20
24

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES

" ACATLAN "

"UN ESTUDIO SOBRE LOS PARTOS POR CESAREA Y SUS CAUSAS USANDO UN MODELO DE REGRESION LOGISTICA MULTIPLE"

T E S I S

QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN MATEMATICAS APLICADAS Y COMPUTACION

P R E S E N T A N:

ANABEL MORENO BALTAZAR

ANDRES HERNANDEZ BALDERAS

ACATLAN, ESTADO DE MEXICO, **1995**

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES ACATLAN

LIC. MATEMATICAS APLICADAS Y COMPUTACION

ANABEL MORENO BALTAZAR ANDRES HERNANDEZ BALDERAS

"UN ESTUDIO SOBRE LOS PARTOS POR CESAREA Y SUS CAUSAS USANDO UN MODELO DE REGRESION LOGISTICA MULTIPLE"

ASESOR ING. ELVIRA BEATRIZ CLAVEL DIAZ

Gracias:

A mi madre por el apoyo que me brindó durante mi carrera, que es un tesoro invaluable. Por ser, más que mi madre, mi amiga.

A mi tía Esther por ser como una segunda madre para mí y por el gran cariño que siempre me ha dado.

A Andrés por apoyarme en los momentos más difíciles de mi carrera y por impulsarme a seguir adelante. Gracias por tu compañía y amor.

A mi familia por el apoyo que, en algún momento de mi vida, me brindaron.

Anabel.

A mis padres por los esfuerzos que han hecho por mí, no por sacrificio, sino por amor.

A mis hermanos por su confianza y ayuda durante mi carrera y a lo largo de toda mi vida.

A Anabel por su cariño, su amistad, su comprensión y por darle un nuevo sentido a mi vida. Gracias por ser la mejor pareja y amiga que puede haber encontrado.

Andrés.

Gracias:

A Dios	por los muchos momentos que nos lleva sobre sus brazos cuando recorremos el difícil camino de la vida. Gracias por darnos la vida para llegar hasta este momento.
A la U.N.A.M.	por darnos la oportunidad de prepararnos en sus aulas.
A nuestros profesores	por las enseñanzas y consejos que nos brindaron a lo largo de nuestros estudios.
A nuestros amigos	por su amistad, su ayuda y la confianza que siempre nos han tenido. Gracias por los buenos momentos que hemos pasado juntos.

Anabel y Andrés.

UN ESTUDIO SOBRE LOS PARTOS POR CESAREA Y SUS CAUSAS USANDO UN MODELO DE REGRESION LOGISTICA MULTIPLE

ANABEL MORENO BALTAZAR
ANDRES HERNANDEZ BALDERAS

PROLOGO

En el transcurso del tiempo el ser humano siempre se ha preocupado por tomar decisiones correctas, por ejemplo: qué tipo de producto lanzar al mercado para que éste se venda, aceptar o no un proyecto de inversión, construir un edificio o no en un determinado terreno, etc. Hay diversos métodos para analizar estos problemas que van desde la simple aplicación de la experiencia hasta el uso de herramientas matemáticas avanzadas.

El problema de interés de esta tesis es la elección del tipo de parto que tendrá una mujer de acuerdo a características especificas de su embarazo. En la actualidad, los médicos toman esta decisión con base a sus conocimientos y experiencia adquirida durante su práctica profesional; sin embargo, se puede considerar que a estos médicos se les podría auxiliar con un modelo probabilístico capaz de describir este fenómeno y con ello poder pronosticar el tipo de parto. En este trabajo se desarrolla este modelo.

De acuerdo a la naturaleza del fenómeno en el cual sólo puede ocurrir uno de dos tipos de parto (cesárea o parto natural) se hace la suposición de que la información obtenida de las características de los embarazos se ajustan a un modelo de Regresión Logística. Este modelo tiene la característica principal de modelar problemas que tengan una respuesta binaria por lo cual será un medio eficaz de diagnóstico del tipo de parto que tendrá una mujer embarazada y se podrán determinar cuáles son las principales causas que ocasionan un nacimiento por cesárea.

El objetivo principal de esta investigación fue el ajustar el modelo mencionado a un conjunto de datos de los embarazos de un grupo de mujeres bajo estudio que fueron atendidas en el Hospital General de México, así como realizar una comparación entre el método clínico usual y el método de diagnóstico por medio del modelo de Regresión Logística para saber si el modelo es eficiente para los pronósticos.

La metodología utilizada para el presente trabajo tiene validez en todos aquellos hospitales o clínicas, tanto públicas como privadas, que se dediquen a la ginecoobstetricia, aunque el modelo obtenido presentará variaciones para cada tipo de institución.

El modelo obtenido en este trabajo no pretende ser de ningún modo un sustituto del médico si no más bien pretende ser una herramienta que le auxilie en su decisión.

Es importante mencionar que para la realización del estudio se tomó una muestra no probabilística pero representativa, esto es, un muestreo sin normas de las pacientes del área de ginecoobstetricia, esto debido a que al obtener la información se restringió el acceso a ésta.

Otro de los propósitos de esta investigación es el mostrar la utilidad que tiene la Regresión Logística para el análisis de datos a pesar de que es una técnica poco utilizada.

Agradecemos la colaboración del Dr. Roberto Flores Guerrero, Subdirector de Servicios Auxiliares de Diagnóstico, Tratamiento y Paramédico y a la Srita. Graciela Nájera Colín, Jefe del Archivo Clínico, ambos del Hospital General de México, por su ayuda para obtener la información clínica necesaria para la realización de esta investigación, así como la de el M. en C. I. Rafael Madrid Ríos, coordinador de postgrado en Estadística del I.I.M.A.S, por sus valiosos consejos y por proporcionarnos un espacio dentro del instituto para poder hacer uso de los paquetes de cómputo SAS y BMDP. Gracias al profesor Noel Melgar Selvas por sus enseñanzas en la materia de seminario de tesis ya que fueron de gran utilidad en esta investigación. Un agradecimiento muy especial a la Ing. Elvira Beatriz Clavel Díaz asesora de esta tesis por su paciencia, sus consejos y su valiosa ayuda para llevar por un mejor camino este trabajo.

INDICE

INTRODUCCION

En la práctica se presentan un gran número de problemas en los cuales se desea determinar el valor de una variable de respuesta, con sólo dos posibles valores, en función de un número determinado de variables predictoras, por ejemplo comprar o no un equipo de cómputo para una empresa, construir o no una carretera, realizar o no una operación a un paciente, dar o no mantenimiento a una máquina, etc.

Muchas investigaciones han demostrado que la Regresión Logística es útil para modelar este tipo de situaciones. En esta tesis se hace uso de este modelo para poder pronosticar el tipo de parto de una mujer de acuerdo a las características de su embarazo. En el primer capítulo se presentan generalidades sobre la Obstetricia, se delimita la muestra y se dan las principales características de ésta así como los casos excluidos para la investigación y las técnicas utilizadas por los médicos para diagnosticar el tipo de parto.

En el capítulo II se hace una presentación general de los modelos de Regresión Logística, se analiza la teoría básica de estos modelos y se explican algunas características de ellos como son: diferencias con otro tipo de regresión, los ámbitos donde se le utiliza, etc.

En la realización de los estudios donde se utiliza la Regresión Logística los cálculos son tan laboriosos que se hace indispensable el uso de software especializado que auxilie en el manejo de la información y la realización de los cálculos. En la actualidad ya hay varios paquetes que manejan la Regresión Logística y en el capítulo III se describen los paquetes utilizados en esta investigación. En este estudio es importante que se obtenga el modelo con el mejor ajuste posible, razón por la cual se utilizan dos paquetes, SAS y BMDP, para tener la oportunidad de comparar resultados. Estos paquetes tienen la característica de ser una poderosa herramienta para el análisis de datos y ser de fácil manejo para cualquier persona que tenga conocimientos en computación y una formación en estadística.

Antes de ajustar el modelo se debe establecer cuáles de las variables son estadísticamente significativas, entendiendo por esto que las variables influyen en la variable de respuesta. En el estudio estas variables representan las causas que ocasionan una cesárea y su incorporación al modelo se hará aplicando el método de selección de variables "Stepwise

Logistic Regression" descrito en el capítulo IV debido a que este método es uno de los más utilizados en la Regresión Logística porque proporciona un medio rápido y efectivo de selección de variables cuando el número de éstas es grande, además de ajustar simultáneamente el modelo.

Una vez que se han obtenido las variables estadísticamente significativas, en el capítulo V se verifica su importancia biológica con los especialistas en Ginecoobstetricia; hecho esto, se procede al ajuste del modelo usando los paquetes estadísticos estudiados y se aplican algunas pruebas de bondad de ajuste para decidir cuál es el mejor modelo.

Una de las características más importantes del modelo de Regresión Logística es que a partir de él se pueden obtener los Odds Ratio o razón de ventajas que son de gran importancia ya que representan una aproximación a lo que en diversas áreas, como la Epidemiología, se conoce como riesgo relativo, bajo algunas restricciones. En el capítulo VI se estudian los riesgos relativos que ocasionan una cesárea utilizando estos Odds Ratio.

Como se mencionó anteriormente, el principal objetivo de la investigación es encontrar un modelo probabilístico capaz de pronosticar el tipo de parto al que es susceptible una mujer. En el capítulo VII se hace uso del modelo de Regresión Logística para realizar pronósticos con él y probar así su eficiencia. El modelo de Regresión Logística tiene como salida solamente probabilidades por lo que en este capítulo se presentan reglas de predicción que indiquen a partir de qué valor de probabilidad se considera un parto por cesárea o bien un parto natural.

De este modo se desarrollará la investigación esperando que sea útil como un apoyo a los médicos ginecoobstetras en sus diagnósticos del tipo de parto que tendrá una paciente, así como también sea útil a las personas que se interesen en el estudio de los modelos de regresión.

CAPITULO I. CONTEXTUALIZACION DEL PROBLEMA

Uno de los acontecimientos de mayor importancia en la vida de una pareja es el nacimiento de un hijo y su mayor preocupación es la conservación de la vida y la salud de éste. Para que los hijos sean sanos es determinante la salud de la madre por lo que los cuidados durante el embarazo y el tratamiento correcto durante el parto tienen enorme trascendencia en el desarrollo y salud del niño.

La mala elección del tipo de parto que se practicará a una mujer a punto de parir puede tener graves consecuencias tanto para ella como para su hijo. Estas consecuencias van desde leves daños físicos hasta la muerte de uno o de ambos. Es por esta razón que el médico encargado de tomar esta decisión debe hacerla tomando en cuenta el mayor número de factores disponibles para que su diagnóstico sea adecuado.

1.1. GENERALIDADES SOBRE LA OBSTETRICIA

Se define a la obstetricia como la parte de la medicina que trata del embarazo, el parto y el puerperio. A continuación se hace una breve descripción de estas tres etapas.

1.1.1. EMBARAZO

Se define como embarazo al desarrollo de un ser en el útero de una mujer desde la fecundación (unión del óvulo con el espermatozoide) hasta el nacimiento. Dura aproximadamente de 270 a 280 días [Mendoza, (1992, p. 118)].

El desarrollo del feto comprende dos períodos: El período embrionario, dura aproximadamente dos meses en los cuales el feto toma forma, volumen y se forman los esbozos de los órganos principales; el período fetal, durante el cual el feto no sufre prácticamente mas que fenómenos de maduración [Tourris, (1974, p.221)]. Al término del período fetal se inicia el proceso de parto.

1.1.2. PARTO

El parto comprende tres períodos [Mendoza, (1992, p.p. 122-123)] los cuales se describen a continuación :

Primer período o de dilatación: se inicia con la regularización de las contracciones las cuales modifican al cuello de la matriz en su posición, consistencia y longitud, y termina con la dilatación completa del mismo.

Segundo período o expulsivo: se inicia con la dilatación completa del cervix y termina con la salida del producto.

Tercer período, placentario o de alumbramiento: Se inicia con la salida del feto y finaliza con la expulsión de la placenta y de las membranas ovulares.

De acuerdo a las dificultades que se presentan, los partos se pueden dividir en *parto normal o eutócico y parto anormal o distócico.*

Parto Normal o Eutócico

Los partos normales o eutócicos son aquellos que no presentan complicaciones durante el mecanismo del parto [Tourris,(1974, p.p. 246-255)]. En este mecanismo intervienen tres elementos:

1. El canal óseo formado por la pelvis que puede compararse a un cilindro acodado abierto por delante y las partes blandas que están constituidas por la vagina y el cuello de la matriz o cervix.

2. El móvil fetal que está constituido ante todo por la cabeza del feto ya que es el único elemento de éste que no puede comprimirse. Para que no haya complicaciones durante el nacimiento la cabeza debe flexionarse.

3. El motor uterino consta de las siguientes etapas: borramiento del canal cervical hasta transformarlo en un anillo fácilmente distendible, dilatar ese anillo hasta un diámetro de 10 cm. y expulsar el feto hacia el tubo uterovaginal fomado por el proceso anterior.

Parto Anormal o Distócico

Los partos distócicos o anormales son los que presentan problemas en sus fases del proceso natural, tanto por parte de la madre como del feto. Las principales causas se describen a continuación:

- **Causas concerniente a los órganos maternos**

a) La pelvis.- Lesiones, deformaciones, estrechez o angostura de la pelvis por el desarrollo incompleto de los huesos o deformaciones de estos que hacen que el canal del parto no dé paso al feto.

b) Vulva.- Deformaciones o estrechez de la vulva congénitas o adquiridas, presentación de tumores o heridas que pueden ser obstáculos para un parto normal.

c) Vagina.- Las lesiones o cicatrices de la vagina producen su estrechez y dificultan el parto, la persistencia del himen o presencia de lazos fibrosos dificultan el parto normal.

d) Utero.- Desprendimiento prematuro de la placenta, adherencia anormal de las paredes del útero a otros órganos genitales, anormalidad de las contracciones, estrechez, falta de elasticidad, debilidad o deformaciones.

- **Causas de origen fetal**

a) La cabeza.- puede presentarse demasiado grande por acumulación de líquidos.
b) Presentación viciada.- Se llama presentación normal o cefálica cuando el feto se coloca en la pelvis materna dispuesto a salir de cabeza hacia fines del embarazo, y presentación viciada cuando se dispone a salir en otra posición.
c) Procedencia del cordón umbilical que a veces se aboca hacia la pelvis materna.

En la realización de este estudio se dividirán a los partos en dos tipos:

1) Natural o fisiológico.- Es el parto que se efectúa por vía vaginal.

Como se mencionó arriba un parto natural es aquel que se produce por vía vaginal. Esto no quiere decir que esté libre de problemas, por ejemplo en partos fisiológicos muchas veces se hace necesario el uso de *fórceps* para ayudar a la extracción del niño.

2) Cesárea.- Cuando el nacimiento del niño se realiza a través de incisiones en las paredes abdominales y uterinas

En el transcurso de este trabajo se hace referencia con mayor frecuencia a las cesáreas que son el principal objetivo de la investigación.

Las cesáreas se realizan desde hace muchos siglos. El término cesárea probablemente proviene del término latino *"Caedere"*, que significa cortar, o de la ley romana *"Lex Caesaria"*, por la que se hacía la extracción abdominal del feto de una madre a punto de morir para salvar al niño. El índice de mortalidad después de una cesárea era muy alto hasta el año de 1882 cuando surgió la suturación de la incisión uterina. Actualmente, a la cesárea se recurre frecuentemente cuando el parto vaginal puede plantear riesgos para la madre, el feto o ambos [Neville,(1989,p.p. 269-270)].

CAUSAS QUE PRODUCEN UNA CESAREA

Las causas principales por las cuales se corre el riesgo de recurrir a una cesárea se pueden resumir en las siguientes:

a) *Desproporción cefalopélvica.* Se refiere a que el diámetro de la pelvis materna es menor que el diámetro cefálico del feto. La mayor parte de las decisiones de operación cesárea son por esta causa.

b) *Embarazo múltiple.* Generalmente cuando hay más de un producto se presentan dificultades en el parto.

c) *Funcionamiento uterino anormal.* Son problemas que se presentan en la matriz de la madre, por ejemplo: cáncer, tumores, etc.

d) *Cirugía uterina previa.* Principalmente por el hecho de haber tenido cesáreas anteriores.

e) *Presentación pélvica del feto.* La presentación pélvica es aquella en la cual el niño se prepara a nacer sentado o de pies.

f) *Situación transversa del feto.* La situación normal del feto debe ser longitudinal, la situación transversa es la que se conoce comúnmente como "atravesada".

g) *Obstrucción del canal del parto.* Se debe principalmente a problemas de la pelvis, la vulva o la vagina.

h) *Complicaciones médicas u obstétricas*, por ejemplo placenta previa, desprendimiento prematura de la placenta, sufrimiento fetal, enfermedades hipertensivas, etc.

i) *Primigesta de edad avanzada*, esto es, la mujer que tiene su primera gestación después de los 35 años.

Las causas anteriores deben tomarse como relativas y no absolutas, por ejemplo, un embarazo múltiple puede tener un parto vaginal sin tener complicaciones que conduzcan a una cesárea, y una mujer que tenga una pelvis amplia puede necesitar esta operación.

Por lo que respecta a las cesáreas previas, en algunos médicos existe la creencia de que a una mujer que haya tenido anteriormente una cesárea se le debe diagnosticar el siguiente parto también como una cesárea. Contra estas creencias se han hecho investigaciones que demuestran lo contrario [Escobar, (1990)].

1.1.3. PUERPERIO

El puerperio es el tiempo que sigue al tercer período del parto, en total tiene una duración aproximada de 6 a 8 semanas. Está caracterizado por el regreso de los órganos reproductores a su estado de preembarazo [Mendoza, (1992,p.127)]. Comprende tres periodos:

Puerperio inmediato que abarca las primeras 24 hrs. después de concluido el parto.

Puerperio mediato el cual se extiende desde el fin del puerperio inmediato hasta los siguientes siete días.

Puerperio tardío el cual comprende desde el fin del puerperio mediato hasta los cuarenta días.

1.2. EL PROBLEMA DE LA DETERMINACION DEL TIPO DE PARTO

Se han tratado de generalizar los factores que determinan el tipo de parto, sin embargo, en la mayoría de los casos prácticos los médicos solamente toman en cuenta sus conocimientos y experiencia previa para decidir si se realizará un parto natural o una cesárea. A continuación se presentan en los cuadros 1.1, 1.2, 1.3 y 1.4 las indicaciones que dan algunos autores para diagnosticar una cesárea.

Mendoza (1992,p.p. 287-288)

1. Absolutos.	a) Desproporción cefalopélvica. b) Placenta previa central. c) Desprendimiento prematuro de placenta. d) Presentaciones y situaciones anormales del feto. e) Inminencia de rotura uterina.
2. Relativos.	a) Toxemia grave. b) Cáncer cervicouterino. c) Primigesta de edad avanzada. d) Presentación pélvica. e) Rotura prematura de membranas.
3. Electivas.	a) Cesárea iterativa. b) Embarazo a término si los embarazos anteriores han terminado en muerte habitual. c) Posmadurez.

Cuadro 1.1.

Gonzalez Merlo (1990 p.p. 29, 97-99)

1. Embarazo gemelar.	a) Si el 1er. producto está en presentación cefálica y el 2o. en presentación transversa. b) Si el 1er. producto está en presentación pélvica se realizará una cesárea electiva si el embarazo es menor a 36 semanas o si el 2o producto está en situación transversa o presentación pélvica c) Si el 1er. producto está en situación transversa siempre realizar cesárea.
2. El feto en presentación pélvica.	a) Historia de partos difíciles o fetos dañados. b) Primigrávida de edad superior a 35 años. c) Infertilidad o esterilidad previas. d) Gestación entre 26 y 33 semanas. e) Feto de peso superior a 3800 g. f) Pelvis no favorable. g) Feto que tiene desarrollo anormal. h) Placenta previa. i) Presentación de pies o brazo en la nuca. j) Parto de evolución lenta o sufrimiento fetal.
3. El feto en situación transversa.	a) Siempre cesárea.
4. Pelvis estrecha.	a) Se realiza cesárea cuando no permite la salida del feto.

Cuadro 1.2

Neville (1989 p. 270)

1. Matemofetales.	a) Desproporción cefalopélvica.
	b) Inducción fallida del parto.
	c) Funcionamiento uterino anormal.
2. Maternas.	a) Diabetes mellitus.
	b) Cardiopatías.
	c) Cáncer del cérvix o cuello de la matriz.
	d) Cesárea previa.
	e) Rotura uterina previa.
	f) Tumores musculares en la pared uterina
	g) Tumores ováricos o fibromas.
3. Fetales.	a) Sufrimiento fetal.
	b) Prolapso del cordón.
	c) Presentación pélvica.
	d) Situción transversa.
4. Placentarias.	a) Placenta previa.
	b) Abruptio placentario.

Cuadro 1.3

Willson (1991, p.p. 572-574)

1. Distocias mecánicas.	a) Desproporción cefalopélvica. b) Tamaño fetal excesivo. c) Tumores uterinos.
2. Trabajo de parto.	a) Cuando éste es disfuncional
3. Placenta.	a) Placenta previa. b) Desprendimiento prematuro de ésta.
4. Malposición fetal.	a) Situación transversa. b) Presentación pélvica.
5. Enfermedades hipertensivas.	
6. Operación cesárea previa.	
7. Indicaciones fetales.	a) Prolapso del cordón. b) Sufrimiento fetal. c) Feto inmaduro.
8. Primigesta mayor de 35 años.	

Cuadro 1.4

Como se puede observar, en la mayoría de los aspectos estos autores concuerdan, sin embargo si el estudio se realizará con datos de un hospital en particular lo conveniente es conocer los criterios que los médicos de esa institución manejan para la elección de una cesárea.

1.3. METODOLOGIA DE LA INVESTIGACION

Antes de iniciar la investigación es importante clasificarla para saber que tipo de estudio se realizará. Según Méndez et. al. (1994, p.p. 11-27) la investigación se puede analizar bajo distintos criterios:

De acuerdo con el periodo en que se capta la información, este estudio es RETROSPECTIVO debido a que la información que se obtendrá se recopilará de los archivos del hospital; esto quiere decir que la información es anterior al momento de la investigación y fue obtenida para otros efectos o propósitos.

De acuerdo a la evolución del fenómeno, este estudio es TRANSVERSAL ya que las variables sólo son medidas una vez en el transcurso de la investigación.

De acuerdo con la comparación de las poblaciones, esta investigación se considera como una mezcla de un estudio DESCRIPTIVO y uno COMPARATIVO. Descriptivo ya que se pretende describir la relación que existe entre el tipo de parto y factores tales como tipo de pelvis, dilatación cervical, posición del feto, etc.; Comparativo por que se hace una comparación entre dos poblaciones, las mujeres con parto natural y las mujeres con parto por cesárea y determinar las causas que intervienen para que una mujer esté en una población o en la otra.

De acuerdo con la interferencia de los investigadores en el fenómeno esta investigación es OBSERVACIONAL ya que en ningún momento se experimenta con alguna de las variables y no se tiene dominio sobre ellas.

De acuerdo a los criterios anteriores la investigación sobre el tipo de parto se clasifica como una mezcla de una ENCUESTA DESCRIPTIVA y una ENCUESTA COMPARATIVA. Algunas de las ventajas de este tipo de estudios son las siguientes:

1. Con él se puede sugerir, apoyar o rechazar una hipótesis de asociación entre variables.
2. Permite encontrar la prevalencia de alguna característica en una población.
3. Es útil para asentar las bases de estudios posteriores de otro tipo como uno longitudinal (realiza un seguimiento para estudiar la evolución de las variables) a fin de contrastar hipótesis.

4. Su diseño y ejecución es de bajo costo y rápido.
5. Es útil en la comparación de métodos de diagnóstico al evaluar sensibilidad y especificidad.

1.3.1. ANTECEDENTES

Este estudio se realizará en el Hospital General de México, el cual está adscrito a la S.S.A. y se encarga de prestar servicios de salud principalmente a gente que no posee algún tipo de seguro médico y que no dispone de los recursos necesarios para asistir a un hospital privado u otra institución.

El hospital cuenta con médicos especialistas en diversas áreas como son: Oftalmología, Odontología, Cardiología, etc. El área de interés de esta investigación es la de Ginecología y Obstetricia. En esta área se encargan de llevar un control del embarazo de las pacientes y de atenderlas durante el parto. Como se mencionó con anterioridad, la gente que recurre al hospital es de escasos recursos, razón por la cual la mayoría de las pacientes no tienen un control de su embarazo por lo que la única información disponible para el estudio es la que se recopila a su ingreso al hospital.

La forma en que los médicos del Hospital General deciden si una paciente necesita una cesárea es bajo los siguientes criterios:

- En caso de desproporción cefalopélvica. Esto es, si la pelvis de la madre no es lo suficientemente amplia para permitir la salida de la cabeza del niño.

- Presentación pélvica del feto al final del embarazo. El feto se encuentra en posición anormal al final del embarazo, se prepara a nacer sentado o de pies.

- Situación transversa del feto al final del embarazo. Es otra malposición del feto en la cual su columna es perpendicular a la de la madre por lo que es prácticamente imposible que se realice un parto natural.

- Sufrimiento fetal. El sufrimiento fetal se debe principalmente a asfixia del niño por lo que se debe precipitar el parto mediante una cesárea.

El resto de las causas que para ellos provocan una cesárea son relativas y son en su mayoría complicaciones médicas.

Cabe señalar que los médicos de este hospital no hacen uso de ningún tipo de modelo probabilístico que les auxilie en sus diagnósticos, por lo que el modelo que esta investigación pretende encontrar puede serles de gran utilidad.

1.3.2. OBJETIVOS

Como se mencionó anteriormente, la toma de decisión del tipo de parto que tendrá una paciente lleva consigo una gran responsabilidad por parte del médico que la toma, por lo cual éste debe hacer uso de todos los medios de diagnóstico que estén a su alcance para que su decisión sea la correcta, ya que es peligroso diagnosticar una cesárea cuando no es indispensable como un parto natural cuando éste no es posible.

Por la razón anterior surge la idea de realizar un modelo probabilístico capaz de pronosticar el tipo de parto que tendrá una paciente, esto con el objetivo de que el médico cuente con otra herramienta, además de sus conocimientos médicos, para tomar una decisión.

