UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

Fundada en 1551

FACULTAD DE CIENCIAS MATEMATICAS

E.A.P. DE ESTADISTICA

"ANÁLISIS ESTADÍSTICO DE LOS FACTORES DE RIESGO QUE INFLUYEN EN LA ENFERMEDAD ANGINA DE PECHO"

MONOGRAFÍA

Para optar el Título Profesional de:

LICENCIADO EN ESTADÍSTICA

AUTOR

LUZ FLORES MANRIQUE

LIMA – PERÚ
2002

A mis padres Luz y Justo por tanto Amor y apoyo, a mi abuela Lucila y a mi hermano Eduardo.

ÍNDICE

INTRODUCCIÓN

CAPITULO I PRELIMINARES

1.1 Objetivos.....2

1.1.1 Objetivo General.....2

1.1.2 Objetivos Específicos.....2

1.2 Justificación.....2

CAPITULO II CONCEPTOS FUNDAMENTALES DE ANGINA DE PECHO.

2.1 Angina de Pecho.....4

2.2 Signos y Síntomas.....5

2.3 Diagnóstico.....6

2.4 Tratamiento.....7

2.5 Factores de riesgo.....8

2.5.1 Tabaquismo.....8

2.5.2 Hipertensión Arterial.....9

2.5.3 Colesterol Alto y Alteraciones.....9
en los Lípidos-Triglicéridos.

2.5.4 Obesidad.....10

2.5.5 Inactividad Física o sedentarismo.....11

2.5.6 Diabetes Mellitus.....12

2.5.7 Sexo.....13

2.5.8 Historia Familiar de enfermedades del Corazón.....13

CAPITULO III MODELOS DE REGRESION LOGISTICA.

3.1 Antecedentes.....14

3.1.1 Definición.....17

3.1.2 Función de Verosimilitud.....18

3.2 Modelo de Regresión Logística Binaria.....19

3.2.1 Función de Verosimilitud.....19

3.2.2 Estimación de Parámetros.....20

3.3 Modelo de Regresión Logística Simple.....21

3.3.1 Estimación de Parámetros.....23

3.3.2 Pruebas de significancia.....24

3.3.3 Bondad de ajuste del Modelo.....26

3.4 Modelo de Regresión Logística Múltiple.....29

3.4.1 Estimación de Parámetros.....30

3.4.2 Pruebas de significancia.....31

3.4.3 Pruebas de Bondad de ajuste.....33

3.4.4 Interpretación de los resultados.....33

CAPITULO IV ANALISIS DE LOS DATOS.

4.1 Descripción de los datos.....36

4.2 Modelo de Regresión Logística.....37

4.3 Ajuste del Modelo de Regresión Logística.....38

4.3.1 Evaluación del Incremento del estadístico.....38

-2LLo. Bondad de ajuste.

4.3.2 Parámetros estimados del modelo.....39
4.3.2.1 Estadística WALD.....40
4.3.2.2 Correlación Parcial (R).....42
4.3.2.3 ODDS Exp (B).....43
4.4 Matriz de Correlación.....43
4.5 Bondad de Ajuste del Modelo 2.....44
4.5.1 Bondad de Ajuste del Modelo 2.....44
4.5.2 Parámetros estimados del Modelo 2.....45
4.5.3 Correlación Parcial (R) y ODDS RATIO (Exp(B))...46
4.6. Método FORWARD.....47

CONCLUSIONES.....51
RECOMENDACIONES.....54
BIBLIOGRAFIA.....55
ANEXOS.....57
Anexo 1.....58
Anexo 2.....65
Anexo 3.....71
Tabla de datos.....77

RESUMEN

En éste trabajo de investigación se presenta la teoría y aplicación de los Modelos de Regresión Logística para conocer cuales son los Factores de Riesgo que influyen en la enfermedad Angina de Pecho. La razón principal de este estudio es identificar los factores más significativos de riesgo y prevención para dicha enfermedad dentro de la población en estudio.

Palabras claves: Regresión Logística, Angina de Pecho, Factores de Riesgo, "ODDS RATIO".

SUMMARY

In this work of investigation one appears Theory and application of the Models of Logistic Regression to know as they are the Risk Factors that influence in the disease chest Angina. The main reason of this study is to identify the significant factors but of risk and prevention for this disease within the population in study.

Key words: Regression Logistic, Angina of chest, Factors of the Risk, “ODDS RATIO”.

INTRODUCCIÓN

Los datos a usarse en éste trabajo provienen de la base de datos del servicio de Cardiología del Policlínico Juan José Rodríguez Lazo en el cuarto trimestre del año 2,001 trata de estudiar la ocurrencia o no ocurrencia de la enfermedad Angina de Pecho en un grupo de 149 pacientes, de los cuales 68 tienen la enfermedad y 78 no la presentan.

El Análisis de Regresión Logística se efectuarán con la ayuda del Software estadístico SPSS v. 9.0 para Windows, algunos cálculos intermedios serán efectuados con el Excel.

En términos generales el objetivo del estudio es conocer los factores que influyen en la ocurrencia de la enfermedad angina de pecho, y así poder establecer mejoras en el tratamiento a paciente en consultorio externo y evitar otras complicaciones que pueden devenir en emergencias que conlleven al fallecimiento del paciente. Más adelante se explicarán de manera más amplia los objetivos de estudio.

Este estudio como se ha mencionado anteriormente se lleva a cabo en el Centro Asistencial (CAS) Juan José Rodríguez Lazo, el cual tiene a su cargo población adscrita de los distritos de Chorrillos y Barranco, así también los datos provienen del área de Consulta Externa específicamente del servicio de Cardiología, en un período de tiempo del cuarto trimestre del año 2,001.

¿Tal vez una pregunta que se haga el lector es porque elegí este trimestre para llevarlo a estudio?, pues fue debido a que éste es un CAS que viene funcionando a la fecha 2 años y medio, y la información hasta fines del año pasado, nos estaba recién mostrando cual era nuestra población a tratar con dicha

enfermedad, era a si mismo por lo tanto la realidad a la cual nos debíamos de dirigir ya con un año y medio de atenciones.

CAPÍTULO I

1.1 OBJETIVOS

1.1.1 Objetivo General

Conocer las factores de riesgo más significativos que influyen en la presencia de la enfermedad Angina de pecho en el servicio de Cardiología de Consultorio Externo, en el Centro Asistencial Juan José Rodríguez Lazo – Chorrillos durante el año 2,001.

1.1.2 Objetivos específicos

a) Comparar la presencia o ausencia de la Angina de pecho en el grupo de observaciones en el cual se hará el estudio.
b) Proponer algunas acciones para mejorar y controlar dicha enfermedad en consulta externa para el servicio de Cardiología.
c) Poder observar el grado de influencia de cada uno de los factores o variables independientes en la presencia o ausencia de la enfermedad.
d) Plasmar la realidad actual del Programa de Atención Integral y de acuerdo a ello buscar las estrategias a seguir para mejorar la atención de los pacientes.

1.2. JUSTIFICACION

Es importante conocer los factores de riesgo que tienen mayor influencia en la presencia de la enfermedad Angina de pecho, la cual podría degenerar en un paro cardiaco, sino se controlan dichos factores.

La mayoría de los pacientes que presentan dicha enfermedad pertenecen a la atención de los llamados pacientes crónicos,

el estudio que se está realizando pretende entonces conocer cual es el grado en que afecta cada factor como son Glucosa, Colesterol, Edad, Obesidad, Sexo, Hipertensión Arterial, en la posible presencia o ausencia de la enfermedad a tratar.

Debido a que el Centro Asistencial Juan José Rodríguez Lazo cuenta con atención tanto en Consulta Externa y dentro de ella el servicio de Cardiología, que es de donde provienen los datos, y el Area Crítica es decir Emergencia, la cual cuenta con las cuatro prioridades que son No urgencias no emergencias, Urgencias, Emergencia y Shock Trauma, las cuales (Consulta Externa y el Area Crítica) se encuentran íntimamente ligadas, me hace realizar la investigación para controlar a los pacientes con la enfermedad en consultorio, a fin de evitar que dichos casos pasen al Area Crítica específicamente a la prioridad I es decir Shock Trauma, que es aquella que compromete seriamente la vida del paciente y que requiere inmediatamente Sala de reanimación, y luego desencadene en otras complicaciones.

Entonces también buscamos de esta manera educar a nuestros pacientes en el cuidado de su salud y poner en marcha el Programa Integral de Atención, ya que es necesario aunar esfuerzos para poder conseguir una meta ya trazada.

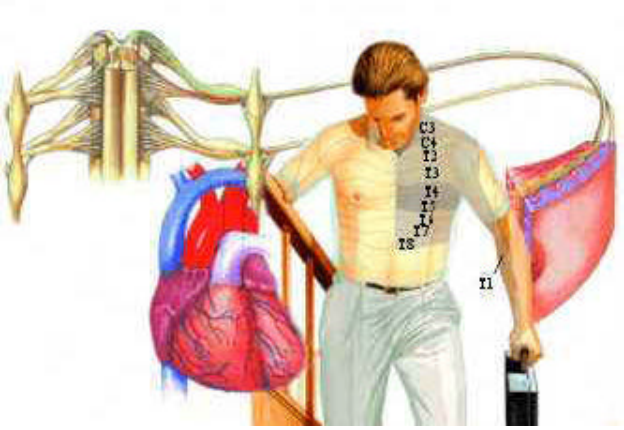
Debido a que la frecuencia de la presencia de la enfermedad Angina de Pecho en Shock Trauma Area Crítica es la más elevada entonces esto nos da la idea que no se está realizando una buena educación al paciente es decir creando conciencia de lo que representa la enfermedad que tienen. Pero la frecuencia de dicha enfermedad en el Area Crítica será tema de otro estudio posterior.

CAPÍTULO II

CONCEPTOS FUNDAMENTALES DE ANGINA DE PECHO

2.1 ANGINA DE PECHO

Se conoce como enfermedad coronaria cualquier trastorno causado por una restricción en el suministro de sangre al músculo cardíaco. Las manifestaciones más corrientes son la angina de pecho y el infarto de miocardio. La angina de pecho es un síndrome clínico de molestia torácica transitoria, con crisis dolorosas súbitas y dificultad respiratoria como resultado del aumento de la demanda miocárdica de oxígeno y de la estenosis coronaria. Típicamente, la angina comienza con un dolor atenazante u opresivo por detrás del esternón, que se puede irradiar hacia el cuello y la mandíbula, o bien hacia el brazo izquierdo. El dolor remite rápidamente con el reposo. Las emociones fuertes o el frío pueden determinar que no sea necesario un gran esfuerzo para provocar la angina.



2.2 SIGNOS Y SINTOMAS

Dolor torácico y sensación de opresión aguda y sofocante, generalmente retrosternal, es decir, centrada detrás del esternón, y a veces extendida (irradiada) a uno u otro brazo. El dolor torácico suele durar desde uno o dos minutos, hasta tanto como 10 ó 15 minutos. A veces se percibe una sensación de pesadez u opresión en el pecho que no llega a dolor.

Los ataques se desencadenan, generalmente, por ejercicio (levantar pesos, deporte, actividad sexual) o stress emocional, y se alivian con el reposo. También pueden desencadenarse por frío extremo o por comidas pesadas.

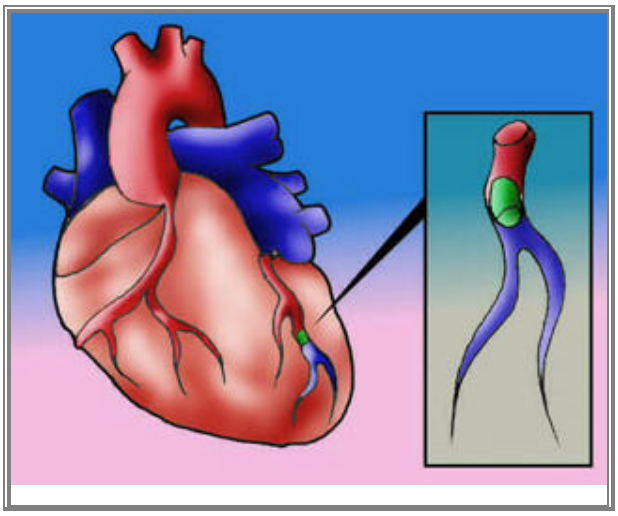
Sentimiento de ansiedad o de muerte inminente.

