parecido en general entre los valores observados y los valores estimados de la variable dependiente. La ventaja (que al mismo tiempo se convierte en una desventaja) de este tipo de medidas globales de distancia consiste en que se utiliza un sólo número para resumir una gran cantidad de información. Por lo tanto, el siguiente paso en la verificación del ajuste del modelo consiste en explorar la contribución de cada par (Y_i , $\hat{Y}_i$), $i = 1,...,n$, en las medidas globales de distancia. Las medidas de diagnóstico utilizadas para este fin serán las siguientes:

1) Nivelaje

2) Cambio en la magnitud de los coeficientes estimados debido a la eliminación de algunos patrones de covariables (exclusión de personas de la muestra)

3) Cambio en la magnitud del estadístico χ^2 cuadrada de Pearson debido a la eliminación de algunos patrones de convariables (exclusión de personas de la muestra)

La gráfica del Nivelaje contra la probailidad estimada para cada uno de los J patrones de variables presentes en la muestra es la siguiente:



Gráfica 3.2. Nivelaje vs probabilidad estimada

Como puede observarse en la gráfica, en el rango de probabilidad estimada de 0.3 a 0.7 existen algunos patrones de covariables que presentan nivelaje mayor que el resto de las observaciones. Por lo cual se procederá a investigar si el modelo es sensible a la eliminación de dichos patrones tanto de manera individual como de manera conjunta.

La manera de investigar si el modelo es sensible a la eliminación de algunos patrones de covariables, consiste en medir el cambio en la magnitud de los coeficientes de regresión estimados después ser de eliminados dichos patrones de covariables. Para esto se trabajará con algunos de los puntos que parecen presentar nivelaje fuera de rango aceptable, los cuales son los encerrados en el círculo de la gráfica 2. La información correspondiente a los patrones de covariables que serán estudiados es la siguiente:

FOLIO	E	LOC	EDAD	ID	NIVELAJE	$\pi(\mathbf{X})$
1361	1	1	26	1	0.054	0.59
1351	1	1	28	1	0.052	0.61
2358	1	0	36	1	0.051	0.52
2390	1	1	37	0	0.050	0.47
2218	1	0	38	1	0.049	0.54
1705	1	1	31	1	0.048	0.64

Tabla 3.16. Información de los patrones de covarible con comportamiento anormal de nivelaje

En los siguientes cuadros se presenta la estimación de los coeficientes así como la variación relativa en dichos coeficientes estimados después de la eliminación del patrón de covariable correspondiente.

Variable	Inicial	sig.	Sin 1361	var.**	sig.	Sin 1351	var.**	sig.	Sin 2358	var.**	sig.
ID	0.98	0.000	0.97	1.0%	0.000	0.97	0.9%	0.000	0.97	1.4%	0.000
E		0.001			0.000			0.000			
E_4	1.47	0.001	1.54	4.8%	0.001	1.53	4.5%	0.001	1.55	5.8%	0.000
E_5	1.28	0.007	1.35	5.9%	0.005	1.35	5.5%	0.005	1.37	7.0%	0.001
E_6	1.53	0.003	1.61	5.2%	0.002	1.60	4.8%	0.002	1.62	6.1%	0.004
E_7	2.13	0.000	2.20	3.6%	0.000	2.20	3.3%	0.000	2.22	4.4%	0.002
LOC	0.73	0.003	0.72	1.3%	0.003	0.72	1.3%	0.003	0.74	2.3%	0.000
EDAD	0.04	0.000	0.05	1.8%	0.000	0.04	1.6%	0.000	0.04	1.6%	0.002
Constant	-2.49	0.000	-2.58	3.8%	0.000	-2.57	3.5%	0.000	-2.60	4.5%	0.000
Var. Promedio*				3.4%			3.2%			4.1%	

*Variación Promedio **Variación Relativa

Tabla 3.17. Variaciones en las estimaciones de la tabla 3.9 debidas a la eliminación de patrones de covariables con comportamiento anormal de nivelaje

Variable	Inicial	sig.	Sin 2390	var.**	sig.	Sin 2218	var.**	sig.	Sin 1705	var.**	sig.
ID	0.98	0.000	1.00	1.7%	0.000	1.00	1.7%	0.000	0.97	0.9%	0.000
E		0.001			0.000			0.001			0.000
E_4	1.47	0.001	1.56	6.2%	0.001	1.37	6.6%	0.003	1.53	4.0%	0.001
E_5	1.28	0.007	1.38	7.6%	0.004	1.18	8.0%	0.015	1.34	4.9%	0.005
E_6	1.53	0.003	1.63	6.5%	0.002	1.42	6.9%	0.006	1.60	4.3%	0.002
E_7	2.13	0.000	2.22	4.5%	0.000	2.02	5.0%	0.000	2.19	2.9%	0.000
LOC	0.73	0.003	0.71	2.4%	0.004	0.71	2.6%	0.004	0.72	1.2%	0.003
EDAD	0.04	0.000	0.04	1.4%	0.000	0.04	1.4%	0.000	0.04	1.1%	0.000
Constant	-2.49	0.000	-2.60	4.6%	0.000	-2.36	5.0%	0.000	-2.56	3.0%	0.000
Var. Promedio*				4.4%			4.6%			2.8%	

