13

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES "ACATLÁN"

288780

ANÁLISIS DE REGRESIÓN LOGÍSTICA PARA EXPLICAR LA RELACIÓN ENTRE LOS NIVELES DE METILACIÓN Y LA PRESENCIA DE LESIONES ESPECÍFICAS POR INGESTIÓN CRÓNICA DE ARSÉNICO.

T E S I S

QUE PARA OBTENER EL TITULO DE:

L I C E N C I A D O E N

MATEMATICAS APLICADAS Y COMPUTACION

P R E S E N T A :

JAVIER DE JESÚS FONSECA MADRIGAL

ASESOR: DOCTORA SILVIA RUIZ VELAZCO

FEBRERO 2001

El presente trabajo se realizó en el laboratorio de metales de la sección de Toxicología Ambiental del Centro de Investigaciones y Estudios Avanzados del IPN, bajo la dirección de la Doctora Luz María Del Razo Jiménez en lo referente a toxicología, y de la Doctora Silvia Ruiz Velazco Acosta en lo concerniente a estadística.

AGRADECIMIENTOS

A mi padre, por brindarme su valioso apoyo

A mi madre, a quien debo el gusto por las matemáticas

A mi hermana que me ha enseñado lo que es la perseverancia

A la Doctora Luz María Del Razo Jiménez, por su paciencia y dedicación

A la Doctora Silvia Ruiz Velasco Acosta, por su certera orientación

Al Doctor Mariano Cebrián, por el respaldo que me brindó

Al Doctor Gonzalo García, por sus interesantes observaciones

A mis profesores, por sus valiosas enseñanzas

A la UNAM y a la ENEP Acatlán, por ofrecer esta interesante carrera

A mi tía Pilos y mi tío Fernando, por su valioso ejemplo y enseñanzas

A mis amigos, con quienes he crecido y compartido una buena parte de mi vida

A Diana González, por su apoyo al término de esta tesis

INDICE

INTRODUCCION

La ciencia de las Matemáticas, desde los inicios de las culturas del hombre, ha evolucionado de forma tan consistente que lejos de caer en obsolescencia ha crecido sobre sus propios principios fundamentales que, como el teorema de Pitágoras, hoy día son aplicados en técnicas recientes muy poderosas para la solución de problemas. Tal vez esto se deba a que las matemáticas no son sino un código que representa algunos de los fundamentos o principios del entorno o naturaleza que rodea al ser humano.

Una gran parte de los últimos avances en matemáticas se debe a la posibilidad de hacer cálculos en computadora que en antaño se llevarían la vida entera de una persona, suponiendo que no se cometieran errores. Como ejemplo de lo anterior están los fractales, la teoría del caos, los métodos numéricos, la simulación, diversas técnicas de análisis multivariado, etc. cuyas aplicaciones impactan sobre prácticamente todas las áreas del conocimiento; investigación, medicina, genética, biología, economía, planeación, audio, ciencias sociales, administración, etc.

El presente trabajo muestra precisamente una aplicación de las matemáticas a un proyecto de investigación en el campo de la toxicología ambiental, mediante el uso de regresión logística, técnica que para estimar sus parámetros debe recurrir a la aproximación por métodos numéricos para resolver un sistema de ecuaciones no lineales que carece de solución analítica, lo cual sería imposible a no ser por el uso de las computadoras.

El contenido del presente trabajo está organizado en tres capítulos, siendo el primero una contextualización al tema de toxicología ambiental que se aborda, el segundo una explicación de los fundamentos matemáticos que sustentan a los modelos de regresión logística, así como de un estadístico muy común en biología conocido como razón de momios y cuya estimación es una de las aplicaciones de la regresión logística, y por último en el capítulo tres se entra de lleno a la estimación de razones de momios para dar respuesta a la hipótesis que se plantea sobre la posible relación entre un proceso de desintoxicación del organismo, conocido como metilación, y la presencia de lesiones por ingestión crónica de arsénico.

LISTA DE ABREVIATURAS

Asi	Arsénico inorgánico
CINVESTAV	Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional
DMA	Arsénico dimetilado
MMA	Arsénico monometilado
MMAIII	Arsénico monometilado trivalente
ppm	Partículas por millón
µg	Micro gramos
µg/l	Micro gramos por litro

CAPÍTULO I
EL ARSENICISMO

1.1 Generalidades

El arsénico es un elemento químico del grupo VA de la tabla periódica, cuyo número atómico es 33. Etimológicamente deriva de la raíz griega arsenikos que significa fuerte, vigoroso. Es un metaloide porque presenta casi todas las características de un metal. Su aspecto físico es sólido de color gris plateado, quebradizo y cristalino que se opaca en el aire húmedo. Se sublima a 613°C, es insoluble en agua y es semiconductor. En estado puro no es tóxico, pero si se encuentra formando compuestos puede ser altamente tóxico.

1.2 Ubicación y distribución geográfica

Al arsénico se le encuentra en el medio ambiente, así como en los volcanes y minas. Es un elemento relativamente movible pues existe en estados gaseosos, soluciones líquidas y sólidos. Puede ser transportado por la acción de la naturaleza o del ser humano.

Actualmente las prácticas humanas han modificado el ciclo global de este metaloide; en los mares ha crecido la erosión debido a los cambios agrícolas, en la tierra se han generado emisiones de carbón y petróleo que contienen arsénico, en la manufactura del cemento y el tueste de minerales de sulfuro también se libera arsénico.

1.2.1 Distribución natural

1.2.1.1 Corteza terrestre y rocas.

El arsénico ocupa el vigésimo lugar de abundancia elemental en la corteza terrestre. Es el elemento constituyente mayor de al menos 245 minerales diferentes. Se le encuentra más frecuentemente en asociación con sulfuros como elemento nativo o formando otros compuestos. El mineral más común es la arsenopirita (FeAsS). Los sulfuros y sales sulfúricas que presentan arsénico se oxidan fácilmente cuando se exponen al aire para producir trióxido de arsénico (arsénico trivalente) y finalmente arseniato (arsénico pentavalente).

1.2.1.2 Suelo.

El arsénico está presente de manera natural en el material rocoso que forma todos los suelos. La cantidad promedio de arsénico varía desde 5 a 6 partículas por millón (ppm), aunque existen suelos con capas de depósitos de minerales sulfúricos, los cuales contienen arsénico en varios cientos de ppm.

El arsénico también puede estar ligado a la materia orgánica en los suelos, en cuyo caso es liberado y queda disponible para ser absorbido por las plantas. La cantidad liberada para la absorción de las plantas depende de las formas químicas y físicas de los compuestos arsenicales.

En los suelos vírgenes el arsénico disponible está presente en cantidades bajas, en promedio es diez veces menor que el arsénico de la mayoría de los suelos cultivados.

1.2.1.3 Agua.

Toda el agua contiene arsénico, la cantidad contenida en lagos, ríos, nacimientos de agua y manantiales varía considerablemente. La mayor parte del arsénico está ahí de manera natural, aunque alguna porción pudo haber sido agregada a través del uso de plaguicidas en los lagos. Aunque el tratamiento de los desperdicios humanos y sus resultantes descargas a los sistemas acuíferos añaden arsénico a los ríos, la mayor parte de este elemento proviene por lixiviación[1] y por erosión de suelos, que transfieren 612 y 2,380 x 10^6 gramos por año a los océanos en forma disuelta o suspendida respectivamente.

En algunos manantiales de aguas termales sobresale el alto contenido de arsénico. Algunas aguas subterráneas de actividad térmica contienen concentraciones extremadamente altas, tal es el caso de los **pozos** perforados en áreas rocosas con alto contenido de arsénico y de aguas con gran cantidad de sales disueltas. Fuera de estos dos casos, es posible que la mayor parte de los otros valores altos reportados en ríos y lagos se deba a la contaminación industrial. Se ha asumido que las aguas de la superficie, como el océano, se purifican a si mismas con respecto al arsénico, puesto que éste se remueve de la solución depositándose en sedimentos, quienes contienen mayor cantidad de arsénico que el agua con la cual están asociados.

[1] Disolución de alguna sustancia en agua.

1.2.1.3 Alimentos.

Los niveles más altos de arsénico se encuentran en los alimentos de origen marino, principalmente en los peces y crustáceos. Estos alimentos son considerados de baja toxicidad, ya que el arsénico se encuentra como compuesto orgánico[2].

1.2.2 Distribución antropogénica.

El hombre, al utilizar recursos naturales relacionados con el arsénico como la minería, la combustión de carbón, o la combustión de gasolinas fósiles, libera arsénico en el aire, agua y tierra, el cual regresa a la tierra u océano en forma de polvo o por precipitación.

1.2.2.1 Fundidoras.

La mayor parte del arsénico producido para fines comerciales se obtiene de los residuos ricos en dicho elemento depositados en la lumbrera de escape de las fundidoras de los minerales de plomo, cobre y oro con la disipación de los gases. Estos gases transportan sustancias valiosas o peligrosas, muchas de las cuales son emitidas como polvos que contienen trióxido de arsénico, partículas de metales y de óxido de éstos. El trióxido de arsénico es volátil, y casi todo es expulsado del mineral por sublimación durante la fundición.

La emisión de gases volátiles de arsénico causada por las fundidoras y minas elevan los niveles atmosféricos de arsénico en las áreas vecinas en forma inversamente proporcional a la distancia; la cantidad emitida depende de la adecuación y funcionamiento de los sistemas colectores de polvos. El

[2] Del Razo 1997, pg. 3

arsénico en la atmósfera es removido por precipitación, de tal forma que los niveles atmosféricos generalmente no incrementan, lo que sí sucede en los suelos vecinos a la fundidora.

1.2.2.2 Combustión.

Existen dos fuentes de arsénico por combustión; el carbón y los derivados del petróleo.

Las diferentes formas de arsénico en el carbón tienen diferentes volatilidades cuando se queman. Una pequeña fracción del arsénico volátil en el carbón escapa del equipo colector de polvo y alcanza la atmósfera, este factor podría ser aplicado a la industria de amplio uso de carbón. En lo concerniente a los derivados del petróleo, el arsénico se puede eliminar mediante un proceso de reducción.

1.2.2.3 Plaguicidas.

Una de las clases de agentes usados para el biocontrol son los pesticidas arsenicales; se les ha utilizado principalmente como insecticidas, herbicidas, defoliantes[3], conservadores de madera, preservativos, etc. Dependiendo de la forma como se aplique, el arsénico puede expandirse en el aire, suelo o ser absorbido por algún organismo, alterando los niveles naturales en aire y suelo. En el caso de la absorción, el arsénico puede trasladarse a otras partes, dependiendo del organismo que lo haya absorbido, o bien regresar al aire o suelo; por ejemplo, el arsénico aplicado a la madera se deposita en el suelo por medio de lixiviación.

[3] Un agente defoliante es aquél cuyo efecto provoca la caída prematura de las hojas.

1.2.2.3 Uso latrogénico pasado y presente

Hace casi doscientos años atrás, la solución de Fowler, la cual contenía cerca de 1% de trióxido de arsénico, era recomendada para curar la fiebre palúdica. Posteriormente se le usó para el tratamiento de leucemia mielocítica crónica[4], psoriasis[5] y asma bronquial. Esta solución era altamente efectiva, sin embargo, si se suspendía su consumo se recaía en la enfermedad. Los síntomas en casos suaves de intoxicación consisten en malestares gastrointestinales o dolor, y en casos más severos, vómito y diarrea. Actualmente se ha retomado su uso en la medicina y se esta usando exitosamente el trióxido de arsénico para combatir la leucemia aguda promielocitica[6].

Varios compuestos orgánicos de arsénico han sido utilizados con fines medicinales; la arsfenamina y neoarsfenamina[7] fueron usadas en el tratamiento de la sífilis. El atoxyl, carbarson glycobiarsol, melarsoprol y triparsamida aún son usados como antiparasitadores. Los efectos nocivos que se han reportado atribuidos a estos compuestos son encefalopatía[8] y atrofia óptica.

[4] Es un tipo de cáncer en los glóbulos blancos
[5] Afección de la piel caracterizada por la aparición de escamas que se levantan fácilmente por el rascado y dejan debajo de ellas una superficie roja que sangra fácilmente
[6] National Research Council (NRC). 1999
[7] Son medicamentos antimicrobianos de origen químico
[8] Conjunto de trastornos cerebrales

1.3 Cinética del arsénico en el organismo

La toxicocinética es la rama de la toxicología, que se encarga del estudio de la distribución de un tóxico a través del organismo. En la presente sección se hará una breve descripción de la toxicocinética del arsénico.

1.3.1 Absorción.

El arsénico puede entrar en contacto con algún organismo por dos vías principalmente: absorción respiratoria y gastrointestinal, y de menor importancia por vía dérmica. Recientemente se está empleando la vía intravenosa para tratamientos iatrogénicos (tratamientos curativos).

1.3.1.1 Absorción respiratoria.

El depósito del arsénico en el sistema respiratorio y su absorción, depende del tamaño de las partículas inhaladas y de su forma química; las partículas pequeñas son absorbidas directamente por los pulmones, mientras que las más grandes son depositadas en los conductos superiores del tracto respiratorio, de donde pasan por la acción ciliar al tracto gastrointestinal, donde son absorbidas de acuerdo a su solubilidad en el jugo gástrico.

1.3.1.2 Absorción gastrointestinal.

La absorción del arsénico inorgánico por esta vía dependerá en gran parte de la solubilidad de los compuestos arsenicales. Por lo general, "más del 90% del arsénico inorgánico dado como solución en agua es absorbido por el tracto gastrointestinal".

1.3.2 Distribución y retención en los tejidos

El arsénico absorbido por el tracto gastrointestinal o por los pulmones, es transportado por la sangre a diferentes órganos en el cuerpo, quienes posteriormente lo regresan a la sangre o bien lo eliminan directamente. Una proporción mayor al 90% de arsénico se limpia de la sangre en un tiempo medio de 1 a 2 horas, el arsénico restante disminuye a una razón mucho menor. Se ha estimado que los tiempos medios para una segunda y tercera fase son de 20 y 200 horas respectivamente.

1.3.2.1 Órganos principales en la distribución del arsénico.

Los órganos principales por los cuales se distribuye el arsénico son hígado, riñón, bilis, cerebro y piel. También se han encontrado altos niveles de arsénico en las uñas y el cabello, los niveles de arsénico en estos tejidos son frecuentemente usados para evaluar su exposición crónica.

Por su complejidad, no es posible realizar estudios de la cinética del arsénico en personas expuestas crónicamente[9]; sin embargo, en animales expuestos experimentalmente de manera continua se ha observado que en un tiempo de dos semanas se incrementan los niveles de arsénico en orina, después del cual, la concentración de arsénico decrece a pesar de la exposición constante. En otros experimentos se ha encontrado que en animales expuestos a arsénico durante un tiempo corto, su excreción de arsénico es significativamente mayor que en aquellos expuestos de manera continua. Estas observaciones pueden indicar la existencia de una fase de adaptación o tolerancia.

[9] Son exposiciones constantes a través del tiempo

1.4 Biotransformación del arsénico

La biotransformación del arsénico es conocida como *metilación;* es un proceso biológico llevado a cabo principalmente en el hígado, aunque también participan otros tejidos del organismo y que consiste en agregar grupos metilo (CH_3) a una sustancia inorgánica (no contiene carbono), convirtiéndola en orgánica. Este es un fenómeno de defensa del organismo contra sustancias extrañas que ingresan a él, y que en la mayoría de los casos las hace menos tóxicas[10]. Se cree que el arsénico orgánico es mucho menos tóxico que el inorgánico, puesto que se elimina rápidamente, permaneciendo poco tiempo dentro del organismo.

El arsénico inorgánico (Asi) generalmente se encuentra en forma pentavalente o trivalente, si está presente como pentavalente se reduce a trivalente por la acción del glutatión, que es una sustancia reductora (libera valencias). Una vez que el arsénico inorgánico se encuentra en estado trivalente, entonces es posible que proceda su metilación, puesto que presenta dos valencias libres.

El arsénico inorgánico es biotransformado en 2 pasos; el primero se lleva a cabo cuando un grupo metilo se le pega formándose el arsénico monometilado (MMA), y el segundo cuando el MMA se vuelve a metilar formándose el arsénico dimetilado (DMA).

En un organismo expuesto a arsénico es posible encontrar como producto de la biotransformación porciones de Asi, MMA y DMA. Es posible identificar estos compuestos al analizar muestras de orina en seres humanos , la

[10] En el caso del mercurio la biotransformación lo hace más tóxico.

excreción urinaria de arsénico a baja dosis de exposición consiste en cerca del 10% de Asi, 10% de MMA y 80% de DMA[11].

1.5 Postulado sobre la saturación de la metilación del arsénico

Se han realizado algunos estudios para evaluar la posible influencia de los niveles de exposición a arsénico sobre los porcentajes de arsénico metilado; la orina de trabajadores de una fundidora, expuestos a altas concentraciones de arsénico inorgánico (74-934 μg/l), presentó 75% de arsénico metilado (MMA+DMA), mientras que en individuos expuestos a bajas concentraciones de arsénico (4-24 μg/l), la proporción de metilados alcanzó casi el 90% del arsénico total urinario[12]; esto sugiere que al aumentar el nivel de exposición al arsénico disminuye en el porcentaje de arsénico metilado con respecto al del inorgánico. Las proporciones de especies de arsénico inorgánico, monometilado y dimetilado respectivamente encontradas en niños que vivían en áreas rurales de Bélgica fueron 6.8%, 14.5% y 78.6% respectivamente, mientras que en la vecindad de una fundidora que emitía arsénico, los valores fueron de 11.8%, 28% y 60.2%; aquí se observa que el porcentaje de arsénico inorgánico y monometilado (MMA) en los niños de la cercanía de la fundidora aumenta, mientras que el dimetilado (DMA) disminuye.

En un estudio piloto realizado por la sección de toxicología ambiental del CINVESTAV en habitantes de la Comarca Lagunera[13], se midieron las especies de arsénico excretado en orina en dos poblados, con diferentes

[11] Vahter (1983) pg. 192.
[12] Buchet, J.P., Lauwerys, R. and Roels, H. (1980) Int. Arch. Occup. Environ. Healt 46, 11-29. Citado por ibidem p.g. 184.
[13] Del Razo (1994). pg. 20

concentraciones de arsénico en el agua usada para beber; un control con niveles de concentración dentro del límite máximo permisible (LMP) y un expuesto con niveles altos, sobrepasando cerca de ocho veces el LMP. Las proporciones encontradas fueron 15%, 9% y 76% de Asi, MMA y DMA respectivamente para el poblado control, y 17%, 20% y 63% para el grupo de alta exposición, observándose que el porcentaje de Asi permanece casi igual, mientras que %MMA aumentó y %DMA disminuyó al comparar el poblado expuesto con respecto al control.

En otro estudio similar[14], también realizado en dos poblaciones de la Comarca Lagunera, las proporciones de arsénico en orina fueron 8.7%, 6.8% y 78.5% (Asi, MMA, DMA) para el poblado control, y 30.6%, 11.3%, 54.1% respectivamente en el expuesto. Como resultado en la comparación del control v.s. el expuesto se encontraron incrementos significativos en las proporciones de Asi y MMA, asi como decrementos significativos en las de DMA.

Si consideramos al proceso de la metilación por partes, en función de la concentración de arsénico, se observa que el paso de Asi a MMA se bloquea o permanece igual, mientras que el de MMA a DMA disminuye considerablemente. Esta observación sugiere la existencia de una saturación o inhibición de la capacidad del organismo al realizar la formación de compuestos metilados, especialmente en el paso de MMA a DMA.

[14] Del Razo (1996). pg. 95.

1.6 Eliminación

1.6.1 Eliminación fecal

Como resultado de una absorción casi completa en el tracto gastrointestinal, se elimina muy poco arsénico en las heces fecales; por esta vía sólo se puede recuperar cerca de 5% de una dosis oral.

Se ha reportado que el arsénico se excreta importantemente en la bilis[15] pero aparentemente esta vía de excreción no influye su eliminación por heces debido a que el arsénico es reabsorbido por el intestino, por lo que vuelve a estar disponible para su excreción a través del riñón.

Se sabe que la bilis excreta arsénico, pero esto no contribuye a su eliminación por heces, puesto que es reabsorbido por el intestino.

1.6.2 Excreción urinaria de arsénico inorgánico

La mayor parte del arsénico inorgánico absorbido es eliminado del cuerpo por medio de la orina a través de los riñones. En condiciones normales se excreta entre 40% y 70% del arsénico ingerido dentro de las primeras 48 horas, aunque la velocidad de eliminación depende de la forma química del compuesto arsenical.

1.6.3 Otras rutas de excreción

Aunque la mayor parte del arsénico es excretado vía los riñones, una menor cantidad es eliminada por otras rutas. Se ha calculado que la pérdida de arsénico por sudoramiento profundo es de alrededor de 2 μg por hora. No existe información sobre la pérdida de arsénico por esta vía bajo condiciones

[15] Del Razo (1997) pg. 8

normales. Como consecuencia de la afinidad del arsénico a la piel, una pequeña cantidad de arsénico es removida a través de la desescamación. También se puede considerar a la acumulación del arsénico en el cabello como una forma de eliminación; se ha estimado un porcentaje máximo de 0.6% de la dosis ingerida. Otra vía es la exhalación, aunque estudios en animales revelan que la cantidad de depuración del arsénico es muy pequeña.

1.7 Efectos biológicos

1.7.1 Reseña histórica sobre los efectos del arsénico

A través de la historia, el arsénico ha adquirido una inigualable reputación como veneno. Desde hace 2,000 años a.C. era posible obtener trióxido de arsénico en la fundición de cobre, el cual se usaba como droga y como veneno. Este compuesto inorgánico es agradable al gusto e inodoro, por lo que constituyó un agente conveniente para efectuar homicidios[16]. Probablemente, una razón de su popularidad consistió en que podía obtenerse muy fácilmente y a bajo costo. En Francia, una tercera parte de los casos de envenenamiento criminal en el siglo XIX fueron atribuidos al arsénico[17].

[16] Leslie (1978) plantea la hipótesis de que a Napoleón lo mataron agregando arsénico a sus alimentos durante su destierro.
[17] Pershagen G. (1983) pg. 200.

1.7.2 Hipótesis sobre la necesidad de consumo de arsénico

El arsénico es conocido por sus efectos tóxicos cuando es administrado en dosis muy altas, sin embargo no se producen efectos dañinos cuando se ingiere en pequeñas cantidades, incluso de manera generalizada, se consume naturalmente; el arsénico en concentraciones muy pequeñas forma parte de las sales minerales del agua potable.

Actualmente se investiga si estas pequeñas cantidades cumplen alguna función útil en la vida de los seres vivos.

Se han realizado estudios en ratas, pollos, cerdos y cabras, donde se observa que la disminución de las concentraciones habituales de arsénico reduce los índices en la concepción, crecimiento, peso de nacimiento y esperanza de vida, así como un aumento en el índice de abortos[18].

Sin embargo, todavía no ha sido documentado suficientemente el uso del arsénico como un elemento esencial, además, aún no se conoce el mecanismo de acción establezca la necesidad del consumo de este elemento, de cualquier manera, esta es una suposición viable; Si algún elemento es necesitado en animales, es probable que en los humanos también sea necesario. Existen estimaciones de cuál sería la cantidad esencial de arsénico para las personas; la mayor parte de ellas caen dentro del intervalo que va desde 10 hasta 30 μg por día. Investigaciones hechas

[18] EPA (1988) pp 33-37.

por FDA[19] han reportado niveles de arsénico en la dieta alimenticia de aproximadamente 46 μg de arsénico por día.

1.7.3 Efectos por exposiciones agudas

Los efectos agudos son aquellos provocados por pocas exposiciones a altas dosis. Los síntomas causados por la exposición a arsénico inician a partir del sistema gastrointestinal e incluyen vómitos y diarrea. Si el envenenamiento es severo, puede desarrollarse un choque a partir de una deshidratación.
Se ha reportado que una dosis fatal de trióxido de arsénico para los adultos varía de 70 a 180 mg[20].
Los efectos agudos y subagudos posteriores a la exposición de compuestos arsenicales inorgánicos pueden observarse a partir de los sistemas cardiovasculares, nervioso y hematopoyético[21], así como en la piel.

Es posible el desarrollo de disturbios periféricos nerviosos de tipo sensor y motor unas semanas después de la exposición inicial cuya recuperación es lenta. Normalmente disminuye la producción de células rojas y blancas, sin embargo, este efecto se revierte durante el mes posterior al cese de la exposición. Puede desarrollarse hiperpigmentación[22] de la piel, especialmente en individuos de complexión oscura. Una característica que se encuentra en las intoxicaciones agudas con arsénico inorgánico es la presencia de líneas blancas transversas a través de las uñas, las cuales aparecen unas semanas después de la exposición.

---

[19] Food and Drug Administration, institución estadounidense que se dedica al monitoreo y análisis de alimentos y fármacos.
[20] Pershagen G. (1983) pp 200-205.
[21] El sistema hematopoyético es aquél en donde se forman los glóbulos sanguíneos.
[22] Cambios en la coloración.

Los compuestos de arsénico inorgánico irritantes tales como trióxido de arsénico, pueden dañar la piel expuesta y las membranas mucosas.

1.7.4 Efectos por exposiciones crónicas

Los efectos crónicos resultan de exposiciones por largas temporadas, lo que dificulta el conocimiento de las dosis de exposición. En muchas situaciones, no es posible hacer alguna afirmación definitiva con relación a los compuestos específicos por medio de los cuales ha ocurrido la exposición. Dada su gran diversidad, a continuación se presenta una descripción de los principales órganos que son afectados por la acción del arsénico.

1.7.4.1 Afecciones en órganos causadas por exposiciones crónicas a compuestos arsenicales.

1.7.4.1.1 Piel

La exposición a compuestos orgánicos e inorgánicos puede causar lesiones en la piel. Cuando la exposición se debe a arsénico inorgánico, las lesiones tienen formas características. Se han encontrado hiperpigmentación e hiperqueratosis palmoplantar en trabajadores de la manufactura de plaguicidas, y en los vinateros, así como en los individuos expuestos excesivamente a arsénico por medio de agua de bebida o medicamentos. En general, los cambios hiperqueratóticos en la piel dependen de la dosis de exposición en los casos donde es muy alta (1 mg por día). Tal parece que la hiperpigmentación inducida por arsénico pueden ser reversibles, mientras que las lesiones de hiperqueratosis de las palmas y plantas del pie siguen un curso más crónico.

Sólo después de un largo tiempo de exposición a compuestos de arsénico inorgánico, es posible notar las lesiones en la piel. Pueden desarrollarse carcinomas[23] de las células basales y de células escamosas. Predominantemente han sido observados cánceres en la piel tanto por vía de ingestión como de inhalación. Como regla, los cánceres en la piel aparecen solamente después de la ingestión de varios gramos de arsénico, a menudo con un tiempo de latencia de diez años o más.

1.7.4.1.2 Pulmones

La exposición ocupacional a compuestos arsenicales irritantes en el aire como trióxido de arsénico, puede provocar lesiones en el tracto respiratorio superior. Se han observado síntomas de rinofaringolaringitis, incluyendo perforación de las fosas nasales, en trabajadores de fundidoras y pesticidas, expuestos principalmente a compuestos de arsénico inorgánico.

Se sabe por estudios en diferentes países que aumenta la mortalidad por cáncer en el pulmón en los trabajadores expuestos a compuestos de arsénico inorgánico. La exposición ocupacional a arsénico y el tabaquismo pueden interactuar en forma multiplicativa induciendo cáncer en el pulmón. Pocos estudios han indicado que la exposición ambiental a arsénico en el aire pueda ser de importancia para la ocurrencia de cáncer pulmonar en la comunidad.