La angina es un síntoma, y no una enfermedad. Es el resultado directo de la falta de sangre en el músculo cardíaco (isquemia). Cuando uno se esfuerza, el corazón necesita más oxígeno para trabajar. Cuando las arterias coronarias están afectadas y no pueden ajustarse al aumento de la demanda de sangre, los nervios del corazón transmiten mensajes dolorosos de aviso urgente al cerebro. El dolor referido o irradiado se debe a que el cerebro, por confusión, siente los impulsos desde localizaciones cercanas como los brazos, el cuello o la mandíbula.

La angina es frecuente. En los hombres ocurre generalmente después de los 30 años de edad, y en las mujeres más tarde. La causa, en la mayor parte de los casos, es la arteriosclerosis.

La angina estable es la que ocurre siempre con el mismo nivel de ejercicio, y la duración de los ataques es similar. Cuando los ataques se hacen más frecuentes y largos o no están ligados a actividad física, los cardiólogos hablan de angina

inestable, que puede ser el aviso de un infarto inminente y necesita tratamiento especial.



2.3 DIAGNÓSTICO

No hay pruebas de laboratorio para el diagnóstico de la angina de pecho. Sin embargo, puede ser útil disponer de algunos análisis para descartar que haya ocurrido daño en el corazón, así como para detectar la presencia de situaciones como **hipertiroidismo** o **anemia**, que pueden forzar al corazón a latir más rápido, usar más oxígeno, y por lo tanto, precipitar la angina.

El ECG sólo detecta alteraciones en el momento preciso del dolor. Con posterioridad, sólo puede confirmar o descartar en algunos casos que se haya producido daño en el corazón.

Por lo tanto, el diagnóstico de la angina es clínico, es decir, no necesita confirmación si los síntomas y la historia clínica general son sugestivos.

2.4 TRATAMIENTO

El tratamiento de la angina de pecho es el de la enfermedad coronaria, y está dirigido a disminuir la carga del corazón y sus necesidades de oxígeno:

- Deje de fumar.
- Pierda los kilos de más.
- Ejercicio. Tener angina no significa que tenga uno que volverse un ser sedentario. De hecho, el ejercicio es parte clave en el manejo de la enfermedad coronaria. Tiene que ser, desde luego, compatible con las limitaciones impuestas por el dolor y por su estado general (ver Salud Cardiovascular).
- Medicación.
- La nitroglicerina (Vernies, Cafinitrina, etc) dilata las arterias coronarias y suele revertir el dolor en minutos. Se toma en pastillas debajo de la lengua o también en spray. Puede dar dolor de cabeza como efecto secundario.
- Los calcio-antagonistas o bloqueantes de los canales del calcio impiden la entrada de calcio en las células del miocardio. Esto disminuye la tendencia de las arterias coronarias a estrecharse y además disminuye el trabajo del corazón y por tanto sus necesidades de oxígeno. También disminuyen la tensión arterial.
- Betabloqueantes. Actúan bloqueando muchos efectos de la adrenalina en el cuerpo, en particular el efecto estimulante sobre el corazón. El resultado es que el corazón late más despacio y con menos fuerza, y por tanto necesita menos oxígeno. También disminuyen la tensión arterial.

- Cirugía. En caso de angina inestable o que resiste al tratamiento con medicamentos, la cirugía puede conseguir corregir la obstrucción de los vasos coronarios, bien mediante by pass (derivación) o en algunos casos mediante la apertura de los vasos estrechados o angioplastia coronaria.

2.5 FACTORES DE RIESGO

2.5.1 TABAQUISMO

El tabaquismo es un problema social de gran importancia en todo el mundo. Las personas que fuman una cajetilla de cigarros al día tienen un riesgo de tres a cinco veces mayor de desarrollar cardiopatía isquémica que los no fumadores y entre más fumen mayor es el riesgo.

El tabaquismo potencia en gran medida otros factores de riesgo relacionados con la enfermedad como son la presión arterial alta, diabetes y colesterol alto aumentado su letalidad.

Fumar pipa y puro también incrementan la frecuencia de isquemia coronaria pero en menor medida que los cigarrillos probablemente secundario a que los primeros inhalan menos humo.

Afortunadamente las persona que deciden dejar de fumar disminuyen su riesgo de desarrollar arteroesclerosis, lamentablemente les toma de 5 a 10 años disminuir el riesgo a los niveles que tienen los no fumadores.

2.5.2 HIPERTENSION ARTERIAL

La hipertensión arterial es un factor de riesgo bien conocido para el desarrollo de arteroesclerosis el cual fue inicialmente identificado por las compañías de seguros y posteriormente por la profesión médica, en forma concluyente, en los estudios realizados en la ciudad Norteamericana de Framingham.

Incrementos tanto en la presión sistólica (primer número) como la diastólica (segundo número) se correlacionan con aumentos en la incidencia de la enfermedad. Una persona con presión arterial de 160/95 tiene un riesgo cinco veces mayor que si tuviera 140/90 o menos. Lo niveles considerados como normales son menores de 160 mmHg para la presión sistólica y menores de 85-90 mmHg para la presión distólica.

La presión arterial alta no solo daña el corazón produce enfermedades serias en el cerebro, riñones y extremidades y como raramente produce síntomas, también es conocida como la "asesina silenciosa".

2.5.3 COLESTEROL ALTO Y ALTERACIONES EN LOS LIPIDOS TRIGLICERIDOS.

El colesterol es una substancia necesaria para el organismo, forma parte de la membrana de las células y es precursor de hormonas y sales biliares. El colesterol se obtiene de los alimentos y es procesado por el hígado, este último en caso necesario lo produce.

El colesterol se une a lípidos y proteínas para ser distribuido por el organismo, en este momento se le conoce

como lipoproteinas. La lipoproteinas de baja densidad (L.D.L.-Colesterol) tienen la función de transportar el colesterol del hígado a los tejidos donde se puede depositar y desarrollar arteroesclerosis por lo tanto son popularmente conocidas como Colesterol Malo. Las lipoproteinas de baja densidad (H.D.L.-Colesterol) transportan el exceso de colesterol de los tejidos al hígado desarrollando una especie de trabajo de limpieza y por lo tanto protegen al organismo contra la ateroesclerosis y son conocidas popularmente como Colesterol Bueno.

Existe una controversia importante en cuanto al papel que juegan los triglicéridos ya que se considera que no son responsables directamente de enfermedad cardiovascular. Los triglicérido altos en ocasiones se acompañan de niveles bajos de lipoproteinas de alta densidad por lo tanto disminuyen los niveles de colesterol bueno aumentando de esta forma el riesgo de ateroesclerosis cuando esto sucede algunos investigadores lo han asociado con un síndrome de resistencia del organismo a la insulina que puede causar otros problemas como intolerancia a los carbohidratos e hipertensión arterial.

2.5.4 OBESIDAD

La obesidad se origina en factores genéticos y ambientales este último se apoya en la alta incidencia de obesidad en países desarrollados.

La obesidad no debe de ser valorada únicamente en términos de peso absoluto, la forma en que la grasa se distribuye y el porcentaje de la misma son los factores determinantes. Las personas que acumulan grasa principalmente en el abdomen (forma de manzana) y no en la cadera (forma de pera) son los

que se encuentran en mayor riesgo de desarrollar arteroesclerosis.

La obesidad se relaciona frecuentemente con otros factores de riesgo como lo son diabetes, hipertensión arterial, colesterol alto y falta de ejercicio. En los años recientes los investigadores han descubierto la presencia de resistencia periférica a la insulina como factor común en individuos que presentan obesidad, hipertensión arterial, alteración en los lípidos e intolerancia a los carbohidratos esto incrementa significativamente el riesgo cardiovascular.

Siempre que se inicie un plan dietético para bajar de peso es indispensable acompañarlo de ejercicio porque de no hacerlo la posibilidad de disminuir la masa muscular del organismo es importante, ocasionando mayores incrementos de peso al suspender la restricción en alimentos (rebote).

2.5.5 INACTIVIDAD FISICA O SEDENTARISMO

La inactividad física es un factor de riesgo bien definido para el desarrollo de ateroesclerosis. La realidad es que el ejercicio es el mejor amigo del corazón principalmente cuando se esta enfermo del corazón.

El ejercicio regular aumenta los niveles de colesterol bueno (HDL-Colesterol), disminuye el sobrepeso, ocasiona el desarrollo de circulación colateral (formación de vasos nuevos de arterias sanas a enfermas) que puede evitar cirugías de corazón, disminuye la presión arterial, mejora el control de la glucosa en diabéticos, normaliza los factores de coagulación disminuyendo la probabilidad de formación de

trombos, disminuye la presión emocional, etc.

En pocas palabras es la mejor medicina para el corazón. Si usted está enfermo del corazón el ejercicio es su mejor aliado pero en un inicio es recomendable que cuente con supervisión médica calificada.

2.5.6 DIABETES MELLITUS

La presencia de diabetes ya sea insulino dependiente (Tipo I) o no (Tipo II) es un factor de riesgo importante para desarrollar enfermedad isquémica coronaria. La diabetes incrementa el riesgo de 3 a 5 veces de lo normal y si se combina con otros factores de riesgo como tabaquismo e hipertensión este aumenta en forma desproporcionada.

El mecanismo como se origina esto no esta totalmente claro pero los niveles de glucosa altos y las alteraciones en el perfil de lípidos que lo acompañan juegan un papel importante. La diabetes produce daños en los pequeños vasos sanguíneos en forma difusa lo cual dificulta enormemente su tratamiento, además produce daños a otros órganos como los riñones donde es responsable de la mayoría de casos de insuficiencia renal, provocando de esta forma grandes dificultades en la administración de medicamentos que requieren del riñón para su excreción.

Afortunadamente el mantener un control estricto de los niveles de glucosa en la sangre pueden disminuir considerablemente estos riesgos, siempre tenga en mente que los daños

relacionados a la diabetes son acumulativos por lo tanto no hay nada mejor que la prevención

2.5.7 SEXO

Las mujeres presentan una menor incidencia de enfermedad cardiovascular en comparación con los hombres pero cuando ellas se enferman su comportamiento es más agresivo.

Esta protección se debe a las hormonas femeninas que producen un perfil de lípidos más favorable con niveles menores de LDL-Colesterol y mayores de HDL-Colesterol lo cual disminuye la incidencia de arteroesclerosis.

Como regla general las mujeres se enferman diez años después que el promedio de sus contrapartes masculinas. En la actualidad toda mujer posmenopáusica con diagnóstico de ateroesclerosis significativa debe iniciar tratamiento con reemplazo hormonal.

2.5.8 HISTORIA FAMILIAR DE ENFERMEDADES DEL CORAZÓN

Solo se considera significativas si se presentan en un padre o hermanos antes de la edad de 45 años o en su madre o hermanas antes de los 55 años de edad. En estos casos siempre es indispensable investigar la presencia de niveles de colesterol muy altos relacionados con padecimientos hereditarios

CAPÍTULO III

MODELOS DE REGRESIÓN LOGÍSTICA

3.1 ANTECEDENTES

La regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización. Dado que el modelo logístico no es lineal, sino exponencial, se utilizan transformaciones logarítmicas para linealizar el modelo y hacen que los coeficientes no pueden interpretarse directamente.

El objetivo del modelo puede ser estimativo, es decir estimar la mejor relación de las variables independientes con la variable dependiente, usado mayormente en estudios etiológicos que consiste en investigar factores causales de una determinada característica de la población y estudiar que factores modifican la probabilidad en la aparición de un suceso determinado; o también predictivo que consiste en predecir lo mejor posible la variable dependiente a través de las independientes, habitualmente es dicotómico (clasifica él valor de la variable respuesta como 1 cuando presenta la característica y con valor 0 cuando no está presente), también puede ser usada para estimar probabilidades de cada una de las posibilidades de un suceso en más de dos categorías (politómico).

La técnica resulta especialmente útil para identificar factores de riesgo y factores de prevención de enfermedades en

muestras prospectivas donde la metodología de la regresión lineal no es aplicable, dado que la variable respuesta sólo presenta dos valores (caso dicotómico) como puede ser presencia/ausencia de un suceso.

Inicialmente, el análisis de Regresión Logística fue sugerido por Cox (1970).
La condición de la existencia de una única solución para la ecuación de verosimilitud fue dada por Albert y Andersson (1984).

El Modelo de Regresión Logística es un caso especial del Modelo Lineal Generalizado como fue propuesto por Nelder y Wedderburn (1972) y ampliamente discutida en McCullagh y Nelder (1983).

