UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

ANÁLISIS DEL ABSTENCIONISMO ELECTORAL 2006, PARA LA PRESIDENCIA DE LOS ESTADOS UNIDOS MEXICANOS, UTILIZANDO REGRESIÓN LOGÍSTICA

TESINA

QUE PARA OBTENER EL TITULO DE
ACTUARIO

PRESENTA

GLORIA OMARA GONZÁLEZ MORALES

Asesor: ACT. MAHIL HERRERA MALDONADO

Fecha: Enero, 2010

AGRADECIMIENTOS

A mis padres por haberme incentivado el gusto por aprender, amor y apoyo.

Al pueblo de México por darme la oportunidad de estudiar en su máxima casa de estudios pública y gratuita.

A la gente que me ha acompañado en el transcurso de mis estudios y que siempre los guardare en mis recuerdos.

ÍNDICE

Capítulo 3 Desarrollo del estudio

Capítulo 4 Conclusiones y observaciones

Resumen

En el presente trabajo se aplicará el modelo de Regresión Logística para analizar el fenómeno del abstencionismo en el proceso de elecciones para Presidente de la República 2006 a través de variables socioeconómicas que miden el nivel de marginación de la población por estado, emitidas por el Consejo Nacional de Población (CONAPO) y el Instituto Nacional de Estadística Geografía e Informática (INEGI) en el año 2005, así como, los resultados electorales emitidos por el Instituto Federal Electoral 2006 (IFE) con la finalidad de investigar si influyen de manera significativa en la no participación de los ciudadanos en las elecciones presidenciales.

Introducción

El contenido del trabajo presenta un análisis de variables de marginación económica que podrían influir en el abstencionismo durante el proceso de elección para Presidente de la República de los Estados Unidos Mexicanos 2006, mediante el método de regresión logística, con el objetivo de identificar si estas variables influyen en el desarrollo de dicho fenómeno social.

Ya que ante la indiferencia del sistema de proveerles un bienestar social con servicios de primera necesidad como educación, servicios públicos o un salario remunerativo, éste sector, podría optar por no tomar interés en sus mecanismos políticos o protestar de forma pasiva de esta manera.

Por lo que la intención es visualizar cuáles son los factores socioeconómicos de riesgo que influyen sobre esta población y en qué medida incrementa la probabilidad de incidencia, así como, plantear un posible modelo de predicción del fenómeno y determinar de esa manera, si es que esto, se

Capítulo primero
El abstencionismo electoral

1.1 Definición del abstencionismo electoral

El abstencionismo es un termino derivado de abstención, del latín *abstencionis,* que significa en su acepción usual inhibición o privación y que, referido a la política, es la expresión doctrinal que define la no participación de los ciudadanos en las tareas públicas o el que no ejerciten un derecho o una función también pública.

En términos generales, se define como la no participación de los ciudadanos en los diferentes eventos de la vida política de un país; se puede manifestar de manera muy concreta cuando aquellos no ejercen su derecho ni cumplen con la obligación cívica de votar en los procesos electorales, o bien mediante una actitud pasiva y apática ante los diferentes actos y actividades políticas. (*Instituto Nacional de Estudios Políticos, AC 2005*).

A su vez existe otra variante del abstencionismo, el cual es "cívico" o "activo", presentándose cuando el elector deposita en las urnas la boleta en blanco o anulada y no contribuye en el éxito de una elección.

Otra forma de interpretación, puede ser una declaración implícita de conformidad con el status quo, debido a la sensación de ser mas una forma de censura que de apoyo, por lo cual deciden no votar, ya que la situación no lo requiere si la votación es reducida, implicando mayor flexibilidad en la actuación de los políticos y contribuyendo a la estabilidad y perpetuidad del sistema. Observado principalmente cuando los abstencionistas en condiciones normales son indiferentes y no manifiestan preferencia alguna, si concurren a las urnas arrastradas por un ambiente real o propagandístico, que los convence que su voto es el de la mayor importancia.

El abstencionismo electoral presenta a su vez las siguientes fuentes, las cuales se combinan o se presentan como único factor:

Demográfica. Los abstencionistas poseen menos recursos personales como ingresos, educación y otros medios que les impulsarían en la participación electoral.

Socio psicológico. Quienes se abstienen manifiestan problemas tales como alineación e insatisfacción política y sentimientos de baja eficiencia de sus acciones políticas.

Contextual. Los abstencionistas no son atraídos por las campañas, ni por la competencia entre los candidatos o por alguna razón no pueden cumplir los requisitos de registro como electores.

Racional. Quienes se abstienen toman una decisión racional que considera los costos y los beneficios de acudir a votar a las urnas y por ejemplo, perciben a la política como deshonesta y rechazan a los políticos profesionales por su doble moral.

Técnica. Obedece a razones de fuerza mayor como enfermedad, ausencia, distancia a la casilla, errores en la inscripción como elector, si la elección tiene lugar en un día festivo o laborable, estado de tiempo y similares.

Indecisión. Entran los electores que desean y pueden votar, que no están descontentos o resignados al sistema político y que tratan de formarse un juicio sobre las opciones que se le ofrecen, pero que son incapaces de encontrar argumentos que lo movilicen en uno u otro sentido, por lo que se abstienen de participar (*Instituto Nacional de Estudios Políticos, AC 2005*).

1.2 Factores de influencia y causas de el abstencionismo

Entre las diversos factores se considera el abstencionismo como un indicador del grado de despolitización que sufre una sociedad, dado que la política se ha desvinculado de lo social, de lo ético y de lo cultural, como lo señalo Marx, *lo político no es otra cosa que una forma protocolizada, institucionalizada y oficial de lo social mismo, forma que cuando se separa y autonomiza de ese sustrato social que ella expresa, comienza a pervertirse y girar en el vacio, pues no hay partido, grupo u organización políticos que valgan, si no tiene detrás de si, como su respaldo y apoyo real, a movimientos, fuerzas o grupos sociales que los soporten, apoyen y retroalimenten constantemente*. (Carlos Aguirre Rojas, Para comprender el mundo actual, una gramática de larga duración).

Sergio Rodríguez Lascano considera que la idea de que la política es un espacio vedado para el ciudadano común posee dos matrices: una de derecha y otra de izquierda.

La de derecha, que busca la despolitización de la sociedad con el objetivo de obligarla a refugiarse al individualismo, en su trabajo individual, a lo más en su familia y en una especie de hedonismo.

La de izquierda, que busca explicar que la política es un espacio para los especialistas, los que entienden la teoría del Estado (en un momento donde el estado vive la peor crisis), donde están solo aquellos que poseen un nivel de conciencia superior y que saben todo sobre la correlación de fuerzas, la táctica y la estrategia.

En el hecho mismo de que la agenda electoral no es puesta por el elector, ni tiene la posibilidad de proponer a los candidatos y sus propuestas que plantea, lo que provoca que sus ofrecimientos no concuerden precisamente con las necesidades de la gente, dando como consecuencia que los votantes no sienta una conexión directa entre su voto individual y el resultado de las elecciones, ya que el sistema no contempla a las minorías que expresaron su decisión.

A su vez, estas dos corrientes se encuentran inmersas entre la sociedad mexicana contemporánea, dado que al analizar las repuestas emitidas a la Encuesta Nacional sobre Cultura y Practicas Ciudadanas 2001 y 2003 de la SEGOB, los resultados al cuestionar sobre su interés por la política en el año 2003 el 44% dijo no hablar sobre esta, mientras que el 56% considero que la política es demasiado complicada y el 78% dijo no haber leído las noticias políticas durante la semana pasada. Mientras que en el año 2001, 51% está en desacuerdo con que la gente es solidaria y para el año 2003 el 56% afirmo lo mismo, lo que da una perspectiva de la individualización que sentimos en nuestra sociedad, esta predisposición negativa inhibe la formación de redes ciudadanas y de capital social que sustente el actual sistema.

Lo anterior, da como consecuencia común del abstencionismo:

a) Desacuerdo y rechazo de la política gubernamental.
b) Carácter antidemocrático del sistema electoral.
c) Propuestas poco interesantes de los diferentes partidos o de sus candidatos.
d) Expresión de protesta y rebeldía ante las condiciones políticas prevalecientes.

Éste también se presenta con mayor intensidad entre sectores de la población que poseen un alto grado de marginación, debido a su falta de inclusión en el sistema, ya que no existe una fuerza política que represente sus intereses, remarcándose en sectores con las siguientes características:

a) Viven en zonas rurales
b) Tiene bajo nivel de escolaridad
c) Son de sexo femenino
d) Son de edad avanzada o muy jóvenes
e) Tiene bajos ingresos o trabajan por cuenta propia

Que en mayor parte son los menos favorecidos por el sistema y cultura de nuestra actual sociedad.

También se señalan cinco factores macro políticos del abstencionismo:

1. Legales, como el voto obligatorio o registro electoral.
2. Sistema de partidos (Número de partidos, competitividad, polarización, etc.)

3. Características de los partidos, como apoyo electoral y segmentación.
4. Sistema político en cuanto al número de cámaras de representantes, estabilidad o inestabilidad.
5. Económicos (desempleo, crisis y prosperidad).

1.3 Abstencionismo electoral en la historia reciente de los Estados Unidos Mexicanos

El abstencionismo dentro de la historia reciente de México, es un fenómeno que depende en gran medida de la situación política que guarda el país, pero sin duda las cifras muestran a su vez la dimensión del fenómeno al pasar de los años.

Es decir, hace 16 años, en 1991, se contaba con una lista nominal de 36 millones de electores, 12 millones de estos no emitieron su voto, el equivalente al 33%, seis años mas tarde, este listado se incremento en 15 millones, sin embargo 22 millones no sufrago, es decir, en esta ocasión fue el 43%, por ultimo en la elección federal del 2003, cuando la lista nominal rebasaba los 64 millones, 37 millones dejaron de ejercer su derecho, esto es, el 58%.

En México, el nivel de la elección es un factor determinante en el índice de abstencionismo, ya que en las elecciones presidenciales, concurren los votantes en mayor numero debido a que tienen una mayor difusión en los medios de comunicación, además de que el grado de competitividad incrementa, demostrando que tan buenas son las propuestas y atractivas las imágenes de los candidatos.

Históricamente se ha encontrado tasas de abstencionismo superiores al 40%, en las elecciones presidenciales de 1988 se encontró que el 50% del padrón se abstuvo, mientras que, en la elección presidencial de 1994 se obtuvo la menor de todas con tan solo el 24.15%, pero nuevamente en los comicios del 2000 se incremento al 36.03%.

Por lo que, a pesar de el enorme gasto que se hace en propaganda televisiva y el uso de estrategias de “marketing político”, los partidos no logran atraer a los electores, siendo un caso extremo el ocurrido en el 2005 en el Estado de México, en el que este llego al 58%.

Analizando la Encuesta Nacional sobre Cultura y Practicas Ciudadanas 2001 y 2003 de la SEGOB, podemos observar el ambiente de los votantes. Por ejemplo, cuando se le pregunto a la gente en el año 2001; ¿Qué tanta confianza le tiene a las instituciones?, tan solo el 5.36% de los encuestados contestaron que tenían mucha confianza en los partidos políticos y el 18.81% contestó que tan solo tenían algo de confianza en ellos. Mientras que en el año 2005 la respuesta a la pregunta en la escala del 0 es nada a 10 es mucho, ¿Qué tanto confía en las siguientes instituciones? Nuevamente los partidos políticos obtuvieron una de las menores calificaciones (6.4), sólo

están por encima de la policía (6.2), lo que demuestra nuevamente el grado de inconformidad de los ciudadanos.

Y al preguntarle a la gente ¿Qué tan satisfecho está usted con la democracia que tenemos hoy en México? En el año 2003 el 60% de los encuestados contestó que se encuentra poco o nada satisfecho y al preguntar que opinan sobre si México vive una democracia, a esta cuestión en 2001 el 37% opina que no existe, es decir un tercio de los encuestados, mientras que en 2003 fue el 22%.

Otro punto a destacar es el manejo de la política, ya que al preguntar ¿Qué es lo que toman en cuenta al elaborar las leyes los diputados? En 2003 el 63% opina que toma en cuenta sólo los intereses de su partido, mientras que en 2001 el 64% considero que era poco importante o nada importante el trabajo de la Cámara de Diputados.

Por ultimo de acuerdo con Transparencia Internacional, la percepción de la sociedad sobre la falta de transparencia del gobierno no ha mejorado durante los últimos diez años. En su Índice de Percepción de la Corrupción 2006, México está ubicado en el lugar 70 de un total de 163 países, con una calificación de 3.3 en una escala de 0 a 10.

1.4 Proceso electoral 2006

La alternancia constituye en las democracias llamadas poliárquicas, competencias entre partidos políticos que viven en situaciones de tendencias iguales para ganar las elecciones, donde los contendientes posean un poder electoral similar en términos de los intereses que representan y el número de votos que obtengan. Pero cuando estos intereses se presentan antagónicos o excluyentes, se dice que los procesos electorales terminan en conflicto y desbordan las reglas institucionales.

Este fue el caso de las elecciones para la presidencia de la república del año 2006, en el que se dice se vivió una polarización entre dos fuerzas que competían por el poder y que a llevado a la mesa de discusión la validez de las reglas electorales existentes, así como, la necesidad de una reforma electoral, ya que por darle legalidad a estas, se perdió la legitimidad, que supone el convencimiento de la población del proceso electoral en su totalidad.

Este fenómeno es producto de la alternancia que se comenzó a vivir en la década de los ochenta, en específico en la reforma electoral de año de 1977, donde el partido dominante PRI, flexibilizo las reglas para permitir el ingreso de nuevas fuerzas a la pelea por el poder político y así legitimar el sistema de partidos existente. Es a partir de tales reformas que cobró importancia el estudio del voto.

Esto provocó entre los electores un proceso continuo de redistribución del voto a favor de nuevas fuerzas políticas a costa del PRI que hasta entonces concentro la totalidad del electorado.

El proceso de redistribución según lo señala Carlos Sirvent (2001) inicio en el año 1989, en el que el Partido Acción Nacional (PAN) triunfa en las elecciones para gobernador en el estado de Baja California.

A partir de ese momento se comienza a vivir un proceso sorprendente de desplazamiento del voto hacia dos fuerzas políticas: el Partido de la Revolución Democrática (PRD) y el Partido Acción Nacional (PAN), que junto con el Partido Revolucionario Institucional formarían el régimen tripartidista.

Es así que Sirvent afirma que la distribución de este voto pasa necesariamente por elecciones críticas, desplazamientos masivos del electorado de unos partidos a otros y por un realineamiento permanente de los votantes.

En el caso de la alternancia, encontramos movimientos masivos de electores que se mantiene estables y otra con una alta volatilidad entre estos,

señalando una descomposición de viejos lazos en los que se funda el voto duro y que se enfoca al fenómeno de realineamiento electoral.

El realineamiento electoral es una herramienta que nos ayuda en el entendimiento de los movimientos de la votación desde el punto de vista de agregación de votos de los ciudadanos, las preferencias electorales, y la volatilidad de las mismas durante periodos específicos. Donde el realineamiento se refiere originalmente a un cambio radical de las preferencias partidarias del electorado, al pasar de un partido a otro en una elección especifica (Carlos Sirvent, 2001). La cual fue introducida por V. O. Key en 1953., donde sobresalen las elecciones criticas, definidas como *"Un tipo de elección en la que ocurre un realineamiento profundo y durable entre partidos"*.

Se entendería entonces que las elecciones del pasado 2 de julio del 2006 podrían considerarse dentro de los patrones de una elección critica debido a que, en ésta, se encontraron factores como el cambio del apoyo del electorado de un partido a otro, marcado por un amplio debate sobre temas de interés social y que definían la polarización de los votantes, debido a que una partido pretendía optar por un sistema de centro izquierda, en las que sobresalían propuestas de protección a los sectores desprotegidos y marginados de la sociedad regresando a un modelo de estado benefactor pero sin un cambio profundo en el modelo económico, como por ejemplo ayuda a personas de la tercera edad y madres solteras entre otros, los cuales se consideraron desde la perspectiva del partido de derecha como medidas populistas que en realidad no solucionaban la brecha social existente.

Mientras que el partido de derecha proponía la continuidad del modelo económico aplicado desde la época de presidentes "Priistas", basado en el neoliberalismo e implementado internacionalmente.

En un ambiente donde la descomposición política se lleva acabo, así como un desprendimiento de lealtades partidarias y alineación hacia nuevas preferencias, revelando una crisis en la esfera política que trae como consecuencia una reformulación de las maneras de participación política y de representación de intereses, así como una replanteamiento de las políticas públicas hacia el favorecimiento de nuevos intereses y nuevos consensos (Schattschneider, 1960). En el que la polarización entre los votantes se incrementó con una fuerza desmesurada, debido al uso de campañas publicitarias agresivas y sin propuesta, que tan solo buscaban el desprestigio de ambas fuerzas y la vacuidad de contenido en las campañas. Aunado a la falta de credibilidad en los resultados electorales, debido a la incertidumbre.

Para Sundquist (Citado por Sirvent, 1973), un realineamiento es un cambio durable en los patrones de comportamiento político. Este estudio establece cinco variables:

a) Existencia de temas específicos en la elección, provenido de amplias y profundas preocupaciones sociales.
b) Capacidad de un tema para provocar resistencia o movilización entre bloques de votantes.
c) Existir un liderazgo capas de promover un cambio electoral de grandes dimensiones y durable.
d) División dentro de las estructuras partidarias de los partidos dominantes que permitan movilidad electoral a lo largo de ellas.
e) Debilidad y fortaleza entre identidades partidarias, misma que es relevante para entender la descomposición de los grupos partidarios de apoyo electoral y su posible recomposición a lo largo de líneas nuevas de identificación partidista.

1.4.1 Abstencionismo en el proceso electoral 2006

Sirvent afirma que durante los últimos años, existe una creciente preocupación de los investigadores sobre lo que ha sido llamado declinamiento y deslizamiento partidista.

Ambos elementos son encontrados cuando un proceso de transición de realineamiento no es conclusivo.

El deslizamiento partidario está caracterizado por un alejamiento de los ciudadanos de los partidos políticos, que tienen repercusiones relevantes particularmente cuando existen elecciones críticas. En un periodo de deslizamiento, los patrones de votación establecidos se vuelen volátiles y se pierden de vista los grupos de apoyo partidario ya previamente identificados, debido a que se hacen mas difusos.

Los electores dejan de mantener vínculos partidarios que los caracterizaban en elecciones pasadas, o tiene la capacidad de modificarlos rápidamente. Incluso puede haber momentos en que entran nuevos grupos de electores sin identidad partidaria o con identidades distintas al padrón, mismos que por su número tiene la capacidad de influir las tendencias de votación, o de no votación, enriqueciéndolas.

Un deslizamiento partidario no necesariamente trae consigo un realineamiento electoral, pues los patrones de identificación partidaria se mantienen, aunque débiles y muy volátiles. Esta es la principal diferencia que existe con el realineamiento electoral además del hecho de que las personas dejen de apoyar a un partido, no implica que apoyen a otro. En estos periodos no únicamente crece la incertidumbre de los participantes, sino principalmente la identificación de los electores y la abstención electoral.

Una de las líneas explicativas más aceptadas con respecto a las causas de un deslizamiento, es en el sentido de la pérdida de las líneas que generaban

patrones de identificación, como condiciones de clase, ideología, o condiciones de vida cambiantes.

Los realineamientos y deslizamientos son fenómenos aunados pero distintos, y aunque no son auto excluyentes entre sí, uno no es sustituto del otro.

Capitulo segundo
Modelo de Regresión Logística

2.1 Introducción

La regresión logística es una técnica multivariante que nos permite estimar la relación existente entre una variable dependiente cualitativa, en particular dicotómica y un conjunto de variables independientes cuantitativas.

El análisis de Regresión Logística tiene la misma estrategia que el Análisis de Regresión Múltiple, el cual se diferencia esencialmente del Análisis de Regresión Logística porque la variable dependiente se considera de tipo binaria. Es decir, que el conjunto de datos consiste de una muestra de tamaño $n = n_1 + n_2$, donde n_1 observaciones son de una clase $c_1\left(y = 1\right)$ y n_2 son de clase $c_2\left(y = 0\right)$.

Esto es, la respuesta toma únicamente dos valores: 1 presencia (con probabilidad p) y 0 ausencia (con probabilidad (1-p)) y las variables explicativas pueden ser cuantitativas o cualitativas; donde la ecuación del modelo no es una función lineal, sino exponencial y con una transformación logarítmica puede hacerla una función lineal.

El objetivo de esta técnica es el modelar la influencia de las variables regresoras en la probabilidad de ocurrencia de un suceso particular.

Sistemáticamente tiene dos objetivos:

1. Investigar como influye en la probabilidad de ocurrencia de un suceso, la presencia o no de diversos factores y el valor o nivel de los mismos.
2. Determinar el modelo más parsimonioso y mejor ajustado que siendo razonable describa la relación entre la variable respuesta y un conjunto de variables regresoras.

