Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ciencias Matemáticas
Unidad de Posgrado

Factores relacionados a las enfermedades cardiovasculares en personas adultas, atendidas en el Hospital PNP. Augusto B. Leguía - 2017

TESIS

Para optar el Grado Académico de Magíster en Bioestadística

AUTOR

Fausto Francisco MATOS URIBE

ASESOR

Mg. Emma Norma CAMBILLO MOYANO

Lima, Perú

2023



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Matos, F. (2023). *Factores relacionados a las enfermedades cardiovasculares en personas adultas, atendidas en el Hospital PNP. Augusto B. Leguía - 2017*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas/Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Fausto Francisco Matos Uribe
Tipo de documento de identidad	DNI
Número de documento de identidad	08758082
URL de ORCID	https://orcid.org/0000-0002-0683-6656
Datos de asesor	
Nombres y apellidos	Emma Norma Cambillo Moyano
Tipo de documento de identidad	DNI
Número de documento de identidad	15377390
URL de ORCID	https://orcid.org/0000-0003-3173-9425
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	José Manuel Huamán Gutiérrez
Tipo de documento	DNI
Número de documento de identidad	71089047
Miembro del jurado 1	
Nombres y apellidos	Carlos Alberto Jaimes Velásquez
Tipo de documento	DNI
Número de documento de identidad	42762905
Miembro del jurado 2	
Nombres y apellidos	Ángel Deciderio Mendoza López
Tipo de documento	DNI
Número de documento de identidad	17824554
Miembro del jurado 3	
Nombres y apellidos	NO HUBO
Tipo de documento	NO HUBO

Número de documento de identidad	NO HUBO
Datos de investigación	
Línea de investigación	No aplica.
Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos y sus coordenadas geográficas Latitud: -12.06027 Longitud: -77.08146
Año o rango de años en que se realizó la investigación	ENERO 2018-ABRIL2023
URL de disciplinas OCDE	**Matemáticas puras** https://purl.org/pe-repo/ocde/ford#1.01.01 **Matemáticas aplicadas** https://purl.org/pe-repo/ocde/ford#1.01.02 **Estadísticas, Probabilidad** https://purl.org/pe-repo/ocde/ford#1.01.03

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE CIENCIAS MATEMÁTICAS
VICEDECANATO DE INVESTIGACIÓN Y POSGRADO
UNIDAD DE POSGRADO

ACTA DE SUSTENTACIÓN VIRTUAL DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN BIOESTADÍSTICA.

Siendo las, 10:00 horas del día Martes veinte de junio de dos mil veintitrés, en la sala virtual meet.google.com/rcz-tswh-iuw, el Jurado de Tesis conformado por los siguientes docentes:

PRESIDENTE	:	Dr. José Manuel Huamán Gutiérrez.
MIEMBRO	:	Mg. Carlos Alberto Jaimes Velásquez.
MIEMBRO EXTERNO	:	Dr. Ángel Deciderio Mendoza López.
ASESORA	:	Mg. Emma Norma Cambillo Moyano

se reunieron para la sustentación de la tesis titulada: **«FACTORES RELACIONADOS A LAS ENFERMEDADES CARDIOVASCULARES EN PERSONAS ADULTAS, ATENDIDAS EN EL HOSPITAL PNP. AUGUSTO B. LEGUIA - 2017»** presentada por el Señor Bachiller **Fausto Francisco Matos Uribe**, para optar el Grado Académico de Magíster en Bioestadística.

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, según tabla adjunta, habiendo obtenido el Señor Bachiller **Fausto Francisco Matos Uribe**, aprobado con el calificativo de **15 (BUENO).**

Habiendo sido aprobada la sustentación de la Tesis, el Jurado Evaluador recomienda para que el Consejo de Facultad apruebe el otorgamiento del Grado Académico de **Magíster en Bioestadística** al Bachiller **Fausto Francisco Matos Uribe**.

Siendo las 10:50 horas, se levantó la sesión, firmando para constancia la presente Acta:

Dr. José Manuel Huamán Gutiérrez
PRESIDENTE

Mg. Carlos Alberto Jaimes Velásquez
MIEMBRO

Dr. Ángel Deciderio Mendoza López
MIEMBRO EXTERNO

Mg. Emma Norma Cambillo Moyano
ASESORA

Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

FACULTAD DE CIENCIAS MATEMÁTICAS

VICEDECANATO DE INVESTIGACIÓN Y POSGRADO

UNIDAD DE POSGRADO

Informe de Evaluación de Originalidad

Vicedecano de Investigación y Posgrado: Dr. Eugenio Cabanillas Lapa

Operador: Mag. José Luis Acuña Guillermo

Tesis: "FACTORES RELACIONADOS A LAS ENFERMEDADES CARDIOVASCULARES EN PERSONAS ADULTAS, ATENDIDAS EN EL HOSPITAL PNP. AUGUSTO B. LEGUÍA - 2017"

Tesista: FAUSTO FRANCISCO MATOS URIBE

Recepción de documento: 06/03/2023

Aplicación del software: 06/03/2023

Software utilizado: Turnitin

Configuración del software:

- Excluir material bibliográfico
- Excluir material citado
- Excluir fuentes pequeñas N° palabras: 40
- Sin depósito

Resultado del análisis con el software: diez por ciento (10%)

Fuentes de originales de las similitudes encontradas:

- Ver abajo

Documento cumple con criterios de originalidad.

- **El documento cumple los criterios de originalidad. <u>Sin observaciones</u>.**

Fecha del informe: 07/03/2023

UNMSM Firmado digitalmente por CABANILLAS LAPA Eugenio FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 07.03.2023 22:19:12 -05:00

Dr. Eugenio Cabanillas Lapa

VICEDECANO DE INVESTIGACIÓN Y POSGRADO

Dedicatoria:

A una persona muy especial que llevo en mi corazón

A mi esposa Marissa Elvira Ríos Moscoso y mi hijo Francisco Fernando Matos Ríos, con amor a ambos.

A mis padres: Francisco Matos Salinas, Faustina Esther Uribe Huapaya; hermano y hermanas, por el apoyo incondicional en la realización del presente trabajo de investigación.

Agradecimiento:

A la Profesora Mg. Emma Norma Cambillo Moyano, por su apoyo como asesora en la realización de la presente tesis.

A todos mis profesores de la maestría en bioestadística por sus enseñanzas (2016-2017).

Pág.

ÍNDICE

Tablas v
Figuras vii
RESUMEN ix
ABSTRACT x
CAPITULO 1: INTRODUCCION 1
1.1 Situación Problemática. 1
1.2 Formulación del Problema. 2
1.3 Justificación de la investigación . 3
1.4 Objetivos de la investigación. 5

CAPITULO 2: MARCO TEÓRICO 6
2.1 Antecedentes de la investigación 6
2.2 Bases teóricas de la enfermedad cardiovascular 12
2.2.1 Factores de riesgo cardiovascular 13
2.2.1.1 Factores de riesgo modificables 13
2.2.1.2 Factores de riesgo no modificables 13
2.3 Método para determinar factores, modelo de regresion logístico binario. 14
2.3.1 Modelo lineal generalizado (MLG). 14
2.3.2 Formulación del modelo lineal generalizado. 14
2.3.3 Formulación del modelo de regresión logístico binario múltiple. 17
2.3.3.1 Suposiciones del modelo logístico binario. 23
2.3.3.2 Estimación de los parámetros β por máxima verosimilitud. 24
2.3.3.3 Interpretación $\hat{\beta}$ y $O\hat{R}$ (Odds ratio) 32
2.3.4 Evaluación del modelo estimado. 35
2.3.4.1 Prueba de bondad de ajuste Estadística G. 35

2.3.4.2 Prueba de significación de los coeficientes, estadística de Wald β_i 38
2.3.4.3 Estimación de β_i por intervalos de confianza. 39
2.3.4.4 Estimación de OR por intervalo de confianza. 39
2.3.4.5 Coeficiente de determinación. 40
2.3.4.6 Prueba de bondad de ajuste de Hosmer-Lemeshow. 41
2.3.4.7 Capacidad Predictiva del modelo. 42
2.3.4.8 Curva ROC. 45
2.3.4.9 Criterio de información de Akaike (AIC) 46
2.3.4.10 Interacción entre las variables 47
2.3.5 Adecuación del modelo de regresión logístico binario. 47
2.3.5.1 Análisis de residuo e influencia. 47
2.3.5.2 Multicolinealidad. 50
2.4 Glosario de términos. 52
2.5 Hipótesis del estudio. 54

CAPITULO 3: METODOLOGIA 55

3.1 Identificación de variables. 55
3.2 Tipo y diseño de investigación. 55
3.3 Población. 56
3.4 Unidad. 56
3.5 Muestra seleccionada. 56
3.6 Diseño muestral 57
3.7 Técnica de recolección de datos. 57
3.8 Análisis e interpretación de la información. 58
3.9 Definición de variables. 59

CAPITULO 4: RESULTADOS Y DISCUSION 61

4.1 Resultados 61

4.1.1 Análisis descriptivo univariado. 61

4.1.2 Análisis bivariado. 66

- Asociación entre la enfermedad cardiovascular y las variables predictoras 66
- Relación lineal de las variables predictoras 78
- Identificación de variables confusoras 79

4.1.3 Modelos de regresion logísticos binarios múltiple. 80

4.1.3.1 Estimación del modelo logístico: Estadística G, Wald, Hosmer-Lemeshow y coeficientes de determinación. 81

4.1.3.2 Interpretación de los coeficientes. 82

4.1.3.3 Evaluación de supuesto del modelo. 85

- Matriz de correlación, multicolinealidad. 86
- Análisis de los residuos e influencia. 86
- Interacción de predictores 88

4.1.3.4 Capacidad predictiva del modelo. 88

4.1.3.5 Prueba ROC, modelo logístico estimado. 89

4.1.3.6 Predicciones. 93

4.2 Discusión. 94

Conclusiones. 96

Recomendaciones. 97

REFERENCIA BIBLIOGRAFIA 99

ANEXOS 105

1. Matriz de consistencia. 106
2. Operacionalización de variables. 108
3. Ficha de Recojo de datos. 112

4. Variables confusoras. 113

5. Salidas de los modelos logísticos binario por factores 116

6. Calculo del factor de inflación de varianzas (FIV) 120

Índice de tablas: Pág.

Tabla 1. Edad (años) relacionado con la hipertensión NCHS (EE.UU) n=15,149. 21
Tabla 2. Grupos etarios y proporción de pacientes con hipertensión arterial NCHS (EE.UU) n=15,149. 22
Tabla 3. Variable explicativa con tres categorías, necesita dos variables dicotómicas. 34
Tabla 4. De contingencia para calcular la estadística de Hosmer-Lemeshow. 42
Tabla 5. Clasificación de valores observados y probabilidad estimada por el modelo, con punto de corte c=0.5. 43
Tabla 6. Definición de variables en el estudio 60
Tabla 7. Factor demográfico de pacientes atendidos por consultas externas PNP augusto B. Leguía - Rímac, Octubre-Diciembre 2017(n=1284). 63
Tabla 8. Factor estilo de vida de pacientes atendidos por consultas externas PNP augusto B. Leguía - Rímac, Octubre-Diciembre 2017(n=1284). 64
Tabla 9. Factor clínico de pacientes atendidos por consultas externas PNP augusto B. Leguía - Rímac, Octubre a Diciembre 2017(n=1284). 64
Tabla 10. Estadísticas de variables cuantitativas relacionadas a las enfermedades cardiovasculares, hospital PNP Augusto B. Leguía, Octubre – Diciembre 2017 (n=1284). 66
Tabla 11. Factor demográficos y las enfermedades cardiovasculares, en pacientes atendidos por consultas externas, hospital PNP Augusto B. Leguía -Rímac, Octubre a Diciembre 2017 (n=1284). 69
Tabla 12. Factor estilo de vida y las enfermedades cardiovasculares, en pacientes atendidos por consultas externas, hospital PNP Augusto B. Leguía -Rímac, Octubre a Diciembre 2017 (n=1284). 73
Tabla 13. Factor clínico y las enfermedades cardiovasculares, en pacientes atendidos por consultas externas, hospital PNP Augusto B. Leguía -Rímac, Octubre a Diciembre 2017 (n=1284). 76
Tabla 14. Correlaciones entre variables predictoras cuantitativas. 80

Tabla 15. Variables relacionadas a las enfermedades cardiovasculares en pacientes adultos atendidos en el hospital PNP Augusto B. Leguía – Rímac, utilizando el modelo logístico (n=1284). 85
Tabla 16. Matriz de correlaciones entre las estimaciones de los parámetros del modelo estimado. 86
Tabla 17. Factor de inflación de varianzas de los predictores del modelo estimado. 86
Tabla 18. Clasificación de pacientes, mediante el modelo logístico estimado, con punto de corte c=0.5 (n=1284). 90
Tabla 19. Clasificación de pacientes mediante el modelo logístico estimado con punto de corte c=0.426 (n=1284). 91
Tabla 20. Clasificación de pacientes, a través de la validación cruzada con punto de corte c=0.5 (n=1284). 92
Tabla 21. Estadística resumen del modelo estimado de regresion logística binaria múltiple. 93
Tabla 22. Variables que se relacionan con las enfermedades cardiovasculares, a través del modelo logístico binario. 95

Índice de figuras: Pág.