*Variación Promedio **Variación Relativa

Tabla 3.18. Continuación de la tabla 3.17

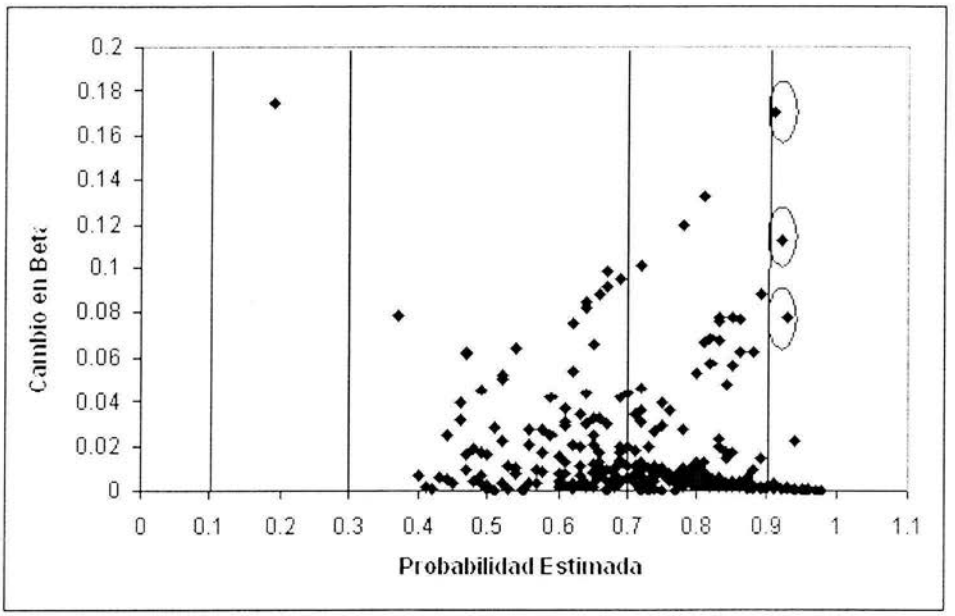
Variable	Inicial	sig.	Sin patrones anteriores	var.**	sig.
ID	0.98	0.000	0.96	1.7%	0.000
E		0.001			0.000
E_4	1.47	0.001	1.82	24.1%	0.000
E_5	1.28	0.007	1.66	29.5%	0.002
E_6	1.53	0.003	1.93	25.8%	0.001
E_7	2.13	0.000	2.51	17.8%	0.000
LOC	0.73	0.003	0.68	6.6%	0.005
EDAD	0.04	0.000	0.05	7.5%	0.000
Constant	-2.49	0.000	-2.95	18.8%	0.000
Var. Promedio*				16.5%	

*Variación Promedio **Variación Relativa

Tabla 3.19. Continuación de la tabla 3.18

Como puede observarse en los cuadros anteriores, las variaciones más grandes en la magnitud de los coeficientes respecto a las estimaciones del modelo original se observan cuando son eliminados de manera conjunta los seis patrones de covariables estudiados, sin embargo, los coeficientes continúan siendo significativamente distintos de cero con un 95% de confianza. Por lo anterior se puede concluir que la estimación de los coeficientes no es sensible a la eliminación de los patrones de covariables que presentan nivelaje más alto.

Otra medida de diagnóstico consiste en el cambio directo en la magnitud de los coeficientes debido a la exclusión de ciertos patrones de covariables (observaciones de la muestra). La gráfica correspondiente es la siguiente:



Gráfica 3.3. $\Delta \hat{\beta}$ vs probabilidad estimada

En la gráfica anterior se puede observar que son tres los valores más atípicos y se encuentran en el rango que va de 0.9 a 1, dado que se esperaría tener valores pequeños en este rango. Por lo cual, se estudiará la influencia de dichos patrones de covariables en la estimación de los parámetros del modelo. El siguiente cuadro presenta la información correspondiente a dichos patrones de covariables:

FOLIO	E	LOC	EDAD	ID	$\Delta \hat{\beta}$	$\pi(\mathbf{X})$
407	1	1	71	1	0.170	0.91
1372	5	0	62	0	0.112	0.92
2192	2	0	60	1	0.078	0.93

Tabla 3.20. Información de los patrones de covarible con comportamiento anormal de $\Delta \hat{\beta}$

En los siguientes cuadros se presenta la estimación de los coeficientes así como la variación relativa en dichos coeficientes estimados después de la eliminación del patrón de covariable correspondiente.

Variable	Inicial	sig.	Sin 407	var.**	sig.	Sin 1372	var.**	sig.	Sin 2192	var.**	sig.
ID	0.98	0.000	1.01	2.5%	0.000	0.97	0.6%	0.000	0.97	1.4%	0.000
E		0.001			0.001			0.000			0.000
E_4	1.47	0.001	1.35	7.8%	0.003	1.50	1.9%	0.001	1.55	5.8%	0.001
E_5	1.28	0.007	1.18	7.8%	0.014	1.32	3.4%	0.005	1.37	7.0%	0.004
E_6	1.53	0.003	1.43	6.2%	0.005	1.58	3.5%	0.002	1.62	6.1%	0.002
E_7	2.13	0.000	2.03	4.5%	0.000	2.23	4.7%	0.000	2.22	4.4%	0.000
LOC	0.73	0.003	0.76	4.6%	0.002	0.73	0.8%	0.003	0.74	2.3%	0.002
EDAD	0.04	0.000	0.05	2.9%	0.000	0.05	4.5%	0.000	0.04	1.6%	0.000
Constant	-2.49	0.000	-2.43	2.1%	0.000	-2.59	4.2%	0.000	-2.60	4.5%	0.000
Var. Promedio*				4.8%			3.0%			4.1%	

*Variación Promedio **Variación Relativa

Tabla 3.21. Variaciones en las estimaciones de la tabla 3.9 debidas a la eliminación de patrones de covariables con comportamiento anormal de $\Delta \hat{\beta}$