Otro método que puede ser empelado es el análisis discriminante (AD) que permite la predicción de pertenencia de la unidad de análisis a uno de los dos grupos preestablecidos, pero se requiere que se cumplan los supuestos de multicolinealidad de las variables regresoras y la igualdad de las matrices de covarianzas de los dos grupos, pueden ser diferentes también para que la regla de predicción sea la optima, Jonson (1982).

La regresión logística requiere mucho menos supuestos que el AD, por ello cuando satisfacen los supuestos requeridos para el AD, la Regresión Logística trabaja bien.

En cambio en el análisis de regresión lineal múltiple cuando la variable respuesta toma solo dos valores, se violan los supuestos necesarios para efectuar inferencias, estos son:

1. La distribución de los errores no es normal.
2. Los valores predictores no pueden ser interpretados como probabilidades, como en la Regresión Logística, ya que no toman valores dentro del intervalo [0,1].

La diferencia básica entre el Modelo de Regresión Lineal Múltiple y la Regresión Logística es la naturaleza entre la variable respuesta y las variables regresoras.

Para el análisis de Regresión Lineal Múltiple, consideremos *y* una variable respuesta cuantitativa y $X_1,\ldots,X_k$ variables regresoras (explicativas); se desea describir la relación que hay entre la variable respuesta y las variables explicativas se espera que:

$$E\left\langle y_i \mid x_1, x_2, ..., x_k \right\rangle = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}; \quad i = 1, 2, ..., n \quad (2.1)$$

Donde:

y_i Es el valor de la variable respuesta cuantitativa para el *i*-ésimo objeto.
$\beta_{n:n=0,1,2,...,k}$ Son los parámetros.

Siendo *n* el número de objetos u observaciones.

Aunque (2.1) no da valores exactos, se espera que varíe linealmente con las variables regresoras, esto es:

$$E\left\langle y_i \mid x_i \right\rangle = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \quad (2.2)$$
$$i = 1, 2, ..., n$$

Siendo $x_i^T = \left(x_{i0}, x_{i1}, x_{i2}, ..., x_{in}\right)$ la *i*-ésima observación, con $x_{i0} = 1$ (2.2) toma valores reales y en forma vectorial es:

$$E\left\langle y_i \mid x_i \right\rangle = x_i^T \beta \quad (2.3)$$

$\beta^T = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ Es el vector de parámetros.

Pero en (2.3) hay otras variables regresoras que pueden influir linealmente sobre y_i, por tanto cada valor de y_i esta variando alrededor de $\mathrm{E}(y_i)$ a esa variación lo denotamos con ε_i, esto es:

$$\begin{aligned} \varepsilon_i &= y_i - E\langle y_i \,|\, x_i \rangle \\ &= y_i - x_i^T \beta \end{aligned} \quad (2.4)$$

De (2.4):

$$y_i = x_i^T \beta + \varepsilon_i \quad (2.5)$$

$$\begin{aligned} &y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_\kappa x_{ik} + \varepsilon_i \\ &i = 1, 2, ..., n \end{aligned} \quad (2.6)$$

ε_i es llamado error aleatorio, que cumple con las siguientes propiedades:

$$\begin{aligned} &E(\varepsilon_i) = 0 \\ &V(\varepsilon_i) = \sigma^2 \\ &Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j \\ &Cov(\varepsilon_i, x_j) = 0 \end{aligned} \quad (2.7)$$

El Modelo de Regresión Lineal Múltiple se generaliza de (2.6), dada por el álgebra matricial:

$$y = X\beta + \varepsilon \quad (2.8)$$

Donde:

$y = (y_1, y_2, ..., y_n)$, es el vector de variables respuesta
$X = (1, x_1, \ldots, x_k)$ matriz de rango completo y con
$x_i^T = (1, x_{i1}, x_{i2}, ..., x_{ik})$
$\beta^T = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$
$\varepsilon^T = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$

y de forma matricial (2.1) es:

$$E(y) = X\beta \quad (2.9)$$

Con el objeto de calcular los parámetros del modelo (2.6), mediante el método de mínimos cuadrados, bajo el supuesto que $\varepsilon \sim N(0, \sigma^2 I_n)$, las observaciones $y_1, y_2, ..., y_n$ son independientes y distribuidas como una Normal n-variada con $E\left(\overset{p}{y}\right) = X\overset{p}{\beta}$ matriz de varianzas y covarianzas $\sigma^2 I_n$.

En la Regresión Logística, se calcula la probabilidad de que un evento acontezca, pues sus variables regresoras toman valores entre 0 y 1 por lo que se define como una probabilidad de que ocurra un evento o no sujeto a control. La relación entre las variables regresoras y la variable respuesta o dependiente no es lineal.

Así comenzaremos por definir la Regresión Logística Simple el cual tiene la forma

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ i &= 1, 2, ..., n \end{aligned} \quad (2.10)$$

si $y = 1, \varepsilon_i = 1 - \beta_0 - \beta_1 x_i$ (2.11)

si $y = 0, \varepsilon_i = -\beta_0 - \beta_1 x_i$ (2.12)

Por lo que ε_i, no tiene distribución normal pues toma valores discretos.

En el Análisis de Regresión Lineal Simple, se emplea un grafico de dispersión de la variable respuesta contra la regresora, pero resulta insuficiente con solo dos valores posibles para la variable respuesta, por lo que se utiliza otro grafico, suavizando los valores de la variable respuesta contra la variable regresora.

La notación que se usará en el presente trabajo para la Regresión Logística es el que emplea Hosmer & Lemeshow (2000).

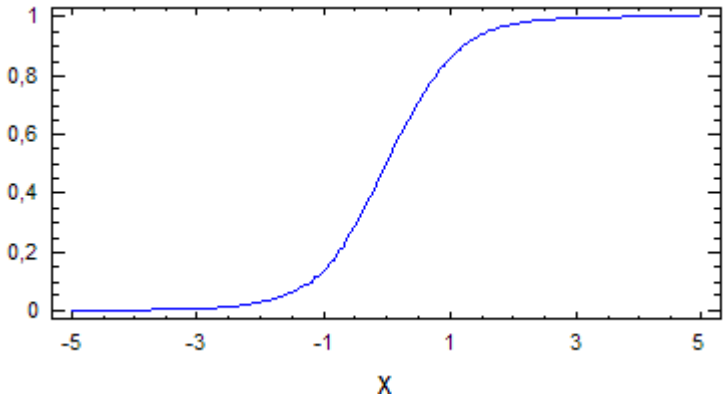
Sea

$$\pi(x) = E\langle y | x \rangle \quad (2.13)$$

Es la media condicional de $y = 1$ dado x, donde $\pi(x)$ representa la probabilidad de que ocurra $y = 1$, en la que existe una relación curvilínea en la que para un valor grande de x, $\pi(x)$ tomará valores cercanos a 1 y para valores pequeños de x, $\pi(x)$ tomara valores cercanos a cero.

El grafico muestra el comportamiento de $\pi(x)$ contra x:

FUGURA 2.1



Curva en forma de S o sigmoideo que tiene las propiedades requeridas para $\pi(x)$ y cumple con propiedades de una función de distribución de probabilidad acumulada:

$(i)\quad F_x(-\infty) \equiv \lim_{x\to-\infty} F_x(x) = 0, y\, F_x(+\infty) \equiv \lim_{x\to+\infty} F_x(x) = 1$

$(ii)\quad F_x(\cdot)$ es una función monótona no decreciente, tal que, $F_x(a) \le F_x(b)$ *donde* $a < b.$

$(iii)\quad F_x(\cdot)$ es continua por la derecha, tal que, $\lim_{0<h\to\infty} F_x(x+h) = F_x(x).$

Para esta se emplea la función de distribución acumulada de la distribución logística dada por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (2.14)$$

(2.14) tiene un grafico similar a la Figura 2.1, cuando $\beta_0 < 0$ y $\beta_1 > 0,$ en el intervalo $[0,1]$.

Si $P[y=1] = 0.5$ el valor de x es: $\frac{-\beta_0}{\beta_1}$

La transformación logit de $\pi(x)$, es:

$$g(x) = Ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \quad (2.15)$$
$$= \beta_0 + \beta_1 x$$

Esta transformación es importante pues tiene propiedades semejantes al Modelo de Regresión lineal simple, dado que es lineal es sus parámetros, puede ser continua y toma valores reales dependiendo del valor de x.

Para el Modelo de Regresión Lineal simple, la variable respuesta, de (2.4) se expresa como:

$$y = E\langle y | x \rangle + \varepsilon \quad (2.16)$$

para la variable respuesta dicotómica lo expresamos como:

$$y = \pi(x) + \varepsilon \quad (2.17)$$

Si $y = 1, \quad \varepsilon_i = 1 - \pi(x)$ y tiene probabilidad $\pi(x)$

Si $y = 0, \quad \varepsilon_i = -\pi(x)$ y tiene probabilidad $1 - \pi(x)$

Entonces ε_i tiene distribución binomial con media cero y varianza $\pi(x)\left[1 - \pi(x)\right]$. Entonces la distribución condicional de la variable respuesta tiene distribución de probabilidad binomial con media $\pi(x)$.

El lado izquierdo de (2.15) se llama logaritmo del consiente de momios (ODDS RADIO) o razón de probabilidades de $y = 1$ contra $y = 0$ o llamada razón de ventaja a favor de éxito.

$$ODDS\ RATIO = \frac{\pi(x)}{1 - \pi(x)} \quad (2.18)$$

2.2 Regresión Logística Múltiple

El Modelo de Regresión Logística Múltiple es una generalización del Modelo de Regresión Simple, pues considera más de una variable regresora, en donde por lo menos una es cuantitativa.

La consideración central del modelo de regresión logística múltiple, será la estimación de los coeficientes en el modelo y probar si estas son significativas.

2.3 Modelo de Regresión Logística Múltiple

Se considera una colección de k variables independientes denotado por el vector de variables regresoras $x^T = (x_1, x_2, ..., x_k)$, donde se asume por el momento que cada una de estas están medidas por lo menos bajo escala intervalar. Sea la probabilidad condicional para que la variable respuesta sea igual a 1, denotado por:

$$P\langle y = 1 | x \rangle = \pi(x) \quad (2.19)$$

el logaritmo del Modelo de Regresión Logística Múltiple es:

$$g(x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad (2.20)$$
$$i = 1, 2, ..., n$$

entonces el Modelo de Regresión Logística Múltiple es:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (2.21)$$

En la que al igual que la Regresión Lineal Múltiple son necesarias las variables categóricas, si tiene c niveles será necesario incorporar c-1 variables facticias o llamadas dummy; así el logit para un modelo con k variables regresoras y una variable j-ésima categórica es:

$$g(x) = \beta_0 + \beta_1 x_{i1} + ... + \sum_{I=1}^{c-1} \beta_{jl} D_{jl} + \beta_k x_{ik} \quad i = 1, ..., n \quad (2.22)$$

2.4 Ajuste del Modelo de Regresión Logística Múltiple

El ajuste se efectúa a través de los métodos de máxima verosimilitud (ML), el cual es una de las varias alternativas de aproximación que estadísticamente se han desarrollado para la estimación de parámetros de un modelo matemático.

En el caso de la regresión logística es empleado el método de Máxima Verosimilitud, ya que los parámetros estimados son los que maximizan la probabilidad de obtener un conjunto de datos observados.

Existen dos alternativas para la aproximación de ML que pueden ser usados en la estimación de los parámetros en el modelo logístico. Estos son el método incondicional o el condicional. Estos dos métodos requieren de diferentes programas los cuales pueden revisarse en el anexo de método condicional e incondicional.

La elección entre ambos métodos depende del numero de de parámetros en un modelo, contra el total de observaciones. En general el método incondicional de ML es preferible si el número de parámetros en el modelo es relativamente más pequeño que el numero de observaciones, por el contrario, si el numero de observaciones es mayor al número de parámetros es empleado el método condicional de ML, para mayor información del tema revisar [Kleimbaum].

Asumiremos una muestra de n observaciones independientes

$$\left(\overset{p}{x_i}, y_i \right), i = 1, 2, ..., n \,; \left(x_i, y_i \right) \quad i = 1, 2, ..., n;$$

Donde y_i toma valores 0 ó 1, para estimar $\beta^T = \left(\beta_0, \beta_1, ..., \beta_k \right)$ que es el vector de parámetros desconocidos.

La función de verosimilitud expresa la probabilidad de los datos observados como una función de parámetros desconocidos. Los estimadores de Máxima Verosimilitud de esos parámetros son aquellos que están en concordancia con los datos observados.

Para la estimación del vector β se requiere hallar el máximo de una función; para lo cual es empleada la primera derivada, esta es llamada función de Score:

$s(\beta) \equiv \frac{\partial}{\partial \beta} \log L(\beta)$ donde la estimación del máximo verosímil de $\hat{\beta}$ es la solución a la función de Score $s(\beta) = 0$ porque se anula en el punto máximo.

La segunda derivada se emplea para calcular las tangentes. En el máximo, la segunda derivada del $\log L(\beta)$ es negativa, esta se define como $I(\beta) \equiv -\frac{\partial^2}{\partial \beta^2} losL(\beta)$, llamada la función de información de Fisher.

Es necesario encontrar la segunda derivada para obtener la matriz de varianzas y covarianzas de los parámetros estimados.

Tomamos un Modelo de Regresión Lineal Múltiple, supongamos que se disponen de n observaciones donde para cada una de ellas existe una respuesta que puede ser.

$y_i = 0 \, o \, y_i = 1$

Sea $\vec{y}^T = (y_1, y_2, ..., y_n)$ donde $y_i \square B(1, \pi_i)$ y sea

$x_i^T = (1, x_{i1}, ..., x_{ik})$ la *i*-ésima observación para las K variables explicativas.

El Modelo de Regresión Logística está dado por la expresión (2.20):

$$P\langle y_i | x_i \rangle = \pi(x_i) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}} \quad (2.23)$$

equivalente

$$P\langle y_i | x_i \rangle = \frac{Exp\left(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij}\right)}{1 + Exp\left(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij}\right)} \quad (2.24)$$

y la probabilidad de que y_i sea igual a cero es:

$P\langle y_i = 0 | x_i \rangle = 1 - P\langle y_i = 0 | x_i \rangle$

Entonces:

$$P\langle y_i = 0 | x_i \rangle = \frac{1}{1 + Exp\left(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij}\right)} \quad (2.25)$$

para facilitar la notación se emplea la variable indicadora
$x_{i0} = 1, i = 1, 2, ..., n.$

Por lo que (2.24) y (2.25) son respectivamente:

$$P\left\langle y_i = 1 \middle| x_i \right\rangle = \pi\left(x_i\right) = \frac{e^{\beta x_i^T}}{1 + e^{\beta x_i^T}} \quad (2.26)$$

$$P\left\langle y_i = 1 \middle| x_i \right\rangle = \pi\left(x_i\right) = \frac{1}{1 + e^{\beta x_i^T}} \quad (2.27)$$

donde $x_i^T = \left(x_{i0}, x_{i1}, ..., x_{ik}\right)$, es el vector que contiene los valores de las variables explicativas

$\beta^T = \beta_0, \beta_1, ..., \beta_k$ es el vector de parámetros a ser estimado.

El i-ésimo logito es:

$$\lambda_i = Ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=0}^{k} \beta_j x_{ij} \quad (2.28)$$

por lo que (2.28) es una función lineal simple del vector de observaciones x_i llama transformación logística de la probabilidad π_i o simplemente Logia o Logito de la ecuación. A la expresión (2.28) también se le llama Modelo Logístico Lineal.

Para obtener la estimación máximo verosímil para el vector β, se escribe la función de densidad de probabilidad del vector y el cual es proporcional a n funciones $B\left(1, \pi_i\right)$, esto es:

$$\begin{aligned} f\left(y_i; \pi_i\right) &= \prod_{i=1}^{n} \pi_i^{y_i} \left(1 - \pi_i\right)^{1 - y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} \left(1 - \pi_i\right) \\ &= \left\{\prod_{i=1}^{n} \left(1 - \pi_i\right)\right\} \left\{\prod_{i=1}^{n} Exp\left[Ln\left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i}\right]\right\} \\ &= \left\{\prod_{i=1}^{n} \left(1 - \pi_i\right)\right\} Exp\left[\sum_{i=1}^{n} y_i Ln\left(\frac{\pi_i}{1 - \pi_i}\right)\right] \end{aligned} \quad (2.29)$$

Reemplazando (2.28) en (2.29), se obtiene:

$$f\left(y_i;\pi_i\right)=\left\{\prod_{i=1}^{n}\left(1-\pi_i\right)\right\}Exp\left\{\sum_{i=1}^{n}y_i\sum_{j=0}^{n}\beta_i x_{ij}\right\}$$
$$=\left\{\prod_{i=1}^{n}\left(1-\pi_i\right)\right\}Exp\left\{\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)\beta_j\right\} \quad (2.30)$$

El logaritmo natural de la función (2.30), llamado función soporte es:

$$l(\pi_i,y_i)=\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)\beta_j-\sum_{i=1}^{n}Ln\left[1-\pi_i\right] \quad (2.31)$$

Pero (2.27):

$$1-\pi_i=\frac{1}{1+Exp\left(\beta^T x_i\right)}$$

$$=\left[1+Exp\left(\beta^T x_i\right)\right]^{-1}$$

entonces $Ln\left(1-\pi_i\right)=-Ln\left[1+Exp\left(\beta^T x_i\right)\right]$

$$Ln\left(1-\pi_i\right)=-Ln\left[1+Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right] \quad (2.32)$$

reemplazando (2.32) en (2.31), se obtiene:

$$l\left(\pi_i;y_i\right)=\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)\beta_j-\sum_{i=1}^{n}Ln\left[1+Exp\sum_{j=0}^{k}\beta_j x_{ij}\right] \quad (2.33)$$

ahora (2.33) es una función que ya no depende de π_i sino de β_j solamente, entonces lo denotamos como:

$$L\left(\beta\right)=\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)\beta_j-\sum_{i=1}^{n}Ln\left[1+Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right] \quad (2.34)$$

Es una función que depende exclusivamente del vector $\overset{p}{\beta}$.

Definimos como:

$$t_j=\sum_{i=1}^{n}y_i x_{ij} \quad (2.35)$$

reemplazando (2.35) en (2.34) se tiene:

$$L\left(\beta\right)=\sum_{j=0}^{k}\beta_j t_j-\sum_{i=1}^{n}Ln\left[1+Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right] \quad (2.36)$$

(2.36) es una función exclusiva del vector de parámetros β, por el Teorema de Factorización de Fisher-Neyman, Bickel y Dukson (1978), se tiene que t_j para $j = 0,1,...,k$ son estadísticas suficientes para los parámetros β_j, para $j = 0,1,...,k$.

La variable aleatoria t_j dada en la expresión (2.36) es la suma de algunos de los términos de la matriz de diseño X, es decir se incluyen en la suma solamente los elementos que corresponden a una respuesta del tipo $y = 1$.

$$\frac{\delta L}{\delta \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \left[\frac{Exp\left(\sum_{j=0}^{k} \beta_j x_{ij} \right)}{1 + Exp\left(\sum_{j=0}^{k} \beta_j x_{ij} \right)} \right] \quad (2.37)$$

las ecuaciones de verosimilitud de (2.37) son:

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \hat{\pi}_i = 0 \quad j = 0,1,2,...,k \quad (2.38)$$

siendo $x_{i0} = 1$, equivalente (2.38) es:

$$\sum_{i=1}^{n} x_{ij} (y_i - \hat{\pi}_i) = 0 \qquad j = 0,1,2,...,k \quad (2.39)$$

donde:

$$\hat{\pi}_i = \frac{Exp\left(\sum_{j=0}^{k} \beta_j x_{ij} \right)}{1 + Exp\left(\sum_{j=0}^{k} \beta_j x_{ij} \right)} \text{ para } i = 1,2,...,n$$

Es el estimador máximo verosímil de π_i y se obtiene mediante $\hat{\beta}_j$ y el vector $\overset{p}{x}_i$.

La expresión (2.38) en su forma matricial es:

$$X^T \left(y - \pi \right) = XS = 0 \quad (2.40)$$

Estas ecuaciones son parecidas a las ecuaciones normales obtenidas para estimar el Modelo de Regresión Lineal Múltiple, pero son no lineales en β, por

lo que se emplean métodos iterativos para determinar los valores del vector β, los cuales se trataran con detenimiento mas adelante.

Se calculara la varianza y covarianza de β.

Sea $X_{(n\times p)}$ la matriz de diseño, con *p= k+1*, con elementos:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12}\cdots & x_{1k} \\ 1 & x_{21} & x_{22}\cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{2n}\cdots & x_{nk} \end{bmatrix}$$

Las ecuaciones de verosimilitud en su forma matricial, de la expresión (2.40)

$$X^T y = X^T \pi \text{ donde } \pi^T = \left(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n\right) \text{ (2.41)}$$

$$\sum_{i=1}^{n} x_{ij}\left(y_i - \pi_i\right) = 0 \text{ (2.42)}$$

El método de estimación de las varianzas y covarianzas se obtiene de la matriz de segunda derivada parcial de (2.42), esta es importante debido a que la información contenida, es usada en los cálculos requeridos para las pruebas de hipótesis y la estimación de los intervalos de confianza.