El libro de McCullagh y Nelder muestra la solución de la ecuación de verosimilitud en el Modelo Lineal Generalizado – y en Regresión Logística – utilizando el método de Newton-Raphson, esta solución puede ser obtenido por un método similar al cuadrado medio ponderado para el Modelo de Regresión Ordinario. El método es llamado método iteractivo del cuadrado Medio Ponderado, el cual también puede ser encontrado , por ejemplo en Andersen (1990) o Agresti (1996).

Residuales Estandarizados y la distancia de Cook en Regresión Logística fueron discutidos en Pregibon (1981).

Análisis de Regresión Logística es también tratado en libros sobre Análisis de Datos Categóricos y en muchos libros sobre

análisis de Regresión Aplicada por ejemplo en Andersen (1990), Agresti (1996) y Weisberg (1985).

Recientemente Hosmer y Lemeshow (1989) publicaron un libro especial en análisis de regresión Logística.
El Modelo de Regresión Logística ha sido utilizado por muchos años; pero no fue hasta que Truett, Cornfield, y Kannel (1967) que aplicaron el Modelo de Regresión Logística utilizando los datos de Framingham, el cual trata de un estudio del corazón, donde se pudo apreciar el poder y la aplicación de estos modelos .

Desde la publicación de este artículo el modelo de regresión logística llega a ser el método estándar para el análisis de regresión de datos dicotómicos en muchas áreas del conocimiento especialmente en las ciencias de la Salud. Luego muchos "journals" como "The American Journal of Epidemiology", "The American Journal of Public Health", "The International Journal of Epidemiology" y "The Journal of Chronic Diseases" publicaron artículos cuyos análisis son basados en el modelo de regresión logística.

Entre los pocos textos que incluyen temas sobre regresión logística se encuentra el libro de Breslow y Day (1980), Cox (1970), Kleinbaum, Kupper y Morgenstern (1982), y Schlesselman (1982). En cada uno de estos textos, el tema central no es regresión logística.

Muchas de las técnicas para aplicar el método e interpretación de los resultados pueden ser solamente encontrados en la

literatura estadística, lo que esta fuera de la comprensión de muchos usuarios potenciales.

Un libro excelente en Regresión Logística aplicada fue escrito por Hosmer y Lemeshow (1989).

El principal objetivo de este libro es dar una introducción en el modelo de regresión logística y utilizar este método para modelar la relación entre la probabilidad de ocurrencia de los resultados de una variable respuesta dicotómica (en general llamada variable dependiente), que normalmente son los términos suceso o fracaso, y las variables explicativas categóricas o continuas (conocidas como variables independientes). La idea básica consiste en establecer una relación lineal entre las variables explicativas (o algunas transformaciones de éstas) y una transformación, denominada logit, de la variable respuesta.

3.1.1 DEFINICIÓN

Sea Y una variable dependiente binaria que toma dos valores posibles etiquetados como 0 y 1.

Sean $X_1,...,X_k$ un conjunto de variables independientes observadas con el fin de explicar y/o predecir el valor de Y.

El objetivo es determinar $P[Y=1/X_1,...,X_k]$, donde P indica probabilidad

por lo tanto

$$P[Y=0/X_1,...,X_k] = 1- P[Y=1/X_1,...,X_k]).$$

se construye un modelo de la forma:

$$P[Y=1/X_1,...,X_k] = p(X_1,...,X_k\ ;\ ß) \qquad (1)$$

donde $p(X_1,...,X_k\ ;\ ß)$: $R^k \quad [0,1]$

es una función que recibe el nombre de función de enlace (función de probabilidad) cuyo valor depende de un vector de parámetros $ß = (ß_1,...,ß_k)'$.

3.1.2 FUNCIÓN DE VEROSIMILITUD

Con el fin de estimar ß y analizar el comportamiento del modelo considerado, observamos una muestra aleatoria simple de tamaño *n* dada por $\{(x_i',y_i);i=1,...,n\}$ donde $x_i = (x_{il},...,x_{ik})'$ es el valor de las variables independientes e $y_i=\{0,1\}$ es el valor observado de Y en el *i-ésimo* elemento de la muestra.

$$Y/(X_1,...,X_k) \sim \text{Binomial}(1,\ p(Y=1/X_1,...,X_k\ ;\ ß))$$

Utilizando el hecho de que la variable dependiente toma sólo dos resultados (*éxito* y *fracaso*), cuando el número de éxitos en *n* repeticiones tiene una distribución binomial B(n,*p*).

La función de verosimilitud es:

$$L(\beta/(x_1',y_1),.....,(x_n',y_n)) = \prod_{i=1}^{n} p_i^{y_1}(1-p_i)^{1-y_1} \qquad (2)$$

donde

$$p_i = p(x_i';\ ß) = p(x_{i1},...,x_{ik};\ ß);\ i=1,...,n$$

3.2 MODELO DE REGRESIÓN LOGÍSTICA BINARIA

Sea

$$p(X_1,...,X_k;\ ß) = G(ß_1X_1 + ... + ß_kX_k) \qquad (3)$$

donde

$$G(x) = \frac{e^x}{1+e^x}$$

es la función de densidades acumuladas que es la función logística, el modelo normalmente conocido es:

$$\log\left(\frac{p(x_1,...,x_k;\beta)}{1-p(x_1,...,x_k;\beta)}\right) = \beta_0 + \beta_1x_1 + ... + \beta_kx_k \qquad \textbf{(4)}$$

llamado *modelo logit*. Cuando la variable cualitativa toma el valor 1 en la expresión:

$$\frac{p\left[Y=1/\ X_1,...,X_k\right]}{p\left[Y=0/X_1,...,X_k\right]} = \frac{p(x_1,...,x_k;\beta)}{1-p(x_1,...,x_k;\beta)} \qquad \textbf{(5)}$$

se conoce con el nombre de factor de riesgo en el mundo de la medicina, donde la variable Y indica habitualmente la presencia de una determinada enfermedad, objeto de estudio y en ausencia toma el valor 0.

3.2.1 FUNCIÓN DE VEROSIMILITUD

Teniendo en cuenta la forma matricial de:

$$p\left(X_1,...,X_k;\beta\right) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \qquad \textbf{(6)}$$

según (2) la función de verosimilitud viene dada por:

$$L\left(\beta/\left(x_1^{'}, y_1\right).....,\left(x_n^{'}, y_n\right)\right)=\prod_{i=1}^{n}\left[\frac{e^{x_i^{'}\beta}}{1+e^{x_i^{'}\beta}}\right]^{y_1}\left[\frac{1}{1+e^{x_i^{'}\beta}}\right]^{1-y_1}$$

3.2.2 ESTIMACIÓN DE LOS PARÁMETROS

El vector de parámetros (ß) se estima mediante el método de máxima verosimilitud que consiste en elegir el valor de ß, como estimador para (ß) para el cual L(ß) es máximo, se toma logaritmo a la función en la ecuación (1) del siguiente modo:

$$L(\beta)=\log L(\beta)=\sum_{i=1}^{n}\log p\left(X_1,....., X_k;\beta\right) \qquad (7)$$

se resuelve mediante la ecuación de verosimilitud

$$\frac{\partial\,\ell(\beta)}{\partial\beta}=\sum_{i=1}^{n}X_i\left(y_i-p_i\right)=0 \qquad (8)$$

donde

$p_i = p\ (X_i;\beta)\quad i=1,...,n$ mediante métodos iterativos.

Este método consiste en maximizar la función de verosimilitud de la muestra en función del parámetro $\hat{b}$.

Este procedimiento es matemáticamente complejo a través del cálculo diferencial, pero lo que importa para el usuario es:

1° El proceso es iterativo, es decir se dan a los coeficientes unos valores arbitrarios (habitualmente, aunque no

necesariamente, el valor 0). Algunos paquetes estadísticos (por ejemplo el PRESTA) preguntan por estos valores, otros (como el SPSS o el SAS) no y asumen 0. La solución final no depende de estos valores, pero sí el tiempo de cálculo y a veces puede ser necesario "jugar" con ellos.

2° A partir de estos valores iniciales y de los valores de la(s) variable(s) independiente(s) se calculan las matrices de varianzas y covarianzas.

3° Y a partir de la inversa de la matriz se calculan los nuevos estimadores, se comprueba si son la solución final se debe parar el proceso y en caso contrario se repite el proceso.

3.3 MODELO DE REGRESIÓN LOGÍSTICA SIMPLE

Para construir el modelo matemático es necesario tener valores numéricos, los cuales se obtienen considerando la probabilidad de que ocurra un suceso determinado P(Y) en relación con la dependencia de que dicha probabilidad no ocurra 1- P(Y).

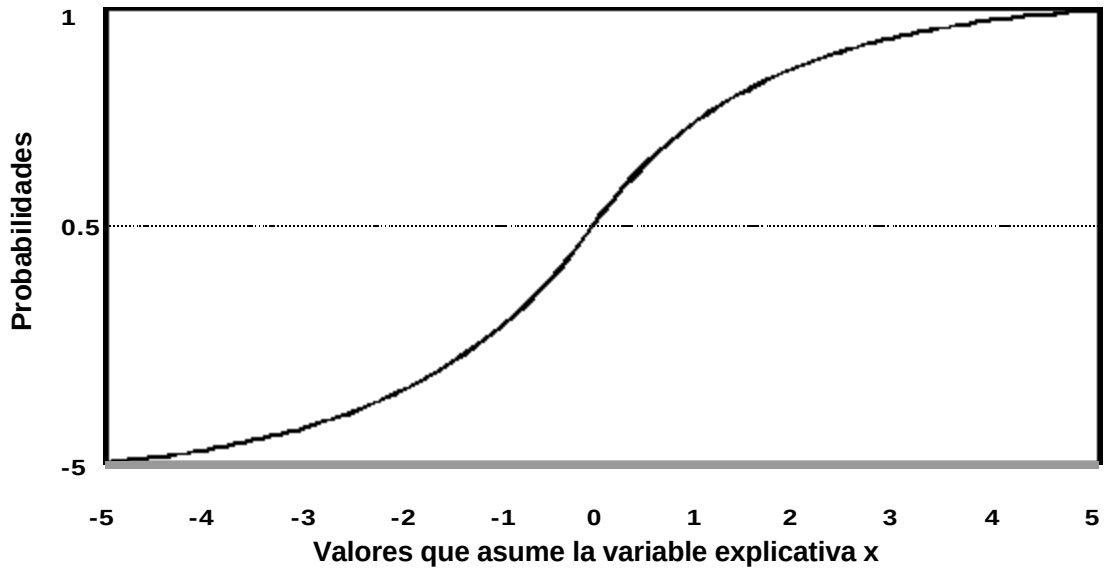
En el primer miembro de la ecuación interesa tener a P y en el segundo miembro la relación funcional con intervención de las variables independientes que son los factores de interés. La probabilidad es un número que oscila entre 0 y 1, que proporciona predicciones consistentes y de fácil interpretación de los resultados en términos de razón de probabilidades llamado "Odds ratio".

Sea la función:

$$f(x) = \frac{1}{1+e^{-x}} \qquad (9)$$

que aparece en otros muchos campos de la matemática aplicada, y cuya gráfica se muestra en la figura N° 1, se le denomina función Logística.

Figura N° 1: FUNCIÓN LOGÍSTICA



Para una única variable independiente X, el modelo de regresión logística de la ecuación (6), toma la forma:

$$p_i = \frac{e^{\beta_0+\beta_1 x}}{1+e^{(\beta_0+\beta_1 x)}} \qquad (10)$$

el modelo Logit será:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X \qquad (11)$$

o simplificando la notación:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X$$

Donde log significa logaritmo en base diez, ß$_0$ y ß$_1$ son constantes y X una variable explicativa que puede ser continua o discreta. El campo de variación de log (p_i/(1- p_i)) es todo el campo real (de - a), mientras para p el campo es sólo de 0 a 1 y para p_i/(1-p_i) de 0 a . Por lo tanto, al modelo logístico no hay que poner restricciones a los coeficientes que sólo complicarían su estimación, lo más importante es que las coeficientes son fácilmente interpretables en términos de independencia o asociación entre las variables.