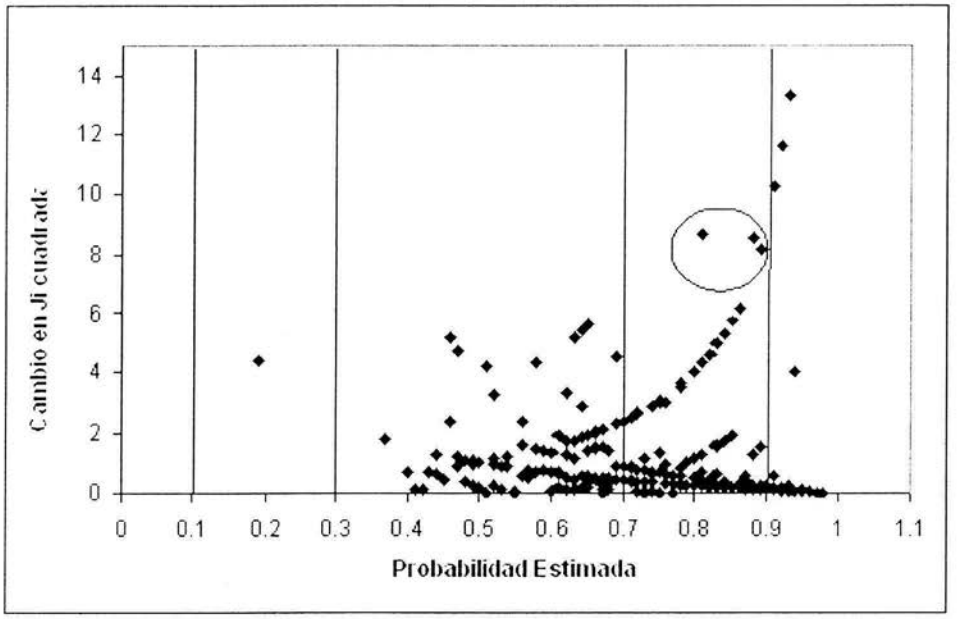
Variable	Inicial	sig.	Sin patrones anteriores	var.**	sig.
ID	0.98	0.000	1.04	6.1%	0.000
E		0.001			0.001
E_4	1.47	0.001	1.44	2.2%	0.002
E_5	1.28	0.007	1.26	1.3%	0.009
E_6	1.53	0.003	1.53	0.2%	0.003
E_7	2.13	0.000	2.17	2.2%	0.000
LOC	0.73	0.003	0.76	4.0%	0.002
EDAD	0.04	0.000	0.05	10.4%	0.000
Constant	-2.49	0.000	-2.63	5.7%	0.000
Var. Promedio*				4.0%	

*Variación Promedio**Variación Relativa

Tabla 3.22. Continuación de la tabla 3.21.

Como puede observarse en los cuadros anteriores el cambio en la magnitud de los coeficientes debido a la eliminación de los patrones de covaribales estudiados se encuentra en promedio alrededor del 4%, además se observa que los coeficientes estimados de todas las variables continúan siendo significativamente distintos de cero para cualquier caso de eliminación de patrones de covariables. Por lo anterior se concluye que la estimación de los parámetros del modelo no es sensible a la eliminación de ciertos patrones de covariables.

La tercera medida de diagnóstico consiste en el cambio directo en la magnitud del estadístico χ^2 de Pearson debido a la exclusión de ciertos patrones de covariables (observaciones de la muestra). La gráfica es la siguiente:



Gráfica 3.4. $\Delta\chi^2$ vs probabilidad estimada

Como puede observarse en la gráfica anterior, existen tres patrones de covariables en el rango de probabilidad estimada de 0.7 a 0.9 cuya influencia sobre el valor del estadístico χ^2 de Pearson parece importante, dado que los valores en este rango deben ser de tamaño moderado y estos parecen ser mayores que el resto. Por lo anterior, se estudiará la influencia que tiene la eliminación de dichos patrones de covariables en el cálculo del estadísitco χ^2 de Pearson. El siguiente cuadro presenta la información correspondiente a dichos patrones de covariables:

FOLIO	E	LOC	EDAD	ID	$\Delta\chi^2$	$\pi(\mathbf{X})$
685	4	1	38	0	8.66	0.81
1928	2	0	45	1	8.55	0.88
916	4	1	31	1	8.18	0.89

Tabla 3.23. Información de los patrones de covarible con comportamiento anormal de $\Delta\hat{\beta}$

En el siguiente cuadro se presenta el valor del estadístico χ^2 de Pearson, la variación relativa y el valor-p asociado a dicha prueba después de eliminado el patrón de covariable correspondiente.

	Inicial	Sin 685	var*	Sin 1928	var*	Sin 916	var*	Sin patrones anteriores	var*
χ^2	367.06	358.54	2.38%	358.54	2.38%	358.98	2.25%	341.96	7.34%
Grados de libertad	375	374		374		374		372	
sig.	0.61	0.71		0.71		0.70		0.87	

*Variación Relativa

Tabla 3.24. Variaciones en los valores de la tabla 3.15 debidas a la eliminación de patrones de covariables con comportamiento anormal de $\Delta\chi^2$

Como puede observarse a partir del cuadro anterior, las variaciones en el estadístico χ^2 de Pearson como resultado de la eliminación de algún patrón de covariable son muy pequeñas, además la conclusión sigue siendo la misma : el modelo ajusta de manera adecuada, por lo tanto el estadístico χ^2 de Pearson no es sensible a la eliminación de dichos patrones de covariables.

3.4 Evaluación del ajuste mediante validación

Con el objeto de verificar si el modelo es aplicable a distintas muestras, en algunas ocasiones es posible excluir una submuestra de las observaciones originales, seleccionada aleatoriamente, y desarrollar un modelo a partir de las observaciones restantes. En otras situaciones es posible obtener una nueva muestra para evaluar el ajuste del modelo previamente desarrollado. En este caso sólo será posible la primera situación, es decir, serán excluidas de manera aleatoria algunas observaciones a partir de la muestra original y se realizará tanto la estimación de los coeficientes como la evaluación del ajuste del modelo con las observaciones restantes. El número de observaciones con el cual se realizará la validación del modelo es de 352. El siguiente cuadro muestra la comparación de los coeficientes estimados utilizando el total de la muestra y la submuestra de validación.