1.3.3. HIPOTESIS

En esta investigación se trabajará bajo la hipótesis de que se puede ajustar un modelo de Regresión Logística a los datos de las variables de investigación, las cuales serán descritas más adelante, con lo cual se podrán hacer pronósticos sobre el tipo de parto de una mujer de acuerdo con las características de su embarazo.

1.3.4. POBLACION BAJO ESTUDIO

La población utilizada para este estudio fueron las pacientes del área de Ginecoobstetricia del Hospital General de México cuyos hijos nacieron en dicha institución durante el mes de agosto de 1994. Según estadísticas, entre el 30 y 35 % de los nacimientos que se atienden en el hospital son por cesárea, la mayoría de las madres son jóvenes (70 % con menos de 25 años), el 80 % tiene una estatura por abajo del 1.60 mts. y aproximadamente el 50 % están en su primera gestación.

1.3.4.1. CRITERIOS DE INCLUSION Y EXCLUSION

Para recolectar la muestra se seguirán los siguientes criterios para incluir y excluir casos de la misma:

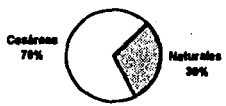
- Se tomarán en cuenta aquellos partos que hayan ocurrido en el hospital general de México durante el mes de agosto de 1994.
- No se tomarán en cuenta partos múltiples ya que éstos no son comunes en el hospital.
- No se tomarán en cuenta partos en los cuales el feto presente situación transversa por que al igual que los partos múltiples se presentan pocas veces.
- No se tomarán en cuenta casos especiales con problemas como: cáncer, sida, enfermedades cardiacas, enfermedades infecciosas, desprendimiento prematuro de placenta, diabetes, etc. ya que es necesario emplear tratamientos muy particulares.
- No se tomarán en cuenta partos post-término y pre-término así como casos con sufrimiento fetal ya que no se contó con la información suficiente en los archivos para incluirlos en el estudio.

1.3.5. MUESTRA

La muestra la componen 100 mujeres que cumplieron con los criterios de inclusión y exclusión. El método de muestreo utilizado fue un *muestreo sin normas* (la muestra se toma sin una regla en particular. Como la población bajo estudio es homogénea este método proporciona una muestra representativa). Este método se utilizó debido a que se encontraron algunas restricciones para acceder a la información de los archivos del hospital.

Algunas de las características de las pacientes incluidas en la muestra son las siguientes:

- 30 de ellas presentaron cesárea.
- 70 presentaron parto natural.



- La edad de las mujeres fue en promedio de 22.89 años con una desviación estándar de 5.63.
- La estatura promedio fue 153 cms. con una desviación estándar de 6.56.
- El peso promedio fue de 66.192 kg. con una desviación estándar de 10.76.

La muestra definitiva se encuentra en el anexo I.

1.3.6. DESCRIPCION DE LAS VARIABLES

Para el estudio se dispondrá de diferentes variables las cuales se describen a continuación:

1) Variable de respuesta o dependiente: Tipo de parto, recordando que para este estudio solo habrá dos tipos de parto, natural y cesárea.

2) Variables potencialmente predictoras o independientes:

- Tipo de pelvis.
- Dilatación cervical.
- Edad de la madre.
- Estatura de la madre.
- Peso de la madre.
- Tensión arterial de la madre.
- Número de partos anteriores.
- Número de cesáreas anteriores.
- Número de abortos anteriores.

- Estado laboral de la madre.
- Perímetro cefálico del feto.
- Posición del feto.

El valor de la variable *tipo de pelvis* determina si la pelvis materna es útil o no para la salida del producto actual, por lo que depende tanto de la pelvis como del perímetro de la cabeza del feto. Existe poca diferencia para la elección del momento para medir la pelvis, siempre y cuando se haga antes de que se inicie el trabajo de parto. Una pelvis anormal puede ser detectada mediante un examen clínico, pero si se desean obtener sus medidas y forma con precisión es necesario un estudio con rayos X [Willson, (1991 p.524)].

La variable *dilatación cervical* indica los centímetros que está dilatado el cervix para dar paso al producto. Se considera que una dilatación de 10 cm. es la más apropiada para un parto sin problemas. La dilatación cervical que se tiene en la muestra corresponde a la medida que se toma a las pacientes a su ingreso al hospital.

Es importante considerar características tales como la *edad, estatura, y peso* de la madre para determinar el tipo de parto puesto que estas variables podrían influir en este fenómeno.

La variable *tensión arterial* se considera debido a que la hipertensión (presión sanguínea elevada) es a menudo un síntoma de la toxemia del embarazo que puede poner en peligro la vida de la madre, el producto o producirles alteraciones permanentes por lo que es necesario precipitar el parto mediante una cesárea.

Las variables tales como *número de partos, número abortos y número de cesáreas* son de importancia puesto que éstas son parte del historial obstétrico de las pacientes y pueden también de algún modo definir el tipo de parto.

Con lo que respecta al *estado laboral de la madre*, éste se incluye para analizar si el hecho de que una madre trabaje o no influye en el tipo de parto.

La variable *perímetro cefálico* indica el perímetro que tiene la cabeza del niño. Los médicos consideran que mientras mayor sea este perímetro la posibilidad de una cesárea es mayor.

La *posición del feto* es una variable de interés para los médicos ya que con esto definen también el tipo de parto. Esta variable representa la "presentación" del feto dentro del útero. Se puede dividir en dos tipos principales a las presentaciones: pélvica, es la presentación en la cual el niño se prepara para nacer de pies o sentado; cefálica, se caracteriza por que el feto se prepara a nacer con la cabeza por delante [Willson, (1991, p. 416)].

La codificación de las variables y los valores que éstas utilizarán a lo largo de este trabajo se dan en el cuadro 1.5.

Nombre de la Variable	Descripción de la Variable	Códigos, Valores y Escalas de Medición
ID	Identificador de caso	1 ... 100
Tparto	Tipo de parto	0 = Parto normal. 1 = Cesárea.
Tpelvis	Tipo de pelvis	0 = Pelvis útil. 1 = Pelvis no útil.
Dilcer	Dilatación Cervical	Centímetros.
Edad	Edad de la madre	Años.
Estatura	Estatura de la madre	Centímetros.
Peso	Peso de la madre	Kilogramos.
Tarteria	Tensión Arterial	0 = Normal. 1 = Baja. 2 = Alta.
Npartos	Partos anteriores	Número de partos previos.
Ncesarea	Cesáreas anteriores	Número de cesáreas previas.
Nabortos	Abortos anteriores	Número de abortos previos.
Edomadre	Estado laboral	0 = No labora. 1 = Labora.
Pcefalic	Perímetro cefálico	Centímetros.
Posfeto	Posición del feto	0 = Presentación cefálica. 1 = Presentación pélvica.

Cuadro 1.5

1.3.7. ACTIVIDADES

En el desarrollo de esta investigación se realizaron las siguientes actividades:

- Obtención de la muestra en los archivos del Hospital General.
- Selección de las variables potencialmente predictoras para elegir aquellas que sean estadísticamente significativas para explicar la variable de respuesta.
- Complementación de las variables seleccionadas con las que los médicos consideren de importancia biológica.
- Ajuste del modelo de Regresión Logística a los datos.
- Diagnóstico del modelo de Regresión Logística.
- Determinación de factores de riesgo utilizando Odds Ratio (Razón de ventajas).
- Prueba del modelo final realizando pronósticos con él

CAPITULO II. EL MODELO DE REGRESION LOGISTICA

2.1. Modelos de Regresión

2.2. Definición del modelo Logístico

2.2.1. Modelo de Regresión Logística simple

2.2.2. Modelo de Regresión Logística múltiple

2.2.3. Casos que requieren el uso de la Regresión Logística

2.3. Regresión Logística vs. Regresión Lineal

2.4. Estimación por máxima verosimilitud

2.5. Inferencias sobre los coeficientes del modelo

2.6. Resumen de característica del modelo de Regresión Logística

2.1. MODELOS DE REGRESION

En la práctica, en un gran número de situaciones, el investigador se interesa en conocer la relación que existe entre una variable dependiente o de respuesta y una o varias variables independientes o predictoras. A la búsqueda de esta asociación se conoce en Estadística como análisis de regresión. Los métodos de Regresión buscan una relación entre las variables de respuesta y predictoras cuando ésta no es perfecta, esto quiere decir que no se tiene un único valor de la variable de respuesta para cada valor de la o las variables predictoras.

Las técnicas de Regresión proporcionan medios apropiados a través de los cuales pueden establecerse asociaciones entre las variables de interés en las cuales la relación usual no es causal [Canavos (1988, p. 445)]. Existen dos significados básicos de regresión los cuales se describen a continuación:

• Regresión basada en la distribución conjunta de probabilidad

El primer significado de Regresión está basado en la distribución conjunta de probabilidad, conocida, de dos o más variables aleatorias. En este caso se desea pronosticar el valor de una variable Y en función de la otra X (en el caso simple). Sin embargo la predicción de una variable en función de otra no es del todo exacta por lo que es preferible tratar de pronosticar el promedio de la variable Y para un determinado valor de X, esto es, para cada valor de X existe una distribución de Y, y lo que se pretende encontrar es la media de esta distribución. La curva de Regresión de Y sobre X o gráfica de la media condicional, E(Y/X), está dada por:

$$E(Y/X)=\int_{-\infty}^{+\infty} Yf(Y/X)dY$$

donde:

$f(Y/X)=f(Y,X)/f_X(X)$ función de densidad de probabilidad condicional.

$f_X(X)=\int_{-\infty}^{+\infty} f(Y/X)dY$ función marginal de probabilidad de x.

$f(X,Y)$ función de densidad de probabilidad conjunta.

Para ilustrar este caso considérese el siguiente ejemplo: Se desean pronosticar las ganancias, Y, de una empresa en función del precio del producto que vende, X, si la función de densidad de probabilidad conjunta está dada por:

$$f(Y,X)=\begin{cases}3X^2 \text{ si } 0<X<Y<1\\ 0 \text{ cualquier otro valor}\end{cases}$$

Para pronosticar las ganancias es necesario obtener la curva de Regresión de Y sobre X por lo que el primer paso es obtener $fx(X)$:

$$fx(X)=\int_Y f(Y,X)dY=\int_X^1 3X^2dY=3X^2Y\Big|_X^1$$

$$fx(X)=3X^2(1-X)$$

Entonces:

$$f(X/Y)=f(X,Y)/fx(X)=\frac{3X^2}{3X^2(1-X)}=\frac{1}{1-X}$$

La curva de Regresión está dada por:

$$E(Y/X)=\int_X^1 \frac{Y}{1-X}dY=\frac{1}{1-X}\frac{Y^2}{2}\Big|_X^1$$

$$E(Y/X)=\frac{1}{1-X}\left(\frac{1}{2}-\frac{X^2}{2}\right)=\frac{1-X^2}{2(1-X)}=\frac{(1-X)(1+X)}{2(1-X)}$$

$$E(Y/X)=\frac{1}{2}+\frac{X}{2}$$

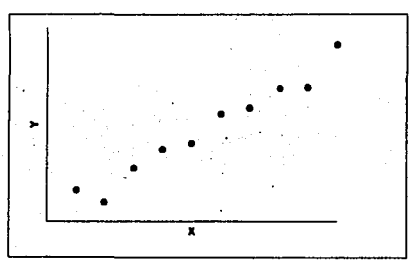
Por lo tanto las ganancias de la empresa en función del precio del producto que vende se pueden pronosticar con una recta con pendiente e intersección igual a 1/2.

• Regresión basada en datos observados

Cuando no es posible conocer la función conjunta de probabilidad, una forma práctica de obtener la curva de regresión, E(Y/X), es a partir de una serie de observaciones del fenómeno de interés.

A partir de las observaciones obtenidas se puede asumir la forma funcional para la curva de regresión y tratar de ajustar ésta a los datos. Y se considera como una variable aleatoria cuyos valores se observan mediante la selección de los valores de las variables predictoras, por lo que éstas no se consideran como variables aleatorias sino como un conjunto de valores fijos [Canavos (1988, p. 446)].

En el caso simple una forma sencilla para seleccionar la curva de regresión es mediante una gráfica de la variable de respuesta contra la variable predictora (gráfica de dispersión) con lo que se sabrá si la curva de regresión es una recta, una función cuadrática, etc. Por ejemplo, considérese la siguiente gráfica:



Es evidente que estos datos tienen una tendencia lineal pero un modelo determinístico de la forma $Y = \beta_0 + \beta_1 X$ no representa con exactitud la relación entre X y Y. Una mejor opción es utilizar un modelo probabilístico $Y = \beta_0 + \beta_1 X + \varepsilon$. A ε se le conoce como error aleatorio, éste representa la variabilidad de Y y se asume que $E(\varepsilon)=0$ por lo que al obtener la media condicional del modelo se obtiene:

$$E(Y/X) = \beta_0 + \beta_1 X$$

En general para todos los modelos de regresión se asume que E(ε)=0 por lo que al obtener la media condicional de Y dado X , E(Y/X), el término del error desaparece.

Después de seleccionarse el modelo, el siguiente paso es obtener los estimadores de los parámetros de la función. Los principales métodos de estimación son: mínimos cuadrados y máxima verosimilitud (de estos métodos se dará una breve descripción más adelante).

EJEMPLOS DE MODELOS DE REGRESION

- *REGRESION LINEAL*

En este modelo se asume que se puede ajustar una línea recta a los datos. Este modelo tiene la forma:

$$E(Y/X) = \beta_0 + \beta_1 X$$

donde β_0 y β_1 son los parámetros que deben ser estimados. El modelo de Regresión Lineal es uno de los más utilizados ya que un gran número de situaciones en la práctica pueden ser descritas por este modelo, de ahí su importancia.

- *REGRESION LOGARITMICA*

En este modelo se asume que se puede ajustar una función logarítmica a los datos. Este modelo tiene la forma:

$$E(Y/X) = \beta_0 + \beta_1 Ln(X)$$

donde β_0 y β_1 son los parámetros que deben ser estimados.

- *REGRESION EXPONENCIAL*

En este modelo se asume que la curva de regresión es una función exponencial. Este modelo tiene la forma:

$$E(Y/X) = \beta_0 \exp(\beta_1 X)$$

donde β_0 y β_1 son los parámetros que deben ser estimados.

• *REGRESION DE POTENCIA*

En este modelo se asume que se puede ajustar una función de potencia de la variable X a los datos. Este modelo tiene la forma:

$$E(Y/X) = \beta_0 X^{\beta_1}$$

donde β_0 y β_1 son los parámetros que deben ser estimados.

Como puede verse, existe una amplia gama de modelos, aquí sólo se han presentado algunos de ellos. El modelo que se utilizará para esta investigación es el modelo de *Regresión Logística* que se describe a continuación.

2.2. DEFINICION DEL MODELO LOGISTICO

La característica principal de los modelos de regresión es que éstos describen la relación entre una variable de respuesta o variable dependiente y una o más variables predictoras o variables independientes.

Entre las situaciones reales pueden presentarse diversos tipos de variables, éstas pueden ser : *continuas* (toman cualquier valor real, por ejemplo edad, estatura, peso), *categóricas* que sólo admiten ciertos valores. Las variables categóricas se dividen en *dicotómicas* (solo dos valores, por ejemplo varón o mujer), *politómicas no ordenadas* (como profesar la religión católica, judía, protestante, etc.) y *politómicas ordenadas* (bachillerato, licenciatura, doctorado,etc) [Guillén (1992, p. 8)].

En el caso donde las variables, de respuesta y predictoras, son continuas lo más común es utilizar un modelo de regresión lineal para encontrar la relación entre ellas. Si la variable de respuesta es del tipo categórico, en especial dicotómica (el fenómeno ocurre, el fenómeno no ocurre) , lo más recomendable es usar un modelo de Regresión Logística el

cual proporciona una probabilidad estimada de que el fenómeno bajo estudio ocurra (El modelo Logístico puede generalizarse para el caso politómico, sin embargo durante este trabajo sólo se tratará con el caso dicotómico). Este modelo tiene el mismo propósito que otras técnicas estadísticas similares: encontrar el modelo que tenga el mejor ajuste de la relación que existe entre la variable de respuesta y las predictoras y que cumpla con el principio de parsimonia, esto es, un modelo sencillo y de pocos parámetros. El modelo de Regresión Logística tiene la característica de que puede modelar variables predictoras tanto continuas como categóricas [Hosmer (1989, p. 1)].

El análisis de variables de respuesta categóricas es uno de los campos menos recurridos de la Estadística pero se trata de uno de los campos de mayor interés y promesa para el futuro. El escaso uso de este análisis se debe en gran parte a que hasta hace poco no se disponía de modelos adecuados que manejaran este tipo de variables de respuesta y que el cálculo manual era complejo y el electrónico mediante computadoras era costoso, por lo cual los investigadores solamente podían dedicarse al estudio de problemas que tenían a lo más dos o tres variables y solo podían obtener algunos estadísticos simples como la chi-cuadrada. Así, desde que se generalizó el uso de las computadoras se hizo más frecuente el uso de la Regresión Logística [Guillén (1992, p. 65)].

2.2.1 MODELO DE REGRESION LOGISTICA SIMPLE

Si se quisiera obtener la curva de regresión de un fenómeno se pensaría en utilizar el modelo más simple, el modelo lineal, de la forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

donde ε es una variable aleatoria, conocida como error aleatorio, que se asume que sigue una distribución $N(0,\sigma^2)$, con σ^2 finita y constante. Esta variable representa la influencia de todos aquellos factores que no se consideran en el modelo. La variable de respuesta Y es una variable aleatoria cuyos valores se observan mediante la selección de los valores de la variable de predicción X la cual se considera una variable matemática, pero no aleatoria.

En general, en cualquier problema de regresión, el resultado de mayor importancia es el valor medio de la variable de respuesta Y dado el valor de la variable predictora X. Esta cantidad es llamada la media condicional E(Y/X) y se lee "El valor esperado de Y dado el

valor X " [Hosmer, (1989, p. 5)]. Por lo tanto, si se obtiene la esperanza de la ecuación anterior se tiene:

$E(Y/X) = E(\beta_0) + E(\beta_1 X) + E(\varepsilon)$ como β_0 , β_1 , X son constantes, entonces:

$$E(Y/X) = \beta_0 + \beta_1 X \qquad [2.0]$$

Como se ve en la ecuación [2.0], ésta puede tomar cualquier valor en el rango $(-\infty,+\infty)$ dependiendo del rango de X.

Cuando la variable de respuesta es dicotómica E(Y/X) debe estar entre 0 y 1 por lo que la función dada en [2.0] no es útil y se debe buscar una expresión para la media condicional que cumpla con la restricción anterior.

Supóngase que se tiene la función lineal $g(X)$ definida como:

$$g(X) = \beta_0 + \beta_1 X$$

El objetivo es realizar alguna transformación sobre $g(X)$ para que ésta cumpla con la restricción de que se encuentre entre 0 y 1. Esta transformación se describe a continuación:

$g(X) = \ln \exp(\beta_0 + \beta_1 X) = \ln [\exp(\beta_0 + \beta_1 X)]$

$$= \ln\left[\frac{\exp(\beta_0+\beta_1 X)[1+\exp(\beta_0+\beta_1 X)]}{1+\exp(\beta_0+\beta_1 X)}\right]$$

$$= \ln\left[\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right] - \ln\left[\frac{1}{1+\exp(\beta_0+\beta_1 X)}\right]$$

$$= \ln\left[\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right] - \ln\left[\frac{1+\exp(\beta_0+\beta_1 X)-\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right]$$

$$= \ln\left[\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right] - \ln\left[\frac{1+\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)} - \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right]$$

$$= \ln\left[\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right] - \ln\left[1 - \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right]$$

$$= \ln\left\{\frac{\left[\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right]}{\left[1 - \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\right]}\right\}$$

Si se hace :

$$\pi(X) = \frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)} \qquad [2.1]$$

Entonces se tiene:

$$g(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] \qquad [2.2]$$

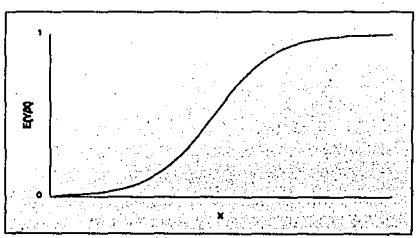
Entonces despejando $\pi(X)$ de [2.2] se llega a:

$$\pi(X) = \frac{exp[g(X)]}{1+exp[g(X)]}$$

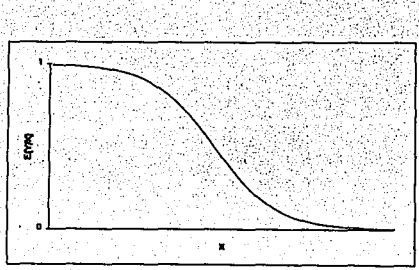
Donde los coeficientes de $g(X)$, β_0 y β_1, son los parámetros a estimar de la función $\pi(X)$ cuando se realiza la regresión. $-\infty < \beta_0, \beta_1 < +\infty$.

La función $\pi(X)$ es conocida como la función de distribución logística y tiene las siguientes propiedades:

Al ser una función de distribución cumple con la restricción de proporcionar valores entre 0 y 1. La función $\pi(X)$ describe una gráfica monótona creciente [Gráfica 2.1 a)] o decreciente [Gráfica 2.1 b)] dependiendo del signo de β_1. Es asintótica a 0 y 1, es decir se aproxima gradualmente a 0 y a 1 en los extremos del rango de X, por lo que está acotada en este intervalo. El coeficiente β_1 indica la rapidez con que la gráfica tiende ya sea a cero o a uno. En investigaciones sobre función $\pi(X)$ se ha demostrado que es aproximadamente lineal en el intervalo [0.2,0.8] [Neter (1989, p. 582)].



Gráfica 2.1 a)



Gráfica 2.1 b)

De acuerdo a lo anterior se puede expresar $E(Y/X) = \pi(X)$ cuando la variable de respuesta es dicotómica y por lo tanto el modelo de Regresión Logística está dado por la ecuación

[2.1]. Al ser $\pi(X)$ la esperanza condicional de Y dado X, esta función proporciona la probabilidad condicional de que Y=1 dado un valor particular de X.

Una característica importante es que la función $\pi(X)$ es fácilmente linealizada mediante la función $g(X)$ en la ecuación [2.2] la cual es conocida como la *transformación logit*. La función $g(X)$ representa, en una escala logarítmica, la diferencia entre las probabilidades de pertenecer a ambas clases (El fenómeno ocurre y el fenómeno no ocurre), debido a que por propiedad de logaritmo se tiene que $g(X) = \ln[\pi(X)] - \ln[1 - \pi(X)]$, [Peña (1987, p. 455)]. $g(X)$ tiene muchas de las propiedades deseables en un modelo de Regresión Lineal: el logit es lineal en sus parámetros, puede ser continua y estar en el rango de [-∞ , +∞] dependiendo del rango de X.

2.2.2 MODELO DE REGRESION LOGISTICA MULTIPLE

Si se tiene una colección de p variables predictoras denotadas por el vector $X' = (X_1, X_2, ..., X_p)$, entonces es necesario utilizar un modelo de Regresión Logística múltiple cuyo logit está definido por la siguiente ecuación:

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \qquad [2.3]$$

Entonces:

$$\pi(X) = \frac{exp[g(X)]}{1 + exp[g(X)]} \qquad [2.4]$$

El modelo de Regresión Logística múltiple tiene las mismas propiedades que el simple: la función $\pi(X)$ es monótona, de forma sigmoidal (forma de S) y es casi lineal cuando está entre [0.2,0.8].

En cuanto al tipo de las variables predictoras, si algunas de éstas son categóricas es necesario hacerles algunas transformaciones ya que los números usados para representar los diferentes niveles de la variable son solo identificadores de estos y no tienen un significado numérico lo cual puede causar problemas al realizar el ajuste. Por la razón anterior es recomendable descomponer dicha variable en varias variables dicotómicas conocidas como

variables de diseño o *variables mudas*, cada una de éstas representa la presencia o ausencia de un determinado nivel, los cuales son mutuamente excluyentes, tomando en cuenta que el primer nivel siempre se codificará como la ausencia de todos los restantes niveles. Para describir estas variables considérese el siguiente ejemplo:

Supóngase que se analiza la variable estado civil, la cual tiene cinco niveles: soltero, casado, divorciado, viudo y unión libre. En este caso se necesitan 4 variables de diseño para representar esta variable categórica de 5 niveles. La codificación se da en el cuadro 2.1.

	VARIABLES DE DISEÑO			
ESTADO CIVIL	D_1	D_2	D_3	D_4
Soltero	0	0	0	0
Casado	1	0	0	0
Divorciado	0	1	0	0
Viudo	0	0	1	0
Unión Libre	0	0	0	1

Cuadro 2.1

Por lo que respecta al cuadro 2.1 se pueden hacer dos observaciones:

1. El primer nivel se representa con el valor de cero en todas las variables de diseño como se mencionó anteriormente (a este nivel se le conoce como nivel basal).

2. Los niveles restantes están representados por un 1 en una de las variables de diseño y ceros en los demás lo que representa la presencia del nivel correspondiente.

De aquí se puede establecer que si la variable predictora categórica tiene k niveles se necesitan k-1 variables de diseño para representarla.

Así, si la j-ésima variable predictora, X_j, tiene k_j niveles y se usan k_j -1 variables de diseño denotadas como D_{ju} con coeficientes β_{ju}, u=1,2...,k -1 entonces el logit para el modelo con p variables y la j-ésima variable categórica será:

$$g(X) = \beta_0 + \beta_1 X_1 + ... + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + ... + \beta_p X_p \qquad [2.5]$$

2.2.3 CASOS QUE REQUIEREN EL USO DE LA REGRESION LOGISTICA

Como ya se ha mencionado en diversas ocasiones, la característica principal de los casos donde se utiliza este modelo son aquellos en los que la variable de respuesta sólo tiene dos resultados posibles: 1,0 (El fenómeno ocurre, el fenómeno no ocurre). El modelo de Regresión Logística proporciona una probabilidad estimada de la ocurrencia del fenómeno dado un valor específico de las variables predictoras. Para determinar la ocurrencia o no del fenómeno a partir de la probabilidad estimada se necesita de alguna regla para determinar a partir de qué valor de probabilidad puede considerarse que el fenómeno ocurre (estas reglas se darán en el capítulo VII).