[23] Cáncer de un epitelio

1.7.4.1.3 Hígado

La exposición a arsénico inorgánico trivalente en concentraciones mayores a 1 mg por día puede provocar hipertensión arterial sin cirrosis hepática. Esta es una condición que requiere varios años de exposición. La cirrosis del hígado también ha sido reportada después de la medicación con la solución de Fowler, sin embargo esta información no ha sido conclusiva.

La exposición ocupacional a arsénico entre los trabajadores de las fundidoras y viñedos se ha asociado con un incremento de la mortalidad por cirrosis en el hígado, aunque no es posible hacer afirmaciones contundentes debido a la presencia de variables confusoras en la exposición como el consumo de alcohol,.

Se ha reportado hemangioendotelioma[24] del hígado debido a la exposición a arsénico por medio de vino contaminado, agua de bebida y solución de Fowler, para los cuales se estima una dosis total ingerida de arsénico en por lo menos en varios gramos.

1.7.4.1.4 Sistema cardiovascular

Se han descrito desórdenes en el sistema vascular periférico, que algunas veces conducen a la gangrena, en personas expuestas a arsénico en agua de bebida en Chile y Taiwan, así como en trabajadores de viñedos.
Se ha observado un incremento moderado de mortalidad por enfermedades cardiovasculares en dos estudios epidemiológicos en trabajadores de

[24] Tumor a veces maligno desarrollado a expensas del endotelio de los capilares.

fundidoras expuestos a arsénico, pero esta información no ha sido confirmada en estudios de otros grupos similares.

1.7.4.1.5 Sistema nervioso

La exposición por tiempos prolongados a compuestos de arsénico en el área de trabajo, por medio de ingestión de ciertas sustancias o agua de bebida, puede provocar neuropatía periférica. Se ha reportado que pueden retroceder algunos síntomas neurológicos después del cese de la exposición.

Los tratamientos con algunos compuestos orgánicos arsenicales como la triparsamida y glicobiarsol pueden provocar efectos laterales serios en el sistema nervioso central, incluyendo encefalopatía y atrofia óptica.

1.7.4.1.6 Sistema hematopoiético y linfático

Los efectos crónicos por arsénico en el sistema hematopoiético son similares a los efectos resultantes por exposiciones de tiempos cortos. Las exposiciones por tiempos largos a arsénico por medio de agua de bebida o medicación han provocado alteraciones en la eritropoyesis[25], con anemia y granulocitopenia[26]. En muchos casos son reversibles las alteraciones en la médula ósea por medio de una terapia adecuada o con el término de la exposición. Se han observado aberraciones cromosómicas en los linfocitos periféricos entre trabajadores de viñedos y fundidoras expuestos a arsénico y en pacientes tratados con compuestos de arsénico inorgánico. Esto datos indican alteraciones potenciales del material genético provocadas por arsénico.

[25] Formación de glóbulos rojos.
[26] Disminución de un tipo particular de glóbulos rojos.

1.7.4.2 Cáncer

El arsénico es un agente carcinogénico. En las personas expuestas a arsénico por vía aérea existe un incremento el la presencia de cáncer de pulmón, mientras que por vía oral es mayor la presencia de cáncer en piel, vejiga, riñón, hígado y pulmón[27].

1.8 Relación entre las proporciones de arsénico metilado y la presencia de lesiones cutáneas

En uno de los estudios piloto antes citado[28], se compararon las proporciones de especies de arsénico entre las personas del poblado expuesto que presentaron lesiones cutáneas y las del mismo poblado pero sin presencia de lesiones en la piel. El resultado fue que en los que presentaron signos cutáneos se encontró una mayor proporción de %MMA y menor de %DMA.

El presente estudio tiene como objetivo identificar el grado de asociación, si existe, entre la cantidad encontrada de arsénico en muestras de orina en sus diferentes niveles de metilación, y la presencia de lesiones en la piel que se sabe son provocadas por la exposición crónica al arsénico, tomando como información la recopilada en una muestra de habitantes de la Comarca Lagunera que están expuestos a altas concentraciones de arsénico a través de la ingestión del agua que usan para beber.

[27] Del Razo (1997) p.g. 14
[28] Del Razo (1994). pg 10.

CAPITULO 2

MODELOS ESTADÍSTICOS PARA EVALUACION DE RIESGO

El presente capítulo comienza por explicar tres modelos de investigación típicos en epidemiología y cuyos resultados suelen interpretarse mediante la estimación de un estadístico conocido como *razón de momios* de uso frecuente en las ciencias biológicas.

A continuación se describen los fundamentos estadísticos para la estimación de razones de momios y sus intervalos de confianza a partir de tablas de contingencia y posteriormente de modelos de regresión logística. Para el segundo caso se presentan estadísticos para evaluar el ajuste con base en el análisis de residuales.

2.1 Estudios analíticos para epidemiología.

La epidemiología es el estudio de la ocurrencia y distribución de enfermedades en las poblaciones.

Los estudios epidemiológicos analíticos inician cuando la información preliminar recolectada de estudios rutinarios (análisis clínicos) y otras fuentes indican la necesidad de realizar algún estudio más detallado para probar

alguna hipótesis. Por la forma de recolección de la información, existen principalmente tres tipos de estudios analíticos; *de cohorte* o *longitudinales* (prospectivos y retrospectivos), *de casos y controles* y *transversales.*

2.1.1 Estudios prospectivos de cohorte.

Se parte de un grupo suficientemente grande de individuos que están expuestos a algo que se sabe causa alguna lesión, según el conocimiento previo obtenido por otros estudios. Después se le hace un seguimiento a estas personas durante algún tiempo considerable, que puede ser de algunas decenas de años. Una vez transcurrido el tiempo del estudio, se analizan las proporciones de las personas que presentan lesiones y están expuestas a aquello que se sospecha es la causa. Con esta información, se obtiene la probabilidad condicional del evento consistente en presentar lesión dado que se está o no expuesto. También se puede evaluar la influencia de otras variables de tipo general como la edad, sexo, etc. sobre la prevalencia de lesiones.

2.1.2 Estudios de casos y controles.

Se toma una muestra de individuos con lesión, y otra de personas sanas. En ambas muestras deberá haber personas expuestas y no expuestas a lo que es un factor de riesgo para la lesión. Las personas con lesiones son los casos y quienes no presenten lesiones serán los controles. En este estudio se trata de evaluar la diferencia en la incidencia de lesiones (variable respuesta) entre las personas no expuestas y las expuestas (variable explicativa).

2.1.3 Estudios transversales o de prevalencia.

La muestra de estos estudios se levanta en un tiempo fijo o en un intervalo corto de tiempo. Se compara la incidencia de lesiones en un grupo de personas expuestas con otro de personas no expuestas o levemente expuestas, y se analizan las diferencias en la prevalencia de lesiones.

2.1.4 Resumen comparativo.

En un estudio transversal primero se selecciona la muestra y después se agrupa a las observaciones de acuerdo al nivel de exposición y de salud que les haya correspondido.

En un estudio de cohorte, los individuos que se hayan seleccionado en un inicio no deberán presentar las lesiones que se desea evaluar, de tal forma que después de algún tiempo considerable, cuando se vuelva a revisar el estado de salud, existirán personas que presentarán la lesión. En este caso se infiere directamente sobre la probabilidad de padecer la enfermedad.

El estudio de casos y controles es muy útil en situaciones donde el presupuesto es limitado y la enfermedad que se desea estudiar es poco común. En esta situación, el investigador puede decidir cuántas personas sanas y enfermas utilizar en la muestra. Si se desea inferir sobre la probabilidad de presentar lesiones, esto se podrá hacer sólo si se conoce la proporción de individuos en la población que padecen dicha enfermedad. en cuyo caso se recurrirá a la probabilidad condicional para que las proporciones obtenidas en la muestra correspondan con las poblacionales.

2.2 Tablas de contingencia y razones de momios.

2.2.1 Tablas de contingencia

Las tablas de contingencia consisten en una tabulación de datos que sirve para identificar el *número de observaciones* que ocurren en los cruces de valores de variables categóricas.

Supóngase que se desea evaluar la relación entre el tiempo de exposición a arsénico, y la presencia de lesiones. Para este ejemplo en particular, se requiere que la variable de exposición sea categórica, y en particular con dos niveles, así que se definirá al nivel de exposición inferior como el tiempo de exposición menor o igual a 15, y el superior para tiempos de exposición mayores a 15 años:

$$t_{exp.\ menor}: t_{exposición} \leq 15$$
$$t_{exp.\ mayor}\ \ t_{exposición} > 15$$

De esta manera obtendremos dos niveles para cada variable; exposición menor o mayor y presencia o ausencia de lesiones específicas del arsénico, resultando 2x2=4 agrupaciones distintas.

Si se agrupan los datos de la muestra entre estas dos variables, y a cada grupo se asigna el *número de observaciones* que caen dentro de él, obtendremos una *tabla de contingencia* o *tabla de cruce de clasificaciones* como la Tabla 1, en la cual además de las frecuencias se han puesto los totales por columna y renglón, y en la esquina inferior derecha el número de personas en la muestra.

TIEMPO DE EXPOSICIÓN POR LESIONES ESPECÍFICAS

	Lesiones Específicas		Total
	si	no	
$t_{exp.\ menor}$	7	84	91
$t_{exp.\ mayor}$	30	59	89
Total	37	143	180

Tabla 1

Nótese que los valores de la variable lesiones se han colocado en las columnas mientras que el tiempo de exposición en los renglones.

A partir de esta tabla, es posible realizar algunos cálculos para identificar la asociación, si es que existe, entre las dos variables. Una forma de hacerlo consiste utilizar una estadística para inferir acerca de la independencia o asociación entre las dos variables. No siempre bastará con decir que existe asociación, en ocasiones habrá que calcular o estimar de alguna forma la fuerza de dicha asociación entre las dos variables.

2.2.2 Independencia y fuerza de asociación

A continuación se introduce la notación empleada para representar a las frecuencias y proporciones en una tabla de contingencia de 2x2.

TABLA DE CONTINGENCIA DE 2x2

	Variable columna		
	y_1	y_2	
x_1	n_{11}	n_{12}	n_{1+}
x_2	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

Tabla 2

La variable columna es el vector X, con los elementos x_1 y x_2. La variable renglón es el vector Y, con y_1 y y_2 , n_{ij} representa el número de observaciones que se encuentran en la celda correspondiente al *i*-ésimo renglón y la *j*-ésima columna. La suma de observaciones se representa con n_{i+} para el renglón i , y con n_{+j} para la columna j. El total de individuos en la muestra se representa con n.

La siguiente expresión es la proporción p_{ij} de observaciones que corresponden a la celda *ij* con respecto al total de individuos en la muestra:

$$p_{ij}=\frac{n_{ij}}{n}$$

Si las variables X, Y son independientes, entonces el valor que tome alguna de ellas no influye sobre los valores de la otra, por lo que la proporción de individuos que presentan la característica y_j será la misma para cualquier valor de la variable renglón x_i y viceversa; la proporción de individuos que

presentan x_i será la misma para cualquier valor de y_j, por lo que X, Y serán independientes si se cumple que:

$$p_{1j} = p_{2j}, \text{ o bien } p_{i1} = p_{i2}$$

en cuyo caso $p_{ij} = p_{i*}\, p_{*j}$ para cualquier valor de i , j.

Esta definición aplica también para el caso en que las variables X y Y tomen más de dos valores:

$$p_{1j} = p_{2j} = \ldots = p_{nj}$$

$$p_{i1} = p_{i2} = \ldots = p_{im}$$

para cualquier valor de i, j donde $i = \overline{1, n}$ $j = \overline{1, m}$ en una tabla de nxm (n renglones y m columnas).

En ambos casos se evalúa si existe independencia entre las dos variables por medio de una prueba de hipótesis para diferencia de proporciones, con algún estadístico como la X^2 de Pearson.

2.2.2.1 Prueba de independencia X^2 de Pearson

Se sabe por la teoría de probabilidades que si dos eventos son independientes, la probabilidad de que ocurran ambos es igual al producto de sus probabilidades. De esta manera podemos obtener la probabilidad de ocurrencia en una celda *ij* bajo el supuesto de independencia, al multiplicar

las probabilidades de cada evento, de tal forma que las variables en la tabla de contingencia serán independientes si se cumple que:

$$p_{ij} = \frac{n_{i+}}{n}\frac{n_{+j}}{n}$$

de donde se obtiene el número de observaciones esperado (E_{ij}) multiplicando por *n*:

$$E_{ij} = np_{ij} = \frac{n_{i+}n_{+j}}{n}$$

de tal forma que $n_{ij} \approx np_{ij}$ si X,Y son independientes.

Para inferir acerca de la independencia de X, Y se puede emplear una prueba de hipótesis basada en una aproximación a la distribución χ^2 propuesta por Pearson, y decidir si la magnitud de las diferencias entre el número observado (n_{ij}) y esperado (E_{ij}) es significativa, en cuyo caso se rechazará la hipótesis de independencia y se concluirá que las variables son dependientes. De lo contrario se aceptará la hipótesis nula.

La prueba consiste en obtener una aproximación a la distribución χ^2 por medio del estadístico:

$$X^2_{gl} \approx \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Este sigue un comportamiento χ^2 con *g.l.*=(r-1)(c-1) grados de libertad[1], donde r = número de renglones y c = número de columnas.

Las hipótesis que se evalúan son las siguientes:

H_o:*X, Y son independientes* (no hay diferencia significativa entre n_{ij} y E_{ij})

H_a:*X, Y no son independientes* (la diferencia es significativa).

Si el valor de X^2 calculado es mayor que χ^2 evaluada en α, donde α es el nivel de significancia establecido, se rechaza la hipótesis nula y se concluye que *X, Y* no son independientes.

Por lo general cuando se evalúa una prueba de hipótesis en la computadora, lo que se obtiene es el estadístico calculado junto con el valor p, que es el nivel de significancia alcanzado en la prueba de hipótesis. Es por esta razón que se rechazará la hipótesis nula, H_o cuando esta probabilidad sea muy pequeña ($p < \alpha$). El valor p se obtiene al evaluar el área bajo la curva en la función de distribución χ^2 que se encuentra entre el lugar que ocupa el valor calculado y la cola derecha.

La importancia del valor p consiste en que proporciona una idea probabilística de qué tanto fundamento existe para rechazar la hipótesis nula.

[1] Número de observaciones independientes de la muestra menos el número de parámetros utilizados

A continuación se muestra un ejemplo donde se evalúa si existe asociación entre las variables tiempo de exposición y lesiones específicas presentadas en la tabla 3, para lo cual se ha definido la siguiente tabla, a la que se añadieron las frecuencias esperadas $E_{ij} = \frac{n_{i+}n_{+j}}{n}$ entre paréntesis.

	Lesiones Específicas		
	Si	No	Total
$t_{exp.\ mayor}$	30	59	89
	(18)	(71)	
$t_{exp.\ menor}$	7	84	91
	(19)	(72)	
Total	37	143	180

Tabla 3

Aplicando el estadístico se obtiene que $X^2 \approx 18.64801$ con un grado de libertad. Si se establece α=0.01 como la región de rechazo, se observa en tablas que χ^2=6.63490 es menor que el valor calculado, inclusive tomando un valor más estricto para α digamos α=0.005, obtenemos que χ^2=7.87944 < 18.64801.

Esto significa que es muy grande la diferencia entre la frecuencia observada y la esperada bajo condiciones de independencia, por lo que se rechaza la hipótesis nula y se concluye que sí existe asociación entre ambas variables.

A continuación se muestra una salida de la estadística calculada en SAS[2] para el mismo problema:

STATISTICS FOR TABLE OF ANOSCUT BY ESPECIF

Statistic	DF	Value	Prob
Chi-Square	1	18.648	0.001

ANOSCUT es la variable discretizada para el tiempo de exposición, y ESPECIF es la variable que indica si se padece lesión especifica o no. Bajo la palabra Prob se indica el valor p, en este caso se interpreta como $p < 0.001$. Esto nos da una idea de la significancia con la que se rechaza la hipótesis nula. Como se puede observar, este valor es muy pequeño y se concluye, como en el ejercicio anterior, que existe asociación entre ambas variables.

Otras pruebas de independencia y consideraciones acerca de χ^2:

Existe un estadístico alternativo, basado en la distribución normal, con el que se puede inferir si las proporciones son diferentes. De hecho, la distribución χ^2 es una suma del cuadrado de variables aleatorias normales, y se puede deducir el estadístico de Pearson a partir del de la distribución normal. Por

[2] Statistical Analisis System o Sistema para análisis estadistico

salirse de los objetivos del presente trabajo, esta demostración no será expuesta, pero puede consultarse la bibliografía para tal efecto[3].

En los casos como el presente, en el cual se ha concluido que las variables no son independientes, surge de manera natural una pregunta: ¿qué tan diferentes son?.

Es posible concebir alguna idea de cuál sería la respuesta por medio del valor p o de la diferencia que existe entre el valor observado de X^2 y el esperado. Intuitivamente se puede inferir que en el ejemplo mostrado, la diferencia es muy grande, pero ¿cómo cuantificarla en términos más claros?

En la siguiente sección se desarrollan algunos estadísticos con los cuales se contestarán estas preguntas.

2.2.3 Cuantificación de las diferencias en una tabla de contingencia.

En epidemiología existen estadísticas específicas para cuantificar la asociación de las variables. Tal es el caso del riesgo relativo y las razones de momios, que a continuación se explican:

2.2.3.1 Riesgo relativo

Es la razón de riesgos de que ocurra una enfermedad durante el transcurso de un tiempo delimitado entre personas sometidas a diferentes niveles de exposición.

---

[3] Collett (1991), pg. 17-35.

El riesgo relativo indica cuántas veces es mayor o menor la proporción de individuos incidentes entre un grupo de mayor exposición con respecto a otro de control. El riesgo relativo está dado por la siguiente expresión:

$$Riesgo\ Relativo = \frac{\frac{p_{11}}{p_{11}+p_{12}}}{\frac{p_{21}}{p_{21}+p_{22}}}$$

donde en p_{ij} j=1 significa presencia de lesiones.

Obsérvese que $p_{11}+p_{12}=1, y\ p_{21}+p_{22}=1$ por lo que la expresión se simplifica a:

$$\text{Riesgo Relativo}=\frac{p_{11}}{p_{21}}$$

para los niveles i=2 (no expuesto) y 1 (expuesto) en p_{ij}.

Un cociente cercano a 1 indicará que la proporción de individuos incidentes se mantiene constante en los diferentes niveles de la variable antecedente (x_i), por lo que no existe relación entre las dos variables. Los cocientes mayores a uno indican el número de veces que es mayor la incidencia en los individuos expuestos con respecto a los no expuestos o control, mientras que

los menores a uno indican que la incidencia es menor en los individuos expuestos.

2.2.3.2 Momios

La siguiente definición está tomada de Agresti[4], quien expone el concepto de *momio* para una tabla de 2x2:

"Dentro del renglón 1, el *momio* de que la respuesta esté en la columna 1 en lugar de la columna 2 está definido como":

$$m_1 = \frac{p_{11}}{p_{12}}$$

"Dentro del renglón 2, el momio correspondiente es igual a":

$$m_2 = \frac{p_{21}}{p_{22}}$$

donde $p_{ij} = \frac{n_{ij}}{n_{i+}}$.

Como puede observarse, un momio es la comparación de dos proporciones que se encuentran en el mismo renglón; en la tabla de 2x2, indica cuántas

[4] Agresti (1990) pg. 14-15

veces es mayor o menor la proporción de casos en y_1 con respecto a y_2 para cada valor de la variable antecedente x_i, $i = \overline{1,2}$. Dicho de otra manera, un momio es el cociente del número de individuos incidentes (y_1) por cada persona sana (y_2).

Por ejemplo, supóngase que se está evaluando una variable que identifica el estado de salud, la cual toma el nivel 1 si existe alguna lesión y 0 en caso contrario. Además existe otra variable explicativa que mide el nivel de exposición a cierta sustancia, y se desea saber si ésta altera el estado de salud.

Si clasificamos una muestra en una tabla de contingencia y calculamos los momios, obtendremos la relación del números de personas que padecen la lesión por cada individuo sano en cada nivel de exposición. La magnitud en la cual estos cocientes difieran entre cada renglón, indicará el grado de asociación de la variable respuesta con la variable explicativa, es decir cuantas veces es mayor o menor el número de personas incidentes por individuo sano entre grupos con diferente nivel de exposición. Una forma de realizar esta comparación consiste en obtener los cocientes de los momios entre cada renglón con respecto al de individuos no expuestos o control y se conoce como *razón de momios*.

2.2.3.3 Razón de momios

En una tabla de 2x2 la razón de momios está dada por la siguiente expresión:

$$Razón\ de\ momios = \frac{m_1}{m_2} = \frac{\frac{p_{11}}{p_{12}}}{\frac{p_{21}}{p_{22}}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Al dividir dos momios se obtendrá el número de veces que es mayor o menor la relación de individuos con lesión respecto a los sanos entre un grupo de individuos con respecto a otro grupo. Si la diferencia es grande, podrá concluirse que existe asociación, puesto que el *momio* de contraer alguna enfermedad es diferente para cada grupo.

Los valores que puede tomar una razón de momios son de cero a infinito. Para una razón de momios del tipo m_1/m_2 (el 1er renglón con respecto al segundo), los valores mayores que 1 indicarán cuantas veces es mayor la razón de incidencia y_1/y_2 en 1 ^{er} renglón que con respecto al segundo. Una razón de momios de 1 indicará que las variables son independientes puesto que se mantiene constante el índice de incidencia, una razón de 4 indicará que el momio de que ocurra la respuesta 1 es cuatro veces mayor en el renglón 1 que en el renglón 2. Una razón de momios de ¼ indicará que el momio de que ocurra la respuesta 1 es cuatro veces menor para el renglón 1 que para el renglón 2. En la medida en que la razón de momios sea mayor que 1, será mayor la ocurrencia de la respuesta 1 para el renglón 1 que para

el renglón 2. Mientras que si esta es menor que 1 y se acerca a cero será menor la ocurrencia de la respuesta 1 para el renglón 1 que para el renglón 2.

2.2.3.5 El logaritmo de la razón de momios

En ocasiones resulta más conveniente utilizar el logaritmo de la razón de momios para su interpretación porque es simétrico[5] respecto al origen; el valor de la razón de momios cuando no existe asociación es 1, mientras que log(1)=0. El logaritmo natural de una razón de momios de 4 es 1.386, mientras que de ¼ es -1.386 aproximadamente; para los números mayores que 1 el logaritmo toma valores positivos, mientras que cuando la razón de momios es menor que la unidad, su logaritmo es negativo. En realidad lo que se obtiene es una transformación F:X→Y uno a uno $f(x) = y$ del conjunto X={x:x ε (0,∞)} en Y={y:y ε (-∞,∞)}, donde se cumple que:

$$\mathbf{Log}\left(\frac{m_1}{m_2}\right) = \mathbf{-Log}\left(\frac{m_2}{m_1}\right)$$

es decir, el logaritmo de una razón de momios es el mismo pero con signo contrario si intercambiamos los renglones. De aquí se desprende que es simétrico con respecto al origen.

[5] Equidistante

2.2.3.6 Intervalos de confianza para razones de momios

Debido a que el logaritmo de la razón de momios se aproxima a una distribución normal[6], es posible la construcción de intervalos de confianza.

El error estándar del logaritmo de la razón de momios está dado por[7]:

$$s.e.\log(\hat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Con lo cual se construyen intervalos de confianza del 100(1-α)% de la forma $\log(\hat{\psi}) \pm z_{\alpha/2} s.e.(\log\hat{\psi})$, donde $\hat{\psi}$ representa la razón de momios estimada, y $z_{\alpha/2}$ es el valor de tablas.

Se calcula el antilogaritmo de los intervalos así obtenidos para interpretar el resultado en términos de razones de momios.

2.3. Modelos de regresión con respuesta binaria

Un modelo de regresión lineal múltiple consiste en una ecuación matemática que relaciona a un conjunto de variables llamadas explicativas con otra variable llamada respuesta, que se supone están relacionados de manera lineal por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon,$$

[6] Collet (1991) p.g. 36

donde $\beta_i \quad i=\overline{0,n}$ son parámetros desconocidos, $x_j \quad j=\overline{1,n}$ son las variables explicativas, y es la variable respuesta y ε_i. es un error aleatorio.

En particular si suponemos que $\varepsilon_i \sim N(0,\sigma^2)$,

$$y_i \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n, \sigma^2)$$

El objetivo en un análisis de regresión consiste en ajustar el mejor modelo posible a la variable respuesta, de tal forma que el error sea mínimo, para tal efecto existen diferentes métodos de estimación de modelos, dependiendo del comportamiento y tipo de la variable respuesta.

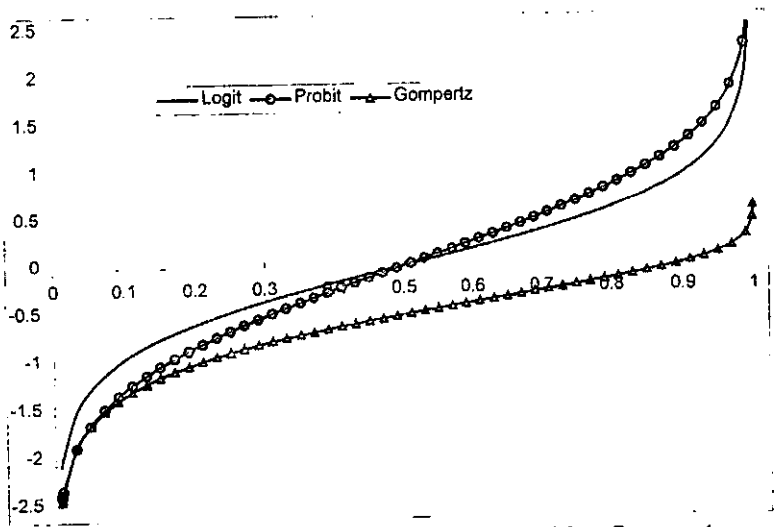
El estimar este tipo de modelo cuando la variable respuesta es binaria trae como consecuencia algunos resultados no deseados, porque para ciertos valores de las variables explicativas resultará que la estimación de la variable respuesta diferirá por mucho a los dos valores posibles de la respuesta binaria; por ejemplo, supóngase que Y=0 ó 1. Si $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$, es posible obtener alguna x_i tal que $\hat{y}$=200 o $\hat{y}$=-200 o $\hat{y}$=0.5. Una alternativa para los valores que pertenecen a (0,1) sería redondearlos y así llegaríamos a 0 ó 1, sin embargo esta técnica no funciona para valores fuera de dicho intervalo, además no es válido suponer, si la variable respuesta es binaria, que:
$y_i \sim N(\mu,\sigma^2)$.

[7] Collet (1991) p.g. 36

Lo anterior se resuelve utilizando una transformación o función de *liga* que realice un mapeo uno a uno del conjunto de los números reales $(-\infty,\infty)$ en el intervalo (0,1) de tal forma que en la medida en que algún número se aleje infinitamente del cero o uno en la ecuación de regresión, éste se acercará infinitamente a cero o a uno respectivamente, estimando así la probabilidad $P(y=1)$.

2.3.1 Funciones de liga logit, probit y gompertz

A continuación se muestra una gráfica donde se aprecia el mapeo de las funciones de liga Logit, Probit y Gompertz:



Gráfica de las funciones de liga Logit, Probit y Gompertz

2.3.1.1 Función logit

Esta es la más comúnmente usada en epidemiología, porque involucra directamente a la razón de momios. Esta función está dada por la siguiente expresión:

$$y = \text{logit}(p) = \log(\frac{p}{1-p})$$

$$p = \text{logit}^{-1}(y) = \frac{e^{y}}{1-e^{y}}$$

donde $y \in (-\infty,\infty)$, $p \in (0,1)$.

Obsérvese que:

a medida que p→0 en $\frac{p}{1-p}$, $\frac{p}{1-p} \rightarrow 0$

a medida que p→1 en $\frac{p}{1-p} \rightarrow \infty$.