Variable	Inicial	sig.	Submuestra	var.**	sig.
ID	0.98	0.000	0.64	53.6%	0.035
E		0.001			0.017
E_4	1.47	0.001	1.63	10.0%	0.006
E_5	1.28	0.007	1.39	7.8%	0.026
E_6	1.53	0.003	1.77	13.5%	0.008
E_7	2.13	0.000	2.20	3.5%	0.001
LOC	0.73	0.003	0.73	0.1%	0.029
EDAD	0.04	0.000	0.04	6.0%	0.000
Constant	-2.49	0.000	-2.49	0.0%	0.003
Var. Promedio*				11.8%	

*Variación Promedio **Variación Relativa

Tabla 3.25. Variaciones en las estimaciones de la tabla 3.9 debidas a la selección aleatoria de una submuestra

Como puede observarse en la tabla anterior, la variación más fuerte en la magnitud de los coeficientes estimados se presenta en la variable Identificación Partidista, sin embargo, los coeficientes de todas las variables continúan siendo significativamente distintos de cero con un 95% de confianza, lo cual comienza a dar evidencia que la estimación de los coeficientes es estable.

Ahora se presentan los resultados de la prueba de hipótesis de Hosmer y Lemeshow:

	Inicial	Submuestra	var*
χ^2	4.47	6.28	40.49%
Grados de libertad	8	8	
sig.	0.812	0.615	

*Variación Relativa

Tabla 3.26. Variaciones en los valores de la tabla 3.14 debidas a la selección aleatoria de una submuestra

El cuadro anterior indica que la variación en el estadístico de χ^2 para la prueba de Hosmer y Lemeshow es del 40%, sin embargo, no se puede rechazar que el modelo estimado a partir de la submuestra ajuste de manera adecuada con un 95 % de confianza.

Como conclusión a partir de la etapa de validación del modelo se obtiene que tanto la estimación de los coeficientes como el ajuste del modelo son estables para una submuestra seleccionada de manera aleatoria.

3.5 Análisis de las tablas de clasificación

A continuación se presentarán los resultados obtenidos después de analizar la tabla de clasificación para distintos puntos de corte de la probabilidad estimada, los resultados son los siguientes:

Punto de corte	% de casos correctamente clasificados	$Pe(\%)$	$pe(\%)$	d	$valor - p$
0.5	72.9	40.0	27.0	7.1	0.000
0.6	69.6	40.0	30.0	5.4	0.000
0.7	65.6	40.0	34.0	3.2	0.001
0.8	55.2	40.0	45.0	2.3	0.020
0.9	41.0	40.0	59.0	9.8	0.000

Tabla 3.27. análisis de la tabla de clasificación para distintos puntos de corte de la probabilidad estimada

Como puede observarse en la tabla anterior, Pe (la proporción de casos clasificados de manera incorrecta sin el modelo) es igual a 40%, la mayor diferencia entre este valor y pe (la proporción de casos clasificados de manera incorrecta con el modelo) ocurre cuando el punto de corte es igual a 0.9, sin embargo, en este caso $pe > Pe$, lo cual indica que el modelo para este punto de corte no es eficiente al momento de pronosticar dado que incrementa el error esperado. Por lo anterior, se debe poner atención en aquellos puntos de corte para los cuales se cumple $Pe > pe$, dado que esto indica que el modelo está reduciendo el error esperado al momento de pronosticar. Así, se tiene que el punto de corte con mayor diferencia entre $Pe > pe$ es igual a 0.5, siendo esta diferencia estadísticamente significativa al 95% de confianza (dado que $valor - p$ es menor a 0.05), lo cual lleva a concluir que este punto de corte es el mejor de ellos. La tabla de clasificación para el punto de corte igual a 0.5 es la siguiente:

	$\hat{Y}= 1$	$\hat{Y}= 0$	
$Y = 1$	448	39	487
$Y = 0$	145	46	191
	593	85	678

Tabla 3.28. Tabla de clasificación para el punto de corte 0.5

El porcentaje de casos clasificados de manera correcta de manera global es igual a 73% para el punto de corte igual a 0.5.

4. Aplicaciones del modelo

4.1 Interpretación de coeficientes

Las razones de momios estimadas a partir del modelo proporcionan información sobre las características de la gente que tiene mayor probabilidad de tener intención de votar el día de la elección. A continación se presenta la interpretación de las razones de momios estimadas para cada una de las variables incluidas en el modelo sociológico - psicológico.

EDUCACIÓN

Cabe recordar que la variable correspondiente a Educación tiene las siguientes categorías:

Categoría	Razón de momios
E_4 = Primaria	4.343
E_5 = Secundaria	3.592
E_6 = Preparatoria	4.618
E_7 = Universidad o más	8.391

Tabla 4.1. Razones de momios para la variable Educación

Teniéndose la categoría No estudió como referencia. De esta forma la interpretación de las razones de momios para cada una de las categorías es respecto a la categoría No estudió. Así tenemos las siguientes conclusiones

1) Una persona con Primaria o Secundaria es 4 veces más probable que tenga intención de votar que una persona sin estudios.

2) Una persona con Preparatoria es 5 veces más probable que tenga intención de votar que una persona sin estudios.

3) Una persona con Universidad o más es 8 veces más probable que tenga intención de votar que una persona sin estudios.

De esta manera podemos concluir que cualquier persona con algún nivel de estudios, sin importar el que este sea, manifiesta mayor intención por asistir a votar el día de las elecciones que alguna sin ningún tipo de instrucción. Además, en general se observa mayor intención de ir a votar el día de las elecciones entre aquellas personas con nivel de educación Universitario o mayor que aquellos cuyo nivel de estudios se encuentra entre Primaria y Preparatoria.

LOCALIDAD

Las categorías para el tipo de localidad en el cual se encuentra el hogar del entrevistado son las siguientes:

Categoría	Razón de momios
1 = Rural/Mixto	2.068

Tabla 4.2. Razón de momios para la variable Localidad

Tomando a la categoría Urbano como la de referencia. De esta forma se tiene la siguiente interpretación para esta variable:

Una persona que vive en una localidad Rural/Mixta es 2 veces más probable que tenga intención de votar que una persona que vive en unca localidad Urbana. Es decir, la gente de localidades Rurales o Mixtas, manifiesta mayor intención por asistir a votar el día de la elección que aquellas personas de localidades Urbanas.