3.3.1 ESTIMACIÓN DE LOS PARÁMETROS

El método más usado es el de máxima verosimilitud que consiste en elegir el valor de $\hat{b}$ (como estimador para ß), tal como se dedujo la ecuación (8) y considerando la ecuación (10), se tiene:

$$L(\beta) = \sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i) \right] = \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) - \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) \right]$$

La primera iteración es la primera derivada de la ecuación de verosimilitud y son las siguientes:

$$\frac{\partial L(b/Y)}{\partial b_0} = \sum_{i=1}^{n}\left(y_i - \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} \right) = \sum_{i=1}^{n}(y_i - p_i)$$

$$\frac{\partial L(\beta/Y)}{\partial \beta_1} = \sum_{i=1}^{n}\left(y_i x_i - \frac{x_i e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) = \sum_{i=1}^{n} x_i (y_i - p_i)$$

La segunda iteración es a través de la segunda derivada por el método Newton Raphson, las ecuaciones son las siguientes:

$$\frac{\partial^2 L(\beta/Y)}{\partial \beta_0^2} = -\sum_{i=1}^{n} \frac{e^{\beta_0 + \beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} = -\sum_{i=1}^{n} p_i (1 - p_i)$$

$$\frac{\partial^2 L(\beta/Y)}{\partial \beta_0 \, \partial \beta_1} = -\sum_{i=1}^{n} \frac{x_i e^{\beta_0 + \beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} = -\sum_{i=1}^{n} x_i p_i (1 - p_i)$$

$$\frac{\partial^2 L(\beta/Y)}{\partial \beta_1^2} = -\sum_{i=1}^{n} \frac{x_1^2 e^{\beta_0 + \beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} = -\sum_{i=1}^{n} x_1^2 p_i (1 - p_i)$$

Las iteraciones deben parar cuando se encuentra la solución y es el máximo estimador de ß.

3.3.2 PRUEBAS DE SIGNIFICANCIA

Una vez estimado los coeficientes del modelo, se tiene que verificar si el modelo predice de manera adecuada a la variable dependiente en un nuevo individuo relacionado con la muestra, donde los valores de las variables explicativas son las probabilidades estimadas cuando P(Y=1) y P(Y=0). Para esto, se formula y prueba la hipótesis estadística, para determinar si la variable independiente influye significativamente en la probabilidad del suceso del modelo relacionado a la variable del resultado del siguiente modo.

H_0 : La variable independiente no influye sobre pi

H_1 : La variable independiente influye sobre pi

Donde:

$$p_i = p(x;\beta_0,\beta_1) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{(\beta_0+\beta_1 x)}}$$

El modelo de regresión logística es válido si ß$_1$, es significativamente diferente de cero y este coeficiente muestral es el estimador de la población (B_1) que va a explicar la variable dependiente p_i y poder hacer posteriormente predicciones con el modelo.

i. EVALUACIÓN ESTADÍSTICA DEL COEFICIENTE : ß$_1$

El coeficiente ß$_1$ debe ser diferente de cero para que tenga influencia significativa en la variable dependiente en las siguientes hipótesis:

H_0 : $B_1 = 0$

H_1 : $B_1 \neq 0$

Estadístico de prueba

$$t = \frac{\beta_1 - B_1}{s_{\beta_1}} \sim t_{\alpha,n-k-1} \qquad (12)$$

donde s_{β_1}, es el error estándar del coeficiente de regresión logística muestral y k es él número de variables independientes, mientras que B_1 es el coeficiente de regresión logística poblacional y B_1 es el coeficiente de regresión

logística muestral, como $B_1=0$ por definición de la hipótesis y k=1 con un nivel de significancia á, entonces (12) será:

$$t = \frac{b_1}{s_{b_1}} \sim t_{a,n-2}$$

Decisión: si |t| > tα, rechazamos Ho

ii. ESTADÍSTICO WALD

Evalúa el coeficiente estimado en la población y se define como un cociente entre el coeficiente y el error estándar del coeficiente en la hipótesis:

$H_0 : B_1 = 0$

$H_1 : B_1 = 0$

Estadístico de prueba

$$WALD = \frac{\beta_1}{s_{\beta_1}} \sim X^2_{\alpha,1} \qquad (13)$$

Decisión: si Wald > $X^2_{\alpha,1}$ rechazamos H_0 con un nivel de significancia á y concluimos que la variable independiente influye en la probabilidad de las características de la variable dependiente. Si la variable independiente es cualitativa los grados de libertad es igual al número dé categorías menos 1.

3.3.3 BONDAD DE AJUSTE DEL MODELO

Para evaluar la bondad del modelo se utiliza el logaritmo del cociente de verosimilitud y la prueba de Hosmer-Lemeshow.

i. EL INCREMENTO DEL ESTADÍSTICO -2logL

El estadístico -2logL mide los cambios que se producen cuando se agrega o se quita una variable, donde L es la función de verosimilitud del modelo estudiado, puede oscilar entre 0 y 1, si el modelo se ajusta perfectamente a la data tiene una verosimilitud igual a 1, de allí que -2logL = 0. Entonces diremos que el modelo se ajusta bien a la data si tiene un valor pequeño de -2logL, que es el logaritmo de la verosimilitud y se distribuye como una X^2(Ji-cuadrado), cuando el modelo incluye sólo la constante los grados de libertad es igual al número de casos menos uno (n-1), y cuando se incluye la variable independiente sigue una distribución X^2 con n-k-1 grados de libertad, en el modelo de regresión logística simple es n-2, la diferencia entre estos dos valores de -2logL se llama ***Devianza***, prueba si la variable x_i es significativa, se define como:

D = -2log (verosimilitud del modelo sin la
variable / verosimilitud del modelo con la variable)

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\log\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \qquad (14)$$

Las hipótesis son:

H_0 : El modelo ajustado es significativo

H_1 : El modelo ajustado no es significativo

Estadístico de prueba

D ~ X^2 con n-k-1 grados de libertad (15)

Decisión

si D < $X^2_{\alpha,(n-k-1)}$ no rechazamos H_0, el modelo ajustado es significativo.

ii. PRUEBA DE HOSMER - LEMESHOW

Evalúa la bondad del modelo construyendo una tabla de contingencia, divide la muestra en aproximadamente 10 grupos iguales a partir de las probabilidades estimadas, para comparar las frecuencias observadas con las esperadas en cada uno de estos grupos a través de la prueba X^2 con j-2 grados de libertad, en donde j es el número de grupos formados.

Se calcula los deciles de las probabilidades estimadas $\hat{P}i$; i = 1,.....,n y D_i,.....,D_9 que son los deciles observados divididos en 10 grupos dados por:

$$A_j = \{i \in \{1,....., n\} / \hat{P}_i \in [D_{j-1}, D_j]\}, j = 1,.....,10$$

donde: $D_0 = 0$, $D_{10} = 1$

sean:

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

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

$$\overline{P}_j = \frac{1}{n_j} \sum_{i \in A_j} \hat{P}_i \; ; j = 1,...10$$

Las hipótesis a contrastar son:

H_0 : El modelo es adecuado

H_1 : El modelo no es adecuado

Estadístico de prueba es:

$$X^2 = \sum_{j=1}^{10} \frac{\left(o_j - n_j \overline{P}_j\right)^2}{\overline{P}_j n_j \left(1 - n_j\right)} \sim X^2_{a,j-2} \qquad (16)$$

Decisión: si $X^2 \geq X^2_{a,j-2}$ rechazamos Ho y concluimos que el modelo no es adecuado a un nivel de significancia á.

3.4 MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE

Es una generalización del modelo simple, relaciona la probabilidad de que ocurra un determinado suceso independiente denotado por el vector X'= (x_1,...,x_k) con probabilidad condicional P(Y=1/X) en función de *k* variables independientes que pueden ser cuantitativas, cualitativas o combinadas según sea el tipo de diseño de estudio.

El modelo logístico múltiple es:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (17)$$

o también:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}} \qquad (18)$$

3.4.1 ESTIMACIÓN DE LOS PARÁMETROS

Sea una muestra de *n* observaciones independientes definido por $(x_{i1}, x_{i2},..., x_{ik}, y_i)$, i=1,...,n y como en el caso univariante se elige el vector ß'= $(ß_0,...,ß_k)$, el método más usado es el de máxima verosimilitud definido en (3.3.2.2) con *k*+1 ecuaciones de verosimilitud que se obtienen derivando el log de la función de verosimilitud respecto a *k*+1 coeficientes. Las ecuaciones de verosimilitud son:

$$\sum_{i=1}^{n}\left[y_i - p_i\right] = 0 \qquad \textbf{y}$$

$$\sum_{i=1}^{n} x_{ij}\left[y_i - p_i\right] = 0\,,\ j = 1,.....k\ ;i = 1,....., n$$

Encontrar la solución a este conjunto de ecuaciones es mediante el cálculo diferencial, hoy en día existen software estadísticos para estimar los parámetros.

Sea $\hat{\beta}$ el estimador de máxima verosimilitud para el sistema de ecuaciones de tal modo $\hat{p}_i$ es el modelo logístico múltiple de la ecuación (18). El método de estimación de varianzas y covarianzas de los coeficientes estimados, es a través del método de máxima verosimilitud con procesos iterativos, consiste en obtener la matriz de la segunda derivada parcial de la función de verosimilitud éstas derivadas parciales tienen la siguiente forma general:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} \frac{x_{ij}^2\, e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = -\sum_{i=1}^{n} x_j^2\, p_i(1-p_i) \qquad (19)$$

$$\frac{\partial^2 L(\beta)}{\partial\beta_j\,\partial\beta_l} = -\sum_{i=1}^{n}\frac{x_{ij}\,x_{il}\,e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = -\sum_{i=1}^{n} x_{ij}\,x_{il}\,p_i(1-p_i) \qquad (20)$$

Sea una matriz (k+1)x(k+1) que contiene los términos negativos en las ecuaciones (19) y (20) denotado por I(B), llamado matriz de información con varianzas y covarianzas de los coeficientes estimados por la inversa de la matriz de la siguiente forma $\Sigma(B) = I^{-1}(B)$, cuyos elementos de la diagonal son $\acute{o}^2(ß_j)$ que es el j-ésimo elemento de la diagonal, la varianza de $\hat{\beta}_j$ y $\sigma(\hat{\beta}_j,\hat{\beta}_l)$ son las covarianzas de $\hat{\beta}_j$ y $\hat{\beta}_l$ para estimar la matriz de información del modelo estimado es $\hat{I}(\hat{b}) = X'VX$, donde $X_{nx(k+1)}$ es la matriz de datos de los sujetos y V_{nxn} es una matriz diagonal cuyo elemento general es $\hat{P}_i(1-\hat{P}_i)$. Las matrices son:

$$X = \begin{bmatrix} 1 & x_{1l} & ..... & x_{1k} \\ 1 & x_{2l} & ..... & x_{2k} \\ ... & ... & ..... & ... \\ 1 & x_{nl} & ..... & x_{nk} \end{bmatrix}_{nx(k+1)} \qquad (21)$$

$$V = \begin{bmatrix} \hat{P}(x_1)\left(1-\hat{P}(x_1)\right) & 0 & ..... & 0 \\ 0 & \hat{P}(x_2)\left(1-\hat{P}(x_2)\right) & ..... & 0 \\ ... & ... & ..... & ... \\ 0 & 0 & ..... & \hat{P}(x_n)\left(1-\hat{P}(x_n)\right) \end{bmatrix}_{nxn} \qquad (22)$$

3.4.2 PRUEBAS DE SIGNIFICANCIA

Como en el caso univariante se prueba la significancia de las variables independientes del modelo mediante la prueba de verosimilitud con la significancia de los k+1 parámetros en la

ecuación (14) bajo la hipótesis para determinar si las variables independientes influyen significativamente en la probabilidad del suceso del modelo relacionado a la variable del resultado del siguiente modo:

H_0 : $ß_1 = ß_2 = ... = ß_k = 0$

H_1 : para algún $ß_i$ 0

Estadístico de prueba

$$D \sim X^2 \text{ con k grados de libertad} \qquad (23)$$

Decisión: si $D > X^2_{\alpha,k}$ rechazamos H_0, entonces al menos uno de los coeficientes es diferente de cero y la variable correspondiente influye en la probabilidad del suceso estudiado.

Una vez encontrado el mejor conjunto de variables explicativas que predicen la variable Y, luego se debe evaluar mediante Wald cada coeficiente para determinar cuál o cuáles ingresan al modelo.

i. ESTADÍSTICO WALD

Evalúa la significancia de los coeficientes se define como el vector matriz de los coeficientes estimados del siguiente modo según las hipótesis:

H_0 : $B_i = 0$

H_1 : algún B_i ‡ 0

Estadístico de prueba

$$W = \hat{b}'\left[\hat{I}\left(\hat{b}\right)\right]^{-1}\hat{b} = \hat{b}'\left(X'VX\right)\hat{b} \sim \mathrm{X}^2_{a,k+1} \qquad (24)$$

donde:

$X_{nx(k+1)}$ y V_{nxn}, son las matrices de las ecuaciones (21) y (22).