Para comprender mejor la aplicación del modelo se pueden mencionar los siguientes casos:

Supóngase que se realiza un estudio para saber si un estudiante abandona o no su carrera. Se puede calcular la probabilidad de que esto ocurra usando Regresión Logística. Este modelo estará en función, por ejemplo, del estado laboral del estudiante, posición económica, estado civil, materias no aprobadas y vocación. La variable de respuesta Y se define como: 1, el estudiante abandona su carrera; 0, el estudiante no abandona su carrera.

Una aplicación útil de la Regresión Logística en ingeniería puede ser ésta: calcular el riesgo de que un edificio se derrumbe ante un temblor; este fenómeno puede estar en función de la situación geográfica del edificio, antigüedad, tipo de cimientos, altura, magnitud del sismo y duración de éste. La variable de respuesta Y se puede definir como sigue: 1, la casa se derrumba ante el temblor; 0, la casa no se derrumba ante el temblor.

Uno de los campos donde la Regresión Logística ha tenido mayor aplicación es en la medicina y algunos casos de estudio donde se puede utilizar este modelo son los siguientes:

En el cálculo de la probabilidad de muerte en quemados críticos se puede encontrar un modelo de Regresión Logística el cual esté en función de la edad, mecanismo de quemadura, síndrome de inhalación, superficie corporal quemada. La variable de respuesta Y estará definida como: 1, el quemado muere ; 0, el quemado no muere.

Otro caso de estudio es el pronosticar si un niño nace con peso bajo como consecuencia de la edad de la madre, tipo de alimentación, si fuma o no durante el embarazo, historia de partos prematuros, etc. Esta relación se puede encontrar utilizando un modelo de Regresión Logística donde la variable de respuesta Y está definida como: 1, el niño nace con peso bajo; 0, el niño no nace con peso bajo.

Un estudio en ginecoobstetricia de gran importancia es el siguiente: calcular la probabilidad de que un niño nazca mediante una cesárea. El modelo de Regresión Logística tendrá como variables predictoras a: el tamaño de la pelvis de la madre, edad, peso y estatura, tamaño de la cabeza del niño, posición de éste en el vientre materno, etc. Así la variable de respuesta Y está definida como: 1, el niño nace por cesárea; 0, el niño no nace por cesárea (parto natural). Este caso es de gran importancia ya que es el estudio que se realizará en esta tesis.

De este modo se podría seguir mencionando diversos casos de estudio en los que la Regresión Logística puede aplicarse, los anteriores solamente se dan para tener un conocimiento de los alcances que este modelo tiene.

2.3. REGRESION LOGISTICA VS. REGRESION LINEAL

Comúnmente, cuando se desea hacer pronósticos u otro tipo de análisis de una serie de datos, donde una variable de respuesta tiene relación con un conjunto de variables predictoras, se hace la suposición que la relación entre ellas es lineal. Del mismo modo se puede pensar que cuando se tiene un fenómeno donde la variable de respuesta es

dicotómica se podría modelar por medio de la Regresión Lineal; sin embargo, esto no es recomendable. A continuación se dan razones para esta afirmación.

La principal diferencia entre un modelo de Regresión Logística y un modelo de Regresión Lineal es que la variable de respuesta en Regresión Logística es binaria o dicotómica y en Regresión Lineal está definida en un intervalo de [-∞ , +∞]. Esta diferencia entre Regresión Logística y Lineal se refleja en ambas en la elección de un modelo y los supuestos establecidos con respecto al error aleatorio [Hosmer (1989,p. 1)].

Cuando en un fenómeno se presenta una variable de respuesta binaria y se busca encontrar la relación entre ella y una serie de variables explicativas el primer intento para analizarlo es utilizar un modelo de Regresión Lineal de la forma:

$$Y = X\beta + \varepsilon$$

Donde $Y' = (Y_1, Y_2, \ldots, Y_n)$ vector de observaciones de Y

$\beta' = (\beta_o, \beta_1, \ldots, \beta_p)$ vector de parámetros.

$\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ errores.

$$X = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix}$$

n = número de observaciones

p = número de variables

Considérese el modelo lineal simple (lo siguiente pueden extenderse para el modelo múltiple):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad [2.6]$$

Uno de los supuestos en Regresión Lineal es que $\varepsilon \sim N(0,\sigma^2)$, donde σ^2 es la varianza la cual es constante y finita.

Ya que $E(\varepsilon_i)=0$, obteniendo la esperanza de la ecuación [2.6] se tiene:

$$E(Y_i) = E(\beta_0) + E(\beta_1 X_i) + E(\varepsilon_i)$$

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad [2.7]$$

Ya que Y_i es una variable aleatoria Bernoulli que sólo toma los valores de 0 ó 1, su distribución de probabilidad es la siguiente:

$$P(Y_i = 1) = p_i \ , \quad P(Y_i = 0) = 1 - p_i \qquad [2.8]$$

Sacando esperanza de Y_i utilizando las probabilidades de la ecuación [2.8] se obtiene:

$$E(Y_i) = 1(p_i) + 0(1 - p_i) = p_i \qquad [2.9]$$

Usando las ecuaciones [2.7] y [2.9] se tiene:

$$p_i = \beta_0 + \beta_1 X_i \qquad [2.10]$$

La ecuación [2.10] indica la probabilidad de que $Y_i = 1$ cuando el nivel de la variable predictora es X_i. Al usar este modelo surgen dos problemas:

1. Por las ecuaciones [2.8.] y [2.9], E(Y/X) debe cumplir $0 < E(Y_i) < 1$. Pero se ve en el lado derecho de la ecuación [2.10] que no necesariamente se cumple esta restricción debido a que X puede ser una variable continua en el intervalo $(-\infty, +\infty)$ y en este caso p_i estará en el mismo intervalo.

2. Conocido el valor de X_i, los únicos valores de Y_i son 0 ó 1. Por lo tanto de las ecuaciones [2.6], [2.8] y [2.10] se sigue que ε tiene distribución Bernoulli con valores:

$\varepsilon_i = 1 - \beta_0 - \beta_1 X_i = 1 - p_i$ con probabilidad p_i , cuando $Y_i = 1$

$\varepsilon_i = -\beta_0 - \beta_1 X_i = -p_i$ con probabilidad $1 - p_i$, cuando $Y_i = 0$

$E(\varepsilon_i) = p_i (1 - p_i) + (1 - p_i)(-p_i) = 0$

Lo anterior indica que ε_i tiene media cero. Si se obtiene la varianza de ε_i se tiene:

$$\text{var}(\varepsilon_i) = E(\varepsilon_i^2) - E(\varepsilon_i)^2 = \sum_{\varepsilon} \varepsilon_i^2 p(\varepsilon_i) - 0$$

$$\text{var}(\varepsilon_i) = (1-p_i)^2 p_i + (1-p_i) p_i^2 = (1-p_i) p_i$$

De la ecuación [2.10] se puede observar que p_i depende del valor de X por lo que la varianza de ε_i no es constante, por lo tanto éste es un modelo heterocedástico (varianza no constante) lo cual no es deseable en un modelo de regresión lineal y no cumple con los supuestos de que $\varepsilon \sim N(0,\sigma^2)$, con σ^2 constante y finita.

Estos dos problemas indican que no es recomendable usar un modelo de Regresión Lineal cuando la variable de respuesta es dicotómica, por lo tanto es indispensable usar otro tipo de modelo. El modelo más usado para estos casos es el de Regresión Logística, el cual tiene propiedades importantes [Hosmer (1989, p. 6)]:

a) Desde un punto de vista matemático es extremadamente flexible y fácil en su manejo debido a que sólo se hace uso de una función exponencial para evaluarla y es fácilmente linealizado mediante la transformación logit.

b) Da en sí mismo una interpretación biológica significativa, ya que el comportamiento de muchas poblaciones puede ser descrita mediante esta función.

c) Ya que $\pi(X)$ es una función de distribución de probabilidad, cumple con la restricción $0<\pi(X)<1$. Esto es debido a la siguiente razón, utilizando la ecuación [2.4] y ya que $g(X)$ se encuentra en el intervalo de $(-\infty,+\infty)$ se tiene que:

$$\text{Si } g(X) \to \infty \text{ entonces } \pi(X) \to 1$$

$$\text{Si } g(X) \to -\infty \text{ entonces } \pi(X) \to 0$$

por lo tanto $0 < \pi(X) < 1$.

2.4. ESTIMACION POR MAXIMA VEROSIMILITUD

El método más usual en regresión lineal para la estimación de los parámetros es el de mínimos cuadrados cuyo objetivo es seleccionar estimadores $\hat{\beta}$ tal que los errores de predicción, $Y-\hat{Y}$, sean mínimos. Este método proporciona estimadores insesgados [$E(\hat{\beta})=\beta$] y eficientes (varianza mínima). El procedimiento es el siguiente: Crear la función $S=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$, obtener las derivadas parciales con respecto a cada uno de los parámetros β , igualar a cero estas ecuaciones y resolver el sistema. Si se intenta utilizar este método cuando la variable de respuesta es categórica se obtienen estimadores que son insesgados pero no eficientes debido a que, como se mostró en el apartado anterior, las varianzas de los errores no son constantes por lo que las estimaciones de las varianzas muestrales no serán correctas y las pruebas estadísticas para efectuar inferencias no serán válidas [Guillén (1992, p.67)].

El propósito principal del método de máxima verosimilitud es encontrar un vector $\hat{\beta}$ que haga máxima la probabilidad de que los valores estimados, $\hat{Y}$, a partir del vector $\hat{\beta}$ y de la matriz de observaciones de las variables predictoras, X, sean lo más parecidas posible a los valores observados de Y.

El método de máxima verosimilitud consiste en maximizar el log de la función de verosimilitud (L) que es la función de probabilidad o función de densidad conjunta. Los estimadores que se obtienen para este método pueden ser sesgados pero generalmente son consistentes (cuando la muestra de tamaño n se incrementa disminuye la varianza), eficientes y suficientes (es aquel que utiliza la mayor cantidad posible de información contenida en la muestra y no depende del parámetro). A continuación se aplicará este método para obtener los estimadores $\hat{\beta}$ del modelo logístico:

Sea una muestra de n elementos caracterizada por un vector de variables $X'=(1,X_1,X_2,...,X_p)$ con coeficientes $\beta'=(\beta_o,\beta_1,...,\beta_p)$ y una variable binaria Y con distribución Bernoulli, entonces:

$$p(Y_i)=\pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad Y_i=0,1 \qquad \pi_i=\frac{exp[g(X)]}{1+exp[g(X)]}$$

La función de verosimilitud es:

$$L=\prod_{i=1}^{n}\pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad [2.11]$$

Obteniendo logaritmo natural:

$$\ln L=\sum_{i=1}^{n}Y_i\ln\left(\frac{\pi_i}{1-\pi_i}\right)+\sum_{i=1}^{n}\ln(1-\pi_i) \qquad [2.12a]$$

$$=\sum_{i=1}^{n}Y_i g_i(X)+\sum_{i=1}^{n}\ln(1-\pi_i)$$

$$\ln L=\sum_{i=1}^{n}Y_i X_i'\beta-\sum_{i=1}^{n}\ln(1+exp(X_i'\beta)) \qquad [2.12b]$$

Obteniendo primeras derivadas parciales a la ecuación [2.12b] e igualando a cero:

$$\frac{\partial \ln L(\beta)}{\partial \beta}=\sum_{i=1}^{n}Y_i X_i-\sum_{i=1}^{n}X_i\left[\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right]=0 \qquad [2.13]$$

Entonces:

$$\sum_{i=1}^{n}Y_i X_i=\sum_{i=1}^{n}X_i\left[\frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}\right]=\sum_{i=1}^{n}\hat{Y}_i X_i \qquad [2.14]$$

Como se puede observar el sistema de ecuaciones [2.14] es no lineal en los parámetros β. Lo recomendable en esta situación es obtener el máximo de [2.12] mediante un método numérico como es el algoritmo de Newton-Raphson. Este método indica que si se quiere obtener el óptimo de una función $f(\theta)$ con primeras y segundas derivadas parciales continuas, cerca del óptimo la función debe ser aproximadamente cuadrática de la forma:

$$f(\theta_{i+1}) \approx f(\theta_i) + G(\theta_i)(\theta_{i+1} - \theta_i) + \frac{1}{2}(\theta_{i+1} - \theta_i)' H(\theta_i)(\theta_{i+1} - \theta_i)$$

Donde: $G(\theta) = \dfrac{\partial f(\theta)}{\partial \theta_i}$ Vector gradiente

$$H(\theta) = \frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j}$$ Matriz Hessiana

Entonces desarrollando [2.13] alrededor de un punto β_o se obtiene:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \frac{\partial \ln L(\beta_o)}{\partial \beta} + \frac{\partial^2 \ln L(\beta_o)}{\partial \beta \partial \beta'}(\beta - \beta_o)$$

Como se busca un máximo se debe cumplir $\dfrac{\partial \ln L(\beta_o)}{\partial \beta} = 0$

Por lo tanto:

$$\beta_o = \beta + \left[-\frac{\partial^2 \ln L(\beta_o)}{\partial \beta \partial \beta'}\right]^{-1}\left[\frac{\partial \ln L(\beta)}{\partial \beta}\right] \qquad [2.15]$$

La ecuación anterior expresa como obtener el punto β_o del máximo a partir de un punto próximo cualquiera β.

La matriz de segundas derivadas de la ecuación anterior puede ser considerada en el caso óptimo como la inversa de la matriz de varianzas y covarianzas de los estimadores de máxima verosimilitud.

Obteniendo la segunda derivada de ln L(β) se obtiene:

$$\hat{M}^{-1} = \left[-\frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta'}\right] = \sum_{i=1}^{n} X_i X_i' \omega_i \qquad [2.16]$$

Donde $\omega_i = \dfrac{exp(-X_i'\beta)}{\left[1+exp(-X_i'\beta)\right]^2} = \dfrac{exp(X_i'\beta)}{\left[1+exp(X_i'\beta)\right]^2} = \pi_i(1-\pi_i)$

Sustituyendo las ecuaciones [2.13] y [2.16] en [2.15] se tiene:

$$\beta_a = \beta + \left[\sum_{i=1}^{n} X_i X_i' \omega_i\right]^{-1}\left[\sum_{i=1}^{n} X_i (Y_i - \pi_i)\right]$$

O en forma matricial:

$$\beta_a = \beta + (X'VX)^{-1} X'(Y - \hat{Y}) \qquad [2.17]$$

Donde V es una matriz diagonal de términos $\pi_i (1-\pi_i)$ y $\hat{Y}$ es el vector de valores esperados de Y.

A partir de la ecuación [2.17] se obtiene un algoritmo iterativo para obtener los estimadores de máxima verosimilitud de β [Peña (1987, p.p. 455-458)]. Este algoritmo se resume en los siguientes pasos:

(1) Fijar un valor arbitrario inicial, $\hat{\beta}_1$ (generalmente 0), y obtener el vector $\hat{Y}_1$ para dicho valor en el modelo.

(2) Definir una variable auxiliar:

$$Z_i = \frac{Y_i - \hat{Y}_i}{\hat{Y}_i (1 - \hat{Y}_i)} = \frac{Y_i - \hat{\pi}_i}{\hat{\pi}_i (1 - \hat{\pi}_i)}$$

O en forma de vector:

$$Z = \hat{V}^{-1} (Y - \hat{Y})$$

(3) Estimar por el método de mínimos cuadrados ponderados una regresión de variable de respuesta Z, variables predictoras X y coeficientes de ponderación $\hat{Y}_i (1 - \hat{Y}_i)$ lo cual produce estimadores $\hat{b}_1$ dados por:

$$\hat{b}_1 = (X'\hat{V}X)^{-1}X'\hat{V}Z = (X'\hat{V}X)^{-1}X'(Y-\hat{Y})$$

(4) Obtener un nuevo estimador de los parámetros $\hat{\beta}$ del modelo logístico mediante:

$$\hat{\beta}_2 = \hat{\beta}_1 + \hat{b}_1$$

(5) Tomar el valor estimado en la etapa anterior, $\hat{\beta}_h$ y usarlo para calcular π_i y así obtener el vector de estimadores $\hat{Y}(\hat{\beta}_h) = \hat{Y}_h$ y utilizando este valor construir la matriz $\hat{}_h$ y la nueva variable Z_h :

$$Z_h = \hat{V}_h^{-1}(Y-\hat{Y}_h)$$

El nuevo valor de $\hat{\beta}_{h+1}$ será:

$$\hat{\beta}_{h+1} = \hat{\beta}_h + (X'\hat{V}_h X)^{-1}X'(Y-\hat{Y}_h)$$

Donde el término de ajuste se calcula regresando por mínimos cuadrados ponderados, Z_h sobre X con ponderaciones $\hat{}_h$.

El proceso se repite hasta obtener convergencia, $\hat{\beta}_{h+1} \approx \hat{\beta}_h$.

Como se puede ver, el proceso es excesivamente laborioso; por esta razón lo más conveniente es utilizar un paquete estadístico que minimice el tiempo de cálculo. La discusión sobre este software se hará en el siguiente capítulo.

2.5. INFERENCIAS SOBRE LOS COEFICIENTES DEL MODELO

Con lo que respecta a la inferencia sobre parámetros, ésta se realiza con el propósito de conocer si una variable en específico es significativa para el modelo o no lo es.

Una vez estimados los parámetros es necesario realizar pruebas de significancia de los mismos. Algunas propiedades de los coeficientes son las siguientes:

1) El signo positivo o negativo de $\hat{\beta}_i$ indica si la probabilidad de que ocurra el hecho descrito por la variable de respuesta aumenta o disminuye ante un aumento o disminución unitaria en la variable predictora.

2) Cuando mayor es el valor absoluto de $\hat{\beta}_i$ más importante es el efecto de la variable predictora correspondiente en el modelo.

Para probar la significancia de un modelo es necesario hacer una comparación entre los valores observados y los pronosticados de la variable de respuesta. Estas comparaciones se hacen en base a la función log verosimilitud definida en la ecuación [2.12a]. Para comprender mejor esta comparación es útil si se piensa que un valor observado de la variable de respuesta es un valor pronosticado a partir de un modelo que tiene tantos parámetros como observaciones hay en la muestra (modelo saturado).

Para conocer la significancia del modelo es necesario utilizar algunos estadísticos, uno de ellos es la Devianza la cual realiza una comparación entre el modelo que se prueba y el modelo que proporciona los pronósticos correctos. La Devianza está definida para un modelo simple como:

$$D = -2\ln\left[\frac{\text{verosimilitud del modelo actual}}{\text{verosimilitud del modelo saturado}}\right]$$

$$D = -2\left[\ln\left(\begin{matrix}\text{verosimilitud del}\\ \text{modelo actual}\end{matrix}\right) - \ln\left(\begin{matrix}\text{verosimilitud del}\\ \text{modelo saturado}\end{matrix}\right)\right]$$

$$D = -2\left\{\sum_{i=1}^{n}\left[Y_i \ln(\hat{\pi}_i) + (1-Y_i)\ln(1-\hat{\pi}_i)\right] - \sum_{i=1}^{n}\left[Y_i \ln(Y_i) + (1-Y_i)\ln(1-Y_i)\right]\right\}$$

$$D = -2\sum_{i=1}^{n}\left[Y_i \ln\left(\frac{\hat{\pi}_i}{Y_i}\right) + (1-Y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-Y_i}\right)\right] \qquad [2.18]$$

Para propósito de evaluar la significancia de una variable predictora hay que comparar el valor de D con y sin la variable predictora en la ecuación. Esto da origen a un nuevo estadístico, G definido como:

$$G = D(\text{modelo sin la variable}) - D(\text{modelo con la variable})$$

$$G = -2\ln\left(\frac{\text{verosimilitud sin la variable}}{\text{verosimilitud con la variable}}\right) \qquad [2.19]$$

Para el caso simple, cuando la variable predictora no está en el modelo, el estimador para β_0 es ln (n_1 / n_0) donde:

$$n_1 = \sum_{i=1}^{n} Y_i \qquad ; \qquad n_0 = \sum_{i=1}^{n} (1 - Y_i)$$

y los valores pronosticados de Y son constantes e iguales n_1 / n_0. En este caso el valor de G es :

$$G = 2\left\{\sum_{i=1}^{n}\left[Y_i \ln(\hat{\pi}_i) + (1 - Y_i)\ln(1 - \hat{\pi}_i)\right] - \left[n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)\right]\right\} \qquad [2.20]$$

Bajo la hipótesis de que $\beta_1 = 0$, el estadístico G tiene una distribución chi-cuadrada con un grado de libertad.

En el caso simple, podemos primero ajustar un modelo que tiene solamente el término constante, después se ajusta un nuevo modelo que contiene una variable predictora y una constante con lo cual se obtienen dos valores para ln L, ln L_1 y ln L_2 respectivamente.

La prueba de razón de verosimilitud se realiza con el valor de G definida por:

$G = -2 [\ln L_1 - \ln L_2]$

Después de obtener G, éste se utiliza para calcular:

valor-p = $P[\chi^2 (1) > G]$

lo cual ayuda a saber si el valor $\hat{\beta}_1$ es significativo o no, recordando que el valor-p es el mínimo nivel de significación, α, para el cual los datos observados indican que se tendrían que rechazar la hipótesis nula [Mendenhall (1986, p. 400)]. Si $\alpha \geq$ valor-p se rechaza Ho ($\beta_1 = 0$); en caso contrario, no se puede rechazar Ho por lo cual valores pequeños de valor-p nos garantizan la significancia de la variable.

Otro de los estadísticos para probar la significancia de los estimadores es W, que se usa para la llamada prueba de Wald. Este estadístico es utilizado generalmente para probar la significancia de una variable en particular. La prueba de Wald trabaja bajo las siguientes hipótesis:

$$\begin{aligned} &\text{Ho: } \beta_1 = 0 \\ &\text{Ha: } \beta_1 \neq 0 \\ &= \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} \end{aligned} \qquad [2.21]$$

Donde la regla de decisión es:

$$\text{si } |W| \leq Z(1-\alpha/2) \text{ se acepta } H_0$$

$$\text{si } |W| > Z(1-\alpha/2) \text{ se acepta } H_a$$

Una regla práctica para esta prueba es aceptar Ha si $|W| > 2$.

El valor de S(β) se obtiene de la siguiente manera, sea la matriz $H = [h_{ij}] \quad i,j = 0,1,\ldots,p$ con elementos definidos por:

$$h_{ij} = \frac{\partial^2 \ln L(\beta)}{\partial \beta_i \, \partial \beta_j} \quad \text{entonces} \quad S^2(\beta) = \left[(-h_{ij})_{\beta = \hat{\beta}} \right]^{-1}$$

La prueba de razón de verosimilitud para probar la significancia de los p coeficientes de las variables predictoras para el caso multivariado es exactamente el mismo que cuando se tiene una sola variable predictora. Esta prueba se basa en el estadístico G dado en

las ecuaciones [2.18] y [2.19]. La única diferencia es que $\hat{\pi}$ tiene p+1 parámetros $\hat{\beta}$. Bajo la hipótesis nula de que los p coeficientes de las variables predictoras son ceros, la distribución de G es una Chi-cuadrada con p grados de libertad. Esta prueba indica si los p coeficientes de las variables predictoras son diferentes de cero.

Si el propósito es el de minimizar el número de variables también se puede aplicar la prueba de razón de verosimilitud. El valor de G es:

G = -2 [ln L(modelo reducido) - ln L(modelo completo)] [2.22]

Se obtiene el valor-p = P [$\chi^2(v) > G$]

donde v = son los grados de libertad que se calculan como la diferencia del número de variables entre el modelo completo y el modelo reducido.

Este procedimiento al igual que la estimación se realizará con el auxilio de paquetes estadísticos.

2.6. RESUMEN DE CARACTERISTICAS DEL MODELO DE REGRESION LOGISTICA

- Al igual que otros modelos de regresión se puede utilizar para la descripción de la naturaleza de la relación entre la variable de respuesta y las variables predictoras, así como realizar predicciones con él.

- El modelo Logístico es considerado como uno de los mejores modelos de regresión para el caso en que la variable de respuesta sea binaria y se puede emplear en una gran variedad de campos ya sea en las ciencias naturales o sociales.

- El modelo de Regresión Logística, para el caso en que la variable de respuesta es dicotómica, proporciona la probabilidad estimada de que el fenómeno bajo estudio ocurra.

- A diferencia de otros modelos similares es más fácil de calcular ya que sólo necesita del uso de la función exponencial.

- El modelo Logístico es fácilmente linealizado mediante la transformación logit.

- Este modelo acepta sin dificultades variables predictoras tanto categóricas como continuas.

- Si se tiene una variable predictora categórica de k niveles es necesario emplear k-1 variables de diseño para representarla en el modelo.

- La estimación de los parámetros es más compleja que en otros modelos de regresión ya que se debe resolver un sistema de ecuaciones no lineales y debido a esto es necesario utilizar un método iterativo; por esta razón es indispensable usar software especializado el cual es escaso.

- Una característica de suma importancia es que los coeficientes estimados del modelo miden el logaritmo de los odds ratio (razón de ventajas) para un cambio de unidad en la variable correspondiente. Así, los exponenciales de los coeficientes de regresión dan odds ratio estimados o riesgos relativos aproximados los cuales son de gran aplicación en estudios epidemiológicos principalmente (una descripción más detallada de los odds ratio se dará en el capítulo VI).

CAPITULO. III SOFTWARE AUXILIAR PARA REALIZAR REGRESION LOGISTICA

Como se explicó en los capítulos anteriores, la estimación de los parámetros y en general el cálculo de los estadísticos que se utilizan en Regresión Logística se hace tan laborioso que es necesario el uso de un paquete estadístico o un programa diseñado para este fin.

Se pueden encontrar algunos paquetes que en la actualidad manejan la Regresión Logística, entre ellos se tienen: BMDP (BioMeDical Program) y SAS (Statistical Analysis System) que son los que se utilizarán para realizar el ajuste del modelo, los cuales son herramientas poderosas para el análisis de datos y son de fácil manejo.

3.1. EL PROGRAMA BMDP (BIOMEDICAL PROGRAM)

El BMDP es un programa que consta de una serie de procedimientos para el análisis estadístico de datos que van desde un simple desplegado y descripción de estos hasta técnicas estadísticas avanzadas.