Esta tiene la forma:

$$\frac{\delta^2 L}{\delta \beta_j^{\,2}} = -\sum_{i=1}^{n} x_{ij}^{\,2}\pi_i\left(1-\pi_i\right) \text{ para } j = 0,1,2,\ldots,k \text{ (2.43)}$$

reemplazando la ecuación para π_i en (2.43)

$$\frac{\delta^2 L}{\delta \beta_j^{\,2}} = -\sum_{i=1}^{n} \frac{x_{ij}^{\,2} Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)}{\left[1+Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right]^2} \text{ para } j = 0,1,2,\ldots,k \text{ (2.44)}$$

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il}\pi_i\left(1-\pi_i\right) \text{ para } j = 0,1,2,\ldots,k \text{ (2.45)}$$

reemplazando:

$$\frac{\partial^2 L}{\partial\beta_j\partial\beta_l} = -\sum_{i=1}^{n} x_{ij}x_{il}\frac{Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)}{\left[1+Exp\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)\right]} \quad (2.46)$$

Tanto (2.44) como (2.45) no son funciones que dependen de y_i, entonces la matriz de observación y la matriz de segunda derivada esperada son idénticas.

La Matriz de información se denota con $I\left(\overset{p}{\beta}\right)$, que es la que contiene el negativo de las ecuaciones (2.44) y (2.46); las varianzas y covarianzas de $\hat{\beta}_j$ se obtienen tomando la inversa de esta matriz, esto es:

$$Cov\left(\beta\right) = I^{-1}\left(\beta\right) \quad (2.47)$$

Los estimadores de la varianza y covarianza, denotada por $Cov(\beta)$, se obtiene evaluando $Cov(\beta)$ en β.

Entonces la matriz de información estimada, matricialmente tiene la forma:

$$\hat{I}\left(\beta\right) = X'VX \quad (2.48)$$

V es una matriz diagonal, esto es:

$V = Diag\left[\hat{\pi}_i\left(1-\hat{\pi}_i\right)\right]$

de tamaño $n\times n$, además (2.48) es:

$$\hat{C}ov\left(\beta\right) = \left(X'VX\right)^{-1} \quad (2.49)$$

y es de tamaño $(k+1)(k+1)$

escribiremos los elementos de la matriz (2.49)

$$\widehat{Cov}\left(\overset{p}{\beta}\right)=\begin{bmatrix}\hat{\sigma}^2\left(\beta_0\right) & \hat{\sigma}\left(\beta_0,\beta_1\right)\cdots\cdots & \hat{\sigma}\left(\beta_0,\beta_k\right)\\ \vdots & \hat{\sigma}^2\left(\beta_1\right)\cdots\cdots & \hat{\sigma}\left(\beta_1,\beta_k\right)\\ \vdots & \vdots & \vdots\\ \vdots & & \hat{\sigma}^2\left(\beta_k\right)\end{bmatrix}$$

donde:

$\hat{\sigma}^2\left(\beta_j\right)$ es la varianza estimada de β_j

$\hat{\sigma}\left(\beta_j,\beta_l\right)$ es la covarianza estimada de β_j y β_l

$\hat{\sigma}\left(\beta_j\right)$ es el error estándar de β_j

2.5 Método de Newton-Raphson para estimar los parámetros del modelo de Regresión Logística

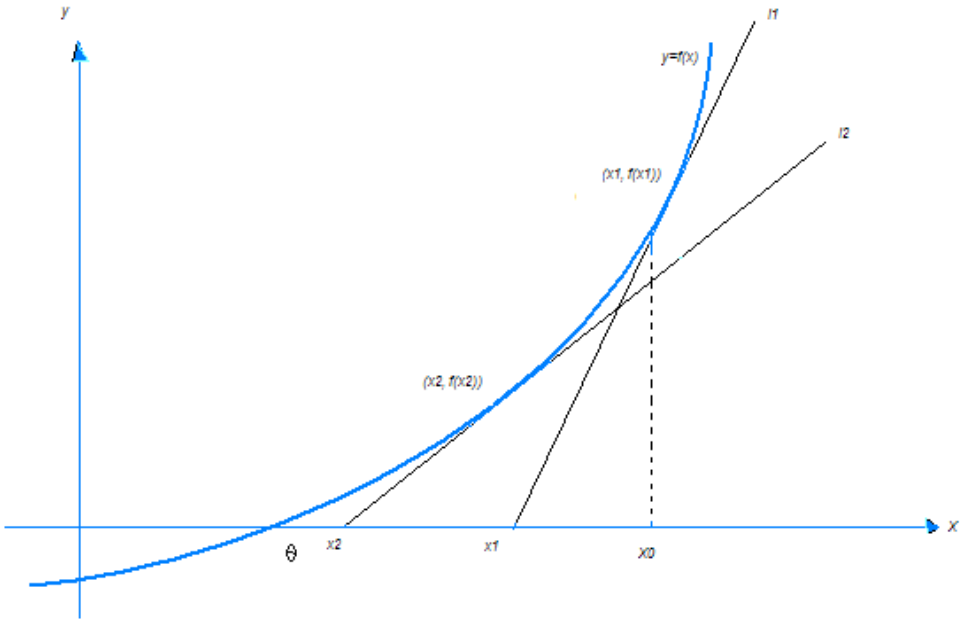
El método de Newton-Rapshon surgió de la formula de aproximación de Taylor $f(x)\approx f(a)+f'(a)(x-a)$, cuando x es cercana a a.

Es un método para resolver ecuaciones no lineales, como las obtenidas en (2.37) o en (2.38), y requieren una solución mediante métodos iterativos para hallar la estimación de los parámetros que es el máximo de la función (2.34).

Se considera la línea tangente l_1 a la grafica $y=f(x)$ en el punto inicial $(x_0, f(x_0))$. Si x_0 es suficientemente cercano a θ, entonces la intersección al eje de las abscisas llamada x_1 de la línea l_1 puede ser cercana a θ, donde l_1 puede ser expresada como $y=f(x_0)+f'(x_0)(x-x_0)$.

Así, x_1 de la línea l_1 puede satisfacer $\theta=f(x_0)+f'(x_0)(x_1-x_0)$ o de forma equivalente $x_1=x_0-\dfrac{f(x_0)}{f'(x_0)}$ si $f'(x_0)\neq 0$. Si se tiene una nueva línea l_2, con una mejor aproximación se repite el proceso, hasta que la diferencia entre aproximaciones consecutivas sea nula o cercana a cero. En la figura 2.2 se ilustra el método.

Figura 2.2 Interpretación Geométrica del Método de Newton Raspón



Se usa el siguiente esquema iterativo:

$$\beta^{(t+1)} = \beta^{(t)} + \left[I\left(\beta^{(t)} \right) \right]^{-1} s\left(\beta^{(t)} \right) \quad (2.50)$$

donde:

$S(\beta)$ y $I(\beta)$ son las funciones de Score y de Información respectivamente.

La función de Score es un vector de tamaño $k+1$, donde el $j-ésimo$ elemento de acuerdo a (2.37) es:

$$\frac{\delta L}{\delta \beta_j} = \sum_{i=1}^{n} \left(y_i - \pi_i^{(t)} \right) x_{ij} \quad (2.51)$$

La cual es similar a la expresión (2.39):

$$\sum_i x_{ij}\left(y_i - \pi_i\right) = 0 \quad j = 1, 2, ..., k$$

La función información es una matriz de tamaño $(k+1)(k+1)$ donde el $i-jésimo$ elemento $(l.j)$ es:

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_l} = -\frac{\delta}{\delta \beta_l}\left[\sum_{i=1}^{n} x_{ij}\left(y_i - \pi_i\right)\right]$$

$$= -\frac{\delta}{\delta \beta_l}\left[\sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right]$$

$$= \sum_{i=1}^{n} x_{ij}\left[\frac{e^{\beta^T x_i} x_{il}\left(1+e^{\beta^T x_i}\right) - e^{\beta^T x_i} x_{il} e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2}\right]$$

$$= \sum_{i=1}^{n} \frac{x_{ij} x_{il} e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2}$$

$$= \sum_{i=1}^{n} x_{ij} x_{il} \pi_i \left(1-\pi_i\right) \quad j = 0, 1, ..., k; l = 0, 1, ..., k \tag{2.52}$$

donde $\pi^{(t)}$, es la *t-ésima* aproximación para π, obtenida de $\beta^{(t)}$ mediante:

$$\pi_i^{(t)} = \frac{Exp\left(\sum_{j=0}^{k} \beta_j^{(t)} x_{ij}\right)}{\left[1+Exp\left(\sum_{j=0}^{k} \beta_j^{(t)} x_{ij}\right)\right]} \tag{2.53}$$

Entonces el próximo valor reemplazando en (2.50) es:

$$\beta^{(t+1)} = \beta^{(t)} - \left\{X^T V^{(t)} X\right\}^{-1} X^T \left(y - \pi^{(t)}\right) \tag{2.54}$$

donde $V^{(t)} = Diag\left[\pi_i^{(t)}\left(1-\pi_i^{(t)}\right)\right]$

La expresión (2.50) se usa para obtener $\pi^{p(t+1)}$ y así sucesivamente.
Después de dar un valor inicial $\beta^{(0)}$, se usa (2.50) para obtener $\pi^{(0)}$ y para $t > 0$ las iteraciones siguientes se efectúan usando (2.50) y (2.51).

En el limite, $\pi^{(t)}$ y $\beta^{(t)}$ converge a los estimadores de máxima verosimilitud π y converge en general en 5 o 6 iteraciones.

El software estadístico de SAS 9.0 emplea el método antes descrito, en el que emplea como máximo de iteraciones 25 para el proceso LOGIST que emplearemos para calcular el modelo y hasta 50 para el proceso GENMOD.

2.6 Interpretación de los coeficientes del modelo estimado

En el modelo de regresión múltiple el valor de un coeficiente significa el cambio en unidades de la variable dependiente por cada unidad de la variable independiente a que se refiere el coeficiente, permaneciendo invariantes los valores del resto de variables independientes del modelo.

A nivel de coeficientes estimados exponencialmente la interpretación es similar, en donde la diferencia estriba en que en la Regresión Logística no es el cambio (Incremento o disminución) de la probabilidad de la variable dependiente por cada unidad de cambio en las independientes, sino del incremento o disminución que se produce en el cociente entre $P(Y=1)/P(Y=0)$

Expresado por:

$$\frac{P(Y=1)}{P(Y=0)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k} \quad (2.55)$$

Más aún, están expresados en logaritmos, por lo que será necesario transformarlos (tomando los valores del antilogaritmo) de tal forma que se evalúe más fácilmente su efecto sobre la probabilidad. Esto lo hacen automáticamente los programas calculando tanto el coeficiente real como el transformado. Utilizar este procedimiento no cambia la forma de interpretar el signo del coeficiente. Un coeficiente positivo aumenta la probabilidad, mientras que un coeficiente negativo disminuye la probabilidad, Así si, β es positivo, su transformación (antilog) será mayor a uno, y el odds ratio aumentará. Este aumento se produce cuando la probabilidad prevista de ocurrencia de un suceso aumenta y la probabilidad prevista de su no ocurrencia disminuye. De la misma forma, si β es negativo, el antilogaritmo es menor que 1 y el odds ratio disminuye. Un valor de cero equivale a un valor de 1, lo que no produce cambio en el odds.

2.7 Prueba de hipótesis para los coeficientes del modelo de Regresión Logística

En el Modelo de Regresión logística se efectúan pruebas con objetivos diferentes, siendo, los cuales son:

1. Determinar si una variable explicativa tiene coeficiente igual a cero.
2. Determinar si un conjunto de variables explicativas tiene coeficientes igual a cero.
3. Determinar la calidad del ajuste global del modelo.

2.7.1 Prueba de Wald

Wald (1943) estudio una prueba asintótica para estimaciones máximos verosímiles, y asevero que los parámetros estimados en los modelos logísticos tienen una distribución normal para muestras grandes.

Esta prueba se usa para evaluar si cada variable explicativa o regresora tiene coeficiente igual a cero.

Sea $\pi^{(t)}$ que converge a los estimadores de máxima verosimilitud de π y $y_1, y_2, ..., y_n$ variables respuesta binaria independientes cuyas probabilidades satisfacen.

$$Logit\left(\pi_i\right) = x^T \beta$$

Donde $\pi_i = P\left[y_i = 1 \middle/ x_i \right]$

Siendo x_i una observación que contiene los valores de las k variables explicativas con $x_i^T = \left(1, x_{i1}, x_{i2}, ..., x_{ik}\right)$.

Sin perdida de generalidad, seleccionamos β_j como el parámetro de interés. Supóngase que las hipótesis son:

$$H_0 : \beta_j = \beta_{j0}$$
$$H_1 : \beta_j \neq \beta_{j0} \quad (2.56)$$

sea $\hat{\beta}_j$ un EMV de β_j y sea:

$I^{-1} = \left(X^T V X \right)^{-1}$ la inversa de la matriz de información muestral, entonces la estadística de Wald para dócimar (2.75) es:

$$W = \frac{\left(\hat{\beta}_j - \beta_{j0}\right)^2}{\sigma^2\left(\beta_j\right)} \quad (2.57)$$

donde $\sigma\left(\hat{\beta}_j\right)$ es la estimación del error estándar de β_j.

Bajo $H_0, W \sim \chi^2_{(1)}$ y para n suficientemente grande se tiene que:

$$z = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\left(\hat{\beta}_j\right)} \sim N\left(\left(\frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\left(\hat{\beta}_j\right)}\right),1\right) \quad (2.58)$$

por tanto $z^2 \sim \chi^2_{(\xi,1)}$ es χ^2 con parámetro de no centralización:

$$\xi = \frac{\left(\hat{\beta}_j - \beta_{j0}\right)^2}{\hat{\sigma}\left(\hat{\beta}_j\right)} \quad (2.59)$$

Pero la estadística W, tiene la probabilidad que cuando el valor absoluto del coeficiente de Regresión es grande, el error estándar también lo es; esta situación hace que la estadística W sea pequeña y por tanto se puede rechazar β_j igual a cero, cuando en realidad no debería rechazarse.

Por tanto, cuando se encuentra que un coeficiente es grande, es preferible no usar la estadística de Wald para efectuar dócima individual. Sino se recomienda construir un modelo con y sin esa variable y basarse en la prueba de hipótesis de la diferencia entre los dos modelos.

Para las hipótesis estadísticas:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \quad (2.60)$$

La estadística (2.57) es:

$$W = \frac{\left(\hat{\beta}_j\right)^2}{\sigma^2\left(\hat{\beta}_j\right)} \quad (2.61)$$

Bajo $H_0, W \sim \chi^2_{(1)}$ y para n suficientemente grande se tiene que:

$$z = \frac{\beta_j}{\sigma(\beta_j)} \sim N\left(\frac{\beta_j}{\sigma(\beta_j)}, 1\right) \quad (2.62)$$

por tanto:

$$\xi = \frac{(\beta_j)^2}{\sigma^2(\beta_j)} \quad (2.63)$$

si la variable explicativa es categórica, los grados de libertad es igual al número de categorías o niveles de la variable menos uno.

2.7.2 Prueba Chi-cuadrado

Sirve para docimar los coeficientes del modelo logístico. Para elegir un modelo, se usa la prueba de razón de verosimilitud, Bickel y Docksum (1977), para probar la hipótesis de que los coeficientes β_j correspondientes a las variables explicativas retiradas, digamos q variables explicativas, del modelo son iguales a cero, siendo la hipótesis estadística:

$H_0 : \beta_j = \beta_2 = ... = \beta_q = 0$

$H_1 : \beta_j \neq 0$

para por lo menos un $j = 1, 2, ..., q$.

Esta prueba se basa en la siguiente estadística:

$$\chi_q^2 = -2\left[LnL_{p-q} - LnL_p \right] \quad (2.64)$$

Bajo la hipótesis de que los coeficientes de las variables retiradas son iguales a cero, la estadística (2.84) tiene la distribución asintótica $\chi^2_{(q)}$.

Valores altos para esta estadística, indican que una o más de las q variables retiradas tienen coeficientes de regresión distintos de cero.

La estadística χ_q^2 se usa también para probar si una variable explicativa determina, x_k muestra una asociación significativa (como factor de riesgo) para con la variable respuesta en la presencia de las demás variables $x_1, x_2, ..., x_{k-1}$.

2.7.3 Estadística Chi-cuadrada de Pearson

Esta estadística sirve para lograr el último objetivo explicado anteriormente, es decir evaluar el modelo ajustado en forma global. La estadística se basa en la comprobación de los valores obtenidos y_i; y sus respectivas probabilidades estimadas π_i.

Las hipótesis estadísticas para usar esta hipótesis estadística son:

$$H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$$
$$H_1 : \beta_j \neq 0$$

para por los menos un conjunto $j = 0,1,2,...,k$.

Esta prueba se basa en la estadística Chi-cuadrada de Pearson, que está dada por:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\pi}_i\right)^2}{\hat{\pi}_i\left(1-\hat{\pi}_i\right)} \quad (2.65)$$

o equivalentemente $\chi^2 = \sum_{i=1}^{n} \frac{r_i^2}{v_{ii}}$ (2.66)

donde:

$$r_i = \left(y_i - \hat{\pi}_i\right)$$
$$v_{ii} = Diag\left(\hat{V}\right) = \hat{\pi}_i\left(1-\hat{\pi}_i\right)$$

Bajo la hipótesis nula de que el modelo se ajusta a los valores observados, la estadística (2.66) tiene distribución asintótica Chi-cuadrada $\chi^2_{(n-(k+1))}$.

Valores altos de la estadística Chi-cuadrada de Pearson indican discrepancias con el modelo teórico.

La estadística (2.66) es inestable cuando $\hat{\pi}_i$ toma valores cercanos a cero o uno.

2.7.4 Estadística Chi-cuadrada de Desvianza

Otra forma de probar el ajuste global del modelo, es mediante la estadística llamada Desvianza, propuesta por Nelder y Wederburn (1982). Es análogo a la suma de cuadrados de los residuales del Modelo de Regresión Lineal Múltiple.

Las hipótesis estadísticas son:

$$D_p = \sum_{i=1}^{n} d_i^2 \quad (2.67)$$

Donde:

$$d_i = \begin{matrix} \sqrt{-2\log \hat{p}_i} & si\ y_i = 1 \\ \sqrt{-2\log\left(1-\hat{p}_i\right)} & si\ y_i = 0 \end{matrix} ;\ j = 1,2,...,n$$

La devianza bajo la hipótesis nula, asintóticamente, es la misma que la distribución Chi-cuadrada de Pearson, es decir se distribuye $\chi^2_{(n-(k+1))}$ y mide la discrepancia o el desvió entre el modelo bajo investigación o actual y el modelo saturado.

La estadística (2.87) para el modelo de regresión logística esta dada por:

$$D = -2\sum\left(y_i \log\left(\hat{\pi}_i\right) + \left(1-y_i\right)\log\left(1-\hat{\pi}_i\right)\right) \quad (2.68)$$

Cuando el modelo bajo investigación es verdadero se compara el valor D con el valor crítico $\chi^2_{(n-p)}$ de una distribución χ^2 a un nivel de significación igual a α, por tanto:

Si $D > \chi^2_{(n-p)}$ el modelo se rechaza y

Si $D < \chi^2_{(n-p)}$ el modelo no se rechaza, donde $p = k+1$

2.7.5 Criterio AIC de Akaike

La bondad de ajuste de los parámetros de un modelo especifico, puede ser medida por el logaritmo del verosímil ($L\left(\hat{\beta}\right)$). Se introduce la media esperada de $L\left(\hat{\beta}\right)$ como una medida de bondad de ajuste del modelo. Esta cantidad es definida como la media, con respecto al dato x, del esperado $L\left(\hat{\beta}\right)$ del máximo verosímil del modelo. El máximo verosímil tiene una tendencia general para el estimador verdadero de la media esperada de $L\left(\hat{\beta}\right)$. Esta tendencia es más exacta para un número grande de parámetros. Si se eligen las medias con el mayor máximo de $L\left(\hat{\beta}\right)$, entonces probablemente no es necesario un número grande de parámetros.

Para cerrar la estimación de la relación entre el mejor estimador y el número de parámetros del modelo, se tiene que:

(Máximo verosímil del modelo)-(numero de parámetros)

Es un estimador imparcial de la media esperada de $L\left(\hat{\beta}\right)$.
Se define por:

$$AIC = D + 2\left(p + 1\right) \quad (2.69)$$

donde $D = -2\left(\text{máximo verosímil}\right)$ y donde p es el número de variables predictoras. Un modelo es mejor que otro si su AIC es más pequeño Sakamoto (1986).