Figura 1. Edad (años) y la hipertensión arterial (presencia o ausencia). 21
Figura 2. Grupos etarios y proporción de pacientes con hipertensión arterial NCHS (EE.UU) n=15,149. 23
Figura 3. Tendencia lineal entre la edad (años) y Logit $(\hat{\pi}_i)$. 23
Figura 4. Proporción de falsos positivos y verdaderos positivos (ROC). 45
Figura 5. Género de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 70
Figura 6. Razas de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 70
Figura 7. Edad de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 71
Figura 8. Grupos etarios de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). Grupos menor de 60 y mayor o igual a los 60 años. 71
Figura 9. Índice de masa corporal (IMC) de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 72
Figura 10. Alcoholismo y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 74
Figura 11. Actividad física y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 74
Figura 12. Nivel de colesterol y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 77
Figura 13. Nivel de triglicéridos y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 77
Figura 14. Diabetes y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 78
Figura 15. Presión arterial sistólica y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 78
Figura 16. Presión arterial diastólica y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284). 79
Figura 17. Residuos estandarizados y observaciones. 87

Figura 18. Distancia de Cook y observaciones. 88
Figura 19. Probabilidad estimada y distancia de Cook (n=1284) 88
Figura 20. Probabilidad estimada y valores de influencia (n=1284). 89
Figura 21. Curva ROC del modelo, tasa de falsos positivos (1-especificidad) y tasa de verdaderos positivos (sensibilidad). 90

RESUMEN

Título del trabajo de investigación: Factores relacionados a las enfermedades cardiovasculares en personas adultas, atendidas en el Hospital PNP. Augusto B. Leguía - 2017

Objetivo general: El presente trabajo de investigación es determinar factores relacionados a las enfermedades cardiovasculares, en personas adultas, atendidas en el Hospital PNP Augusto B. Leguía Rímac. Lima-Perú, factores que influyen en la probabilidad de ocurrencia de tener o no la enfermedad. **Métodos**: El estudio fue observacional y retrospectivo y analítico. El diseño muestral, es aleatorio simple, se tomó una muestra al azar de 1284 pacientes e historia clínica entre los meses de octubre a diciembre 2017, se usó la regresión logística binaria múltiple. **Resultados**: De un total de 12 predictores (clasificados factores: demográficos, estilo de vida y clínicos), resultaron significativo al 5% para predecir la enfermedad cardiovascular los predictores: edad (factor demográfico), alcoholismo (factor estilo de vida), nivel de colesterol, nivel de triglicéridos, presión arterial sistólica y presión arterial diastólica (factor clínico); la prueba de bondad de ajuste de Hosmer-Lemeshow resultó no significativo ($p>0.05$), indicando que el modelo logístico se ajusta bien a los datos de la muestra; los coeficientes de determinación de Cox-Snells (47.5%), Mc Fadden (48.8%), Nagelkerke (64.8%) todos aceptables; la precisión o capacidad predictiva del modelo estimado fue de 85.6% (detectar con la enfermedad o sin la enfermedad cardiovascular) y el área bajo la curva ROC fue de 0.920, tiene un poder discriminante excelente. **Conclusión**: La regresión logística binaria demostró ser un buen método para determinar factores que se relacionan con la enfermedad cardiovascular y además realizar estimaciones.

Palabras claves: Enfermedades cardiovasculares, factores demográficos, estilo de vida, clínicos y regresion logística binaria.

ABSTRACT

Title of the research work: Factors related to cardiovascular diseases in adults, treated at the PNP Hospital. Augusto B. Leguía – 2017

General objective: This research work is to determine factors related to cardiovascular diseases in adults treated at the PNP Augusto B. Leguía Rímac Hospital. Lima-Peru, factors that influence the probability of occurrence of having or not the disease. Methods: The study was observational and retrospective and analytical. The sample design is simple random, a random sample of 1284 patients and clinical history was taken between the months of October to December 2017, multiple binary logistic regression was used. Results: Of a total of 12 predictors (classified factors: demographic, lifestyle and clinical), the predictors were significant at 5% to predict cardiovascular disease: age (demographic factor), alcoholism (lifestyle factor), level of cholesterol, triglyceride level, systolic blood pressure and diastolic blood pressure (clinical factor); the Hosmer-Lemeshow goodness-of-fit test was not significant ($p>0.05$), indicating that the logistic model fits the sample data well; Cox-Snells (47.5%), Mc Fadden (48.8%), Nagelkerke (64.8%) coefficients of determination were all acceptable; the precision or predictive capacity of the estimated model was 85.6% (detect with or without cardiovascular disease) and the area under the ROC curve was 0.920, it has excellent discriminant power. Conclusion: Binary logistic regression proved to be a good method to determine factors that are related to cardiovascular disease and also make estimates.

Keywords: Cardiovascular diseases, demographic factors, lifestyle, clinical and binary logistic regression.

CAPITULO 1: INTRODUCCION

1.1. Situación Problemática

La Organización Mundial de la Salud (OMS, 2020) refiere que:

> Las enfermedades cardiovasculares (ECV) son la principal causa de muerte en todo el mundo. Desde el año 2000, las muertes se han incrementado en lo que respecta a la cardiopatía isquémica, incrementándose de más de 2 millones de defunciones en 2000 a 8.9 millones en 2019, siendo esta enfermedad, responsable del 16% del total de muertes en el mundo, y en segundo lugar de causas de muertes en el mundo es accidente cerebrovascular que representa el 11% del total de muertes. (p.1)

La Organización Panamericana de la Salud (OPS), “en el Día Mundial del Corazón destaca la importancia de la prevención y el tratamiento de las enfermedades cardiovasculares (ECV), la principal causa de muerte en las Américas, que cobra 2 millones de vidas cada año”. (p.1).

> A nivel mundial, cada año mueren más personas por enfermedades cardiovasculares que por cualquier otra causa, principalmente enfermedad cardíaca isquémica y accidente cerebrovascular. Más de tres cuartas partes de estas muertes se producen en países de bajos y medianos ingresos, donde los casos siguen aumentando (OPS, 2021).
>
> En América Latina y el Caribe, la hipertensión es una condición muy común que conduce tanto a la enfermedad cardíaca como al accidente cerebrovascular, y datos de la red NCD Risk Factor Collaboration (NCD-RisC) muestran que 28% de las mujeres y el 43% de los varones desconocen su condición de hipertensos (OPS, 2021).
>
> El riesgo de cardiopatía y de accidente cerebrovascular se ve incrementado por las dietas poco saludables, sobre todo las que tienen un alto contenido en sal, grasas y azúcares refinadas, y por los bajos niveles de actividad física. El consumo de tabaco es también uno de los principales factores de riesgo, ya que contribuye a cerca del 10% de todas las muertes por ECV (OPS, 2021).
>
> Para quienes viven con ECV, la pandemia de COVID-19 ha supuesto un motivo de preocupación urgente. Las personas con problemas de salud subyacentes, como las enfermedades cardíacas y los accidentes cerebrovasculares, corren un mayor riesgo de desarrollar formas graves de enfermedad y muerte relacionadas con la COVID-19. Muchas de estas personas han evitado incluso

> buscar atención médica de rutina y de emergencia, que podría salvarles la vida, por temor a contagiarse (OPS, 2021).

Según el INEI (2012),

> Las enfermedades cardiovasculares continúan siendo la principal causa de muerte en el mundo; así como, de morbilidad y pérdida de calidad de vida relacionada con la salud. En el caso del Perú, las enfermedades isquémicas del corazón y las enfermedades cerebrovasculares se constituyen como segunda y tercera causa de mortalidad en el adulto mayor, respectivamente; además, presentan importantes diferencias por edad, sexo, nivel de educación, quintiles de riqueza entre otras. Entre los principales factores de riesgo que se presentan en este estudio destacan el colesterol alto, triglicérido alto, hipertensión arterial, enfermedades del corazón y diabetes. La cuantificación de estos factores constituye un aspecto esencial para comprender la dimensión de este problema a escala poblacional y para una planificación de los recursos sanitarios.

El Hospital PNP Augusto B. Leguía, se encuentra ubicado en el distrito del Rímac, Lima-Perú, los pacientes atendidos por consultas externas son miembros de la Policía Nacional del Perú, en situación de actividad, retiro y familiares directos, la procedencia de los pacientes son de los distritos de Lima, Callao y provincias, el promedio de atenciones de pacientes en todas las áreas de consultas es de 400 pacientes diarios, las atenciones se realiza solo en horas de la mañana, siendo la atención del servicio de emergencia las 24 horas del día. Durante el año 2019, se registró 110,438 atenciones, de las cuales 9,442 (8.55%) corresponde al área de cardiología.

1.2. Formulación del problema.

¿Cuáles son los factores relacionados a las enfermedades cardiovasculares en personas adultas, atendidos en el Hospital PNP Augusto B. Leguía, Lima-Perú 2017?

1.3. Justificación de la investigación

1.3.1 Justificación teórica

La Organización Mundial de la Salud (OMS, 2020) refiere que:

> Las enfermedades cardiovasculares son la principal causa de defunción en el mundo y, según estimaciones, se cobran 17,9 millones de vidas cada año. Estas enfermedades constituyen un grupo de trastornos del corazón y los vasos sanguíneos que incluyen cardiopatías coronarias, enfermedades cerebrovasculares y cardiopatías reumáticas. Más de cuatro de cada cinco defunciones por enfermedades cardiovasculares se deben a cardiopatías coronarias y accidentes cerebrovasculares, y una tercera parte de esas defunciones ocurren prematuramente en personas menores de 70 años. (p.1)

Padecer alguna enfermedad cardiovascular, limita a una gran cantidad de personas a desenvolverse de una manera normal en sus actividades diarias, llevando muchas veces a la muerte, sino logras controlar de una forma conveniente los factores de riesgos; lo que busca este estudio es determinar los factores de riesgos de mayor frecuencia que se relacionan a esta enfermedad, con la finalidad de reducir de los que no la tienen y la muerte a los que la tienen. La investigación de esta enfermedad, podrá llamar la atención para planear programas de prevención desde una perspectiva integral, vinculando a todos los entes responsables de la salud de la población adulta.

El Hospital PNP Augusto B Leguía es un lugar donde se puede tener acceso a las historias clínicas de pacientes por consultas externas en la especialidad de cardiología. Existe un área de prevención donde se realizan charlas de sensibilización a diversas enfermedades, la cual también podría dar una capacitación con los resultados obtenidos de la investigación (Misión y Funciones del Hospital PNP. Augusto B. Leguía, 2022).

Abordar de manera correcta la prevención de la enfermedad cardiovascular, la cual requiere una identificación de los factores de riesgos.

1.3.2 Marco metodológico

Para determinar los factores que se relacionan con la enfermedad cardiovascular en personas adultas de 20 años a más en el Hospital de la PNP Augusto B. Leguía, Lima-Perú, se consultó literaturas científicas sobre las variables que se relacionan con la aparición de la enfermedad cardiovascular, se registró 12 variables predictoras clasificadas en factores y como la variable dependiente en estudio, la enfermedad cardiovascular es cualitativa dicotómica, el mejor modelo a priori que se ajusta a los datos de la muestra es el modelo logístico binario múltiple, modelo que permite determinar factores que se relacionan con la enfermedad, a través de la evaluación de la bondad de ajuste y la evaluación de la adecuación del modelo estimado.

1.3.3 Justificación práctica

Determinar los factores (variables cualitativas y/o cuantitativas) que se relacionan o influyen en la aparición de las enfermedades cardiovasculares

1.4. Objetivos de la investigación

1.4.1 Objetivo General

Determinar factores relacionados a las enfermedades cardiovasculares, en personas adultas, atendidas en el Hospital PNP Augusto B. Leguía. Lima-Perú, mediante la regresión logística binaria múltiple.

1.4.2 Objetivos específicos

a. Identificar factores relacionados a las enfermedades cardiovasculares más frecuentes, en personas adultas, atendidas en el Hospital PNP Augusto B. Leguía. Lima-Perú.

b. Evaluar factores relacionados a las enfermedades cardiovasculares, en personas adultas, atendidos en el Hospital PNP Augusto B. Leguía-Rímac. Lima-Perú.