Al aplicar la función logaritmo ocurre que:

cuando x→0, log(x) →-∞

cuando x→∞, log(x) →∞

Con lo que habremos obtenido el mapeo deseado.

2.3.1.2 Función probit

Esta función está basada en la correspondencia del área bajo la curva de la función de densidad normal acumulada, de donde se obtiene el mapeo de p $\in$ (0,1) en x $\in$ $(-\infty,\infty)$ por medio de la siguiente función:

$$p = F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}v^2} dx$$

2.3.1.3 Función gompertz

También es conocida como *transformación complementaria log-log*. Está dada por:

$$log[-log(1-p)].$$

Esta función se diferencia de las dos anteriores en que no es simétrica alrededor de *p*. Esto se puede observar si se grafica el valor de *p* con su correspondiente transformación.

2.3.2 Modelos de regresión logística

Los modelos lineales de regresión logística están dados por la expresión:

$$\text{logit}(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n$$

donde $\text{logit}(\hat{p})$ es la función logit.

Hay algo que siempre es posible obtener de este modelo, sin importar que se trate de un estudio de cohorte, transversal o de casos y controles; se le

utiliza para estimar razones de momios ajustadas por variables confusoras[8] en un análisis multivariado. En otras palabras, los modelos de regresión logística son especialmente útiles para estimar las razones de momios debidas a los incrementos o decrementos de cualquier combinación lineal de variables explicativas.

2.3.3 Razones de momios ajustadas

El cálculo de las razones de momios, una vez que el modelo de regresión logística se ha estimado, consiste en lo siguiente: se sustituyen en éste primeramente los que se consideren valores iniciales o de referencia para obtener $\text{logit}(\hat{p}_0)$, posteriormente se sustituyen los nuevos valores de las variables explicativas cuyo efecto se desea evaluar sobre la variable respuesta, de tal forma que se obtiene $\text{logit}(\hat{p}_1)$. La razón de momios está dada por:

$$Log(\hat{\psi}_{10}) = \text{logit}(\hat{p}_1) - \text{logit}(\hat{p}_0)$$

$$\hat{\psi}_{10} = e^{\text{logit}(\hat{p}_1) - \text{logit}(\hat{p}_0)}$$

según se demuestra a continuación:

Dado que $\text{logit}(\hat{p}_i) = \log(\frac{\hat{p}_i}{1-\hat{p}_i})$

[8] Son variables explicativas adicionales a las que miden exposición

$$\text{logit}(\hat{p}_1) - \text{logit}(\hat{p}_0) = \log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \log\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right) = \log\left(\frac{\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right)}{\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right)}\right) = \log(\hat{\psi}_{10})$$

Este razonamiento puede verse de la siguiente manera:

$$\text{Sean } \log it(\hat{p}_1) = \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{12} + \hat{\beta}_3 x_{13}$$

y

$$\log it(\hat{p}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \hat{\beta}_3 x_{03}$$

la diferencia $\log it(\hat{p}_1) - \log it(\hat{p}_0)$ estará dada por:

$$\log it(\hat{p}_1) - \log it(\hat{p}_0) = \hat{\beta}_1(x_{11} - x_{01}) + \hat{\beta}_2(x_{12} - x_{02}) + \hat{\beta}_3(x_{13} - x_{03})$$

Si se desea conocer la importancia de la variable explicativa x_1 el resto de las variables explicativas permanecerán constantes por lo que las diferencias $(x_{12}-x_{02})$ y $(x_{13}-x_{03})$ serán cero y $\log it(\hat{p}_1) - \log it(\hat{p}_0)$ se reduce a $\hat{\beta}_1(x_{11} - x_{01})$, de donde $\hat{\psi}_{10} = e^{\hat{\beta}_1(x_{11}-x_{01})}$. En este caso, dada la linealidad del predictor, no importa el valor inicial sino el incremento en x_1, por lo que la razón de momios estimada para el incremento en x_1 será expresada como

$\hat{\psi}_{10} = e^{\Delta_{x_1}\hat{\beta}_1}$. Por lo tanto, la razón de momios ajustada para el incremento en la variable x_i está dada por: $\hat{\psi}_i = e^{\Delta_{x_i}\hat{\beta}_i}$

2.3.4 Modelos de regresión logística con términos independientes polinomiales o con interacciones entre ellos

En ocasiones, es posible que el comportamiento del modelo que se desea estimar requiera evaluar el efecto combinado de dos o más variables (interacción) o de alguna transformación de las variables explicativas.

Supóngase que se desea estimar el siguiente modelo:

$$\log it(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_1^3 + \hat{\beta}_6 x_2 x_3$$

Es posible definir transformaciones tales que:

$$z_4 = x_1^2 \qquad z_5 = x_1^3 \qquad z_6 = x_2 x_3$$

obteniendo el modelo:

$$\log it(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 z_4 + \hat{\beta}_5 z_5 + \hat{\beta}_6 z_6$$

Al calcular una razón de momios para este modelo, se deberá tomar en cuenta que los parámetros $\hat{\beta}_4$ y $\hat{\beta}_5$ corresponden a las variables transformadas.

Deshaciendo la transformación de variables y reescribiendo el modelo en términos de los incrementos de las variables explicativas, se obtiene la siguiente expresión:

$$\log(\hat{\psi}) = \hat{\beta}_0 + \hat{\beta}_1\Delta x_1 + \hat{\beta}_2\Delta x_2 + \hat{\beta}_3\Delta x_3 + \hat{\beta}_4\Delta(x_1^2) + \hat{\beta}_5\Delta(x_1^3) + \hat{\beta}_6\Delta(x_2x_3)$$

En este modelo, el valor de la razón de momios estimada será diferente por cada valor inicial distinto en las variables explicativas.

Es posible definir un gran número de combinaciones para formar modelos distintos, el cual se multiplicará en la medida en que se agreguen variables al modelo. En la práctica resulta inoperante ajustar todos los modelos posibles para determinar si éste es el comportamiento que más se asemeja a la realidad. Incluso si se desea experimentar esto se corre el riesgo de sugerir modelos muy elaborados que por casualidad resulten significativos y que poco tienen que ver con la realidad. Por esta razón resulta más conveniente recurrir a la experiencia del investigador y sólo considerar este tipo de modelos cuando existan razones suficientes para suponer la existencia de comportamientos no lineales. Como ejemplo de esto, si se conoce que el efecto producido por alguna variable explicativa depende de los valores que tome otra variable, se trata de un caso de interacción entre variables. Ésta se puede evaluar ajustando un término de tipo multiplicativo: x_1x_2.

2.4 Estimación de parámetros en modelos de regresión logística

2.4.1 Función de verosimilitud

Se desea estimar un modelo de regresión logística partiendo de un conjunto de observaciones con determinadas características, identificadas por las combinaciones lineales de las variables explicativas X_j. Estas observaciones se pueden agrupar en una tabla de contingencia, de tal forma que se obtenga el número de casos y de no casos para cada combinación posible en términos de las variables explicativas.

Por ejemplo, supóngase que se desea encontrar la relación entre dos variables explicativas x_1 y x_2, donde cada una puede tomar 2 y 3 valores distintos respectivamente. En este caso existirán 2x3=6 combinaciones posibles en términos de las variables explicativas. Al agrupar a las observaciones con base a estas variables, podrá sumarse el número de casos y de no casos encontrado en cada situación y obtener así la proporción de individuos que presentan o no presentan la característica o respuesta que se está evaluando.

En términos estadísticos es posible construir una distribución binomial para cada una de las combinaciones lineales existentes en la muestra mediante la agrupación de observaciones con los mismos valores en las variables explicativas. Distinguiéndose cada combinación por el subíndice i, se podrán obtener los parámetros n_i, número de individuos incluidos en el grupo i, así como y_i número de individuos que presentan la respuesta.

Una vez agrupados los datos de esta manera, será definida una distribución binomial para cada grupo i, con parámetros n_i, y_i, cuya probabilidad a estimar será p_i.

$$p(y = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

La función de verosimilitud para este modelo estará dada por la multiplicación de todas las binomiales de la siguiente manera:

$$L(p) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

2.4.2 Maximización de la función de verosimilitud

Como primer paso se obtendrá la derivada de siguiente función:

1.- $$L(p) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

Puesto que contiene productos y potencias, se procederá a obtener el logaritmo en ambos lados de la ecuación para simplificar el proceso de derivación:

$$\ln L(p) = \sum_{i=1}^{n} \left\{ \ln \binom{n_i}{y_i} + y_i \ln(p_i) + (n_i - y_i) \ln(1 - p_i) \right\}$$

$$\ln L(p) = \sum_{i=1}^{n} \left\{ \ln \binom{n_i}{y_i} + y_i \ln(p_i) - y_i \ln(1-p_i) + n_i \ln(1-p_i) \right\}$$

$$\ln L(p) = \sum_{i=1}^{n} \left\{ \ln \binom{n_i}{y_i} + y_i \left(\ln(p_i) - \ln(1-p_i)\right) + n_i \ln(1-p_i) \right\}$$

2.- $$\ln L(p) = \sum_{i=1}^{n} \left\{ \ln \binom{n_i}{y_i} + y_i \ln \frac{p_i}{1-p_i} + n_i \ln(1-p_i) \right\}$$

Si se despeja p_i de la expresión $logit(p_i) = \eta_i$, donde

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

se obtiene que:

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) = \eta_i,$$

$$\frac{p_i}{1-p_i} = e^{\eta_i}$$

$$p_i = e^{\eta_i}(1-p_i)$$

$$p_i = e^{\eta_i} - e^{\eta_i} p_i$$

$$p_i + e^{\eta_i} p_i = e^{\eta_i}$$

$$p_i(1 + e^{\eta_i}) = e^{\eta_i}$$

3.- $$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Aplicando **3.-** a los argumentos de la función logaritmo **2.-** y simplificando resulta:

4.- $$\frac{p_i}{1 - p_i} = \frac{\frac{e^{\eta_i}}{1 + e^{\eta_i}}}{1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}}} = \frac{\frac{e^{\eta_i}}{1 + e^{\eta_i}}}{\frac{1 + e^{\eta_i} - e^{\eta_i}}{1 + e^{\eta_i}}} = \frac{\frac{e^{\eta_i}}{1 + e^{\eta_i}}}{\frac{1}{1 + e^{\eta_i}}} = e^{\eta_i}$$

5.- $$1 - p_i = 1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{1 + e^{\eta_i} - e^{\eta_i}}{1 + e^{\eta_i}} = \frac{1}{1 + e^{\eta_i}}$$

Sustituyendo **4.-** y **5.-** en **2.-** y reescribiendo en términos de los parámetros de regresión β_j, $j = \overline{0,k}$,

$$B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}$$

y simplificando se llega a la siguiente función de verosimilitud:

$$\ln L(B) = \sum_{i=1}^{n} \left\{ \ln \binom{n_i}{y_i} + y_i \ln e^{\eta_i} + n_i \ln \frac{1}{1+e^{\eta_i}} \right\}$$

6.- $$\ln L(B) = \sum_{i=1}^{n} \left\{ \ln \binom{n_i}{y_i} + y_i \eta_i - n_i \ln(1+e^{\eta_i}) \right\}$$

El siguiente paso consiste en derivar la función de verosimilitud con respecto a cada una de las betas:

7.- $$\frac{\partial \ln L(B)}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ y_i \frac{\partial \eta_i}{\partial \beta_j} - n_i \frac{\partial \ln(1+e^{\eta_i})}{\partial \beta_j} \right\} \qquad j = 0,1,\ldots,k$$

Substituyendo el valor de $\eta_i = \beta_j x_{ji}$ para resolver las derivadas parciales se obtiene que:

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_{j=0}^{k} \beta_j x_{ji}}{\partial \beta_j} = x_{ji}$$

y

$$\frac{\partial \ln(1+e^{\eta_i})}{\partial \beta_j} = \frac{\partial \ln(1+e^{\sum_{j=0}^{k} \beta_j x_{ji}})}{\partial \beta_j} = \frac{1}{1+e^{\sum_{j=0}^{k} \beta_j x_{ji}}} e^{\sum_{j=0}^{k} \beta_j x_{ji}} x_{ji} = x_{ji} \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

al sustituir estos resultados en 7.- resulta:

$$\frac{\partial \ln L(B)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} n_i x_{ji} \frac{e^{\sum_{i=0}^{k} \beta_j x_{ji}}}{1 + e^{\sum_{i=0}^{k} \beta_j x_{ji}}} \qquad j = 0,1,...,k$$

Por simplicidad se reescribirá el resultado en términos de η_i:

$$\frac{\partial \ln L(B)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ji} - \sum_{i=1}^{n} n_i x_{ji} \frac{e^{\eta_i}}{1 + e^{\eta_i}} \qquad j = 0,1,...,k$$

Al igualar las derivadas parciales con cero se llega a un sistema de ecuaciones simultáneas no lineales el cual carece de solución analítica, por lo que se deberá hacer uso de métodos numéricos para resolverlo.
Las incógnitas del sistema de ecuaciones son precisamente los parámetros que se desea estimar β_j donde $j = 0,1,...,k$, por lo que al resolver éste se podrá sustituir directamente la solución obtenida en el modelo de regresión logística.

A continuación se procederá a explicar el fundamento de los modelos utilizados en métodos numéricos para resolver sistemas de ecuaciones simultáneas no lineales.

2.4.3 Método de Newton para resolver sistemas de ecuaciones simultáneas no lineales

Existen diferentes modelos en métodos numéricos para resolver sistemas de ecuaciones no lineales y todos ellos son variaciones al método de Newton. La razón de ello consiste en que presentan mayor eficiencia en situaciones específicas; son modelos especializados para resolver cierto tipo de problemas. Puesto que el método básico es el de Newton, será éste el expuesto en este trabajo.

Para facilitar la explicación del método para sistemas de ecuaciones, se expondrá la deducción del método para encontrar la solución de una sola ecuación no lineal y después se generalizará para un sistema de ecuaciones.

Partiendo de un valor inicial y de la función *f* a evaluar, el método consiste en los siguientes pasos:

1. Obtener la ecuación de la recta que pasa por $(x_i, f(x_i))$ con pendiente $f'(x_i)$
2. Evaluar en la ecuación de la recta el valor de la abcisa x (haciendo $y=0$) para obtener el nuevo valor inicial de x
3. Con el nuevo valor inicial, repetir el primer paso hasta que $f(x_i) \simeq 0$.

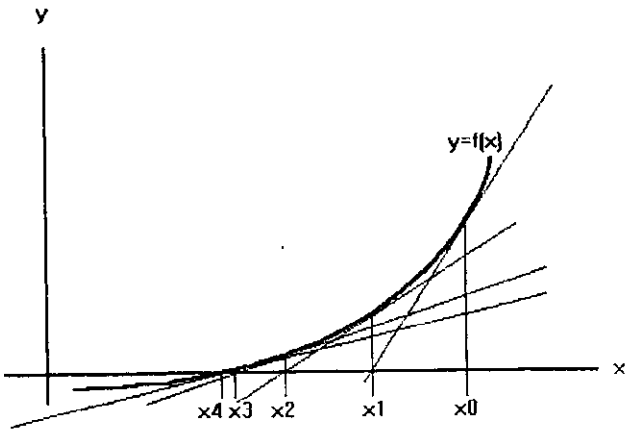
La fórmula está dada por la siguiente ecuación:

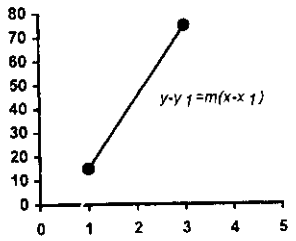
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Deducción:

Sea una función y=f(x) como la que se muestra en la gráfica. Se desea encontrar el valor de x que hace que f(x)=0.

Método de Newton para resolver ecuaciones no lineales





Dado un valor inicial x_0 y una función $f(x)$, es posible obtener la ecuación de la recta tangente a la función $f(x)$ en el punto x_0 utilizando la forma punto pendiente de la ecuación de una recta; $y-y_1=m(x-x_1)$. El punto (x_1,y_1) estará dado por $(x_0,f(x_0))$ y la pendiente m por $f'(x_0)$. Sustituyendo dichos valores la ecuación queda así:

$$y-f(x_0)=f'(x_0)(x-x_0)$$

Para encontrar el intercepto de la recta con el eje de las equis basta con despejar x e igualar y con cero:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

donde x es el nuevo valor inicial, x_1.

Puesto que cada nuevo valor inicial estará dado en función del valor de la iteración anterior, se reescribe la fórmula de la siguiente manera:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

que es la fórmula de Newton que se quería demostrar.

Generalizando el modelo se obtiene el método de Newton para resolver sistemas de ecuaciones no lineales en términos matriciales[9]:

$$\mathbf{x}_{n+1}=\mathbf{x}_n-F(\mathbf{x}_n)^{-1}\,\mathbf{f}(\mathbf{x}_n)\quad n \geq 0$$

donde:

$$\mathbf{x}_n=\begin{bmatrix} x_{1,n} \\ x_{2,n} \\ \cdot \\ \cdot \\ \cdot \\ x_{k,n} \end{bmatrix} \qquad \mathbf{x}_{n+1}=\begin{bmatrix} x_{1,n+1} \\ x_{2,n+1} \\ \cdot \\ \cdot \\ \cdot \\ x_{k,n+1} \end{bmatrix} \qquad \mathbf{f}(\mathbf{x}_n)=\begin{bmatrix} f_1(x_{1,n}, x_{2,n}, \cdots, x_{k,n}) \\ f_2(x_{1,n}, x_{2,n}, \cdots, x_{k,n}) \\ \cdot \\ \cdot \\ \cdot \\ f_m(x_{1,n}, x_{2,n}, \cdots, x_{k,n}) \end{bmatrix}$$

$$\mathbf{F}(\mathbf{x}_n)=\begin{bmatrix} \dfrac{\partial f_1(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_1} & \dfrac{\partial f_1(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_2} & \cdots & \dfrac{\partial f_1(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_k} \\ \dfrac{\partial f_2(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_1} & \dfrac{\partial f_2(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_2} & \cdots & \dfrac{\partial f_2(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \dfrac{\partial f_m(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_1} & \cdots & & \dfrac{\partial f_m(x_{1,n}, x_{2,n}, \cdots, x_{k,n})}{\partial x_k} \end{bmatrix}$$

$F(\mathbf{x}_n)$ es el jacobiano de $\mathbf{f}(\mathbf{x}_n)$, y $F(\mathbf{x}_n)^{-1}$ es la inversa de $F(\mathbf{x}_n)$

[9] Atkinson 1989 103-110

El sistema de ecuaciones que se resuelve con este sistema es:

$$\mathbf{f(x)=0}$$

donde $\mathbf{f(x)\approx 0}$

2.5 Evaluación del ajuste de modelos de regresión logística

Puesto que los modelos estadísticos son estimados a partir de variables aleatorias, siempre existirá un error asociado a ellos, el cual será menor en la medida en que mejor sea el ajuste del modelo a los datos. Por esta razón es de vital importancia evaluarlo, lo que nos dará una idea de la validez del modelo para inferir sobre el fenómeno que se está investigando.

La evaluación del ajuste de los modelos de regresión logística se lleva a cabo de manera similar a la evaluación de modelos de regresión ordinarios, la diferencia se debe a que la distribución de la variable respuesta binaria es binomial y no normal. Existen estadísticos específicos para esto, mismos que serán presentados en esta sección.

2.5.1 Cálculo de residuales para regresión logística

Un residual es un estadístico que mide las diferencias entre los valores observados y los valores estimados.

El cálculo de los residuales más simples está dado por la diferencia $y_i - \hat{y}_i$ (i-ésimo residual "crudo"); el valor real de la i-ésima observación menos su correspondiente estimado.

En regresión logística este resultado es engañoso, porque la precisión[10] de $\hat{y}_i$, está dada en función del número de observaciones n_i y de la probabilidad $\hat{p}_i$; mientras más grande sea n_i, será más preciso el valor de $\hat{y}_i$.

Un estimador de las diferencias entre el valor observado, y el estimado, será aquel que pondere el efecto de la precisión; esto se logra dividiendo a la diferencia $y_i - \hat{y}_i$ por el error estándar de y_i, con lo cual se obtiene el estadístico X_i:

$$X_i = \frac{y_i - \hat{y}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$$

Estos residuales son conocidos como residuales de *Pearson*, porque la suma de sus cuadrados es igual a la estadística X^2 de *Pearson*: $X^2 = \sum X_i^2$.

Debido a que no se considera la variación inherente a los valores estimados $\hat{y}_i$, estos residuales no se aproximan a una varianza unitaria, para lo cual un mejor procedimiento consiste en dividirlos por su error estándar ($y_i - \hat{y}_i$) en lugar de hacerlo por el de y_i.

---

[10] Notese que: $\hat{p}_i = \frac{\hat{y}_i}{n_i}$, de donde $\hat{y}_i = \hat{p}_i n_i$ y el error estándar de $\hat{y}_i$ está en función del error de $\hat{p}_i$

La fórmula del error estándar de $y_i - \hat{y}_i$ está dada por la siguiente expresión[11]:

$$\text{s.e.}\ (\ y_i - \hat{y}_i) = \sqrt{\hat{v}_i(1 - h_i)}$$

donde $\hat{v}_i = n_i \hat{p}_i(1 - \hat{p}_i)$, y h_i es el i-ésimo elemento de la diagonal de la matriz de "apalancamiento" o matriz "sombrero" H, llamada así porque es la que "le pone el sombrero a la variable estimada". En Collet (1991)[12] se expone con más detalle esta matriz y se presenta la siguiente expresión para obtener los valores de la diagonal: $h_i = n * \hat{p} * (1 - \hat{p}) * s^2$, donde s es la desviación estándar del predictor lineal $\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$.

Dividiendo los residuales $y_i - \hat{y}_i$ entre s.e.$(\ y_i - \hat{y}_i)$ se obtienen los residuales de Pearson estandarizados:

$$r_{P_i} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{v}_i(1 - h_i)}}$$

Se han desarrollado otros estadísticos alternativos para el análisis de residuales, ya que los de Pearson presentan el inconveniente de que su aproximación a la distribución normal disminuye cuando son relativamente grandes y las probabilidades fijadas son cercanas a cero o uno[13]. Una mejor opción consiste en utilizar los residuales de la devianza; éstos se forman a

[11] Collett (1991) p.g. 122
[12] Ver Collet (1991) p.g. 122, 328
[13] Collett (1991) p. g. 125-126

partir de la función de máxima verosimilitud del modelo estimado. El cálculo de ellos está dado por la siguiente expresión[14]:

$$d_i = \text{sgn}(y_i - \hat{y}_i)\sqrt{2y_i \log(\frac{y_i}{\hat{y}_i}) + 2(n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{y}_i})}$$

Los residuales de la devianza se estandarizan por[15]:

$$r_{D_i} = \frac{d_i}{\sqrt{(1-h_i)}}$$

Bajo ciertas suposiciones, el estadístico $D = \sum d_i^2$ se distribuye asintótica o aproximadamente como una X^2_{n-p}, donde *n* es el número de observaciones y *p* es el número de parámetros estimados; al igual que la X^2 de Pearson es una medida de bondad de ajuste general.

En McCullagh (1983) p.g. 118-119 se describen las suposiciones paramétricas para este estadístico y se plantea bajo qué condiciones es válida la aproximación a la distribución χ^2. También se establece que cuando se utiliza esta estadística para evaluar la significancia de la diferencia de desviaciones debido a la inclusión de términos a un modelo de regresión logística, el método es válido incluso cuando no se cumplen las suposiciones de aproximación para *D*.

[14] En McCullagh (1983) p.g. 118 se deduce la función de desviacion.
La expresión que aquí se presenta está tomada de Collett(1991) p.g. 122
[15] Collett (1991) p.g. 123

2.5.2 Análisis de residuales

El análisis de las gráficas de residuales sirve para detectar malos ajustes. Estos pueden ser ocasionados por valores atípicos en las variables explicativas o en la variable respuesta.

En caso de encontrar atipicidades, se procederá a identificar si éstas se deben a errores de escritura, en cuyo caso serán corregidos. En caso contrario se llevarán a cabo ajustes **adicionales** eliminando a las observaciones atípicas. Con esto no se estará excluyendo al modelo original que incluye a las atipicidades (es incorrecta la práctica de eliminar observaciones sin justificación alguna), de hecho se utilizará la información proporcionada por ambos modelos en la elaboración de las conclusiones.

Existen diferentes opciones para el cálculo de residuales en regresión logística, dependiendo de la manera como sean calculados; residuales de Pearson, de la devianza y de Anscombe. Los terceros son una suma ponderada de los dos primeros que se aproxima mejor a normalidad.

Los residuales de Pearson difieren considerablemente cuando los valores de la probabilidad estimada por el modelo de regresión logística son cercanos a uno o a cero, y los valores de los residuales son relativamente grandes[16]. Los residuales de la devianza en cambio, no presentan tal inconveniente, y además se ajustan mejor a la distribución normal, por lo que es más conveniente su uso en modelos de regresión logística.

[16] Collett p.g. 125

Existe una gran variedad de gráficas de residuales especiales para regresión logística. Por razones prácticas se seleccionó un grupo de ellas a manera de obtener la mayor información posible evitando la redundancia de información y simplificando el análisis. Las gráficas de residuales seleccionadas son las siguientes: "Residuales índice[17]", "Delta-Beta", "Residuales de apalancamiento[18]", "residuales Ci" y "Residuales parciales[19]".

Para el valor de los residuales *estandarizados* se cumple la regla empírica que establece que aproximadamente el 68% de las observaciones estarán contenidas en ± una desviación estándar, el 95% de las observaciones estarán contenidas en ± dos desviaciones estándar y casi todas las observaciones estarán contenidas en ± tres desviaciones estándar.

Este razonamiento aporta información adicional en la detección de atipicidades o valores extremos al establecer que la mayoria de los residuales estarán contenidos en el intervalo[-2,2], y que los valores que estén fuera del intervalo [-3,3] serán muy probablemente valores atípicos; por ejemplo un residual de seis indica que existe una atipicidad.

2.5.3 Residuales índice

Consisten en la gráfica del número o "índice" de cada observación contra los residuales de la devianza, de tal forma que se vea una nube de puntos en la que sobresalgan los residuales extremos o atípicos o se observe alguna seriación.

---

[17] Traducido del inglés "Index residuals"
[18] Traducido del ingles "Hat Residuals"
[19] Del inglés "Partial Residual".

2.5.4 Residuales Delta-Beta o DFBetas[20]

Evalúan el efecto sobre el valor de cada parámetro estimado $\hat{\beta}_j$ debido a la eliminación de la observación *i*, donde $j = \overline{1,p}$ para *p* parámetros estimados, $i = \overline{1,n}$ para *n* observaciones. Dicho en otras palabras, evalúa el cambio en los parámetros si se estimara un modelo por cada i-ésima observación eliminada; la estadística Delta-Beta es la diferencia estandarizada en cada parámetro estimado debido a la eliminación de la i-ésima observación.

La estadística delta-beta está dada por la siguiente expresión[21]:

$$\Delta_i \hat{\beta}_j = \frac{(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}_{j+1}\, \mathbf{x}_i (y_i - \hat{y}_i)}{(1-h_i)\, s.e.(\hat{\beta}_j)}$$

Donde $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}_{j+1}$ es el (*j*+1) ésimo renglón de la matriz de varianzas y covarianzas de los parámetros estimados, $\mathbf{x}_i$ es el vector de las variables explicativas para la i-ésima observación y h_i es el i-ésimo elemento de la diagonal de la matriz H[22].