2.7.6 Tabla de Clasificación

También llamada Matriz de Confusión, es una forma sencilla de evaluar el ajuste del Modelo de Regresión Logística, no es tan objetiva pero se usa como indicador de bondad de ajuste.
Es una tabla sencilla de 2×2, en el cual se muestra la distribución de los objetos que permite evaluar a las categorías 1 y 2, es decir cuando $y = 0$ y cuando $y = 1$, conjuntamente con la clasificación a cualquiera de las dos categorías de acuerdo a la probabilidad estimada.
Para interpretarla se hace mediante el porcentaje de objetos bien clasificados, esto es, aquellos que mediante la probabilidad estimada permanecen en su respectiva categoría. También se interpreta mediante la probabilidad de objetos mal clasificados, esto es, aquellos que mediante la probabilidad estimada se asignan a categorías deferentes del que fueron observados.

Figura 2.3 Tabla de clasificación

Grupo actual	Grupo estimado		Total marginal
	0	1	
0	n_{11}		$n_{11}+n_{12}$
1	n_{21}		$n_{21}+n_{22}$
Total marginal	$n_{11}+n_{21}$		n

Es el porcentaje de objetos bien clasificados mediante el Modelo de Regresión Logística estimado.

$$\frac{n_{11}+n_{22}}{n}\times 100\%$$

Por tanto, lo que se debe de esperar es que el porcentaje sea lo mas alto posible, a fin de que el modelo obtenido clasifique bien.

a) Índices de correlación por rangos

Describe la asociación entre los datos observados y los predichos en la tabla de clasificación. Estas medidas son apropiadas para variables ordinales, y ellas clasifican los pares de observaciones como concordantes o discordantes.

Un par es concordante si la observación con un valor grande de $\hat{y}$ también tiene un valor grande de y. Mientras un par discordante es si la observación con un valor grande de $\hat{y}$ tiene un valor pequeño de y.

Gamma de Goodman-Kruskal: es el cociente entre la diferencia del número de concordancias y el número de discordancias en el numerador, y la suma de concordancias y discordancias en el denominador. Este proporciona un valor que normalmente es superior al valor dado por otros índices.

Asume una escala ordinal, que la relación entre $\hat{y}$ e y es simétrica, y oscila entre 0 y 1. Un valor muy cercano a 1 indica una fuerte relación, mientras que los valores cercanos a 0 indica poca o nula relación, por lo que son independientes.

Gamma es estimada por $G=\left[(C-D)/(C+D)\right]$

Tau-a de Kendall (Tau-a): es el cociente entre la diferencia del número de concordancias y el número de discordancias en el numerador, y el número total de parejas en el denominador $\left(C-D/n\right)$. Su valor oscila entre 0 y 1 (valores mas próximos a 1 indican fuerte relación).

D de Sommers: Es un estadístico similar a la Gamma de Gamma de Goodman-Kruskal en cuanto a los valores que toma y su interpretación.

Es obtenida de la siguiente manera $D=\frac{C-D}{C+D+T}$

c es calculada de la siguiente manera

$c=.5\left(1+D \text{ de } sommers\right)$

Donde C es el número de pares concordantes, D el número de pares discordantes, T el número de pares y N el número total de pares.

2.8 Contraste de Bondad de Ajuste de Hosmer-Lemenshow

Este contraste evalúa la bondad del modelo, es decir el grado en que la probabilidad predicha coincide con la observada, construye una tabla de contingencia a la que aplica un contraste $\div^2$. Para ello calcula los deciles de las probabilidades estimadas $\left(\square_i; i=1,2,...,n\right)$, $D_1, D_2, ..., D_9$ y divide los datos observados en 10 categorías dadas por:

$$A_j=\left\{\square_i \in \left[D_{j-1}, D_j\right) \mid i \in \left\{1,2,...,n\right\}\right\};\quad j=1,2,...,10$$

donde $D_0=0,\quad D_{10}=1,$

Sean:

$n_j=$ número de casos en $A_j; j=1,2,...,10$

$o_j=$ número de $y_i=1$ en $A_j; j=1,2,...,10$

$$\overline{p_j}=\frac{1}{n_j}\sum_{i\in A_j}\square_i; j=1,2,...,10$$

El estadístico del contraste viene dado por:

$$T = \sum_{j=1}^{10} \frac{\left(o_j - n_j \overline{p_j}\right)^2}{n_j \overline{p_j}\left(1-\overline{p_j}\right)} \quad (2.70)$$

y el p-valor del contraste es $P\left[\chi_8^2 \geq T_{obs}\right].z$

2.9 Diagnóstico del modelo

Es la evaluación de la bondad de ajuste caso por caso mediante el análisis de los residuos del modelo y de su influencia en la estimación del vector de parámetros del mismo, se realiza usando:

Residuos estandarizados son el cociente entre residuales y una estimación de la desviación estándar.

$$z_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i\left(1-\hat{p}_i\right)}}; \quad i = 1,2,...,n \quad (2.71)$$

Residuos studentizados son el cambio en el valor de la desviación del modelo si el caso es excluido.

$$st_i = \frac{y_i - \hat{p}_{(i)}}{\sqrt{\hat{p}_{(i)}\left(1-\hat{p}_{(i)}\right)}}; \quad i = 1,2,...,n \quad (2.72)$$

donde $\hat{p}_{(i)}$ es la estimación P_i obtenida eliminando la observación i de la muestra.

Residuos Desvianza para cada observación la desviación se calcula:

$$d_i = \begin{array}{ll} \sqrt{-2\log\left(\hat{p}_i\right)} & si\ y_i = 1 \\ \sqrt{-2\log\left(1-\hat{p}_i\right)} & si\ y_i = 0 \end{array}; \quad j = 1,2,...,n \quad (2.73)$$

Todos estos residuos se distribuyen aproximadamente como una $N(0,1)$, si el modelo ajustado es correcto.

2.10 Medidas de influencia

Cuantifican la influencia que cada observación ejerce sobre la estimación del vector de parámetros o sobre las predicciones hechas a partir del mismo, de modo que, cuando más grande son, mayor es la influencia que ejerce una observación en la estimación del modelo.

Medida de apalancamiento de (Leverage)

Se utiliza para detectar observaciones que tienen un gran impacto en los valores predichos por el modelo.

Se calcula a partir de la matriz $H = W^{1/2} X \left(X'WX \right)^{-1} X'W^{1/2}$ donde $W = diag\left| p_i \left(1 - p_i\right) \right|$. El apalancamiento para la observación i-ésima viene dado por el elemento i.ésimo de la diagonal principal de H, h_{ii}, y toma valores entre 0 y 1 con un valor medio de p/n.

Las dos medidas siguientes miden el impacto que tiene una observación en la estimación de $\hat{a}$.

Distancia de Cook mide la influencia en la estimación de $\hat{a}$.

$$(2.74)\; COOK_i = \frac{1}{p}\left(\hat{a} - \hat{a}_{(i)} \right)' X'WX \left(\hat{a} - \hat{a}_{(i)} \right)$$

DFBETA mide la influencia en la estimación de una componente de $\hat{a}, \hat{a}_i$

$$Dfbeta1_i = \frac{\hat{a}_1 - \hat{a}_{1(i)}}{std\left(\hat{a}_1 \right)}$$

donde $\hat{a}_1, \hat{a}_{1(i)}$ denotan las estimaciones del modelo logístico de $\hat{a}$ y $\hat{a}_1$, eliminando la i-ésima observación de la muestra y $std\left(\hat{a}_1 \right)$ el error estándar en la estimación de $\hat{a}_1$.

2.11 Interacción y Confusión

Se denomina factor de confusión a una variable que se encuentra relacionada con la variable respuesta, así como, con otra variable regresora que es considerada factor de riesgo o significativa, y que además no se encuentra en la cadena causal (no sea significativa) entre este factor de riesgo y la variable dependiente.

Cuando a la asociación entre dos variables difiere significativamente si se considere, o no, otra variable, a esta última variable se le llama *variable de confusión* para la asociación. Se le llama así, puesto que al valorar la relación puede generar confusión sobre el valor real del factor de riesgo.

El modelo más sencillo para estudiar la asociación entre una variable respuesta y las variables regresoras es:

$$g(x) = \beta_0 + \beta_1 x \quad (2.75)$$

donde $g(x)$ es la transformación logit de $\pi(x)$ de 2.15, donde β_1 cuantifica la asociación: e^{β_1} es el odds ratio por unidad de cambio en x_1. Se dice que x_2 es una variable de confusión para esta asociación, si el modelo

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (2.76)$$

produce una estimación para β_1 diferente del modelo anterior.

Esta definición se puede amplia a un conjunto de variables, donde se dice que las variables $x_2, ..., x_k$ son variables de confusión si la estimación de β_1 obtenida por el modelo

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

es diferente de la obtenida en el modelo simple. En ambos casos se dice que la estimación de β_1 obtenida en los modelos múltiples está controlada o ajustada por $x_2, ..., x_k$.

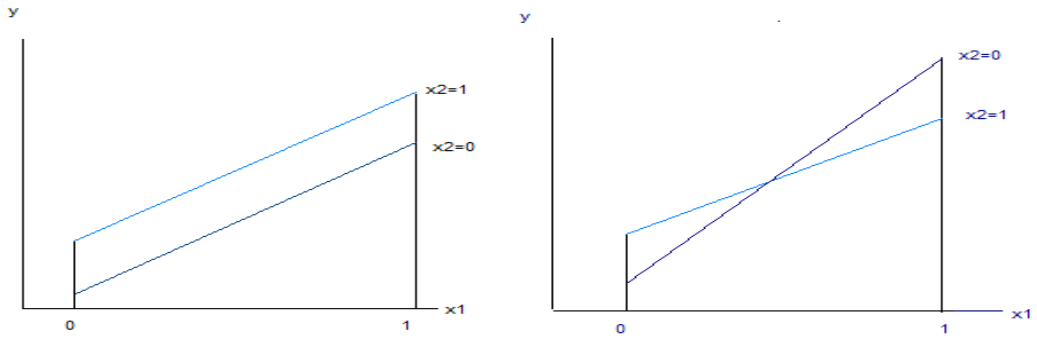
Por lo que para contrastar la existencia de confusión se requiere comparar los coeficientes de regresión obtenidos en dos modelos diferentes y si hay diferencia, existe la condición, en cuyo caso la mejor estimación es la ajustada. Para dicha comparación no se precisa realizar un contraste de hipótesis estadístico ya que aunque la diferencia encontrada sea debido al azar, representa una distorsión que la estimación ajustada corrige.

Un criterio para determinar la existencia de confusión es cuando la exponencial del coeficiente (el Odds Ratio) cambia en más del 10%.

Existe interacción cuando la asociación entre dos variables varía según los diferentes niveles de otra u otras variables, el factor de riesgo y la variable regresora son independientes, pero el factor de riesgo sobre la variable dependiente es diferente según el valor de dicha variable regresora (Llamado modificador del efecto).

En ocasiones esta diferenciación no es nítida, pudiendo existir confusión e interacción al mismo tiempo.

Figura 2.4 Ejemplo de Interacción



En la figura 2.3 muestra la existencia o no de interacción pura, donde se tiene una variable dependiente y, y dos covariables, x_1, x_2; la imagen de la izquierda muestra que la relación entre x_1 e y no varia según los valores de x_2 (La pendiente se mantiene constante, o que para cada valor de x_1 existe equidistancia entre las dos rectas). En este caso se dice que no existe interacción. En cambio en la imagen de la derecha si la existe, puesto que la relación entre x_1 e y se modifica según el valor de x_2.

El modelo correspondiente, debería de contener un tercer término (de interacción), que exprese la interacción entre las dos covariables.

El modelo más sencillo que hace explicita la interacción entre dos variables en donde se denota al factor de riesgo como F, la covariable como X y su interacción como $F \times X$.

El logit para este modelo evaluando a $F = f$ y $X = x$ es

$$g(f, x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f \times x \quad (2.77)$$

para una unidad de cambio en F donde $F = f_1 = x_1 + 1$ y $F = f_0 = x_1$ con $X = x$.

En este modelo el logit para f_0, x de F, X es

$$\begin{aligned} g(f_0, x) &= \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x + \beta_3 x_1 \times x \end{aligned} \quad (2.78)$$

para una unidad de cambio en el logit del factor de riesgo se tiene

$$\begin{aligned} g(f_1, x) &= \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x \\ &= \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x + \beta_3 (x_1 + 1) \times x \\ &= \beta_0 + \beta_1 x_1 + \beta_1 + \beta_2 x + \beta_3 x_1 \times x + \beta_3 x \end{aligned} \quad (2.79)$$

restando ambas se encuentra el cambio en el logit por una unidad de cambio en F

$$\begin{aligned} \ln\left[OR(F = f_1, F = f_0, X = x)\right] &= g(f_1, x) - g(f_2, x) \\ &= (\beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x) - (\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x) \\ &= \beta_1 (f_1 - f_0) + \beta_3 x (f_1 - f_0) \\ &= \beta_1 + \beta_3 x \end{aligned} \quad (2.80)$$

donde $f_1 - f_0 = (x_1 + 1) - x_1 = 1$

obtenemos el odds ratio por unidad de cambio en F

$$OR = \mathrm{e}^{\beta_1 + \beta_3 x} = e^{\beta_1} e^{\beta_3 x} = e^{\beta_1} (e^{\beta_3})^x \quad (2.81)$$

que es diferente para cada valor de la covariable X de x. De esta forma, cuando existe interacción, la OR del factor de riesgo no es constante e^{β_1}, sino que depende del valor que tome la covariable X. Si X es 0, la OR será e^{β_1}, pero si X es 1, la OR será $e^{\beta_1 + \beta_3 x}$.
Del mismo modo el OR del logit por unidad de cambio de la covariable es

$$e^{\beta_2+\beta_3 f} = e^{\beta_2}\left(e^{\beta_3}\right)^f \quad (2.82)$$

Por lo tanto, contrastar la existencia de interacción entre $F \times X$ es contrastar si el coeficiente β_3 es cero (no hay interacción), o distinto de cero (existe interacción).

En primer término se debe contrastar la interacción y después, en caso de que no exista, la confusión.

Los intervalos de confianza que presentan los programas asumen que no existe interacción y por tanto, son sólo parcialmente validos si existe interacción.

2.12 Selección automatizada de variables

La idea de este método es elegir el mejor modelo pero incluyendo o excluyendo una sola variable predictiva en cada paso de acuerdo a ciertos criterios. Esta secuencia termina cuando una regla de parada se satisface.

a) Paso a Paso hacia adelante (Forward)

Se comienza con aquella de mayor significancia, contrasta el modelo con la variable regresora frente al modelo sólo con la constante y la mantiene si la prueba de razón de verosimilitud es significativa.

Continúa evaluando a cada una de las variables regresoras restantes e incorpora aquéllas con mayor significación.

Este paso se repite una y otra ves hasta que no quedan covariables que incluir por no aportar significación. Si esta primera variable no es significativa se para el proceso y se considera el modelo $\hat{Y} = Y$, toda variable que es añadida al modelo ya no puede salir.

b) Paso a paso hacia atrás (Backward)

El algoritmo comienza incluyendo en el modelo todas las covariables y elimina en cada paso aquella variable regresora que menos contribuya a la significación del modelo, hasta mantener a todas las covariables que aportan significación al modelo. Donde todas las variables eliminadas ya no vuelven al modelo.

CATULO TERCERO
DESARROLLO DEL ESTUDIO

3.1 Diseño

El objetivo del análisis, es investigar cómo influye en la probabilidad de ocurrencia de la no participación de la gente en la última elección presidencial 2006, la presencia o ausencia de marginación en la población, mediante variables socioeconómicas que indican su grado de marginación por entidad federativa.

Las variables socioeconómicas de estudio serán los índices de marginación por entidad federativa, (CONAPO 2005) con base en el II Conteo de Población y Vivienda 2005, y la Encuesta nacional de Ocupación y Empleo (ENOE) 2005, IV Trimestre.

Tabla 3.1 Definición de variables de estudio

Definición	Nombre	Descripción
% Población analfabeta de 15 años y más	ANALFABETA	Representa el porcentaje de la población mayor a 15 años que no cuenta con educación básica por estado.
% Población sin primaria completa de 15 años y mas	PRIMARIA	Es el porcentaje de habitantes por estado que no han terminado la educación primaria.
% ocupantes en viviendas sin drenaje ni servicios sanitarios	DRENAJE	El porcentaje de habitantes en viviendas que no cuentan con servicios sanitarios, ni drenaje.
% Ocupantes en viviendas sin energía eléctrica	ENERGIA	Es el porcentaje de personas que habitan en viviendas que no cuentan con energía eléctrica.
% Ocupantes en viviendas sin agua entubada	AGUAENTUB	Representa el porcentaje de habitantes por estado que no cuentan con el servicio de agua entubada.
% Viviendas con algún nivel de hacinamiento	HACIANAMIENTO	Es el porcentaje de viviendas por estado que poseen algún nivel de hacinamiento en viviendas, esto es, que viva un número elevado de personas en una misma casa.
% Ocupantes en viviendas con piso de tierra	PISOTIERRA	Porcentaje de personas que habitan en viviendas con piso de tierra.

% Población en localidades con menos de 5000 habitantes	MENHABIT	Es el porcentaje de personas que viven en localidades en la que su número de habitantes es menor de 5,000 habitantes.
% Población ocupada con ingresos de hasta dos salarios mínimos	SALARMINM	Porcentaje de personas que percibe como máximo 2 salarios mínimos.
Índice de marginación	INDIMAR	Mide por estado el grado de marginación en base a la cobertura de servicios en la población e ingreso económico, este entre mas se aleje este del cero en sentido positivo, mas alta será la marginación del estado, por el contrario cuando el signo es positivo indica a los estados que cubre en buena proporción los servicios públicos a la población que habite en este.
Lugar que ocupa en el contexto nacional	LUGARNAC	Enumera del 1 al 32 a los estados de manera descendente, indicando en el primer lugar al estado con el índice de marginación mas elevado, estos son, los que muestran porcentajes preocupantes para solventar sus necesidades básicas de supervivencia y que se encuentran habitando en condiciones muy precarias.
Población que se abstuvo de votar	ABS	Es la variable dependiente del modelo, la cual indica si el estado tuvo un alto abstencionismo o voto.

La variable dependiente será la variable definida como ABS, esta se construyo dividiendo el número total de abstencionistas en el estado entre el total de empadronados por estado, lo que resulta un porcentaje de abstencionismo por estado.

Del porcentaje de abstencionismo por estado, se calculó su media con la fórmula:

$$\frac{\sum_{i=1}^{n} X_i}{n} = \hat{x}$$

Donde:

x_i es el porcentaje de abstencionismo por estado, con $i = 1,...,32$

n es el número total de estados

El resultado es el promedio general de abstencionismo a nivel nacional. Este resultado se comparo nuevamente contra cada porcentaje de abstencionismo por estado x_i

Si $x_i < \hat{x}$ entonces ABS es igual a 0, donde cero indica que el estado no se abstuvo.
Si $x_i > \hat{x}$ entonces ABS es igual a 1, donde uno indica que el estado se abstuvo.