CAPITULO 2: MARCO TEORICO

2.1 . Antecedentes de la investigación

Luis (2021), realizó un trabajo de investigación que tiene por objetivos, determinar el riesgo cardiovascular en adultos que acuden a consulta médica a un policlínico privado en el distrito de Carabayllo-Lima-Perú; el método que utilizo fue descriptivo-transversal, la población en estudio estuvo constituida por 75 adultos varones que tienen la enfermedad cardiovascular; los resultados del estudio, ocupando el primer puesto el riesgo moderado (42,7%), seguido del riesgo bajo (22,7%), alto (18,7%), muy alto (8%) y crítico (8%), la edad mínima fue 40 años y la máxima 79 años, la edad promedio 53.6 años; en cuanto al colesterol total, se impuso el nivel elevado (65,3%), normal (34,7%); en cuanto a presión arterial, la sistólica elevada (70,7%), la sistólica normal (29.3%) ; consumo de tabaco el 52% respondió que “si” y el 48% dijo que “no”, diabetes mellitus (85,3%); la conclusión fue: el riesgo moderado fue el más frecuente, en segundo lugar riesgo bajo, alto, muy alto y crítico.

Álvarez et al. (2020), el objetivo fue evaluar los factores de riesgos cardiovascular (FRCV) modificables en trabajadores del Hospital Universitario La Paz Madrid-España con relación a la categoría profesional, edad, sexo, así como estimar el riesgo cardiovascular según el modelo SCORE para los trabajadores con edades mayor o igual a los 40 años y con el riesgo relativo en menores de 40 años; método utilizado fue observacional descriptivo retrospectivo en una muestra al azar de 687 trabajadores del Hospital Universitario, se eligieron variables clínicas, antropométricas y sociodemográficas; los resultados encontrados fue: mujeres (70,8%), y varones (29,2%), la prevalencia de tabaquismo (21%), no fumadores (60%) y exfumadores (19%), dislipemia (58,8%), hipertensión arterial (24,3%), inactividad física (35,7%), obesidad/sobrepeso (36,5%). DM (4,5%), alcohol (37,8%), en el cálculo del riesgo cardiovascular se encontró un SCORE

alto/ muy alto: 15,2% y RR=1 (83%), RR=2 (16%) y RR=3 (1%). Conclusiones: se encontró mayor prevalencia de factores de riesgo modificables en personal de gestión y servicios y en técnicos sanitarios que en diplomados y licenciados sanitarios, y conforme al aumento de edad y en el sexo varón.

Berdú et al. (2020), realizaron una investigación que tiene por objetivo evaluar la influencia de factores riesgo en el desarrollo de enfermedades cardiovasculares en personas con edades menor o igual a los 50 años; el método utilizado es un estudio analítico retrospectivo de casos y control en pacientes con el diagnóstico de enfermedad cardiovascular (ECV) que ingresaron al Hospital Provincial Universitario “Carlos Manuel de Céspedes” del municipio Bayamo, Granma; los resultados de la investigación fueron: el género varón y alcoholismo resultaron no significativo para el riesgo de ECV; todo lo contrario para el hábito de fumar aumento el riesgo a más de cinco veces, la comorbilidad con mayor fuerza de asociación lo mostró la HTA mientras que la proteína C reactiva fue el biomarcador de mayor asociación con el riesgo de desarrollar una ECV; el análisis multivariado de la regresion logística binaria, mostró que los factores de mayor influencia independiente para desarrollar una ECV fue la proteína C reactiva al elevar el riesgo, seguida en orden de importancia la HTA ambos de forma muy significativa, la conclusión del trabajo de investigación, se obtuvo un modelo logístico para estimar el riesgo de desarrollar una enfermedad cardiovascular útil como instrumento de vigilancia clínica y epidemiológica, al identificar a sujetos con mayor probabilidad de enfermar.

Hernández-Martínez et al. (2019), el objetivo de la investigación fue determinar los factores asociados a la enfermedad cardiovascular con el ausentismo laboral de los trabajadores de una entidad pública en la ciudad de Bogotá; el método y material utilizado en ese orden fue un estudio es descriptivo de corte transversal y una muestra al azar de 214 trabajadores y las variables consideradas fueron

sociodemográficas, clínicas-cardiovasculares y laborales (Edad, tabaquismo, alcoholismo, sedentarismo, IMC, PAS, PAD, genero, estado civil, ingreso y tiempo de servicios). Se efectuó un análisis de las variables predictoras y la variable de estudio, y se planteó un modelo de regresión logística, teniendo como variable dependiente el ausentismo laboral; los resultados, fue la prevalencias del tabaquismo (13.1%), consumo de alcohol (35.1%), sedentarismo(39.7%); trabajadores con sobrepeso u obesidad (54.2%) y colesterol total elevado (, 43.9%); las conclusiones, el modelo de regresión logística encontró que las variables estado civil casado, obesidad, tratamiento para la hipertensión, estadio de presión arterial diastólica y tiempo de servicio influyen en el ausentismo laboral.

Chambergo-Michilot (2020), realizó un trabajo de investigación el cual tiene por objetivo conocer la Prevalencia de enfermedades cardiovasculares en el Hospital Nacional Dos de Mayo Lima-Perú; material y método utilizado fue: observacional descriptivo y la muestra fue de 446 pacientes que ingresaron al servicio de cardiología; los resultados fue las enfermedades cardíacas más prevalentes fue la falla cardíaca (51.57%), cardiopatías congénitas (23.99%), fibrilación auricular (17.49%) e infarto miocárdico agudo (14.57%); las conclusiones del estudio, reporta las enfermedades y comorbilidades cardiovasculares en un servicio de cardiología.

Zuni-Chávez et al. (2019) realizaron un trabajo de investigación, el objetivo es, determinar los factores asociados a riesgo cardiovascular en pacientes del Departamento de Medicina del Hospital Sergio E. Bernales (HSEB), Lima-Perú; el método, es transversal, descriptivo y observacional; material la muestra de 83 pacientes elegidos probabilísticamente de una población de 105 hospitalizados; los resultados: el 53% de los pacientes fueron mujeres y la edad promedio fue 54,5 años, el sobrepeso 30,1%, la obesidad 13,3%, el Índice cintura, cadera muy alto en mujeres 54,2%, antecedente familiar de infarto agudo al miocardio 38,5% e HTA 34,6%, diabetes tipo II 65,5%,

colesterol Alto 34,4%, sedentarismo 50%, consumo carbohidratos 56.6%, lípidos 32.5%, comida rápida 44,6% y como grado de ansiedad menor un 54.2% y mayor el 32,5%; conclusión: mayor riesgo de evento coronario afecta más a las mujeres mayores de 50 años que en los varones, teniendo como factores de riesgo cardiovascular principales la dieta, el sedentarismo, el sobrepeso, las presiones normal- alta, la ansiedad menor, y comorbilidades asociadas como la diabetes y colesterol alto.

Barrera (2015), realizo un trabajo de investigación, que tuvo como objetivo determinar los factores de riesgo de la enfermedad cardiovascular en choferes de la empresa de transporte público Unidos Vitarte S. A; el estudio fue descriptivo y corte transversal, la muestra consistió de 50 choferes. Se aplicó un cuestionario de preguntas relacionado al riesgo de la enfermedad cardiovascular y además algunas medidas antropométricas a través de la entrevista. Los resultados en cuanto al riesgo de la enfermedad cardiovascular fueron: 54% de los encuestados pertenecen a la dimensión estilo de vida (mala alimentación, falta de ejercicios y otros), 28% a la Biología Humana familiares (antecedentes familiares con factor de riesgo, obesidad, mayor de 45 años y otros), 11% al sistema sanitario (ausencia de control médico) y el 7% al medio ambiental (horario de trabajo mayor a las 8 horas y ambiente laboral conflictivo). Las conclusiones a las que se llegó son que todos los choferes tienen factores de riesgos, siendo la más resaltante fue la variable estilos de vida.

Patiño-Villada et al. (2011) realizaron una investigación, el objetivo fue evaluar la prevalencia de los factores de riesgo cardiovascular en una población urbana entre 25 y 50 años del municipio de Santa Rosa de Osos, Antioquia 2009-Colombia; se realizó un estudio de prevalencia, la muestra utilizada fue de 357 persona sin enfermedad cardiovascular, la muestra fue obtenida usando el muestreo aleatorio bietápico, se elaboró un cuestionario de preguntas diseñada para evaluar la prevalencia de los factores de riesgo, además

se midieron en sangre colesterol, lipoproteínas de baja y alta densidad, triglicéridos y glucosa, se tomaron valores de presión arterial, perímetro abdominal y índice de masa corporal; los resultados son: baja actividad física (56,6 %), obesidad central (52,7 %), dislipidemias (35,3 %), tabaquismo (19,0 %), síndrome metabólico (19,6 %) y obesidad por Índice de Masa Corporal (17,4 %). Los factores de riesgo con mayor riesgo atribuible poblacional fueron la hipertensión arterial (17,1 %), Índice colesterol total/lipoproteína de alta densidad (16,4 %) y baja actividad física (15,7 %); conclusión del estudio, se encontró una alta prevalencia de los factores de riesgo cardiovascular lo que puede llevar a un aumento de las enfermedades cardiovasculares a mediano y largo plazo.

Rivas (2016), realizo un trabajo de investigación sobre recomendaciones para la Atención Primaria de Salud en Cuba, para todos aquellos pacientes que padecen de alguna enfermedad del corazón, a través de un programa de ejercicios físicos que comprende tres fases (hospitalaria, convalecencia y mantenimiento), para diseñar un régimen de ejercicios hay que tener en tipo de ejercicios, intensidad, duración y frecuencia. El principal objetivo de un programa de este tipo es aumentar la calidad y cantidad de vida del paciente con cardiopatía, y para comprobar su efectividad se pueden realizar evaluaciones periódicas a cualquier nivel de atención. Conclusión, principales beneficios consiste en que la mayoría de los pacientes ya no se sentirá limitado por la enfermedad y volver a sus actividades laborales o sociales, además una disminución en el consumo de medicamentos.

Sacanella et al. (2019). Realizaron un estudio de investigación relacionado a la prevención de la enfermedad cardiovascular y bebidas alcohólicas fermentadas, llegando a la conclusión de que existe mucha evidencia que sugiere una relación en J entre el consumo moderado de alcohol y la salud cardiovascular, se ha referido una reducción de alrededor del 20% en la mortalidad y de hasta el 40% en la cardiopatía

isquémica en los bebedores moderados respecto a los abstemios absolutos.

Cedeño-Zambrano et al. (2016), realizaron un estudio de investigación relacionado con el consumo de alcohol y la enfermedad cardiovascular, el consumo de alcohol tiene un comportamiento ambivalente con el sistema cardiovascular con efectos tanto beneficiosos como nocivos; que el consumo excesivo de alcohol, puede ocasionar miocardiopatía alcohólica, infartos al corazón. Su consumo frecuente eleva los niveles de sodio y calcio en la célula y puede producir hipertensión arterial; cuando se reduce el consumo en personas con hipertensión se logra disminuir la enfermedad, la ingesta no moderada de alcohol, puede producir arritmias graves en personas sanas.

Cisneros & Carrazana (2013), objetivo del estudio es actualizar a los médicos de la atención primaria de salud sobre factores de riesgos de la cardiopatía isquémica; el método de trabajo consistió en la revisión de literatura científica relacionado al tema, llegando a la siguiente conclusiones: se encontró factores de riesgo no modificables, tales como, la herencia genética, edad, sexo y factores modificables: dislipidemia, tabaquismo, alcoholismo, Sedentarismo, obesidad, hipertensión arterial, diabetes mellitus y otros factores de riesgo, es de suma importancia conocer cómo influye cada factor de riesgo cardiovascular en la aparición de la Cardiopatía Isquémica para tomar medidas que coadyuve al control de la enfermedad.

Rojas et al. (2016), realizaron un estudio de investigación para determinar factores de riesgos que más se relacionan con la hipertensión arterial, utilizando el modelo de regresion logística binaria, los factores estudiados fueron: edad, índice de masa y cintura hipertrigliceridemia (CHT), la muestra para el estudio fue de 202 pacientes de 18 años a más, las mediciones tomadas fueron: edad, peso, talla, perímetro de la cintura o circunferencia abdominal, presión

arterial sistólica y diastólica, además se definió la obesidad abdominal 102 cm en varones y 88 cm en mujeres, glucosa, ácido úrico, colesterol y triglicéridos. Una primera regresion logística con factor, edad, resulto significativo, la segunda regresion con las variables edad y fenotipo CHT, ambos resultaron significativo al 5% y el tercer modelo logístico estimado, con las variable edad, CHT e índice de masa corporal, resultaron significativa todas para Wad; conclusión que los factores de riesgo antes mencionado pueden contribuir al diagnóstico y la atención primaria de la hipertensión arterial.

2.2. Bases teóricas de las enfermedades cardiovasculares

> La enfermedad cardiovascular es un concepto amplio para problemas con el corazón y los vasos sanguíneos, estos problemas a menudo se deben a la aterosclerosis, ocurre cuando la grasa y el colesterol se acumulan en las paredes de las arterias, esta acumulación grasa se llama placa, con el correr del tiempo estrecha los vasos sanguíneos y causa problemas en todo el cuerpo, si una arteria resulta obstruida, esto puede llevar a que se presente un ataque cardíaco o un accidente cerebrovascular (MedlinePlus, 2021, p.1).