Decisión: si W > $X^2_{\alpha,k}$ rechazamos H_0 con un nivel de significancia fijado á, concluimos que la variable independiente influye en la probabilidad del suceso.

3.4.3 PRUEBAS DE BONDAD DE AJUSTE

Para evaluar la bondad de ajuste del modelo se utiliza la prueba de Hosmer-Lemeshow, consiste en calcular para cada observación del conjunto de datos las probabilidades de la variable dependiente que predice el modelo, se agrupa en aproximadamente 10 grupos iguales a partir de las probabilidades esperadas y se compara con las frecuencias observadas mediante una prueba X^2 con j-2 grados de libertad, donde j es el número de grupos formados como se explicó en el modelo simple (3.3.3.3). El modelo se ajusta bien si no hay evidencias para rechazar la hipótesis nula.

3.4.4 INTERPRETACIÓN DE LOS RESULTADOS

La interpretación de los resultados obtenidos se realiza a partir de la interpretación de los coeficientes del modelo. Para ello basta tener en cuenta que si el modelo ajustado es bueno, entonces se dice que el modelo es significativo, pero

además se debe analizar el grado de asociación estadística que existen en sus parámetros, a partir de la ecuación (4) se tiene:

$$\log\left[\frac{p(X_1,\ldots,\ X_k;\beta)}{1-p(X_1,\ldots,\ X_k;\beta)}\right] = \beta_0 + \beta_1 X_1 + \ldots.. + \beta_k X_k$$

donde el "odds ratio" que es el factor de riesgo está dado por la razón de esta expresión:

$$\frac{p(X_1,\ldots,X_k;\beta)}{1-p(X_1,\ldots,X_k;\beta)} = e^{\beta_0+\beta_1 X_1+\ldots.+\beta_k X_k}$$

entonces:

$$\frac{\dfrac{p(X_1+1,\ldots,X_k;\beta)}{1-p(X_1+1,\ldots,X_k;\beta)}}{\dfrac{p(X_1,\ldots,X_k;\beta)}{1-p(X_1,\ldots,X_k;\beta)}} = e^{\beta_1}$$

Por lo tanto, e^{β_1} es el factor de cambio en el "odds ratio" (OR) de riesgo si el valor de la variable X_1 cambia en una unidad. Así, si $ß_1 > 0$ (ó $ß_1 < 0$) el factor será mayor que 1 y $p(X_1,\ldots,X_k; ß)$ aumentará (disminuirá). Si $ß_1 = 0$ la variable X_1 no ejerce ningún efecto sobre p_i.
$ß_0$ es un ajuste de escala. Su mejor interpretación se obtiene calculando el valor de $p(X_1,\ldots,X_k;ß)$ en los valores medios de $X_1,\ldots,X_k$ y usar como variables explicativas sus valores estandarizados.

En regresión logística la medida de asociación más empleada es el OR debido que el número *e* es la base de los logaritmos

neperianos y elevados a un coeficiente de regresión logística del factor, si es mayor que 1 supone un aumento unitario, indica que el factor de riesgo es mayor.

Si el modelo de Regresión Logística es significativo y una de las variables independientes es dicotómica con valores de 0 y 1, el número *e* elevado al coeficiente de regresión logística es el OR, denominado factor de riesgo o protección que implica un aumento unitario de la variable independiente. En el caso de una variable cuantitativa, *e* elevado a $ß_1$ es el número de veces que aumenta la probabilidad de padecer una enfermedad por cada unidad de aumento de la variable independiente, o dicha de otra manera, cuántas veces es más probable que padezca la enfermedad una persona que presenta síntomas relacionadas a ella.

CAPÍTULO IV

ANALISIS DE LOS DATOS

4.1 DESCRIPCIÓN DE LOS DATOS

Como ya se mencionó anteriormente, el tamaño de la población con la que se está trabajando es de 149 observaciones, 69 de ellas presentan la enfermedad Angina de Pecho y el resto no.

Las variables observadas son:

Angina (Y_1) : Presencia de la enfermedad Angina de Pecho en el paciente.
0 : No se presenta.
1 : Sí se presenta.

Edad (X_1) : Edad en años cumplidos del paciente.

Sexo (X_2) : Sexo del paciente.
0 : Femenino.
1 : Masculino.

Colesterol (X_3) : Cantidad de colesterol que contiene la sangre del paciente.
Unidad de Medida : mg/dl.
Valor Normal : 140 – 200 mg/dl.

Triglicéridos (X_4) : Cantidad de triglicéridos que contiene la sangre del paciente.
Unidad de Medida : mg/dl.
Valor Normal : 45 – 150 mg/dl.

Glucosa (X_5) : Cantidad de glucosa en la sangre del paciente.
Unidad de Medida : mg%.
Valor Normal : 70 – 110 mg%.

HTA (X_6) : El paciente Hipertenso. Información proviene de la Historia Clínica.
0 : No la presenta.
1 : Sí la presenta.

Obesidad (X_7) : Paciente con diagnóstico de obesidad. Información proviene de la Historia Clínica.
0 : No tiene.
1 : Sí tiene.

4.2 MODELO DE REGRESIÓN LOGÍSTICA

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7)}}$$

Hipótesis:

H_0 : $ß_1 = ß_2 = ... = ß_7 = 0$
H_1 : $ß_i \neq 0$ para $i = 1,...,7$
0 equivalente
H_0 : Las variables independientes (X_1, X_2,...,X_7) no influyen significativamente sobre la enfermedad Angina de Pecho (Y_1)
H_1 : Las variables independientes (X_1, X_2,...,X_7) influyen significativamente sobre la enfermedad Angina de Pecho (Y_1)

4.3 AJUSTE DEL MODELO DE REGRESIÓN LOGÍSTICA

4.3.1 EVALUACIÓN DEL INCREMENTO DEL ESTADÍSTICO –2LLO – BONDAD DE AJUSTE

Este ajuste se efectúa mediante el software estadístico SPSS vs. 9.0.

La Bondad de Ajuste del modelo 1: los resultados se muestran en el Anexo N° 1.

Para el modelo que contiene sólo el término constante, -2 Log Likelihood, es igual a 205,42219, que se distribuye como $X^2_{(n-1)} = X^2_{(149-1)} = X^2_{(148)}$

Cuando las 7 variables explicativas –2 Log Likelihood es igual a 98.797, para este estadístico, las hipótesis son:

H_0 : El modelo se ajusta perfectamente
H_1 : El modelo no se ajusta perfectamente

o equivalentemente

H_0 : -2 Log Likelihood = 1
H_1 : -2 Log Likelihood ‡ 1

Esta estadística se distribuye como un $X^2_{(n-p)}$, donde p = k+1, y se rechaza la hipótesis nula (H_0), si su valor es mayor que $X^2_{(n-p)}$ (á), con á de nivel de significación (0.05) =169.71

Como se observa 98.797 no es mayor que 169.71, por lo tanto no se rechaza la hipótesis nula; entonces, el modelo se ajusta perfectamente.

La estadística "Modelo Chi-Square" toma valor igual a 106.625 que es igual a la diferencia entre –2 Log Likelihood que contiene sólo a la constante y el modelo con las 7 variables explicativas. En este caso, las hipótesis estadísticas son:

H_0 : $ß_1 = ... = ß_7 = 0$

H_1 : $ß_i \neq 0$ para por lo menos un i = 1,...,7

En este caso el valor crítico es $X^2_{(7)}$ (0.05) = 14.067, como se observa 106.625 es mayor que 14.067, por tanto se rechaza la hipótesis nula; entonces, los coeficientes de las 7 variables explicativas del modelo son diferentes de cero excepto la constante.

A continuación se observa una tabla 2x2; en dicha tabla se muestran los casos observados de la enfermedad Angina de Pecho frente a los casos estimados de la enfermedad. En esta tabla se enfrentan valores estimados y observados de la enfermedad, calculando el porcentaje de coincidencias. Para el presente trabajo, hay un 84.56% de coincidencias.

4.3.2 PARÁMETROS ESTIMADOS DEL MODELO

En la Tabla de Variables in the equation se nos muestran los parámetros del modelo, tomando en cuenta 5 iteraciones que es lo que nos arrojó por defecto el paquete SPSS vs. 9.0.

Los parámetros del modelo son:

$ß_0$ = -12.4117, $ß_1$ = 0.0830, $ß_2$ = 0.4296, $ß_3$ = 0.0348, $ß_4$ = 0.0123, $ß_5$ = 0.0021, $ß_6$ = -2.3912, $ß_7$ = -0.0611

Entonces, la ecuación de regresión logística para predecir la ocurrencia de la Enfermedad Angina de Pecho es:

$$P = \frac{1}{1+e^{-Z}}$$

siendo:

$$Z = -12.4117 + 0.0838\ X_1 + 0.4296\ X_2 + 0.0348\ X_3 + 0.0123\ X_4 + 0.0021\ X_5 - 2.3912\ X_6 - 0.0611\ X_7$$

4.3.2.1 ESTADÍSTICO WALD

En cuanto a la prueba de hipótesis para coeficientes individuales, se efectúa mediante la estadística de WALD.

Las hipótesis son las siguientes:

H_0 : $ß_i = 0$

H_1 : algún $ß_i \neq 0$ para por lo menos un $i = 1,...,7$

Esta estadística se distribuye como una $X^2_{(1)}$ si la variable explicativa es cuantitativa y si la variable explicativa es de tipo categórica, se distribuye como una $X^2_{(C-1)}$, donde C es igual al número de categorías que toma la variable; para nuestro caso C=2, entonces $X^2_{(C-1)} = X^2_{(2-1)}$, = $X^2_{(1)}$.

El valor crítico al 7% de nivel de significancia es 3.76, por lo tanto se rechaza la hipótesis nula si el valor de la estadística WALD es mayor que el valor crítico indicado.

Para efectuar la dócima, tomamos los valores que están debajo de la columna etiquetada como WALD (Variables in the Equatión – Anexo 1).

Veamos qué pasa con las variables explicativas.

i) EDAD no se rechaza la hipótesis nula, por tanto es una variable significativa, también llegamos al mismo resultado usando los valores de las probabilidades asociados a la estadística WALD y éstas se encuentran debajo de la columna etiquetada con SIG.

ii) Las variables Sexo, Glucosa y Obesidad en la columna WALD presentan los siguientes valores en el mismo orden 0.6247, 0.0327, 0.0137, debido a que este valor es menor que el valor crítico (3.76), entonces diremos que no se rechaza la hipótesis nula; por lo tanto, dichas variables no son significativas.

iii) Con respecto a las variables explicativas Colesterol, Triglicéridos, Hipertensión ocurre lo mismo que con la variable Edad, es decir, rechazamos la hipótesis nula, entonces dichas variables son significativas.

A continuación, el modelo será el siguiente:

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5)}}$$

4.3.2.2 CORRELACIÓN PARCIAL (R)

La contribución de cada variable explicativa es difícil de calcular debido a que la influencia depende del resto de variables; por ello usaremos la estadística que nos permita conocer aproximadamente la correlación parcial de cada variable. Ésta se encuentra en la columna etiquetada con R.

Si el valor de la estadística de WALD es menor que 2k, entonces R es igual a cero, lo que indica que la contribución de la variable explicativa correspondiente en el modelo es pequeña.

R puede tomar valores entre 0 y 1.

Ahora, para las variables Sexo, Glucosa y Obeso, los resultados obtenidos en R es cero; por lo cual, no contribuyen en el modelo.

Lo cual corrobora lo anteriormente mencionado en las dócimas de la estadística de WALD.

4.3.2.3 ODDS RATIO [Exp (ß)]

El ODDS RATIO cambia cuando la i-ésima variable explicativa regresora se incrementa en una unidad; así:

i) Si $\hat{\beta}_j > 1$, significa que el ODDS RATIO se incrementa.

ii) Si $\hat{\beta}_j < 1$, significa que el ODDS RATIO decrece.

iii) Si $\beta_j = 0$, significa que el factor es igual a uno, lo cual hace que ODDS RATIO no varía.

Entonces, en la columna Exp (ß) se observa que:

- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_6 > 1$, por lo tanto, las variables Edad, Sexo, Colesterol, Triglicéridos, Glucosa, son factores que incrementan la probabilidad de ocurrencia de la enfermedad Angina de Pecho.

- Para β_5 y $\beta_7 < 1$, por lo tanto, las variables Hipertensión y Obeso son factores que disminuyen la probabilidad de la ocurrencia de la enfermedad Angina de Pecho.