Los procedimientos incluidos en el BMDP realizan análisis de datos como son los siguientes:

- Listados, histogramas y estadísticas descriptivas.
- Pruebas t, análisis de varianza y pruebas no paramétricas.
- Diagramas de dispersión, correlaciones y regresión lineal.
- Tablas de frecuencia, análisis de correspondencia y modelos log-lineales.
- Regresión no lineal.
- Estimación por máxima verosimilitud.
- Análisis multivariado.
- Análisis de agrupamiento.
- Análisis de supervivencia.
- Análisis de series de tiempo.

Por lo que respecta a los datos, éstos deben incluirse en un archivo separado que puede crearse usando un procesador de textos, una base de datos o una hoja de cálculo, o bien incluirse los datos al final de las instrucciones del programa BMDP.

3.1.1 BLOQUES DEL PROGRAMA BMDP

Para realizar el análisis de un conjunto de datos en el paquete BMDP es necesario elaborar un programa o una serie de instrucciones que están compuestas por bloques y comandos. Un ejemplo de estas instrucciones pueden observarse en el siguiente problema:

PROBLEMA 3.1. Supóngase que se pretende obtener las estadísticas básicas (media, desviación estándar, coeficiente de variación, rango, etc) de las calificaciones de un conjunto de 13 alumnos que toman el mismo curso con tres diferentes profesores. Además de obtener las estadísticas para el conjunto de 13 estudiantes, también se desea obtenerlas para los alumnos de cada uno de los 3 profesores por separado. Para hacer este análisis se necesita hacer un programa y usar un procedimiento del BMDP como es el 1D que se encarga de obtener estadísticas simples. El programa se encuentra en el cuadro 3.1.

En estas instrucciones el bloque INPUT le dice al programa que hay 3 variables y que el formato de ellas es libre. El bloque VARIABLE indica el nombre de las variables y que los datos están agrupados por la variable Prof. El bloque GROUP proporciona los códigos y nombres de cada uno de los valores de las variables Prof y Turno. El bloque PRINT indica que se deben desplegar las estadísticas calculadas. El bloque END indica el fin de las instrucciones. Después del bloque END se encuentra la matriz de datos correspondientes al problema. La primer columna corresponde a los valores de la variable Prof, la segunda a Turno y la tercera a Calif.

Estos son sólo algunos de los bloques y comandos que se pueden usar en los distintos procedimientos de BMDP, más adelante se describirán los bloques y comandos que se utilizarán para el desarrollo de esta tesis.

BLOQUES	COMANDOS
/ INPUT	VARIABLES=3.
	FORMAT IS FREE.
/ VARIABLE	NAMES ARE Prof,turno,calif.
	GROUPING IS Prof.
/ GROUP	CODES(Prof)=1,2,3.
	NAMES(Prof)=Juan,Paco,Pedro.
	CODES(Turno)=1,2.
	NAMES(Turno)=am,pm.
/ PRINT	DATA.
/END.	
1 1 69	
1 2 70	
1 1 79	
1 2 55	
2 1 89	
2 2 90	
2 1 75	
2 2 69	
3 1 95	
3 2 70	
3 1 75	
3 2 80	
3 1 70	

Cuadro 3.1

3.1.2 EL PROCEDIMIENTO LR (LOGISTIC REGRESION)

El procedimiento LR calcula los estimadores de máxima verosimilitud de los parámetros del modelo de Regresión Logística descrito en el capítulo II. Este procedimiento incorpora o remueve las variables independientes del modelo de una manera "por pasos", esto es, en cada iteración incorpora una variable al modelo o la deja fuera de él. Esto permite hacer una selección de las variables al mismo tiempo que se ajusta un modelo a los datos.

Como se explicó en el capítulo II, cuando en el modelo hay una variable categórica con k categorías se necesitan k-1 variables de diseño para representarla. El procedimiento LR permite generar variables de diseño de una forma automática.

Además de las variables de diseño y estimadores de máxima verosimilitud de los parámetros, el procedimiento calcula una serie de estadísticas importantes para el análisis de los datos como son: media, desviación estándar, sesgo y curtosis para las variables continuas; estadísticas de bondad de ajuste, matriz de correlación de los coeficientes, desviación estándar de los estimadores, odds ratios, etc.

A continuación se hace una descripción de los bloque y comandos que integran al procedimiento LR.

BLOQUE / PROBLEM.

Este bloque se utiliza para especificar el tiempo máximo en segundos en el cual LR debe hacer el cálculo de los estimadores de los parámetros. Esto se realiza con el siguiente comando:

TIME=No. de segundos.

Si no se especifica este bloque, el programa asume que el tiempo es ilimitado.

BLOQUE / INPUT.

Este bloque es común en todos los procedimientos del BMDP. Consta de 3 comandos principales cuya sintaxis es la siguiente:

VARIABLE=No. de variables.

Indica el número de variables para cada sujeto (caso) bajo estudio.

FORMAT=Formato

Especifica la presentación de los valores de los datos para cada caso. Los formatos disponibles son los siguientes:

FREE Las variables están separadas por uno o más espacios en blanco y/o una coma, cada caso debe ocupar un renglón.

STREAM Permite más de un caso por renglón. Las variables son separadas por blancos o comas.

SLASH Permite más de un caso por renglón. Las variables son separadas y los casos son separados por una diagonal después de la última variable.

BINARY Se usa cuando los datos están en un archivo en forma binaria.

Otro formato es el formato tipo FORTRAN, por ejemplo:

FORMAT='2A4,2F6,F1,3X,F5.2/5X,F6'.

El programa lee 2 campos alfanuméricos de cuatro caracteres, dos números de 6 dígitos, un número de 1 dígito, salta 3 espacios en blanco y lee un número de 5 dígitos con dos decimales, después pasa al siguiente renglón, salta 5 espacios en blanco y lee un número de 6 dígitos.

FILE='Nombre del archivo.extensión'. o UNIT=No. de unidad.

Identifica la localización del archivo de datos. Se usa FILE cuando se trabaja en una VAX, PDP-11, IBM PC y compatibles, y sistema UNIX; UNIT se usa para Mainframe IBM.

BLOQUE / VARIABLE.

El bloque VARIABLE al igual que el INPUT es parte del conjunto básico de bloques usados por todos los procedimientos BMDP. Los comandos usados con mayor frecuencia son los siguientes:

NAMES=Lista de nombres.

Proporciona la lista de nombres de cada una de las variables. Asigna los nombres en el mismo orden en que las variables aparecen en los registros. Los nombres constan de a lo más ocho caracteres y deben estar encerradas entre apóstrofos si el primer caracter es un blanco, un símbolo o un número.

USE=Lista de variables.

Proporciona la lista de nombres o números de las variables que serán usadas en el análisis. Si se omite el comando, el programa usa todas las variables.

BLOQUE / GROUP.

Muchos procedimientos del BMDP utilizan el bloque GROUP para:

- Clasificar casos en grupos para pruebas t, análisis de varianza, etc.
- Definir categorías para tablas de frecuencia.
- Especificar símbolos de identificación para miembros de un mismo grupo en gráficas de puntos.

Se pueden formar grupos de dos formas:

- Especificando códigos para las categorías de una variable discreta (por ejemplo sexo).
- Especificando puntos de corte para dividir una variable continua en intervalos (por ejemplo edad).

Los comandos del bloque GROUP más comunes son los siguientes:

CODES(j)=Lista de categorías.

Proporciona la lista de los códigos para las categorías de una de una variable discreta j, donde j es el nombre o número de la variable correspondiente.

NAMES(j)=Lista de nombres.

Lista los nombres para las categorías o intervalos de la variable j, donde j es el nombre o número de la variable correspondiente. Los nombres son asignados en el orden de las categorías o intervalos.

BLOQUE / REGRESS.

Se utiliza para especificar cual es la variable dependiente, las variables independientes y su tipo; el método a usar para generar las variables de diseño, etc. Los comandos de este bloque son los siguientes:

PARA LA VARIABLE DEPENDIENTE.

DEPENDENT=Nombre de variable.

Especifica el nombre de la variable dependiente.

PARA LAS VARIABLES INDEPENDIENTES.

INTERVAL=Lista de variables.

Especifica la lista de los nombres de las variables continuas.

CATEGORICAL= Lista de variables.

Especifica los nombres de las variables categóricas. El procedimiento LR genera un conjunto de variables de diseño para cada variable especificada en esta lista.

DVAR=Método.

Especifica el método a utilizar para generar las variables de diseño. Existen tres métodos:

MARGINAL.- Se le asigna a cada categoría, excepto a la primera, un 1 a la variable de diseño que la representa y 0 a las restantes. Para la primera categoría, LR asigna un -1 a todas las variables de diseño.

PARTIAL.- Es igual que el método anterior con la diferencia de que la primera categoría se le asigna 0 a todas las variables de diseño. Este método es apropiado si se tiene definido un grupo de referencia y si se desean estimar los odds ratio.

ORTHOGONAL: Con este método LR genera componentes de polinomios ortogonales. La primer variable de diseño es el componente lineal, la segunda el componente cuadrático, etc. Se utiliza cuando se tienen categorías ordenadas.

EL MODELO

MODEL=Lista de términos.

Especifica los términos que deben ser incluidos en el modelo. Pueden incluirse interacciones de variables, por ejemplo: ESTATURA*PESO. Solamente se ajusta el modelo con las variables de la lista de MODEL.

START=Lista de OUT ó IN.

Para cada término especificado en MODEL se indica si éste es incluido o no al inicio del proceso por pasos. Cuando MODEL es especificado, por omisión el valor de START es IN para cada término en MODEL. Cuando MODEL no se especifica, el valor de START es OUT para cada término.

CONSTANT=OUT ó IN.

Indica si el término constante es incluido al inicio del proceso por pasos. Si se omite, su valor es IN.

MOVE=Lista del no. de movimientos de cada variable.

Para cada término en MODEL, indica cuantas veces el término puede moverse dentro o fuera del modelo. Si se omite, cuando MODEL es especificado MOVE=0 para cada término; cuando MODEL no se especifica, MOVE=2 para cada término.

CMOVE=No. de movimientos.

Especifica el número de veces que el término constante puede moverse dentro o fuera del modelo. Por omisión no se le permite moverse.

ENTER= No.1, No.2

Indica el valor-p para los valores χ^2 o F usados para controlar la entrada de una variable al modelo. Para que la variable entre el valor-p debe ser menor que los valores de ENTER. No.1 se utiliza para el paso hacia adelante y No.2 para el paso hacia atrás.

REMOVE= No.1, No.2

Indica el valor-p para los valores χ^2 o F usados para controlar la salida de una variable al modelo. Para que la variable salga el valor-p debe ser mayor que los valores de REMOVE. Al igual que en ENTER No.1 se utiliza para el paso hacia adelante y No.2 para el paso hacia atrás.

EL METODO DE SELECCION Y CONTROLES DE CALCULO

METHOD=Método.

Especifica el método para incorporar o remover términos en cada paso. Hay 2 métodos disponibles:

MLR.- Método de máxima verosimilitud. Para remover o incorporar términos al modelo utiliza el valor-p de una $\chi^2 = 2\left|\ln\left[\frac{L(\beta_{actual})}{L(\beta_{candidato})}\right]\right|$.

ACE.- Estimación por covarianza asintótica. Hace lo mismo que MLR pero con el valor-p de un estadístico F que se calcula a partir de un estimador de la matriz de covariazas asintótica de β.

RULE=Regla.

Especifica la regla para incorporar o remover variables e interacciones de variables. Hay 3 reglas disponibles:

NONE.- Cualquier término puede ser incorporado o removido.

SINGLE.- Sólo se puede mover un término en cada paso. Un término se incorpora sólo si sus interacciones de orden menor están dentro del modelo o removido sólo si sus interacciones de orden mayor están fuera.

MULTIPLE.- Es similar a la regla SINGLE con la diferencia de que se pueden mover más de un término a la vez.

ITERATION=No. de iteraciones.

Especifica el máximo número de iteraciones para maximizar la función de verosimilitud. Por omisión son 10 iteraciones.

COST=costo 1, costo 2, costo 3, costo 4.

Los cuatro números designan el costo relativo para la clasificación correcta o incorrecta de los dos grupos (1 y 0, éxito y fracaso, etc.) según la siguiente matriz:

		valores pronosticados	
		grupo 1	grupo2
valores observados	grupo 1	costo 1	costo 3
	grupo 2	costo 2	costo 4

Estos costos se utilizan para generar puntos de corte para las probabilidades que da el modelo, y con estos puntos hacer pronósticos (los puntos de corte se analizarán en el capítulo VII). Por omisión los valores de los costos son 0,-1,-1,0 respectivamente.

BLOQUE / PRINT.

Los comandos de este bloque son los siguientes:

LINESIZE=Tamaño.

LINESIZE determina el tamaño de las líneas en la salida.

CELLS=Patrón.

Controla la formación de celdas en el reporte. Una celda es un patrón (valores particulares) diferente de las variables independientes. Los patrones son los siguientes:

USED.- Celdas formadas usando un patrón de todas las variables consideradas por el modelo.

MODEL.- Usa un patrón distinto solamente para aquellas variables incluidas en el modelo.

BOTH.- Utiliza las dos anteriores.

NONE.- No imprime un reporte.

SORT=Ordenamiento.

Especifica la forma en que las celdas son ordenadas en el reporte. Hay 4 tipos de ordenamientos:

NONE.- No se ordenan.

PROP.- Se ordenan por proporción de éxitos.

VAR.- Se ordenan por los valores de las variables independientes.

BOTH.- Se ordenan por las dos anteriores.

PLOT.

Imprime un diagrama de dispersión de la proporción observada del primer grupo contra la proporción pronosticada y la proporción observada del primer grupo contra los log odds pronosticados.

HISTOGRAM.

Para ambos grupos, se imprimen histogramas de las probabilidades pronosticadas de cada sujeto que están en el primer grupo. Tienen gran utilidad cuando la variable independiente es continua.

COST.

Imprime la tabla de clasificaciones correctas e incorrectas y las pérdidas de cada punto de corte utilizando los costos especificados en el bloque REGRESS. Este comando también imprime una gráfica que muestra el porcentaje de clasificaciones correctas para cada grupo como una función de los puntos de corte.

COVA.

Imprime la matriz de covarianzas de los parámetros.

CORR.

A menos que se especifique NO CORR, LR imprime la matriz de correlaciones de los parámetros.

BLOQUE /END.

Se utiliza para finalizar las instrucciones del programa.

El orden de los bloques debe ser el mismo en el cual se explicaron. Los bloques INPUT, REGRESS y END son indispensables para el procedimiento LR. Para ampliar la explicación del uso de estos bloques y comandos se recomienda consultar el manual "BMDP statistical software" de Dixon (1990).

3.2. EL PROGRAMA SAS (STATISTICAL ANALYSIS SYSTEM)

El sistema SAS es un software para el análisis de datos y creación de reportes, es un grupo de programas de cómputo.

Con el software base de SAS se pueden crear "bases de datos SAS", modificarlas, calcular estadísticas simples y crear reportes. Otra parte del software proporciona gráficas, pronósticos y estadísticas sofisticadas como por ejemplo:

- Histogramas y estadísticas descriptivas.
- Modelos log-lineales.
- Pruebas t y análisis de varianza.
- Modelos de regresión lineal y no lineal.
- Estimación por máxima verosimilitud.
- Análisis multivariado.
- Análisis de supervivencia.
- Análisis de series de tiempo.

3.2.1. LOS PROCEDIMIENTOS DE SAS

De la misma manera que en BMDP, para realizar el análisis de un conjunto de datos es necesario elaborar un programa o una serie de instrucciones que están compuestas por procedimientos y comandos. Para ejemplificar estos procedimientos tomaremos en cuenta nuevamente al problema 3.1. Un programa en SAS equivalente se encuentra en el cuadro 3.2.

DATA se utiliza para crear una base de datos SAS, en este ejemplo la base de datos se nombrará como "Clase". El comando INPUT es para leer las variables que en este caso son tres: prof, turno y calif, a continuación de INPUT está la instrucción CARDS que indica que los datos a analizar están en el cuerpo del programa. Después de los datos se encuentra que el comando RUN que se utiliza para decirle al programa que inicie la creación de la base de datos.

El procedimiento que se utiliza para este ejemplo es el procedimiento MEANS el cual calcula estadísticas básicas. Para indicar que se utiliza un procedimiento se utiliza la palabra reservada PROC. DATA=calif es para ordenar que se utilicen los datos de la base "calif". VAR señala la variable a la cual se calcularán las estadísticas y BY es para formar grupos de ellas por profesor y turno.

```
DATA clase;
      INPUT prof,turno,calif;
      CARDS;
           1 1 89
           1 2 70
           1 1 79
           1 2 55
           2 1 89
           2 2 90
           2 1 75
           2 2 69
           3 1 95
           3 2 70
           3 1 75
           3 2 80
           3 1 70
RUN;

PROC MEANS DATA=clase;
  VAR calif
  BY prof turno
RUN;
```

Cuadro 3.2

Este es sólo uno de los procedimientos que se pueden usar en el programa SAS. Para la presente investigación se utilizará principalmente al procedimiento LOGISTIC.

3.2.2. EL PROCEDIMIENTO LOGISTIC

El procedimiento LOGISTIC ajusta un modelo de Regresión Logística a datos con variable de respuesta binaria u ordinal por el método de máxima verosimilitud. Una característica importante de este procedimiento es que posee varios métodos de selección de variables y que para una variable de respuesta binaria puede desplegar estadísticos y gráficos útiles para el diagnóstico de regresión. La estructura y principales comandos del procedimiento LOGISTIC se describen a continuación:

PROC LOGISTIC

Señala el inicio del procedimiento y sus principales opciones son las siguientes:

ORDER=Orden

Especifica el orden de las categorías de la variable de respuesta. El orden es de suma importancia, ya que el programa realiza todos los cálculos en base a la primera categoría de la variable. Si se especifica ORDER=data, el programa toma como primera categoría a la primera que aparece en la base de datos (es la más recomendable).

OUTEST=conjunto de datos

Crea una nueva base de datos que contiene los estimadores de los parámetros finales y opcionalmente sus covarianzas.

COVOUT

Incluye la matriz de covarianza al conjunto de datos especificado en OUTEST.

MODEL

El comando MODEL contiene el nombre de la variable de respuesta y las variables predictoras:

Respuesta=Lista de variables predictoras;

"Respuesta" es el nombre de la variable de respuesta o dependiente seguida del signo igual y la lista de variables predictoras.

Las principales opciones del comando MODEL son las siguientes:

NOINT

Suprime la estimación del término constante del modelo de Regresión Logística.

SELECTION=Método de selección

Especifica el método de selección de variables. Hay tres métodos principales:

BACKWARD para una eliminación hacia atrás.
FORWARD para una selección hacia adelante.
STEPWISE para una selección por pasos.

DETAILS

Imprime una salida detallada del proceso de selección de variables.

FAST

Utiliza un algoritmo computacional para calcular una aproximación de primer orden de los estimadores para cada subsecuente eliminación. FAST es extremadamente eficiente porque el modelo no se reajusta para cada variable removida.

INCLUDE=n

Incluye las primeras n variables especificadas en el comando MODEL en el ajuste de todos los modelos.

SLENTRY=valor

Especifica el nivel de significancia para entrar al modelo. Para que una variable entre al modelo, su valor-p debe ser menor que el valor especificado en SLENTRY.

SLSTAY=valor

Especifica el nivel de significancia para salir del modelo. Para que una variable salga del modelo, su valor-p debe ser mayor que el valor especificado en SLSTAY.

CORRB

Imprime la matriz de correlación de los parámetros estimados.

CTABLE

Imprime la tabla de clasificación del modelo final donde se incluye, para cada nivel de probabilidad, los aciertos y errores de pronósticos del modelo.

INFLUENCE

Despliega medidas de diagnóstico para determinar observaciones influenciales en el caso de una variable de respuesta binaria. Incluye para cada observación las medidas de diagnóstico de regresión desarrolladas por Pregibon (1981).

IPLOTS

Imprime una gráfica para cada estadístico de diagnóstico de regresión.

OUTPUT

Crea un nuevo conjunto de datos SAS que contiene todas las variables de entrada, el estimador de la probabilidad de respuesta, los límites de confianza para esta probabilidad y los estadísticos de diagnóstico de regresión. Su sintaxis es la siguiente:

OUTPUT OUT=base datos

"Base" es el nombre de la nueva base de datos, "datos" especifica los datos y estadísticos a incluir en "base", los estadísticos más importantes son los siguientes:

H elementos de la diagonal de la matriz Hat.
L límite de confianza inferior de la probabilidad estimada de la respuesta.
U límite de confianza superior de la probabilidad estimada de la respuesta.
P probabilidad pronosticada de la variable de respuesta.
RESCHI residuales de Pearson.

RESDEV residuales de la devianza.

(Los residuales y la matriz Hat serán tratados con mayor detalle en el capítulo V).

RUN

Indica que se ejecuten las instrucciones del procedimiento LOGISTIC (y en general de cualquier procedimiento).

Para ampliar la explicación del uso de estos comandos se recomienda consultar los manuales "SAS Guide for personal computers" del SAS Institute (1988).

CAPITULO IV. CAUSAS QUE ORIGINAN EL PARTO POR CESAREA. (SELECCION DE VARIABLES)

4.1. La importancia de la selección de variables

4.2. Selección por pasos (Stepwise Logistic Regression)

4.3. Resultados de la selección de variables

4.1. LA IMPORTANCIA DE LA SELECCION DE VARIABLES

La selección de variables predictoras tiene gran importancia en cualquier problema de regresión. El principal objetivo de la selección de variables es encontrar el modelo más parsimonioso que describa mejor a la variable de respuesta. La selección de variables predictoras se hace generalmente después de seleccionar el modelo de regresión.

Los métodos de selección de variables eligen un subconjunto de una serie de variables potencialmente predictoras determinando cuáles de éstas tienen mayor influencia sobre la variable de respuesta.

Algunos de los problemas que se pueden presentar cuando no se realiza una selección de variables son los siguientes:

a) Si hay un número excesivo de variables predictoras, el modelo resultante puede ser difícil de manejar y puede resultar costoso.

b) Se pueden producir estimadores numéricamente inestables, lo cual se refleja en coeficientes y errores estándar estimados excesivamente grandes. Este problema es consecuencia, principalmente, de la presencia de variables altamente correlacionadas.

c) Si el número de variables predictoras es alto, la estimación del modelo requerirá de un tiempo de cómputo más largo, lo cual resulta más costoso.

Como puede verse, es más conveniente manejar un modelo con el menor número de variables; sin embargo, recientemente en epidemiología se recomienda que sean incluidas en el modelo todas aquellas variables que tengan relevancia científica en el fenómeno bajo estudio, independientemente de si son o no estadísticamente significativas, esto con el propósito de que la selección de variables no produzca resultados absurdos.

4.2. SELECCION POR PASOS (STEPWISE LOGISTIC REGRESSION)

Uno de los métodos de selección de variables más utilizados es la selección por pasos. Greenland (1989, p. 344) resume al método de selección por pasos como sigue:

1. Dividir las variables predictoras candidatas en dos clases: variables fijas y variables no fijas dentro del modelo. Las variables fijas son incluidas en el modelo inicial y no pueden ser removidas de él; las variables no fijas son sujetas a una selección o eliminación por el procedimiento "por pasos".

2. El ciclo de selección-eliminación se realiza de la siguiente manera:

Selección: Entre todas las variables candidatas que no están en el modelo actual, seleccionar aquella que tenga el estadístico chi-cuadrada más grande y cuyo valor-p sea menor que algún nivel de significancia establecido.

Eliminación: Entre todas las variables que están en el modelo actual eliminar aquella que tenga el estadístico chi-cuadrada más pequeño y un valor-p mayor que algún nivel de significancia establecido.

El algoritmo continúa hasta que ya no se pueda seleccionar o eliminar variable alguna.

Un algoritmo más detallado de la selección por pasos lo da Hosmer (1989, p.p, 106-110):

Sean p variables potencialmente predictoras
PE nivel de significancia para que entre una variable.
PR nivel de significancia para remover una variable.

PASO 0

a) Ajustar un modelo que contiene únicamente al término constante y evaluar su log verosimilitud L_0 definida, por la ecuación [2.12a].

b) Ajustar los p posibles modelos de Regresión Logística univariados con las p variables potencialmente predictoras y calcular su log verosimilitud $L_j^{(0)}$ (log verosimilitud del modelo que contiene a X_j en el paso 0).

c) Calcular el estadístico G de la prueba de razón de verosimilitud para el modelo que contiene a X_j contra el modelo que contiene sólo al término constante. $G_j^{(0)}=2(L_j^{(0)} - L_0)$. Calcular el valor-p denotado por $p_j^{(0)}$ = Pr [χ^2 (v) > $G_j^{(0)}$], donde v=1 si X_j es continua y v=k-1 si X_j es categórica con K niveles.

d) La variable más importante es aquella que tenga el valor-p mas pequeño y se denota como X_{e1} , entonces $p_{e1}^{(0)}$ = min($p_j^{(0)}$). Esta variable entra al modelo y se ejecuta el paso 1 si $p_{e1}^{(0)}$ < PE, de otro modo el algoritmo termina.

PASO 1

a) Ajustar el modelo de Regresión Logística que contenga X_{e1} y su log verosimilitud $L_{e1}^{(1)}$.

b) Ajustar p-1 modelos de Regresión Logística que contengan a X_{e1} y X_j , j=1,2,3,...,p. j≠ e1. Calcular la log verosimilitud de estos modelos denotadas por $L_{e1j}^{(1)}$.

Calcular el estadístico de la prueba de razón de verosimilitud del modelo que contiene X_{e1} y X_j contra el que contiene solamente a X_{e1} denotado por $G_j^{(1)}=2(L_{e1j}^{(1)} - L_{e1}^{(1)})$. Calcular el valor-p, $p_j^{(1)}$.

d) Denotar como X_{e2} a la variable que se relacione con el $p_j^{(1)}$ más pequeño. Si este valor es menor que PE, entonces X_{e2} entra al modelo y se ejecuta el paso 2, de otro modo detenerse.