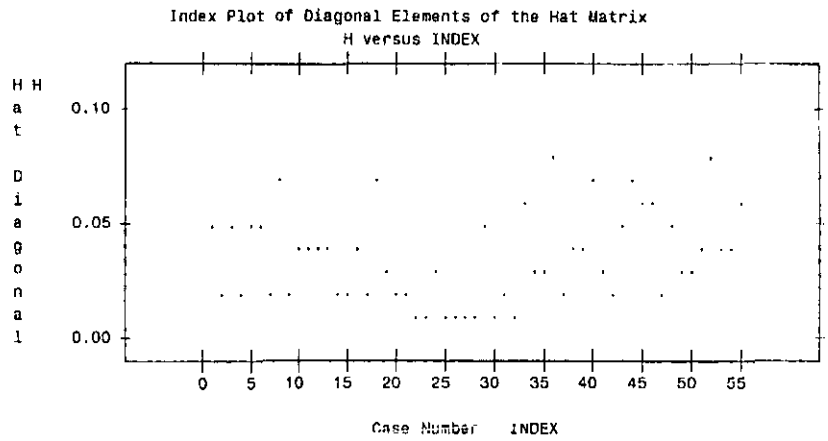
2.5.5 Residuales de apalancamiento

Esta gráfica sirve para detectar valores extremos en las variables explicativas. El eje vertical corresponde a la diagonal de la matriz H (h_i) y el horizontal al "*número índice*". Es importante la identificación de estos valores porque provocan atipicidades en el ajuste del modelo. En McCullagh (1983)

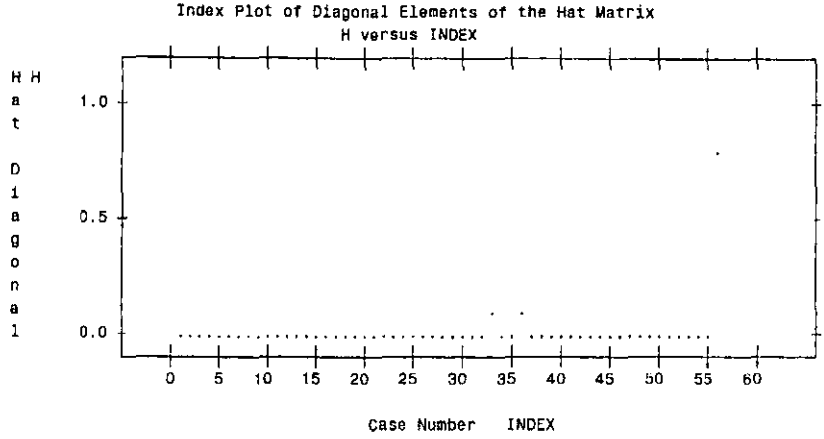
[20] SAS/STAT Volume 2 p. 1094
[21] Collett p.g. 152

p.g. 405 se sugiere que los valores de $h_i > \frac{2p}{n}$ indican puntos de alto nivel de apalancamiento. En la práctica se ha encontrado que no son importantes los valores extremos de esta gráfica, pero si se llegara a encontrar algún residual que se aleje considerablemente del resto, la gráfica indica la existencia de una observación que por ser atípica en la variable explicativa será influyente en la estimación de los parámetros, y que muy probablemente proviene de un error.

A continuación se muestran dos gráficas de residuales de apalancamiento en las que se ajustó la variable peso. En la primera no se observa ningún punto sobresaliente, ocurriendo lo contrario en la segunda, ya que se ha creado un error ficticio en los datos para proporcionar un ejemplo:



[72] Esta matriz fue explicada en la sección 2.5.1, Cálculo de residuales para regresión logística



La observación sobresaliente de la segunda gráfica se debe a que se asignó un valor de 500 kg al peso de una de las personas para crear un dato atípico en esta variable.

2.5.6 Residuales C_i y $\bar{C}_i$

Son diagnósticos de desplazamiento que proporcionan mediciones escalares de la influencia de las observaciones individuales sobre el **conjunto de parámetros** estimados $\mathbf{b}$.

Las ecuaciones de estos residuales están dadas por[23]:

$$C_i = \frac{\chi_i^2 h_i}{(1-h_i)^2}$$

$$\bar{C}_i = \frac{\chi_i^2 h_i}{(1-h_i)}$$

Donde χ_i son los residuales de Pearson y h_i es el i-ésimo elemento de la diagonal de la matriz H

Esta gráfica es útil para identificar una o algunas observaciones atípicas que influyen fuertemente sobre la significancia del modelo estimado en su conjunto.

2.5.7 Residuales DifDev

Sirven para detectar observaciones mal ajustadas por el modelo; observaciones que difieren fuertemente de los valores predecidos por el modelo ajustado.

DifDev es el cambio en la devianza debido a la eliminación de la i-ésima observación

[23] SAS/Stat Volume 2 p. 1094

Las ecuaciones para el cálculo de DifDev son respectivamente[24]:

$$\Delta_i D = d_i^2 + \overline{C}_i$$

Donde d_i es el i-ésimo residual de la devianza,

$\overline{C}_i$ es el la estadística definida en la sección anterior

2.5.8 Residuales parciales

Sirven para evaluar si es necesario aplicar alguna transformación a las variables explicativas continuas. Están dados por la siguiente expresión[25]:

$$\frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}} + \hat{\beta}_j x_{ij}$$

El valor de esta expresión se grafica contra el de la variable explicativa x_{ij}, generando una gráfica para cada j-ésimo parámetro.

En esta gráfica se identifica si el comportamiento de la variable continua es no lineal, al mismo tiempo que se sugiere qué tipo de comportamiento no lineal es el que sigue; si la gráfica muestra una tendencia lineal, la variable no deberá transformarse, en caso contrario la tendencia de la gráfica indicará qué tipo de transformación deberá aplicarse v.g. logarítmica, exponencial, polinomial, etc.

[24] SAS/Stat Volume 2 p. 1095
[25] Collett p.g. 135

2.6 Significancia por inclusión de variables independientes adicionales

El método de reducción de la devianza por inclusión de variables adicionales consiste en restar a la suma de los cuadrados de la devianza del modelo inicial (D_A) la del modelo con términos independientes agregados (D_B); (D_A-D_B). La diferencia así obtenida se aproxima a la distribución X^2 con: *n - p - (número de términos adicionales)* grados de libertad. En general se utiliza esta estadística para decidir si se deben o no incluir variables adicionales.

2.7 Intervalos de confianza y pruebas de significancia

2.7.1 Significancia de los parámetros estimados

Serán utilizados dos métodos alternativos para medir la significancia de los parámetros estimados: uno está basado en la reducción de la devianza del modelo por inclusión de variables (sección anterior, 2.6) y el otro en una prueba de hipótesis conocida como estadística χ^2 de Wald.

2.7.1.1 Estadística χ^2 de Wald:

La estadística χ^2 de Wald está dada por el cuadrado de la razón del parámetro estimado sobre su error estándar estimado[26]:

$$\chi^2 = \left(\frac{\hat{\beta}}{s.e.(\hat{\beta})} \right)^2$$

Esta estadística evalúa la hipótesis nula $\beta_i = 0$ para el parámetro deseado utilizando una distribución χ^2 con un grado de libertad.

---

[26] Agresti 1990, pg 89

La salida de los procedimientos en SAS que estiman modelos de regresión logística incluye un valor *p* que está calculado con base a la estadística χ^2 de Wald y bajo la misma hipótesis nula, por lo que ésta será utilizada para evaluar la significancia de los predictores estimados.

2.7.2 Intervalos de confianza para razones de momios ajustadas

i.- En la sección 2.3.3 (*razones de momios ajustadas*) se estableció que la razón de momios $\hat{\psi}_i$ está dada por: $\hat{\psi}_i = e^{\Delta x_i \hat{\beta}_i}$

Si se conoce $s.e.\{\hat{\beta}_i\}$, es posible establecer límites de confianza para el parámetro estimado $\hat{\beta}_i$ y obtener los intervalos inferior y superior de la ecuación de regresión $\Delta x_i \hat{\beta}_i$ que al sustituirse en $\hat{\psi}_i = e^{\Delta x_i \hat{\beta}_i}$ proporcionen a su vez los intervalos de confianza para $\hat{\psi}_i$

ii.- El error estándar $s.e.\{\hat{\beta}_i\}$ está dado por $\hat{\mathbf{x}}' Cov(\hat{\mathbf{x}}) \hat{\mathbf{x}}$, donde $Cov(\hat{\mathbf{x}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ es la matriz estimada de covarianzas de los parámetros estimados, y $\mathbf{x}$ es el vector de parámetros independientes[27]. Este dato, $s.e.\{\hat{\beta}_i\}$ se obtiene directamente de la salida de SAS junto con los valores de los parámetros estimados.

[27] SAS/STAT User's guide vol 2 p. 1091, Agresti 1990, p. 89

iii.- Por lo anterior, el intervalo de confianza al $100(1-\alpha)\%$ de la razón de momios $\hat{\psi}_i$ estará dado por:

$$\left[e^{\Delta x_i\left(\hat{\beta}_i - z_{\alpha/2} s.e.\left\{\hat{\beta}_i\right\}\right)}, e^{\Delta x_i\left(\hat{\beta}_i + z_{\alpha/2} s.e.\left\{\hat{\beta}_i\right\}\right)} \right]$$

o bien aplicando las propiedades distributiva y de los exponentes y sustituyendo $e^{\Delta x_i \hat{\beta}_i}$ por $\hat{\psi}_i$:

$$\left[\frac{\hat{\psi}_i}{e^{\Delta x_i z_{\alpha/2} s.e.\left\{\hat{\beta}_i\right\}}}, \hat{\psi}_i e^{\Delta x_i z_{\alpha/2} s.e.\left\{\hat{\beta}_i\right\}} \right]$$

CAPITULO 3
ESTIMACIÓN DE RAZONES DE MOMIOS RESPECTO A LA VARIABLE RESPUESTA *LESION*

En este capítulo serán estimados modelos de regresión logística ajustados a la variable de respuesta binaria *lesion*, la cual toma valores iguales a uno en las personas que presentaron síntomas de arsenicismo, e iguales a cero en los demás casos. El objetivo de este modelo es identificar y dimensionar la posible relación existente entre los niveles de metilación de arsénico[1] y la presencia de lesiones.

Una vez identificadas las especies de metilación de arsénico que resulten significativas, serán calculadas a partir de los modelos de regresión logística estimados, las razones de momios para evaluar el grado de asociación entre los niveles de metilación del arsénico y la presencia de lesiones.

El contenido del presente capítulo consta de cuatro partes principales: *Descripción de la información utilizada, Metodología para la selección de variables, Estimación de los parámetros de regresión logística* y *Cálculo de razones de momios e intervalos de confianza.* En la primera se describe el contexto del estudio así como las variables utilizadas. La segunda está enfocada a explicar las acciones consideradas para seleccionar de todo el conjunto de información disponible, a las variables que finalmente fueron

utilizadas en los modelos estimados. La tercera muestra cada etapa de la estimación de parámetros y por último en la cuarta parte se calculan las razones de momios y sus intervalos de confianza a partir de los parámetros de regresión estimados.

3.1. Descripción de la información utilizada

La información fue recopilada por investigadores de la sección de toxicología ambiental en el departamento de farmacología y toxicología ambiental del Centro de Investigaciones y Estudios Avanzados (CINVESTAV) del Instituto Politécnico Nacional. Consiste en la aplicación de una encuesta y de un análisis de orina a una muestra de los habitantes de un poblado de la Comarca Lagunera llamado Ampueros que están crónicamente expuestos a arsénico porque el agua que utilizan para beber, que procede de la perforación de pozos, presenta una alta concentración de arsénico ($393 \mu g/l$). Esto se debe a que la tierra de manera natural es muy rica en este mineral.

La información recopilada se capturó en una base de datos de 180 observaciones (cada persona corresponde a una observación). En ella se cuenta con información clínica, de hábitos alimenticios y farmacológicos, de fuentes alternas de exposición a arsénico y actividades que por sus características pueden incrementar el grado de exposición o bien alterar el proceso de metilación del mismo en el organismo, así como de los resultados del análisis de las muestras de orina, que serán identificadas como **especies de arsénico**, y por último se indica a las personas que

[1] La metilación es un proceso de desintoxicación del organismo que consiste hacer orgánica una sustancia inorgánica al añadirle carbonos

presentaron síntomas que se sabe son consecuencia de la exposición a arsénico.

3.1.1 Variable respuesta

Es la variable *lesion* que identifica cuales son las personas que presentaron síntomas de arsenicismo.

En principio se planeaba hacer una distinción entre personas con síntomas específicos e inespecíficos; los primeros provocados exclusivamente por el arsénico mientras que los otros además del arsénico por otras causas. Inicialmente se hicieron estimaciones para ambas variables respuesta, llegando a obtener los mismos resultados.

La explicación está dada por la siguiente tabla donde se observa que solamente existe una persona con síntomas inespecíficos que no presenta síntomas específicos. Por esta razón se considera únicamente a la variable *lesion* que considera como casos a las personas que hayan presentado ya sea síntomas específicos o inespecíficos.

Síntomas específicos	Síntomas inespecíficos	Número de personas
0	0	142
0	1	1
1	0	25
1	1	12

Nota: El número 0 significa que la persona no presentó síntomas, mientras que el 1 es lo contrario

3.1.2 Variables de identificación

Es necesario contar con alguna forma de identificar a las personas que conforman cada observación, en caso de que se requiera hacer alguna verificación

Para tal efecto existen las variables *numo* y *fam* correspodientes al código de observación por persona y por familia respectivamente, asi como *nombre* y *poblado*.

Variable	*Descripción*
Numo	Codigo de observación
Fam	Código por familia
Nombre	Nombre de la persona
Poblado	Código del poblado

3.1.3 Especies de arsénico

La metilación del arsénico consiste en un proceso de desintoxicación mediante el cual el organismo asocia carbonos al arsénico inorgánico, convirtiéndolo en arsénico orgánico, el cual se supone es menos tóxico y se elimina más rápidamente en la medida en que cuente con mayor número de carbonos. En el presente estudio se consideran tres especies de arsénico, el inorgánico y dos orgánicos; el monometilado con un carbono y el dimetilado con dos según se muestra en la siguiente tabla:

Especie	*Nombre de Variable*	*Carbonos*
Arsénico inorgánico	ASI	0
Arsénico monometilado	MMA	1
Arsénico dimetilado	DMA	2

Cada especie se encuentra registrada en tres tipos de unidades:

Unidad	*Nombre de variables*
Cantidad registrada en μg/l (micro gramos por litro)	ASI, MMA, DMA
Especie dividida entre a la cantidad de creatinina[2] encontrada	ASIG, MMAG, DMAG
Valor porcentual de cada especie con respecto al total de arsénico encontrado (orgánico+inorgánico)	ASIP, MMAP, DMAP

Adicionalmente a éstas, se evaluará el efecto de los siguientes cocientes de especies de arsénico:

Cociente	*Nombre de variable*
MMA / ASI	MMAASI
DMA / ASI	DMAASI
ASO / ASI	ASOASI
DMA / MMA	DMAMMA

donde ASO es el arsénico orgánico (MMA + DMA)

3.1.4 Agrupación de variables confusoras y especies de arsénico

Para facilitar el proceso de selección e identificación de variables, éste se llevará a cabo mediante la formación de grupos, analizando primeramente por separado a cada variable perteneciente a un grupo para después evaluarlas de manera conjunta.

[2] Este ajuste se llevó a cabo para hacer comparables las cantidades de arsénico detectadas en la orina debido a las diferencias en las cantidades de agua ingeridas

Se considera variable confusora a toda aquella que ejerce una relación sobre la variable respuesta adicionalmente a la variable explicativa que se desea estudiar; Si la variable objetivo en el estudio deja de ser significativa al añadir alguna variable confusora, es importante considerar las siguientes opciones:

- Existe una alta correlación entre ambas y por ello se restan importancia entre si.
- Existe alguna relación de dependencia.

Exceptuando a las variables Edad, Sexo, Peso, Estatura y las relacionadas con la alimentación, todas son iguales a uno dependiendo de que se cumpla o no la característica que representan.

Las variables referidas a los hábitos alimenticios se codificaron de la siguiente manera:

0 = No consime el alimento
1 = Si lo consumre
2 = Se consumio recientemente

A continuación se muestra la agrupación de las variables que se definió para el proceso de selección:

Grupo 1.- Características físicas de las personas:

Variable	*Descripción*
Edad	Edad
Sexo	Sexo
Peso	Peso
Estatura	Estatura

Grupo 2.- Fuentes alternas de exposición a arsénico:

Variable	*Descripción*
AnosResi	Años de residir en el poblado
TWE	Tiempo de residir en el poblado
Fertiliz	Exposicióna fertilizantes
Insect	exposición a insecticidas
Herbicid	exposición a herbicidas
Cult_Uva	trabaja en cultivo de uva
Maquilad	trabaja en maquiladora

Grupo 3.- Tabaquismo y alcoholismo:

Variable	Descripción
Tabaco	Hábito de fumar
Alcohol	Hábito de beber alcohol

Grupo 4.- Hábitos de consumo:

Variable	Descripción
PescMar	Pescados o mariscos
Higado	Higado
Huevos	Huevos
CarneRes	Carne de Res
CarnePue	Carne de Puerco
CarnesFr	Carnes Frías
Frijoles	Frijoles
Chicharo	Chicharo
Cereal	Cereal
Pan	Pan
Zanahori	Zanahoria
VerdLeg	Verd y Legumbres
Melon	Melon
Papa	Papa
Calabaza	Calabaza
Vitamina	Vitamina
Sulfas	Consumo de sulfas
Anticonc	Consumo de anticoneptivos
Anticonv	Consumo de anticonvulsionantes
OtrosMed	Consumo de otros medicamentos

Grupo 5.- Cuadro Clínico:

Variable	Descripción
Gripa	Gripa
Sarampio	Sarampión
Rubeola	Rubéola
Anginas	Anginas
Paperas	Paperas
Pulmonia	Pulmonia
Bronquit	Bronquitis
Diarrea	Diarrea
AntF	Antecedentes familiares
AntP	Antecedentes personales

Grupo 6.- Especies de arsénico en la orina:

Creat ASIG MMAG DMAG ASOG SUMESG	Ajustadas por creatinina
ASIP MMAP DMAP ASOP	Porcentaje de especies (ASOP=MMAP+DMAP)
ASI MMA DMA SumEsp	Cantidad sin ajustar
MMAASI DMAASI ASOASI DMAMMA	Razones de especies

3.2 Metodología utilizada para la selección de variables

Esta se llevó a cabo mediante los siguientes pasos:

1. Selección de variables significativas al agregar cada una al modelo con únicamente el intercepto
2. Formación de grupos con las variables afines que al juntarse conserven su significancia
3. Unión de grupos y selección de variables que aún conserven su significancia

La selección de variables en los pasos 2 y 3 se llevó a cabo eliminando y en algunos casos reincorporando variables con base a la estadística wald de significancia y por los métodos automáticos "Setpwise" y "Backward".

El primer método (el no automático) es más intuitivo y consiste en ir eliminando variables partiendo de las de menor significancia, y ocasionalmente evaluando si alguna variable eliminada que se considera muy importante vuelve a ser significativa.

A continuación se describen los métodos de selección automáticos que se ejecutaron en SAS:

- Backward simplemente comienza con todas las variables y las elimina una por una, comenzando por la de menor significancia.
- Stepwise va añadiendo y evaluando la significancia de cada variable una por una (método Forward), comenzando por la más importante y con la ventaja de que cada vez que se añade una variable se ejecuta uno o varios pasos Backward de eliminación. El proceso termina hasta que no puedan ser añadidas más variables.

Adicionalmente a los pasos de selección, se revisaron las siguientes gráficas de residuales: ***índice***, ***Delta-Beta*** *o* ***DFBetas***, ***de apalancamiento***, C_i *y* $\overline{C}_i$, ***DifDev***, y para las variables continuas se analizaron adicionalmente la gráficas de ***residuales parciales*** con el objetivo de identificar si es necesario llevar a cabo alguna transformación.

En los ajustes se dio especial importancia a las especies de arsénico. Debido a que existe una relación entre ellas, éstas fueron consideradas por separado, y sólo cuando resultaron conjuntamente significativas se evaluaron en un mismo modelo.

A partir de los modelos resultantes de la selección y estimación se procedió a calcular las razones de momios para las especies que resultaron significativas.

3.3 Estimación de los parámetros de regresión logística

En esta sección se da seguimiento al análisis que se llevó a cabo para seleccionar las variables, verificar la existencia de errores o datos atípicos que afectaran las estimaciones y por último la obtención del modelo de regresión logística con los parámetros y errores estándar estimados que serán utilizados para calcular las razones de momios.

Al final de esta sección se muestra un cuadro resumen comparativo con los resultados obtenidos por cada método.

3.3.1 Macros auxiliares en SAS

Debido a que se evaluó un número considerable de variables, se crearon los siguientes macros o programas en SAS:
%FitLogit(DataIn,y,X,Grupo,Graficas,Cond)
y %FitXVar(Data,y,X,XFijas,Grupo,Graficas,Cond)
los cuales facilitan los procesos de selección de variables y de estimación de parámetros.

El apéndice III contiene el listado del código de estos macros así como una breve descripción del uso de cada parámetro. A continuación se explicará en términos generales lo que hace cada macro.

3.3.1.1 Macro %FitLogit(DataIn,Y,X,Grupo,Graficas,Cond)

Sus principales funciones son:

- Estimar el modelo de regresión logística Y=X (donde Y es la variable respuesta y X un listado de las variables explicativas)

- A partir de un archivo de entrada con observaciones bernoulli (en la base de datos del presente trabajo cada observación es una persona), contar el número de casos y no casos (casos:y=1, no casos:y=0) para cada combinación de valores de las variables explicativas (X), de tal forma que los datos bernoulli sean agrupados para formar conjuntos de distribuciones binomiales[3].
- Calcular los residuales de la devianza
- Generar los archivos de datos SAS de salida "Ajuste" y "Residual". El primero contiene la suma de las devianzas cuadradas del modelo y el segundo los residuales (estos archivos serán utilizados para calcular algunas gráficas y estadísticos).
- Generar un listado de las observaciones bernoulli que conforman a cada binomial para que en caso de que sea necesario, identificar a que observación (o personas) pertenece cada punto de los mostrados en las gráficas de residuales.
- Eliminar fácilmente a las observaciones que se desee por medio del parámetro &Cond, en caso de que se requiera evaluar el efecto de eliminar alguna observación.

Este macro fue utilizado en los pasos 2 y 3.

[3] Inicialmente se trabajó con observaciones bernoulli, y se hicieron pruebas para identificar si había alguna diferencia al hacerlo con los datos en forma binomial. Nada cambió excepto la inteligibilidad de las gráficas de residuales, pues era considerablemente mayor el número de observaciones en el caso bernoulli, resultando más clara la interpretación de las gráficas procedentes de datos binomiales.

3.3.1.1.1 Ejemplo

A continuación se muestra un ejemplo de la salida de este macro, en el cual se estimará el modelo:

$$Logit\left(P\left[Lesion = 1\right]\right) = \beta_0 + \beta_1 Edad + \beta_2 Alcohol$$

donde *Principa* es el nombre del archivo de datos SAS con la base de datos

Invocación al macro: %FitLogit(Principa,Lesion, Edad Alcohol,,,)

Salida en SAS:

Salida 1:

AJUSTE DE: Lesion = Edad Alcohol

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.7774475	0.737974	41.90918	0.0001	Intercept
EDAD	1	0.11574503	0.020325	32.42929	0.0001	Edad
ALCOHOL	1	0.36253555	0.55799	0.422132	0.5159	

Salida 2:

AJUSTE DE: Lesion = Edad Alcohol

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-4.7774	0.7380	41.9092	0.0001	.	.
EDAD	1	0.1157	0.0203	32.4293	0.0001	1.142601	1.123
ALCOHOL	1	0.3625	0.5580	0.4221	0.5159	0.074729	1.437

Se han mostrado dos salidas, la primera generada con el procedimiento SAS "Proc Probit" y la segunda con "Proc Logistic". Esto se debe a que el macro %FitLogit los utiliza de manera alternativa según convenga de acuerdo a lo que se explica a continuación:

Proc Probit:

Fue diseñado para realizar análisis de regresión logística utilizando la función de liga *probit*, aunque si se especifica por medio de un parámetro en su lugar utiliza la función *logit*.

Ventajas: Para estimar los parámetros utiliza una versión estabilizadora del algoritmo de *Newton-Raphson*, el cual en la práctica no presentó problemas de convergencia bajo situaciones de *sobreparametrización*.

Desventajas: Este procedimiento no genera gráficas

Proc Logistic:

Fue diseñado para realizar análisis de regresión logística utilizando la función de liga *logit*.

Ventajas: Mediante los parámetros *iPlot* o *influence* genera un conjunto de gráficas de residuales especialmente diseñadas para regresión logística

Desventajas: Para estimar los parámetros utiliza un algoritmo iterativo conocido como "*Mínimos cuadrados reponderados*", el cual en la práctica mostró no converger en situaciones de sobreparametrización.

Por las razones anteriormente expuestas, se incluyó una condición en el macro %FitLogit para que utilizara el "*Proc Logistic*" únicamente en el caso en el que se especificara ya sea la opción *iplots* o bien *influence* en el parámetro &Graficas. De esta forma, en caso de presentarse algún problema en la estimación de parámetros utilizando *Proc Logistic*, podrá analizarse mediante los resultados del *Proc Probit* la causa. En algunas corridas se detectó que estos problemas ocurrían cuando se presentaron situaciones de *sobreparametrización*, identificándose estos casos en la salida de *Proc Probit*, ya que muestra valores exageradamente atípicos en el error estándar.

3.3.1.2 Macro %FitXVar(Data,y,X,XFijas,Grupo,Graficas,Cond)

Este macro fue creado para calcular el valor p de significancia debido a la reducción de la devianza por el incremento de variables explicativas, considerando las hipótesis nula Ho: no existe diferencia en la reducción de la devianza por el incremento de las variables, y Ha: existe una diferencia en la reducción de la devianza al añadir las variables. La reducción de la devianza se calcula restando a la devianza del modelo ajustado al intercepto, la resultante de añadir cada variable que se desea evaluar.

Su funcionamiento consiste en simplemente invocar dos veces al macro %FitLogit por cada variable a evaluar, la primera evaluando únicamente al intercepto (opcionalmente a las variables indicadas en el parámetro XFijas) y la siguiente añadiendo la variable, para posteriormente calcular la diferencia de las devianzas entre el modelo reducido y aumentado con la(s) variable(s) especificada(s) en el parámetro X.

En el presente trabajo se utilizó este macro para la selección del paso1; identificación de variables significativas al añadirlas por separado al modelo con únicamente el intercepto.

3.3.1.2.1 Ejemplo

A continuación se muestra con un ejemplo la invocación al macro y la salida que éste genera. Para tal efecto se evaluará la significancia de las variables explicativas *Edad Alcohol* sobre la variable respuesta *Lesion*:

Invocación al macro: %FitXVar(Principa,Lesion,Edad Alcohol,,,,)

Salida en SAS:

Analisis de la devianza para regresión logistica
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	Edad	54.1834	54	106.601	55	.0000000	56	2
	Alcohol	0.0000	0	9.890	1	.0016620	2	2

En la salida se observan bajo las columnas DevFijo y DFFijo los valores correspondientes a la devianza y grados de libertad del modelo fijando únicamente al intercepto, y bajo DevModel y DFModel los correspondientes valores una vez que se ha añadido por separado cada una de las variables explicativas X (Edad Alcohol). Se incluyeron también las columnas NOBS y NVAR para indicar el número de observaciones binomiales así como de variables estimadas, de donde se observa fácilmente el cálculo de los grados de libertad (56-1 variable para el intercepto y 56-2 variables para el modelo añadido).

Por último, el valor p proviene de evaluar la diferencia de las devianzas DevFijo-DevModel en una distribución χ^2 con DFFijo-DFModel grados de libertad.

En cuanto a la columna XFijas que se muestra, corresponde al parámetro &XFijas, el cual se diseñó pensando en el caso en el que se desee evaluar la significancia al añadir variables a un modelo que considerara no solamente al intercepto.