IDENTIFICACION PARTIDISTA

Las categorías para la variable Identificación Partidista son las siguientes:

Categoría	Razón de momios
1 = Muy Identificados con algún partido político	2.666

Tabla 4.3. Razón de momios para la variable Identificación Partidista

Tomando a la categoría poco o algo identificados con algún partido político como la de referencia. De esta forma se tiene la siguiente interpretación para esta variable:

Una persona Muy identificada con algún partido político es 3 veces más probable que tenga intención de votar que una persona poco o algo identificada identificada con algún partido político, lo cual confirma la hipótesis sobre la cual está basada la escuela psicológica, es decir, es más probable que acudan a votar aquellos muy identificados con alguna fuerza política que aquellos que no están muy identificados.

EDAD

Dado que la estimación puntual y por intervalo de la razón de momios de una variable continua depende del valor de c_1, a continuación al momento de interpretar el coeficiente para la variable *EDAD* el valor de c_1 será fijado con un valor de 10, es decir, un incremento de 10 años de *EDAD*.

De esta forma, la siguiente tabla muestra la estimación tanto puntual como de intervalo de la razón de momios para la variable continua *EDAD*.

	$\hat{\beta}_{10}$	$10\hat{\beta}_{10}$	$\exp(10\hat{\beta}_{10})$	LI95%	LU95%
EDAD	0.044	0.439	1.551	1.323	1.818

Tabla 4.4. Razón de momios para la variable Edad

En lo que respecta a la interpretación de la razón de momios para la variable EDAD, podemos concluir que si una persona es 10 años mayor que otra, será 1.551 veces más probable que tenga intención de votar el día de la elección que la persona que es 10 años menor.

CONCLUSIONES

A partir de los resultados derivados de la aplicación de la regresión logísitica se concluye que la combinación de las teorías sociológica y psicológica explica de una manera más adecuada la participación en la elección presidencial del año 2000. La teoría racional también fue probada, sin embargo, por sí sola no puede explicar porqué los individuos asisten a votar o no. Es importante señalar que esto no significa que la teoría racional no funcione, ya que para este caso, la operacionalización de los elementos de la ecuación que determina la uitlidad de votar de un individuo puede no ser la más conveniente.

Se puede concluir que los principales factores que explican la participación en México en las elecciones presidenciales del año 2000 son el nivel de educación, el tipo de localidad (Urbano, Rural/Mixto), el nivel de identificación con algún partido político y la edad, variables explicativas de las teorías psicológica y sociológica. Así, a partir de esta investigación se sugiere que conforme mayor sea el nivel de educación alcanzado, la edad, la identificación partidista y se habite en una localidad Rural/Mixta se tendrá mayor intención de asistir a votar el día de la elección.

Una buena interpretación al hecho de que la gente más jóven tengan menos intención de votar es que este grupo no ha desarrollado ningún tipo de interés político. Los intereses de los más jóvenes son inmediatos y rara vez piensan en el mediano o largo plazo. Por otro lado, una persona con mayor nivel de educación está más conciente de las implicaciones que se derivan de una elección, dado que dichas decisiones afectan su entorno social y económico. Tradicionalmente la gente que habita localidades rurales/mixtas ha asisitido a votar ya sea por acarreo o por costumbre, mientras que la gente que habita zonas urbanas tiene otras actividades que ocupan su tiempo. Es natural que la gente que se identifica con algún partido político tenga mayor intención de votar ya que su prinicpal interés radica en apoyar a su partido.

La participación electoral es muy importante para el buen funcionamiento de toda democracia. Es por esto que es fundamental investigar los principales factores que la determinan ya que así se podrán conocer aquellos grupos de la población menos interesados en votar y de esta forma evitar el desinterés en eventos políticos tan importantes como el que es estudiado en este trabajo y que le dan sentido a la democracia en nuestro país.

ANEXO A. Niveles Socioeconómicos AMAI

A.1 Introducción

La AMAI (Asociación Mexicana de Agencias de Investigación de Mercados) representa el órgano regulador de las actividades de las agencias de investigación en México. Desde su creación, una de sus preocupaciones y ocupaciones funadamentales fue la de definir un conjunto de Niveles Socioeconómicos que se convirtieran en el estándar de la industria, así como la de establecer una regla que permitiera asignar a cualquier hogar el Nivel Socioeconómico que le corresponda, sin tener que ver físicamente el hogar.

Al paso de los años las reglas de asignación desarrolladas por el Comité de Niveles Socioeconómicos de la AMAI se han venido utilizando cada vez más, hasta lograr convertirse en un referente obligado para los practicantes de la investigación de mercado y de opinión pública del país y también para las agencias de investigación y empresas extranjeras que llevan a cabo estudios de mercado en nuestra república.

A.2 Preguntas necesarias para la determinación del NSE AMAI

Los niveles socioeconómicos de la AMAI son determinados mediante un árbol de clasificación a partir de las siguientes 13 preguntas:

1. Pensando en el Jefe de Familia de su hogar, ¿cuál fue el último año de estudios que completó? (espere respuesta, y pregunte) ¿Realizó otros estudios? (reclasificar en caso necesario).