PASO 2

a) Ajustar un modelo que contenga a X_{e1} y X_{e2} . El paso 2 incluye un chequeo para la eliminación hacia atrás de una de las variables que están en el modelo. Sea $L_{-ej}^{(2)}$ la log verosimilitud con X_{ej} removida se obtiene $G_{-ej}^{(2)}=2(L_{e1e2}^{(2)} - L_{-ej}^{(2)})$ y el valor-p, $p_{-ej}^{(2)}$. La forma de elegir a la variable que será removida es a partir del valor-p. Si se denota a la variable candidata a ser removida como X_{r2} , entonces $p_{r2}^{(2)}$ = max($p_{-e1}^{(2)}$, $p_{-e2}^{(2)}$). Para decidir si X_{r2} debe ser removida, el programa compara $p_{r2}^{(2)}$ con PR: Si $p_{r2}^{(2)}$ > PR entonces

la variable es removida del modelo, de lo contrario X_{r2} permanece en éste y se procede a la fase de selección.

b) Ajustar p-2 modelos de Regresión Logística que contienen X_{e1}, X_{e2} y X_j, j=1,2,3,...,p. j≠ e1,e2. Evaluar la log verosimilitud para cada uno de estos modelos.

c) Calcular G para el modelo que contiene X_{e1}, X_{e2} y X_j contra el que contiene solamente a X_{e1} y X_{e2} y determinar su correspondiente valor-p.

d) Sea X_{e3}, la variable que contiene el valor-p más pequeño, si éste es menor que PE entonces X_{e3} se incluye en el modelo y se ejecuta el paso 3, de otro modo detenerse.

PASO 3

El procedimiento para el PASO 3 es idéntico al PASO 2: el programa ejecuta el chequeo para eliminación hacia atrás seguido de una selección hacia adelante. Este proceso se repite de la misma manera hasta el último paso, PASO S.

PASO S

Este paso se ejecuta cuando: 1) las p variables entraron al modelo, o cuando, 2) todas las variables que están en el modelo tienen su valor-p menor que PR por lo que no pueden ser eliminadas y todas las que no están en el modelo tienen su valor-p mayor que PE por lo que no pueden entrar. El algoritmo termina.

El mismo Hosmer (1978) desarrolló un programa para el algoritmo anterior. Un procedimiento de selección de variables por pasos para modelos de regresión no lineal puede encontrarse en Peduzzi (1980).

4.3. RESULTADOS DE LA SELECCION DE VARIABLES

El programa SAS difiere en algunas cosas del programa BMDP pero los resultados finales son los mismos. La corrida que aquí se presenta es la producida por SAS ya que es más detallada que la que produce el BMDP. La selección por pasos de SAS difiere del algoritmo

que se dio en el apartado anterior pero en esencia hacen lo mismo (Los programas para realizar la selección de variables y el ajuste del modelo se encuentran en el anexo IV). A continuación se da una breve explicación de los estadísticos en los que se basa SAS para la selección de variables:

En la selección y eliminación de variables no se utiliza el estadístico de razón de verosimilitud; éstas se hacen en base al estadístico *Chi-cuadrada score* que se define de la siguiente manera para cada variable X_j de las t que están fuera del modelo:

$$X^2 = U'(\hat{\beta})I^{-1}(\hat{\beta})U(\hat{\beta})$$

donde $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{j-1}, 0, \hat{\beta}_{j+1}, ..., \hat{\beta}_t)$ es el vector de estimadores de los parámetros sin tomar en cuenta a X_j.

$U(\beta)$ es el vector de derivadas parciales de la función log verosimilitud con respecto al vector de parámetros.

$I(\beta)$ es la matriz de los negativos de las segundas derivadas parciales de la función log verosimilitud con respecto al vector de parámetros.

Este estadístico tiene una distribución χ^2 asintótica con 1 grado de libertad.

Después de calcular la Chi-cuadrada score para cada X_j se encuentra su valor-p y se compara con el nivel de significancia PE para saber si la variable entra al modelo de acuerdo al algoritmo de Hosmer.

Para la eliminación de las variables que están dentro del modelo, se sigue un procedimiento similar, con la diferencia de que el valor-p se compara con el nivel de significancia PR para saber si alguna variable sale del modelo.

Después de seleccionar alguna variable, el programa SAS hace una evaluación del ajuste del modelo con las variables seleccionadas, utilizando los siguientes estadísticos (cabe señalar que para efectos de esta investigación se utilizarán otros estadísticos, los cuales serán descritos en el capítulo V):

a) $-2 \log L = -2\sum_{i=1}^{n} \ln(\hat{\pi}_j)$.

b) AIC = -2 log L + 2(k+s) (Akaike Information Criterion)

c) SC = -2 log L + (k+s) ln (n) (Schwartz Criterion)

d) Score que sirve para probar la significancia conjunta de las variables predictoras que están dentro del modelo.

Donde s es el número de variables predictoras dentro del modelo y
k es el número de niveles para la variable de respuesta, para esta investigación k es igual a 2.

En general, cuando se comparan modelos, los valores más pequeños de estos estadísticos indican un mejor ajuste.

Después de cada paso, SAS proporciona un estadístico Chi-cuadrada residual para evaluar la significancia del modelo. Este estadístico se calcula de la siguiente forma:

Sea s el número de variables de interés
t el número de variables del modelo reducido (t<s).
$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1,...,\hat{\beta}_t,...,\hat{\beta}_s)$ el vector de parámetros estimados.

La Chi-cuadrada residual es el estadístico Chi-cuadrada score evaluado en $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1,...,\hat{\beta}_t,0,...,0)$. La chi-cuadrada residual tiene una distribución Chi-cuadrada asintótica con s-t grados de libertad. Mientras más pequeño es este valor, mejor es el ajuste del modelo.

Al realizar una selección por pasos de las variables potencialmente predictoras dadas en el cudro 1.5., utilizando los programas SAS y BMDP con PE = PR = 0.20, se dieron los siguientes resultados:

SELECCION POR PASOS

PASO 0. Ninguna variable en el modelo.

Chi-cuadrada residual = 60.3443 con 12 GL (p=0.0001)

Análisis de las variables que no están en el modelo

Variable	Score Chi-Cuadrada	Valor-p
TPELVIS	12.2500	0.0005
DILCER	24.4854	*0.0001
POSFETO	0.3333	0.5637
PCEFALIC	16.0274	0.0001
EDAD	15.2751	0.0001
ESTATURA	16.2564	0.0001
PESO	15.9793	0.0001
TARTERIA	5.5538	0.0184
NPARTOS	13.5692	0.0002
NCESAREA	1.9231	0.1655
NABORTOS	1.9231	0.1655
EDOMADRE	1.8000	0.1797

PASO 1. Entra la variable DILCER al modelo:

Criterios para valuar el ajuste del modelo

Criterio	Sin las variables	Con las variables	Chi-cuadrada para las variables
AIC	138.629	113.956	
SC	138.629	116.561	
-2 LOG L	138.629	111.956	26.674 con 1 GL (p=0.0001)
Score			24.485 con 1 GL (p=0.0001)

Chi-Cuadrada residual = 43.6076 con 11 GL (p=0.0001)

Análisis de variables que no están en el modelo

Variable	Score Chi-Cuadrada	Valor-p
TPELVIS	30.1182	*0.0001
POSFETO	1.7451	0.1865
PCEFALIC	2.5720	0.1088
EDAD	1.3696	0.2419
ESTATURA	2.0924	0.1480
PESO	1.6487	0.1991
TARTERIA	0.2536	0.6146
NPARTOS	1.9135	0.1666
NCESAREA	8.7252	0.0031
NABORTOS	0.2138	0.6438
EDOMADRE	0.0634	0.8012

PASO 2. Entra la variable TPELVIS al modelo:

Criterios para valuar el ajuste del modelo

Criterio	Sin las variables	Con las variables	Chi-cuadrada para las variables
AIC	138.629	84.609	
SC	138.629	89.820	
-2 LOG L	138.629	80.609	58.020 con 2 GL (p=0.0001)
Score			48.507 con 2 GL (p=0.0001)

Chi-Cuadrada residual = 18.3695 con 10 GL (p=0.0490)

Análisis de las variables que no están en el modelo

Variable	Score Chi-Cuadrada	Valor-p
POSFETO	3.1176	0.0775
PCEFALIC	0.0097	0.9214
EDAD	0.0948	0.7582
ESTATURA	0.0512	0.8210
PESO	0.1024	0.7490
TARTERIA	0.0392	0.8431
NPARTOS	1.7797	0.1822
NCESAREA	5.2548	*0.0219
NABORTOS	1.0882	0.2969
EDOMADRE	0.5865	0.4438

PASO 3. Entra la variable NCESAREA al modelo:

Criterios para valuar el ajuste del modelo

Criterio	Sin las variables	Con las variables	Chi-cuadrada para las variables
AIC	138.629	82.038	
SC	138.629	89.853	
-2 LOG L	138.629	76.038	62.592 con 3 GL (p=0.0001)
Score			50.619 con 3 GL (p=0.0001)

Chi-Cuadrada residual = 13.7556 con 9 GL (p=0.1313)

Análisis de las variables que no están en el modelo

Variable	Score Chi-Cuadrada	Valor-p
POSFETO	3.7770	0.0520
PCEFALIC	0.1088	0.7415
EDAD	0.5111	0.4746
ESTATURA	0.1657	0.6840
PESO	0.1890	0.6638
TARTERIA	0.0867	0.7684
NPARTOS	4.7056	*0.0301
NABORTOS	1.5123	0.2188
EDOMADRE	0.4889	0.4844

PASO 4. Entra la variable NPARTOS al modelo:

Criterios para valuar el ajuste del modelo

Criterio	Sin las variables	Con las variables	Chi-cuadrada para las variables
AIC	138.629	76.909	
SC	138.629	87.330	
-2 LOG L	138.629	68.909	69.720 con 4 GL (p=0.0001)
Score			54.506 con 4 GL (p=0.0001)

Chi-Cuadrada residual = 9.0564 con 8 GL (p=0.3376)

Análisis de las variables que no están en el modelo

Variable	Score Chi-Cuadrada	Valor-p
POSFETO	3.9482	*0.0469
PCEFALIC	0.2132	0.6443
EDAD	0.7317	0.3923
ESTATURA	0.1250	0.7236
PESO	0.0937	0.7595
TARTERIA	0.1449	0.7034
NABORTOS	0.9479	0.3303
EDOMADRE	0.0411	0.8394

PASO 5. Entra la variable POSFETO al modelo

Criterio para el ajuste del modelo

Criterio	Sin las variables	Con las variables	Chi-cuadrada para las variables
AIC	138.629	75.655	
SC	138.629	88.681	
-2 LOG L	138.629	65.655	72.975 con 5 GL (p=0.0001)
Score			56.248 con 5 GL (p=0.0001)

Chi-Cuadrada residual = 5.2751 con 7 GL (p=0.6264)

Análisis de las variables que no están en el modelo

Variable	Score Chi-Cuadrada	Valor-p
PCEFALIC	0.0149	0.9027
EDAD	0.2828	0.5949
ESTATURA	0.0002	0.9877
PESO	0.0047	0.9456
TARTERIA	0.7870	0.3750
NABORTOS	0.8415	0.3590
EDOMADRE	0.0176	0.8944

PASO S. Las variables que están en el modelo no tienen un valor-p que rebase el nivel de significancia PR y las que están fuera del modelo no tienen un valor-pel cuál sea menor a el nivel de significancia PE.

Como conclusión, las variables estadísticamente importantes a los niveles de significancia PE = PR = 0.20 son las siguientes, a reserva de la opinión de los médicos:

RESUMEN DE LA SELECCION POR PASOS

PASOS	Variable Entra	Removida	Score Chi-Cuadrada	Valor-p
1	DILCER	--	24.4854	0.0001
2	TPELVIS	--	30.1182	0.0001
3	NCESAREA	--	5.2548	0.0219
4	NPARTOS	--	4.7056	0.0301
5	POSFETO	--	3.9482	0.0469

CAPITULO V. AJUSTE DEL MODELO

5.1. Verificación de la selección de las variables con médicos

5.2. Estimación de los parámetros

5.3. Diagnósticos del modelo de Regresión Logística
- 5.3.1. Diagnóstico informal
- 5.3.2. Chi-cuadrada de Pearson, Devianza y Matriz Hat
- 5.3.3. Método gráfico

5.4. Modelo final

5.1. VERIFICACION DE LA SELECCION DE LAS VARIABLES CON MEDICOS

Como se mencionó en el capítulo anterior, es importante no separarse de la opinión de los especialistas del área de ginecoobstetricia; es por ello, que es indispensable consultar con los médicos para verificar si en la selección de variables no se tomó en cuenta alguna variable biológicamente importante.

Después de haber hecho la selección de variables en el capítulo anterior y verificando con los médicos las variables que resultaron de importancia estadística para el modelo, se concluyó que entre estas variables se encuentran las que para ellos son de importancia biológica las cuales son: posición del feto y tipo de pelvis, es por esto que el ajuste del modelo se hará con las variables obtenidas en el capítulo anterior.

5.2. ESTIMACION DE LOS PARAMETROS

Después de hacer la selección por pasos de las variables predictoras y ponerlas a consideración de los médicos, el siguiente paso es proceder a ajustar un modelo de Regresión Logística múltiple a los datos con variable de respuesta binaria Y definida como:

Y=1, Cesárea.
Y=0, Parto natural.

y variables predictoras:

Continuas: Dilatación cervical, número de partos anteriores y número de cesáreas anteriores.
Binarias: Tipo de pelvis y posición del feto.

El modelo de regresión logística es el siguiente:

$$\pi(X) = \frac{exp[g(X)]}{1+exp[g(X)]} \quad \text{donde:}$$

$g(X) = \beta_0 + \beta_1$ tpelvis $+ \beta_2$ dilcer $+ \beta_3$ posfeto $+ \beta_4$ npartos $+ \beta_5$ ncesarea

β_0 ,β_1 ,β_2 ,β_3 ,β_4 ,β_5 son parámetros desconocidos a estimar.

La estimación de estos parámetros se realizó con la ayuda de los paquetes SAS y BMDP donde, de manera general, se obtuvieron los mismos resultados que se describen en la tabla 5.1.

ESTIMADORES DE MAXIMA VEROSIMILITUD

VARIABLE	ESTIMADOR	ERROR ESTANDAR	χ^2 DE WALD	VALOR-P
TPELVIS	4.081	1.156	12.454	0.0004
DILCER	-0.332	0.084	15.621	0.0001
POSFETO	2.373	1.384	2.942	0.0863
NPARTOS	-1.395	0.726	3.691	0.0547
NCESAREA	3.183	1.164	7.473	0.0063

Tabla 5.1

El error estándar de los parámetros estimados se calcula como la raíz cuadrada del elemento diagonal correspondiente de la matriz de covarianzas estimada. La Chi-cuadrada de Wald es el cuadrado del cociente del parámetro entre su error estándar. La Chi-cuadrada de Wald sigue una distribución Chi- cuadrada con un grado de libertad; a partir de esto, se calcula el valor-p y se utiliza para saber si la variable es significativa o no (mientras menor sea el valor-p, más significativa es la variable).

5.3. DIAGNOSTICOS DEL MODELO DE REGRESION LOGISTICA

Al obtener la muestra, no se puede saber si alguna de las observaciones puede influir para que el ajuste de el modelo no sea bueno, por lo que es recomendable realizar un diagnóstico al modelo ajustado.

Algunos de estos diagnósticos se hacen utilizando estadísticos de bondad de ajuste como la Chi-cuadrada de Pearson, la Devianza y otros derivados de éstos. Además, es posible hacer algunos diagnósticos en base a gráficas.

5.3.1. DIAGNOSTICO INFORMAL

Un primer intento para probar el buen ajuste del modelo es hacer un diagnóstico informal para observar si la gráfica descrita por el modelo ajustado es monótona y tiene forma sigmoidal.

Un procedimiento para este diagnóstico informal se puede encontrar en Neter (1989, p.p. 594-595) y se resume en los siguientes pasos:

1. Obtener la función $\hat{g}(X)$.
2. Ordenar los valores de la función $\hat{g}(X)$ y su correspondiente valor Y en forma ascendente.
3. Agrupar en clases a los casos con similares valores $\hat{g}(X)$ y aproximadamente el mismo número de casos en cada clase.
4. Obtener el punto medio de los $\hat{g}(X)$ en cada clase.
5. Obtener el valor p_j para cada clase. p_j es la suma de todas las Y=1 de la clase j dividido por el total de casos en la clase, N_j.
6. Graficar el punto medio de las $\hat{g}(X)$ de la clase j contra p_j.
7. La gráfica debe ser monótona y de forma sigmoidal.

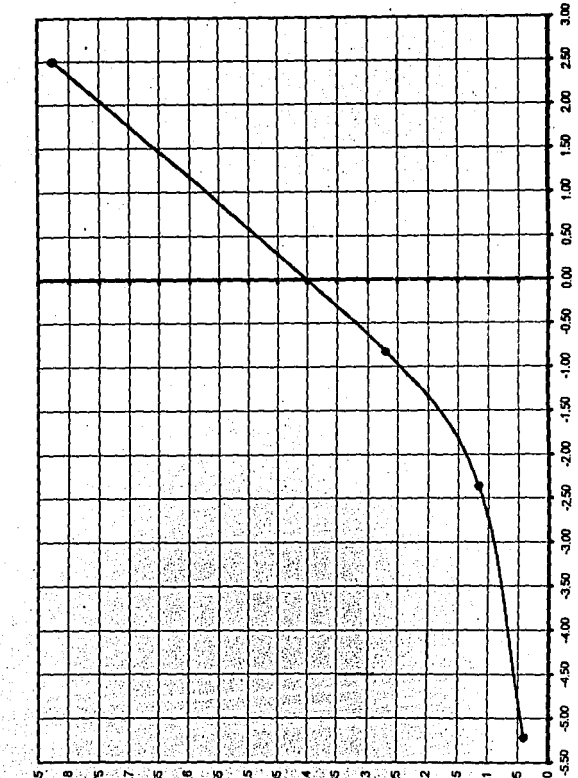
Para el estudio de los tipos de partos en particular se obtuvieron para este diagnóstico la tabla 5.2 y la gráfica 5.1. Como se puede observar en la gráfica el ajuste del modelo es relativamente bueno, pero vale la pena hacer diagnósticos más formales que éste.

TABLA PARA EL DIAGNOSTICO INFORMAL

CLASE	PUNTO MEDIO	P_j	SUMA DE 1'S	N_j
1	-5.214	0.040	1	25
2	-2.357	0.115	3	26
3	-0.830	0.269	7	26
4	2.500	0.826	19	23

Tabla 5.2

GRAFICA PARA EL DIAGNOSTICO INFORMAL



Gráfica 5.1

5.3.2. CHI-CUADRADA DE PEARSON, DEVIANZA Y MATRIZ HAT

Para realizar el diagnóstico del modelo de Regresión Logística se utilizarán algunos estadísticos como son la Chi Cuadrada de Pearson, la Devianza y la matriz Hat. Para definir a éstos es importante hacer notar que es posible que en la muestra recolectada existan observaciones repetidas, por lo que vale la pena utilizar otra notación para el diagnóstico del modelo:

Sea J el número de patrones de las variables predictoras (un patrón es un conjunto de valores particulares de las variables predictoras, esto es, una observación diferente a las demás), $J \leq n$, donde n es el número de observaciones y m_j el número de casos en el patrón j, en este caso, Y_j representa el total de casos con salida igual a uno en el patrón j. El valor ajustado $\hat{Y}_j$ se calcula como:

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

CHI-CUADRADA DE PEARSON

La chi-cuadrada de Pearson está definida como:

$$X^2 = \sum_{j=1}^{J} r\left(Y_j, \hat{\pi}_j\right)^2 \qquad [5.1]$$

donde $r\left(Y_j, \hat{\pi}_j\right)$ se conoce como el j-ésimo residual de Pearson definido como:

$$r\left(Y_j, \hat{\pi}_j\right) = \frac{Y_j - m_j\hat{\pi}_j}{\left[m_j\hat{\pi}_j\,(1-\hat{\pi}_j)\right]^{1/2}} \qquad [5.2]$$

La distribución del estadístico X^2 bajo la hipótesis de que el modelo ajustado es el correcto se asume que sigue una distribución Chi-cuadrada con grados de libertad igual a J-(p+1) donde p+1 es el número de parámetros estimados.

DEVIANZA

La Devianza se calcula como

$$D=\sum_{j=1}^{J} d\left(Y_j,\hat{\pi}_j\right)^2 \qquad [5.3]$$

donde $d\left(Y_j,\hat{\pi}_j\right)$ es conocida como el j-ésimo residual de la Devianza los cuales están definidos como:

$$d\left(Y_j,\hat{\pi}_j\right)=\pm\left\{Y_j \ln\left(\frac{Y_j}{m_j\hat{\pi}_j}\right)+(m_j-Y_j)\ln\left(\frac{m_j-Y_j}{m_j(1-\hat{\pi}_j)}\right)\right\}^{1/2} \qquad [5.4]$$

el signo es el mismo de la diferencia entre el valor observado y el valor pronosticado, $Y_j-m_j\hat{\pi}_j$.

Del mismo modo que la Chi-cuadrada de Pearson, D sigue una distribución Chi cuadrada con J-(p+1) grados de libertad. La Devianza tiene el mismo papel que tiene la suma de cuadrados de los residuales en regresión lineal.

En general se considera que valores pequeños de D y X^2 indican un buen ajuste, sin embargo es más recomendable hacer los diagnósticos utilizando los residuales definidos en las ecuaciones [5.2] y [5.4]. Como puede verse estos estadísticos son residuales, los cuales han sido divididos entre sus errores estándar (a causa de que tienen varianza no constante), aunque no es tan obvio en la Devianza [Hosmer (1989, p. 150)]. Así se espera que bajo la hipótesis de que el modelo de Regresión Logística es correcto las ecuaciones [5.2] y [5.4] deben tener aproximadamente una media igual a cero y una varianza igual a uno. En la práctica, se considera que una observación no está bien ajustada si sus residuales de Pearson y de la Devianza son mayores que 2 en valor absoluto [González de Rivera (1993, p. 109)].

MATRIZ HAT

Un estadístico importante para el diagnóstico del modelo en Regresión Lineal es la matriz Hat y los valores "palanca" (elementos de la diagonal de la matriz Hat) derivados de ella. Esta matriz proporciona, en regresión lineal, valores "palanca" que son proporcionales a la distancia que hay entre los valores de las variables X, para un patrón específico, y la media de todos los datos. La matriz Hat se calcula para Regresión Lineal como:

$H = X(X'X)^{-1}X'$ donde

X es una matriz de Jx(p+1) que contiene los valores para los J patrones de las variables predictoras.

Pregibon (1981) generalizó los diagnósticos utilizados en Regresión Lineal al modelo de Regresión Logística y define a la matriz Hat como:

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

V es una matriz diagonal de JxJ con $v_j = m_j\hat{\pi}(X_j)\left[1-\hat{\pi}(X_j)\right]$.

Los elementos de la matriz Hat son:

$$h_j = m_j\hat{\pi}(X_j)\left[1-\hat{\pi}(X_j)\right](1, X_j')(X'VX)^{-1}(1, X_j')'$$

La suma de todos los elementos de la diagonal de la matriz Hat es igual al número de parámetros estimados y están entre 0 y 1.

En Regresión Lineal se considera que si el elemento h es muy grande ($h_j > 2p/n$) entonces el patrón j puede ser considerado como un punto extremo que perjudica el buen ajuste de el modelo. Sin embargo en Regresión Logística no se sigue la misma regla, la forma de analizar estos valores se verá más adelante.

Si se encuentran observaciones que pueden afectar el buen ajuste del modelo, es necesario tomar la decisión de eliminar o no esa observación de la muestra para después

reestimar los parámetros y ver si el ajuste es mejor que el anterior. Algunos estadísticos que son útiles en este caso son los siguientes:

$$\Delta\hat{\beta}_j = \frac{r_{s,j}^2 h_j}{1-h_j} \qquad [5.5]$$

donde $r_{s,j} = \frac{r_j}{(1-h_j)^{1/2}}$ es el residual de Pearson estandarizado.

$$\Delta X_j^2 = \frac{r_j^2}{1-h_j} = r_{s,j}^2 \qquad [5.6]$$

$$\Delta D_j = \frac{d_j^2}{1-h_j} \qquad [5.7]$$

$\Delta\hat{\beta}_j$ representa el cambio en los coeficientes estimados al borrar el patrón j. De manera similar ΔX_j^2 y ΔD_j representan el decremento en la Chi-cuadrada de Pearson y en la Devianza respectivamente después de borrar el patrón j.

Hosmer (1989, p. 157) proporciona una tabla, la cual es reproducida en la tabla 5.3, en la cual muestra los posibles valores para cada uno de los estadísticos anteriores de acuerdo al valor $\hat{\pi}_j$ (los valores de esta tabla representan lo que se puede esperar en general; mas no lo que debe pasar para un estudio en particular).

	PROBABILIDAD ESTIMADA				
	[0.0-0.1)	[0.1-0.3)	[0.3-0.7)	[0.7-0.9)	[0.9-1.0]
ΔX_j^2	Grande o moderada	Moderada	Moderada a pequeña	Moderada	Grande o pequeña
$\Delta\hat{\beta}_j$	Pequeña	Grande	Moderada	Grande	Pequeña
h_j	Pequeña	Grande	Moderada a pequeña	Grande	Pequeña

tabla 5.3

5.3.3. METODO GRAFICO

El diagnóstico con los estadísticos ΔX_j^2 , ΔD_j y $\Delta \hat{\beta}_j$ se hace de manera general en forma gráfica. Las gráficas de mayor importancia para el diagnóstico son las siguientes:

a) ΔX_j^2 contra $\hat{\pi}_j$

b) ΔD_j contra $\hat{\pi}_j$

La observaciones que son pobremente ajustadas en las dos gráficas anteriores pueden ser identificadas por que caen ya sea en la esquina superior izquierda o en la esquina superior derecha y se encuentran muy retiradas de las demás.

c) $\Delta \hat{\beta}_j$ contra $\hat{\pi}_j$

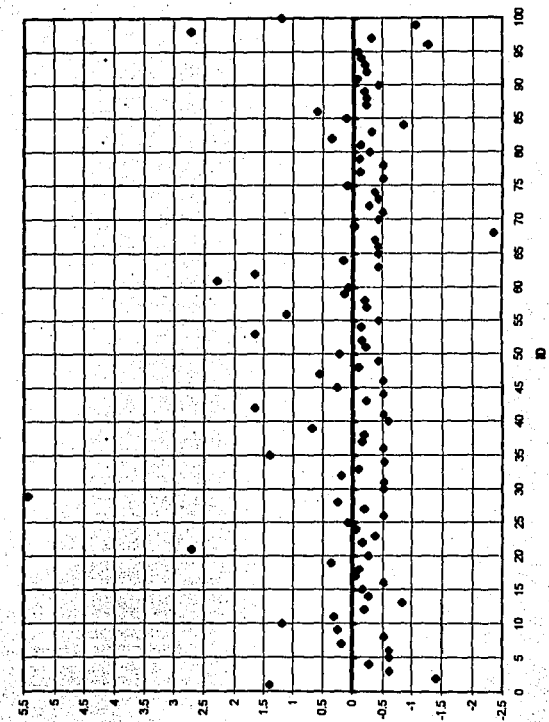
Esta gráfica se puede analizar tomando como mal ajustadas aquellas observaciones que están más alejadas del eje horizontal.