4.4 MATRIZ DE CORRELACIÓN

En esta misma corrida podemos ver la matriz de correlación para las variables explicativas (X_1, X_2,...,X_7) y la variable dependiente (Y_1 = Enfermedad Angina de Pecho).

La Matriz de Correlaciones nos muestra correlaciones bajas entre las variables.

4.5 BONDAD DE AJUSTE DEL MODELO 2

Ahora veamos que se observa quitando las variables explicativas Sexo (X_2), Glucosa (X_6) y Obeso (X_7), que no son significativas bajo la corrida anterior, el cual llamaremos Modelo 2.

Estos datos los observamos en el Anexo 2.

4.5.1 BONDAD DE AJUSTE DEL MODELO 2 CON 4 VARIABLES EXPLICATIVAS

Para este modelo que contiene sólo el término constante, -2 Log Likelihood, es igual a 205,42219 y se distribuye como una $X^2_{(148)}$; cuando contiene las 4 variables explicativas – 2 Log

Likelihood, es igual a 99,550, las hipótesis son iguales a las utilizadas en el Modelo 1.

El valor crítico es $X^2_{(144)}$ (0.05) =173.004, como se observa 99,550 no es mayor que 173.004, por lo tanto, no se rechaza la hipótesis nula; entonces el Modelo 2 se ajusta perfectamente.

La estadística "Goodness of Fit" toma el valor 116.952, las hipótesis estadísticas son iguales que en Modelo 1, la estadística se distribuye como una $X^2_{(144)}$. El valor crítico es $X^2_{(144)}$ (0.05) = 173.004.

Como se observa 116.952 no es mayor que 173.004, por lo tanto no rechazamos la hipótesis nula; entonces el modelo se ajusta perfectamente.

La estadística Model Chi-Square, toma valor igual a 105,872.

En este caso, las hipótesis estadísticas son:

$H_0 : ß_1 = ß_2 = ß_3 = ß_4 = 0$

$H_1 : ß_j \neq 0$ para un i = 1,...,4

En este caso, K=4, y el valor crítico es $X^2_{(4)}$ (0.05) = 9.488. Como se observa 116.952 es mayor que 9.488, por tanto, rechazo la hipótesis nula; entonces los coeficientes de las 4 variables explicativas del Modelo 2 son diferentes de cero excepto la constante.

4.5.2 PARÁMETROS ESTIMADOS DEL MODELO 2

Consideramos los resultados de variables en The Equation (Anexo 2), que contienen los coeficientes estimados y se encuentran debajo de la columna B. Así, la ecuación de regresión logística para producir la probabilidad de ocurrencia de Enfermedad Angina de Pecho es similar que la presentada en el Modelo 1.

$$Z = -12.3817 + 0.0875\ X_1 + 0.0344\ X_2 + 0.0133\ X_3 - 2.3962\ X_4$$

X_1 = Edad
X_2 = Colesterol
X_3 = Triglicéridos
X_4 = Hipertensión

En cuanto a la prueba de hipótesis para coeficientes individuales, se efectúa con la estadística WALD; ésta se distribuye en nuestro caso tanto para la variable cuantitativa como categórica, como una $X^2_{(1)}$, ya habiéndose éste explicado en el anterior modelo.

Ahora observemos los valores debajo de la columna WALD y docimamos con un valor crítico al 7% de nivel de significación es 3.76; por tanto, se rechaza la hipótesis nula si dicho valor de la estadística WALD es mayor que el valor crítico indicado.

Veamos debajo de la columna WALD; vemos que pata las 4 variables explicativas Edad, Colesterol, Triglicérido y HTA, el valor de la estadística de Wald es mayor a 3.76 por tanto,

se rechaza la hipótesis nula; luego, son variables explicativas significativas.

Llegamos al mismo resultado usando los valores de probabilidades asociados a la estadística de WALD, encontrándose debajo de la columna SIG.

4.5.3 CORRELACIÓN PARCIAL (R) Y ODDS RATIO (EXP (B))

En cuanto al valor de R, es muy baja dada variable explicativa.

Interpretemos los coeficientes significativos.

Entonces, en la columna Exp (B) observamos que:

Exp($\hat{b}_1, \hat{b}_2, \hat{b}_3) > 1$, por lo tanto, las variables Edad, Colesterol y Triglicéridos son factores para que incremente la probabilidad de ocurrencia de la Enfermedad Angina de Pecho.

Exp($\hat{b}_4) < 1$, por lo tanto, la variable Hipertensión es un factor que disminuye la probabilidad de ocurrencia de la Enfermedad Angina de Pecho.

El comportamiento es idéntico para este Modelo 2 con respecto al Modelo 1.

Por último, en cuanto a la Tabla de Clasificación, vemos que 70 pacientes no sufren de la Enfermedad Angina de Pecho y 55 sí la presentan. Los elementos fuera de la diagonal indican el número de pacientes mal clasificados que en total son 24.

De los pacientes que no presentan la enfermedad Angina de Pecho, un 86.42% fueron correctamente clasificados y de los pacientes con enfermedad Angina de Pecho, un 80.88% se clasificaron correctamente.

Del total 149, un 83.89% fueron correctamente clasificados.

4.6 METODO FORWARD

Hasta el momento trabajé con el Método Enter, método que por defecto nos da el programa spss 9.0, con el cual he hallado cada uno de los estadísticos en los puntos anteriores.

Ahora veamos cómo que nos indica el Método Forward (WALD), el cual lo observamos en el anexo No.3. "Beginning Block Number".

El sistema analiza el estadístico chi-square y su grado de significación de las 4 variables (Colesterol, Hta, Edad y Triglicéridos).

Este sistema se inicia analizando en el primer paso la variable aplicativa Colesterol con 1 grado de libertad, cuyo estadístico es 45.798 y será comprado con $X^2_{(1)}(0.05) = 3.841$, siendo las hipótesis:

$H_0 : B_1 = 0$

$H_1 : B_1 \neq 0$

Entonces $X^2_{(1)}(0.05) = 3.841 < 45.798$, por lo tanto rechazo Ho, la variable explicativa *Colesterol* es significativa, lo cual podemos corroborar con la columna SIG, que es igual a 0.000 menor que 0.05, es decir, la variable entra en el modelo.

Además debe de mencionar una clasificación correcta del 69.80%.

Para el paso 2, considera el modelo con 2 variables Colesterol y HTA, aquí como en el caso anterior, el modelo será:

$H_0 : B_1 = B_2 = 0$

$H_1 : B_1 = B_2 \neq 0$

La estadística que nos da la corrida es 70.453 que es mayor que $X^2_{(1)}(0.05) = 5.991$, y con un SIG = 0.000 < 0.05, con lo cual concluimos que se rechaza la hipótesis nula, es decir, tanto la variable colesterol como HTA son significativas, con una clasificación correcta del 78.52%.
En el paso 3, tenemos la estadística chi-square igual a 96.269, con al columna SIG = 0.000 y en 81.83% correcta de clasificación de las variables explicativas Colesterol, HTA, Edad.

La estadística $X^2_{(1)}(0.05) = 7.815$, que como observamos es menor que la mostrada en la corrida del SPSS (96.269), es decir, rechazamos la hipótesis nula, las variables mencionadas anteriormente son significativas.

Paso 4:

Este es el último paso, lo que vemos es que las cuatro variables explicativas son significativas, las tres mencionadas anteriormente y la última Triglicéridos.

La estadística chi-square que arroja la corrida del SPSS vs 9.0 es 105.872 que es mayor a la estadística $X^2_{(1)}(0.05) =$

9.488. con un SIG = 0.000 menor al 5% con el que se está haciendo la comparación. Las hipótesis:

H_0 : $B_i = 0$, para i = 1,...,4.
H_1 : $B_i \neq 0$, para i = 1,...,4.

Para este último Paso, rechazamos la hipótesis nula, es decir las variables explicativas Colesterol, HTA, Edad y Triglicéridos son significativas y son las que se mantendrán en el modelo.

El Modelo es:

$$P=\frac{1}{1+e^{-Z}}$$

$$P=\frac{1}{1+e^{-(b_0+b_1X_1+b_3X_3+b_4X_4+b_6X_6)}}$$

donde:

Z = -12.3817 + 0.0875 edad + 0.0344 colesterol + 0.0133 triglicéridos – 2.3962 hipertensión.

Ejemplo:

Considerando las variables del paciente No. 14 y aplicándo al modelo obtenido, tenemos:

Edad = 77 años.

Colesterol = 169 mg/dl.

Triglicéridos = 236 mg/dl.

HTA = 1

Z = 0.912

Entonces el valor es P = 0.7134, donde

La probabilidad de que el paciente N. 14 presente la enfermedad Angina de Pecho es de 0.7134, es decir que este paciente presenta factores de riesgo elevados.

CONCLUSIONES

1. El análisis de Regresión Logística es muy utilizado en el campo de la Salud y nos brinda un método excelente para poder tomar en cuenta variables explicativas cuantitativa y categóricas.

2. No cabe duda que la Regresión Logística es una herramienta estadística con mejor capacidad para el análisis de datos en investigación clínica y epidemiológica, de ahí su amplía utilización.

3. Esta técnica nos sirve para identificar factores de riesgo y factores de prevención en enfermedades de muestras prospectivas.

4. Para determinar Factores de Riesgo que influyen en la Enfermedad Angina de Pecho, se utilizó el método ya mencionado líneas arriba; para el Modelo 1 incluyendo las 7 variables explicativas, edad, sexo, colesterol, triglicéridos, glucosa, hipertensión y obesidad, encontramos que de todas ellas, las variables significativas y que aportan en la influencia de la presencia de la Enfermedad Angina de Pecho, fueron Edad ($p=0.003$), colesterol ($p=0.008$), triglicéridos ($p=0.0143$) e hipertensión ($p=0.000$), a un nivel de significancia del 5%.

5. El Modelo 2 consta sólo de las últimas cuatro variables mencionadas en el punto anterior y ocurre exactamente lo mismo.

6. También se utilizó el Método FORWARD, tomando en cuenta todas las variables y el resultado arrojó que las mismas variables Edad, colesterol, triglicéridos e hipertensión, son los factores de riesgo que influyen en la Enfermedad Angina de Pecho.

7. Descartamos, por ende, las variables explicativas Sexo (p=0.4293), obesidad (p=0.9067) y glucosa (p=0.8565), las cuales no aportan gran influencia en la presencia de la Enfermedad Angina de Pecho, considerando nivel de significancia del 5%.

8. El ODDS Ratio para la variable edad es 1.0914, para la variable colesterol es igual a 1.0350 y para la variable triglicéridos es 1.0134 (ver anexo #3), que son mayores a 1, por lo tanto son factores que incrementan la probabilidad de ocurrencia de la enfermedad Angina de Pecho.

9. Mientras que para la variable hipertensión (HTA)presenta un ODDS Ratio igual a 0.0911 menor que 1, dicha variable es un factor que disminuye la probabilidad de ocurrencia de la enfermedad Angina de Pecho.

10. En la tabla de clasificación 70 pacientes no sufren de la enfermedad que significa que el 86.42%, (ver anexo #3), de estos están correctamente clasificados.

11. Para los pacientes que si sufren esta enfermedad 55 de ellos están correctamente clasificados esto es el 80.80%. (ver anexo #3).

12. La suma de los elementos que se encuentran fuera de la diagonal es 24, y nos indica el número de pacientes mal clasificados.

13. Del total de la tabla de clasificación 125 pacientes es decir 83.89% fueron correctamente clasificados.

RECOMENDACIONES

1. Debemos enfocar nuestra ayuda a los pacientes en estudio, en base a un mejor control de dichos exámenes, como colesterol y triglicéridos.

2. Educar y concientizar a los pacientes que sufren de la Enfermedad Angina de Pecho de los riesgos que influyen en su enfermedad.

3. Promover charlas grupales que hablen de los riesgos de dicha Enfermedad, enriqueciendo los conocimientos del público asistente.

4. Seguimiento constante de los pacientes que presentan dicha enfermedad para un mejor control de los factores de riesgo.

5. Para un estudio más completo de este tipo, se recomienda obtener información de éstos factores de riesgo y de otros que no han sido considerados en éste estudio para poder determinar con más exactitud los principales factores de riesgo que influyen en la presencia de la enfermedad Angina de Pecho.

6. No se consideró los factores como Tabaquismo, Inactividad física, Diabetes Mellitus, Historia familiar de enfermedades del corazón, entre otras, debido no se contaba con dicha información, y la que se pudo hallar no se encontraba completa, se recomienda al médico que se encuentra atendiendo, mayor énfasis en la recolección de éstos datos.