1. No estudió

2. Primaria incompleta

3. Primaria completa

4. Secundaria incompleta

5. Secundaria completa

6. Carrera comercial

7. Carrera técnica

8. Preparatoria incompleta

9. Preparatoria completa

10. Licenciatura incompleta

11. Licenciatura completa

12. Diplomado o Maestría

13. Doctorado

14. NS/NC

2. ¿Cuál es el total de piezas y/o habitaciones con que cuenta su hogar?, por favor no incluya baños, medios baños, pasillos, patios y zotehuelas. (Si el entrevistado pregunta específicamente si cierto tipo de pieza pueda incluirla o no, debe consultarse la referencia que se anexa)

1. Uno

2. Dos

3. Tres

4. Cuatro

5. Cinco

6. Seis

7. Siete o más

Sí cuentan: recámaras, sala, cocina, comedor, cuarto de lavado, cuarto de TV, biblioteca, cuarto de servicio si está dentro de su vivienda, tapancos, sótano y el garage o cochera sólo si está techado y rodeado de paredes y puertas que impidan mirar al interior del mismo.

No cuentan: cobachas, tienditas que estén dentro de la vivienda, garages o cocheras que no tengan techo ni tres paredes y una puerta que impida ver al interior de ellos.

3. ¿Cuántos baños completos con regadera y W.C.(excusado) hay para uso exclusivo de los integrantes de su hogar?

0. Cero

1. Uno

2. Dos

3. Tres

4. Cuatro o más

4. En su hogar ¿cuenta con calentador de agua o boiler?

0. No

1. Sí

5. Contando todos los focos que utiliza para iluminar su hogar, incluyendo los de techos, paredes y lámparas de buró o piso, dígame ¿cuántos focos tiene su vivienda?

1. Cinco o menos

2. Entre seis y diez

3. Entre once y quince

4. Entre dieciséis y veinte

5. Veintiuno o más

6. ¿El piso de su hogar es predominantemente de tierra, o de cemento, o de algún otro tipo de acabado?

1. Tierra

2. Cemento (firme de)

3. Otro tipo de material o acabado

7. ¿Cuántos automóviles propios, excluyendo taxis, tienen en su hogar?

0. Ninguno

1. Uno

2. Dos

3. Tres y más

8. ¿Cuenta su hogar con aspiradora que funcione?

0. No

1. Sí

9. ¿Cuenta su hogar con lavadora de ropa que lave y enjuage automáticamente que funcione?

0. No

1. Sí

10. ¿Cuenta su hogar con horno de microondas que funcione?

0. No

1. Sí

11. ¿Cuenta su hogar con tostador eléctrico de pan que funcione?

0. No

1. Sí

12. ¿Cuenta su hogar con videocassetera que funcione?

0. No

1. Sí

13. ¿Cuenta su hogar con Computadora Personal propia que funcione?

0. No

1. Sí

A.3 Descripción de los NSE AMAI

NIVEL A/B

Este es el estrato que contiene a la población con el más alto nivel de vida e ingresos del país.

- Perfil Educativo y Ocupacional del Jefe de Familia

En este segmento el Jefe de Familia tiene en promedio un nivel educativo de Licenciatura o mayor. Los jefes de familia de nivel AB se desempeñan como grandes o medianos empresarios (en el ramo industrial, comercial y de servicios); como gerentes, directores o destacados profesionistas. Normalmente laboran en importantes empresas del país o bien ejercen independientemente su profesión.

- Perfil del Hogar

Los hogares de las personas que pertenecen al nivel AB son casas o departamentos propios de lujo que en su mayoría cuentan con 6 habitaciones o más, dos 2 ó 3 baños completos, el piso de los cuartos es de materiales especializados distintos al cemento y todos los hogares de este nivel, tienen boiler. En este nivel las amas de casa cuentan con una o más personas a su servicio, ya sean de planta o de entrada por salida. Los hijos de estas familias asisten a los colegios privados más caros o renombrados del país, o bien a colegios del extranjero.

- Artículos que posee

Todos los hogares de nivel AB cuentan con al menos un auto propio, regularmente es del año y algunas veces de lujo o importado, y tienden a cambiar sus autos con periodicidad de aproximadamente dos años. Los autos usualmente están asegurados contra siniestros.

- Servicios

En lo que se refiere a servicios bancarios, estas personas poseen al menos una cuenta de cheques (usualmente el jefe de familia), y tiene más de 2 tarjetas de crédito, así como seguros de vida y/o de gastos médicos particulares.

- Diversión/Pasatiempos

Las personas de este nivel asisten normalmente a clubes privados. Suelen tener casa de campo o de tiempo compartido. Además, más de la mitad de la gente de nivel alto ha viajado en avión en los últimos 6 meses, y van de vacaciones a lugares turísticos de lujo, visitando al menos una vez al año el extranjero y varias veces el interior de la república. La televisión ocupa parte del tiempo dedicado a los pasatiempos, dedicándole menos de dos horas diarias.

- Ingreso Mensual Familiar

Al menos $77,000.00

NIVEL C+

En este segmento se consideran a las personas con ingresos o nivel de vida ligeramente superior al medio.

- Perfil educativo del Jefe de Familia

La mayoría de los jefes de familia de estos hogares tiene un nivel educativo de licenciatura y en algunas ocasiones cuentan solamente con educación preparatoria. Destacan jefes de familia con algunas de las siguientes ocupaciones: empresarios de compañías pequeñas o medianas, gerentes o ejecutivos secundarios en empresas grandes o profesionistas independientes.

- Perfil del Hogar

Las viviendas de las personas que pertenecen al Nivel C+ son casas o departamentos propios que cuentan con 5 habitaciones o más, 1 ó 2 baños completos. Uno de cada cuatro hogares cuenta con servidumbre de planta o de entrada por salida. Los hijos son educados en primarias y secundarias particulares, y con grandes esfuerzos terminan su educación en universidades privadas caras o de alto reconocimiento.

- Artículos que posee

Casi todos los hogares poseen al menos un automóvil, aunque no tan lujoso como el de los adultos de nivel alto. Usualmente tiene un auto familiar y un compacto. Normalmente, sólo uno de los autos está asegurado contra siniestro. En su hogar tiene todas las comodidades y algunos lujos; al menos dos aparatos telefónicos, equipo modular, compact disc, dos televisores a color, videocassettera, horno de microondas, lavadora, la mitad de ellos cuenta con inscripción a televisión pagada y PC. Uno de cada tres tiene aspiradora. En este nivel las amas de casa suelen tener gran variedad de aparatos electrodomésticos.