Cabe hacer notar que las observaciones que reflejen mayores valores en las gráficas son solamente observaciones candidatas a salir de la muestra por lo que es necesario posteriormente hacer un análisis de acuerdo a la tabla 5.3 para saber si en verdad es recomendable eliminarlas.

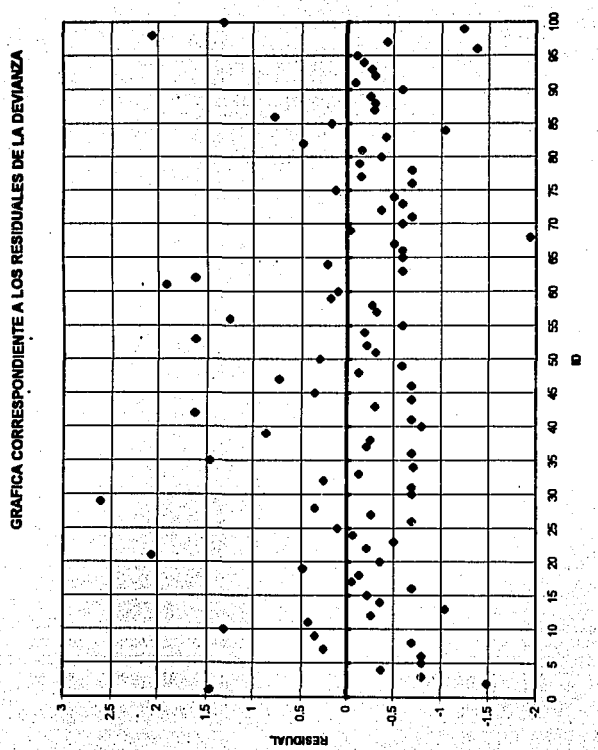
5.4. MODELO FINAL

En el problema que se estudia se puede observar que el número de patrones resulta ser igual al número de observaciones ya que no se encuentra ninguna observación repetida y por consecuencia se tienen 100 patrones y en el caso de que se borre uno o varios patrones se borrarán observaciones.

Para analizar si el modelo del tipo de parto está bien ajustado se realizaron las gráficas 5.2 y 5.3 que corresponde a los residuales de Pearson y de la Devianza respectivamente. Se puede notar en la 5.2 que los residuales para las observaciones 21, 29 y 98 son muy grandes, del mismo modo en la gráfica 5.3 la observaciones 21, 29, 61, 68, y 98 presentan grandes valores en la de los residuales de la devianza por lo que es recomendable analizarlas con mayor detalle.


GRAFICA CORRESPONDIENTE A LOS RESIDUALES DE PEARSON
RESIDUAL
ID
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

Gráfica 5.2


GRAFICA CORRESPONDIENTE A LOS RESIDUALES DE LA DEVIANZA
RESIDUAL
ID
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

Gráfica 5.3

El siguiente paso fue realizar las gráficas descritas en el apartado anterior y resultaron las siguiente conclusiones:

La gráfica 5.4 es la que representa la ΔX_j^2 contra $\hat{\pi}_i$, en ésta se puede observar claramente que sobresalen 4 puntos los cuales pertenecen a 5 observaciones (dos de ellas están sobrepuestas) éstas son la 21, 29, 61, 68, 98.

La gráfica 5.5 es la que representa $\Delta\hat{\beta}_j$ contra $\hat{\pi}_j$, se puede observar que sobresalen la observaciones 2, 29, 68, y 100 puesto que estas sobrepasan los valores de las demás.

La gráfica 5.6 es la que representa ΔD_j contra $\hat{\pi}_j$, en esta gráfica se puede ver que los valores más alejados son los que corresponden a las observaciones 21, 29, 61 68 y 98. En esta gráfica también se puede observar que dos observaciones están sobrepuestas.

Se puede notar que las observaciones que más problemas presentan en todas las gráficas analizadas anteriormente son: 21, 29, 61, 68, y 98, por esto, serán las que se analizarán con la tabla 5.3.

La tabla 5.4 contiene las observaciones 21, 29, 61, 68 y 98 con sus respectivos valores $\hat{\pi}_j$, $\Delta\hat{\beta}_j$, ΔX_j^2 y h_j (la tabla con los residuales y los estadísticos anteriores para todas las observaciones se encuentran en el anexo II).

OBSERVACIONES CANDIDATAS A SER ELIMINADAS					
	$\hat{\pi}_j$	ΔX_j^2	ΔD_j	$\Delta\hat{\beta}_j$	h_j
21	0.120	7.532	4.357	0.207	0.027
29	0.033	30.159	6.976	0.597	0.019
61	0.160	5.387	3.757	0.131	0.024
68	0.847	7.290	4.950	2.330	0.242
98	0.120	7.532	4.357	0.207	0.027

tabla 5.4

GRAFICA DE ΔX² CONTRA π
ΔX²
π
32.5
30
27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
2.5
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1

Gráfica 5.4

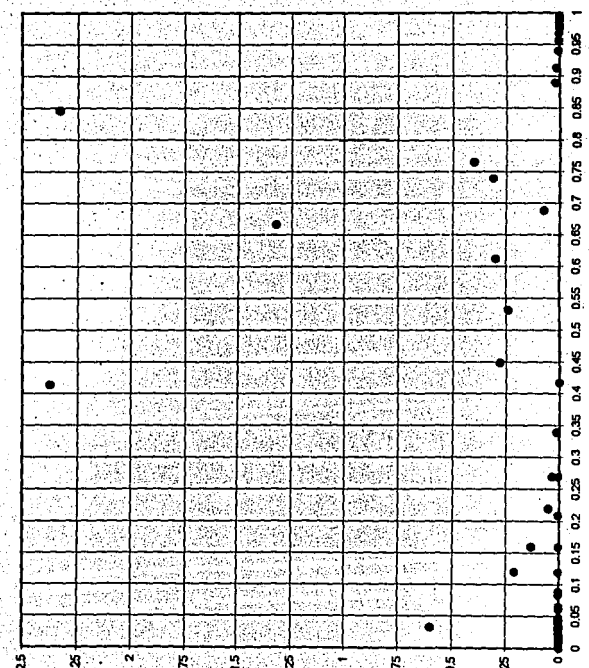
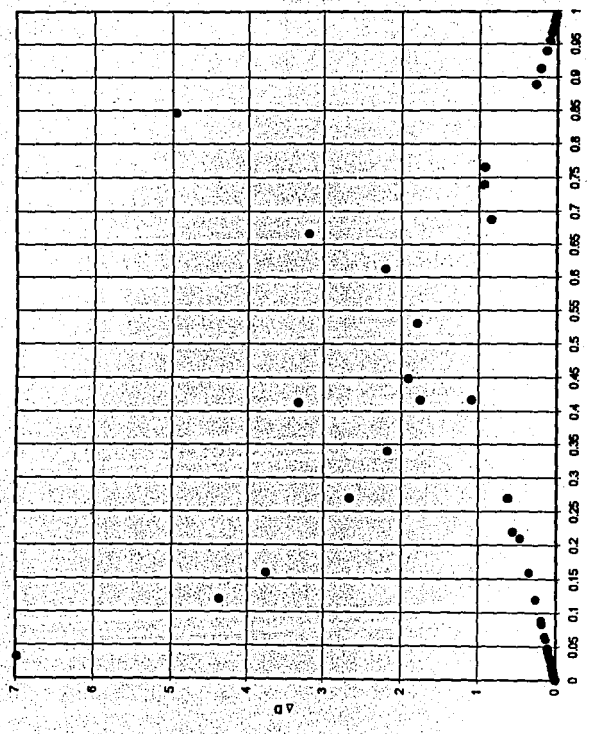

GRAFICA DE Δβ CONTRA π
Δβ
π
2.5
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0.25
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1

Gráfico 5.5


GRAFICA DE ΔD CONTRA π
ΔD
π
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
2
3
4
5
6
7

Gráfica 5.6

Si se analiza la tabla 5.4 se verifica que las observaciones que no cumplen con los supuestos de la tabla 5.3 son la 21, 61, 68 y 98; sin embargo, como se mencionó anteriormente, no se deben tomar como absolutas estas reglas, lo más recomendable es ajustar nuevamente el modelo sin cada una de las observaciones candidatas a ser eliminadas para saber cual es el modelo que más conviene. En la tabla 5.5 se hace una comparación de los siete modelos posibles: el primero contiene todas las observaciones de la muestra, al segundo se le ha eliminado la observación 21, al tercero la 29, al cuarto la 61, al quinto la 68, al sexto la 98 y al último se le han eliminado todas las anteriores.

Variable	Todos	Valor-p	sin 21	Valor-p	sin 29	Valor-p	sin 61	Valor-p
Tpelvis	4.081	0.0004	4.230	0.0003	4.447	0.0004	4.197	0.0003
Dilcer	-0.332	0.0001	-0.373	0.0000	-0.384	0.0001	-0.364	0.0001
Posfeto	2.373	0.0863	2.485	0.0727	2.775	0.0697	2.460	0.0756
Npartos	-1.395	0.0547	-1.316	0.0721	-2.049	0.0468	-1.332	0.0681
Ncesarea	3.183	0.0063	3.220	0.0062	4.123	0.0062	3.211	0.0062
Pearson	87.136		88.742		63.338		89.289	
Devianza	65.656		61.203		58.088		61.854	

Variable	sin 68	Valor-p	sin 98	Valor-p	sin anteriores	Valor-p
Tpelvis	4.246	0.0005	4.230	0.0003	6.301	0.0019
Dilcer	-0.336	0.0001	-0.373	0.0000	-0.588	0.0000
Posfeto	9.724	0.6984	2.485	0.0727	13.315	0.7373
Npartos	-1.783	0.0566	-1.316	0.0721	-3.579	0.0683
Ncesarea	3.681	0.0079	3.220	0.0062	6.772	0.0116
Pearson	94.660		88.742		35.706	
Devianza	59.321		61.203		34.020	

tabla 5.5

Si se observan los valores de los estimadores así como su valor-p se puede notar que el borrar la observación 68 causa problemas en los modelos quinto y último; esto se nota en el valor extremadamente grande del estimador de la variable Posfeto, por lo que no será eliminada de la muestra. De los modelos restantes se puede observar que el que presenta el

mejor ajuste es al que se le ha borrado la observación 29, por lo que se trabajará con este modelo en los siguientes capítulos. Debido a la razón anterior el modelo final es el siguiente:

$$\pi(X) = \frac{exp[g(X)]}{1+exp[g(X)]} \quad \text{donde:}$$

$g(X) = 4.477\ \text{tpelvis} - 0.364\ \text{dilcer} + 2.775\ \text{posfeto} - 2.049\ \text{npartos} + 4.123\ \text{ncesarea}$

CAPITULO VI. OBTENCION E INTERPRETACION DE LOS ODDS RATIOS

6.1. Definición de Odds Ratio (razón de ventajas)

6.1.1. Odds Ratio para el modelo simple

6.1.2. Odds Ratio para el modelo múltiple

6.2. Los Odds Ratio de los parámetros estimados

6.1. DEFINICION DE ODDS RATIO (RAZON DE VENTAJAS)

Si se considera a los términos "enfermedad " como el fenómeno de interés de una investigación el cual tiene dos categorías (ocurre, no ocurre) y a "la exposición" como los valores de las variables que influyen en la ocurrencia o no ocurrencia del fenómeno se tienen las siguientes definiciones de odds ratio:

Last (1989, p. 145) define a los odds ratio (OR) o razón de ventajas como el cociente de la .probabilidad de adquirir la enfermedad si se produce la exposición, dividida por la que existe si no hay exposición.

González de Rivera (1993, p.99) los define como una medida de asociación entre un factor de riesgo teórico y la presencia o desarrollo de una enfermedad.

Los OR son de sumo interés en los estudios epidemiológicos ya que éstos representan una aproximación del riesgo relativo de la presencia del fenómeno de interés, bajo la condición de que la probabilidad de que se presente la enfermedad si se presenta la exposición sea aproximadamente igual a la probabilidad de que se presente la enfermedad si no hay exposición. El OR indica cuanto más probable o frecuente es la presencia de la enfermedad, o fenómeno de interés, dados cada uno de los valores de las variables explicativas [Hosmer, (1989, p. 42)]. Para efectos de esta investigación, la definición que utilizaremos será esta última.

6.1.1. ODDS RATIO PARA EL MODELO SIMPLE

Considérese primero al modelo logístico simple donde la variable predictora es binaria con valores 0 y 1. Se pueden definir a los odds o ventaja como:

$$\frac{\pi(1)}{1-\pi(1)} \quad \text{para } X = 1 \qquad\qquad \frac{\pi(0)}{1-\pi(0)} \quad \text{para } X = 0$$

El log de estas ventajas, de acuerdo a la ecuación [2.2], son:

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

$$\ln\left[\frac{\pi(0)}{1-\pi(0)}\right] = g(0) \qquad \text{para } X = 0$$

El OR, denotado por ψ, para este caso es igual a :

$$\psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \qquad [6.1]$$

$$\psi = \frac{exp[g(1)]}{exp[g(0)]} = exp[g(1) - g(0)]$$

entonces:

$$\ln \psi = g(1) - g(0) \qquad [6.2]$$

La ecuación [6.2] se conoce como la diferencia logit o el log de los OR. Sustituyendo los valores para π(0) y π(1) en la ecuación [6.1] se tiene:

$$\pi(X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$$

$$\psi = \frac{\left[\dfrac{\left(\dfrac{\exp(\beta_0 + \beta_1)}{1+\exp(\beta_0 + \beta_1)}\right)}{\left(1 - \dfrac{\exp(\beta_0 + \beta_1)}{1+\exp(\beta_0 + \beta_1)}\right)}\right]}{\left[\dfrac{\left(\dfrac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)}{\left(1 - \dfrac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)}\right]}$$

$$\psi = \frac{exp(\beta_0 + \beta_1)}{exp(\beta_0)} = exp(\beta_1)$$

La ecuación anterior implica que para el caso simple y cuando la variable predictora es binaria, el OR se calcula con la siguiente expresión:

$$\psi = exp(\beta_1) \qquad [6.3]$$

La ecuación [6.3] representa que tan probable o con que frecuencia se presenta la salida Y=1 cuando se presenta la característica x=1 con respecto a x=0.

Por ejemplo, si se está estudiando la presencia o ausencia de Osteoporosis en una persona dado que sea del sexo masculino o femenino. Este fenómeno se codifica como:

Y = 1	se presenta la Osteoporosis.
Y = 0	no se presenta la Osteoporosis.
X = 1	sexo femenino.
X = 0	sexo masculino.

Supóngase que ψ = 3, esto indica que la Osteoporosis tiene una frecuencia 3 veces mayor de ocurrencia en las mujeres que en los hombres.

Si la variable predictora es de tipo categórico con más de dos niveles se puede extender el método anterior para la obtención de los OR. Como se mencionó en el capítulo II, cuando la variable es categórica con k niveles, es necesario utilizar k-1 variables de diseño para representarla en el modelo; por lo tanto, al realizar el ajuste se obtendrán k-1 parámetros estimados denotados como $\hat{\beta}_{11}, \hat{\beta}_{12}, \ldots, \hat{\beta}_{1(k-1)}$. En este caso la interpretación de los OR se hace en base al contraste del j-ésimo nivel contra el nivel basal (se llama nivel basal al que tiene todas las variables de diseño igual a 0).

Supóngase que la variable X tiene k categorías; las variables de diseño serán las siguientes:

VARIABLES DE DISEÑO

X	D_1	D_2	...	D_{K-1}
NIVEL1	0	0	...	0
NIVEL 2	1	0	...	0
NIVEL 3	0	1	...	0
...	...	...	...	...
NIVEL K	0	0	...	1

Para obtener el OR para el *j*-ésimo nivel con respecto al nivel 1 (lo más común es utilizar siempre al nivel 1 como referencia, aunque es posible utilizar otro nivel), lo primero es obtener la diferencia logit:

$$\ln\left[\hat{\psi}\ (nivel\ j, nivel\ 1)\right] = \hat{g}(nivel\ j) - \hat{g}(nivel\ 1) \qquad [6.4]$$

Utilizando la ecuación [2.5], es posible reescribir la ecuación [6.4] como:

$$\hat{g}(nivel\ j) = \hat{\beta}_0 + \hat{\beta}_{1\,1}(D_1 = 0) + \ldots + \hat{\beta}_{1\,j}(D_1 = 1) + \ldots + \hat{\beta}_{1\,(k-1)}(D_{(k-1)} = 0)$$

$$\hat{g}(nivel\ 1) = \hat{\beta}_0 + \hat{\beta}_{1\,1}(D_1 = 0) + \ldots + \hat{\beta}_{1\,j}(D_1 = 0) + \ldots + \hat{\beta}_{1\,(k-1)}(D_{(k-1)} = 0).$$

Por lo tanto;

$$\ln\left[\hat{\psi}\ (nivel\ j, nivel\ 1)\right] = \hat{g}(nivel\ j) - \hat{g}(nivel\ 1)$$
$$= \hat{\beta}_{1j},$$

por lo que el OR se puede calcular como:

$$\hat{\psi}\ (nivel\ j, nivel\ 1) = exp(\hat{\beta}_{1j}) \qquad [6.5]$$

Este OR indica que cuando se posee la característica X con un nivel j, la salida Y=1 se presenta $exp(\hat{\beta}_{1j})$ veces más que cuando se tiene la característica X a un nivel 1.

Por ejemplo, si se estudia la aparición de enfisema pulmonar en función del número de cajetillas de cigarros fumadas al día representado por X. La codificación de este problema es la siguiente:

Y = 1 se presenta el enfisema pulmonar.
Y = 0 no se presenta el enfisema pulmonar.
X = 0 cero cajetillas al día.
X = 1 una o dos cajetillas al día.
X = 2 tres o más cajetillas al día.

Las variables de diseño serán de la siguiente manera:

X	D_1	D_2
0	0	0
1	1	0
2	0	1

y supóngase que $\hat{\psi}(1,0) = exp(\hat{\beta}_{11}) = 4$; esto indica que el enfisema pulmonar es 4 veces más frecuente entre los que fuman una o dos cajetillas al día que entre los que no fuman. Supóngase que $\hat{\psi}(2,0) = exp(\hat{\beta}_{12}) = 8$; esto indica que el enfisema pulmonar es 8 veces más frecuente entre los que fuman 3 cajetillas de cigarros o más al día que entre los que no fuman.

Para el caso en que la variable predictora es continua, la diferencia logit puede ser expresada como:

$$\ln\left[\hat{\psi}(X+1, X)\right] = \hat{g}(X+1) - \hat{g}(X)$$

Por lo tanto

$$\hat{\psi}(X+1, X) = exp(\hat{\beta}_1) \qquad [6.6]$$

Para este caso, $\hat{\psi}$ indica como aumenta la frecuencia en la ocurrencia de Y=1 ante un incremento unitario de la variable X. Por ejemplo, supóngase que se quiere determinar la probabilidad de que una persona presente una enfermedad del corazón debido a la edad. X denota la edad, Y = 0 no se presenta la enfermedad y Y=1 se presenta la enfermedad. Supóngase que $\hat{\psi}$ = 1.5, esto indica que con cada año de vida la frecuencia con que ocurra una enfermedad de corazón aumenta 1.5 veces.

6.1.2. ODDS RATIO PARA EL MODELO MULTIPLE

La forma de interpretar los OR en un modelo de Regresión Logística múltiple es muy compleja debido a que la diferencia logit está en función de todas las variables predictoras las cuales pueden ser categóricas, continuas o una mezcla de ambas. Lo común en esta situación es analizar el OR de una variable en particular suponiendo a las demás constantes [Friedman (1983, p. 201)]. A continuación se presenta un ejemplo dado en Rey Calero (1989, p. 194) para ilustrar mejor esta situación:

Sea un estudio para determinar la muerte (Y=1) o sobrevivencia (Y=0) de una persona que padece una enfermedad cardiovascular. Las variables predictoras son: edad, tabaco (1 fuma, 0 no fuma) y HTA (alta tensión arterial 1 si y 0 no). El logit estará dado por:

$$g(X) = \beta_0 + \beta_1 Edad + \beta_2 Tabaco + \beta_3 HTA$$

Los estimadores de los parámetros resultaron ser los siguientes:

β_0 = -6614
β_1 = 0.075
β_2 = 0.312
β_3 = 0.018

Por lo tanto la muerte es exp(0.312) = 1.4, veces más frecuente en un fumador con respecto a uno que no fuma.

Es posible también hacer comparaciones tomando en cuenta a dos variables o más, por ejemplo si se desea saber la posibilidad de muerte de una persona de 50 años que fuma respecto a una de 25 que no fuma éste se calcularía como:

exp[0.075(50-25)+0.312(1-0)]=9

Con lo que la muerte es nueve veces más frecuente entre los fumadores de 50 años con respecto a los no fumadores de 25 años.

De este modo se pueden analizar los OR de una manera sencilla aunque no del todo exacta.

6.2. LOS ODDS RATIO DE LOS PARAMETROS ESTIMADOS

Después de obtener los estimadores de los parámetros, el siguiente paso es estimar e interpretar los OR. Los estimadores y sus correspondientes OR se dan en la tabla 6.1.

VARIABLE	ESTIMADOR	ODDS RATIO
Tpelvis	4.477	87.970
Dilcer	-0.364	0.695
Posfeto	2.775	16.039
Npartos	-2.049	0.129
Ncesarea	4.123	61.744

Tabla 6.1

La interpretación de los OR para las variables empleadas en este estudio es la siguiente:

Con respecto al tipo de pelvis, su OR estimado es 88.0 lo cual indica que una cesárea es 88.0 veces más frecuente entre las mujeres que tiene pelvis no útil que entre las que tienen pelvis útil. Como se puede observar la diferencia es muy alta, de ahí la importancia de que una mujer tenga pelvis útil para tratar de evitar una cesárea.

El OR puede tomar tres tipos de valores: entre 0 y 1, indica que la variable está inversamente relacionada al resultado, lo que implica una disminución en la frecuencia con que ocurre el fenómeno de interés al aumentar el valor de la variable; 1, indica que el frecuencia es la misma independientemente del valor de la variable; mayor de 1, indica que la variable está directamente relacionada con el resultado lo que implica un aumento en la frecuencia conforme aumenta el valor de la variable [González de Rivera (1993, p. 105)]. El OR de la dilatación cervical es 0.7, se encuentra en el primer caso, lo cual indica una disminución de la frecuencia con que ocurre una cesárea conforme la dilatación cervical aumenta. A partir de esto se deduce que entre mayor sea la dilatación cervical la frecuencia con la que ocurre una cesárea es menor.

Con respecto a la posición del feto éste tiene un OR de 66.0 lo cual indica que una mujer cuyo producto esté en presentación pélvica tiene 66.0 más posibilidad de tener una cesárea que una mujer cuyo hijo está en presentación cefálica. Al igual que el tipo de pelvis, el OR en este caso es muy alto, por lo cual la posición del feto se puede considerar como esencial para determinar el tipo de parto.

El número de partos anteriores tiene un OR de 0.1, por lo que se encuentra en el mismo caso que la dilatación cervical, esto indica que con cada parto que se tenga la posibilidad de tener una cesárea disminuye.

Un caso especial es el número de cesáreas. Su OR es de 61.7 que es muy alto, esto indicaría que con cada cesárea que se tuviera el siguiente parto tendría una posibilidad 61.7 veces más de tener una cesárea que en el parto anterior. Como se explicó en el capítulo I, la creencia de los médicos de la cesárea iterativa sigue influyendo en la determinación del tipo de parto y aquí se puede corroborar. Sin embargo, es importante que se tenga en cuenta que

esta creencia debe ser abandonada ya que algunos estudios han demostrado que una cesárea no implica que el siguiente parto sea también por cesárea.

Como se ha podido observar, la interpretación de los OR tiene coherencia con lo investigado en el capítulo I; por esto, se podría pensar que el modelo ajustado es útil para el diagnóstico del tipo de parto. Una forma de evaluar la utilidad del modelo es realizar diagnósticos con él y probar su eficacia lo cual se hará en el siguiente capítulo.

CAPITULO VII. PRONOSTICOS

7.1. La elección del punto de corte de probabilidad

7.2. Tabla de clasificación

7.2.1. Sensibilidad y Especificidad

7.3. Pronóstico del tipo de parto

7.1. LA ELECCION DEL PUNTO DE CORTE DE PROBABILIDAD

Si el principal objetivo de la investigación es realizar pronósticos con el modelo de Regresión Logística, o utilizarlo como un auxiliar para hacer diagnósticos del tipo de parto, la primer duda que surgiría es saber a partir de qué valor de $\pi(X)$ se puede determinar si un parto será natural o cesárea. A esta probabilidad se le conoce como *punto de corte*.