3.3.2 Selección de variables

A continuación será llevado a cabo el proceso de selección de variables de acuerdo a los pasos ya indicados y los grupos previamente definidos.

3.3.2.1 Grupo 1 Características físicas de las personas:

Variables:

Edad Sexo Peso Estatura

A continuación se muestra la tabla de contribuciones a la reducción de la devianza por variable:

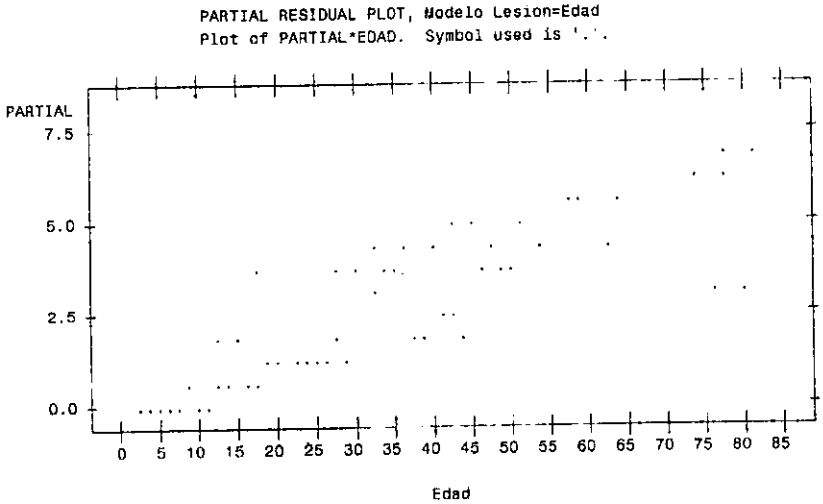
Análisis de la devianza para regresión logistica
Valor p para el modelo fijado a Lesion

X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
Edad	54.1834	54	106.601	55	0.00000	56	2
Peso	50.0636	53	64.913	54	0.00012	55	2
Estatura	40.8056	46	52.675	47	0.00057	48	2
Sexo	0.0000	0	0.010	1	0.92225	2	2

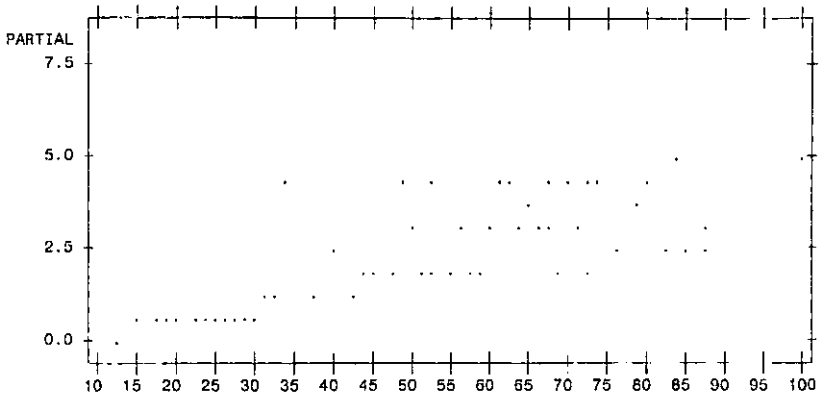
Como puede observarse, la variables Sexo no es significativa.

Residuales

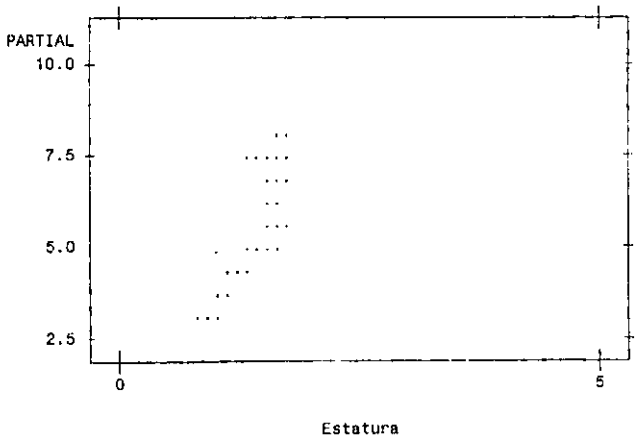
Debido a que las variables Edad Peso y Estatura no son categóricas, se generó la *gráfica de residuales parciales* para identificar si es necesario hacer alguna transformación, encontrándose que tanto para Edad como para Peso esta no se requiere, sin embargo la variable Estatura muestra un patrón que sugiere realizar alguna transformación. A continuación se muestras las gráficas mencionadas:



PARTIAL RESIDUAL PLOT, Modelo Lesion=Peso
Plot of PARTIAL*PESO. Symbol used is '.'.



PARTIAL RESIDUAL PLOT, Modelo Lesion=Estatura
Plot of PARTIAL*ESTATURA. Symbol used is '.'.



Para evaluar si realmente es necesario realizar alguna transformación se probó la disminución de la devianza para cada una de las siguientes transformaciones:

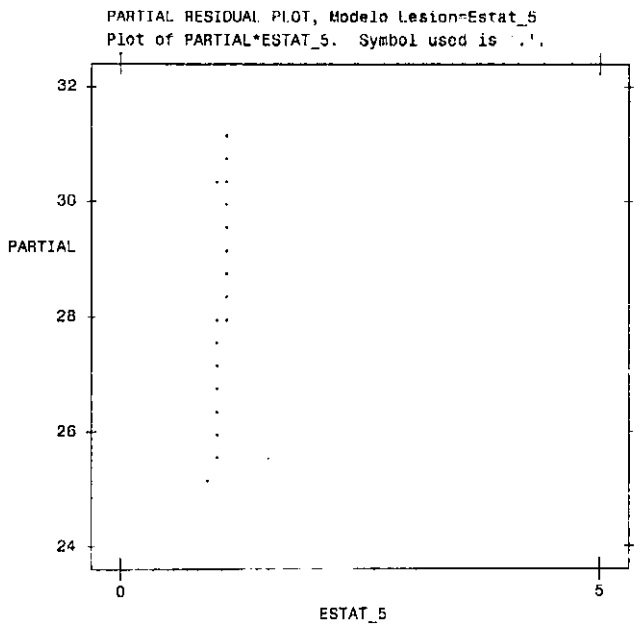
Transformación	Nombre de variable
e^x	EstatE
x^{-5}	Estat_5
x^{-4}	Estat_4
x^{-3}	Estat_3
x^{-2}	Estat_2
x^{-1}	Estat_1
$Ln(x)$	EstatL
x^2	Estat2
x^3	Estat3
x^4	Estat4
x^5	Estat5
Sin transformar	Estatura

A continuación se muestra la tabla de reducción de las devianzas para las transformaciones, ordenadas de acuerdo al nivel de significancia:

Análisis de la devianza para regresión logística
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	EstatL	39.6309	46	52.6747	47	.0003386	48	2
	Estat_5	40.0181	46	52.6747	47	.0003743	48	2
	Estat_4	40.0656	46	52.6747	47	.0003839	48	2
	Estat_3	40.1453	46	52.6747	47	.0004006	48	2
	Estat_2	40.3067	46	52.6747	47	.0004368	48	2
	Estatura	40.8056	46	52.6747	47	.0005707	48	2
	Estat2	41.8522	46	52.6747	47	.0010027	48	2
	EstatE	42.3018	46	52.6747	47	.0012788	48	2
	Estat3	42.9247	46	52.6747	47	.0017933	48	2
	Estat4	43.9788	46	52.6747	47	.0031693	48	2
	Estat5	44.9799	46	52.6747	47	.0055382	48	2

Como puede observarse, los mejores ajustes corresponden a las variables EstatL, Estat_5 y Estat_4. La gráfica de residuales parciales para EstatL tiene cierta curvatura, mientras que para Estat_5 y Estat_4 es casi idéntica y una línea recta bastante bien definida como se muestra a continuación:



No obstante que se logró un mejor ajuste con las variables transformadas, éste es apenas de orden de 0.0005707 - 0.0003839 = 0.0001868, debido a que la disminución en la devianza es muy pequeña y a que no existe algún argumento adicional, por conocimientos previos, que sugiera transformar la variable Estatura, la variable no será transformada.

Por lo anterior, serán consideradas las variables Edad, peso y Estatura.

Antes de empezar a estimar el modelo conjunto para estas variables, es importante considerar que debe existir una buena correlación entre ellas, puesto que el peso y la estatura se incrementa con la edad en la etapa de crecimiento.

Para obtener una idea de lo anterior, a continuación se mostrarán las correlaciones entre éstas variables:

Correlation Analysis

Pearson Correlation Coefficients

/ Prob > |R| under Ho: Rho=0
/ Number of Observations

	EDAD	PESO	ESTATURA
EDAD	1.00000	0.61934	0.57895
	0.0	0.0001	0.0001
	180	112	113
PESO	0.61934	1.00000	0.88083
	0.0001	0.0	0.0001
	112	112	112
ESTATURA	0.57895	0.88083	1.00000
	0.0001	0.0001	0.0
	113	112	113

Como puede observarse, efectivamente existe correlación, siendo la más importante para Peso y Estatura, muy probablemente habrá que seleccionar de entre estas variables, o de lo contrario podrían dejar de ser significativas.

Una vez considerado lo anterior, se analizará la siguiente tabla de parámetros estimados:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-4.1058	3.3481	1.5039	0.2201	.	.
EDAD	1	0.1001	0.0245	16.7391	0.0001	0.976197	1.105
PESO	1	0.00686	0.0301	0.0520	0.8196	0.075562	1.007
ESTATURA	1	-0.2412	2.8643	0.0071	0.9329	-0.032478	0.786

Como puede observarse, las variables peso y estatura perdieron significancia. Esto se debe a que o bien por estar tan fuertemente correlacionadas se están quitando significancia una a otra, o bien la variable edad es la que las hace ser significativas. Para analizar esto, será generada la misma tabla eliminado respectivamente a las variables Estatura y Peso.

A continuación se muestran las tablas resultantes:

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.3700083	1.217023	12.89338	0.0003	Intercept
EDAD	1	0.09990238	0.024277	16.93418	0.0001	Edad
PESO	1	0.00509628	0.021533	0.056012	0.8129	Peso

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.4219664	3.022711	2.140118	0.1435	Intercept
EDAD	1	0.10266016	0.023496	19.08968	0.0001	Edad
ESTATURA	1	0.17404375	2.070763	0.007064	0.9330	Estatura

Con esto se observa que debido a lo importante que es la variable Edad, pierden significancia las variables Peso y Estatura, por lo tanto sólo será seleccionada la variable Edad.

3.3.2.2 Grupo 2 Fuentes de exposición a arsénico:

Variables:

AnosResi AnosFuer TWE
Fertiliz Insect Herbicid Cult_Uva Maquilad

A continuación se muestra la tabla de reducción de la devianza por variable:

Analisis de la devianza para regresión logística
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	TWE	47.3787	42	94.9304	43	0.00000	44	2
	AnosResi	47.4519	42	94.4922	43	0.00000	44	2
	Fertiliz	0.0000	0	9.3091	1	0.00228	2	2
	Cult_Uva	0.0000	0	2.6804	1	0.10159	2	2
	Herbicid	0.0000	0	1.6549	1	0.19829	2	2
	Maquilad	0.0000	0	0.6388	1	0.42415	2	2
	Insect	0.0000	0	0.5426	1	0.46136	2	2
	AnosFuer	11.1394	4	11.1405	5	0.97364	6	2

Se observa que los niveles de reducción de la devianza, grados de libertad y valor de significancia para las variables TWE y AnosResi son muy similares. Esto se debe a que parte de la información de la variable TWE proviene de AnosResi; TWE=(AnosResi-AnosFuer)*393. Por lo tanto, serán consideradas por separado.

Residuales

En al apéndice I se muestran las gráficas de residuales correspondientes al ajuste de la variable AnosResi. En las gráficas 4A, 4B y 5 se distigue la observación 35, y cabe preguntarse si corresponde a un valor atípico. Para responder a esta pregunta se muestra a continuación una tabla donde se indica el número de cada observación binomial (obs), el valor de AnosResi, el número de personas con lesiones (Y) y totales (N), y la razón R=Y/N

IDENTIFICACION DE OBSERVACIONES
Modelo: y=Lesion x=AnosResi

Obs	ANOSRESI	Y	N	R
1	2	0	2	0
2	3	0	2	0
3	4	0	3	0
4	5	0	7	0
5	6	0	9	0
6	7	0	4	0
7	8	0	6	0
8	9	0	3	0
9	10	1	12	0.08
10	11	0	5	0
11	12	1	11	0.09
12	13	2	6	0.33
13	14	1	6	0.17
14	15	2	6	0.33
15	16	0	4	0
16	17	1	3	0.33
17	18	0	1	0
18	19	0	1	0
19	20	1	2	0.5
20	21	0	1	0
21	22	3	7	0.43
22	23	0	1	0
23	24	0	2	0
24	25	0	3	0
25	26	0	1	0

26	27	2	3	0.67
27	28	0	1	0
28	29	0	2	0
29	30	3	5	0.6
30	32	1	3	0.33
31	33	2	3	0.67
32	34	3	4	0.75
33	35	1	1	1
34	37	2	2	1
35	38	0	3	0
36	39	0	1	0
37	40	1	2	0.5
38	42	0	1	0
39	43	3	3	1
40	44	4	4	1
41	46	0	1	0
42	47	1	1	1
43	49	1	1	1
44	60	2	2	1

En la observación 35 existen tres personas con 38 años de residir en el poblado y que no presentan lesiones cuando la mayoría de las personas sí las padecen. No obstante existen otras observaciones (36, 38 y 41) que están en la misma situación pero que no sobresalieron tanto porque solo existe una observación binomial.

Estas observaciones no corresponden a errores en el levantamiento de la información, sino a variaciones no explicadas por la variable AnosResi. Por tal motivo se concluye que ninguna de estas observaciones es lo suficientemente sobresaliente como para que altere significativamente el valor de los parámetros estimados.

Por otra parte, la gráfica de residuales parciales (gráfica 6 del apéndice 1) mostró una tendencia lineal, por lo que no es necesario aplicar transformación alguna.

Para el resto de variables no es posible hacer análisis de residuales por el número tan pequeño de observaciones binomiales que se generan.

A continuación se muestran las tablas de parámetros estimados que resultaron para las variables significativas

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-3.5891586	0.527938	46.21886	0.0001	Intercept
TWE	1	0.00028526	0.000049	33.32469	0.0001	

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-3.568175	0.525096	46.17592	0.0001	Intercept
ANOSRESI	1	0.10962941	0.019121	32.87355	0.0001	

Como puede observarse la variable Fertiliz, que por sí sola había resultado significativa, no quedó seleccionada. Esto se debe a que la varianza que era explicada por Fertiliz se explica mejor con las variables TWE y AnosResi.

3.3.2.3 Grupo 3 Tabaquismo y Alcoholismo

Variables:

Tabaco Alcohol

A continuación se muestra la tabla de reducción de las devianzas

Análisis de la devianza para regresión logística
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	Alcohol	2.0645E-14	0	9.88981	1	0.001662	2	2
	Tabaco	2.1316E-14	0	4.63856	1	0.031261	2	2

Después de realizar la selección de variables, resultó insuficiente la significancia de la variable Tabaco, por lo que sólo será considerada la variable Alcohol, resultando los siguientes parámetros estimados para este grupo:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.4351	0.2225	41.5854	0.0001	.	.
ALCOHOL	1	1.4351	0.4510	10.1267	0.0015	0.295814	4.200

3.3.2.4 Grupo 4 Dieta Alimenticia:

Variables:

PescMar Higado Huevos CarneRes CarnePue CarnesFr Frijoles
Chicharo Cereal Pan Zanahori VerdLeg Melon Papa
Calabaza Vitamina Medicina Sulfas Anticonc Anticonv OtrosMed

A continuación se muestra la tabla resultante de reducción de las devianzas:

Análisis de la devianza para regresión logística
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	OtrosMed	0.00000	0	9.84739	1	0.00170	2	2
	Higado	0.00000	0	6.29638	1	0.01210	2	2
	Medicina	0.00000	0	5.58108	1	0.01816	2	2
	Vitamina	0.00000	0	3.89255	1	0.04850	2	2
	Pan	2.37587	1	5.87734	2	0.06131	3	2
	VerdLeg	2.82099	1	3.99501	2	0.27858	3	2
	CarnesFr	0.13956	1	1.10022	2	0.32702	3	2
	PescMar	0.00000	0	0.95452	1	0.32857	2	2
	CarneRes	0.24327	1	0.98840	2	0.38802	3	2
	Anticonv	0.00000	0	0.58685	1	0.44364	2	2
	Papa	0.46448	1	0.97561	2	0.47465	3	2
	Huevos	1.01783	1	1.43782	2	0.51694	3	2
	Melon	0.04999	1	0.31923	2	0.60384	3	2
	Calabaza	0.22839	1	0.42904	2	0.65419	3	2
	Frijoles	1.41735	1	1.54123	2	0.72486	3	2
	Cereal	0.00069	1	0.11765	2	0.73236	3	2
	Anticonc	0.00000	0	0.07712	1	0.78124	2	2
	Chicharo	0.00000	0	0.03567	1	0.85021	2	2
	CarnePue	0.12818	1	0.13038	2	0.96255	3	2
	Zanahori	0.14965	1	0.15032	2	0.97943	3	2
	Sulfas	0.00000	-1	0.00000	0	1.00000	1	2

El resultado de juntar y eliminar una por una a las variables que resultaron significativas en la tabla anterior y que ya no lo fueron en el modelo conjunto se muestra a continuación:

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-1.5321665	0.247594	38.29406	0.0001	Intercept
OTROSMED	1	1.46545376	0.471365	9.665619	0.0027	
HIGADO	1	1.38518884	0.619143	5.005373	0.0252	

Debido a que en este modelo sólo se cuenta con cuatro observaciones binomiales, no es posible hacer análisis de residuales para este modelo.

3.3.2.5 Grupo 5 Cuadro Clínico

Variables: Gripa Sarampio Rubeola Anginas Paperas Pulmonia Bronquit Diarrea AntF AntP

Como puede observarse en la tabla de significancia por reducción de la devianza sólo fueron importantes las variables AntP y Bronquit

Análisis de la devianza para regresión logística
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	AntP	.00000000000000	0	9.40655	1	0.00216	2	2
	Bronquit	.00000000000001	0	4.39663	1	0.03601	2	2
	AntF	.00000000000001	0	2.43859	1	0.11838	2	2
	Gripa	.00000000000001	0	1.38973	1	0.23845	2	2
	Sarampio	.00000000033510	0	1.21750	1	0.26985	2	2
	Pulmonia	.00000000016757	0	0.60623	1	0.43621	2	2
	Diarrea	.00000000001277	0	0.15510	1	0.69371	2	2
	Anginas	.00000000001534	0	0.04387	1	0.83410	2	2
	Rubeola	.00000000001132	-1	0.00000	0	1.00000	1	2
	Paperas	.00000000001132	-1	0.00000	0	1.00000	1	2

Al estimar el modelo conjunto de variables significativas la variable Bronquit perdió significancia, por lo que el modelo seleccionado solo incluye a la variable AntP con los siguientes parámetros estimados:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.5305	0.2048	55.8479	0.0001	.	.
ANTP	1	1.6483	0.5273	9.7713	0.0018	0.266508	5.198

3.3.2.6 Grupo 6 Especies de arsénico en la orina:

3.3.2.6.1 Creat ASIG MMAG DMAG ASOG SUMESG (Ajustadas por creatinina)

A continuación se muestra la tabla de reducción de la devianza por variable

Analisis de la devianza para regresión logistica
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	ASIG	175.838	176	182.780	177	0.00841	178	2
	SUMESG	181.305	178	185.553	179	0.03929	180	2
	DMAG	182.432	178	185.553	179	0.07726	180	2
	ASOG	182.702	177	185.553	178	0.09134	179	2
	MMAG	184.683	178	185.553	179	0.41303	180	2
	Creat	175.314	155	175.509	156	0.65852	157	2

Sólo las dos primeras variables son significativas

En las gráficas de residuales no se encontró nada anormal, aunque para la variable ASIG existe una observación sobresaliente (ASIG = 380.82) lo cual

es debido a que esta persona no presentó lesiones, mientras que el resto con niveles altos y similares de arsénico inorgánico si las presentaron.

Eliminando esta observación del análisis se obtiene la siguiente significancia por reducción de la devianza:

Analisis de la devianza para regresión logística
Valor p para el modelo fijado a Lesion
Condición de selección: asig ne 380.82

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	ASIG	170.051	175	179.649	176	.0019487	177	2

La cual no difiere mucho del dato anterior. Por tal motivo y porque no existen razones para suponer que se trata de un error, la observación no será eliminada del análisis.

Debido a que el valor de Asig depende de SUMESG, estas variables serán consideradas por separado. A continuación se muestran las tablas correspondientes de parámetros estimados para cada caso:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.7985	0.2640	9.1450	0.0025	.	.
ASIG	1	-0.00483	0.00206	5.5177	0.0188	-0.36244	0.995

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.7395	0.3488	4.4925	0.0340	.	.
SUMESG	1	-0.00085	0.000474	3.1935	0.0739	-0.275341	0.999

En la tabla correspondiente a la variable SUMESG se observa que con el criterio de la estadística de Wald, la variable deja de ser significativa.
Debido a que ésta no es una variable de interés y a que de cualquier forma no es muy buena su significancia por reducción de la devianza, ya no será considerada en el análisis, quedando únicamente seleccionada la variable ASIG en este grupo.

3.3.2.6.2 ASIP MMAP DMAP ASOP (Porcentaje de especies)

A continuación se muestra la tabla de significancias por reducción de la devianza:

Analisis de la devianza para regresión logistica
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	MMAP	174.719	168	180.008	169	0.02147	170	2
	ASIP	176.128	176	180.008	177	0.04888	178	2
	ASOP	176.128	176	180.008	177	0.04888	178	2
	DMAP	180.003	175	180.008	176	0.94623	177	2

En las gráficas de residuales no se observaron anormalidades

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.9724	0.3538	31.0714	0.0001	.	.
MMAP	1	0.0425	0.0184	5.3183	0.0211	0.221982	1.043

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.8058	0.3104	6.7374	0.0094	.	.
ASIP	1	-0.0344	0.0182	3.5779	0.0586	-0.211819	0.966

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-4.2458	1.5798	7.2228	0.0072	.	.
ASOP	1	0.0344	0.0182	3.5779	0.0586	0.211819	1.035

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	1.4296	0.4130	11.9813	0.0005	.	.
MMAP	1	0.0478	0.0189	6.3572	0.0117	0.249467	1.049
ASIP	1	-0.0417	0.0185	4.5871	0.0322	-0.256827	0.959

Obsérvese que la significancia de las variables ASIP y ASOP son las mismas y lo único que cambia son es el signo del parámetro estimado. Esto se debe a que como ASOP=MMAP+DMAP y ASIP+MMAP+DMAP=1, resulta que ASOP=1-ASIP. La diferencia en el signo es importante, pues indica que a mayor proporción de arsénico inorgánico existe un mayor padecimiento de lesiones y viceversa.

Excepto por MMAP, apenas son significativas las demás variables. Se observa que la variable ASIP logra significancia al considerarse junto con MMAP.

Solo serán considerados los modelos con MMAP y MMAP ASIP

3.3.2.6.3 ASI MMA DMA SumEsp (Especies sin ajustar)

A continuación se muestra la tabla de reducción de la devianza:

Analisis de la devianza para regresión logistica
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	ASI	181.671	176	185.553	177	0.04881	178	2
	SumEsp	183.435	178	185.553	179	0.14554	180	2
	DMA	183.611	178	185.553	179	0.16345	180	2
	MMA	185.530	178	185.553	179	0.87835	180	2

La única variable que resultó significativa fue ASI.

A continuación se muestra la tabla de parámetros ajustados para ASI

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.9083	0.2706	11.2672	0.0008	.	.
ASI	1	-3E-6	1.615E-6	3.4615	0.0628	-0.218875	1.000

Aunque con la estadística de Wald dejó de ser significativa, será considerada la variable ASI debido a su importancia en la investigación y a que se encuentra en el límite de la región de rechazo.

3.3.2.6.4 MMAASI DMAASI ASOASI DMAMMA (Razones de especies)

Como puede observarse en la tabla de reducción de significancias, ninguna de éstas variables es significativa:

Analisis de la devianza para regresión logistica
Valor p para el modelo fijado a Lesion

XFIJAS	X	DEVMODEL	DFMODEL	DEVFIJO	DFFIJO	P	NOBS	NVAR
	MMAASI	146.216	127	147.416	128	0.27334	129	2
	DMAMMA	171.154	161	171.690	162	0.46415	163	2
	ASOASI	176.132	168	176.189	169	0.81175	170	2
	DMAASI	185.546	170	185.553	171	0.93346	172	2

Por lo tanto no serán seleccionadas

3.3.3 Unión de grupos

3.3.3.1 Grupo 1 y 2

A continuación se muestran las tablas de parámetros estimados para la unión de los grupos:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-4.9916	0.7977	39.1539	0.0001	.	.
EDAD	1	0.0969	0.0236	16.8524	0.0001	0.960766	1.102
TWE	1	0.000096	0.000066	2.1095	0.1464	0.268762	1.000

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	4.9970	0.7981	39.2032	0.0001	.	.
EDAD	1	0.0982	0.0236	17.3706	0.0001	0.972737	1.103
ANOSRESI	1	0.0357	0.0257	1.9297	0.1648	0.257745	1.036

Debido a que la variable Edad explica mejor la varianza, queda ésta como única variable seleccionada en estos dos grupos.

3.3.3.2 Grupos 1, 2 y 3

Al añadir la variable alcohol, ésta dejó de ser significativa, por lo que no será seleccionada:

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-4.7774	0.7380	41.9092	0.0001	.	.
EDAD	1	0.1157	0.0203	32.4293	0.0001	1.142601	1.123
ALCOHOL	1	0.3625	0.5580	0.4221	0.5159	0.074729	1.437

3.3.3.3 Grupos 1, 2, 3 y 4

En este grupo dejó de ser significativa la variable OtrosMed, esto no ocurrió con la variable Hígado, por lo que será seleccionada:

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.4593258	0.656309	46.16596	0.0001	Intercept
EDAD	1	0.09356759	0.016487	32.20967	0.0001	Edad
HIGADO	1	1.9893222	0.729462	7.437132	0.0064	

3.3.3.4 Grupos 1, 2, 3, 4 y 5

La variable AntP dejó de ser significativa por lo que será eliminada:

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.4505719	0.65776	45.78225	0.0001	Intercept
EDAD	1	0.09279193	0.016919	30.08116	0.0001	Edad
HIGADO	1	1.97903615	0.731259	7.324276	0.0068	
ANTP	1	0.13890141	0.690193	0.040502	0.8405	

3.3.3.5 Especies de arsénico

3.3.3.5.1 Variables ASI y ASIG

Las variables ASI y ASIG perdieron significancia. Dada la importancia de éstas, se consideró el ajustarlas únicamente con Edad, resultando igualmente nula su significancia:

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.5090532	0.768102	34.46136	0.0001	Intercept
HIGADO	1	1.99035304	0.729424	7.445612	0.0064	
EDAD	1	0.09403054	0.016913	30.90895	0.0001	Edad
ASI	1	2.3421E-7	1.852E-6	0.015989	0.8994	

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.577162	0.808453	32.05408	0.0001	Intercept
EDAD	1	0.09493627	0.017424	29.68763	0.0001	Edad
HIGADO	1	2.02047777	0.741577	7.423272	0.0064	
ASIG	1	0.00060594	0.002347	0.066673	0.7962	Arsénico Inorgánico

3.3.3.5.2 Variables ASIP y MMAP

La variable ASIP que había logrado significancia al unirla con MMAP, dejó de serlo, sin embargo MMAP logró mantenerse como variable significativa:

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-5.5997101	1.004763	31.06018	0.0001	Intercept
EDAD	1	0.0969671	0.017659	30.15347	0.0001	Edad
HIGADO	1	1.99351849	0.739568	7.265806	0.0070	
MMAP	1	0.05927391	0.024566	5.821602	0.0158	MMAP
ASIP	1	0.00341312	0.022493	0.023025	0.8784	ASIP

Debido a lo anterior, queda como modelo resultante el compuesto por las variables EDAD, Higado y MMAP con los parámetros estimados que a continuación se muestran:

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-5.5199153	0.848686	42.30293	0.0001	Intercept
EDAD	1	0.09618852	0.016822	32.69454	0.0001	Edad
HIGADO	1	1.99211115	0.738884	7.268995	0.0070	
MMAP	1	0.05937986	0.024607	5.823421	0.0158	MMAP

3.3.4 Cuadro resumen de variables seleccionadas

A continuación se muestra un cuadro resumen de las variables seleccionadas y sus significancias, al cual se añadieron los resultados de las selecciones automáticas con los métodos *stepwise* y *backward*. Como pude observarse, los resultados finales son muy similares, aunque con el método Backward no fue seleccionada la variable MMAP.