- Servicios

En cuanto a servicios bancarios, las personas de nivel C+ poseen un par de tarjetas de crédito, en su mayoría nacionales, aunque pueden tener una internacional.

- Diversiones/Pasatiempos

Las personas que pertenecen a este segmento asisten a clubes privados, siendo éstos un importante elemento de convivencia social. La televisión es también un pasatiempo y pasan en promedio poco menos de dos horas diarias viéndola. Vacacionan generalmente en el interior del país, y a lo más una vez al año salen al extranjero.

- Ingreso Mensual Familiar

Varía desde $30,000 hasta $76,999

NIVEL C

En este segmento se considera a las personas con ingresos o nivel de vida medio.

- Perfil Educativo del Jefe de Familia

El jefe de familia de estos hogares normalmente tiene un nivel educativo de preparatoria y algunas veces secundaria. Dentro de las ocupaciones del jefe de familia destacan pequeños comerciantes, empleados de gobierno, vendedores, maestros de escuela, técnico y obreros calificados.

- Perfil de Hogares

Los hogares de las personas que pertenecen al nivel C son casa o departamentos propios o rentados que cuentan en promedio con 4 habitaciones y 1 baño completo. Los hijos algunas veces llegan a realizar su educación básica (primaria/secundaria) en escuelas privadas, terminando la educación superior en escuelas públicas.

- Artículos que posee

Dos de cada tres hogares de clase C sólo posee al menos un automóvil, regularmente es para uso de toda la familia, compacto o austero, y no de modelo reciente; casi nunca está asegurado contra siniestros. Cuentan con algunas comodidades: 1 aparato telefónico, equipo modular, 2 televisores, y videocassettera. La mitad de los hogares tiene horno de microondas y uno de cada tres tiene televisión pagada y PC. Muy pocos cuentan con servidumbre de entrada por salida.

- Servicios

En cuanto a instrumentos bancarios, algunos poseen tarjetas de crédito nacionales y es poco común que usen tarjeta internacional.

- Diversión/Pasatiempos

Dentro de los principales pasatiempos destacan el cine, parques públicos y eventos musicales. Este segmento usa la televisión como pasatiempo y en promedio la ve diariamente por espacio de dos horas. Gustan de los géneros de telenovela, drama y programación cómica. Estas familias vacacionan en el interior del país, aproximadamente una vez por año van a lugares turísticos accesibles (poco lujosos).

- Ingreso Mensual Familiar

Varía desde $10,000 hasta $29,999

NIVEL D+

En este segmento se consideran a las personas con ingresos o nivel de vida ligeramente por debajo del nivel medio, es decir es el nivel bajo que se encuentra en mejores condiciones (es por eso que se llama bajo/alto o D+).

- Perfil Educativo del Jefe de Familia

El jefe de familia de estos hogares cuenta en promedio con un nivel educativo de secundaria o primaria completa. Dentro de las ocupaciones se encuentran taxistas (choferes propietarios del auto), comerciantes fijos o ambulantes (plomería, carpintería), choferes de casas, mensajeros, cobradores, obreros, etc.Suelen existir dentro de esta categoría algunos jefes de familia que tienen mayor escolaridad pero que como resultado de varios años de crisis perdieron sus empleos y ahora se dedican a trabajar en la economía informal.

- Perfil del Hogar

Los hogares de las personas que pertenecen a este nivel son, en su mayoría, de su propiedad, aunque algunas personas rentan el inmueble. Cuentan en promedio con 3 o más habitaciones y 1 baño completo. Algunas viviendas son de interés social. Los hijos asisten a escuelas públicas.

- Artículos que posee

En estos hogares uno de cada cuatro posee automóvil propio, por lo que en su mayoría utilizan los medios de transporte público para desplazarse. Cuentan con: un aparato telefónico, 1 televisor a color, y 1 equipo modular barato. La mitad de los hogares tiene videocassettera y línea telefónica. No tienen aspiradora y muy pocos llegan a contar con PC.

- Servicios

Los servicios bancarios que poseen son escasos y remiten básicamente a cuentas de ahorros, cuentas o tarjetas de débito y pocas veces tienen tarjetas de crédito nacionales.

- Diversión/pasatiempos

Generalmente las personas de este nivel asisten a espectáculos organizados por la delegación y/o por el gobierno, también utilizan los servicios de poli-deportivos y los parques públicos. La televisión también es parte importante de su diversión y atienden preferentemente a las telenovelas y a los programas de concurso. Este grupo tiende a ver televisión diariamente por un espacio algo superior a dos horas.

- Ingreso Mensual Familiar

Varía de $6,000 a $9,999

NIVEL D

El nivel D está compuesto por personas con un nivel de vida austero y bajos ingresos.

- Perfil Educativo del Jefe de Familia

El jefe de familia de estos hogares cuenta en promedio con un nivel educativo de primaria (completa en la mayoría de los casos). Los jefes de familia tienen actividades tales como obreros, empleados de mantenimiento, empleados de mostrador, choferes públicos, maquiladores, comerciantes, etc.

- Perfil del Hogar

Los hogares de nivel D son inmuebles propios o rentados. Las casas o departamentos cuentan con al menos dos habitaciones y 1 baño que puede ser completo o medio baño. La mitad de los hogares cuenta con boiler (calentador de agua) y lavadora. Estas casas o departamentos son en su mayoría de interés social o de rentas congeladas (tipo vecindades). Los hijos realizan sus estudios en escuelas del gobierno.

- Artículos que posee

Las personas de este nivel suelen desplazarse por medio del transporte público, y si llegan a tener algún auto es de varios años de uso. La mayoría de los hogares cuenta con un televisor y/o equipo modular barato. Uno de cada cuatro hogares tienen videocassettera y línea telefónica.