BIBLIOGRAFÍA

1. Abanto, A.(2000). Modelos de regresión logística y neuronales. Escuela de Ingeniería UNI Lima, Perú.

2. Abraira,V.(1996). Métodos multivariados en bioestadística.Ed. Centro de estudios Ramón Areces. Madrid, España.

3. Alvarez, Rafael (1995).Estadística multivariante y no paramétrica con SPSS. Aplicación a las ciencias de la salud.Diaz de Santos S.A Madrid, España.

4. Andersen. E.B.(1990). Introducción to the statistical. analysis of Categorical data. New York, Estados Unidos.

5. Hosmer, D. Lemeshow,S.(1989). Applied logistic Regression. John Wiley &Sons. New York, U.S.A.

6. Juez, P., Díez,F.J(1996). Probabilidad y Estadística en Medicina. Madrid, España.

7. Molinero, L.(2000).Regresión logística. Almirall Prodesfarma.Madrid, España.

8. Nolberto, V. A. (2000). Medidas de adecuación para un modelo de regresión logística. U.N.M.S.M. Lima, Perú.

9. Silva, A.(1990). Excursión a la regresión logística en Ciencias de la salud. Madrid, España.

10. Visauta, B.(1998). Análisis estadístico con SPSS para Windows. McGraw Hill / interamericana de España,S.A.U. Madrid. España.

ANEXO # 1

```
      Total number of cases:      149 (Unweighted)
      Number of selected cases:   149
      Number of unselected cases: 0

      Number of selected cases:                 149
      Number rejected because of missing data:  0
      Number of cases included in the analysis: 149


Dependent Variable Encoding:

Original       Internal
Value          Value
0              0
1              1
_


                                    Parameter
                  Value   Freq      Coding
                                      (1)
OBESO
                   ,00     78    1,000
                  1,00     71     ,000
HTA
                   ,00     86    1,000
                  1,00     63     ,000
SEXO
              M            64    1,000
              F            85     ,000
_


Dependent Variable..   Y1

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   205,42219

* Constant is included in the model.
```

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number

1.. EDAD
COLESTER
TRIGLICE
GLUCO
SEXO
HTA
OBESO

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than ,01 percent.

Iteration History:

Iteration	Log Likelihood	Constant	EDAD	COLESTER	TRIGLICE
1	-57,832895	-5,544597	,03427607	,01559815	,00642020
2	-50,812073	-9,115945	,06011700	,02532791	,00957683
3	-49,481123	-11,539049	,07748980	,03217290	,01151356
4	-49,398910	-12,345439	,08266913	,03457428	,01219596
5	-49,398443	-12,411738	,08303616	,03478291	,01226040

GLUCO	SEXO(1)	HTA(1)	OBESO(1)
,00080699	,45769440	-1,2446629	-,05257534
,00124831	,49908976	-1,8617311	-,06886971
,00173541	,45164387	-2,2544964	-,06255724
,00208281	,43138221	-2,3809864	-,06090549
,00212583	,42959375	-2,3911580	-,06105541

-2 Log Likelihood	98,797
Goodness of Fit	119,320
Cox & Snell - R^2	,511
Nagelkerke - R^2	,683

	Chi-Square	df	Significance
Model	106,625	7	,0000
Block	106,625	7	,0000
Step	106,625	7	,0000

---------- Hosmer and Lemeshow Goodness-of-Fit Test----------

	Y1 = 0		Y1 = 1		
Group	Observed	Expected	Observed	Expected	Total
1	15,000	14,896	,000	,104	15,000
2	14,000	14,497	1,000	,503	15,000
3	14,000	13,868	1,000	1,132	15,000
4	13,000	12,644	2,000	2,356	15,000
5	10,000	10,532	5,000	4,468	15,000
6	8,000	7,701	7,000	7,299	15,000
7	5,000	4,615	10,000	10,385	15,000
8	2,000	1,559	13,000	13,441	15,000
9	,000	,587	15,000	14,413	15,000
10	,000	,101	14,000	13,899	14,000

	Chi-Square	df	Significance
Goodness-of-fit test	1,7052	8	,9888

Classification Table for Y1
The Cut Value is ,50

		Predicted 0	Predicted 1	Percent Correct
Observed				
0	0	71	10	87,65%
1	1	13	55	80,88%
	Overall			84,56%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R
EDAD	,0830	,0230	12,9817	1	,0003	,2312
COLESTER	,0348	,0104	11,2324	1	,0008	,2120
TRIGLICE	,0123	,0050	6,0011	1	,0143	,1396
GLUCO	,0021	,0118	,0327	1	,8565	,0000
SEXO(1)	,4296	,5435	,6247	1	,4293	,0000
HTA(1)	-2,3912	,5671	17,7781	1	,0000	-,2771
OBESO(1)	-,0611	,5210	,0137	1	,9067	,0000
Constant	-12,4117	2,5904	22,9586	1	,0000	

_

		95% CI for Exp(B)	
Variable	Exp(B)	Lower	Upper
EDAD	1,0866	1,0386	1,1368
COLESTER	1,0354	1,0145	1,0567
TRIGLICE	1,0123	1,0025	1,0223
GLUCO	1,0021	,9793	1,0255
SEXO(1)	1,5366	,5295	4,4590
HTA(1)	,0915	,0301	,2781
OBESO(1)	,9408	,3389	2,6118

Correlation Matrix:

	Constant	EDAD	COLESTER	TRIGLICE	GLUCO	SEXO(1)	HTA(1)
Constant	1,00000	-,53164	-,67606	-,20018	-,38112	,01709	,23268
EDAD	-,53164	1,00000	-,01367	,07755	,00979	-,19644	-,30345
COLESTER	-,67606	-,01367	1,00000	-,05295	-,03218	,08634	-,25998
TRIGLICE	-,20018	,07755	-,05295	1,00000	-,03891	-,19365	-,06464
GLUCO	-,38112	,00979	-,03218	-,03891	1,00000	-,09748	,04624

SEXO(1)	,01709	-,19644	,08634	-,19365	-,09748	1,00000	-,02719
HTA(1)	,23268	-,30345	-,25998	-,06464	,04624	-,02719	1,00000
OBESO(1)	-,15399	,01809	,02438	,06406	-,03683	,16688	,04921

	OBESO(1)
Constant	-,15399
EDAD	,01809
COLESTER	,02438
TRIGLICE	,06406
GLUCO	-,03683
SEXO(1)	,16688
HTA(1)	,04921
OBESO(1)	1,00000

Observed Groups and Predicted Probabilities

```
      20


          0

F         0

R     15  0

E         0

Q         0 0
1
U         0 0
1
E     10  0 0
1
N         0 0
1
C         0 0
1   111
Y         0 0   1
1   111
       5  0 0 0 0
1   111
          0 0 010    1                    1
1 1 111
          000 000  000 10 0 1    11 1110 101
11   0111111
          00000000 0000000000   000 10000000 1  1 011  0
101110111111
Predicted

  Prob:   0           ,25           ,5            ,75
1
  Group:
00000000000000000000000000000001111111111111111111111111111111
```

Predicted Probability is of Membership for 1
The Cut Value is ,50

```
          Symbols: 0 - 0
                   1 - 1
          Each Symbol Represents 1,25 Cases.
```

—

CASE	Observed Y1	Pred	PGroup	Resid	ZResid
13	S 1 **	,0957	0	,9043	3,0738
19	S 1 **	,1064	0	,8936	2,8986
52	S 1 **	,0378	0	,9622	5,0446
70	S 0 **	,8966	1	-,8966	-2,9454
140	S 0 **	,8983	1	-,8983	-2,9717

S=Selected U=Unselected cases
** = Misclassified cases

* Cases with studentized residuals greater than 2 are listed.
The Cut Value is ,50

ANEXO #2

```
      Total number of cases:      149 (Unweighted)
      Number of selected cases:   149
      Number of unselected cases: 0

      Number of selected cases:                 149
      Number rejected because of missing data:  0
      Number of cases included in the analysis: 149


Dependent Variable Encoding:

Original       Internal
Value          Value
0              0
1              1
_

                                 Parameter
                  Value   Freq   Coding
                                   (1)
HTA
                    ,00     86  1,000
                   1,00     63   ,000
_

Dependent Variable..   Y1

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   205,42219

* Constant is included in the model.
```

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number
1.. EDAD
COLESTER
TRIGLICE
HTA

Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than ,01 percent.

Iteration History:

Iteration	Log Likelihood	Constant	EDAD	COLESTER	TRIGLICE
1	-58,680179	-5,401437	,03604383	,01511577	,00705557
2	-51,256827	-9,065615	,06434327	,02464811	,01053141
3	-49,859967	-11,520665	,08203726	,03169375	,01256355
4	-49,775590	-12,317793	,08710706	,03421798	,01325085
5	-49,775128	-12,381687	,08745556	,03443740	,01331383

HTA(1)
-1,2360176
-1,8510864
-2,2555498
-2,3858674
-2,3961672

-2 Log Likelihood	99,550
Goodness of Fit	116,952
Cox & Snell - R^2	,509
Nagelkerke - R^2	,680

	Chi-Square	df	Significance
Model	105,872	4	,0000
Block	105,872	4	,0000
Step	105,872	4	,0000

---------- Hosmer and Lemeshow Goodness-of-Fit Test----------

	Y1 = 0		Y1 = 1		
Group	Observed	Expected	Observed	Expected	Total
1	15,000	14,905	,000	,095	15,000
2	14,000	14,463	1,000	,537	15,000
3	14,000	13,816	1,000	1,184	15,000
4	13,000	12,495	2,000	2,505	15,000
5	9,000	10,627	6,000	4,373	15,000
6	9,000	7,830	6,000	7,170	15,000
7	5,000	4,628	10,000	10,372	15,000
8	2,000	1,558	13,000	13,442	15,000
9	,000	,591	15,000	14,409	15,000
10	,000	,088	14,000	13,912	14,000

	Chi-Square	df	Significance
Goodness-of-fit test	2,7699	8	,9480

Classification Table for Y1
The Cut Value is ,50

		Predicted 0	Predicted 1	Percent Correct
Observed 0	0	70	11	86,42%
1	1	13	55	80,88%
		Overall		83,89%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R
EDAD	,0875	,0225	15,1495	1	,0001	,2530
COLESTER	,0344	,0104	11,0513	1	,0009	,2099
TRIGLICE	,0133	,0049	7,2742	1	,0070	,1602
HTA(1)	-2,3962	,5644	18,0269	1	,0000	-,2793
Constant	-12,3817	2,3632	27,4507	1	,0000	

		95% CI for Exp(B)	
Variable	Exp(B)	Lower	Upper
EDAD	1,0914	1,0444	1,1405
COLESTER	1,0350	1,0142	1,0563
TRIGLICE	1,0134	1,0036	1,0233
HTA(1)	,0911	,0301	,2753

_

Correlation Matrix:

	Constant	EDAD	COLESTER	TRIGLICE	HTA(1)
Constant	1,00000	-,58716	-,75892	-,23989	,28772
EDAD	-,58716	1,00000	,00792	,04353	-,32330
COLESTER	-,75892	,00792	1,00000	-,03720	-,26285
TRIGLICE	-,23989	,04353	-,03720	1,00000	-,06794
HTA(1)	,28772	-,32330	-,26285	-,06794	1,00000

Observed Groups and Predicted Probabilities

```
      16  0
          0
1
          0
1
F         0
1
R     12  0
1
E         0
1
Q         0
1
U         0
1
E      8  0 10
1
N         0 00
1
C         0 00   0
1    1
Y         0000   0
1  111
       4  0000   0     0                1
11 111
          0000 0 0    00       1        1
1 111111
          00000010 1  0010   110    0 1 01 0   1   1     1
1 11 11111
          00000000100000001100001   010000001  101 1 0  10
1110 011111
Predicted

  Prob:   0           ,25           ,5            ,75
1
  Group:
00000000000000000000000000000111111111111111111111111111111

          Predicted Probability is of Membership for 1
          The Cut Value is ,50
          Symbols: 0 - 0
```

1 - 1
Each Symbol Represents 1 Case.