Se usan los siguientes comandos de SAS para leer los datos y llevarlos a un conjunto de datos SAS e imprimirlos:

```
title 'Regresion Logistica';
DATA Abstencionismo;
INPUT  ABS $ ANALFABETA PRIMARIA DRENAJE ENERGIA AGUAENTUB
HACINAMIENTO PISOTIERRA MENHABIT SALARMINIM INDIMAR LUGARNAC $ ;
CARDS;
ABS   4.16  17.82 1.68  0.85  1.79  34.34 2.36  24.67 32.79 -0.95352    28
NOABS 3.08  15.02 0.56  1.49  4.89  32.26 3.95  9.27  14.24 -1.25336    30
NOABS 3.62  16.49 1.84  2.88  11.28 35.16 8.27  19.84 24.07 -0.71946    24
ABS   10.2  26.96 9.85  4.85  11.15 51.42 8.77  32.01 55.1  0.55876      8
NOABS 3.29  14.6  1.65  0.77  2.17  33.71 2.51  12.33 32.97 -1.13709    29
ABS   6.42  21.58 0.8   0.67  1.66  34.92 8.4   15.78 42.66 -0.73788    25
NOABS 21.35 42.76 8.07  5.88 25.9  60.2  32.99 58.46 78.14  2.326460     2
NOABS 4.42  18.81 3.29  4.28  6.45  31.93 6.45  17.54 27.25 -0.68411    23
ABS   2.59  9.7   0.16  0.15  1.51  29.31 1.12  0.52  33.04 -1.50487    32
NOABS 4.84  22.92 8.51  3.52  8.63  36.09 11.34 38.64 50.04 -0.01884    15
NOABS 10.44 28.8  9.87  1.93  6.14  40.99 8.95  34.46 44.69  0.09191    14
NOABS 19.88 35.98 27.18 6.33  31.34 55.06 35.69 50.51 64.97 2.41213     1
ABS   12.8  27.5  8.98  3.9   12.21 42.69 12.78 57.28 61.63 0.75057     5
ABS    5.56 21.3  2.42  1.12  5.91  33.27 5.35  17.4  34.74 -0.76871    27
ABS    5.32 16.24 4.76  0.96  6.04  41.46 6.04  18.63 41.18 -0.62211    21
NOABS 12.58 33.48 5.66  2.11  9.97  40.01 15.67 40.51 55.79  0.45654    10
ABS    8.13 21.01 3.1   0.81  7.84  38.63 10.71 22.88 37.77 -0.44346    20
NOABS  8.02 26.05 6.78  4.38  8.35  37.73 9.93  41.67 51.73  0.19052    12
ABS    2.78 12.7  0.54  0.56  3.48  33.49 2.34  7.13  23.55 -1.32611    31
ABS   19.35 38.49 6.84  7.21  26.29 53.06 35.17 61.27 69.65  2.12936     3
ABS   12.71 29.02 5.45  2.19  14.03 49.16 15.86 39    61.34  0.63482     7
ABS    8.14 20.03 9.95  2.99  5.76  37.6  8.59  38.34 39.23 -0.14165    17
NOABS  6.58 19.42 5.19  2.59  4.66  49.9  8.41  17.1  36.78 -0.31569    19
ABS    9.92 27.42 5.72  5.58  16.97 38.6  19.67 41.18 56.11  0.65573     6
NOABS  6.42 23.42 5.14  1.92  6.24  43.08 9.78  35.3  44.95 -0.14817    18
NOABS  3.73 17.21 1.92  1.87  4.01  38.73 9.67  18.29 31.76 -0.74955    26
ABS    8.57 25.1  3.99  1.95  22.94 47.11 9.11  55.78 51.97  0.46224     9
NOABS  4.52 18.61 0.84  2.88  4.26  39.24 5.39  14.92 37.56 -0.68338    22
ABS    6.68 18.78 4.84  1.11  2.03  47.93 6.26  40.3  62.59 -0.12922    16
ABS   13.42 32.9  4.18  4.67  23.32 45.02 22.77 46.68 58.36  1.07674     4
ABS   10.89 29.99 17.96 2.61  3.03  48.36 4.81  27.42 63     0.43144    11
NOABS  7.2  30.83 10.53 1.91  6.72  37.06 6.29  50.59 54.25  0.15999    13
```

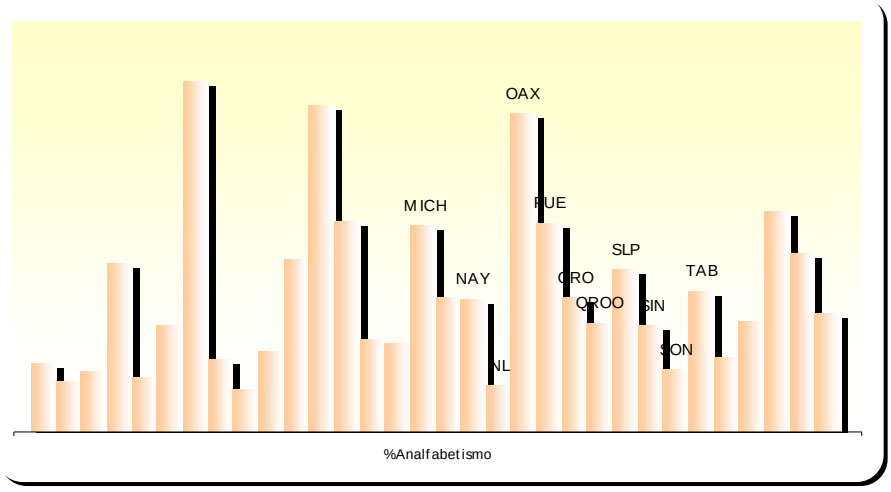
```
PROC PRINT;
run;
```

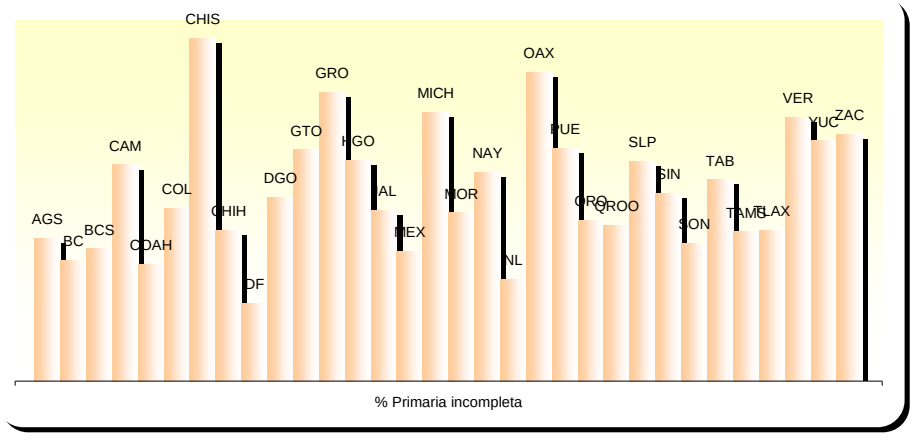
Nota: Los datos empleados para los cálculos no contemplan los votos emitidos en el extranjero.

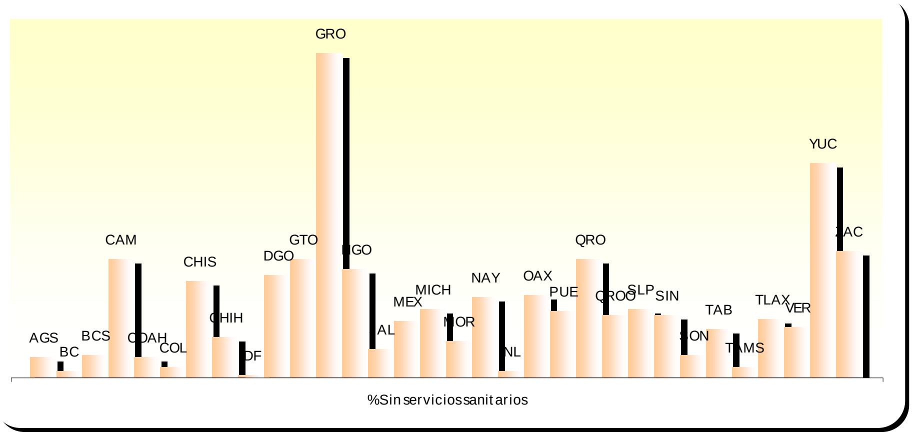
Ahora los datos están listos para ser analizados con el procedimiento LOGISTIC de SAS.

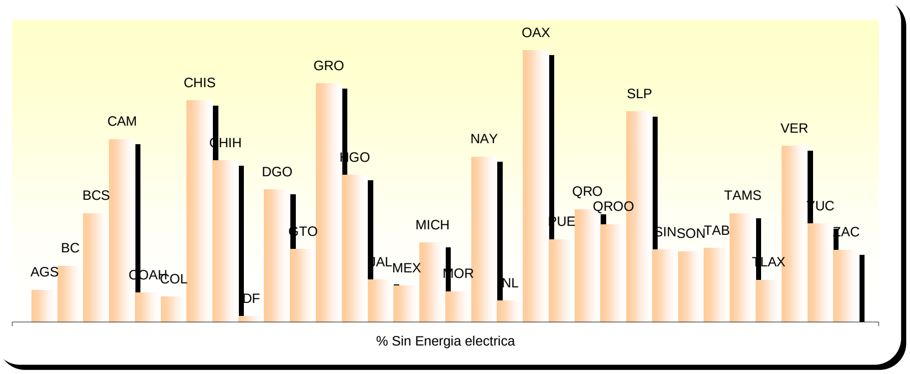
3.2 Selección del modelo

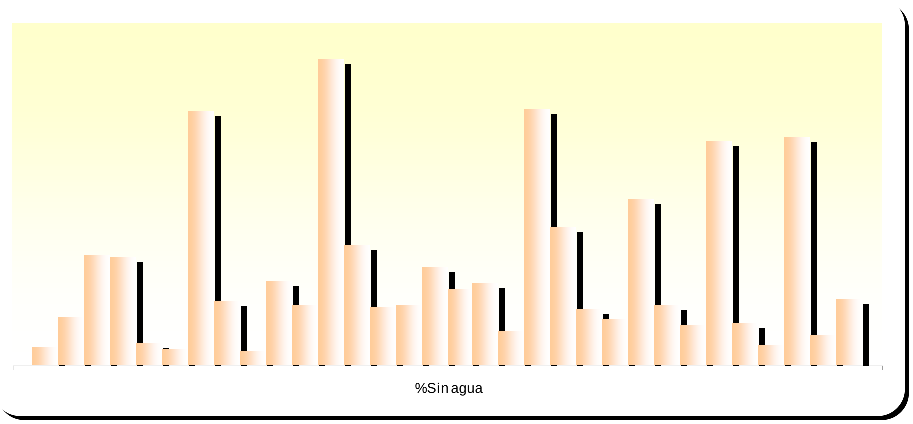
a) Análisis exploratorio previo

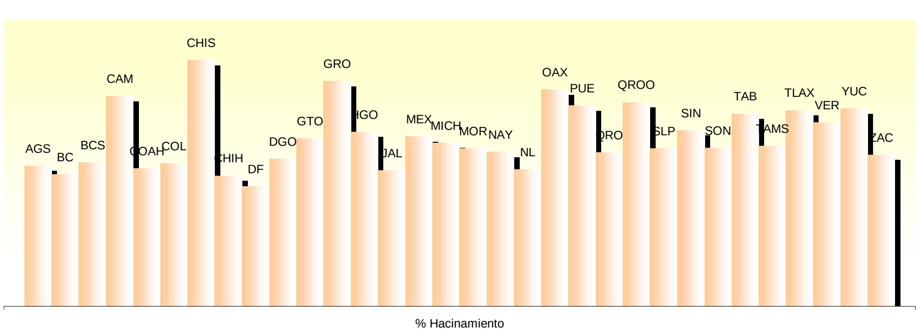


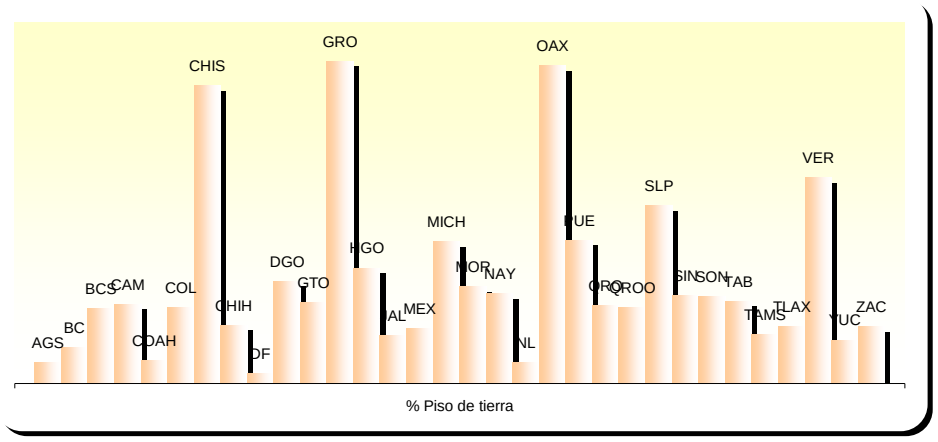


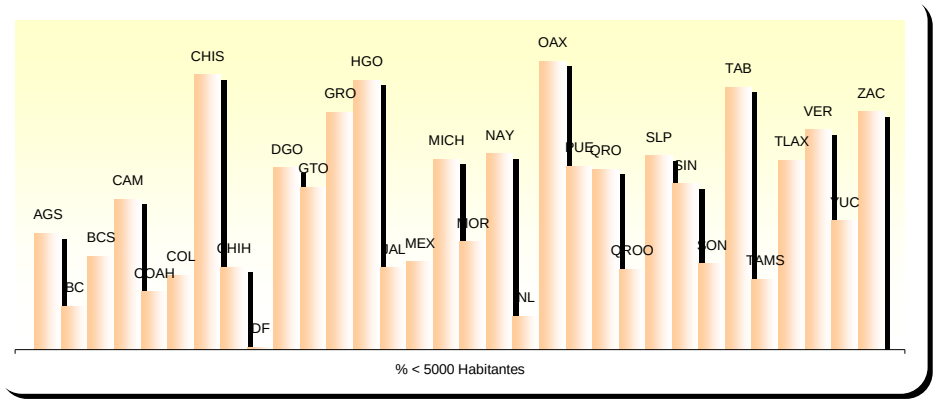

GRO
YUC
ZAC
CAM
GTO
QRO
HGO
DGO
CHIS
NAY
OAX
MICH
PUE
QROO
SLP
SIN
MEX
TAB
TLAX
VER
MOR
CHIH
AGS
BC
BCS
COAH
COL
JAL
SON
TAMS
DF
NL
%Sin serviciossanitarios

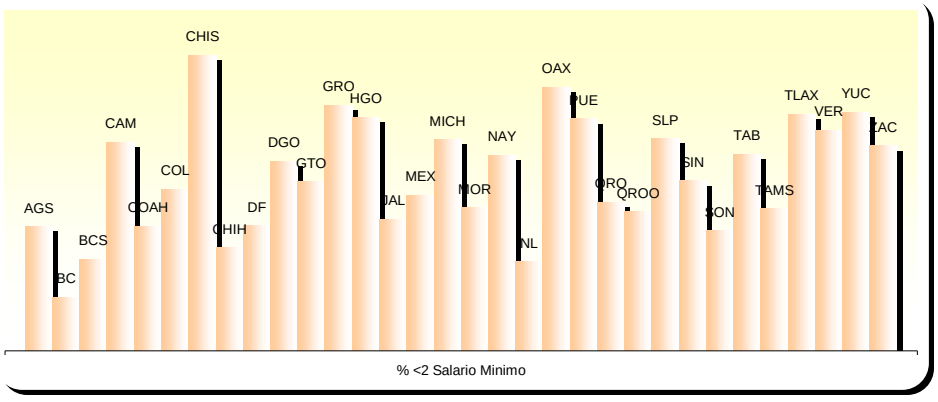

OAX
GRO
CHIS
SLP
CAM
NAY
VER
CHIH
HGO
DGO
BCS
QRO
TAMS
QROO
YUC
PUE
MICH
GTO
SIN
SON
TAB
ZAC
BC
TLAX
JAL
MEX
AGS
COAH
COL
MOR
NL
DF
% Sin Energia electrica


%Sin agua


CHIS
GRO
OAX
CAM
PUE
QROO
TAB
TLAX
YUC
VER
SIN
GTO
HGO
MEX
MICH
MOR
NAY
QRO
SLP
SON
TAMS
DGO
AGS
BCS
COAH
COL
ZAC
BC
CHIH
JAL
NL
DF
% Hacinamiento







Se realizará un análisis exploratorio previo para ampliar nuestros conocimientos sobre los datos y sus posibles relaciones entre las variables. Lo que nos ayuda a la especificación y refinamiento del modelo, así como una perspectiva para la interpretación de los resultados con el fin de encontrar posibles problemas, como: datos ausentes, valores extremos o atípicos (outliers).

Como un primer paso se presentan las estadísticas descriptivas

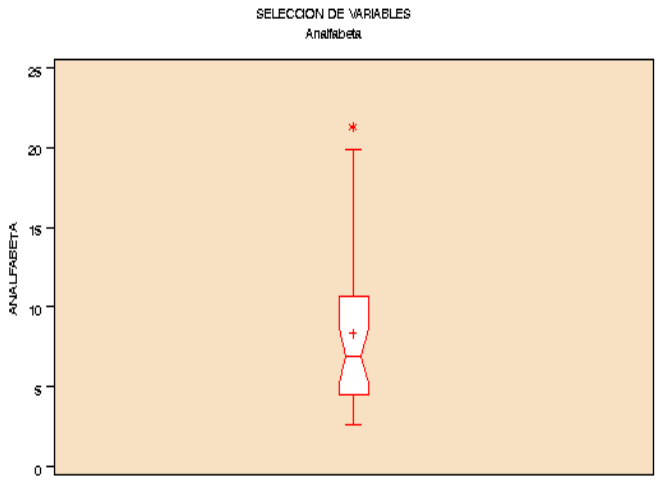
ANALISIS EXPLORATORIO

Procedimiento MEANS

Variable	Media	Desviación estándar	Varianza	Mínimo	Máximo	Rango
ANALFABETA	8.3628125	5.0007994	25.0079951	2.5900000	21.3500000	18.7600000
PRIMARIA	23.7793750	7.8029948	60.8867286	9.7000000	42.7600000	33.0600000
DRENAJE	5.8828125	5.4987050	30.2357564	0.1600000	27.1800000	27.0200000
ENERGIA	2.7162500	1.8590906	3.4562177	0.1500000	7.2100000	7.0600000
AGUAENTUB	9.592812	8.1245634	66.0085305	1.5100000	31.3400000	29.8300000
HACINAMIENTO	41.172500	7.6004957	57.7675355	29.3100000	60.2000000	30.8900000
PISOTIERRA	11.106250	9.0590709	82.0667661	1.1200000	35.6900000	34.5700000
MENHABIT	31.428125	16.5671719	274.4711835	0.5200000	61.2700000	60.7500000
SALARMINIM	46.059375	15.1086369	228.2709093	14.2400000	78.1400000	63.9000000
INDIMAR	9.375E-7	0.9999999	0.9999998	-1.5048700	2.4121300	3.9170000
LUGARNAC	16.50000	9.3808315	88.0000000	1.0000000	32.0000000	31.0000000

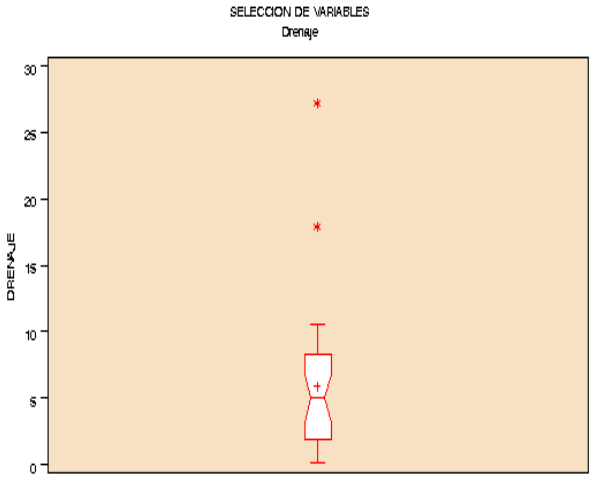
Emplearemos los gráficos de cajas, el cual ayudara a visualizar como se comportan los datos en cada variable y sus valores extremos.

Figura 3.1



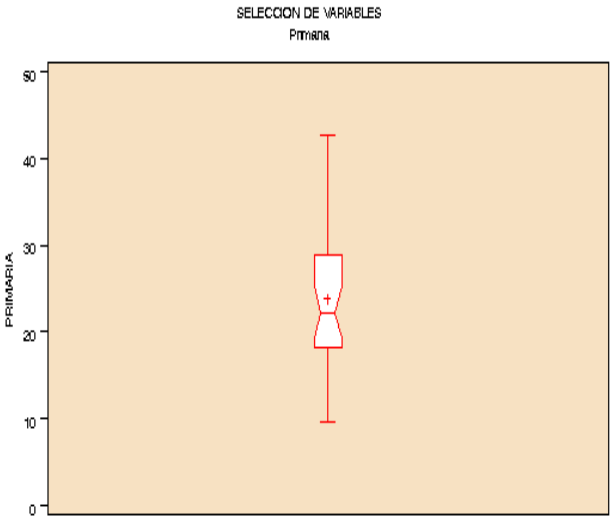
El analfabetismo nacional varía entre 2.44% y 10.77%. Su media es de 8.36%, presenta una distribución asimétrica, con una dispersión moderada, lo cual indica un analfabetismo similar entre los estados, a excepción de Chiapas que presenta un analfabetismo de 21.35%.

Figura 3.2



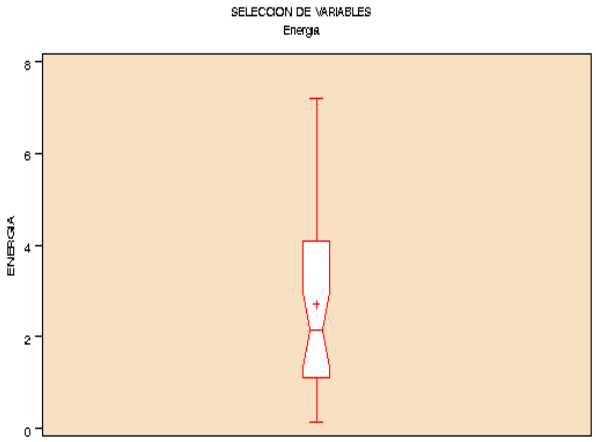
Este servicio no es proporcionado entre 1.86% y 8.4%. Su media es de 5.88%, presenta una distribución poco asimétrica y una dispersión mínima, a excepción de Yucatán con un porcentaje de falta del servicio de 17.96% y guerrero con 27.18%.