"*Los tipos de enfermedad cardiovascular se encuentran: La cardiopatía coronaria, la insuficiencia cardiaca, las arritmias, la arteriopatía periférica, la hipertensión, el accidente cerebrovascular y la cardiopatía congénita*" (MedlinePlus, 2021, p.1).

2.2.1. Factores de riesgo cardiovascular.

> "*Los factores de riesgo cardiovascular son los que se asocian a una mayor probabilidad de sufrir una enfermedad cardiovascular: colesterol, diabetes, hipertensión, tabaquismo, herencia genética,*

estrés, obesidad y la frecuencia cardiaca" (Fundación española del corazón, 2022, p.1).

2.2.1.1 Factores de riesgo modificable.

Son aquellos que pueden ser corregidos o eliminados a través de cambios en el estilo de vida.

A. Hipertensión arterial
B. Hipercolesterolemia
C. Obesidad
D. Estrés
E. Consumo de alcohol
F. Consumo de tabaco
G. Actividad física
H. Diabetes mellitus

2.2.1.2 Factores de riesgo no modificables.

Son constitutivos de la persona, la que siempre tendrá ese factor de riesgo y no es posible revertirlo o eliminarlo.

A. Edad
B. Género
C. Historia familiar de enfermedad coronaria
D. Característica étnica

2.3 Métodos para determinar factores, modelo de regresión logístico binario.

2.3.1 Modelo lineal generalizado (MLG).

Los modelos lineales generalizados (MLG) aglutina a los modelos de regresión lineal y no lineal, la variable dependiente, solo necesita ser miembro de la familia exponencial, que comprende a casi

todas las distribuciones discretas y continuas. Por ejemplo en el modelo de regresión logístico binario es un modelo no lineal, la variable respuesta (dicotómica) es Bernoulli, solo tiene dos posibles resultados presencia y ausencia de una cualidad ($Y_i = 1$, o $Y_i = 0$) su varianza no es homogénea (Nelder & Wedderburn, 1972).

2.3.2 Formulación del modelo lineal generalizado.

El modelo lineal generalizado (MLG) es una extensión del modelo lineal general que contiene tres componentes:

Sea el modelo, expresado en su forma matricial

$$Y = X\beta + \varepsilon \qquad (1)$$

Siendo:

$$Y'_{1xn} = (y_1, y_2, ..., y_n)$$

$$\beta'_{1x(k+1)} = (\beta_0, \beta_1, \ . \ . \ , \beta_k)$$

$$X_{nx(k+1)} = \begin{bmatrix} 1 & x_{11}, & x_{12}, & . & . & .x_{1k} \\ 1 & x_{21}, & x_{22}, & . & . & .x_{2k} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ 1 & x_{n1}, & x_{n2}, & . & . & .x_{nk} \end{bmatrix}$$

$$\varepsilon'_{1xn} = (\varepsilon_1, \varepsilon_1, \ . \ . \ , \varepsilon_n)$$

Supuesto $E(\varepsilon_i) = 0$

Modelo (1) presenta 3 componentes:

a. Componente aleatoria Y : La variable aleatoria Y con observaciones independientes $(y_1, y_2, ..., y_n)$, cuya distribución de probabilidades es miembro de la familia exponencial de distribución de probabilidades.

La familia exponencial en su forma canónica:

$$f(y_i,\theta,\phi) = Exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\} \qquad i = 1,2,...,n \qquad (2)$$

En (2) θ , parámetro de localización; Y es una estadística suficiente, la función $b(\theta)$, depende sólo del parámetro θ y la función $c(y_i,\varphi)$ depende del parámetro ϕ y de la variable aleatoria Y

Por ser Y una variable aleatoria, tiene esperanza y varianza matemática condicionales a las variables explicativas.

La esperanza poblacional condicional

$$E(Y / X_1 = x_1, X_2 = x_2, X_3 = x_3,..., X_k = x_k) = \mu$$

La varianza poblacional condicional

$$V(Y / X_1 = x_1, X_2 = x_2, X_3 = x_3,..., X_k = x_k) = V$$

.

b. Componente sistemática: Formada por una combinación lineal entre las variables predictoras $X_1, X_2, X_3,..., X_k$ que se suponen fijas, medidas sin error y conocidas (cualitativas o cuantitativas) que influyen a Y, y un conjunto de parámetros desconocidos $(\beta_0, \beta_1, \beta_2,..., \beta_k)$, combinación lineal que determinan el valor de Y para $X_1 = x_{i1}, X_2 = x_{i2}, X_3 = x_{i3},..., X_k = x_{ik}$, simbolizada matemáticamente por η_i:

$$\eta_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_j x_{ij} + ... + \beta_k x_{ik} = X_i^{'}\beta = \sum_{j=0}^{k} \beta_j x_{ij} \qquad (3)$$

$$Nota: \quad x_{i0} = 1$$

x_{ij} , es el valor que toma la i-ésima observación para el j-ésimo predictor. $i - esima \quad observacion \quad i = 1,2,...,n$

$j - esimo \quad predictor \quad j = 1,2,...,k$

c. Componente de enlace: Una función real $g(.)$, monótona con inversa y derivable, que enlaza a la esperanza matemática de la componente aleatoria Y , con la componente sistemática.

$$x_i^{'}\beta \xrightarrow{g(.)} E(Y/x_1,x_2,...,x_k) \Rightarrow g(E(y/x_1,x_2,...,x_k)) = x_i^{'}\beta$$
$$\Rightarrow g^{-1}(x_i^{'}\beta) = E(y/x_1,x_2,...,x_k)$$

Ejemplo

Sea Y una variable aleatoria con distribución Bernoulli $B(1,\pi_i)$, tal que su función de distribución es dado por:

$$f_i(y_i,\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad y_i = 0, \quad 1 \quad ; \quad i = 1,...,n$$

Se puede expresar como miembro de la familia exponencial:

$$f(y_i,\theta_i,\phi) = Exp\left\{\ln \pi_i^{y_i}(1-\pi_i)^{1-y_i}\right\}$$
$$f(y_i,\theta_i,\phi) = Exp\left\{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\right\}$$
$$f(y_i,\theta_i,\phi) = Exp\left\{y_i \ln \pi_i + \ln(1-\pi_i) - y_i \ln(1-\pi_i)\right\}$$
$$f(y_i,\theta_i,\phi) = Exp\left\{y_i \ln \pi_i - y_i \ln(1-\pi_i) + \ln(1-\pi_i)\right\}$$
$$f(y_i,\theta_i,\phi) = Exp\left\{y_i \ln \frac{\pi_i}{1-\pi_i} + \ln(1-\pi_i)\right\}$$

De (4) Identificando elementos de la familia exponencial

$$f(y_i,\theta,\phi) = Exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}$$

$$\theta = \ln(\frac{\pi_i}{1-\pi_i}) \rightarrow \pi_i = \frac{e^{\theta}}{1+e^{\theta}}$$

$$b(\theta) = -\ln(1-\pi_i) = -\ln(1-\frac{e^{\theta}}{1+e^{\theta}}) = -\ln(\frac{1}{1+e^{\theta}}) = \ln(1+e^{\theta})$$

$$a(\phi) = 1$$

$$c(y_i,\phi) = 0$$

Luego, La distribución de probabilidad Bernoulli es miembro de la familia exponencial.

Para la distribución de Bernoulli

$$x_i^{'}\beta \xrightarrow{g(.)} E(Y/x_1,x_2,...,x_k) \Rightarrow g(E(y/x_1,x_2,...,x_k)) = x_i^{'}\beta$$
$$\Rightarrow g^{-1}(x_i^{'}\beta) = E(y/x_1,x_2,...,x_k)$$

$E(Y/x_1,x_2,...,x_n) = \pi_i(x)$, dependiendo de k variables explicativas

$$g^{-1}(x_i^{'}\beta) = \frac{e^{x^{'}\beta}}{1+e^{x^{'}\beta}}$$

$$\rightarrow \pi_i(x_1,x_2,...x_k) = \frac{e^{x_i^{'}\beta}}{1+e^{x_i^{'}\beta}} = P(Y=1/x_1,x_2,...x_k)$$

Modelo logístico binario múltiple.

2.3.3 Formulación del modelo de regresión logístico binario múltiple.

Un modelo de regresión logístico binario múltiple, es un caso particular de un modelo lineal generalizado, en la cual se estudia la relación funcional entre una variable dependiente Y con distribución Bernoulli, que solo puede tomar dos valores: $Y=1$, lo cual indica la presencia de la característica de interés, con probabilidad de ocurrencia π e $Y=0$, que indica la ausencia de la característica de interés con probabilidad de ocurrencia $1-\pi$ y un conjunto de variables predictoras $(X_1, X_2, ..., X_k)$ que influyen en la probabilidad de ocurrencia de Y.

Supóngase que el modelo de regresión logístico binario múltiple tiene la forma:

$$y_i = x_i^{'}\ \beta + \varepsilon_i \tag{4}$$

Siendo $x_i^{'} = (1, x_{i1}, x_{i2}, x_{i3}, ..., x_{ik})$ $\beta^{'} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$

Como la variable Y es el resultado de un experimento de Bernoulli, las observaciones son independientes, entonces la variable aleatoria Y tiene distribución Bernoulli, con función de probabilidad:

$$P(Y=y_i) = \pi_i^{\,y_i}(1-\pi_i)^{1-y_i} \quad y_i = 0\,,1 \qquad i = 1,2,3,...,n$$

$$P(Y=1) = \pi_i \tag{5}$$

$$P(Y=0) = 1-\pi_i \tag{6}$$

La esperanza y varianza de la variable Y con distribución Bernoulli

La esperanza

$$E(Y_i) = \sum_{y_i=0}^{1} y_i.P(Y=y_i) = 0.P(y_i=0) + 1.P(y_i=1)$$

$$E(Y_i) = 0.(1-\pi_i) + 1(\pi_i) = \pi_i \tag{7}$$

La varianza

$$V(Y_i) = \sum_{y_i=0}^{1} [Y_i - E(Y_i)]^2.P(Y_i=y_i) = [0-\pi_i]^2.P(Y_i=0) + [1-\pi_i]^2.P(Y_i=1)$$

$$V(Y_i) = \pi_i^2(1-\pi_i) + [1-\pi_i]^2\pi_i = \pi_i(1-\pi_i)[\pi_i + (1-\pi_i)] = \pi_i(1-\pi_i) \tag{8}$$

Suponiendo $E(\varepsilon_i)=0$ el modelo (4), el valor esperado de la variable respuesta es:

$$E(y_i)=x_i^{'}\beta \qquad (9)$$

De (5, 7 y 9)

$$E(y_i)=x_i^{'}\beta=\pi_i(x)=P(y_i=1) \qquad (10)$$

(10), indica que cualquiera de los elementos $E(y_i)$, $X_i^{'}\beta$, π_i, es la probabilidad de que la variable respuesta tome el valor 1.

Del modelo (4), como la respuesta es binaria, entonces el término de error ε_i, solo puede tomar dos valores, que son:

$$\varepsilon_i=1-x_i^{'}\beta \qquad para \quad y_i=1$$
$$\varepsilon_i=-x_i^{'}\beta \qquad para \quad y_i=0$$

Situación que indica, que no es posible que los errores aleatorios del modelo (4) tenga una distribución normal.

En el modelo (4), $V(y_i)=V(\varepsilon_i)$ ya que $V(x_i^{'}\beta)=0$, lo que indica que la varianza de las observaciones Y_i es igual a la varianza de los errores, además de la expresión (8):

$$V(y_i)=V(\varepsilon_i)$$

$$V(y_i)=V(\varepsilon_i)=\pi_i(1-\pi_i)$$

$$V(y_i)=V(\varepsilon_i)=\pi_i(1-\pi_i)=E(y_i)(1-E(y_i))$$

Luego se concluye que la varianza del error no es constante (no homocedasticidad) y depende del valor esperado de la variable respuesta. Además, hay una restricción para la función respuesta, de (10).

$$0\le P(Y=1)=\pi_i(x)=x_i^{'}\beta\le 1$$

$$0\le E(y_i)=\pi_i(x)\le 1$$

Esta restricción puede causar graves problemas en la elección de una función matemática de respuesta lineal, como se ha supuesto

en (4). Ya que es posible predecir para un modelo lineal valor negativo o mayores a 1. Es posible ajustar un modelo a los datos, para los cuales los valores predichos de la respuesta pertenecen al intervalo $[0,1]$.

En general, cuando la variable respuesta es binaria, hay muchas pruebas empíricas que indican que la forma de la función matemática de respuesta $\pi_i(x)$, debe ser no lineal. Una función monótona (creciente o decreciente), con inversa y derivable en forma de S o de S invertida, como de la figura 2, es la forma que se acostumbra emplear; esta función matemática se llama función logística y tiene la forma (Montgomery et al. 2004, pp. 399-414).

$$E(Y_i) = \pi_i = P(Y_i = 1/x_1, x_2, ..., x_k) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k}} \qquad (11)$$

Siendo $\pi_i(x_1, x_2, ..., x_k)$, la probabilidad de éxito de una distribución Bernoulli y puede ser interpretada como la probabilidad de que ocurra el evento de interés, dependiendo de un conjunto de variables explicativas (independientes o predictoras), asociados a un conjunto de parámetros, que van a ser estimados.