—

CASE	Observed Y1	Pred	PGroup	Resid	ZResid
13	S 1 **	,1076	0	,8924	2,8800
19	S 1 **	,1390	0	,8610	2,4891
52	S 1 **	,0397	0	,9603	4,9178
70	S 0 **	,9157	1	-,9157	-3,2957
140	S 0 **	,8817	1	-,8817	-2,7304

S=Selected U=Unselected cases
** = Misclassified cases

* Cases with studentized residuals greater than 2 are listed.
The Cut Value is ,50

ANEXO #3

```
      Total number of cases:      149 (Unweighted)
      Number of selected cases:   149
      Number of unselected cases: 0

      Number of selected cases:                 149
      Number rejected because of missing data:  0
      Number of cases included in the analysis: 149


Dependent Variable Encoding:

Original       Internal
Value          Value
0              0
1              1
_


                                 Parameter
                  Value   Freq   Coding
                                   (1)
SEXO
              M             64  1,000
              F             85   ,000
HTA
                   ,00      86  1,000
                  1,00      63   ,000
_


Dependent Variable..   Y1

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   205,42219

* Constant is included in the model.
```

Beginning Block Number 1. Method: Forward Stepwise (WALD)

Step	Improv. Chi-Sq.	df	sig	Model Chi-Sq.	df	sig	Correct Class %	Variable
1	45,798	1	,000	45,798	1	,000	69,80	IN: COLESTER
2	24,655	1	,000	70,453	2	,000	78,52	IN: HTA
3	25,816	1	,000	96,269	3	,000	81,88	IN: EDAD
4	9,603	1	,002	105,872	4	,000	83,89	IN: TRIGLICE

No more variables can be deleted or added.

End Block Number 1 PIN = ,0500 Limits reached.

Final Equation for Block 1

Estimation terminated at iteration number 5 because Log Likelihood decreased by less than ,01 percent.

Iteration History:

Iteration	Log Likelihood	Constant	EDAD	COLESTER	TRIGLICE
1	-58,680179	-5,401437	,03604383	,01511577	,00705557
2	-51,256827	-9,065615	,06434327	,02464811	,01053141
3	-49,859967	-11,520665	,08203726	,03169375	,01256355
4	-49,775590	-12,317793	,08710706	,03421798	,01325085
5	-49,775128	-12,381687	,08745556	,03443740	,01331383

HTA(1)
-1,2360176
-1,8510864
-2,2555498

-2,3858674
-2,3961672

-2 Log Likelihood	99,550
Goodness of Fit	116,952
Cox & Snell - R^2	,509
Nagelkerke - R^2	,680

	Chi-Square	df	Significance
Model	105,872	4	,0000
Block	105,872	4	,0000
Step	9,603	1	,0019

---------- Hosmer and Lemeshow Goodness-of-Fit Test----------

	Y1 = 0		Y1 = 1		
Group	Observed	Expected	Observed	Expected	Total
1	15,000	14,905	,000	,095	15,000
2	14,000	14,463	1,000	,537	15,000
3	14,000	13,816	1,000	1,184	15,000
4	13,000	12,495	2,000	2,505	15,000
5	9,000	10,627	6,000	4,373	15,000
6	9,000	7,830	6,000	7,170	15,000
7	5,000	4,628	10,000	10,372	15,000
8	2,000	1,558	13,000	13,442	15,000
9	,000	,591	15,000	14,409	15,000
10	,000	,088	14,000	13,912	14,000

	Chi-Square	df	Significance
Goodness-of-fit test	2,7699	8	,9480

Classification Table for Y1
The Cut Value is ,50

		Predicted		
		0	1	Percent Correct
		0	1	
Observed				
0	0	70	11	86,42%

1	1	13	55	80,88%
			Overall	83,89%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R
EDAD	,0875	,0225	15,1495	1	,0001	,2530
COLESTER	,0344	,0104	11,0513	1	,0009	,2099
TRIGLICE	,0133	,0049	7,2742	1	,0070	,1602
HTA(1)	-2,3962	,5644	18,0269	1	,0000	-,2793
Constant	-12,3817	2,3632	27,4507	1	,0000	

_

		95% CI for Exp(B)	
Variable	Exp(B)	Lower	Upper
EDAD	1,0914	1,0444	1,1405
COLESTER	1,0350	1,0142	1,0563
TRIGLICE	1,0134	1,0036	1,0233
HTA(1)	,0911	,0301	,2753

Correlation Matrix:

	Constant	EDAD	COLESTER	TRIGLICE	HTA(1)
Constant	1,00000	-,58716	-,75892	-,23989	,28772
EDAD	-,58716	1,00000	,00792	,04353	-,32330
COLESTER	-,75892	,00792	1,00000	-,03720	-,26285
TRIGLICE	-,23989	,04353	-,03720	1,00000	-,06794
HTA(1)	,28772	-,32330	-,26285	-,06794	1,00000

Observed Groups and Predicted Probabilities

```
      16  0
          0
1
          0
1
F         0
1
R     12  0
1
E         0
1
Q         0
1
U         0
1
E      8  0 10
1
N         0 00
1
C         0 00   0
1    1
Y         0000   0
1  111
       4  0000   0     0                1
11 111
          0000 0 0    00       1        1
1 111111
          00000010 1  0010   110    0 1 01 0   1   1     1
1 11 11111
          00000000100000001100001   010000001  101 1 0  10
1110 011111
Predicted

  Prob:   0           ,25           ,5            ,75
1
  Group:
00000000000000000000000000000011111111111111111111111111111111

          Predicted Probability is of Membership for 1
          The Cut Value is ,50
          Symbols: 0 - 0
```

1 - 1
Each Symbol Represents 1 Case.

—

CASE	Observed Y1	Pred	PGroup	Resid	ZResid
13	S 1 **	,1076	0	,8924	2,8800
19	S 1 **	,1390	0	,8610	2,4891
52	S 1 **	,0397	0	,9603	4,9178
70	S 0 **	,9157	1	-,9157	-3,2957
140	S 0 **	,8817	1	-,8817	-2,7304

S=Selected U=Unselected cases
** = Misclassified cases

* Cases with studentized residuals greater than 2 are listed.
The Cut Value is ,50

BASE DE DATOS

Dato	y1	Edad	Sexo	Colesterol	Triglicéridos	Hipertensión	Glucosa	Obesidad
1	1	45	M	226	337	1	105	1
2	1	69	F	158	174	0	112	0
3	1	64	M	235	266	0	87	0
4	1	66	M	240	214	1	77	0
5	1	68	F	152	138	1	86	1
6	1	75	M	152	138	0	86	1
7	1	67	M	225	302	0	120	0
8	1	50	M	173	159	1	69	0
9	1	77	F	170	46	1	85	0
10	1	72	M	224	98	0	82	1
11	1	68	M	217	181	1	79	1
12	1	71	M	203	163	1	84	0
13	1	54	F	183	123	0	94	1
14	1	77	F	169	236	1	207	0
15	1	69	M	170	148	0	78	0
16	1	53	M	220	188	1	152	1
17	1	79	M	208	247	0	78	0
18	1	63	M	179	147	1	82	1
19	1	76	F	153	78	0	69	0
20	1	64	M	153	78	1	94	0
21	1	69	F	274	84	1	84	1
22	1	76	M	214	114	1	73	0
23	1	71	F	177	183	1	81	1
24	1	73	F	212	185	1	156	0
25	1	78	F	258	106	1	101	1
26	1	81	F	272	146	0	92	0
27	1	82	M	237	98	0	87	1
28	1	58	F	340	256	0	87	1
29	1	46	M	289	142	1	97	1
30	1	73	F	212	294	0	76	1
31	1	47	M	148	112	1	78	0
32	1	79	F	191	64	1	95	1
33	1	75	F	256	97	1	77	0
34	1	71	F	284	212	1	75	1
35	1	55	M	162	104	1	202	0
36	1	78	M	185	66	0	95	1
37	1	56	F	182	114	1	81	0
38	1	55	M	267	292	0	88	1
39	1	65	M	200	115	1	173	1
40	1	79	M	182	169	1	74	1
41	1	56	F	186	75	1	76	1

42	1	83	F	240	302	1	155	0
43	1	85	F	274	120	1	156	0
44	1	78	M	160	221	1	83	1
45	1	64	F	266	234	1	238	0
46	1	69	M	205	312	1	93	1
47	1	71	M	168	170	1	92	1
48	1	63	M	197	187	1	92	0
49	1	71	F	240	250	1	80	0
50	1	70	M	192	270	0	87	1
51	1	73	F	200	52	1	72	1
52	1	46	F	185	90	0	98	1
53	1	88	M	236	105	1	91	0
54	1	86	M	185	125	0	85	1
55	1	73	M	265	191	1	76	0
56	1	64	M	166	67	1	82	1
57	1	72	M	202	254	0	105	0
58	1	57	F	240	112	1	109	1
59	1	67	F	135	181	1	66	0
60	1	67	F	222	421	0	102	1
61	1	66	F	306	105	1	93	1
62	1	49	M	243	110	0	86	0
63	1	81	M	224	102	0	88	0
64	1	63	F	224	110	1	92	0
65	1	81	F	198	70	1	91	1
66	1	73	M	174	182	0	105	1
67	1	62	F	218	138	1	99	0
68	1	77	F	220	43	0	90	0
69	0	73	M	174	182	0	105	1
70	0	62	F	218	138	1	99	0
71	0	77	F	220	43	0	90	0
72	0	68	F	150	45	0	110	0
73	0	41	M	160	46	0	70	0
74	0	51	F	185	130	0	95	1
75	0	68	F	170	125	0	85	0
76	0	59	F	199	126	0	95	1
77	0	67	M	190	130	0	72	0
78	0	37	M	177	68	0	105	0
79	0	58	F	145	19	1	110	0
80	0	66	F	122	70	0	85	0
81	0	79	M	165	149	0	93	1
82	0	28	F	187	135	0	110	1
83	0	1	M	174	126	0	163	0
84	0	43	F	168	11	1	130	0
85	0	62	F	159	105	0	162	1

86	0	55	F	16	106	1	111	0
87	0	46	F	185	117	0	75	0
88	0	52	F	192	123	0	79	1
89	0	55	M	147	124	0	92	0
90	0	54	F	200	130	0	93	1
91	0	52	F	172	90	0	65	1
92	0	49	F	165	91	0	82	0
93	0	36	F	175	56	0	74	0
94	0	69	M	182	45	0	74	0
95	0	60	M	160	85	0	82	1
96	0	30	F	162	150	0	96	1
97	0	56	F	196	132	0	106	1
98	0	66	F	194	116	0	108	0
99	0	72	M	168	118	0	116	1
100	0	17	M	163	82	0	92	1
101	0	59	F	152	65	0	66	0
102	0	51	F	199	92	0	86	0
103	0	10	F	146	75	0	76	0
104	0	29	F	156	65	0	88	0
105	0	69	F	152	45	0	92	1
106	0	1	M	195	80	0	85	1
107	0	62	M	168	90	1	72	1
108	0	39	M	165	63	0	75	1
109	0	48	F	197	52	0	92	1
110	0	45	F	165	45	0	85	0
111	0	56	M	164	64	0	72	0
112	0	48	F	189	65	0	94	1
113	0	58	M	187	149	0	98	1
114	0	27	F	174	165	0	95	0
115	0	29	F	163	132	0	95	1
116	0	66	F	192	126	0	75	0
117	0	69	F	165	137	0	85	1
118	0	52	F	162	145	0	73	0
119	0	74	M	195	117	0	75	1
120	0	65	F	185	115	0	95	0
121	0	56	F	200	123	0	85	0
122	0	42	F	198	106	0	74	0
123	0	46	F	165	156	0	66	0
124	0	44	F	165	98	1	75	1
125	0	66	F	168	78	1	85	0
126	0	49	M	165	56	1	95	0
127	0	79	F	198	98	0	110	0
128	0	35	F	187	55	0	105	1
129	0	61	M	166	68	0	86	1

130	0	66	F	155	78	1	82	0
131	0	58	M	143	65	0	92	0
132	0	44	F	165	69	1	72	0
133	0	66	M	156	58	0	96	0
134	0	72	F	189	162	0	81	0
135	0	39	F	199	156	0	85	1
136	0	58	M	156	189	1	86	1
137	0	52	F	154	147	1	77	0
138	0	29	F	165	162	1	105	1
139	0	63	M	167	156	1	109	0
140	0	74	M	165	168	1	102	1
141	0	72	F	198	46	0	99	1
142	0	70	M	199	89	0	88	1
143	0	43	F	158	92	1	65	1
144	0	84	M	140	102	0	98	0
145	0	84	M	172	106	0	89	1
146	0	34	F	165	111	1	87	0
147	0	23	F	195	125	1	95	1
148	0	61	M	195	125	1	96	0
149	0	55	F	144	150	0	87	0