Figura 3.3



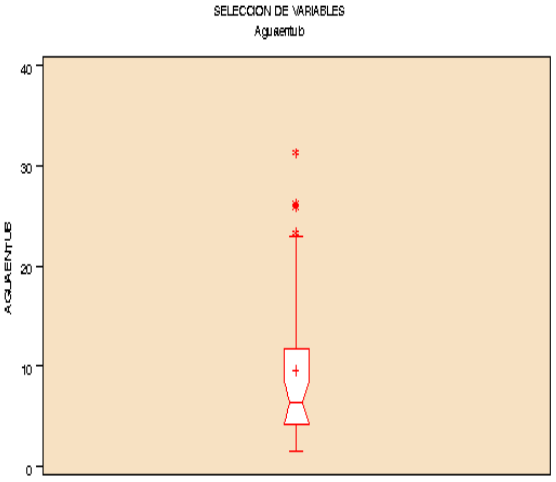
La población sin primaria completa varia entre 18% y 28.9%, la cual presenta una distribución asimétrica, ya que, existen un gran número de estados que presentan en gran porcentaje este problema, con una dispersión amplia lo que indica una gran desigualdad entre estados, pues el máximo se presenta en un 42.7% y el mínimo en 9.7%, con una media de 23.77%.

Figura 3.4



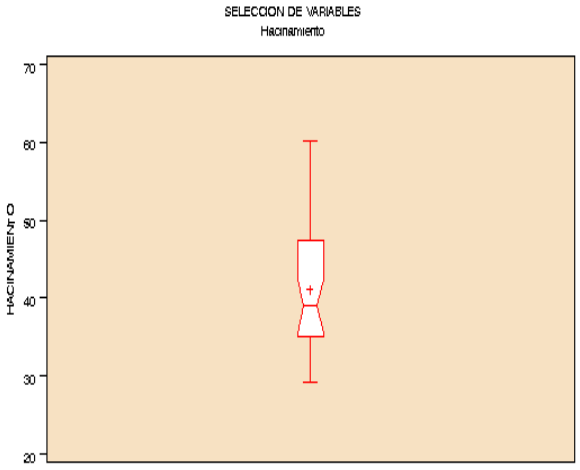
La falta de este servicio varia entre 1.11% y 4.18%, la cual presenta una distribución asimétrica debido a que un gran número de estados presentan un mayor porcentaje de falta de este servicio, con una alta dispersión que indica una gran desigualdad entre estados, ya que su máximo se presenta en 7.21% en Oaxaca, un mínimo en .15% en el DF, con una media de 2.7%

Figura 3.5



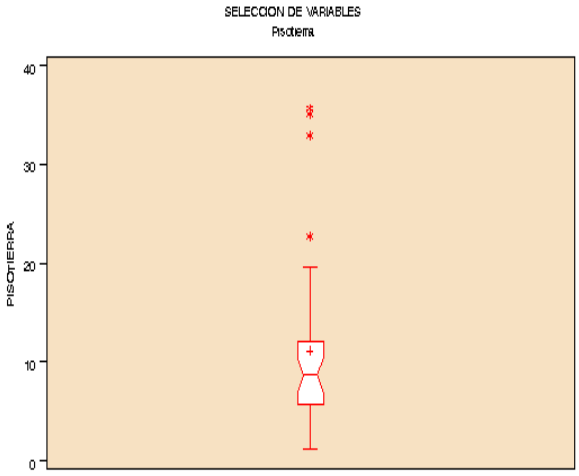
El servicio no es proporcionado entre un 4.07% y 11.98%, con una distribución asimétrica ya que un mayor numero de estados no cuenta con el servicio, con una dispersión baja ya que la mayoría se encuentra entre los rangos anteriores, a excepción de dos estados que muestran datos atípicos en Chiapas con un 26.29% y Guerrero con 31.34%, con una media del 6.3%

Figura 3.6



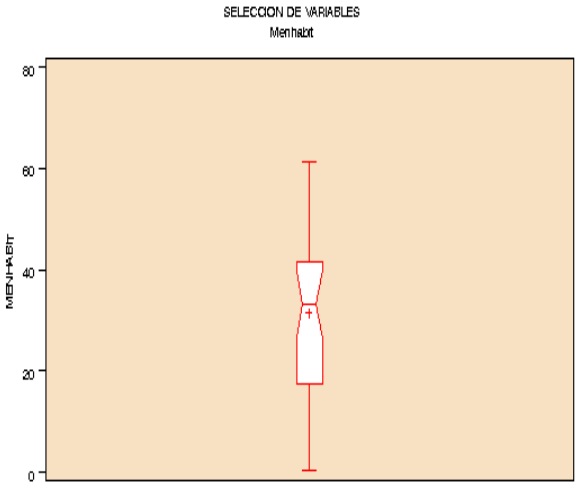
El hacinamiento varia entre 34.9% y 47.7%, con una distribución asimétrica, pero con una dispersión moderada ya que la mayoría no se aleja de la media, la cual es de 38.98%, por lo que se puede concluir que en general en todos los estados hay un alto grado de hacinamiento.

Figura 3.7



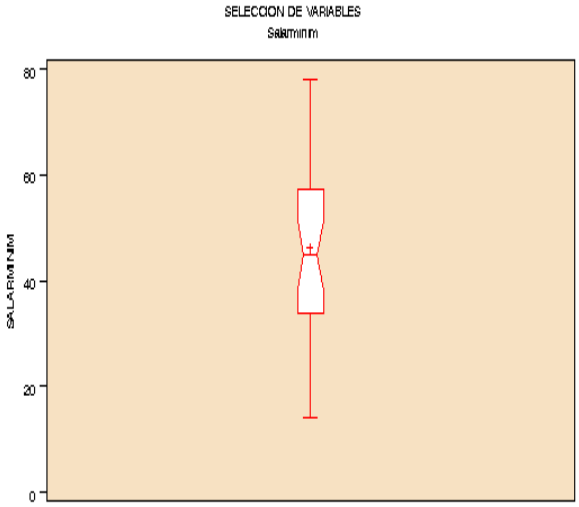
Los estados que poseen algún porcentaje de viviendas con piso de tierra varia entre 5.5% al 12.4%, tiene una distribución simétrica, ya que la mayoría de los estados se encuentra en este rango, pero su media se encuentra alta con un valor de 11.1%, debido a que existen tres datos atípicos, los cuales son Veracruz con un 22.7%, Chiapas con un 32.99% y Oaxaca con un 35.69%

Figura 3.8



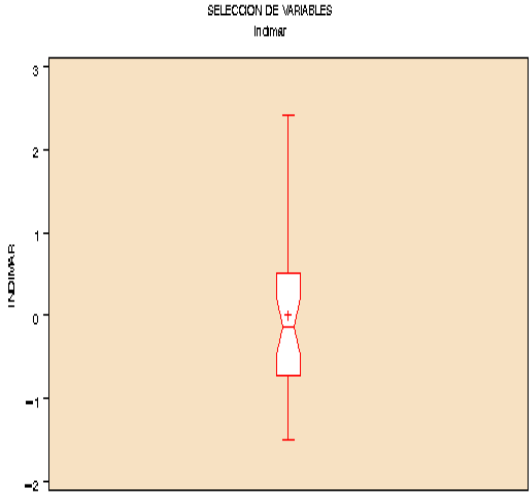
Los estados con un porcentaje de poblaciones con un numero de habitantes menor a 5000 personas presenta una distribución asimétrica, ya que son pocos los estados que presentan un alto grado de este porcentaje, pero con una dispersión amplia ya que la diferencia entre estados es grande la cual varia entre 17.4 a 41.5, con una media de 33.2%

Figura 3.9



El porcentaje de estados con población que gana menos de dos salarios mínimos varia entre 33.4% y 57.7%, con una distribución simétrica, ya que los estados presentan similares proporciones en el numero de estados con un alto grado y los que presentan uno menor, con una dispersión amplia debido a que existen estados con contrastes de desigualdad, ya que el mínimo es de 14.2% y el máximo es de 78.4%,con una media de 46.05%

Figura 3.10

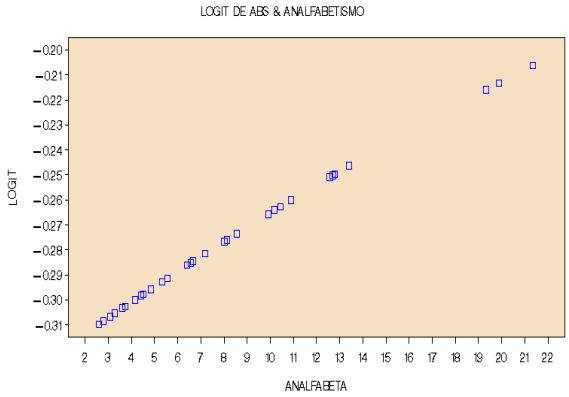


El índice de marginación presenta una distribución simétrica, ya que existen similares proporciones entre estados con altos índices y los que presentan menores, pero presentándose en mayor número los que poseen indicadores positivos, con una dispersión moderada, ya que la mayor parte varían entre 0.53% a -0.7%, con una media de -0.13%.Donde porcentajes positivos indican mayor marginación.

El lugar nacional indica la posición con respecto a la marginación que ocupan a nivel nacional los estados.

El logit de la distribución logística es una relación lineal entre el logaritmo del Odds Ratio y una combinación lineal de las variables reagresoras, se observará su comportamiento para determinar si se cumple esta relación en las variables a analizar.

Figura 3.11



Se observa una relación lineal entre el Logit y el analfabetismo, esta relación es más fuerte en bajos porcentajes de analfabetismo. Donde se puede observar que a menor porcentaje de analfabetismo, se incrementa la probabilidad de abstencionismo en la población.

Figura 3.11

Existe una relación lineal entre el logit del abstencionismo y poblaciones sin primaria completa donde a mayor porcentaje de población sin primaria completa, se incrementa la probabilidad de abstencionismo.

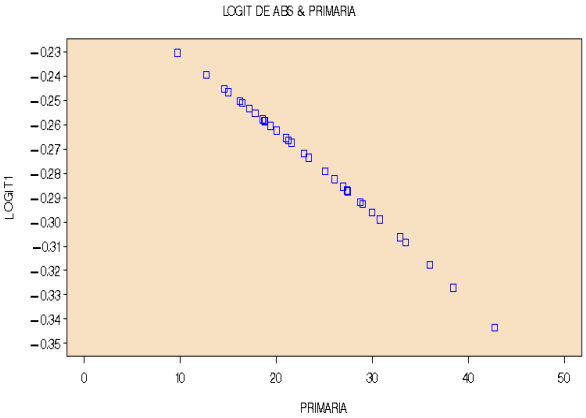
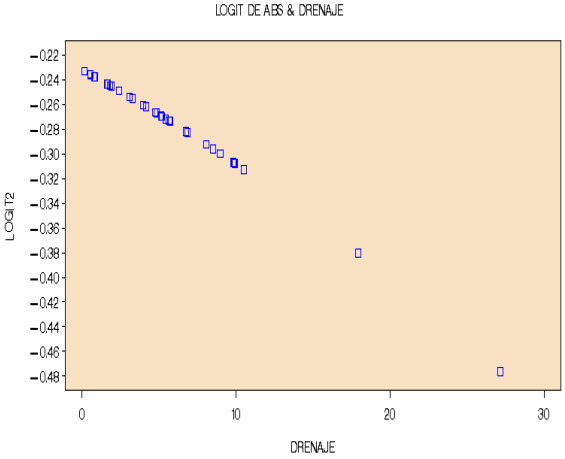


Figura 3.12



La relación lineal entre la falta del servicio de drenaje y el logit del abstencionismo, es que, entre mayor es la falta del servicio, incrementa la probabilidad de abstencionismo.

Figura 3.13

Se aprecia que no existe relación lineal entre la falta de servicio de energía y el logit de abstencionismo.

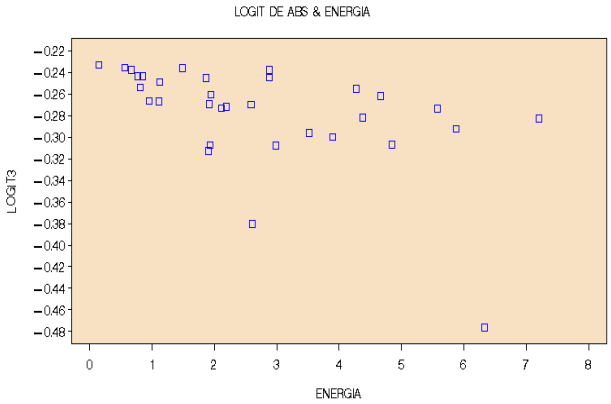
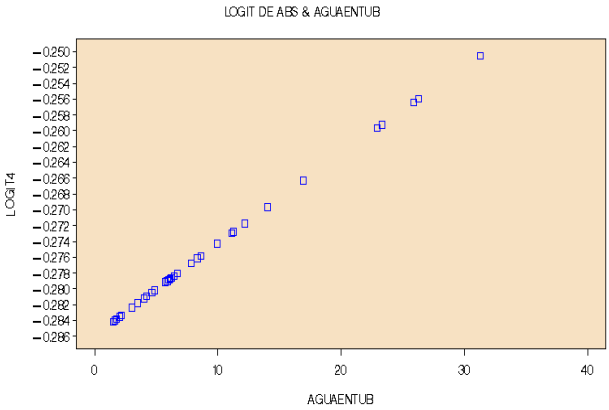


Figura 3.14



Existe una relación lineal entre viviendas sin agua entubada y el logit del abstencionismo donde a mayor porcentaje de población sin el servicio, disminuye la probabilidad de abstencionismo.

Figura 3.15

La relación lineal entre el hacinamiento en viviendas y el logit del abstencionismo, es que a mayor porcentaje de población con algún grado de hacinamiento en su vivienda, disminuye la probabilidad de que se abstengan.

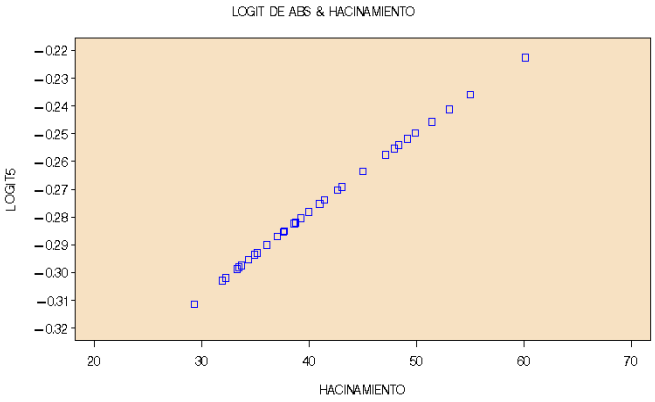
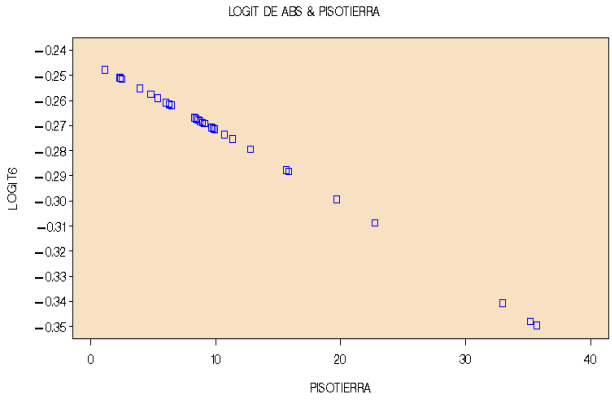
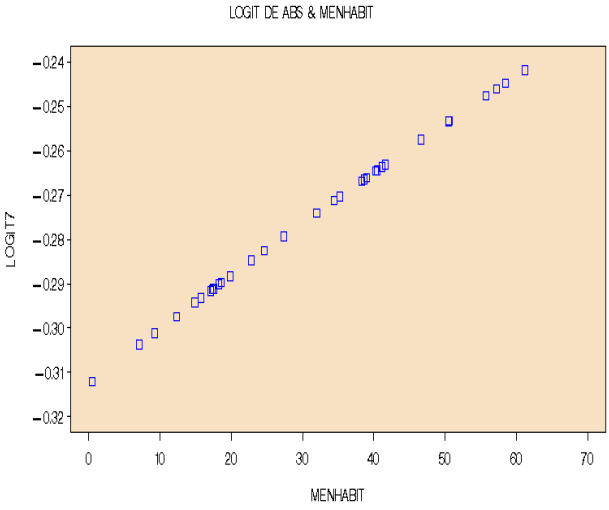


Figura 3.16



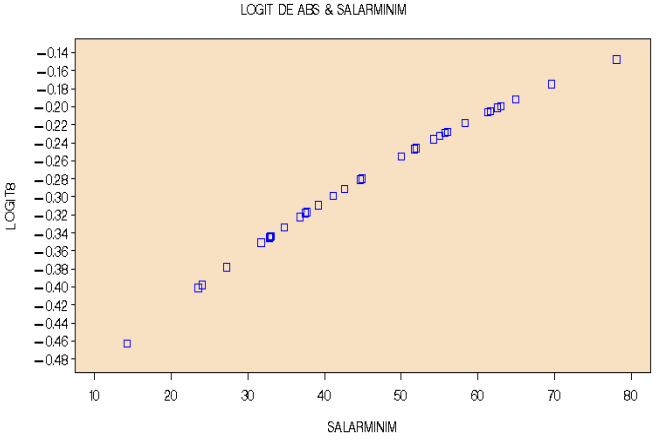
La relación lineal entre viviendas con piso de tierra y el logit del abstencionismo, es que a mayor porcentaje de población con piso de tierra, incrementa la probabilidad de que se abstenga.

Figura 3.17



La relación lineal entre poblaciones con menos de 5000 habitantes y el logit del abstencionismo, es que a mayor porcentaje de poblaciones con esta característica, menor es la probabilidad de abstencionismo.

Figura 3.18



Como se observa, la relación lineal entre la población que sólo percibe como ingreso hasta dos salarios mínimos, es que a menor porcentaje de población con bajos ingresos, se incrementa la probabilidad de que estos se abstengan.

Figura 3.19

La relación lineal entre el índice de marginación y el logit del abstencionismo, es que a menor grado de marginación, mayor es la probabilidad de abstenerse.

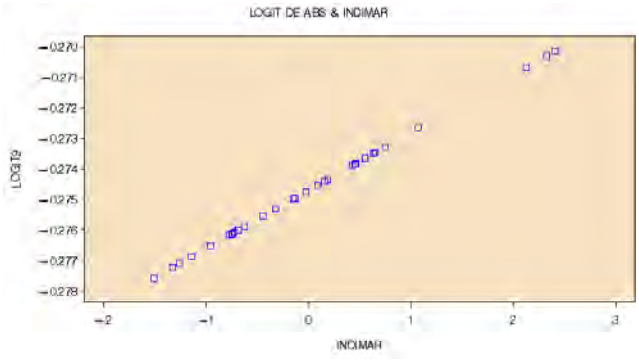
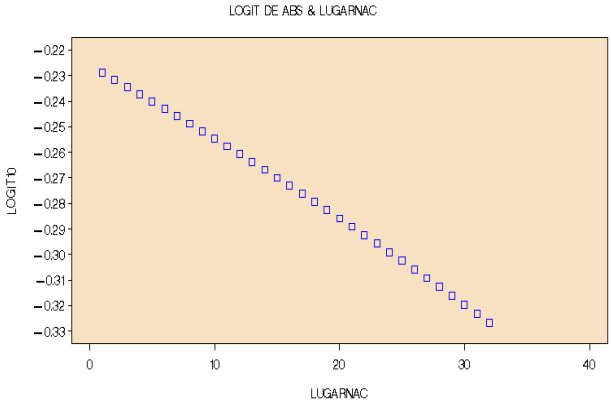


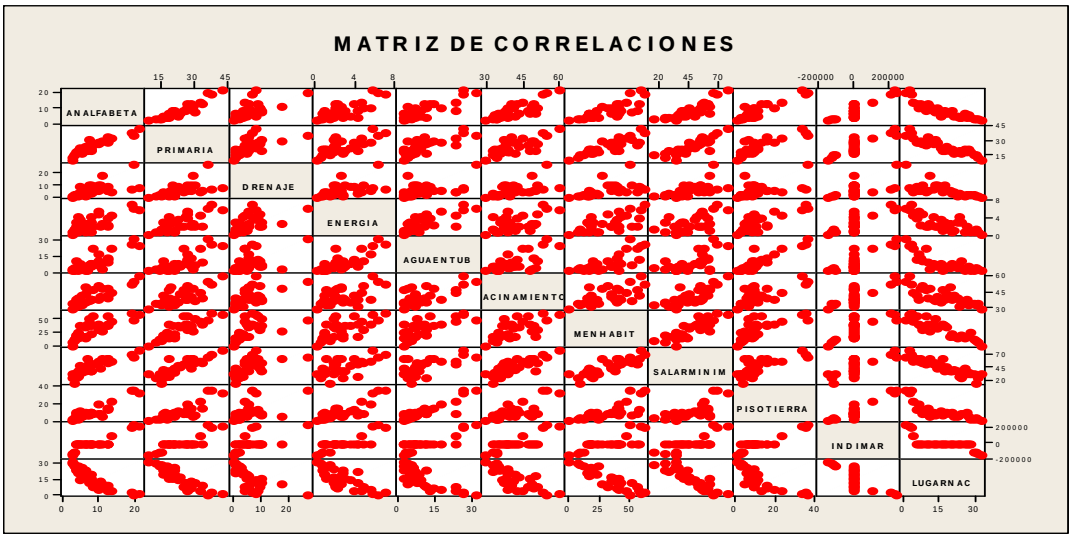
Figura 3.20



La relación lineal existente entre el lugar que ocupan en cuanto a marginación en el contexto nacional, es que se comporta de forma ascendente, puesto que a menor pobreza en el estado, incrementa la probabilidad de abstenerse.