Modelo logit

Se obtiene al linealizar (11), se realizó la siguiente transformación:

$$P(Y = 1/x_1, x_2, ..., x_k) = \pi_i = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k}}$$

$$Ln\left(\frac{P(Y=1/x_1,x_2,...,x_k)}{1-P(Y=1/x_1,x_2,...,x_k)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_kx_k = \log it(x_1, x_2, ..., x_k) \qquad (12)$$

Ejemplo

El Centro Nacional de Estadística de Salud (NCHS) de los EE. UU, anualmente realiza el NHANES III (Encuesta nacional de examen

de salud y nutrición III), para obtener datos vías exámenes físico, clínicos y prueba de laboratorio en una muestra de 15,149 personas con edades entre 20 a 94 años. El objetivo de este estudio es determinar factores asociados a la enfermedad hipertensión arterial. A fin de modelar la variable respuesta hipertensión que toma solo dos valores 1=tiene la hipertensión y 0=no tiene la enfermedad, se hace uso de la variable predictora edad (años) del paciente como factor relacionado a la enfermedad (tabla 1).

Tabla 1

Edad (años) relacionado con la hipertensión arterial(n=15,149).

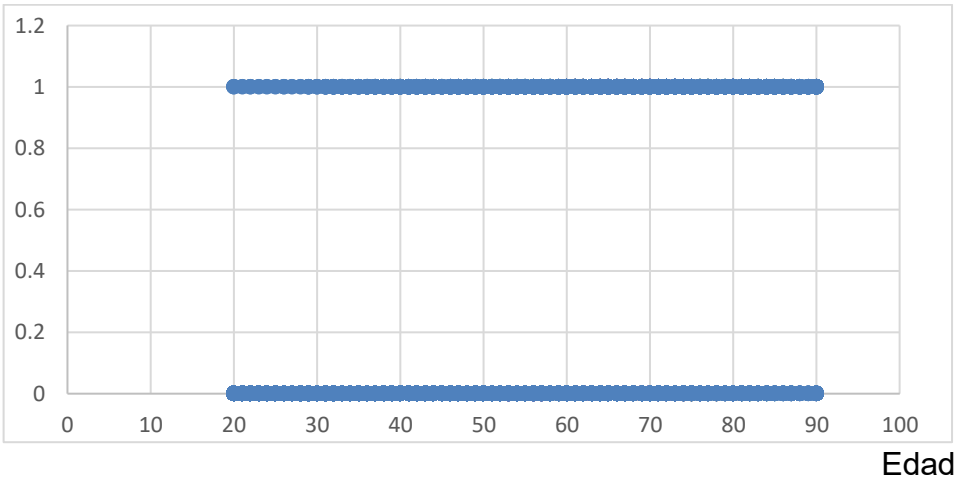
Pacientes	X=Edad(años)	Y=Hipertensión arterial
		Y=1=hipertenso, y=0 no hipertenso
1	52	0
2	25	0
3	42	0
4	84	1
5	81	1
.	.	.
.	.	.
.	.	.
15,149	23	1

Nota: Centro nacional de estadística de salud (NCHS) EEUU. 1988-1994.

Dado que no tenemos idea de cuál puede ser el modelo que se ajuste a los datos, exploraremos los datos y trataremos de identificar un modelo matemático. Con la finalidad de conocer una posible relación entre la variable respuesta hipertensión arterial (Y) y edad del paciente (X), se presenta la dispersión de los datos de la tabla 1, tal como se presenta en la figura 1.

Figura 1.

Edad (años) y la hipertensión arterial (presencia o ausencia).



El figura 1, muestran solamente dos líneas paralelas, representando la presencia (y=1) y la ausencia (y=0) de hipertensión arterial con la edad (años); no se muestra ningún otro tipo de relación entre ambas variables. Para eliminar la variabilidad se clasifica a las edades en grupos etarios y para cada grupo, cuantos no tienen y cuantos si tienen la hipertensión y luego la proporción de pacientes con la hipertensión arterial, esto se presenta en la tabla 2 y figura 2.

Tabla 2.

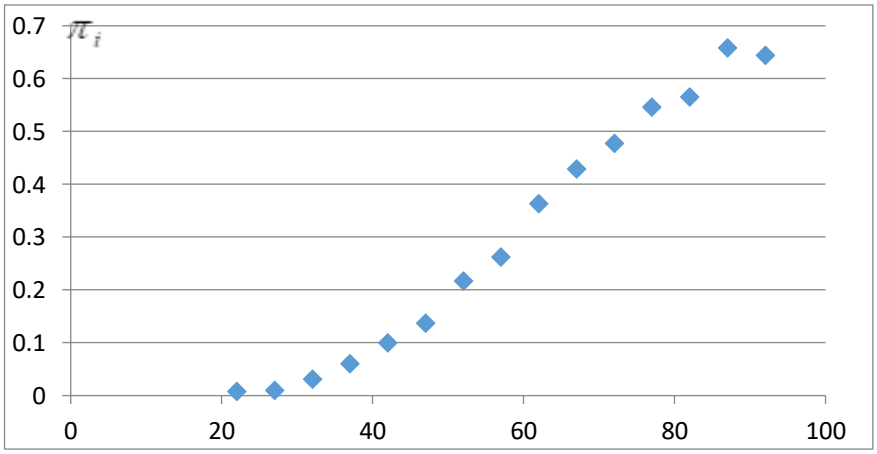
Grupos etarios y proporción de pacientes con hipertensión arterial

Edad del paciente (años)	**Hipertensión arterial**			***Proporción de pacientes con hipertensión*** π_i ***(prevalencia de hipertensión arterial)***
	No	Si	Total	
20-24(1)	1582	12	1594	0.00753
25-29(2)	1524	15	1539	0.00975
30-34(3)	1537	49	1586	0.03090
35-39(4)	1372	88	1460	0.06027
40-44(5)	1298	144	1442	0.09986
45-49(6)	850	135	985	0.13706
50-54(7)	699	194	893	0.21725
55-59(8)	630	224	854	0.26230
60-64(9)	721	412	1133	0.36364
65-69(10)	583	439	1022	0.42955
70-74(11)	515	470	985	0.47716
75-79(12)	276	332	608	0.54605
80-84(13)	307	399	706	0.56516
85-89(14)	91	175	266	0.65789
90-94(15)	27	49	76	0.64474
Total	**12012**	**3137**	**15149**	

Nota. Centro nacional de estadística de salud (NCHS) EEUU. 1988-1994

Figura 2.

Grupos etarios (años) y proporción de pacientes con la hipertensión

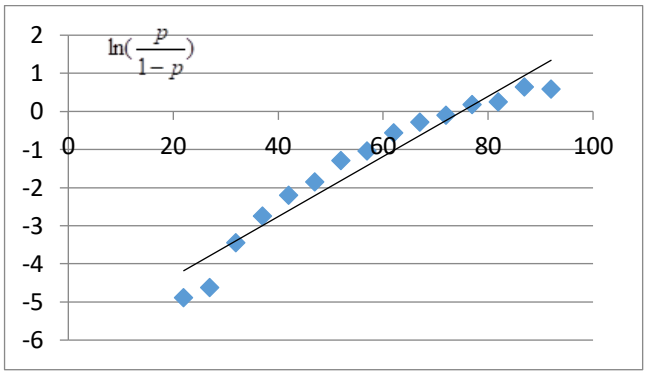


En la figura 2, se observa que la función logística (función matemática), es la que mejor se ajusta (se acomoda) a la nube de puntos. La cual indica que a mayor edad la proporción de pacientes con la enfermedad aumenta y viceversa.

Además, la tabla y figura, indica que para cualquier valor de x, la proporción π_i tomara valores entre 0 y 1. La transformación $\log it(\pi_i) = \ln(\frac{\pi_i}{1-\pi_i})$ con la variable explicativa x, tiene la tendencia aproximadamente lineal (figura 3).

Figura 3

Tendencia lineal entre edad y $\log it(\pi_i)$



2.3.3.1 Suposiciones del modelo de regresión logística binaria múltiple.

Sea el modelo lineal

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

La variable respuesta Y tiene distribución Bernoulli, bajo el supuesto $E(\varepsilon) = 0$, entonces $E(Y) = X\beta = \pi$, luego

$$Y = \pi + \varepsilon \qquad (13)$$

Supuestos:

a. La variable respuesta Y es dicotómica y se distribuye como una Bernoulli con media π_i y varianza $\pi_i(1-\pi_i)$.

b. Las observaciones de la variable respuesta Y son independientes.

c. $V(\varepsilon_i) = \pi_i(1-\pi_i)$, no constante (heteroscedasticidad)

d. Independencia de los errores, las observaciones de la variable Y no deben estar relacionados entre si

e. Existe una relación lineal entre cada variable predictora continua y el logit de la esperanza de Y.
f. El conjunto de variables predictoras, no presentan multicolinealidad, es decir, las variables predictoras no deben estar altamente correlacionadas
g. Con respecto al tamaño de la muestra, se recomienda muestras mayores de 400 casos (Hosmer & Lemeshow, 2000).

2.3.3.2 Estimación de los parámetros por máxima verosimilitud.

En la regresión logística binaria, la variable respuesta Y solo toma dos valores como por ejemplo: (1=presencia de un evento de interés o 0=ausencia de un evento), esta variable aleatoria Y se distribuye como una Bernoulli con media π_i y varianza $\pi_i(1-\pi_i)$, siendo π_i la probabilidad de que ocurra el evento de interés.

Donde las observaciones Y_i son variables aleatorias independientes de Bernoulli, cuyos valores esperados son:

$$E(y_i) = \pi_i(x_{i1},...,x_{ik}) = \frac{e^{\beta_0+\beta_1 x_{i1}+....+\beta_b x_k}}{1+e^{\beta_0+\beta_1 x_{i1}+....+\beta_b x_k}} \qquad (14)$$

Para estimar los parámetros del modelo (14), se usó el método de máxima verosimilitud, que selecciona los coeficientes que hacen más probable que los valores observados ocurran.

Se aplica el método de máxima verosimilitud por ser (14) un modelo no lineal y además el estimador obtenidos por máxima verosimilitud para el parámetro β es óptimo (consistente y eficiente).

Función de verosimilitud

Para estimar los parámetros de (14), se elige una muestra al azar de tamaño n, se tiene para cada observación $(y_i, x_{i1}, x_{i2},..., x_{ik})$

donde y_i, toma valores 0 o 1, y los x_{ij}, toman valores cualitativos o cuantitativos. La función de verosimilitud (Fisher, 1922), es la función de probabilidad conjunta de la variable aleatoria y_i dada por:

$$f(y_1, y_2, ..., y_n, \pi_1, \pi_2, ..., \pi_n) = f(y_1, \pi_1).f(y_1, \pi_1)....f(y_1, \pi_1)$$

Por ser las observaciones independientes e idénticamente distribuidas. A la función de verosimilitud se denota por: $L(\pi)$

$$L(\pi) = f(y_1, \pi_1).f(y_1, \pi_1)....f(y_1, \pi_1)$$

Por ser la función f de probabilidad monótona creciente, un equivalente es usar logaritmos, tomado logaritmos neperianos a ambos lados, se tiene:

$$Ln(L(\pi)) = Ln(f(y_1, \pi_1).f(y_1, \pi_1)....f(y_1, \pi_1))$$

$$Ln(L(\pi)) = \sum_{i=1}^{n} Ln(f(y_i, \pi_i))$$

$$Ln(L(\pi)) = \sum_{i=1}^{n} Ln\,\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$Ln(L(\pi)) = \sum_{i=1}^{n} \left[y_i Ln\,\pi_i + (1-y_i) Ln(1-\pi_i) \right]$$

$$Ln(L(\pi)) = \sum_{i=1}^{n} \left[y_i Ln(\frac{\pi_i}{1-\pi_i}) + Ln(1-\pi_i) \right] \qquad (15)$$

Como:

$$\pi_i(x_{i1}, ..., x_{ik}) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_b x_k}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_b x_k}}$$

Sus equivalentes:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_b x_k} \rightarrow Ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_b x_k \qquad (16)$$

También

$$1 - \pi_i = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_b x_k}} \rightarrow Ln(1-\pi_i) = -Ln(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_b x_k}) \qquad (17)$$

Haciendo los reemplazos (16) y (17) en (15), la función de verosimilitud, tiene la forma:

$$Ln(L(\pi)) = \sum_{i=1}^{n}\left[y_i(\beta_0 + \beta_1 x_{i1} + ...\beta_k x_{ik}) - Ln(1 + e^{\beta_0+\beta_1 x_{i1}+...\beta_k x_{ik}})\right] \qquad (18)$$