Existen varias técnicas para determinar el punto de corte algunas de ellas se pueden encontrar en Neter (1989, p.p. 609-610), éstas son las siguientes:

a) Usar 0.5 como punto de corte, por lo que si $\hat{\pi}_i \geq 0.5$ se pronostica $Y_j = 1$; de lo contrario, se pronostica $Y_j = 0$. Esta regla es válida cuando en la población bajo estudio es igualmente probable que ocurra Y = 0 ó Y = 1 y el costo de predecir incorrectamente cualquiera de las dos respuestas sea el mismo.

b) Encontrar el mejor punto de corte a partir de los datos que se usaron para el ajuste del modelo de Regresión Logística. Esto implica evaluar diferentes puntos de corte y elegir aquel que proporcione la mínima proporción de predicciones incorrectas. Por lo general se considera que el mejor punto de corte se encuentra alrededor de la proporción de salidas Y = 1 que hay en la muestra, por ejemplo, si en la muestra de 99 observaciones se tienen 29 cesáreas (recuérdese que en el capítulo V se tomó la decisión de borrar una de las observaciones), entonces el punto de corte se encuentra alrededor de 0.293.

c) Utilizar probabilidades de la salida Y = 1 en la población y costos de predicción incorrectas conocidas para determinar el mejor punto de corte.

7.2. TABLA DE CLASIFICACION

De las técnicas anteriores la más práctica es la segunda, ya que no en todas las poblaciones ambas salidas tienen la misma ocurrencia y por lo general su probabilidad de ocurrencia es desconocida.

Como lo dice la técnica b), se deben evaluar diferentes puntos de corte para determinar cuál de ellos es el mejor de acuerdo a la proporción de pronósticos correctos que produce. La tabla que contiene esta información es conocida como *tabla de clasificación*.

7.2.1. SENSIBILIDAD Y ESPECIFICIDAD

Al realizar pronósticos hay cuatro situaciones entre lo que se pronostica y lo que se observa en la muestra:

	Observado	Pronosticado
1	Y = 1	Y = 1
2	Y = 1	Y = 0
3	Y = 0	Y = 1
4	Y = 0	Y = 0

Como se puede ver en las situaciones 1 y 4 el pronóstico fue correcto y en la 2 y 3 fueron incorrectos.

Se conoce como *sensibilidad* a la proporción de pronósticos correctos de la salida Y = 1 del total de la muestra y *especificidad* a la proporción de pronósticos correctos de la salida Y=0. Se elige como punto de corte a aquel que tenga aproximadamente la misma sensibilidad y especificidad y que a su vez éstas sean lo más altas posibles.

Además de tomar en cuenta a la sensibilidad y especificidad como criterios para determinar el mejor punto de corte, se puede utilizar una matriz de costos asociada a los pronósticos incorrectos para determinar el costo total que produciría el utilizar un punto de corte determinado. Esta matriz de costos se puede definir de la siguiente manera:

		valores pronosticados	
		Y = 1	Y = 0
valores observados	Y = 1	a	b
	Y = 0	c	d

Así, a representa el costo por pronosticar Y=1 y observarse Y=1, b es el costo por pronosticar Y=0 y observarse Y=1, c es el costo por pronosticar Y=1 y observarse Y=0, d es el costo por pronosticar Y=0 y observarse Y=0. El costo total está dado por la siguiente ecuación:

C = a (suma de las salidas Y=1 pronosticadas correctamente)+
b (suma de las salidas Y=1 pronosticadas incorrectamente)+
c (suma de las salidas Y=0 pronosticadas incorrectamente)+
d (suma de las salidas Y=0 pronosticadas correctamente)

7.3. PRONOSTICO DEL TIPO DE PARTO

El primer paso para poder pronosticar el tipo de parto que tendrá una mujer, según el modelo elegido, es encontrar el mejor punto de corte. La forma de elegir el punto de corte es la siguiente: elegir aquel punto en la tabla de clasificación que tenga la más alta sensibilidad y especificidad y que a su vez tenga el mínimo costo posible (entendiendo como costo al riesgo al que se expone a la madre y su hijo). La matriz de costos para el modelo del tipo de parto es la siguiente:

		valores pronosticados	
		CESAREA	NATURAL
valores observados	CESAREA	0	2
	NATURAL	1	0

La explicación de estos valores es porque si se pronostica correctamente alguna de las salidas no se produce ningún costo, pero en el caso de pronosticar incorrectamente se corre un riesgo: al pronosticar una cesárea dado que la paciente en realidad puede tener un parto normal se tiene un costo de 1 ya que el pronóstico o diagnóstico es incorrecto. Pero si se pronostica un parto natural, dado que la paciente requiere de una cesárea, se corre un riesgo mayor, ya que se pone en peligro tanto la vida de la madre como del niño, es por esto que se le asigna un costo de 2. Estos costos no involucran un valor específico, sólo se asignan para dar un peso de importancia a cada situación.

Al analizar la tabla 7.1, tabla de clasificación, se observa que existen varios puntos de corte que pueden ser considerados como los mejores y para poder analizarlos más detalladamente se puede generar una gráfica en la cual se grafiquen el punto de corte contra la sensibilidad y especificidad (gráfica 7.1). En esta gráfica se observa que existe un rango en el cual se puede considerar que la sensibilidad y especificidad son las más altas y aproximadamente iguales, el rango es entre 0.192 y 0.251 de acuerdo a la tabla 7.1. En este rango el costo total es de 16 el cual es el mínimo, esto indica que cualquiera de estos valores puede ser útil como punto de corte.

El tener todo un rango como punto de corte es poco práctico, por lo que hay que elegir sólo uno de todos los puntos. En este estudio se encontró un modelo de Regresión Logística que calcula la probabilidad de que una mujer tenga una cesárea, por lo que es lógico pensar que a una mujer a la cual el modelo le dé una alta probabilidad será necesario diagnosticarle una cesárea, es por esta razón que es recomendable elegir como punto de corte al punto más grande del rango antes mencionado el cual es 0.251. Utilizando este punto de corte se obtuvieron los siguientes resultados:

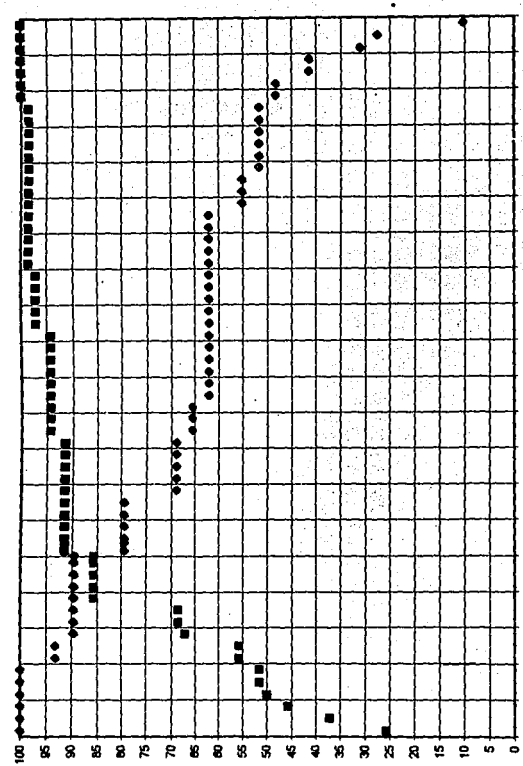
		valores pronosticados		
		CESAREA	NATURAL	
valores observados	CESAREA	26	3	29
	NATURAL	10	60	70
		36	63	

El punto de corte 0.251 proporciona una sensibilidad, porcentaje de cesáreas pronosticadas correctamente, de 89.7 % y una especificidad, porcentaje de partos naturales pronosticados correctamente, de 85.7 %, con un porcentaje total de pronósticos correctos de 86.9 % y un costo total de 16.

TABLA DE CLASIFICACION

P.CORTE	CCESAREA	CNORMAL	CTOTAL	CCESARE	PCNORMAL	PCTOTAL	ICESAREA	INATURAL	ITOTAL	GANANCIA
0.000	29	18	47	100.0	26.7	47.0	0	52	52	0
0.020	29	29	58	100.0	37.1	58.0	0	44	44	0
0.041	29	32	61	100.0	48.7	61.0	0	39	39	0
0.061	29	36	65	100.0	55.0	64.0	0	35	35	0
0.081	29	36	65	100.0	61.4	65.7	0	34	34	0
0.101	29	36	65	100.0	61.4	65.7	0	34	34	0
0.121	27	36	63	93.1	66.7	65.7	2	31	33	0
0.125	27	38	65	93.1	66.7	65.7	2	31	33	0
0.143	26	47	73	90.7	67.1	73.7	3	23	26	0
0.159	26	48	74	90.7	69.8	74.7	3	22	25	0
0.178	26	48	74	90.7	69.8	74.7	3	22	25	0
0.192	26	50	76	90.7	66.7	76.8	3	10	13	0
0.206	26	50	76	90.7	66.7	76.8	3	10	13	0
0.220	26	50	76	90.7	66.7	76.8	3	10	13	0
0.242	26	50	76	90.7	66.7	76.8	3	10	13	0
0.261	26	50	76	90.7	66.7	76.8	3	10	13	0
0.280	23	44	67	78.3	61.4	67.0	6	6	12	0
0.240	23	44	67	78.3	61.4	67.0	6	6	12	0
0.270	23	44	67	78.3	61.4	67.0	6	6	12	0
0.292	23	44	67	78.3	61.4	67.0	6	6	12	0
0.304	23	44	67	78.3	61.4	67.0	6	6	12	0
0.326	23	44	67	78.3	61.4	67.0	6	6	12	0
0.342	20	64	84	69.0	61.4	64.0	9	9	18	0
0.360	20	64	84	69.0	61.4	64.0	9	9	18	0
0.378	20	64	84	69.0	61.4	64.0	9	9	18	0
0.393	20	64	84	69.0	61.4	64.0	9	9	18	0
0.408	20	64	84	69.0	61.4	64.0	9	9	18	0
0.425	19	66	85	65.5	64.5	65.0	10	4	14	0
0.442	19	66	85	65.5	64.5	65.0	10	4	14	0
0.460	19	66	85	65.5	64.5	65.0	10	4	14	0
0.478	19	66	85	62.1	64.5	64.0	11	4	15	0
0.492	18	66	84	62.1	64.5	64.0	11	4	15	0
0.508	18	66	84	62.1	64.5	64.0	11	4	15	0
0.526	18	66	84	62.1	64.5	64.0	11	4	15	0
0.542	18	66	84	62.1	64.5	64.0	11	4	15	0
0.560	18	66	84	62.1	64.5	64.0	11	4	15	0
0.575	18	68	86	62.1	67.1	66.0	11	2	13	0
0.592	18	68	86	62.1	67.1	66.0	11	2	13	0
0.608	18	68	86	62.1	67.1	66.0	11	2	13	0
0.625	18	68	86	62.1	67.1	66.0	11	2	13	0
0.642	18	68	86	62.1	67.1	66.0	11	2	13	0
0.658	18	68	86	62.1	67.1	66.0	11	2	13	0
0.675	18	68	86	62.1	68.8	67.0	11	1	12	0
0.692	18	68	86	62.1	68.8	67.0	11	1	12	0
0.708	18	68	86	62.1	68.8	67.0	11	1	12	0
0.725	18	68	86	62.1	68.8	67.0	11	1	12	0
0.742	18	69	86	58.2	68.8	66.0	12	1	13	0
0.758	18	69	86	58.2	68.8	66.0	12	1	13	0
0.775	18	69	86	58.2	68.8	66.0	12	1	13	0
0.792	18	69	86	51.7	68.8	64.0	14	1	15	0
0.808	18	70	86	51.7	70.6	64.0	14	1	15	0
0.825	18	69	86	51.7	68.8	64.0	14	1	15	0
0.842	18	69	86	51.7	68.8	64.0	14	1	15	0
0.858	18	69	86	51.7	68.8	64.0	14	1	15	0
0.875	18	69	86	51.7	68.8	64.0	14	1	15	0
0.892	14	70	84	48.3	100.0	64.0	15	0	15	0
0.908	14	70	84	48.3	100.0	64.0	15	0	15	0
0.925	12	70	82	41.4	100.0	62.0	17	0	17	0
0.942	12	70	82	41.4	100.0	62.0	17	0	17	0
0.958	8	70	78	31.0	100.0	78.0	20	0	20	0
0.975	8	70	78	27.6	100.0	78.0	21	0	21	0
0.992	3	70	73	10.3	100.0	73.7	26	0	26	0

Tabla 7.1


GRAFICA DE LA SENSIBILIDAD Y LA ESPECIFICIDAD
SENSIBILIDAD Y ESPECIFICIDAD
PUNTO DE CORTE
SENSIBILIDAD
ESPECIFICIDAD

Gráfica 7.1

Después de elegir el punto de corte ya se está en posibilidad de pronosticar, lo cual se hace de acuerdo a los siguientes pasos:

a) Medir las siguientes características en la paciente:

1. Tipo de pelvis: 0 pelvis útil.
   1 pelvis no útil.
2. Dilatación cervical en centímetros.
3. Posición del feto: 0 presentación cefálica.
   1 presentación pélvica.
4. Número de partos anteriores.
5. Número de cesáreas anteriores.

b) Sustituir los datos obtenidos en el modelo final:

$$\pi(X)=\frac{exp[g(X)]}{1+exp[g(X)]} \quad \text{donde:}$$

$$g(X)=4.477\ \text{tpelvis} - 0.364\ \text{dilcer} + 2.775\ \text{posfeto} - 2.049\ \text{npartos} + 4.123\ \text{ncesarea}$$

c) La condición para pronosticar el tipo de parto es la siguiente:

$$\hat{\pi}(X)\begin{cases}\geq 0.251 & \text{PRONOSTICAR CESAREA}\\ < 0.251 & \text{PRONOSTICAR PARTO NATURAL}\end{cases}$$

Una forma equivalente a la anterior es realizar el diagnóstico en base a la función $\hat{g}(X)$, esto porque es más fácil evaluar en una función lineal como $\hat{g}(X)$ que en una no lineal como $\hat{\pi}(X)$. De acuerdo a la transformación logit dada en la ecuación [2.2] se tiene que:

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

sustituyendo el punto de corte se tiene que:

$$\hat{g}(X)=\ln\left[\frac{.251}{1-.251}\right]=-1093$$

Entonces las condiciones para pronosticar el tipo de parto de acuerdo a la función $\hat{g}(X)$ son:

$$\hat{g}(X)\begin{cases} \geq -1.093 & \text{PRONOSTICAR CESAREA} \\ < -1.093 & \text{PRONOSTICAR PARTO NATURAL} \end{cases}$$

Para probar la eficacia del modelo ajustado se obtuvo una muestra adicional de 16 casos los cuales se muestran en el anexo III, 10 de ellos fueron partos naturales y 6 cesáreas. La tabla 7.2 muestra los valores de las cinco variables que se utilizan en el modelo y sus respectivos valores de $\hat{\pi}(X)$ y $\hat{g}(X)$ así como los diagnósticos que generan cada valor de $\hat{\pi}(X)$ y $\hat{g}(X)$ para cada observación.

PRONOSTICOS DEL TIPO DE PARTO PARA LA MUESTRA ADICIONAL

ID	TPARTO	TPELVIS	DILCER	POSPETO	NPARTOS	NCESAREA	π (X)	g(X)	PRONOSTICO
101	NATURAL	0	4	0	0	0	0.189	-1.456	NATURAL
102	NATURAL	0	7	0	3	0	0.000	-8.895	NATURAL
103	CESAREA	1	1	0	0	0	0.984	4.113	CESAREA
104	NATURAL	0	5	0	0	0	0.139	-1.820	NATURAL
105	CESAREA	1	1	0	1	1	0.998	6.187	CESAREA
106	NATURAL	0	3	0	2	0	0.006	-5.190	NATURAL
107	CESAREA	0	2	0	2	0	0.006	-4.826	NATURAL
108	NATURAL	0	6	0	0	0	0.101	-2.184	NATURAL
109	NATURAL	0	6	0	0	0	0.101	-2.184	NATURAL
110	NATURAL	0	4	0	0	0	0.189	-1.456	NATURAL
111	NATURAL	0	3	0	1	0	0.041	-3.141	NATURAL
112	CESAREA	0	3	0	0	0	0.251	-1.092	CESAREA
113	NATURAL	0	6	0	1	0	0.014	-4.233	NATURAL
114	CESAREA	0	4	1	0	0	0.789	1.319	CESAREA
115	CESAREA	1	1	0	2	2	1.000	8.261	CESAREA
116	NATURAL	0	6	0	3	0	0.000	-8.331	NATURAL

Tabla 7.2

Para la muestra adicional la tabla 7.2 dió los siguientes resultados:

valores pronosticados

		CESAREA	NATURAL	
valores observados	CESAREA	5	1	6
	NATURAL	0	10	10
		5	11	

El porcentaje de cesáreas pronosticadas correctamente fue del 83.3 % y el porcentaje de partos naturales pronosticados correctamente fue del 100 %, con un porcentaje total de diagnósticos correctos de 93.8 %.

Esto puede indicar que el modelo es útil para pronosticar el tipo de parto excluyendo aquellos casos que se mencionaron en el capítulo I.

CONCLUSIONES

A lo largo de este trabajo se ha podido ver que los modelos de Regresión Logística son útiles para modelar problemas cuya característica es que la variable de respuesta es de tipo categórica, en especial binaria.

La hipótesis que se manejó en este trabajo fue que la información obtenida de las características de los embarazos de las mujeres bajo estudio se podían ajustar a un modelo de Regresión Logística, el cual debería ser eficaz para diagnosticar el tipo de parto que tendría una mujer en particular. Si se analiza el modelo final, aquel que se le ha borrado la observación 29, se ve que los datos se ajustan lo suficientemente bien al modelo logístico: bajo la hipótesis de que el modelo de regresión logística es el correcto, el valor de la Devianza fue 58.088 con 95 g.l. y un valor-p superior a 0.95, de la misma manera el valor de la Chi-cuadrada de Pearson fue de 63.338 con 95 g.l. y un valor-p también superior a 0.95 lo que hace ver que es un buen ajuste.

Con lo que respecta a los diagnósticos obtenidos en el último capítulo, el modelo pronosticó de una forma eficiente la muestra adicional de 16 casos teniendo un porcentaje de aciertos correctos de 93.8 %. Sin embargo, es importante hacer notar que sería conveniente realizar y evaluar un nuevo modelo el cual contenga información suficiente sobre los casos que fueron excluidos en este estudio, ya que la muestra utilizada tuvo algunas restricciones en ese aspecto.

Otro de los objetivos de la investigación fue el determinar las principales causas que ocasionan una cesárea. En el capítulo IV se encontró que las causas principales fueron: el tipo de pelvis, la posición del feto, la dilatación cervical, el número de partos anteriores y el número de cesáreas anteriores. De las características arriba mencionadas, las que representan un mayor riesgo para tener una cesárea, de acuerdo a los Odds Ratio, son el tipo de pelvis y la posición del feto.

Con lo que respecta al número de cesáreas anteriores, es un tema que se debe tratar con mayor cuidado, ya que como se mencionó anteriormente, en la mayoría de los casos los médicos diagnostican una cesárea cuando la paciente ha tenido cesáreas anteriores, independientemente de un examen con mayor profundidad que determine la posibilidad de un parto normal.

El modelo desarrollado en esta investigación tiene varias limitantes, no es útil en los siguientes casos:

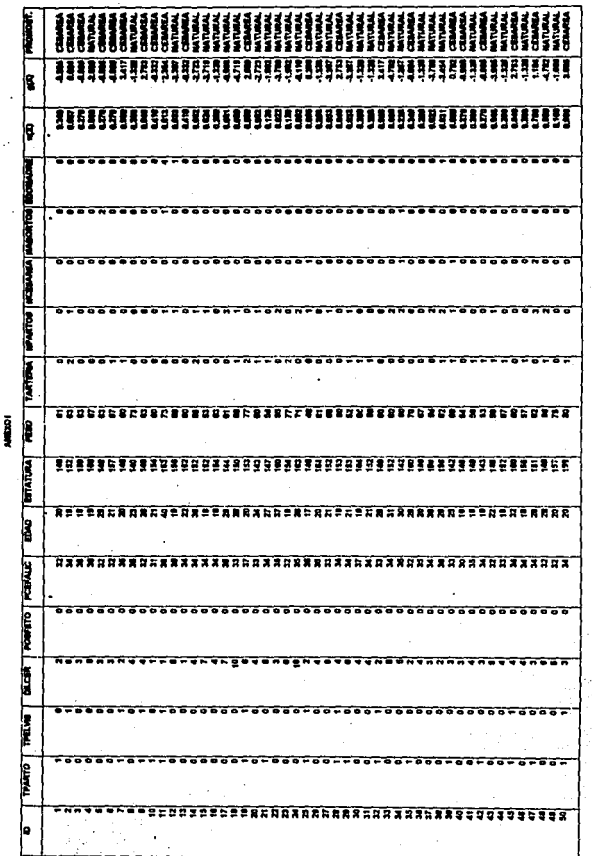
- En el embarazo múltiple.
- En situaciones transversas del feto.
- En problemas que complican el parto tales como cáncer, sida, enfermedades cardiacas, problemas con la placenta etc.
- Partos postérmino y pretérmino.
- Sufrimiento fetal.

La principal causa que motiva estas limitaciones fue la falta de información así como la restricción de acceso a ésta, pero esto no indica que este tipo de características no puedan influir en el tipo de parto, por lo que es necesario hacer estudios más detallados para poder incluir estos factores.

Este trabajo pretende ser un primer intento para considerar algunos modelos probabilísticos en la Ginecoobstetricia

En resumen, los datos de los embarazos de las mujeres sí se pueden ajustar a un modelo de Regresión Logística múltiple, se pueden analizar factores de riesgo que determinan una cesárea como son el tipo de pelvis y la posición del feto, y diagnosticar con el modelo. Por estas razones se puede corroborar la hipótesis que se sustentó al inicio de esta investigación y se considera que los objetivos se cumplieron.

ANEXOS


ANEXO I
MUESTRA
ID
TPARTO
TPELVIS
DILEVI
PONPTO
PCEFALIC
EDAD
ESTATURA
PESO
TARTERIA
EPARTOS
HCESAREA
HABORTOS
ESCOLARI
NCO
GAN
PRESENT.

ANEXO I (Continuación)

ID	TPARTO	TPELVIS	DILCER	POSFETO	PCEFALIC	EDAD	ESTATURA	PESO	TARTERIA	NPARTOS	NCESAREA	NABORTOS	EDOMADRE	π(X)	g(X)	PRONOST.
51	0	0	5	0	33	27	152	74	1	1	0	0	0	0.048	-3.000	NATURAL
52	0	0	3	0	33	38	148	64	0	2	0	0	0	0.022	-3.780	NATURAL
53	1	0	3	0	33	19	151	72	2	0	0	0	0	0.270	-0.998	CESAREA
54	0	0	0	0	33	29	148	80	0	1	0	0	0	0.017	-4.001	NATURAL
55	0	0	5	0	34	20	150	61	0	0	0	0	0	0.100	-1.000	NATURAL
56	1	0	6	0	32	23	146	90	0	1	1	0	0	0.440	-0.204	CESAREA
57	0	0	8	0	36	27	150	88	0	0	0	0	1	0.048	-2.998	NATURAL
58	0	0	10	0	36	18	155	80	0	0	0	0	0	0.030	-3.330	NATURAL
59	1	1	0	0	37	24	160	74	0	0	0	0	0	0.983	4.001	CESAREA
60	1	1	2	0	33	23	160	83	0	1	1	0	0	0.995	5.300	CESAREA
61	1	0	5	0	34	27	152	60	0	0	0	0	0	0.100	-1.000	NATURAL
62	1	0	3	0	36	18	160	111	0	0	0	0	0	0.270	-0.998	CESAREA
63	0	0	8	0	34	18	151	68	0	0	0	0	0	0.100	-1.000	NATURAL
64	1	1	1	0	34	18	160	72	0	0	3	0	0	0.977	3.740	CESAREA
65	0	0	6	0	32	18	150	68	0	0	0	0	0	0.100	-1.000	NATURAL
66	0	0	5	0	34	17	157	83	1	0	0	0	0	0.100	-1.000	NATURAL
67	0	0	8	0	34	17	157	71	0	0	0	0	0	0.120	-1.992	NATURAL
68	0	0	2	1	30	24	160	98	1	0	0	0	0	0.847	1.700	CESAREA
69	0	0	1	0	33	44	156	60	0	5	0	0	0	0.001	-7.307	NATURAL
70	0	0	0	0	36	18	153	80	0	0	0	0	0	0.100	-1.000	NATURAL
71	0	0	4	0	34	18	157	64	1	0	0	0	0	0.200	-1.330	NATURAL
72	0	0	6	0	33	20	158	83	0	0	0	0	0	0.008	-2.000	NATURAL
73	0	0	5	0	36	18	158	80	1	0	0	0	0	0.100	-1.000	NATURAL
74	0	0	8	0	34	27	162	82	2	0	0	1	0	0.120	-1.992	NATURAL
75	1	1	3	0	34	18	147	80	1	1	1	0	0	0.992	4.573	CESAREA
76	0	0	4	0	36	20	158	68	1	0	0	0	0	0.200	-1.330	NATURAL
77	0	0	5	0	34	23	154	88	1	2	0	0	0	0.012	-4.400	NATURAL
78	0	0	4	0	33	18	157	68	0	0	0	0	0	0.200	-1.330	NATURAL
79	0	0	10	0	33	19	162	86	0	1	0	0	0	0.009	-4.715	NATURAL
80	0	0	8	0	33	22	151	60	0	0	0	0	0	0.080	-2.000	NATURAL
81	0	0	0	0	34	21	154	72	0	1	0	0	0	0.012	-4.383	NATURAL
82	1	1	0	0	33	22	140	60	1	0	0	0	0	0.999	7.000	CESAREA
83	0	0	3	0	33	18	148	51	0	1	0	0	0	0.088	-2.301	NATURAL
84	0	0	1	0	36	18	163	80	1	0	0	0	0	0.418	-0.332	CESAREA
85	1	1	5	0	36	25	140	83	0	1	1	1	0	0.986	4.280	CESAREA
86	1	0	4	1	37	18	154	63	2	0	0	0	0	0.740	1.046	CESAREA
87	0	0	1	0	33	28	157	73	0	2	0	1	0	0.042	-3.122	NATURAL
88	0	0	6	0	31	28	146	64	1	1	0	0	0	0.046	-3.008	NATURAL
89	0	0	2	0	33	22	158	82	0	2	0	0	0	0.031	-3.404	NATURAL
90	0	0	5	0	34	23	160	73	0	0	0	2	0	0.100	-1.000	NATURAL
91	0	0	4	0	37	29	144	96	0	3	0	0	0	0.004	-5.813	NATURAL
92	0	0	8	0	31	16	148	80	0	1	0	0	0	0.046	-3.008	NATURAL
93	0	0	10	0	34	32	163	73	0	0	0	0	1	0.035	-3.330	NATURAL
94	0	0	4	0	36	32	148	73	2	2	0	0	0	0.016	-4.118	NATURAL
95	0	0	7	0	36	23	140	98	0	2	0	0	0	0.000	-8.114	NATURAL
96	0	0	4	0	32	19	148	90	1	1	1	0	0	0.613	0.400	CESAREA
97	0	0	7	0	38	18	163	68	0	0	0	0	0	0.098	-2.334	NATURAL
98	1	0	8	0	36	19	143	68	0	0	0	0	0	0.120	-1.992	NATURAL
99	0	0	5	0	34	31	140	98	2	1	1	0	0	0.532	0.120	CESAREA
100	1	0	4	1	34	23	153	67	0	1	0	0	0	0.413	-0.380	CESAREA