Para definir cual es la interrelación entre las variables se empleara el análisis de la matriz de correlaciones.

Figura 3.21



Como se puede observar la variable LUGARNAS (Lugar nacional) mantiene un colinealidad perfecta, ya que se puede determinar a partir de una ecuación lineal de todas las variables, por lo que es irrelevante su presencia en el modelo.

Para el análisis de multicolinealidad entre las variables y así obtener las variables más relevantes en el estudio, se observará el posible problema de la siguiente manera; tomaremos en primer lugar una covariable cualquiera como variable dependiente, dejando de lado por el momento ha la variable dependiente abstencionismo (ABS).

Realizaremos modelos de regresión lineal. A cada uno de estos modelos podemos calcularle su R^2. Se denomina tolerancia al complemento de $\mathrm{R}^2\left(1-\mathrm{R}^2\right)$, factor de inflación de la varianza (FIV) al inverso de la tolerancia $\left[1/\mathrm{R}^2\left(1-\mathrm{R}^2\right)\right]$. Cuando existe estrecha relación entre covariables la tolerancia tiende a ser 0, y por tanto FIV tiende al infinito. Como regla general nos deberían de preocupar tolerancias menores de 0.1 y FIV mayores de 10. SAS nos ofrece la matriz de correlaciones, pero no aporta índices de multicolinealidad para la regresión logística. Por lo que para ello emplearemos el paquete estadístico MINITAB para calcular el FIV (Tabla 3.).

ANALFA	14,6	20,0	20,8	21,4	18,8	21,2	21,0	15,3	20,9	21,5
8,7	PRIMA	12,6	12,8	12,4	11,1	12,7	12,8	12,5	11,2	11,4
2,0	2,1	DRENA	2,2	2,1	2,1	2,1	2,2	2,0	2,0	2,0
4,7	4,8	4,8	ENERGIA	4,8	4,7	4,6	4,5	4,4	4,5	3,4
8,2	8,0	8,1	8,2	AGUAENT	8,1	7,8	6,5	6,1	8,2	7,3
4,5	4,5	5,1	5,1	5,1	HACINA	4,6	5,0	4,8	3,9	4,6
7,6	7,7	7,6	7,4	7,3	6,9	MENHAB	7,3	7,6	6,9	5,9
9,6	9,8	9,8	9,1	7,8	9,4	9,3	SALARM	9,6	9,8	7,9
12,4	17,1	16,4	15,8	13,0	16,1	17,1	17,0	PISOT	13,9	17,3
8,7	7,9	8,2	8,3	9,0	6,8	8,0	9,0	7,2	INDIM	7,0
25,1	22,5	23,1	17,8	22,4	22,2	19,4	20,0	24,9	19,7	LUGN

Se puede observar que las variables marcadas en rojo, señalan un VIF superior al tolerable, además de que, esto se presenta en todas las covariables, esto indica una alta relación lineal con las demás, por esto dichas covariables pueden ser explicadas con cualquiera de las variables restantes, retirando del modelo ANALFABETA (Porcentaje de analfabetismo), PISOTIERRA (Porcentaje de viviendas con piso de tierra) y LUGARNAC (Lugar nacional de marginación). Mientras que la variable PRIMARIA (Porcentaje de habitantes sin primaria completa) permanece en el modelo debido a que sus valores del VIF son menores que los anteriores casos a pesar de que salen de la tolerancia.

Retirando dichas variables del modelo, se observa un gran cambio entre las variables de estudio restantes.

RIMARIA	6,8	6,9	6,8	6,9	6,7	5,7	6,2
1,8	RENAJE	1,7	1,7	1,7	1,8	1,8	1,8
3,1	3,0	NERGIA	2,7	3,1	3,1	3,1	2,9
4,2	4,2	3,7	GUAENTUB	4,0	3,7	4,0	4,0
3,8	3,7	3,8	3,6	ACINAMIEN	3,6	2,8	3,4
4,9	5,2	5,2	4,5	4,9	ENHABIT	3,8	5,2
5,8	7,2	7,2	6,7	5,2	5,3	ALARMINIM	7,1
4,2	4,7	4,4	4,4	4,2	4,7	4,7	NDIMAR

Como se ve la presencia de valores altos del FIV a desaparecido entre las variables de estudio, por lo que dichas variables serán las consideradas para detectar cuales son significativas en el modelo.

3.3 Análisis del modelo de Regresión Logística

Los cálculos y gráficos se realizaron a través del paquete estadístico SAS for Windows versión 9.0.

Como primer paso se especificara el modelo máximo, el cual establece todas las variables que van a ser consideradas. El modelo saturado (el máximo que se puede considerar) tiene *n-1* variables, pero este no es de gran interés, por lo que el modelo máximo debe de tener menos variables independientes que el saturado (un criterio habitual es incluir como máximo una variable cada 10 observaciones).

En el presente estudio se cuentan con 11 variables de estudio por lo que nuestro modelo saturado será de 10 variables, que nos ayudara a minimizar la probabilidad de error tipo II, que consiste en no considerar una variable que tiene un coeficiente de regresión distinto de cero.

Ahora, ya que se cuenta con 32 observaciones correspondientes al total de estados que componen la república mexicana, se puede concluir que nuestro modelo máximo será de 3 variables el cual también debe de considerar los términos de interacción que se van a introducir.

Este modelo máximo al ser más reducido que el anterior, minimiza la probabilidad de error tipo I o sobré ajuste (Incluir en el modelo una variable independiente cuyo coeficiente de regresión sea cero).

Del modelo de regresión se retirarán las variables que presentaron alta multicolinealidad (ANALFABETA, PISOTIERRA Y LUGARNAC), así como la variable ENERGIA pues no presento relación lineal con el Logit de ABS.

a) Ajuste inicial del modelo de Regresión Logística

Se dan los comandos SAS recomendados para llevar acabo la regresión logística. Las instrucción modelo que se usa en estos comandos debe explicarse por si misma por que la variable dependiente está en el primer miembro del signo igual y todas las variables predictoras se listan en el segundo miembro. Se utiliza la instrucción UOPUT para predecir un conjunto de datos de salida llamados PDICTS además de contener aquellas variables que se encontraban en el conjunto original de datos, este ultimo contiene una variable llamada PHAT, que da la probabilidad estimada de que un estado pertenezca a la primera población (El grupo de riesgo ABS).

```
PROC LOGISTIC DATA=ABSTENCIONISMO;
MODEL ABS = PRIMARIA DRENAJE AGUAENTUB HACINAMIENTO  MENHABIT
SALARMINIM INDIMAR;
OUTPUT OUT=PDICTS PREDICTED=PHAT;
RUN;
DATA; SET PDICTS;
IF PHAT>.5 THEN PREDICT='ABS'; ELSE PREDICT='NOABS';
IF PHAT=. THEN DELETE;
PBAD=PHAT; PGOOD=1-PHAT;
PROC PRINT;
RUN;
PROC FREQ;
TABLE ABS*PREDICT;
RUN;
```

Con el fin de estimar los valores de los coeficientes de las β´s se ajusta un modelo de regresión logística binaria teniendo como primer resultado con todas las variables independientes el siguiente:

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	0.8526	4.0440	0.0445	0.8330
PRIMARIA	1	-0.3630	0.2059	3.1082	0.0779
DRENAJE	1	-0.0880	0.0972	0.8203	0.3651
AGUAENTUB	1	0.1174	0.1155	1.0336	0.3093
HACINAMIENTO	1	-0.0658	0.1147	0.3285	0.5665
MENHABIT	1	-0.0374	0.0650	0.3310	0.5651
SALARMINIM	1	0.2463	0.1106	4.9617	0.0259
INDIMAR	1	-0.00294	0.0111	0.0703	0.7909

Al correr la regresión con todas las variables incluyendo la β_0 se observa la prueba de Wald.

Estos datos nos permiten verificar las siguientes hipótesis:

H_0: $\beta_j=0$ Para todo *j*
H_1: $\beta_j=0$ para algún *j* donde *j = 1,…, k*

Como se vio con mas detalle en el capitulo 2.7.1, se observa que la variable SALAMINIM (Porcentaje de habitantes con menos de dos salarios mínimos) es significativa, así como, la variable PRIMARIA (Porcentaje de habitantes sin primaria concluida), ya que se puede rechazar la hipótesis nula con una confianza mayor del 90%.

Al correr la regresión sin el coeficiente β_0 se observan los siguientes resultados:

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
PRIMARIA	1	-0.3677	0.1846	3.9673	0.0464
DRENAJE	1	-0.0730	0.1036	0.4969	0.4809
AGUAENTUB	1	0.1376	0.1316	1.0928	0.2958
HACINAMIENTO	1	-0.0709	0.0906	0.6116	0.4342
MENHABIT	1	-0.0345	0.0635	0.2952	0.5869
SALARMINIM	1	0.2634	0.1130	5.4334	0.0198
INDIMAR	1	-0.6453	1.0718	0.3625	0.5471

Retirando las variables sin relevancia para el modelo, esto es, donde la prueba de Wald no se puede rechazar la hipótesis nula de que el estadístico es cero pues no se tiene suficientes pruebas estadísticas, se obtiene el siguiente resultado

Estos datos nos permiten verificar las siguientes hipótesis:

H_0: $\beta_j=0$ Para todo j
H_1: $\beta_j=0$ para algún j donde $j = 1,\ldots, k$

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
PRIMARIA	1	-0.3452	0.1508	5.2436	0.0220
SALARMINIM	1	0.1840	0.0797	5.3350	0.0209

El abstencionismo (ABS) como la variable dependiente y porcentaje de habitantes sin primaria terminada (PRIMARIA), porcentaje de habitantes que perciben menos de dos salarios mínimos (SALARMINIM) como las variables independientes. Las demás variables no son consideradas debido a que no son de relevancia en el modelo.

Con nuestra muestra de tamaño 32, los resultados que entrega SAS (versión 9.0) son:

Estadísticos de ajuste del modelo

Criterio	Sin variable adicional	Con variable adicional
AIC	44.361	39.875
SC	44.361	42.807
-2 LOG L	44.361	35.875

Debido a que el AIC ha disminuido de 44.361 a 39.875, además de que -2 LOG L de la máxima verosimilitud disminuyo de 44.361 a 35.875 incluyendo todas las variables en el modelo lo cual significa que la verosimilitud ha aumentado.

Prueba de la hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	8.4863	2	0.0144
Puntuación	6.7697	2	0.0339
Wald	5.3398	2	0.0693

Estos resultados sirven para verificar las siguientes hipótesis:

H_0: Todos los coeficientes son igual a cero
H_1: Por lo menos un coeficiente es distinto de cero

Lo que se interpreta que se puede rechazar la hipótesis nula de que los coeficientes son iguales a cero.

Así también otro estadístico para evaluar el ajuste global del modelo es χ^2 □ *cuadrada* de Person, la cual se basa en los residuos del modelo. Donde la ausencia de significancia indica que el ajuste del modelo es bueno, como se vio con mas detalle en el capitulo 2.7.3

Test chi-cuadrado residual

Chi-cuadrado	DF	Pr > ChiSq
4.0394	6	0.6713

Esto indica un buen ajuste del modelo.

Otra forma de evaluar el modelo, es a través de tablas de clasificación en la que es posible determinar las tasa global de clasificaciones correctas, la sensibilidad, la especificad, el valor predicativo positivo y el valor predicativo negativo.

Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje	74.9	D de Somers	0.498
Discordancia de porcentaje	25.1	Gamma	0.498
Porcentaje ligado	0.0	Tau-a	0.256
Pares	255	c	0.749

En le modelo estudiado, la tasa global de concordancia de predicciones es de 74.9%.

- Gamma mide la asociación entre Y_{pred} e Y_{obs}, un valor cercano a +1 ó a -1 indica una fuerte relación, por lo que se puede observar una relación moderada.
- Tau-a de Kendall, oscila entre 0 y 1donde los valores próximos a 1 indica fuerte relación, en este caso se distingue poca relación.
- D de Sommers es similar a la Gamma por lo que se concluye lo mismo que en el caso de la Gamma.

Procedimiento FREQ

Tabla de ABS por PREDICT

ABS	PREDICT		
Frequencia Porcentaje Pct fila Pct col	ABS	NOA	Total
ABS	11 34.38 64.71 68.75	6 18.75 35.29 37.50	17 53.13
NOA	5 15.63 33.33 31.25	10 31.25 66.67 62.50	15 46.88
Total	16 50.00	16 50.00	32 100.00

El porcentaje de abstencionismo con clasificación correcta del total de estados que se abstuvieron fue 68.75% y para los no abstencionistas fue de 62.5%.

Los estados mal clasificados como ABS fueron: Colima, Guanajuato, Nuevo León, Sonora y Tlaxcala.

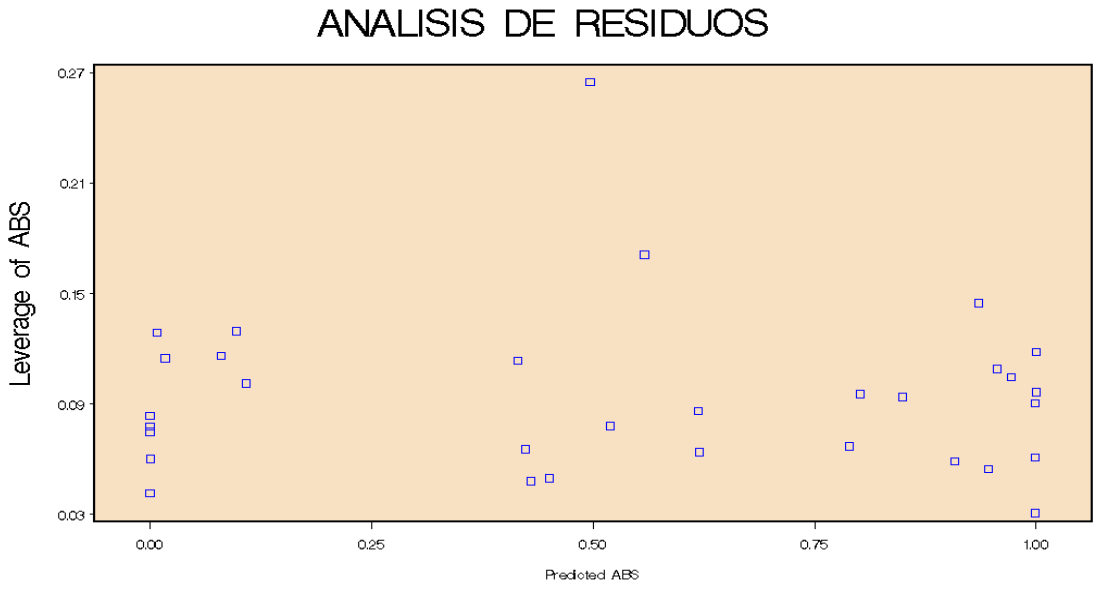
Mientras los estados mal clasificados como NOABS fueron: Baja California Norte, México, Nayarit, Oaxaca, Puebla y Yucatán.

Las causas de esto puede deberse a que los factores de marginación no son de influencia en el abstencionismo de dichos estados y que la población se vea motivada a votar o no por otras causas no estudiadas en este trabajo.

b) Análisis de residuos

Para lo que emplearemos la medida de influencia (O Leverage, estadístico h). Es una medida de como valores individuales pueden afectar a los resultados del modelo; tal que valores de probabilidad estimada por debajo de 0.1 o por encima de 0.9 es esperable que tengan siempre valores de influencia tendientes a 0, por lo que los valores entre 0.1 y 0.9 pueden dar una idea de distanciamiento o influencia, SAS nos permite guardar los valores de H (Vea Anexo), la cual nos muestra que las observación 7 con .2652 es la que presenta mayor probabilidad, esta observación pertenece al estado de Chiapas como se observa en la grafica (3.22)

Figura 3.22



c) Ajuste de regresión Logística retirando datos con residuales altos

Con una muestra de tamaño 31, retirando el dato del estado de Chiapas. Los resultados que entrega SAS (anexo) son:

Procedimiento LOGISTIC
Estadísticos de ajuste del modelo

Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	44.684	40.321
SC	46.118	44.623
-2 LOG L	42.684	34.321

Se observa que -2 LOG L de la máxima verosimilitud ha disminuido de 42.684 de la constante a 34.321 incluyendo todas las variables en el modelo lo cual significa que la verosimilitud ha aumentado, incluso mas que el modelo inicial era 35.814.
Y con los estadísticos de la chi-cuadrada con:

Prueba de la hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	8.3631	2	0.0153
Puntuación	6.7010	2	0.0351
Wald	5.2549	2	0.0723

Lo que quiere decir que se puede rechazar la hipótesis nula de que los coeficientes del modelo son iguales a cero, pero con menor fuerza que el modelo original.

Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje	75.2	D de Somers	0.508
Discordancia de porcentaje	24.4	Gamma	0.511
Porcentaje ligado	0.4	Tau-a	0.260
Pares	238	c	0.754

Procedimiento FREQ

Tabla de ABS por PREDICT

ABS	PREDICT		
Frequencia Porcentaje Pct fila Pct col	ABS	NOA	Total
ABS	12 38.71 70.59 66.67	5 16.13 29.41 38.46	17 54.84
NOA	6 19.35 42.86 33.33	8 25.81 57.14 61.54	14 45.16
Total	18 58.06	13 41.94	31 100.00

Mientras que la tasa global de concordancia de predicciones aumenta de 74.1% del modelo original a 75.2%. Así como los índices de concordancia de porcentaje incrementan en general.

Pero esto no indica que sea un mejor modelo, ya que al retirar la observación 7, se clasifican mejor ya que esta pertenecía a un grupo mal clasificado ya que su dato observado es ABS, mientras el predicho fue NOABS, por sus altos porcentajes en las variables observadas.

Después del análisis anterior el modelo 2 a pesar de tener un incremento en los índices y la tasa global de concordancia de predicciones, no es de gran trascendencia, así como la verosimilitud ya que es menor que el modelo inicial, y su χ *cuadrada* rechaza con menos fuerza que los coeficientes sean cero, por lo tanto el modelo inicial esta mejor ajustado a los datos.

3.4 Interpretación de los Coeficientes del Modelo

$\beta_1 = -0.3452$ Es el coeficiente de PRIMARIA y significa el cambio en unidades de la razón de probabilidades de incidencia y no incidencia, cuando todas las variables regresoras excepto PRIMARIA, permanezcan invariantes. Dado que es negativo quiere decir que su transformación (antilog) será un valor menor que 1, entonces el odds disminuirá, por lo que la probabilidad de no abstencionismo es más alta que la de abstencionismo, esto es, que en este modelo los valores altos de estados con porcentajes de personas sin primaria completa disminuirá la probabilidad de abstenerse.

$\beta_2 = .1840$ Es el coeficiente de SALARMINIM y significa el cambio en unidades de la razón de probabilidades de incidencia y no incidencia, cuando todas las variable regresoras excepto SALARMINIM, permanezcan invariantes. Puesto que es un coeficiente positivo quiere decir que su transformación (antilog) será un valor mayor que 1 por lo que el odds aumentará, por lo tanto la probabilidad de abstenerse es más alta que la de no abstenerse, esto es, que en el modelo valores altos de estados con porcentajes elevados de personas que ganan hasta dos salarios mínimos, la probabilidad de que se abstengan aumentará

Capitulo cuarto
Conclusiones y Observaciones

4.1 Conclusiones

Para observar el fenómeno del abstencionismo se analizó a la población más marginada de México, ésta es, la más segregada de la sociedad en todos los aspectos, desde educativa hasta económicamente.

En este análisis se puede concluir que las variables socioeconómicas que miden el grado de marginación de la población, son un factor de riesgo que influye de manera significativa sobre el fenómeno del abstencionismo. Estas variables consideradas factor de riesgo fueron: poblaciones sin primaria concluida, así como, poblaciones que ganan hasta dos salarios mínimos. Con una tasa global correcta de clasificación del 74.9%.