La función de verosimilitud depende de los parámetros desconocidos, luego:

$$LL(\beta) = \sum_{i=1}^{n}\left[y_i(\beta_0 + \beta_1 x_{i1} + ...\beta_k x_{ik}) - Ln(1 + e^{\beta_0+\beta_1 x_{i1}+...\beta_k x_{ik}})\right] \qquad (19)$$

Para maximizar la función de verosimilitud (19), se deriva parcialmente con respecto a cada parámetro β_j e igualando a cero.

$$\frac{\partial(LL(\beta))}{\partial\beta_j} = \sum_{i=1}^{n}\left[y_i x_{ij} - \frac{e^{\beta_0+\beta_1 x_{i1}+...\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+...\beta_k x_{ik}}} x_{ij}\right] = 0 \qquad (20)$$

Nota: $x_{i0} = 1$ para todo $i = 1,2,...,n \qquad j = 0,1,2,...,k$

(20) también se escribe:

$$\frac{\partial(LL(\beta))}{\partial\beta_j} = \sum_{i=1}^{n}\left[y_i x_{ij} - \pi_i x_{ij}\right] = \sum_{i=1}^{n}\left[(y_i - \pi_i)x_{ij}\right] = 0 \qquad j = 0,1,2,...,k$$

(20) se puede escribir como una función de los parámetros:

$$f(\beta_0,\beta_1,...,\beta_k) = \sum_{i=1}^{n}(x_{ij}y_i - x_{ij}\frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_j x_{ij}+..+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_j x_{ij}+..+\beta_k x_{ik}}}) = 0 \qquad (21)$$

Se tiene $(k+1)$ ecuaciones no lineales, las cuáles permitió la estimación de los $(k+1)$ parámetros, resolver este sistema de ecuaciones no lineal, no es sencillo, tampoco imposible, gracias al método matemático iterativo de Newton-Raphson (Chapra & Canale, 2006) y la programación, resolvieron este problema calculando los coeficientes del modelo logístico binario, la $(t+1)$ aproximación para la función (21), está dada por:

$$\begin{pmatrix} \beta_0^{t+1} \\ \beta_1^{t+1} \\ . \\ . \\ . \\ \beta_k^{t+1} \end{pmatrix}_{kx1} = \begin{pmatrix} \beta_0^{t} \\ \beta_1^{t} \\ . \\ . \\ . \\ \beta_k^{t} \end{pmatrix} - \begin{pmatrix} \frac{\partial f_0}{\partial \beta_0} & \frac{\partial f_0}{\partial \beta_1} & . & . & . \frac{\partial f_0}{\partial \beta_k} \\ \frac{\partial f_1}{\partial \beta_0} & \frac{\partial f_1}{\partial \beta_1} & . & . & . \frac{\partial f_1}{\partial \beta_k} \\ . & & & & \\ . & & & & \\ \frac{\partial f_k}{\partial \beta_0} & \frac{\partial f_k}{\partial \beta_1} & . & . & . \frac{\partial f_k}{\partial \beta_k} \end{pmatrix}^{-1} \begin{pmatrix} f_0(x_0, y_0, ..z_0) \\ f_1(x_0, y_0, ..z_0) \\ . \\ . \\ f_k(x_0, y_0, ..z_0) \end{pmatrix}$$

Los programas estadísticos, utilizan como valores iniciales para los coeficientes estimados el vector $\beta^t = (0, \quad 0,...,0)$.

El algoritmo se detiene cuando $f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k) \to 0$

O también los $\hat{\beta}_j$ en las sucesivas iteraciones son casi las mismas.

El modelo logístico binario múltiple estimado:

$$\hat{\pi}_i(x_1, x_2, ..., x_k) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k}} \tag{22}$$

$$\hat{\pi}_i(x_1, x_2, ..., x_k) \to P(Y_i = 1/x_1, x_2, ..., x_k) = E(Y = 1/x_1, x_2, ...x_k) \tag{23}$$

el modelo estimado vectorialmente es

$$\hat{\pi}_i(x_1, x_2, ..., x_k) = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}}$$

siendo:

$x' = (1,, x_1, ..., x_k)$ y $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$, vector que estima al parámetro $\beta' = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$,

El modelo (22) se puede expresar como:

$$P(Y = 1/x_1, x_2, ..., x_k) = \hat{\pi}_i(x_1 . x_2, ..., x_k) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)}} \tag{24}$$

$$\frac{\hat{\pi}_i(x_{i1}, x_{i2}, ..., x_{ik})}{1 - \hat{\pi}_i(x_{i1}, x_{i2}, ..., x_{ik})} = \frac{P(Y = 1/x_{i1}, x_{i2}, ..., x_{ik})}{1 - P(Y = 1/x_{i1}, x_{i2}, ..., x_{ik})} == e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_b x_{ik}} \tag{25}$$

De (25) aplicando logaritmo neperiano a ambos miembros de la igualdad, se obtiene un modelo lineal

$$Ln(\frac{\hat{\pi}_i(x_{i1},x_{i2},...,x_{ik})}{1-\hat{\pi}_i(x_{i1},x_{i2},...,x_{ik})}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_b x_{ik} \quad (26)$$

También de (19), se tiene las siguientes propiedades

$$LnL(\beta) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_{i1} + ...\beta_k x_{ik}) - Ln(1 + e^{\beta_0 + \beta_1 x_{i1} + ...\beta_k x_{ik}}) \right]$$

Propiedades de la función de verosimilitud:

- $0 < L(\beta) < 1$
- $LnL(\beta) < 0$
- $-2LL(\beta) \geq 0$
- $-2LL(\beta) \approx \chi_{(n-k-1)gl}$
- $LnL(\hat{\beta}) = \sum_{i=1}^{n} \{ y_i Ln\hat{\pi}_i + (1-y_i) Ln(1-\hat{\pi}_i) \}$

El valor de esta última expresión es similar a la suma de cuadrados de los residuos de la regresión lineal múltiple, si $y_i \approx \hat{\pi}_i$ el modelo logístico estimado se ajusta bien a los datos observados (variable respuesta), en otras palabras, el modelo estimado recibe mucha información de la variable respuesta, todo lo contrario, ocurre si son muy diferentes $y_i \neq \hat{\pi}_i$. Además se demuestra que el estadístico $-2LnL(\hat{\beta})$ sigue una distribución asintótica (para n grande) Chi-Cuadrado con (n-k-1) gl (McCullagh & Nelder, 1983).

La función Score

Se define a la función Score como la derivada parcial del logaritmo de la función de verosimilitud respecto al parámetro (Evans & Rosenthal, 2013).

$$S(y,\theta) = \frac{\partial LnL(\theta / x)}{\partial \theta}$$

Propiedades:

Teorema de Cramer-Rao (Evans & Rosenthal, 2013).

$$E\left[S(y,\theta) = \frac{\partial LnL(\theta / x)}{\partial \theta} \right] = 0$$

La matriz de información de Fisher (Evans & Rosenthal, 2013).

$$I(\theta) = -E\left[\frac{\partial^2 LnL(\theta / x)}{\partial \theta \partial \theta'} \right]$$

De la función de verosimilitud (19)

$$LL(\beta) = \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_{i1} + ... \beta_k x_{ik}) - Ln(1 + e^{\beta_0 + \beta_1 x_{i1} + ... \beta_k x_{ik}}) \right]$$

$$\frac{\partial (LL(\beta))}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{e^{\beta_0 + \beta_1 x_{i1} + ... \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... \beta_k x_{ik}}} x_{ij} \right] \qquad j = 0,1,2,...,k$$

La matriz de información de Fisher

Su valor se define como la esperanza matemática del negativo de la matriz de segunda derivada del logaritmo de la función de verosimilitud (Evans & Rosenthal, 2013).

$$I(\beta) = -E\left[\frac{\partial^2 (LL(\beta))}{\partial \beta_j \partial \beta_m} \right] = \qquad j, m = 0,1,2,...,k \qquad (27)$$

Calculamos la segunda derivada, para $m = j$

$$-\frac{\partial^2 (LL(\beta))}{\partial \beta_j^2} = \sum_{i=1}^{n} \left[\pi_i (1 - \pi_i) x_{ij}^2 \right] \qquad j = 0,1,2,...,k \qquad (28)$$

También, para $m \neq j$

$$-\frac{\partial^2 (LL(\beta))}{\partial \beta_j \partial \beta_m} = \sum_{i=1}^{n} \left[x_{ij} x_{im} \pi_i (1 - \pi_i) \right] \qquad j; m = 0,1,2,...,k \qquad (29)$$

Con los elementos de las segundas derivadas parciales 28 y 29, se obtiene la matriz de información de Fisher, de orden $(k+1)x(k+1)$, denotada por $I(\hat{\beta})$:

$$I(\hat{\beta}) = \begin{bmatrix} \frac{\partial^2 LL(\beta)}{\partial \beta_0^2} & \frac{\partial^2 LL(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & , \frac{\partial^2 LL(\beta)}{\partial \beta_0 \partial \beta_k} \\ \frac{\partial^2 LL(\beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 LL(\beta)}{\partial \beta_1^2} & \cdots & , \frac{\partial^2 LL(\beta)}{\partial \beta_1 \partial \beta_k} \\ . & & & \\ . & & & \\ . & & & \\ \frac{\partial^2 LL(\beta)}{\partial \beta_k \partial \beta_0} & \frac{\partial^2 LL(\beta)}{\partial \beta_k \partial \beta_1} & \cdots & , \frac{\partial^2 LL(\beta)}{\partial \beta_k^2}. \end{bmatrix} \quad (30)$$

También

$$(I(\hat{\beta})) = X^i V X$$

La inversa de la matriz de información de Fisher (24), proporciona la matriz varianza y covarianza de $\hat{\beta}$

$$(I(\hat{\beta}))^{-1} = V(\hat{\beta}) = \begin{bmatrix} \frac{\partial^2 LL(\beta)}{\partial \beta_0^2} & \frac{\partial^2 LL(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & , \frac{\partial^2 LL(\beta)}{\partial \beta_0 \partial \beta_k} \\ \frac{\partial^2 LL(\beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 LL(\beta)}{\partial \beta_1^2} & \cdots & , \frac{\partial^2 LL(\beta)}{\partial \beta_1 \partial \beta_k} \\ . & & & \\ . & & & \\ . & & & \\ \frac{\partial^2 LL(\beta)}{\partial \beta_k \partial \beta_0} & \frac{\partial^2 LL(\beta)}{\partial \beta_k \partial \beta_1} & \cdots & , \frac{\partial^2 LL(\beta)}{\partial \beta_k^2}. \end{bmatrix}^{-1} \quad (31)$$

También se calcula la matriz de varianza y covarianza por:

$$V(\hat{\beta}) = (I(\beta))^{-1} = (X^{'} V X)^{-1} \quad (32)$$

siendo

$$X_{nx(k+1)} = \begin{bmatrix} 1 & x_{11}, & x_{12}, & . & . & .x_{1k} \\ 1 & x_{21}, & x_{22}, & . & . & .x_{2k} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ 1 & x_{n1}, & x_{n2}, & . & . & .x_{nk} \end{bmatrix}$$

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

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

$$V(\hat{\beta}) = \begin{bmatrix} \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i) & \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{i1}... & \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{ik} \\ \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{i1} & \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{i1}^2... & \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{i1}x_{ik} \\ . & & \\ . & & \\ . & & \\ \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{ik} & \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{i1}x_{ik}... & \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_{ik}^2 \end{bmatrix}^{-1}$$

Se demuestra que el estimador de máxima verosimilitud $\hat{\beta}$ es asintóticamente insesgado

$$E(\hat{\beta}) = \beta \qquad (33)$$

Propiedades del estimador máximo verosímiles.

Sea $\hat{\beta}$, el estimador obtenido por el método de máxima verosimilitud para el parámetro β, tiene las siguientes propiedades para muestras grandes:

- **Consistencia**

Los estimadores obtenidos por el método de máxima verosimilitud son consistentes, converge en probabilidad al valor verdadero, cuando el tamaño de la muestra tiende al infinito.

$$PLim\,\hat{\beta} = \beta$$

- **Normalidad asintótica**

Para un tamaño de muestra grande, la distribución muestral del estimador $\hat{\beta}$ es asintóticamente normal.

$$\hat{\beta} \xrightarrow{a} N(\beta;\ \ [I(\beta)]^{-1})$$

Siendo:

$[I(\beta)]^{-1}$, la inversa de la matriz de información de Fischer.

- **Eficiencia asintótica.**

El estimador $\hat{\beta}$ obtenidos por el método de máxima verosimilitud es eficientes, dentro de la clase de estimador consistente, asintóticamente insesgado y asintóticamente normales.

- **Invarianza**

Si $\hat{\beta}$ es un estimador obtenido por el método de máxima verosimilitud y $w = g(\beta)$ es un función continua de β , entonces $\hat{w} = g(\hat{\beta})$ es un estimador máximo verosímil de w.