MUESTRA

ANEXO II

ID	R.PEARSON	R.DEVIANZA	MAT.HAT	ELTA CHI-C	DELTA DEV.	DELTA BETA
1	1.394	1.469	0.006	1.955	2.172	0.012
2	-1.415	-1.483	0.312	2.909	3.195	1.319
3	-0.608	-0.793	0.013	0.374	0.637	0.005
4	-0.265	-0.368	0.028	0.072	0.140	0.002
5	-0.608	-0.793	0.013	0.374	0.637	0.005
6	-0.608	-0.793	0.013	0.374	0.637	0.005
7	0.181	0.254	0.039	0.034	0.067	0.001
8	-0.515	-0.686	0.019	0.270	0.479	0.005
9	0.253	0.352	0.071	0.069	0.133	0.005
10	1.181	1.321	0.002	1.396	1.749	0.002
11	0.308	0.426	0.122	0.108	0.207	0.015
12	-0.184	-0.258	0.019	0.034	0.068	0.001
13	-0.847	-1.040	0.002	0.719	1.084	0.001
14	-0.256	-0.357	0.031	0.068	0.131	0.002
15	-0.156	-0.219	0.016	0.025	0.049	0.000
16	-0.515	-0.686	0.019	0.270	0.479	0.005
17	-0.039	-0.055	0.007	0.002	0.003	0.000
18	-0.095	-0.134	0.008	0.009	0.018	0.000
19	0.352	0.483	0.127	0.142	0.267	0.021
20	-0.256	-0.357	0.031	0.068	0.131	0.002
21	2.707	2.059	0.027	7.532	4.357	0.207
22	-0.151	-0.212	0.044	0.024	0.047	0.001
23	-0.369	-0.506	0.027	0.140	0.263	0.004
24	-0.047	-0.067	0.005	0.002	0.004	0.000
25	0.074	0.105	0.011	0.006	0.011	0.000
26	-0.515	-0.686	0.019	0.270	0.479	0.005
27	-0.184	-0.258	0.019	0.034	0.068	0.001
28	0.253	0.352	0.071	0.069	0.133	0.005
29	5.438	2.615	0.019	30.159	6.976	0.597
30	-0.515	-0.686	0.019	0.270	0.479	0.005
31	-0.515	-0.686	0.019	0.270	0.479	0.005
32	0.181	0.254	0.039	0.034	0.067	0.001
33	-0.092	-0.129	0.017	0.009	0.017	0.000
34	-0.531	-0.704	0.137	0.328	0.575	0.052
35	1.394	1.469	0.006	1.955	2.172	0.012
36	-0.515	-0.686	0.019	0.270	0.479	0.005
37	-0.151	-0.212	0.044	0.024	0.047	0.001
38	-0.178	-0.250	0.060	0.034	0.066	0.002
39	0.673	0.865	0.123	0.517	0.852	0.073
40	-0.608	-0.793	0.013	0.374	0.637	0.005
41	-0.515	-0.686	0.019	0.270	0.479	0.005
42	1.645	1.619	0.013	2.742	2.654	0.035
43	-0.217	-0.304	0.024	0.048	0.094	0.001
44	-0.515	-0.686	0.019	0.270	0.479	0.005
45	0.253	0.352	0.071	0.069	0.133	0.005
46	-0.515	-0.686	0.019	0.270	0.479	0.005
47	0.553	0.731	0.428	0.535	0.933	0.400
48	-0.092	-0.129	0.017	0.009	0.017	0.000
49	-0.436	-0.590	0.024	0.195	0.357	0.005
50	0.214	0.299	0.052	0.048	0.094	0.003

ESTADISTICOS PARA EL DIAGNOSTICO

ANEXO II (Continuación)

ID	R.PEARSON	R.DEVIANZA	MAT.HAT	ELTA CHI-C	DELTA DEV.	DELTA BETA
51	-0.217	-0.304	0.024	0.048	0.094	0.001
52	-0.151	-0.212	0.044	0.024	0.047	0.001
53	1.645	1.619	0.013	2.742	2.654	0.035
54	-0.132	-0.186	0.013	0.018	0.035	0.000
55	-0.436	-0.590	0.024	0.195	0.357	0.005
56	1.108	1.265	0.160	1.460	1.906	0.278
57	-0.225	-0.314	0.026	0.052	0.101	0.001
58	-0.190	-0.267	0.024	0.037	0.073	0.001
59	0.130	0.183	0.022	0.017	0.034	0.000
60	0.074	0.105	0.011	0.006	0.011	0.000
61	2.293	1.915	0.024	5.387	3.757	0.131
62	1.645	1.619	0.013	2.742	2.654	0.035
63	-0.436	-0.590	0.024	0.195	0.357	0.005
64	0.153	0.216	0.029	0.024	0.048	0.001
65	-0.436	-0.590	0.024	0.195	0.357	0.005
66	-0.436	-0.590	0.024	0.195	0.357	0.005
67	-0.369	-0.506	0.027	0.140	0.263	0.004
68	-2.350	-1.937	0.242	7.290	4.950	2.330
69	-0.026	-0.037	0.009	0.001	0.001	0.000
70	-0.436	-0.590	0.024	0.195	0.357	0.005
71	-0.515	-0.686	0.019	0.270	0.479	0.005
72	-0.265	-0.368	0.026	0.072	0.140	0.002
73	-0.436	-0.590	0.024	0.195	0.357	0.005
74	-0.369	-0.506	0.027	0.140	0.263	0.004
75	0.088	0.124	0.015	0.008	0.015	0.000
76	-0.515	-0.686	0.019	0.270	0.479	0.005
77	-0.108	-0.152	0.023	0.012	0.024	0.000
78	-0.515	-0.686	0.019	0.270	0.479	0.005
79	-0.095	-0.134	0.008	0.009	0.018	0.000
80	-0.265	-0.368	0.026	0.072	0.140	0.002
81	-0.112	-0.158	0.010	0.013	0.025	0.000
82	0.352	0.483	0.127	0.142	0.267	0.021
83	-0.303	-0.419	0.039	0.095	0.182	0.004
84	-0.847	-1.040	0.002	0.719	1.084	0.001
85	0.122	0.172	0.027	0.015	0.030	0.000
86	0.593	0.776	0.361	0.550	0.943	0.311
87	-0.210	-0.294	0.083	0.048	0.094	0.004
88	-0.217	-0.304	0.024	0.048	0.094	0.001
89	-0.178	-0.250	0.060	0.034	0.066	0.002
90	-0.436	-0.590	0.024	0.195	0.357	0.005
91	-0.064	-0.090	0.018	0.004	0.008	0.000
92	-0.217	-0.304	0.024	0.048	0.094	0.001
93	-0.190	-0.267	0.024	0.037	0.073	0.001
94	-0.128	-0.180	0.032	0.017	0.033	0.001
95	-0.078	-0.110	0.012	0.006	0.012	0.000
96	-1.258	-1.378	0.139	1.838	2.204	0.295
97	-0.313	-0.432	0.028	0.101	0.192	0.003
98	2.707	2.059	0.027	7.532	4.357	0.207
99	-1.066	-1.232	0.151	1.339	1.789	0.239
100	1.191	1.329	0.470	2.675	3.331	2.368

ESTADISTICOS PARA EL DIAGNOSTICO

ANEXO III

ID	TPARTO	TPELVIS	DILCER	POSFETO	PCEFALIC	EDAD	ESTATURA	PESO	TARTERIA	NPARTOS	NCESAREA	NABORTOS	EDOMADRE
101	0	0	4	0	35	20	156	69	1	0	0	0	0
102	0	0	7	0	36	32	148	61	0	3	0	0	0
104	0	0	5	0	33	23	152	69	0	0	0	0	0
106	0	0	3	0	36	23	152	72	0	2	0	0	0
108	0	0	6	0	33	18	156	68	1	0	0	0	0
109	0	0	6	0	34	19	159	52	0	0	0	0	0
110	0	0	4	0	36	18	147	54	1	0	0	0	0
111	0	0	3	0	31	30	155	66	0	1	0	0	0
113	0	0	6	0	37	21	162	62	0	1	0	0	0
116	0	0	6	0	35	37	168	75	0	3	0	1	0
103	1	1	1	0	33	22	157	72	1	0	0	0	0
105	1	1	1	0	33	19	152	72	0	1	1	0	0
107	1	0	2	0	35	34	146	70	1	2	0	0	0
112	1	0	3	0	35	19	160	60	2	0	0	0	0
114	1	0	4	1	37	19	154	63	2	0	0	0	0
115	1	1	1	0	34	31	153	66	1	2	2	0	0

MUESTRA ADICIONAL

ANEXO IV

```
/PROBLEM TIME = 120.

/INPUT VARIABLES = 14.
       FORMAT IS FREE.
       FILE = DATOS.DAT

/VARIABLE NAME = ID, TPARTO, TPELVIS, DILCER, POSFETO,
                 PCEFALIC, EDAD, ESTATURA, PESO,
                 TARTERIA, NPARTOS, NCESAREA,
                 NABORTOS, EDOMADRE.

USE = TPARTO TO EDOMADRE.

/GROUP   CODES(TPARTO) = 1, 0.
         NAMES(TPARTO) = CESAREA, NORMAL.

/REGRESS DEPENDENT = TPARTO.
         INTERVAL = DILCER, PCEFALIC, EDAD, ESTATURA,
                    PESO, NPARTOS, NCESAREA, NABORTOS.
         MODEL = TPELVIS, DILCER, POSFETO, PCEFALIC,
                 EDAD, ESTATURA, PESO, TARTERIA,
                 NPARTOS, NCESAREA, NABORTOS, EDOMADRE,
         START = IN, IN, IN, OUT, OUT, OUT, OUT, OUT,
                 OUT, OUT, OUT, OUT.
         MOVE = 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2.
         CMOVE = 2.
         ENTER = 0.2, 0.2.
         REMOVE = 0.2, 0.2.
         METHOD = ACE.
         ITERATION = 25.

/PRINT   LINESIZE = 75.
       CELL = BOTH.
       COST.
       COVA.
       PLOT.
       HIST.

/END
```

PROGRAMA BMDP

ANEXO IV (Continuación)

```
DATA SASUSER.PARTOS;
INFILE 'A:\DATOS.DAT';

RUN;

TITLE3 'REGRESION POR PASOS PARA PARTOS';
PROC LOGISTIC ORDER=DATA OUTEST=ESTIMADO COVOUT;
              MODEL TPARTO = TPELVIS DILCER POSFETO PCEFALIC EDAD
                             ESTATURA PESO TARTERIA NPARTOS NCESAREA
                             NABORTOS EDOMADRE
               /NOINT
               SELECTION = STEPWISE
               DETAILS
               FAST
               INCLUDE= 0
               SLENTRY= 0.20
               SLSTAY = 0.20
               CORRB
               CTABLE
               INFLUENCE
               IPLOTS;
              OUTPUT OUT=DIAGNOS H=HAT LOWER=INF P=PROB
                     UPPER=SUP RESCHI=RPEARSON RESDEV=RDEVIANC;
RUN;

PROC PRINT DATA=ESTIMADO;
     TITLE3 'ESTIMADORES Y MATRIZ DE COVARIANZAS';
RUN;

PROC PRINT DATA=DIAGNOS;
     TITLE3 'PROBABILIDADES PREDECIDAS';
RUN;
```

PROGRAMA SAS

LISTA DE SIMBOLOS

Y	Variable de Respuesta.
X_i	i-ésima variable predictora.
$E(Y/X)$	Valor esperado de Y dado el valor X.
β_0	Término constante del modelo de Regresión Logística.
β_i	Coeficiente de la i-ésima variable predictora del modelo de Regresión Logística.
$\pi(X)$	Probabilidad condicional del modelo de Regresión Logística.
$g(X)$	Transformación Logit.
p	Número de variables predictoras en el modelo de Regresión Logística.
k_i	Niveles o categorías de la i-ésima variable categórica.
D_{ju}	Variable de diseño de la j-ésima variable predictora, u = 1,2,...,k -1.
β_{ju}	Coeficiente de D_{ju} .
n	Número de observaciones de la muestra.
ε_j	Error aleatorio de la j-ésima observación, j=1,2,...,n.
$N(\mu, \sigma^2)$	Distribución Normal con media μ y varianza σ^2.
S	Suma del cuadrado de las diferencias entre valores observados y estimados de la variable de respuesta.
$\hat{\beta}_i$	Estimador del coeficiente de la i-ésima variable predictora.
$\hat{Y}_j$	Estimador de la variable de respuesta para la j-ésima observación.
L	Función de Verosimilitud.
$G(\theta)$	Vector Gradiente.
$H(\theta)$	Matriz Hessiana.
D	Devianza.
G	Estadístico de la prueba de razón de Verosimilitud.
Valor-p	Mínimo valor de significación, α , para el cual los datos observados indican que se tendría que rechazar la hipótesis nula.
W	Estadístico de Wald.
H_0	Hipótesis nula.
Ha	Hipótesis alternativa.
$S(\hat{\beta}_i)$	Error o desviación estándar del estimador del i-ésimo parámetro.

$\hat{\pi}(X)$	Estimador de la probabilidad condicional del modelo de Regresión Logística.
$\chi^2(\nu)$	Distribución Chi-cuadrada con n grados de libertad.
PE	Nivel de significancia para que una variable ingrese al modelo.
PR	Nivel de significancia para que una variable salga del modelo.
$L^{(h)}_{el\ ej-}$	Log verosimilitud del modelo que contiene a X_{e1} , X_{ej} ,... al paso h.
$G^{(h)}_{el\ ej-}$	Estadístico G al paso h para el modelo que contiene la variable X_{e1} , X_{ej} ,...
$\hat{g}(X)$	Estimador de la transformación Logit.
J	Número de patrones en la muestra de n observaciones, J ≤ n.
m_j	Número de observaciones en el patrón j.
$r(Y_j, \hat{\pi}_j)$	Residuales de Pearson.
X^2	Chi-cuadrad de Pearson.
$d(Y_j, \hat{\pi}_j)$	Residuales de la Devianza.
H	Matriz Hat.
h_j	Elemento j de la matriz Hat.
r_{sj}	Residual de Pearson estandarizado.
$\Delta\hat{\beta}_j$	Cambio en los coeficientes estimados después de borrar el patrón j.
ΔX^2_j	Cambio en la Chi-cuadrada de Pearson después de borrar el patrón j.
ΔD_j	Cambio en la Devianza después de borrar el patrón j.
ψ	Odds Ratio, Razón de ventajas.
a	Costo por pronosticar Y=1 dado que se observó Y=1.
b	Costo por pronosticar Y=0 dado que se observó Y=1.
c	Costo por pronosticar Y=1 dado que se observó Y=0.
d	Costo por pronosticar Y=0 dado que se observó Y=0.
C	Costo total de pronóstico.

GLOSARIO

MEDICO

Borramiento	Dilatación total del cervix.
Cervix	Cuello de la matriz.
Cesárea	Es la intervención quirúrgica que tiene por objeto extraer al producto de la concepción a través de la incisión abdominal y uterina.
Dilatación cervical	Indica los centímetros que está dilatado el cervix o cuello de la matriz.
Embarazo	El desarrollo de un ser en el útero de una mujer desde la fecundación hasta el nacimiento.
Fecundación	Unión del óvulo con el espermatozoide.
Feto	Producto del desarrollo del embrión en su último periodo.
Fórceps	Instrumento en forma de tenazas utilizado en los partos para auxiliar al producto a nacer.
Obstetricia	Parte de la medicina que trata del embarazo, el parto y el puerperio.
Parto	Expulsión del producto cuando el embarazo tiene más de 22 semanas y el feto pesa más de 500 gramos.
Parto distócico	También conocido como parto anormal es aquel que presenta problemas en sus fases del proceso natural, tanto por parte de la madre como del feto.
Parto eutócico	También conocido como parto normal es quel que no presenta complicaciones durante el mecanismo del parto.

Parto natural	En este estudio se considera a un parto natural aquel que se efectúa por vía vaginal.
Pelvis	La parte más inferior del tronco. Se utiliza para designar el armazón óseo compuesto por el sacro y los dos huesos ilíacos.
Perímetro cefálico	Es el perímetro de la cabeza del feto.
Placenta	Organo del aporte sanguíneo y nutricio de la madre al feto al mismo tiempo que elimina el anhídrido carbónico y otros productos de desecho del feto.
Presentación	Indica la parte del niño que se prepara a salir primero durante el parto.
Presentación cefálica	El feto se prepara a nacer de cabeza.
Presentación pélvica	El feto se prepara a nacer de pies o de nalgas.
Primigesta	Mujer que tiene su primera gestación.
Primigrávida	Mujer que tiene su primer parto.
Prolapso del cordón	Consiste cuando el cordón umbilical se encuentra por debajo del feto con membranas rotas.
Puerperio	Tiempo que sigue al tercer periodo de parto. Está caracterizado por la involución de los órganos reproductores a su estado de preembarazo.
Raquitismo	Mala alimentación, falta de minerales o vitaminas.
Situación	Es la relación que existe entre la columna vertebral de la madre y del producto.
Situación longitudinal	La columna vertebral del feto es paralela a la de la madre

Situación transversal La columna vertebral del feto es perpendicular a la de la madre.

Sufrimiento fetal Estado crítico del feto debido a una alteración en el intercambio metabólico entre éste y la madre.

Toxemia Serie de complicaciones del embarazo que aparecen generalmente al término del mismo.

Trabajo de parto Serie de procesos mediante los cuales la madre expulsa al producto viable de la gestación a través de las vías genitales.

MATEMATICO

Estimador consistente Un estimador es consistente si al incrementar el tamaño de la muestra, n, disminuye la varianza.

Estimador eficiente Un estimador es eficiente si su varianza es mínima.

Estimador insesgado Un estimador $\hat{\theta}$ es insesgado si $E(\hat{\theta}) = \theta$,ésto es, que la esperanza del estimador es igual al parámetro.

Estimador suficiente Un estimador es suficiente si utiliza la mayor cantidad de información contenida en la muestra y el parámetro no depende de ésta.

Logit Transformación mediante la cual se linealiza el modelo de Regresión Logística.

Máxima verosimilitud Es el método de estimación que tiene como propósito encontrar estimadores $\hat{\theta}$ que hagan máxima la probabilidad de que los valores estimados, $\hat{Y}$, a partir de los estimadores $\hat{\theta}$ y de la matriz de observaciones de las variables predictoras, X, sea lo más parecidos posibles a los valores observados de Y.

Mínimos cuadrados	**Método de estimación que tiene como propósito seleccionar estimadores $\hat{\theta}$ tal que los errores de predicción de la variable de respuesta sean mínimos. Una variante es el método de mínimos cuadrados ponderados en el cual el cuadrado de las diferencias entre el valor pronosticado y el estimado es multiplicado por un peso o ponderación.**
Modelo de Regresión	**Modelo que describe la relación entre una variable de respuesta y una o más variables predictoras.**
Parámetro	**Medida numérica que describe una característica específica de una población.**
Parsimonia	**Al construir un modelo debe tratarse de elegir el más sencillo de todos los posibles, cuidando que éste sea capaz de describir a la variable de respuesta.**
Regresión Lineal	**Modelo de regresión en el que se supone que la variable de respuesta y las variables predictoras tienen una relación lineal.**
Regresión Logística	**Modelo de regresión en el que se supone que la variable de respuesta y las variables predictoras tienen una relación descrita por la función Logística. Se caracteriza porque la variable de respuesta es categórica, principalmente binaria.**
Variable categórica	**Variable que solamente puede tomar valores discretos.**
Variables de diseño	**Se utilizan para representar una variable categórica en el modelo de Regresión Logística.**
Variables predictora	Son aquellas que son capaces de describir la variable de respuesta.
Variable respuesta	También llamada variable dependiente, es aquella que representa los posibles valores del fenómeno de interés.

BIBLIOGRAFIA

BIBLIOGRAFIA BASICA

Canavos, G. C. *Probabilidad y Estadística. Aplicaciones y Métodos.* México, Mc Graw Hill, 1988, p.p. 200, 251-268, 443-448.

Dixon, W. *BMDP statistical Software.* Berkley, University of California, 1990. p.p. 1013-1046

Escobar Habeica, R. *Eutocia después de cesárea (Tesis de posgrado: especialista de Ginecología y Obstetricia).* Facultad de Medicina, U.N.A.M., 1990.

Friedman, G. *Primer of Epidemiology, Third Edition.* New York, Mc.Graw Hill, 1983. p. p. 106-107, 200-205.

González-Merlo, J. *Avances en Ginecología y Obstetricia.* Barcelona, Salvat, 1980. p.p. 15-17, 28-29, 97-99.

González de Rivera, J. L., Rodríguez, F., Sierra, A. *El método Epidemiológico en Salud Mental.* Barcelona, Masson-Salvat, 1993. p. p. 99-115.

Greenland, S. "Modelling variable Selection in Epidemiologic analysis". *American Journal of Public Health.* Vol. 79, 1989, p.p. 340-349.

Guillén, M. *Análisis de Regresión Múltiple.* Madrid, Centro de Investigaciones Sociológicas, 1992. p. p. 8-67

Hosmer, D. W., Wang, C. Y., Lemeshow, S. "A computer program for stepwise logistic regression using maximum likelihood". *Computer Programs in Biomedicine.* Vol 8, 1978, p.p. 121-134.

Hosmer, D. W., Lemeshow, S. *Applied Logistic Regression.* New York, Wiley, 1989, p.p. 1-175

Landwehr , J. P., Pregibon, D., Shoemaker, A. C. "Graphical Methods for Assessing Logistic Regression Models". *Journal of the American Statistical Association.* Vol. 79, p. p. 61-71.

Last, J. M. *Diccionario de Epidemiología.* México, Salvat Editores, 1989. p.p. 136-139, 144-147, 152-155.

Mendenhall, W., Scheaffer, R., Wackerly, D. *Estadística Matemática con aplicaciones.* México D. F., Grupo Editorial Iberoamericana, 1986. p.p. 297-344, 381-502.

Méndez, I., Namihira, D., Moreno, L., Sosa, C. *El protocolo de la investigación, lineamientos para su elaboración y análisis, 2a. ed.*México D.F., Trillas, 1994. p.p. 11-27, 85-108

Mendoza Aréstegui, I. *Ginecoobstetricia, guías.* México, D.F., Manual Moderno, 1992. p. p. 118-288.

Neter, J., Wasserman, W., Kutner, M. *Applied Linear Regression Models.* Boston, Irwin, 1989. p. p. 578-616.

Neville, F. *Compendio de Ginecología y Obstetricia.* México, Interamericana - Mc Graw Hill, 1989. p. p. 269-272.

Peduzzi, P. N., Hardy, R. J., Holford, T.R. "A Stepwise Variable Selection procedure for nonlinear Regresion Models". *Biometrics.* Vol. 36, p. p. 511-516.

Peña Sánchez de Rivera, D. *Estadística, Modelos y Métodos.* Vol. 2. México, 1987. p. p. 452-461, 584-587.

Pregibon, D. "Logistic Regression Diagnostics". *Annals of Statistics.* Vol. 9, 1981, p. p. 705-724.

Rey Calero, Juan. *Método epidemiológico y salud de la comunidad.* Madrid, McGraw-Hill-Interamericana de España, 1989. p. p. 191-198.

SAS Institute Inc. *SAS Guide for Personal Computers, version 6.03.* Cary, N. C., SAS Institute Inc., 1988. p. p. 1071-1126.

Tourris, H. *Manual de Ginecología y Obstetricia.* Barcelona, Toray-Masson, 1974. p. p. 221-255.

Wilson, J. *Ginecoobstetricia.* México, D.F., El Manual Moderno, 1991. p. p. 416-423, 524-537, 572-578,

BIBLIOGRAFIA COMPLEMENTARIA

Aguirre Hernández, R. *Técnicas de Diagnóstico para el modelo Logístico (Tesis de maestría: Estadística e Investigación de Operaciones).* UACP y P del CCH, U.N.A.M., 1991.

Armitage, P., Berry, G. *Estadística para la investigación Biomédica.* Barcelona, Doyma, 1992, p.p. 452-466, 536-539.

Breslow, N. E., and Day, N. E. *Statistical Methods in Cancer Research-Vol I- The analysis of case-control studies.* Lyon, International Agency on Cancer, 1988, p.p. 192-210, 233-242.

Feinstein, Alvar R. *Clinical Epidemiology. The architecture of Clinical Research.* London, W. B. Saunders Company, 1985, p.p. 124-125, 423-434.

Hosmer, D. W., Lemeshow, S. "A goodness-of-fit test for the multiple logistic regression model". *Communications in Statistics.* Vol. A10, 1980, p. p. 1043-1069.

Jennings, D. E. " Judging inference adequancy in logistic regression". *Journal of the American Statistical Association.* Vol 81, 1986, p.p. 471-476.

Kleinbaum , D. G., Kupper, L. L., Morgenstern, H. *Epidemiologic Research: principles and quantitative methods.* New York, Van Nostrand Reinhold, 1982. p. p. 242-245, 420-429, 477-491.

Lemeshow, S., Hosmer, D. W., "Estimation of odds ratios with categorical scaled covariates in múltiple Logistic Regression analysis". *American Journal of Epidemiology.* Vol. 119, 1983, p. p. 147-151.

McCullagh, P., and Nelder, J. A. *Generalized Linear Models 2a. Ed.* London, Chapman Hall, 1983, p.p. 107-124.