		Tipo de seleccion					
		Manual		StepWise		Backward	
Grupo	Descripción	Variable	p > chi	Variable	p > chi	Variable	p > chi
Grupo 1	Caracteristicas fisicas	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
Grupo 2	Fuentes de exposición	TWE	0.0001	TWE	0.0001	ANOSRESI	0.0001
		ANOSRESI	0.0001				
Grupo 3	Tabaquismo y alcoholismo	ALCOHOL	0.0015	ALCOHOL	0.0016	ALCOHOL	0.0016
Grupo 4	Dieta alimenticia	HIGADO	0.0253	HIGADO	0.0105	HIGADO	0.0105
		OTROSMED	0.0019				
Grupo 5	Cuadro clinico	ANTP	0.0018	ANTP	0.0018	ANTP	0.0018
Grupo 6	Especies ajustadas	ASIG	0.0188	ASIG	0.0188	NINGUNA	
Grupo 7	Porcentaje de especies	ASIP	0.0322	ASIP	0.0322	ASIP	0.0024
		MMAP	0.0117	MMAP	0.0117	DMAP	0.0117
		MMAP	0.0211				
Grupo 8	Especies sin ajustar	NINGUNA		NINGUNA		NINGUNA	
Grupo 9	Razones de especies	NINGUNA		NINGUNA		NINGUNA	

Unión de grupos	Tipo de selección					
	Manual		StepWise		Backward	
	Variable	p > chi	Variable	p > chi	Variable	p > chi
Grupo 1, 2	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
Grupo 1, 2, 3	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
Grupo 1, 2, 3, 4	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
	HIGADO	0.0064	HIGADO	0.0064	HIGADO	0.0064
Grupo 1, 2, 3, 4, 5	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
	HIGADO	0.0064	HIGADO	0.0064	HIGADO	0.0064
Grupo 1, 2, 3, 4, 5, 6	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
	HIGADO	0.0064	HIGADO	0.0064	HIGADO	0.0064
Grupo 1, 2, 3,4, 5, 6, 7	EDAD	0.0001	EDAD	0.0001	EDAD	0.0001
	HIGADO	0.0070	HIGADO	0.0070	HIGADO	0.0064
	MMAP	0.0158	MMAP	0.0158		

3.3.5 Parámetros estimados para el modelo seleccionado

A continuación se muestra la salida de SAS con los parámetros estimados para el modelo resultante:

AJUSTE DE: Lesion = MMAP Edad Higado

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	5.5199153	0.848686	42.30293	0.0001	Intercept
MMAP	1	0.05937986	0.024607	5.823421	0.0158	MMAP
EDAD	1	0.09618852	0.016822	32.69454	0.0001	Edad
HIGADO	1	1.99214115	0.738884	7.268995	0.0070	

Debido al objetivo del estudio, también se muestran los parámetros estimados para las especies de arsénico que fueron significativas por sí solas, sin agregar variables adicionales:

AJUSTE DE: Lesion = ASI

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-0.9082686	0.270587	11.26719	0.0008	Intercept
ASI	1	-3.0039E-6	1.615E-6	3.461442	0.0628	

AJUSTE DE: Lesion = ASIG

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-0.7985	0.2640	9.1450	0.0025
ASIG	1	-0.00483	0.00206	5.5177	0.0188

AJUSTE DE: Lesion = MMAP

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-1.9724	0.3538	31.0714	0.0001
MMAP	1	0.0425	0.0184	5.3183	0.0211

AJUSTE DE: Lesion = ASIP

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-0.8058	0.3104	6.7374	0.0094
ASIP	1	-0.0344	0.0182	3.5779	0.0586

3.3.5.1 Resumen de los modelos seleccionados

A continuación se muestran los modelos que finalmente fueron seleccionados

(1) *Lesion = MMAP Edad Higado*

$$Logit(p) = -5.52 + 0.059MMAP + 0.096Edad + 1.992Higado$$

(2) *Lesion = ASI*

$$Logit(p) = -0.9083 - (3.0039E-6)ASI$$

(3) *Lesion = ASIG*

$$Logit(p) = -0.7985 - 0.0048ASIG$$

(4) *Lesion=MMAP*

$$Logit(p) = -1.9724 + 0.0425MMAP$$

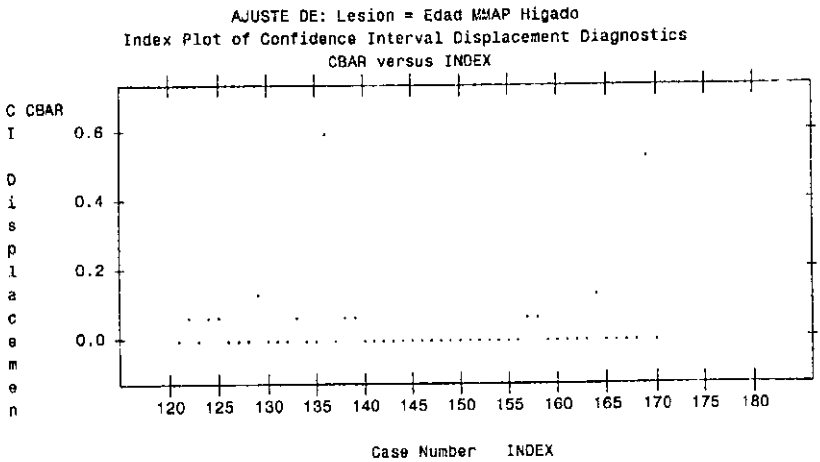
(5) *Lesion = ASIP*

$$Logit(p) = -0.8058 - 0.0344ASIP$$

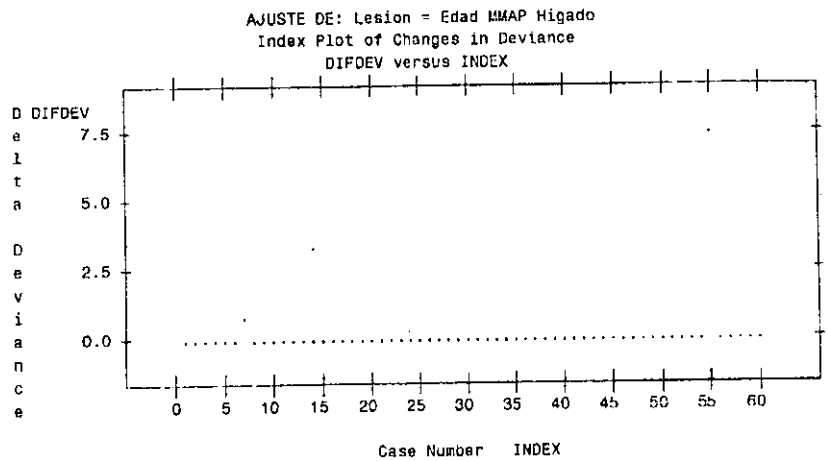
3.3.6 Graficas de residuales para el modelo ajustado a *Lesion = Edad Higado MMAP*

En el apéndice II se muestran las gráficas de residuales correspondientes a este modelo. En general no se observan atipicidades, solo algunos puntos sobresalientes en las gráficas C_i, $\bar{C}_i$ y **DifDev**. Las dos últimas se muestran a continuación:

Gráfica $\bar{C}$



Gráfica DifDev



Como puede observarse en la gráfica $\bar{C}_1$, sobresalen las observaciones 136 y 169, mientras que en DifDev la 55.

La observación 136 corresponde a una persona de 42 años con el más alto nivel de MMAP y que no presenta lesiones, cuando la mayoría de la personas con estos niveles sí las presenta

La 169 corresponde a una persona de 80 años que no presenta lesiones siendo una de las de mayor edad. Llama la atención que ésta persona presenta niveles muy bajos de MMAP.

La observación 55 es el caso contrario a las anteriores porque se trata de la persona con menor edad (12 años) que presenta lesiones a su temprana edad.

Ninguno de estos casos corresponde a una situación irreal, por lo que no se trata de error alguno, sino de personas que presentan un comportamiento diferente al de la mayoría. Por esta razón no se considerará la posibilidad de excluirlos.

3.4 Cálculo de razones de momios e intervalos de confianza

Las razones de momios fueron calculadas conforme a lo expuesto en el capítulo dos, donde la razón de momios para el incremento de la variable *x* está dada por: $\hat{\psi}_{\Delta x_i} = e^{\Delta x_i \hat{\beta}_i}$, y el intervalo de confianza (α=0.05) para los

valores mínimo y máximo respectivamente por:

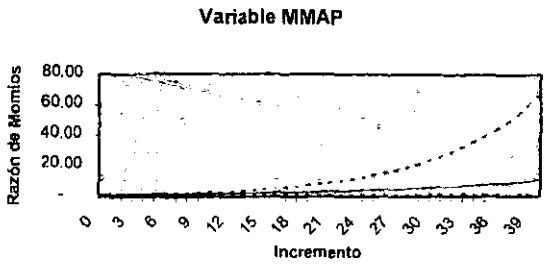
$$e^{\log(\hat{\psi}_{\Delta x_i}) \pm z_{\alpha/2} s.e.\{\Delta x_i \hat{\beta}_i\}}$$

o bien por:

$$\left(\hat{\psi}_{\Delta x_i} e^{z_{\alpha/2} s.e.(\Delta x_i \hat{\beta}_i)}, \frac{\hat{\psi}_{\Delta x_i}}{e^{z_{\alpha/2} s.e.(\Delta x_i \hat{\beta}_i)}} \right)$$

Antes de proceder al cálculo de las razones de momios, es necesario establecer el valor del incremento en x (Δx) que será utilizado para calcular las razones de momios correspondientes a las *variables numéricas*, puesto que dependiendo de la magnitud de las unidades de cada variable, un incremento de una sola unidad sería despreciable y como consecuencia las razones de momios se interpretarían como una no-relación.

Como ejemplo de lo anterior se muestra una gráfica en la que se aprecian los cambios en la razón de momios y sus respectivos intervalos de confianza, para diferentes valores de incremento en MMAP y considerando al resto de variables fijas. Las líneas punteadas corresponden a los intervalos y la continua a la razón de momios.



Por lo tanto se tomarán incrementos de un décimo de la diferencia entre el valor máximo y el valor mínimo de cada variable continua como se expresa en la siguiente fórmula: Δx=(Valor Máximo - Valor mínimo)/10.

Para las variables categóricas el incremento será igual a uno, puesto que en estos casos el modelo estima su presencia o ausencia.

A continuación se muestran una tabla resumen para cada uno de los 4 modelos considerados, donde se muestran las razones de momios con sus intervalos de confianza y los datos con los que se estimaron:

Modelo / Variable	*Parámetro Estimado*	*Error Estándar*	*Z (Alpha=0 .05)*	*Max*	*Min*	*Incre-mento*	*Razón de momios*	*Intervalo de Confianza (95%) Inf*	*Sup*
MMAP	0.0594	0.0246	1.96	50.81	1.03	5.0	1.34	1.06	1.71
EDAD	0.0962	0.0168	1.96	81	3.	7.8	2.12	1.64	2.74
HIGADO	1.9921	0.7389	1.96	-	-	1	7.33	1.72	31.20
ASI	-0.00395E-6	1.615E-6	1.96	662004.	2000	66000.4	1.00	0.81	1.23
ASIG	-0.0048	0.0021	1.96	652.54	1.74	65.1	0.73	0.56	0.95
MMAP	0.0425	0.0184	1.96	50.81	1.03	5.0	1.24	1.03	1.48
ASIP	-0.0344	0.1820	1.96	48.54	0.63	4.8	0.86	0.15	4.68

CONCLUSIONES

A continuación se muestra la tabla con los resultados que se presentaron en el capítulo 3:

# Mod	Modelo / variables indepen.	Parámetro Estimado	Error Estándar	Valor p	Z (Alpha= 0.05)	Razón de momios	Intervalo de Confianza (95%)	
							Inf	Sup
1	MMAP	0.0594	0.0246	0.0158	1.96	1.34	1.06	1.71
	EDAD	0.0962	0.0168	0.0001	1.96	2.12	1.64	2.74
	HIGADO	1.9921	0.7389	0.0070	1.96	7.33	1.72	31.20
2	ASI	-0.0039E-6	1.615E-6	0.0628	1.96	1.00	0.91	1.23
3	ASIG	-0.0048	0.0021	0.0188	1.96	0.73	0.56	0.95
4	IMMAP	0.0425	0.0184	0.0211	1.96	1.24	1.03	1.48
5	ASIP	-0.0344	0.1820	0.0586	1.96	0.85	0.15	4.68

Primer modelo:

El primer modelo que se muestra (compuesto por las variables independientes MMAP, EDAD e HIGADO) es el que resultó del análisis de selección y unión de variables, lo que significa que éstas no dejaron de ser significativas a pesar de considerarlas junto con una variable tan importante como es EDAD ($p<0.0001$). Nótese que todos los parámetros estimados son mayores que cero (no negativos), lo que significa que a medida que se incrementa el valor de cualquiera de estas variables crece la proporción de personas con presencia de lesiones en la piel. Lo anterior se observa en los valores que toman las razones de momios, pues todas ellas son mayores a uno, incluso en los límites inferiores de confianza.

Es lógico que haya resultado seleccionada la variable EDAD por su alto nivel de significancia. Lo interesante de ésta es que cuando se llevó a cabo la selección de variables resultó que EDAD fue más importante que la variable AnosResi (tiempo de residir en el poblado), lo que sugiere que además del tiempo de exposición a arsénico la edad es una variable relacionada con el riesgo de aparición de lesiones cutáneas.

En lo que corresponde a la variable HIGADO, llama la atención el nivel de significancia (p=0.007). Al parecer el consumo de hígado incrementa el riesgo de padecer lesiones en 7.33 veces más incidencias para las personas que lo consumen. No obstante lo anterior, es importante mencionar que sólo 14 personas de las 180 personas entrevistadas (14/180=0.78) indicaron que consumen este alimento, siendo un número muy pequeño como para hacer alguna inferencia sobre esta variable:

LESION	HIGADO	No. de personas
0	0	135
	1	7
1	0	31
	1	7

Es posible que la importancia de esta variable se deba a que la proporción de personas que consumen hígado y no presentan lesiones cutáneas (7/135=0.05) es mucho menor que la de quienes consumen hígado y presentan alteraciones en la piel (7/31=0.23). Es pues necesario contar con

información más robusta con respecto a esta variable que permita elaborar alguna conclusión.

Respecto a la variable MMAP, llama la atención que el signo del parámetro estimado sea positivo, pues esto significa que a mayor cantidad de MMAP se incrementa el riesgo de que una persona presente alteraciones en la piel. Este dato contradice los supuestos iniciales donde se indica que las especies de arsénico metiladas (MMA y DMA) son menos tóxicas que el arsénico inorgánico.

En investigaciones más recientes se han reportado estudios sobre la toxicidad de las especies de arsénico considerando su valencia, y se ha encontrado que las especies metiladas trivalentes (MMAIII y DMAIII) son significativamente mas tóxicas que el arsénico inorgánico[1]. Por otra parte se ha reportado que las especies metiladas trivalentes son formadas en el metabolismo de arsénico en humanos[2].

Desafortunadamente en este trabajo no se realizó la diferenciación entre los estados de oxidación trivalente y pentavalente de las formas metiladas de arsénico, debido a que tradicionalmente se había asumido que la especies metiladas trivalentes eran intermediarios inestables con poca importancia en el metabolismo de arsénico. De haberse considerado esta diferenciación, hubiera sido posible evaluar el papel de la especie monometilada trivalente (MMAIII) en la presencia de lesiones en la piel provocadas por la exposición a arsénico.

[1] Lin y col. (1999), Styblo y col. (2000), Petrick y col (2000).

Estos hallazgos recientes sugieren que es posible que el aumento de la variable MMAP, que en este trabajo fue evaluada como la presencia de MMA total (trivalente y pentavalente), sea un reflejo de un aumento de la especie trivalente de MMA.

Segundo y tercer modelo:

Nuevamente llama la atención que el signo del correspondiente parámetro estimado para las variables (ASI y ASIG) sea negativo y por consiguiente la razón de momios menor a la unidad, pues esto indica que cuando es mayor la cantidad de ASI y ASIG es menor el riesgo de padecer lesiones y viceversa, lo cual en caso de considerar a esta variable contradice la hipótesis de que el riesgo es mayor cuando existe una mayor cantidad de arsénico sin metilar.

Una posible explicación de lo anterior, considerando la reciente información que muestra que las especies metiladas trivalentes pueden estar relacionadas con los efectos tóxicos que produce la exposición a arsénico, consiste en que al existir mayor cantidad de arsénico inorgánico disminuyen las especies metiladas y por lo tanto el MMA. En otras palabras, que esta relación en realidad esté reportando indirectamente un efecto del incremento en MMA.

Se observa que la significancia de la variable ASIG fue de $p = 0.0188$ sin ajustar variables adicionales. No obstante perdió su significancia al juntarse

[2] Del Razo y col. (2000) y Le y col. (2000)

con Edad e Higado pues su valor *p* resultante fue de 0.7962. Como dato adicional se menciona que al juntarse con únicamente la variable Edad su valor *p* fue de 0.8711, lo que parece indicar que la edad explica de mejor manera la varianza explicada por ASIG. Algo similar ocurrió con la variable ASI, con la diferencia de que su significancia es mucho menor, incluso el parámetro estimado para esta variable es muy cercano al cero y la razón de momios es prácticamente uno (0.99974).

Tercer modelo:

En este modelo se ajustó únicamente la proporción de arsénico monometilado MMAP. Se observa la misma relación directamente proporcional al padecimiento de lesiones que en el modelo uno. Las consideraciones arriba mencionados en relación a la variable MMAP también son válidos para este modelo.

Cuarto modelo:

Este modelo se muestra debido a que el nivel de significancia para la variable ASIP, aunque mayor a 0.05, fue muy cercano a la región de rechazo α=0.05 (p=0.0586). Los intervalos de confianza para la razón de momios del incremento de esta variable (0.15, 4.68) contienen a la unidad, por lo que se concluye que no hay suficiente evidencia para rechazar la hipótesis nula de que existe asociación entre la proporción de arsénico inorgánico (ASIP) y la presencia de lesiones.

Arsénico Dimetilado

Esta variable quedó excluida de cualquier análisis debido a la falta de significancia como se muestra en la siguiente tabla:

Variable	Valor p
DMAG	0.077
DMAP	0.946
DMA	0.163

Conclusiones generales

Cuando se inició el presente estudio, se pensaba que existía un mayor riesgo de padecer lesiones en la piel asociadas a la exposición de arsénico, en la medida en que disminuyera la capacidad de metilación del arsénico, bien fuera por una inhibición en el proceso de metilación producida por concentraciones altas de arsénico inorgánico o por una saturación del proceso de metilación, debido a la exposición crónica a este metaloide. La disminución en el proceso de metilación del arsénico provocaría un aumento (acumulación) en la cantidad del arsénico inorgánico.

Sin embargo, los resultados del presente trabajo muestran que el riesgo a presentar lesiones en la piel incrementa cuando la proporción de arsénico monometilado aumenta, y por otra parte, aunque con menos significancia, que la presencia de lesiones en la piel disminuye cuando aumenta la cantidad de arsénico inorgánico.

Recientemente se han publicado algunas investigaciones en las que se ha encontrado que las especies metiladas trivalentes, especialmente la forma monometilada de arsénico (MMAIII) presenta mayor capacidad de producir efectos adversos que los producidos por el arsénico inorgánico[3]. Este novedoso hallazgo podría explicar la relación directamente proporcional existente entre la presencia de lesiones en la piel y los parámetros estimados para MMAP.

Por otra parte, se tienen evidencias experimentales y epidemiológicas que muestran que en condiciones donde la exposición a arsénico es mayor, la proporción de la especie monometilada (MMA) se incrementa significativamente, mientras que la proporción de arsénico inorgánico se incrementa apenas unos cuantos puntos porcentuales, ambos incrementos se realizan en función del decremento en la proporción de la especie de arsénico dimetilada (DMA)[4].

Conclusiones personales

Este trabajo se inició con la intención de usar un ejemplo de una acción real, en este caso de un efecto biológico que pudiera ser estudiado satisfactoriamente mediante el empleo de las matemáticas aplicadas. Así fue como se logró el contacto con los investigadores de la sección de toxicología ambiental del CINVESTAV, quienes recibieron muy bien la propuesta de colaboración con un egresado de la carrera de matemáticas aplicadas y

[3] Styblo y col., (2000), Del Razo y col. (2000), Le y col. (2000)

computación para realizar un proyecto de tesis, que se concretó con el análisis de regresión logística aplicado a un efecto hipotetizado en una área biológica.

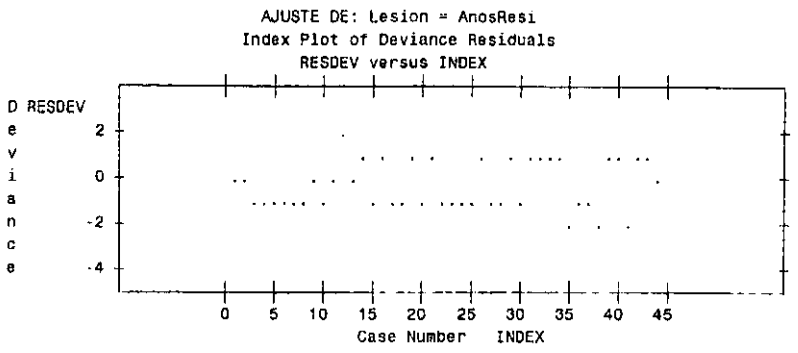
La experiencia fue muy positiva, ya que se logró apoyar satisfactoriamente a la sección de toxicología ambiental mediante el empleo de métodos estadísticos que son posibles gracias a los últimos desarrollos de la computación, que en este caso se llevaron a cabo mediante el programa SAS.

Es muy interesante y amplio el campo de desarrollo de las matemáticas aplicadas; tan sólo considerando la estadística aplicada a las ciencias biológicas, existe un gran número de modelos bioestadísticos que han sido desarrollados con objetivos específicos, lo cual no solo implica un amplio campo de desarrollo para la estadística, sino también una gran especialización de ésta en el ramo, pero que vale la pena apoyar por su contribución desde la elaboración de los diseños experimentales hasta la interpretación de la información recopilada.

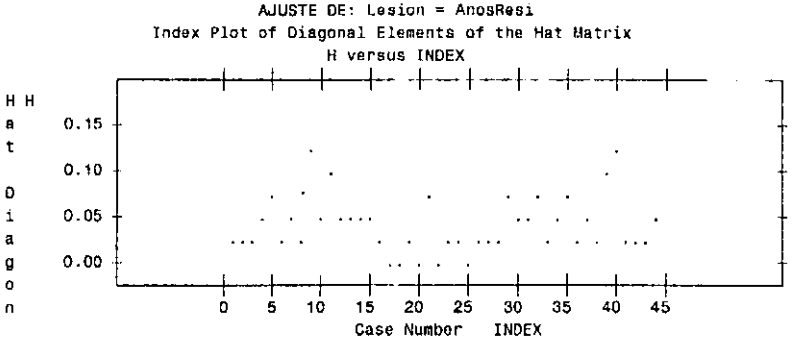
[4] Del Razo y col. 1997, Hughes y col, 2000

APENDICE I
EJEMPLO DE GRAFICAS DE RESIDUALES

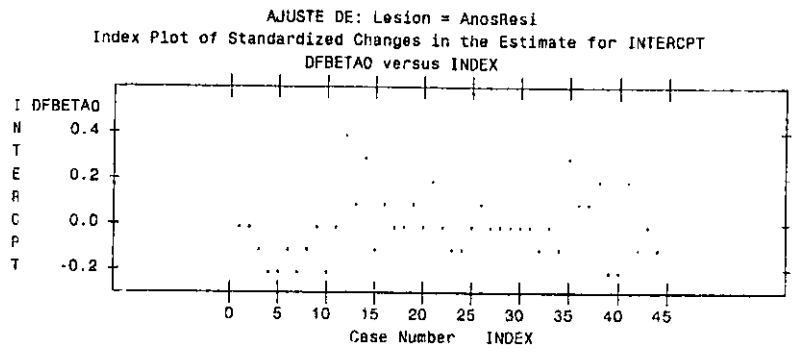
Grafica 1, Residuales Indice



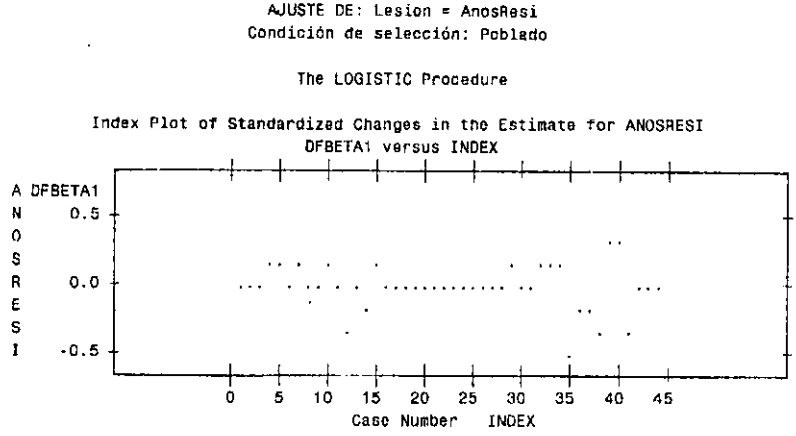
Gráfica 2, Residuales de apalancamiento



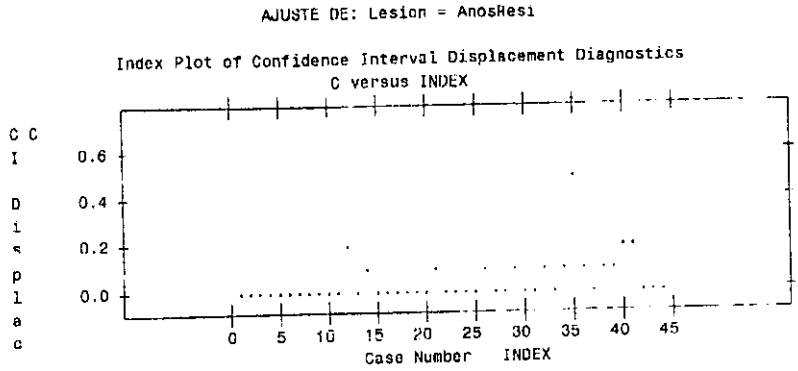
Gráfica 3A, Residuales Delta Beta para el intercepto