2.3.3.3 Interpretación $\hat{\beta}_j$ y $O\hat{R}_j$ (Odds ratio).

La interpretación de los coeficientes del modelo logístico estimado, se hace más fácil, si se presenta en términos de Odds (cociente de la probabilidad de que ocurra el evento de interés y la probabilidad de que no ocurra el evento de interés).

Del modelo (25), escrito en términos de Odds.

$$Odds(x_1; x_2,...,x_j,...,x_k) = \frac{P(Y=1/x_1,x_2,...,x_j...,x_k)}{1-P(Y=1/x_1,x_2,...x_j,..x_k)} = e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+...+\beta_j x_j+...+\hat{\beta}_b x_{ik}} \quad (34)$$

Supongamos, que al predictor x_j , se incrementa en $x_j + c$ manteniendo fijos a los otros predictores.

$$Odds(x_1,x_2,...,x_j+c...,x_k) = \frac{P(Y=1/x_1,x_2,...,x_j+c...,x_k)}{1-P(Y=1/x_1,x_2,...x_j+c,..x_k)} = e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+...+\beta_j(x_j+c)+...+\hat{\beta}_b x_{ik}} \quad 35)$$

Dividiendo las expresiones (35) y (34), se obtiene el Odds-ratio=OR

$$OR = \frac{Odds(x_1,...x_j+c,...,x_k)}{Odds(x_1,...x_j,...,_k)} = \frac{\dfrac{P(y=1/x_j+c)}{1-P(y=1/x_j+c)}}{\dfrac{P(y=1/x_j)}{1-P(y=1/x_j)}} = e^{c\hat{\beta}_j} \quad (36)$$

Supongamos, que el incremento es la unidad, entonces $c=1$, y el $OR=e^{\hat{\beta}_j}$, en otras palabras los coeficientes del modelo estimado al ser exponente de la base de los logaritmos neperianos, es la razón de Odds, llamados los Odds ratio=OR, que se obtiene al incrementar en una unidad al predictor x_j, permaneciendo fijos los otros predictores, su valor se interpreta dependiendo del signo de $\hat{\beta}_j$.

Si $\hat{\beta}_j>0$, entonces $OR=e^{\hat{\beta}_j}>1$, indica que si el predictor x_j se incrementa en una unidad, permaneciendo fijos los otros predictores, la posibilidad de que ocurra el evento $P(Y=1/x_j)$ se incrementa en $OR=e^{\hat{\beta}_j}$, veces mayor; en el campo epidemiológico se dice que la variable predictora x_j es una variable de riesgo para la enfermedad.

Si $\hat{\beta}_j<0$, entonces $OR=e^{\hat{\beta}_j}<1$, indica que si el predictor x_j se incrementa en una unidad, permaneciendo fijos los otros predictores, la posibilidad de que ocurra el evento $P(Y=1/x_j)$, disminuye en $OR=e^{\hat{\beta}_j}$, veces menor; en este caso la variable predictora x_j cumple un efecto protector para la enfermedad; cuando $\hat{\beta}_j<0$ es mejor interpretar el inverso del OR.

Si $\hat{\beta}_j=0$, entonces $OR=e^{\hat{\beta}_j}=1$, indica que si el predictor x_j se incrementa en una unidad, permaneciendo fijos los otros predictores no produce cambio en la posibilidad de ocurrencia del evento $P(Y=1/x_j)$, este caso la variable predictora x_j no es factor riesgo ni protector, para la enfermedad en el campo de la medicina.

Cuando la variable predictora es cualitativa

Cuando una variable predictoras es cualitativas (nominal u ordinal), con t categorías, se tiene t-1 variables dicotómicas como variables predictoras: Supongamos que una variable cualitativa tiene 3

categorías, entonces se necesita t-1=3-1=2 variables dicotómicas $X_{i1} \quad y \quad X_{i2}$, tal como se detalla en el siguiente tabla 3.

Tabla 3

Variable predictora con tres categorías, necesita dos variables dicotómicas.

	Variables dicotómicas	
Categorías	X_{i1}	X_{i2}
1	1	0
2	0	1
3	0	0

$X_{i1} = 1$, si el dato pertenece a la categoría 1 y 0 si el dato no pertenece a categoría 1.

$X_{i2} = 1$, si el dato pertenece a la categoría 2, 0 si el dato no pertenece a categoría 2.

$X_{i3} = 0$, si el dato pertenece a la categoría 1 y categoría 2 (variable dummy de referencia con la cual se realizan las comparaciones)

$$P(Y = 1/x_1, x_2) = \frac{e^{\beta_0 + \beta_0 x_{i1} + \beta_2 x_{i2}}}{1 + e^{\beta_0 + \beta_0 x_{i1} + \beta_2 x_{i2}}}$$

$$Odds(x_1, x_2) = \frac{P(Y = 1/x_1, x_2)}{1 - P(Y = 1/x_1, x_2)} = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}$$

$$= \log it(Odds(x_1, x_2)) = Ln\left[\frac{P(Y = 1/x_1, x_2)}{1 - P(Y = 1/x_1, x_2)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

$$Odds_1 = \frac{p(Y = 1/x_1 = 1, x_2 = 0)}{1 - p(Y = 1/x_1 = 1, x_2 = 0)} = e^{\beta_0 + \beta_1}$$

$$Odds_2 = \frac{p(Y = 1/x_1 = 0, x_2 = 1)}{1 - p(Y = 1/x_1 = 0, x_2 = 1)} = e^{\beta_0 + \beta_2}$$

$$Odds_3 = \frac{p(Y = 1/x_1 = 0, x_2 = 0)}{1 - p(Y = 1/x_1 = 0, x_2 = 0)} = e^{\beta_0} \qquad (referencia)$$

Calculo de los Odds-ratio (OR):

El cociente de $odds_1$ con $odds_3$ (referencia)

$$OR = \frac{Odds_1}{Odds_3} = e^{\beta_1}$$

Si $\beta_1 > 0$, la categoría 1 tiene e^{β_1} veces mayor posibilidad de ocurrir el evento de interés $P(Y=1)$, comparado con la categoría 3 de referencia, manteniendo fijos las otras variables.

Si $\beta_1 < 0$, la categoría 1 tiene e^{β_1} veces menor posibilidad de ocurrir el evento de interés $P(Y=1)$, comparado con la categoría 3 de referencia, manteniendo fijos las otras variables. Cuando ocurre esto es preferible calcular el inverso de e^{β_1} y la interpretación seria la categoría 3 tiene $1/e^{\beta_1}$ veces mayor posibilidad de ocurrir el evento de interés $P(Y=1)$, comparado con la categoría 1.

Si $\beta_1 = 0$, entonces $e^{\beta_1} = 1$, indica que la categoría 1 no ejerce ningún efecto sobre la posibilidad de ocurrir el evento de interés $P(Y=1)$, comparado con la categoría 3, es decir la categoría 1 y 3 de referencia tienen el mismo efecto.

De igual forma también el cociente del $odds_2$ y $odds_3$ (referencia)

$$OR = \frac{Odds_2}{Odds_3} = e^{\beta_2}$$

2.3.4 Evaluación del modelo estimado

2.3.4.1 Prueba de bondad de ajuste estadística G

Al igual que la prueba F, se usa para probar la significación de las variables predictoras en el modelo de regresión múltiple a través de sus parámetros. El modelo logístico binario, se basa en la prueba de razón de máxima verosimilitud, que es un procedimiento para muestras grandes y la teoría asintótica. El método de la razón de verosimilitud (RV) conduce a un estadístico llamado desviación o desvianza (D). La desvianza es la distancia (lejanía) entre -2 veces el logaritmo de la función de verosimilitud del modelo saturado (aquel modelo que tiene tanto parámetros como observaciones n) y -2 veces el logaritmo de la verosimilitud del modelo propuesto (modelo en estudio con k+1 parámetros)

$$D = -2Ln\left[\frac{Verosimilitud \quad del \quad \text{mod}elo \quad propuesto}{Verosimilitud \quad del \quad \text{mod}elo \quad saturado}\right] \tag{37}$$

(37), es una estadística, cuya distribución de probabilidad se aproxima asintóticamente a una distribución Chi-Cuadrado con $(n-k-1)$ .grados de libertad.

La desvianza sólo con la constante con 1 parámetro (intercepto) y el saturado con cero grados de libertad.

$$D_1 = -2Ln\left[\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_n)}\right] \approx \chi^2_{((n-1)-(n-n))gl} = \chi^2_{(n-1)gl}$$

La desvianza con el modelo propuesto con k+1 parámetros y el modelo saturado con n parámetro:

$$D_2 = -2Ln\left[\frac{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_n)}\right] \approx \chi^2_{(n-k-1-(n-n))gl} = \chi^2_{(n-k-1)gl}$$

La estadística G , la diferencia de $D_1 \quad y \quad D_2$

$$G = D_1 - D_2 = -2Ln\left[\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_n)}\right] - \left\{-2Ln\left[\frac{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_n)}\right]\right\} \approx \chi^2_{(k)gl}$$

La diferencia indica la mejora del modelo al agregar las k variables predictoras. Si la desvianza del modelo que solo tiene la constante disminuye considerablemente cuando se usa el modelo con las k variables, se puede decir que el modelo propuesto es adecuado a los datos de la muestra.

$$RV = G = -2Ln\left[\frac{L(\hat{\beta}_0)}{L(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k)}\right] \approx \chi^2_{(k)gl} / H_o \quad cierta \tag{38}$$

$$RV = G = -2LnL(\hat{\beta}_0) - \left\{-2LnL(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_k)\right\} \approx \chi^2_{(k)gl}$$

También puede calcularse como:

$$RV = G = -2\left[n_o \ln(n_0) - n\ln(n) + (n-n_0)\ln(n-n_0)\right]$$
$$-\left[-2\sum_{i=1}^{n}\left[y_i \ln\hat{\pi}_i + (1-y_i)\ln(1-\hat{\pi}_i)\right]\right] \quad , \quad n_o = \sum_{i=1}^{n} y_i$$

Dado que: Bajo Ho cierta, la estadística G (Hosmer e Lemeshow (2000), sigue una distribución Chi-Cuadrado con k grados de libertad, siendo k el número de variables predictoras.

La estadística G, definida por cociente de verosimilitudes del modelo solo con la constante (β_0) y no contiene predictores, y el modelo propuesto con los $k+1$ coeficiente $(\beta_0, \beta_1, \beta_2, ..., \beta_k)$ asociados a k predictores. La hipótesis nula indica que todos los parámetros del modelo son nulo a excepción el intercepto, esto significa que las variables predictoras asociadas no se relacionan con la probabilidad de ocurrencia del evento $P(Y=1)$, si la hipótesis nula se rechaza con un nivel de significación α, significa al menos algunas de las variables predictoras $(X_1, X_2, ..., X_k)$, sirve para explicar la probabilidad de ocurrencia del evento $P(Y=1)$.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_a : por \quad lo \quad menos \quad a\lg un \quad \beta_j \neq 0 \quad j = 1,2,..,k$$

$Si \quad G > \chi^2_{(\alpha;\, k)}$, se rechaza la hipótesis nula con un nivel de significación del α; es decir existe por lo menos una variable predictora que se relaciona o influye con la probabilidad de ocurrencia del evento de interes; en caso contrario no se rechaza la hipótesis nula.

Para probar la significación de un subconjunto de variables explicatorias del modelo estimado (Uriel & Aldas, 2005).

$$Ln\frac{P(Y=1/X)}{1-P(Y=1/X)} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_m X_m + \hat{\beta}_{m+1} X_{m+1} + ... + \hat{\beta}_k X_k$$

Supongamos que con un nivel de significación α se desea probar la hipótesis:

$$H_0 : \beta_{m+1} = \beta_{m+2} = ... = \beta_k = 0$$

$$H_a : \text{Algún elemento de } \mathrm{H}_0 \text{ no es nulo}$$

El estadístico para la prueba:

$-2LL(\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_m) \approx \chi^2_{(n-m-1)gl}$, la desvianza del logaritmo de la función de la verosimilitud que se obtiene al estimar el modelo con los m+1 parámetros, bajo la hipótesis nula cierta:

$-2LL(\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_k) \approx \chi^2_{(n-k-1)gl}$, la desvianza del logaritmo de la función de la verosimilitud que se obtiene al estimar el modelo en estudio con los k+1 parámetros.