Observando que poblaciones con un alto porcentaje de personas que no concluyeron sus estudios de primaria, implicó que la probabilidad de abstención disminuyera, pues este sector es mas vulnerable a ser manipulado.

Mientras que estados con poblaciones con altos porcentajes de personas que tienen como percepción máxima hasta dos salarios mínimos mensuales, aumenta la probabilidad de que el estado obtenga altos porcentajes de abstencionismo.

Ya que el abstencionismo se plantea como la no participación de los ciudadanos en la vida política. Ésta no participación responde a factores que provocan la indiferencia de la sociedad que ejerce su derecho a no participar dentro de un sistema que no responde á sus necesidades de proveerles un bienestar social con servicios de primera necesidad y que los mantiene excluida de la misma, ya que al percibir los ingresos más bajos, no tienen acceso a muchos de estos servicios.

Por lo que los mecanismos de elección de sus representantes son irrelevantes para su vida cotidiana ya que no existe una fuerza política que represente sus intereses, lo que provoca un alto grado de despolitización, debido a su falta de inclusión en el sistema.

4.2 Observaciones

El método de regresión logística es una herramienta de gran ayuda en el análisis de la geopolítica nacional, la cual puede ser también empleada para el análisis de otros fenómenos sociales, la implementación un método matemático en un problema social especifico que ayuda a vislumbrar de manera más precisa los motivos de este fenómeno.

El estudio se puede realizar de manera más detallada por municipios, debido a que esta información es muy amplia dado que existen estados que poseen más de trescientos municipios como el caso de Oaxaca, así como, la organización de los distritos a los que pertenece cada municipio para la emisión del voto de los habitantes. Esto analizaría de manera precisa el comportamiento de la población por estado, identificando de forma particular la problemática, puesto que en el modelo existieron once estados clasificados de manera incorrecta, lo que podría ayudar a encontrar otros factores no contemplados o ampliarlos.

En general, el abstencionismo es un fenómeno social que debe ser tomado con mas seriedad en su estudio, pues el gran gasto empleado en publicidad para las campañas políticas, puede en realidad no ser de relevancia para que la gente acuda a las urnas, sino mas bien se debe a factores ajenos a los medios de comunicación y que son parte de la vida cotidiana de la gente, como fue el caso de los indicadores de marginación.

La CONAPO contempló en el año 2005, doce indicadores de marginación, pero el presente trabajo solo contempló once, debido a que el indicador Grado de Marginación se explicada ya por la variable Índice de Marginación, además de que la variable Grado de Marginación es de tipo categórica, requería de 4 variables dummy, provocando problemas en el cálculo del máximo verosímil, lo que ocasionaba que la validación del modelo fuera cuestionable. Por lo que está se retiro del estudio, además de encontrarse estrechamente correlacionada con el resto de las variables.

Los datos del abstencionismo no toman en cuenta los votos emitidos en el extranjero.

Anexos

Anexo 1

Los datos producidos por el procedimiento PRINT se muestran continuación:

Analisis Discriminante Usando Regresion Logistica

Obs	ABS	POBTOTAL	ANALFABETA	PRIMARIA	DRENAJE	ENERGIA	AGUANTUB	HACINAMIENTO	PISOTIERRA	MENHABIT	SALARMIN	INDIMAR	GRADMARG
1	abs	1.030	4.16	17.82	1.68	0.85	1.79	34.34	2.36	24.67	32.79	-0.954	Bajo
2	noabs	2.750	3.08	15.02	0.56	1.49	4.89	32.26	3.95	9.27	14.24	-125.336	Muy
3	noabs	0.495	3.62	16.49	1.84	2.88	11.28	35.16	8.27	19.84	24.07	-0.719	Bajo
4	abs	0.730	10.20	26.96	9.85	4.85	11.15	51.42	8.77	32.01	55.10	0.559	Alto
5	noabs	2.420	3.29	14.60	1.65	0.77	2.17	33.71	2.51	12.33	32.97	-113.709	Muy
6	abs	0.550	6.42	21.58	0.80	0.67	1.66	34.92	8.40	15.78	42.66	-0.738	Bajo
7	noabs	4.160	21.35	42.76	8.07	5.88	25.90	60.20	32.99	58.46	78.14	232.646	Muy
8	noabs	3.140	4.42	18.81	3.29	4.28	6.45	31.93	6.45	17.54	27.25	-0.684	Bajo
9	abs	8.450	2.59	9.70	0.16	0.15	1.51	29.31	1.12	0.52	33.04	-150.487	Muy
10	noabs	1.460	4.84	22.92	8.51	3.52	8.63	36.09	11.34	38.64	50.04	-0.019	Medio
11	noabs	4.740	10.44	28.80	9.87	1.93	6.14	40.99	8.95	34.46	44.69	0.092	Medio
12	noabs	3.020	19.88	35.98	27.18	6.33	31.34	55.06	35.69	50.51	64.97	241.213	Muy
13	abs	2.270	12.80	27.50	8.98	3.90	12.21	42.69	12.78	57.28	61.63	0.751	Alto
14	abs	6.540	5.56	21.30	2.42	1.12	5.91	33.27	5.35	17.40	34.74	-0.769	Bajo
15	abs	13.560	5.32	16.24	4.76	0.96	6.04	41.46	6.04	18.63	41.18	-0.622	Bajo
16	noabs	3.840	12.58	33.48	5.66	2.11	9.97	40.01	15.67	40.51	55.79	0.457	Alto
17	abs	1.560	8.13	21.01	3.10	0.81	7.84	38.63	10.71	22.88	37.77	-0.443	Bajo
18	noabs	0.920	8.02	26.05	6.78	4.38	8.35	37.73	9.93	41.67	51.73	0.191	Medio
19	abs	4.070	2.78	12.70	0.54	0.56	3.48	33.49	2.34	7.13	23.55	-132.611	Muy
20	abs	3.400	19.35	38.49	6.84	7.21	26.29	53.06	35.17	61.27	69.65	212.936	Muy
21	abs	5.210	12.71	29.02	5.45	2.19	14.03	49.16	15.86	39.00	61.34	0.635	Alto
22	abs	1.550	8.14	20.03	9.95	2.99	5.76	37.60	8.59	38.34	39.23	-0.142	Medio
23	noabs	1.100	6.58	19.42	5.19	2.59	4.66	49.90	8.41	17.10	36.78	-0.316	Bajo
24	abs	2.330	9.92	27.42	5.72	5.58	16.97	38.60	19.67	41.18	56.11	0.656	Alto
25	noabs	2.530	6.42	23.42	5.14	1.92	6.24	43.08	9.78	35.30	44.95	-0.148	Medio
26	noabs	2.320	3.73	17.21	1.92	1.87	4.01	38.73	9.67	18.29	31.76	-0.750	Bajo
27	abs	1.930	8.57	25.10	3.99	1.95	22.94	47.11	9.11	55.78	51.97	0.462	Alto

28	noabs	2.930	4.52	18.61	0.84	2.88	4.26	39.24	5.39	14.92	37.56	-0.683	Bajo
29	abs	1.030	6.68	18.78	4.84	1.11	2.03	47.93	6.26	40.30	62.59	-0.129	Medio
30	abs	6.890	13.42	32.90	4.18	4.67	23.32	45.02	22.77	46.68	58.36	107.674	Alto
31	abs	1.760	10.89	29.99	17.96	2.61	3.03	48.36	4.81	27.42	63.00	0.431	Alto
32	noabs	1.320	7.20	30.83	10.53	1.91	6.72	37.06	6.29	50.59	54.25	0.160	Medio

Anexo 2

A continuación se muestran las salidas de SAS del procedimiento LOGISTIC, con el intercepto:

Analisis Discriminante Usando Regresion Logistica

Procedimiento LOGISTIC

Información del modelo

Conjunto de datos	WORK.ABSTENCIONISMO
Variable de respuesta	ABS
Número de niveles de respuesta	2
Número de observaciones	32
Modelo	logit binario
Técnica de optimización	Puntuación de Fisher

Perfil de respuesta

Valor ordenado	ABS	Frecuencia total
1	abs	17
2	noabs	15

La probabilidad modelada es ABS='abs'.

Estado de convergencia del modelo

Convergence criterion (GCONV=1E-8) satisfied.

Estadísticos de ajuste del modelo

Criterio	Sólo términos independientes	Términos independientes y Variables adicionales
AIC	46.236	49.407
SC	47.702	61.133
-2 LOG L	44.236	33.407

Prueba de la hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	10.8291	7	0.1463
Puntuación	8.8668	7	0.2624
Wald	6.4088	7	0.4929

Analisis Discriminante Usando Regresion Logistica 41
18:21 Thursday, September 16, 2008

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	0.8526	4.0440	0.0445	0.8330
PRIMARIA	1	-0.3630	0.2059	3.1082	0.0779
DRENAJE	1	-0.0880	0.0972	0.8203	0.3651
AGUAENTUB	1	0.1174	0.1155	1.0336	0.3093
HACINAMIENTO	1	-0.0658	0.1147	0.3285	0.5665
MENHABIT	1	-0.0374	0.0650	0.3310	0.5651
SALARMINIM	1	0.2463	0.1106	4.9617	0.0259
INDIMAR	1	-0.00294	0.0111	0.0703	0.7909

Estimadores de cocientes de disparidad;

Efecto	Estimador del punto	95% Wald Límites de confianza	
PRIMARIA	0.696	0.465	1.041
DRENAJE	0.916	0.757	1.108
AGUAENTUB	1.125	0.897	1.410
HACINAMIENTO	0.936	0.748	1.172
MENHABIT	0.963	0.848	1.094
SALARMINIM	1.279	1.030	1.589
INDIMAR	0.997	0.976	1.019

Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje	81.6	D de Somers	0.631
Discordancia de porcentaje	18.4	Gamma	0.631
Porcentaje ligado	0.0	Tau-a	0.325
Pares	255	c	0.816

A continuación se muestran las salidas de SAS del procedimiento LOGISTIC sin el intercepto:

LOG ODSS DE ABS POR ANALFABETA

Procedimiento LOGISTIC

Información del modelo

Conjunto de datos	_PROJ_.PREDIT
Variable de respuesta	ABS
Número de niveles de respuesta	2
Número de observaciones	32
Modelo	logit binario
Técnica de optimización	Puntuación de Fisher

Información del modelo

Valores predichos y estadísticos de diagnóstico
ABS

Perfil de respuesta

Valor ordenado	ABS	Frecuencia total
1	abs	17
2	noabs	15

La probabilidad modelada es ABS='abs'.

Estado de convergencia del modelo

Convergence criterion (GCONV=1E-8) satisfied.

Estadísticos de ajuste del modelo

Criterio	Sin variable adicional	Con variable adicional
AIC	44.361	47.322
SC	44.361	57.582
-2 LOG L	44.361	33.322

Prueba de la hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	11.0391	7	0.1369
Puntuación	8.8773	7	0.2616
Wald	6.2876	7	0.5066

LOG ODSS DE ABS POR ANALFABETA

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
PRIMARIA	1	-0.3677	0.1846	3.9673	0.0464
DRENAJE	1	-0.0730	0.1036	0.4969	0.4809
AGUAENTUB	1	0.1376	0.1316	1.0928	0.2958
HACINAMIENTO	1	-0.0709	0.0906	0.6116	0.4342
MENHABIT	1	-0.0345	0.0635	0.2952	0.5869
SALARMINIM	1	0.2634	0.1130	5.4334	0.0198
INDIMAR	1	-0.6453	1.0718	0.3625	0.5471

Estimadores de cocientes de disparidad;

Efecto	Estimador del punto	95% Wald Límites de confianza	
PRIMARIA	0.692	0.482	0.994
DRENAJE	0.930	0.759	1.139
AGUAENTUB	1.147	0.887	1.485
HACINAMIENTO	0.932	0.780	1.113
MENHABIT	0.966	0.853	1.094
SALARMINIM	1.301	1.043	1.624
INDIMAR	0.524	0.064	4.286

Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje	81.6	D de Somers	0.631
Discordancia de porcentaje	18.4	Gamma	0.631
Porcentaje ligado	0.0	Tau-a	0.325
Pares	255	c	0.816

Anexo 4

A continuación se muestran las salidas de SAS del procedimiento LOGISTIC sin intercepto:

LOG ODSS DE ABS POR ANALFABETA

Procedimiento LOGISTIC

Información del modelo

Conjunto de datos	_PROJ_.PREDIT
Variable de respuesta	ABS
Número de niveles de respuesta	2
Número de observaciones	32
Modelo	logit binario
Técnica de optimización	Puntuación de Fisher

Información del modelo

Valores predichos y estadísticos de diagnóstico
ABS

Perfil de respuesta

Valor ordenado	ABS	Frecuencia total
1	abs	17
2	noabs	15

La probabilidad modelada es ABS='abs'.

Estado de convergencia del modelo

Convergence criterion (GCONV=1E-8) satisfied.

Estadísticos de ajuste del modelo

Criterio	Sin variable adicional	Con variable adicional

AIC	44.361	47.322
SC	44.361	57.582
-2 LOG L	44.361	33.322

Prueba de la hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	11.0391	7	0.1369
Puntuación	8.8773	7	0.2616
Wald	6.2876	7	0.5066

LOG ODSS DE ABS POR ANALFABETA

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
PRIMARIA	1	-0.3677	0.1846	3.9673	0.0464
DRENAJE	1	-0.0730	0.1036	0.4969	0.4809
AGUAENTUB	1	0.1376	0.1316	1.0928	0.2958
HACINAMIENTO	1	-0.0709	0.0906	0.6116	0.4342
MENHABIT	1	-0.0345	0.0635	0.2952	0.5869
SALARMINIM	1	0.2634	0.1130	5.4334	0.0198
INDIMAR	1	-0.6453	1.0718	0.3625	0.5471

Estimadores de cocientes de disparidad;

Efecto	Estimador del punto	95% Wald Límites de confianza	
PRIMARIA	0.692	0.482	0.994
DRENAJE	0.930	0.759	1.139
AGUAENTUB	1.147	0.887	1.485
HACINAMIENTO	0.932	0.780	1.113
MENHABIT	0.966	0.853	1.094
SALARMINIM	1.301	1.043	1.624
INDIMAR	0.524	0.064	4.286

Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje	81.6	D de Somers	0.631
Discordancia de porcentaje	18.4	Gamma	0.631
Porcentaje ligado	0.0	Tau-a	0.325
Pares	255	c	0.816

Anexo 5

A continuación se muestran las salidas de SAS del procedimiento LOGISTIC con PRIMARIA y SALARMINIM:

LOG ODSS DE ABS POR ANALFABETA

Procedimiento LOGISTIC

Información del modelo

Conjunto de datos	_PROJ_.PREDIT
Variable de respuesta	ABS
Número de niveles de respuesta	2
Número de observaciones	32
Modelo	logit binario
Técnica de optimización	Puntuación de Fisher

Información del modelo

Valores predichos y estadísticos de diagnóstico
ABS

Perfil de respuesta

Valor ordenado	ABS	Frecuencia total
1	abs	17
2	noabs	15

La probabilidad modelada es ABS='abs'.

Estado de convergencia del modelo

Convergence criterion (GCONV=1E-8) satisfied.

Estadísticos de ajuste del modelo

	Sin variable	Con variable

Criterio	adicional	adicional
AIC	44.361	39.875
SC	44.361	42.807
-2 LOG L	44.361	35.875

Prueba de la hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	8.4863	2	0.0144
Puntuación	6.7697	2	0.0339
Wald	5.3398	2	0.0693

LOG ODSS DE ABS POR ANALFABETA

Procedimiento LOGISTIC

Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
PRIMARIA	1	-0.3452	0.1508	5.2436	0.0220
SALARMINIM	1	0.1840	0.0797	5.3350	0.0209

Estimadores de cocientes de disparidad;

Efecto	Estimador del punto	95% Wald Límites de confianza	
PRIMARIA	0.708	0.527	0.951
SALARMINIM	1.202	1.028	1.405

Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje	74.9	D de Somers	0.498
Discordancia de porcentaje	25.1	Gamma	0.498
Porcentaje ligado	0.0	Tau-a	0.256
Pares	255	c	0.749

Anexo 6

Método condicional e incondicional:

Como se menciona en el Capitulo 2 los programas empleados para el método condicional e incondicional son distintos, a continuación se mencionan, así como, en que situación se recomienda implementar uno u otro:

Incondicional:

- SAS (LOGIST)
- BMDP
- GLIM
- SPSS
- EGRET
- SPISA
- S+

Es Preferible emplearlos cuando el número de variables es menor al número de observaciones bajo estudio.

Condicional:

- SAS (PECAN)
- SAS (PHREG)
- EGRET
- SPIDA
- S+

Este es preferible emplearlo si el número de variables es mayor al número de observaciones en el estudio.

BIBLIOGRAFIA

Hosmer & Lemanshow, "Applied Logistic Regression". 2nd Edition. Wiley Intercience, 2000.

Dallas E. Johnson. "Métodos multivariados aplicados al análisis de datos". Internacional Thomson Edition, 2000.

Paul D. Allison. "Logistic Regression using SAS system: theory and application".Cary North Carolina: SAS institute, 1999.

Asha Seth kapadia, Wenyaw Chan, Lemuel Moye. "Mathematical Statistic whit application". Chapman and Hall/CRC, Taylor and Francis Group, 2005.

David G. Kleimbaum; whit contributions by Erica Rihl Pryor. "Logistic Regression: A self-learning text". 2nd Edition. New York: Springer, Series Statistics for Biology and health, 2002.

Cook, D.R. & Weisberg S. Residual and Influence in Regression. Chapman Hall. London, 1982.

Y. Sakamoto, M. Ishiguro and G. Kitagawa. "Akaike Information Criterion Satatistics". D. Reidel Publishimg Company, Series Mathematics and its applications, 1986.

Yudi Pawitan. "In All Likelihood: Stadistical Modelling and Inference Using Likelihood", Oxford science publications, 2001.

Celia Mercedes Salcedo Poma. "Estimación de la Ocurrencia de Incidencias en declaraciones de Pólizas de Importación", Tesis digital de la Universidad Nacional Mayor de San Marcos [Citado: 23/07/2007] Disponible en: http://sisbib.unmsm.edu.pe/BibVirtual/bibvirtual.asp

Carlos Sirvent, "Alternancia y distribución del voto en México Estudio de 7 casos", Editorial GERNIKA, Ciudad de México, 2001

Raúl Zibechi, "Genealogía de la Revuelta; Argentina: La sociedad en movimiento", Ediciones FZLN, Ciudad de México 2004

Carlos Antonio Aguirre Rojas, "Para comprender el mundo actual, una gramática de larga duración", Centro de Investigación y Desarrollo de la Cultura Cubana Juan Marinello, Editorial Linotipia Bolívar, Colombia 2003

Edgar Acuña Fernández, "Análisis de Regresión", Departamento de Matemáticas Universidad de Puerto Rico, 2003.

Modelo de regresión logística incondicional. Publicación electrónica de la Sociedad Andaluza de Enfermedades Infecciosas [Citado: 25/10/07]. Disponible en: http://saei.org/hemero/epidemiol/nota4.html

Modelo de regresión logística. Publicación electrónica del Hospital Universitario Ramón y Cajal, Comunidad de Madrid [Citado: 27/10/07]. Disponible en: http://www.hrc.es/bioest/M_docente.html#tema9

El abstencionismo electoral. Publicación electrónica del Instituto Nacional de Estudios Políticos AC. [Citado: 06/11/2007]. Disponible en: http://www.inep.org/content/view/524/120/

Abstencionismo en México. Publicación electrónica de Fuerza Ciudadana AC. [Citado: 06/11/2007]. Disponible en: http://www.fuerzaciudadana.org.mx/materiales/abstencionismo.pdf

Encuesta Nacional sobre Cultura Política y Prácticas Ciudadanas. Publicaciones de la Secretaria de Gobernación. [Citado: 07/11/2007]. Disponible en: http://www.gobernacion.gob.mx/encup/

Estado de Derecho, Plan Nacional de Desarrollo. Publicación electrónica de la Secretaria de Gobernación. [Citado: 07/11/2007]. Disponible en: http://pnd.calderon.presidencia.gob.mx/index.php?page=certezajuridica

Elecciones Federales 2006. Publicación electrónica de Instituto Federal Electoral. [Citado: 07/11/2007]. Disponible en: http://www.ife.org.mx/portal/site/ife/menuitem.16a169478ded8d1df997c170241000a0/

Índices de Marginación. Publicaciones electrónicas del Consejo Nacional de Población. [Citado: 04/06/2007]. Disponible en: http://www.conapo.gob.mx/publicaciones/margina2005/AnexoA.pdf

Alberto Aziz Nacif, Nuevo mapa electoral. Publicación electrónica de el periódico El Universal, Editoriales. [Citado: 09/11/2007]. Disponible en: http://www.el-universal.com.mx/editoriales/34800.html