- Servicios

Se puede decir que las personas de nivel D prácticamente no poseen ningún tipo de instrumento bancario.

- Diversión/Pasatiempos

Asisten a parques públicos y esporádicamente a parques de diversiones. Suelen organizar fiestas en sus vecindades. Toman vacaciones una vez al año en excursiones a su lugar de origen o al de sus familiares. Cuando ven televisión su tipo de programación más favorecida son las telenovelas y los programas dramáticos. En promedio ven televisión diariamente por espacio de dos y media horas.

- Ingreso Mensual Familiar

Varía de $2,250 a $5,999

NIVEL E

El nivel E se compone de la gente con menores ingresos y nivel de vida en todo el país.

- Perfil Educativo del Jefe de Familia

El jefe de familia de estos hogares cursó, en promedio, estudios a nivel primaria sin completarla, y generalmente tiene subempleos o empleos eventuales.

- Perfil del Hogar

Estas personas usualmente no poseen un hogar propio (sobre todo en la Cd. de México), teniendo que rentar o utilizar otros recursos para conseguirlo (paracaidistas). En un solo hogar suele vivir más de una generación. Sus viviendas poseen 1 ó 2 cuartos en promedio, mismos que utilizan para todas las actividades (en ellos duermen, comen, etc.). La mayoría de los hogares no tienen baño completo propio (dentro de su casa). No poseen agua caliente (calentador de agua), ni drenaje. Los techos son de lámina y/o asbesto y el piso muchas veces

es de tierra. Difícilmente sus hijos asisten a escuelas públicas y existe un alto nivel de deserción escolar.

- Artículos que posee

Estos hogares son muy austeros, tienen un televisor y un radio y en pocos casos videocassettera. La mitad de los hogares de clase E poseen refrigerador.

- Servicios

Este nivel no cuenta con ningún servicio bancario o de transporte propio.

- Diversión/Pasatiempos

Su diversión es básicamente la radio y la televisión. Dentro de este último medio, la programación de telenovelas, drama y concursos son los que más atienden. En promedio ven televisión diariamente por espacio de casi tres horas.

- Ingreso Mensual Familiar: menor de $2,250

ANEXO B. Encuestas preelectorales

Las encuestas prelectorales se refieren al levantamiento de una encuesta a los empadronados, con cierta anticipación (algunas semanas) antes de alguna contienda electoral. Es decir, se entrevista a una muestra representativa de empadronados, con el objeto de conocer si votarán o no, si ya decidieron por quién votarán, en específico por qué partido votarán, las razones asociadas a su decisión, y ciertas características demográficas del votante. Además, se pueden incluir aspectos relacionados con imagen de candidatos, votos anteriores, opinión pública, e impacto de campañas políticas.

Esta encuesta, si se realiza dentro de una muestra representativa de secciones electorales, es útil para conocer las preferencias actuales del electorado, y, se han utlizado para pronosticar los resultados de la votación, así como para determinar el voto por segmentos de la población, el "perfil" de los votantes de cada partido, y las motivaciones asociadas al voto. No obstante, su elaboración tiene aspectos analíticos particulares, ya que generalmente existirá un grupo de "indecisos" que si votarán pero en el momento de la encuesta aún no saben o prefieren no declarar por que partido. Además, los resultados de este tipo de encuestas deben tomarse con cuidado, ya que la opinión de los votantes en el momento de responder la entrevista no es necesariamente la misma que tendrán en el momento de ejercer su voto.

Para seleccionar la muestra de entrevistados, primero se selecciona un conjunto de secciones electorales, con probabilidad proporcional al tamaño de la sección. Posteriormente, se visita un número determinado de hogares dentro de la sección electoral, realizando arranque aleatorio y salto sistemático de hogares para que sean seleccionados aleatoriamente, y cada hogar dentro de la sección tenga la misma probabilidad de selección. Dentro del hogar, también se selecciona aleatoriamente al entrevistado.

En este momento, el entrevistador registra el sexo del entrevistado y le aplica las preguntas (con respuestas previamente codificadas) de una primera parte del cuestionario de la encuesta preelectoral. Estas preguntas, básicamente, se refieren a datos demográficos, opinión de la situación personal y del país o de su estado, e imagen de los candidatos. Después de realizar estas preguntas al votante, se le entrega el cuestionario, el cual tiene impresa una réplica de la boleta electoral. Se le pide que marque el cuestionario de la misma forma en que espera hacerlo con la boleta electoral oficial, de una manera privada. También se le pide que responda en secreto otras preguntas de opinión especialmente sensibles. Después de esto, el entrevistado deposita el cuestionario en una urna especial. En todo momento su voto permanece secreto.

BIBLIOGRAFIA

(1989) Hosmer David W. and Lemeshow Stanley, Applied Logistic Regression. Wiley Series in Probability and Mathematical Statistics.

(1997) Ryan P. Thomas, Modern Regression Methods. Wiley Series in Probability and Mathematical Statistics.

(2000) Pampel Fred C. Logistic Regression. Series: Quantitative Aplications in the Social Sciences. A Sage University Paper.(132).

(2000) Menard Scott, Applied Logistic Regression Analysis. Series: Quantitative Aplications in the Social Sciences. A Sage University Paper.(106).

(1992) Kleinbaum David G. Logistic Regression, A self learning text. Statistics in the Health Science. Springer.

(1999) Hutcheson Graeme and Sofroniou Nick. The Multivariate Social Scientist. Sage Publications.

(1987) Harrop Martin and Miller William. Elections and voters. McMillan. Chaptet 6

(1993) Aldrich John. Rational Choice and turnout. American Journal of Political Science. Vol. 37

(1957) Downs Anthony. An economic theory of democracy. Harper Collins

(1968) Riker William H. and Ordeshook Peter. A theory of the calculus of voting. American Political Science Review

(1981) Gravelle and Reef. Microeconomics. London: Longman