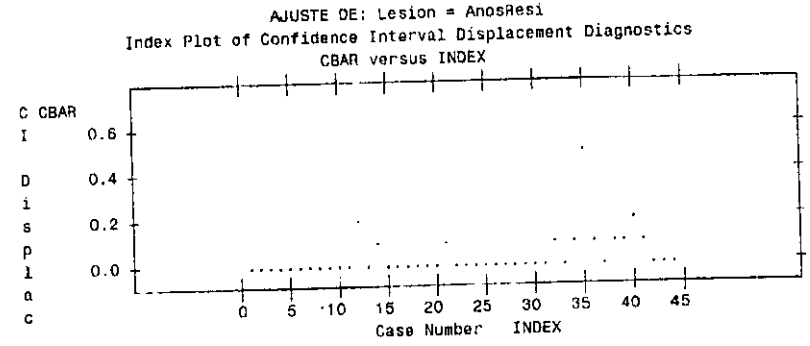
Gráfica 3B, Residuales Delta Beta para AnosResi



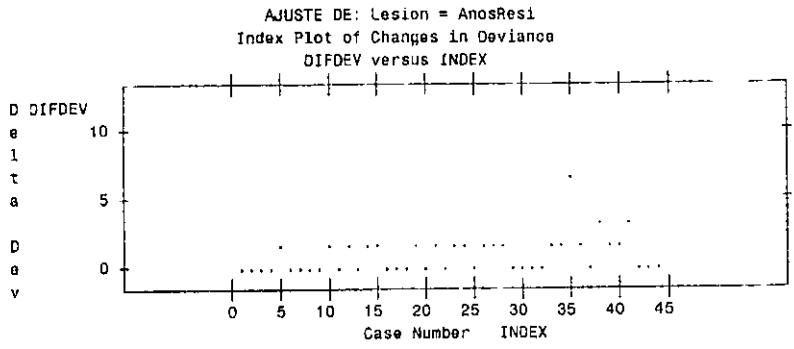
Gráfica 4A, Residuales C_i



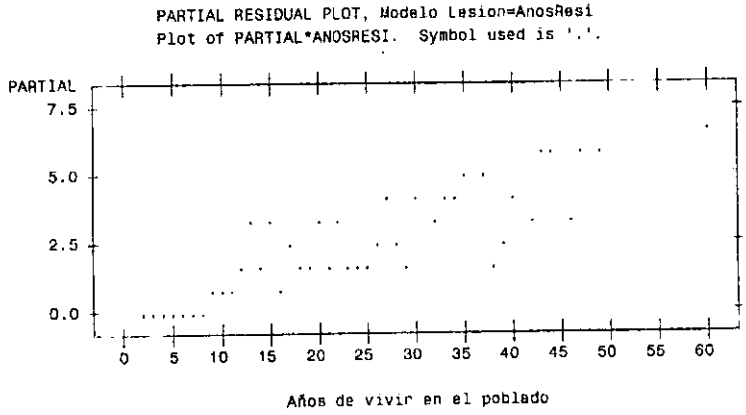
Gráfica 4B, Residuales $\bar{C}_i$



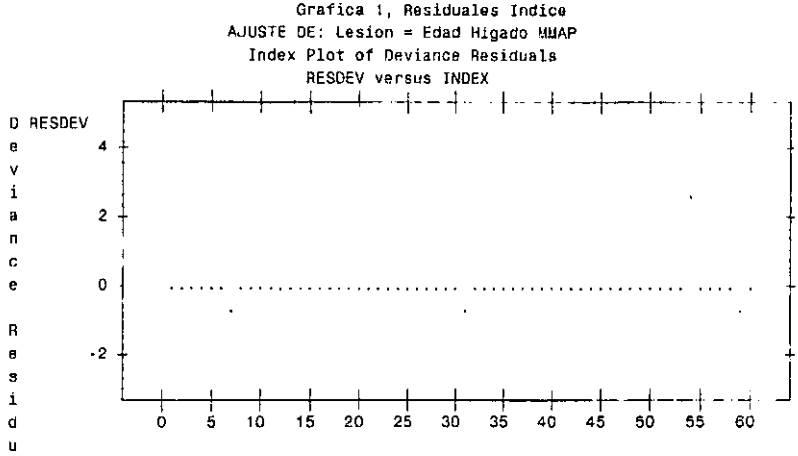
Gráfica 5, Residuales DifDev o Desviacion Delta

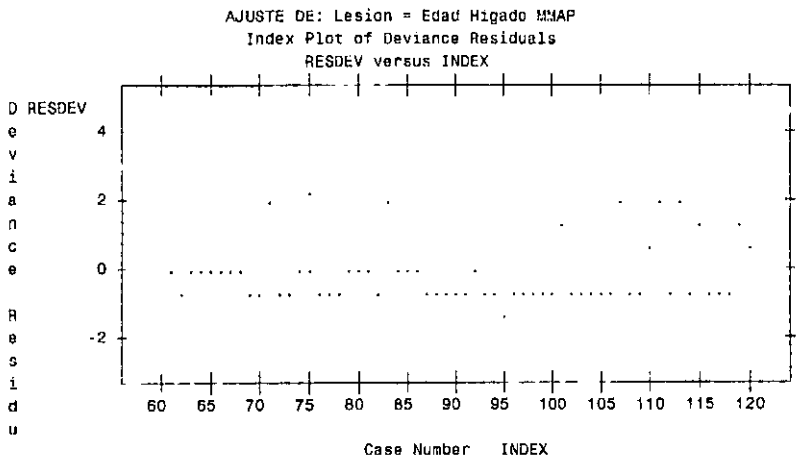


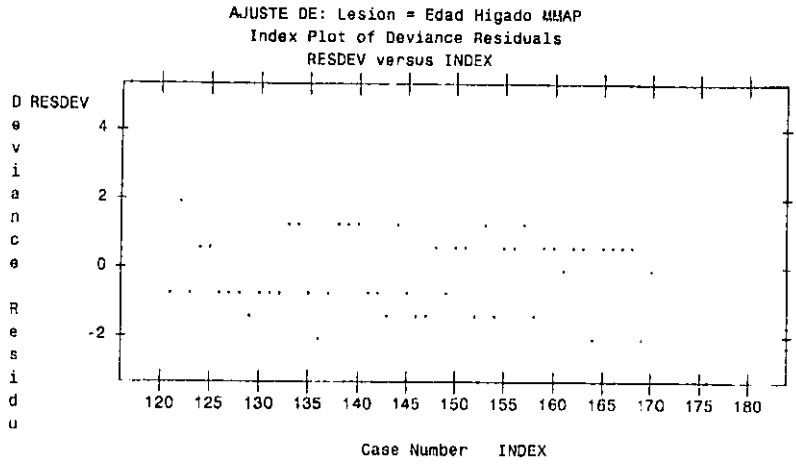
Gráfica 6, Residuales Parciales



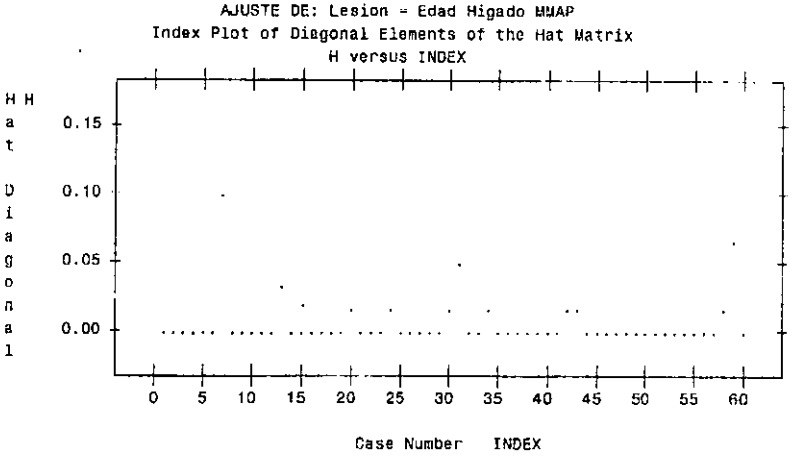
APENDICE II
GRAFICAS DE RESIDUALES DEL MODELO ESTIMADO



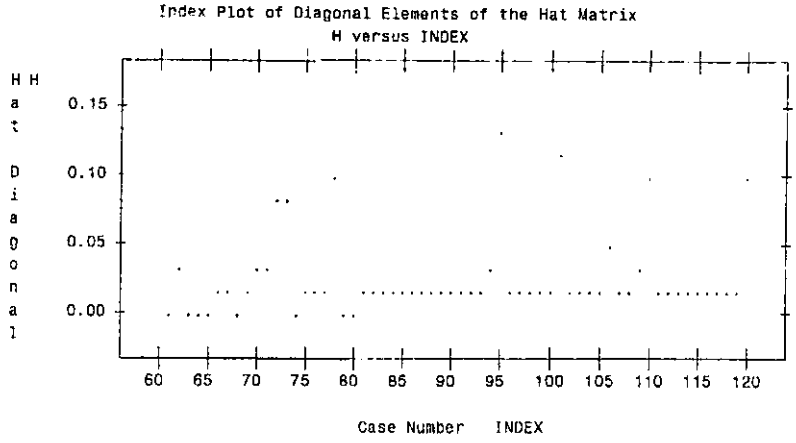




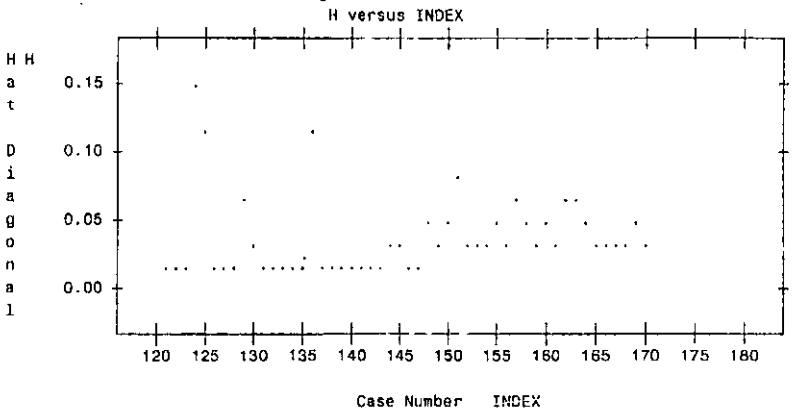
Grafica 2, Residuales de apalancamiento



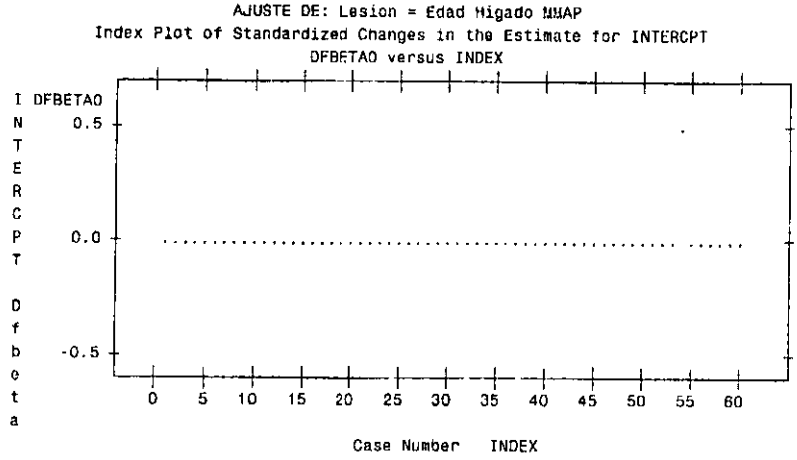
AJUSTE DE: Lesion = Edad Higado MMAP
Condición de selección: Poblado

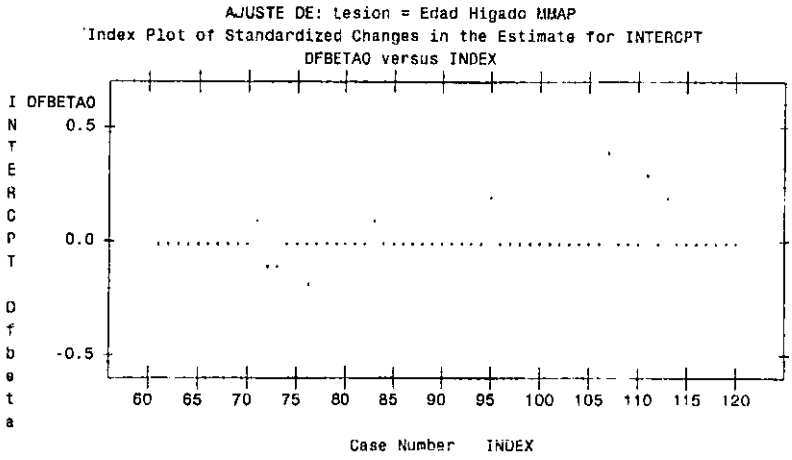


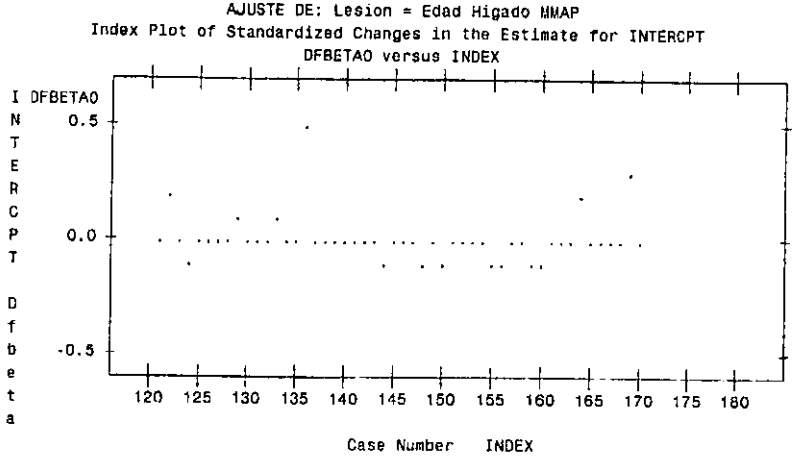
AJUSTE DE: Lesion = Edad Higado MMAP



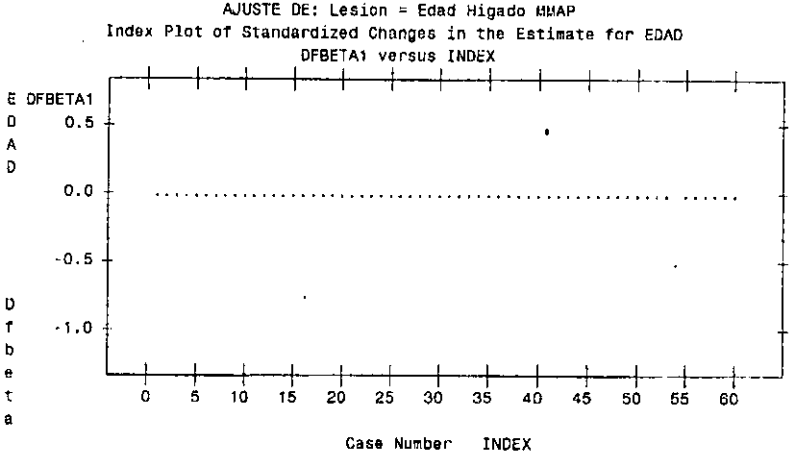
Grafica 3A, Delta Beta para el intercepto

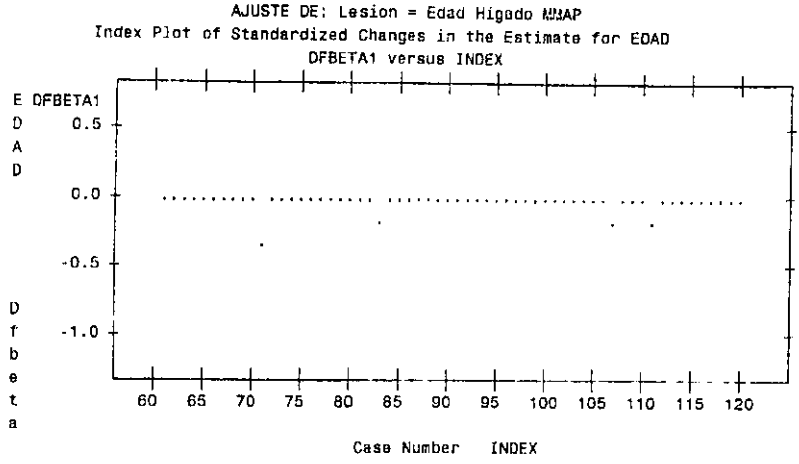


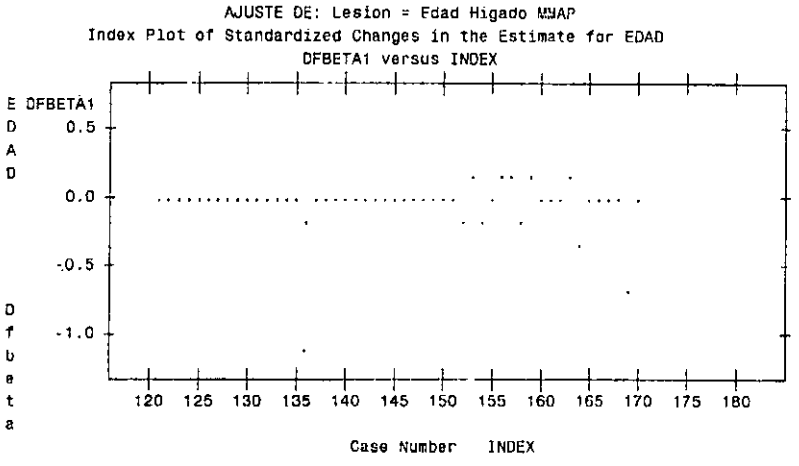




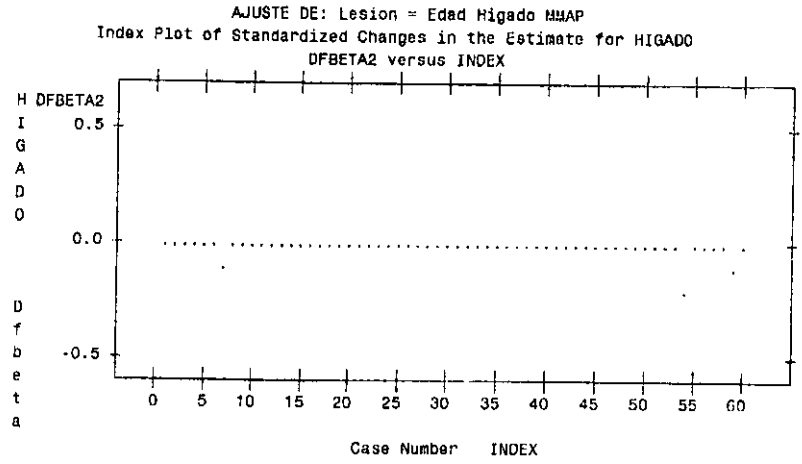
Grafica 3B, Delta Beta para EDAD

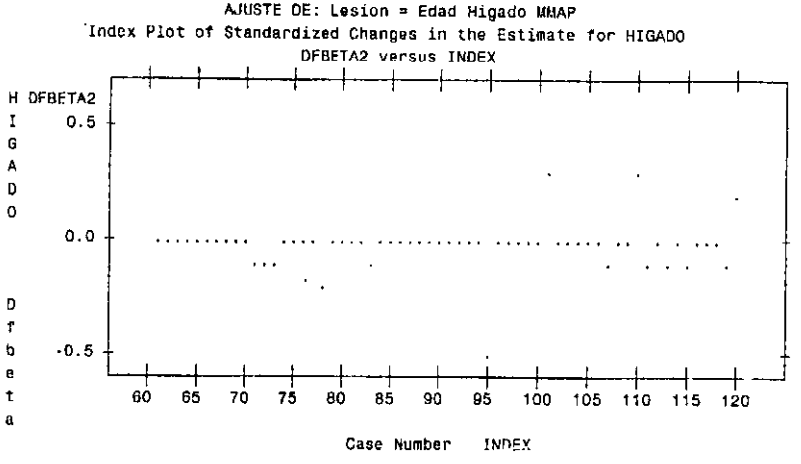


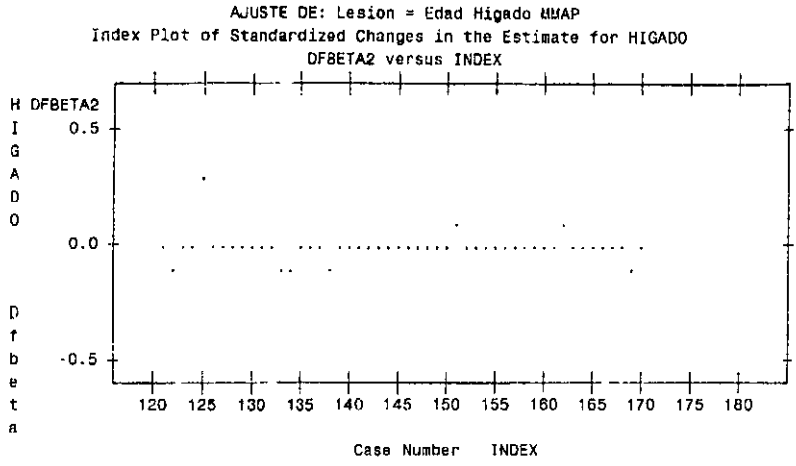

AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Standardized Changes in the Estimate for EDAD
DFBETA1 versus INDEX
E DFBETA1
D 0.5
A
D
0.0
-0.5
D
f -1.0
b
e
t
a
60 65 70 75 80 85 90 95 100 105 110 115 120
Case Number INDEX


AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Standardized Changes in the Estimate for EDAD
DFBETA1 versus INDEX
E DFBETA1
D 0.5
A
D
0.0
-0.5
D
f -1.0
b
e
t
a
120 125 130 135 140 145 150 155 160 165 170 175 180
Case Number INDEX

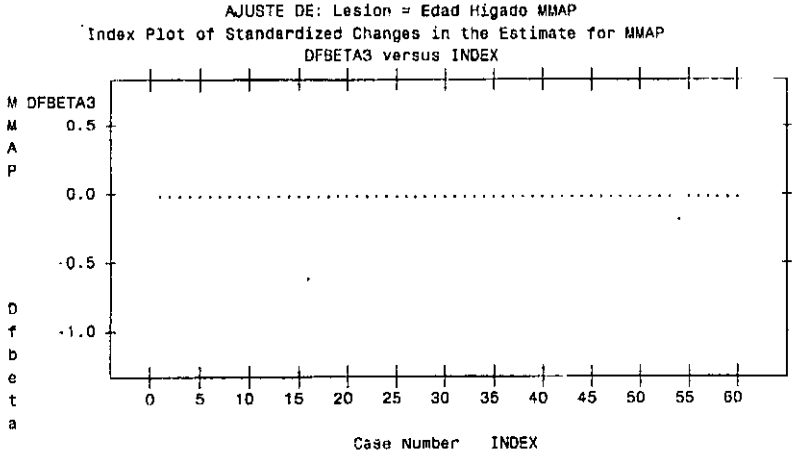
Grafica 3C, Delta Beta para HIGADO

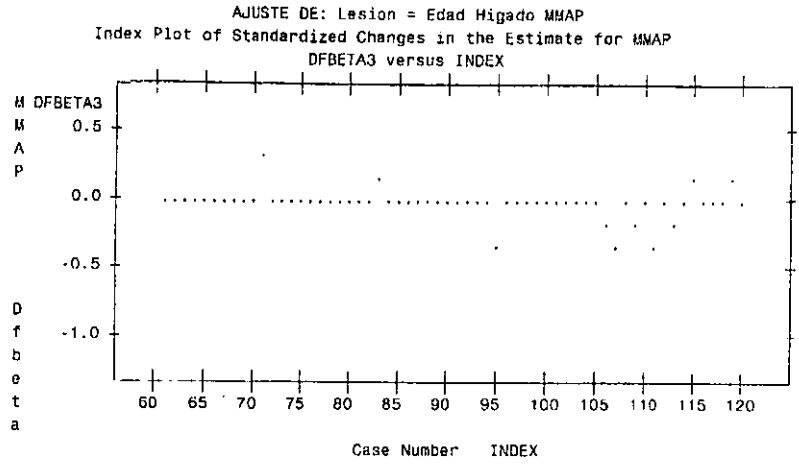


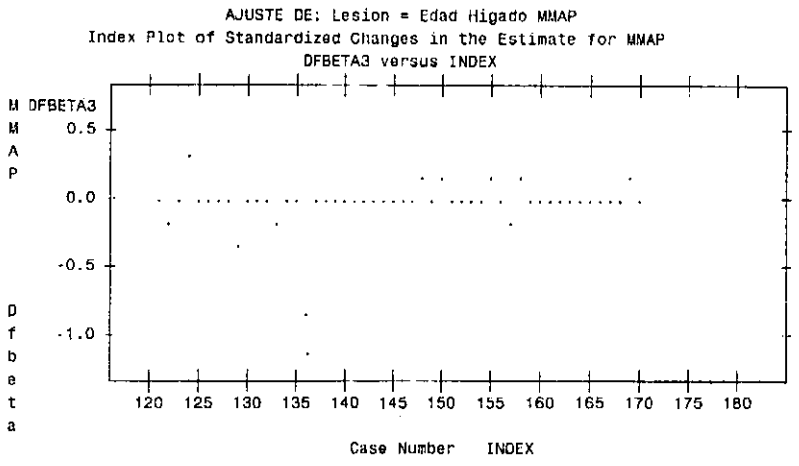




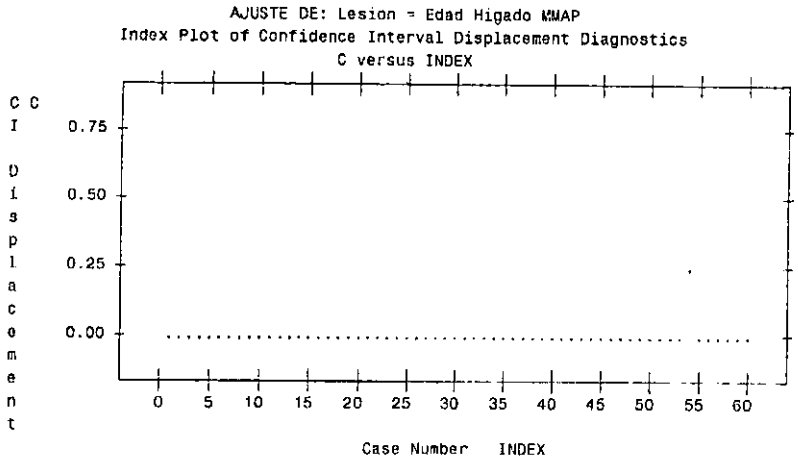
Grafica 3C, Delta Beta para MMAP

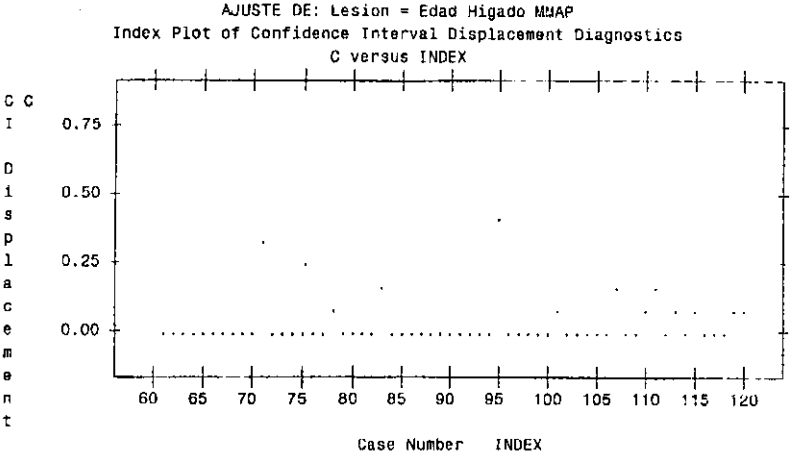


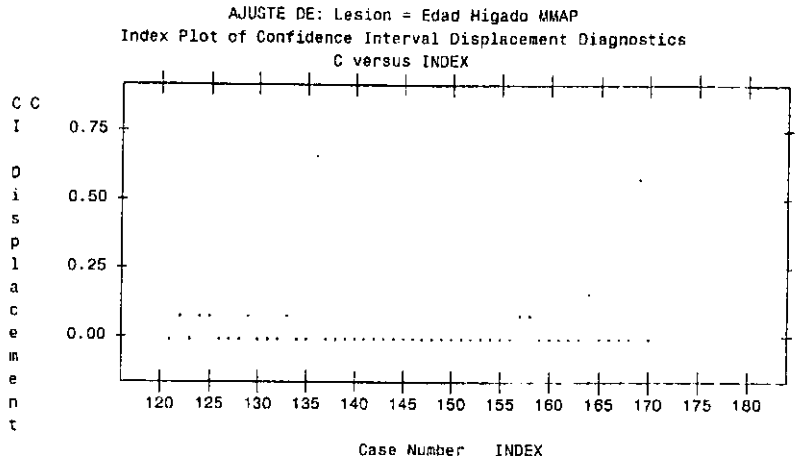

AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Standardized Changes in the Estimate for MMAP
DFBETA3 versus INDEX
M DFBETA3
M
A
P
0.5
0.0
-0.5
-1.0
D
f
b
e
t
a
60 65 70 75 80 85 90 95 100 105 110 115 120
Case Number INDEX


AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Standardized Changes in the Estimate for MMAP
DFBETA3 versus INDEX
M DFBETA3
M
A
P
0.5
0.0
-0.5
-1.0
D
f
b
e
t
a
120 125 130 135 140 145 150 155 160 165 170 175 180
Case Number INDEX

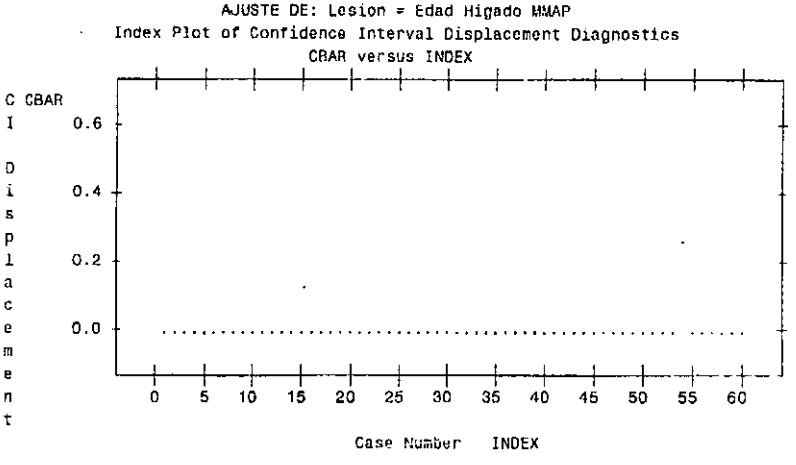
Grafica 4A, Residuales C_i

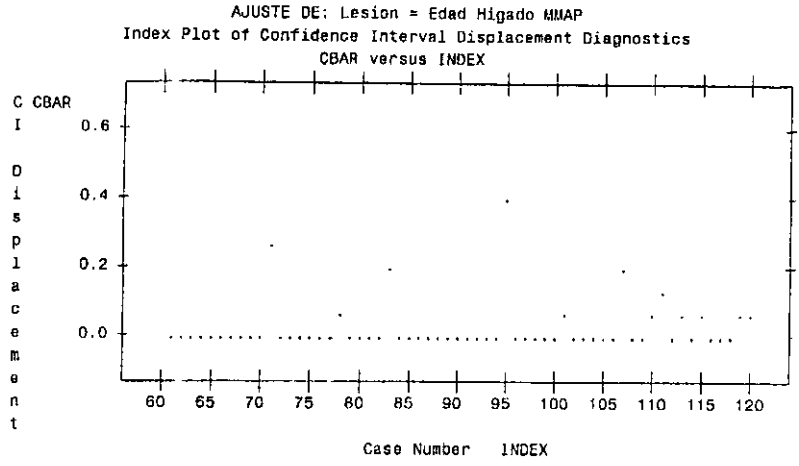


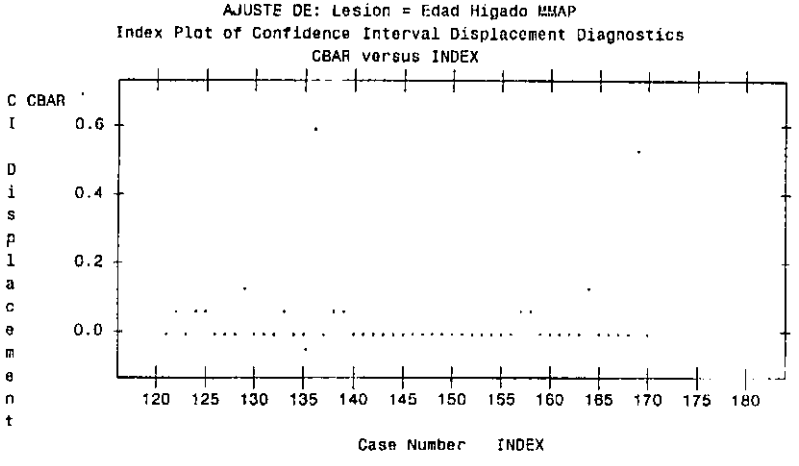




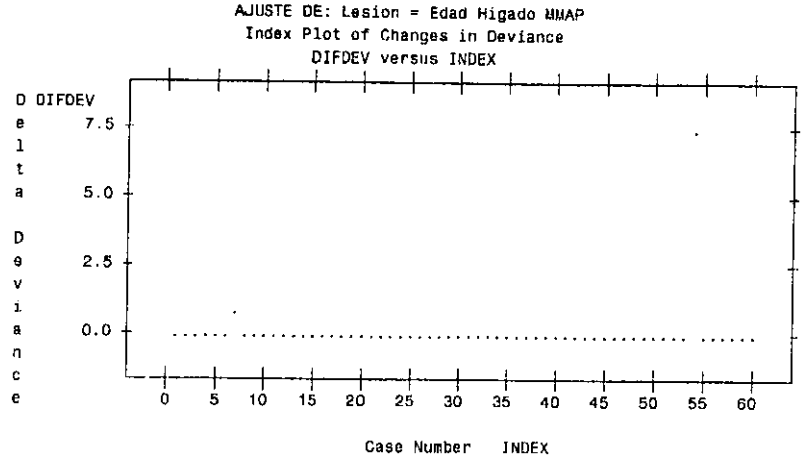
Grafica 4B, Residuales $\overline{C}_i$

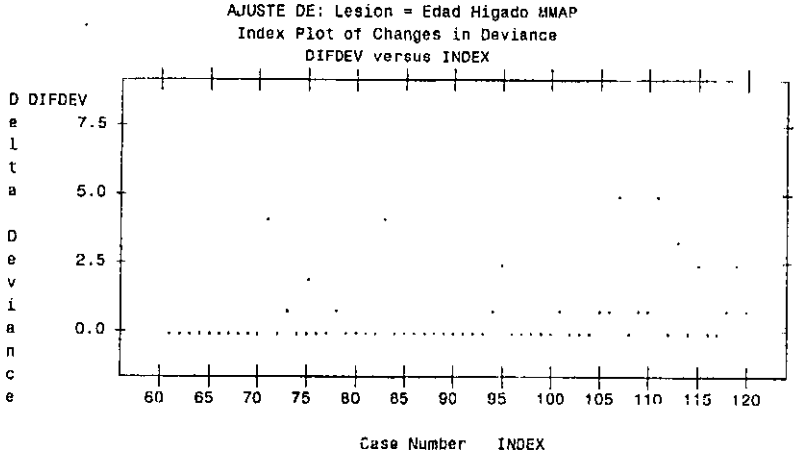


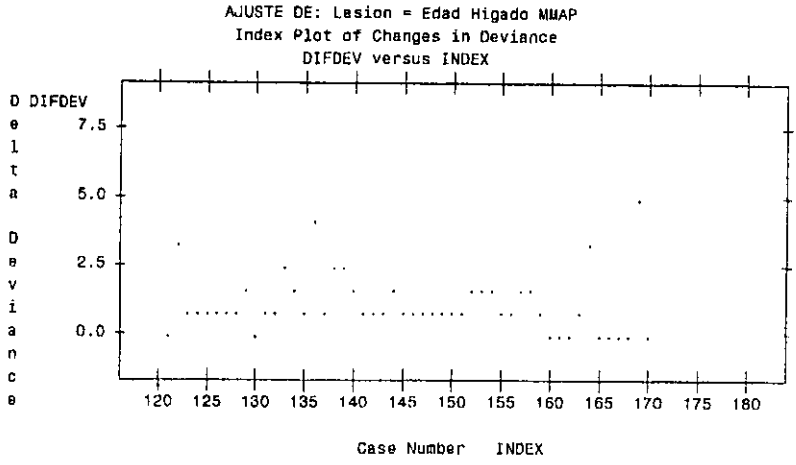

AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Confidence Interval Displacement Diagnostics
CBAR versus INDEX
C CBAR
I
D
i
s
p
l
a
c
e
m
e
n
t
0.6
0.4
0.2
0.0
60 65 70 75 80 85 90 95 100 105 110 115 120
Case Number INDEX


AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Confidence Interval Displacement Diagnostics
CBAR versus INDEX
C CBAR
I
D
i
s
p
l
a
c
e
m
e
n
t
0.6
0.4
0.2
0.0
120 125 130 135 140 145 150 155 160 165 170 175 180
Case Number INDEX

Grafica 5, DifDev o Desviación Delta






AJUSTE DE: Lesion = Edad Higado MMAP
Index Plot of Changes in Deviance
DIFDEV versus INDEX
D DIFDEV
e
l
t
a
D
e
v
i
a
n
c
e
7.5
5.0
2.5
0.0
120 125 130 135 140 145 150 155 160 165 170 175 180
Case Number INDEX

APENDICE III

MACROS AUXILIARES

```
*** Genera gráficas ***;
%Macro Plot(Data,y,x);
       Proc Plot Data=&Data;
            Plot &y*&x='.' / Box HAxis=By 5;
       Quit;
%Mend Plot;

*** Genera variable SAS id para identificar observaciones bernoulli ***;
%Macro Id(Data,X,Grupo,Cond);
       %Local i;
       Proc Sort Data=&Data Out=Id;
            %If %Scan(&Cond,1) ne %Then Where &Cond;;
            By &x &Grupo;
       Run;
       Data Id;
            Set Id;
                By &x &Grupo;
            If
               %Let i=1;
               %Do %While(%Scan(&x &Grupo,&i) ne);
                   %If &i=1 %Then First.%Scan(&x &Grupo,&i);
                   %Else and First.%Scan(&x &Grupo,&i);
                   %Let i=%Eval(&i+1);
               %End;
               Then id+1;
       Run;
%Mend Id;
```

```
*** Calcula residuales de desviación ***;
%Macro LogitRes(data,n,y);
       Data &Data;
            Set &Data;
            yhat=&n*phat;
            w=&n*phat*(1-phat);
            h=w*stderror**2;
            if phat=0 or phat=1 then Xi=0;
            else Xi=(&y-&n*phat)/(&n*phat*(1-phat))**0.5;
            if &y=0 then
               di=-(2*&n*log(&n/(&n-yhat)))**0.5;
            else if &y=&n then
                   di=(2*&y*log(&y/yhat))**0.5;
                 else
                     di=sign(&y-yhat)*Abs(2*&y*log(&y/yhat)+2*(&n-&y)*log((&n-&y)
                        /(&n-yhat)))**0.5;
            rpi=Xi/Abs(1-h)**0.5;
            rDi=di/Abs(1-h)**0.5;
            rLi=sign(&y-yhat)*(h*rpi**2+(1-h)*rDi**2)**0.5;
            rDiAbs=abs(rdi);
            di2=di**2; * Residuales de desviación al cuadrado ;
            k=1; * Truco para hacer "merge" con una sola observacion ;
            * Nota: Los valores absolutos al interior de las raices cuadradas están
                  para evitar números negativos provocados por redondeos de la
                  máquina;

     Run;
%Mend LogitRes;

*** Macro %FitLogit Ajusta un modelo de regresión logística ***;
%Macro FitLogit(DataIn,y,X,Grupo,Graficas,Cond);

      ********** Declaración de variables **********;
      %Local i;        * Contador ;
      %Local k;        * Contador ;
      %Local Exp;      * Listado de nombres de variables explicativas ;
      %Local Cat;      * Listado de nombres de variables de clasificación ;
      %Local NVar;     * Número de variables independientes (&Exp+&Cat);
      %Local NObs;     * Número de observaciones sin valores faltantes en
                         cualquiera de las variables independientes ;
      %Local Opc;      * Opción para graficas generadas por SAS ;
      %Local Print;    * Bandera para suprimir la salida impresa del SAS ;
      %Local Proc;     * Bandera para identificar si se utiliza el Proc Logistic
                                o el Proc Probit;
```

```
* Construye la entrada de datos binomiales ;
Proc Summary NWay Data=&DataIn(%If %Scan(&Cond,1) ne %Then Where=(&Cond););
     Class &x &Grupo;
     Var &y;
     Output Out=Binomial (Drop=_Type_ _Freq_) Sum=y n=n;
Run;

* Numera a las observaciones binomiales;
Data Binomial;
     Set Binomial;
     r=y/n;
     i=_n_;
Run;

* Elimina las observaciones con alguna variable sin valor ;
Data Temp;
     Set Binomial;
     Array Vars y n &x &Grupo;
     NoMiss=1;
     Do Over Vars;
        If Vars = . Then NoMiss=0;
     End;
     If NoMiss Then Output;
     Drop r;
Run;

* Número de parámetros a estimar mas 1 (el intercepto) &NVar;
%Let NVar=1;
%Do %While(%Scan(&X,&NVar) ne);
    %Let NVar=%Eval(&NVar+1);
%End;

* Número de observaciones binomiales;
Data _Null_;
     Set Binomial Nobs=Nobs;
     Call Symput('Nobs',Nobs);
Run;

Proc DataSets NoList;
     Delete OuTest;
Run;
Quit;

* Estimación del modelo de regresión logística ;
%Put AJUSTANDO &y = &x;
%Put G.L.=%Eval(&Nobs-&Nvar);
```

```
%Let k=1;
%Let Opc=;
%Let Print=;
%Let Proc=;
%Do %While(%UpCase(%Scan(&Graficas,&k)) ne);
    %If %UpCase(%Scan(&Graficas,&k)) eq INFLUENCE %Then
                                                %Let Opc=&Opc Influence;
    %If %UpCase(%Scan(&Graficas,&k)) eq IPLOTS %Then %Let Opc=&Opc Iplots;
    %If %UpCase(%Scan(&Graficas,&k)) eq NOPRINT %Then %Let Print=NoPrint;
    %If &Opc ne %Then %Let Proc=Logistic;
    %Let k=%Eval(&k+1);
%End;

Title 'AJUSTE DE: &y = &x';
%If %Scan(&Cond,1) ne %Then Title2 Condición de selección: &Cond;;
%If %Scan(&Grupo,1) ne %Then Title3 Variable de agrupamiento: &Grupo;;

%If &Proc=Logistic %Then %Do;
    %Put *****Proc Logistic *****;
    Proc Logistic OuTest=OuTest &Print;
         Model  y/n = &x / &Opc;
         OutPut Out=Residual Prob=Phat StdXBeta=StdError XBeta=XBeta;
    Run;
%End;
%Else %Do;
      %Put *****Proc Probit *****;
      Proc Probit OuTest=OuTest &Print;
           Model  y/n = &x /d=logistic;
           OutPut Out=Residual Prob=Phat Std=StdError XBeta=XBeta;
      Run;
%End;

* Calcula residuales de desviación ;
%LogitRes(Residual,n,y)

* Suma residuales cuadrados ;
Proc Univariate Data=Residual NoPrint;
     Var di2;
     OutPut Out=Ajuste Sum=Deviance;
Run;
```

```
      * Asigna grados de libertad y valores para presentar el resultado ;
        Data Ajuste;
             Set Ajuste;
             Y = "&y";
             Length X $80;
             X = "&X";
             NObs=&NObs;
             NVar=&NVar;
             DF=NObs-NVar;
             MeanDev=Deviance/DF; * Mean Deviance o Desviacion Media;
        Run;

        * Genera grafica de residuales parciales;
        %Let k=1;
        %Do %While(%UpCase(%Scan(&Graficas,&k)) ne);
           * Opcion PR : Partial Residual Plot ;
           *** NOTA: Para hacer esta grafica se requiere que no se hayan definido
                     variables de clasificacion;
           %If %UpCase(%Scan(&Graficas,&k)) eq PARTIAL and &X ne %Then %Do;
               %Put Partial Residual Plot Bo + B1*(&X);
               Title "PARTIAL RESIDUAL PLOT, Modelo &y=&X";
               Data _Null_;
                    Set OuTest;
                    Call Symput('Beta',&X);
                    Put 'Model=' Intercep '+' &X "&X";
               Run;
               Data Plot;
                    Set Residual;
                    Partial=(y-n*phat)/(n*phat*(1-phat))**0.5+&Beta*(&X);
               Run;
               %Plot (Plot,Partial,&X)
           %End;
           %Let k=%Eval(&k+1);
        %End;

        * Identificación de observaciones;
        %If %Index(%UpCase(&Graficas),IDNOITER) ne 0 %Then %Do;
            Title "IDENTIFICACION DE OBSERVACIONES";
            Title2 "Modelo: y=&y x=&x Grupo=&Grupo";
            Proc print data=Binomial;
                 Id i;
                 Var &x &Grupo y n r;
                 Format r 4.2;
            Run;
        %End;
        Proc DataSets NoList;
             Delete Temp Plot;
        Run;
        Quit;
%Mend FitLogit;
```

```
*** Macro %FitXVar Estima un modelo para cada variable y calcula su desviación ;
%Macro FitXVar(Data,y,X,XFijas,Grupo,Graficas,Cond);
   %Local j;
   %Local VarX;

   Proc DataSets NoList;
      Delete Deviance;
   Run;
   Quit;

   *** Estima un modelo para cada variable y compara con el intercepto ;
   %Let j=1;
   %Do %While(%scan(&X,&j) ne);
       %Let VarX=;

       *** Selecciona a la variable actual ;
       %Let VarX=%Scan(&X,&j);

       *** Estima un modelo para el intercepto (y variables fijas);
       %FitLogit(&Data,&y,&XFijas,&VarX &Grupo,NoPrint,&Cond);
       Proc DataSets NoList;
            Delete FitInt Resid_0;
            Change Ajuste=FitInt Residual=Resid_0;
       Run;
       Quit;

       *** Estima el modelo para la &j ésima variable;
       %FitLogit(&Data,&y,&XFijas &VarX,&Grupo,&Graficas,&Cond);
       Data Temp;
            Merge FitInt(Rename=(Deviance=DevFijo DF=DFFijo MeanDev=MDFijo))
                  Ajuste(Rename=(Deviance=DevModel DF=DFModel MeanDev=MDModel));
            XFijas="&XFijas";
            p=1-ProbChi(DevFijo-DevModel,DFFijo-DFModel);
       Run;
       Proc Append Base=Deviance Data=Temp;
       Run;

       *** Impresión de valores ;
           %If %Index(%UpCase(&Graficas),INFLUENCE) ne 0 %Then %Do;
               Proc Print Data=Di;
                    Id i;
                    Var &XFijas &VarX &Grupo Rdi_0 Rdi Dif_Rdi;
               Run;
           %End;
       %End;
```

```
        * Identificación de observaciones;
        %If %Index(%UpCase(&Graficas),ID) ne 0 %Then %Do;
            Title "IDENTIFICACION DE OBSERVACIONES";
            Title2 "Modelo: y=&y x=%scan(&X,&j) Grupo=&Grupo";
            Proc print data=Binomial;
                 Id i;
                 Var &XFijas %scan(&X,&j) &Grupo y n r;
                 Format r 4.2;
            Run;
        %End;

        %Let j=%Eval(&j+1);
    %End;
    Proc Sort Data=Deviance;
        By p MDModel;
    Run;
    Title1 'Analisis de desviación para regresión logística';
    Title2 "Valor p para el modelo fijado a &y";
    %If %Scan(&Cond,1) ne %Then Title3 Condición de selección: &Cond;;
    %If %Scan(&Grupo,1) ne %Then Title4 Variable de agrupamiento: &Grupo;;
    Proc Print Data=Deviance NoObs;
         Var XFijas X DevModel DfModel DevFijo DFFijo p NObs NVar ;
    Run;

    Proc DataSets NoList;
         Delete Temp;
    Run;
    Quit;
%Mend FitXVar;
```

```
*** DESCRIPCIÓN DE PARAMETROS ***;
*************************************************************************************
**********
%FitLogit(DataIn,y,X,Grupo,Graficas,Cond)
          PARAMETROS:
               DataIn      : Archivo SAS de entrada
               y           : Variable binaria dependiente
               X           : Vector de variables independientes
               Grupo       : Vector de variables de agrupamiento
               Cond        : Condición de selección de observaciones
               Graficas    : Descripción de las gráficas que se desea generar
                             NoPrint
                             IPlots
                             Influence
                             PR
                             IDNoIter
          ARCHIVOS DE SALIDA:
               Ajuste      : Nombre del archivo de datos SAS que contendrá el resumen
del ajuste
               Residuales  : Nombre del archivo de datos SAS con los residuales de
desviación
*************************************************************************************
**********
%FitXVar(Data,y,X,XFijas,Grupo,Graficas,Cond)
          PARAMETROS:
               Data        : Archivo SAS de entrada
               y           : Variable binaria dependiente
               X           : Vector de variables independientes
               XFijas      : Vector de variables fijas
               Grupo       : Vector de variables de agrupamiento
               Graficas    : Descripción de las gráficas que se desea generar
                             NoPrint
                             iPlots
                             Influence
                             PR
               Cond        : Condición de selección de observaciones
*************************************************************************;
```

BIBLIOGRAFIA

Agresti, A. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, inc, 1984.

Atkinson, Kendall E. *An Introduction to Numerical Analysis*. John Wiley & Sons, Inc. Second Edition. Singapore 1989.

Breslow N.E. & Day N.E. *Statistical Methods in cancer research, Volume 1, The analysis of case-control studies*. IARC Scientific Publications No. 32, LYON 1980.

Cebrian M. E., Albores A., Aaguilar M. & Blakely E. (1983). Chronic arsenic poisoning in the North of México. *Human Toxicol.*, **2**: 121-133.

Challenger, F. (1945) Chem. Rev. **36**, 315.

Collett D. *Modelling Binary Data*. Ed. Chapman & Hall, first edition, New York, 1991.

Daniel H. Freeman. *Applied Categorical Data Analysis*. New York.

Davis, W. E. et al. (1971) National inventory of sources of emissions. Arsenic, beryllium, manganese, mercury, and vanadium (1968) in: W. W. Davis et al. (eds)., Leawood, Kansas. 51 pp

Del Razo, L.M., Styblo, M., Thomas, D.J. Determination of Trivalent Methylated Arsenic Species in Water, Cultured Rat Hepatocytes, and Human Urine. 4th International Conference on Arsenic Exposure and Health Effects. San Diego, California. June 2000.

Del Razo L.M. et. al. Urinary excretion of arsenic species in a human population chronically exposed to arsenic via drinking water. A pilot study. in: Arsenic, Exposure and Heald (Science and technology letters). Ed. Chappell W.L. et. al.Laws & Stimson Associates, Northwood, 1994.

Del Razo L.M. et. al. Altered profile of urinary arsenic matabolites in patients with chronic arsenicism, a pilot study. Arch. Toxicol. 1996 (en prensa).

Del Razo María de la Luz. Biotransformación del arsénico y su relación con las lesiones de piel en individuos expuestos crónicamente al metaloide. Departamento de farmacología y toxicología. Sección de toxicología ambiental. CINVESTAV. México, D.F. 1997.

Edelstein-Keshet, Leah. (1988). *Mathematical models in biology*. Birkhäuser Mathematic Series. First Edition. Random House, New York. 586 pags.

Environmental Protection Agency (1984). Special report on ingested inorganic arsenic: Skin cancer, nutritional essenciality. EPA 625/3-87/013. U.S. Environmental Protection Agency, Washington, D.C.

EPA. Special report on ingested inorganic arsenic. Skin cancer; Nutritional Essenciality. Risk EPA/625/3-87/013. Risk Assessment Forum W;S: Environmental Protection Agency. Washington, D.C. 20460 1988

Hughes, M.F., Del Razo, L.M., Kenyon E.M.,. Dose-dependent effects on tissue and subcellular distribution and metabolism of dimethylarsinic acid in the mouse after intravenous administration.Toxicol.143, 155-166 (2000)

Kelsey L. J., Thompson W. D., Evans A.S. (1986). Methods in Observational Epidemiology. Oxford University Press, New York. pp. 366

Le, X.C., Ma, M, Lu, X., Cullen, W.R., Aposhian, H.V. and Zheng, B.. Determination of monomethylarsonous acid (MMAIII), a key arsenic methylathion intermediate,in human urine. Environ. Health Perspect.108 1015-1018 (2000)

Leslie A. C. D. and Smith H. (1978) Napoleon Bonaparte's exposure to arsenic during 1816. *Arch. Toxicol.* **41** : 163-167.

Lin, S., Cullen W.R. and Thomas D.J. (1999). Methylarsenicals and arsinothiols are potent inhibitors of mouse liver thioredoxin reductase. Chem. Res. Toxicol. 12,924-930 (1999)

McCullagh, P., and J. A. Nelder. 1983, 2nd edn. 1989. *Generalized Linear Models*. London: Chapman and Hall.

National Research Council (NRC). 1999. Arsenic in Drinking Water. National Academy Press; Washington, D.C.

Pershagen, G., Wall, S., Taube, A. and Linnman, L. (1981) Scand. J. *Work Environ. Health* **7**, 302-309.

Pershagen Göran (1983). *The epidemiology of human arsenic exposure. In: Biological environmental effects of arsenic (Topics in Environmental Health; v.6.).* Ed. Bruce A. Fowler. Elsevier Science Publisher, New York. pp 199-232.

Petrick, J.S., Ayala-Fierro, F., Cullen, W.R., Carter, D.E., Aposhian, H.V.,. Monomethylarsonous acid (MMAIII) is more toxic than arsenite in Chang human hepatocytes. Toxicol. Appl. Pharmacol 163, 203-207 (2000).

Ronald P. Cody and Jeffrey K.Smith. Applied Statistics and the SAS Programming Language, third edition, Prentice Hall, New Yersey, 1991.

SAS Institute Inc. SAS / STAT User's Guide, version 6, fourth edition, volume 1,2, SAS Institute Inc. Cary, NC, USA, 1989.

Styblo M., Del Razo L.M., Libia Vega, Dori R. Germolec, Edward L. LeCluyse, Geraldine A. Hamilton, William Reed, Changqing Wang, William R. Cullen and David J Thomas. Comparative toxicity of trivalent and pentavalent inorganic and methylated arsenicals in rat and human cells.. Arch Toxicol 74: 289-299 (2000)

Vallee B. L., Ulmer D. D. & Wacher W. E. C.. (1960). Arsenic toxicology and biochemistry. *A. M. A. Arch. ind. Healt.* **21**:132-151.

Vahter, M. (1983). *Metabolism of arsenic. In: Biological environmental effects of arsenic (Topics in Environmental Health; v.6.).* Ed. Bruce A. Fowler. Elsevier Science Publisher, New York. pp.171-198.

Woolson E. A. (1983). *Emissions, cycling and effects of arsenic in soil ecosystems. In: Biological environmental effects of arsenic (Topics in Environmental Health; v.6.).* Ed. Bruce A. Fowler. Elsevier Science Publisher, New York. pp. 51-139.