2.3.4.2 Prueba de significación de los coeficientes, estadística de Wald.

Si los datos proporcionan evidencia significativa para rechazar la hipótesis nula, a través de la estadística G, el siguiente paso es evaluar la significancia individual de cada parámetro del modelo, se formula las siguientes hipótesis:

$$H_o : \beta_j = 0 \Leftrightarrow H_0 : OR = 1$$
$$H_o : \beta_j \neq 0 \Leftrightarrow H_0 : OR \neq 1$$

Aplicando la teoría asintótica de los estimadores máximos verosímiles, se demuestra para un tamaño de muestra grande, la distribución aproximada de $\hat{\beta}$, obtenido por el método de máxima verosimilitud, es asintóticamente normal. Wald (1943), es decir

El estadístico:

$$\hat{\beta}_{(k+1)x1} \approx N(\beta_{(k+1)x1} \quad ; \quad (X^{'}VX)^{-1}_{(k+1)x(k+1)})$$

$$V_{nxn} = diag[\hat{p}_1(1-\hat{p}_1),\ \hat{p}_2(1-\hat{p}_2),..., \hat{p}_n(1-\hat{p}_n)]$$

Luego:

$$Z = \frac{\hat{\beta}_j - \beta_o}{\sigma_{\beta_j}} \xrightarrow{asintoticamente} \approx N(0,1)$$

Bajo la hipótesis nula cierta

$$W = (\frac{\hat{\beta}_j}{\hat{S}_{\hat{\beta}_j}})^2 \approx \chi^2_{(1)gl}$$

Nota: $\hat{S}^2_{\hat{\beta}_j}$ es elemento de la diagonal principal de la matriz $[X'VX]^{-1}$

$Si \quad W > \chi^2_{(1,\alpha)gl}$

Se rechaza la hipótesis nula a un nivel de significación de α, lo cual significa que el predictor x_j influye en la probabilidad de ocurrencia de la variable respuesta.

2.3.4.3 Estimación de β_j por Intervalos de confianza

Límite de confianza $(1-\alpha)100\%$, para el parámetro β_j, es dado por:

$$LC(\beta_j) = \hat{\beta}_j \pm Se(\hat{\beta}_j).Z_{1-\frac{\alpha}{2}} \qquad j = 1,2,3,...,k \qquad (39)$$

Interpretación: existe una confianza del $100(1-\alpha)\%$ de que el intervalo contenga al verdadero valor de β_j.

2.3.4.4 Estimación de OR por Intervalos de confianza

En la práctica, es más significativo obtener intervalos de confianza para OR ajustado:

$$LC(OR) = e^{\hat{\beta}_j \pm Se(\hat{\beta}_j).Z_{1-\frac{\alpha}{2}}} \qquad j = 1,2,3,...,k \qquad (40)$$

Interpretación: existe una confianza del $100(1-\alpha)\%$ de que el intervalo contenga al verdadero valor de OR.

2.3.4.5 Coeficiente de determinación de McFadden, Cox & Snell y Nagelkerke, como medidas de bondad de ajuste.

Al igual que en los modelos de regresión lineal múltiple, en la regresión logística binaria múltiple, es necesario medir la bondad de ajuste del modelo estimado a través de los pseudos(parecidos) coeficientes de determinación, ellos miden la variabilidad de los datos observados de la variable respuesta, explicado por las variables predictoras del modelo logístico binario estimado.

El R^2 de McFadden (1974), mide la variación de los valores que toma la variable respuesta explicada por el modelo logístico binario estimado, su valor oscila entre 0 y aproximadamente 1.

$$R^2_{McFadden} = 1 - \frac{LL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)}{LL(\hat{\beta}_0)} \tag{41}$$

$LL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k) =$ Logaritmo neperiano de la función de verosimilitud del modelo en estudio

$LL(\hat{\beta}_0) =$ Logaritmo neperiano de la función de verosimilitud del modelo nulo.

$k =$ Numero de predictores del modelo estimado.

El R^2_{CS} de Cox & Snell (1989), también estima la proporción de la variabilidad de la variable dependiente (datos de y) explicada por las variables predictoras (independientes). Su valor se basa en la comparación de la desvianza para el modelo solo con el intercepto y la desvianza para un modelo en estudio, su valor oscila entre 0 y aproximadamente 1, incluso para un modelo “perfecto”.

$$R^2_{CS} = 1 - \left(\frac{e^{-2LL(\hat{\beta}_0)}}{e^{-2LL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)}}\right)^{-\frac{1}{n}} = 1 - e^{-\left(\frac{G}{n}\right)} \tag{42}$$

El R^2_N de Nagelkerke (1991), es una versión corregida del R^2 de Cox & Snell (1989), el cual tiene un valor máximo inferior a la unidad. La El R^2 de Nagelkerke, corrige la escala del estadístico para cubrir el rango completo de $[0, \quad 1]$

$$R^2_N = \frac{1 - \left(\frac{e^{-2LL(\hat{\beta}_0)}}{e^{-2LL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)}}\right)^{-\frac{1}{n}}}{1 - \left(e^{-2LL(\hat{\beta}_0)}\right)^{-\frac{1}{n}}} = \frac{R^2_{CS}}{1 - \left(e^{(G + -2LL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k))}\right)^{-\frac{1}{n}}} \tag{43}$$

2.3.4.6 Prueba de la bondad de ajuste de Hosmer- Lemeshow (2000).

Evalúa la bondad de ajuste del modelo de la regresión logística binario estimado, Hosmer–Lemeshow (2000), su estrategia es una vez ajustado, se puede calcular las probabilidades estimadas $\hat{\pi}_i$ para cada observación, a partir de estos resultados se pueden formar los llamados grupos de riesgos que se consiguen dividiendo el conjunto ordenado de esas probabilidades estimadas en g-grupos (deciles).
Hosmer–Lemeshow (2000), prueba la siguiente hipótesis:

H_o : El modelo se ajusta adecuadamente a las observaciones

H_1 : El modelo no se ajusta adecuadamente a las observaciones

El estadístico de Prueba:

$$\chi_c^2 = \sum_{i=0}^{1}\sum_{j=1}^{g}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \approx \chi^2_{(g-2)gl} \qquad (44)$$

$\chi_c^2 > \chi^2_{1-\alpha}(g-2)$, rechazamos la hipótesis nula con un nivel de significación α, es decir el modelo no se ajusta adecuadamente a las observaciones.

$\chi_c^2 < \chi^2_{1-\alpha}(g-2)$, no rechazamos la hipótesis nula, es decir el modelo se ajusta adecuadamente a las observaciones. El estadístico de prueba, no se computa para frecuencias esperadas menores a 5. La tabla 4 presenta las frecuencias observadas y las esperadas.

Tabla 4

Tabla de contingencia para calcular la estadística de Hosmer–Lemeshow

Grupos	Y=0		Y=1		total
	observado	**Esperado**	**observado**	**esperado**	
1	o_{01}	$e_{01} = n_{01}\overline{\pi}_{01}$	o_{11}	$e_{11} = n_{11}\overline{\pi}_{11}$	$o_{.1}$
2	o_{02}	$e_{02} = n_{02}\overline{\pi}_{02}$	o_{12}	$e_{12} = n_{12}\overline{\pi}_{12}$	$o_{.2}$
3	o_{03}	$e_{03} = n_{03}\overline{\pi}_{03}$	o_{13}	$e_{13} = n_{13}\overline{\pi}_{13}$	$o_{.3}$
.	.	.	.	.	.
g	o_{0g}	$e_{0g} = n_{0g}\overline{\pi}_{0g}$	o_{1g}	$e_{1g} = n_{1g}\overline{\pi}_{1g}$	$o_{.g}$
Total	$\sum_{j=1}^{g} o_{0j} = \sum_{j=1}^{g} e_{oj} = \alpha_1$		$\sum_{j=1}^{g} o_{1j} = \sum_{j=1}^{g} e_{1j} = \alpha_2$		$n = \alpha_1 + \alpha_2$

2.3.4.7 Capacidad predictiva del modelo estimado

El modelo de regresión logístico estimado permite establecer una regla para clasificar las observaciones en cada una de las categorías $(y_i = 1, o\ \ y_i = 0)$. La regla de clasificación se establece al determinar un punto de corte (c) entre $(0 \le P(Y_i) \le 1)$. Si para una observación la probabilidad de que $y_i = 1$, es mayor que el punto de corte, se considera que esta observación pertenece a la clase de interés para $y_i = 1$, de lo contrario la observación pertenece a la clase determinada para $y_i = 0$. Existen técnicas clásicas del desempeño clasificatorio del modelo estimado, se presenta a continuación:

Tabla de confusión

Es una tabla de doble entrada de 2x2, la cantidad de elementos de cada casilla provienen de comparar lo observado por la variable respuesta $(y_i = 1, o\ \ y_i = 0)$ con la probabilidad $\hat{\pi}_i$ obtenida por el modelo de

regresion logístico estimado y teniendo como criterio de clasificación el punto de corte elegido, que por lo general es $c = 0.5$

$$Y_i = 1 \begin{cases} \hat{\pi}_i \geq 0.5 \rightarrow \in & Y_i = 1 \\ \hat{\pi}_i < 0.5 \rightarrow \notin & Y_i = 1 \end{cases} \qquad Y_i = 0 \begin{cases} \hat{\pi}_i \geq 0.5 \rightarrow \notin & Y_i = 0 \\ \hat{\pi}_i < 0.5 \rightarrow \in & Y_i = 0 \end{cases}$$

Tabla 5

Clasificación de valores observado y probabilidad estimada por el modelo, con punto de corte c=0.5

Valores observados en la variable respuesta Y	Probabilidad estimada del modelo de regresión logística (predicción)		**Total**
	$P(Y=1/X) \geq 0.5$	$P(Y=0/X) < 0.5$	
1=Ocurrió el evento de interés	VP	FN	$VP+FN$
0=No ocurrió el evento de interés	FP	VN	$FP+VN$
Total	$VP+FP$	$FN+VN$	n

VP =Total de predicciones correctas, para cuando $Y=1 \quad y \quad P(Y=1) \geq 0.5$ cada predicción se llama verdadera positiva.

FN =Total de predicciones incorrectas, para cuando $Y=1 \quad y \quad P(Y=1) < 0.5$ cada predicción se llama falsa negativa.

FP =Total de predicciones incorrectas, para cuando $Y=0 \quad y \quad P(Y=1) \geq 0.5$ cada predicción se llama falsa positivas.

VN =Total de predicciones correctas, para cuando $Y=0 \quad y \quad P(Y=1) < 0.5$ cada predicción se llama verdadera negativas.

De la tabla se definen las siguientes medidas:

Acierto total del modelo de regresión logístico binario múltiple estimado.

Precisión (A) de acierto total, del modelo estimado:

$$A = \frac{VP+VN}{VP+FN+FP+VN} = \frac{VP+VN}{n} * 100\%$$

Corresponde al porcentaje de observaciones bien clasificados por el modelo de regresión logística binario múltiple estimado, se espera que este porcentaje sea lo más altos posible, de esta forma se concluye que el modelo estimado clasifica bien a las observaciones. Al valor

$(1-A)*100\%$, corresponde al porcentaje de observaciones de la muestra mal clasificado por el modelo estimado, este porcentaje debe ser mínimo para considerarse adecuado el modelo (error aparente).

Tasa verdadera positiva (TVP) del modelo de regresión logístico estimado (sensibilidad del modelo)

El modelo estimado a través de sus valores pronosticados, clasifica correctamente a lo observado para lo cual y_i toma el valor 1.

$$TVP = \frac{VP}{VP+FN} * 100\%$$

Tasa falsa negativa (TFN) del modelo de regresión logística estimada.

El modelo estimado a través de sus valores pronosticados, clasifica incorrectamente a lo observado para lo cual y_i toma el valor 1.

$$TFN = \frac{FN}{VP+FN}$$

Tasa falsa positiva (TFP) del modelo de regresión logístico estimado

El modelo estimado a través de sus valores pronosticados, clasifica incorrectamente a lo observado para lo cual y_i toma el valor 0.

$$TFP = \frac{FP}{FP+VN}$$

Tasa verdadera negativa (TVN) del modelo de regresión logístico estimado (especificidad del modelo).

El modelo estimado a través de sus valores pronosticados, clasifica correctamente a lo observado para lo cual y_i toma el valor 0

$$TVN = \frac{VN}{FP+VN}$$

Para un modelo estimado de regresión logística binaria múltiple se considera que el modelo es adecuado para clasificar a las observaciones y_i, cuando tiene la máxima sensibilidad y la máxima especificidad.

Nota

$$1 - TVN = 1 - especificidad = TFP(tasa \quad de \quad falsos \quad positivos)$$

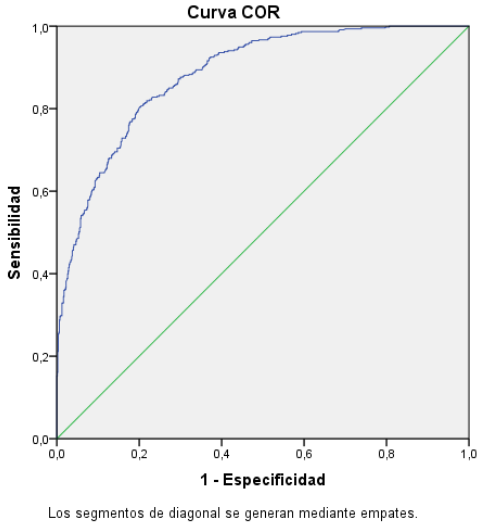
2.3.4.8 La curva ROC

La curva ROC (curva característica de operación) se obtiene al graficar en el eje horizontal (abscisa) la tasa de falsos positivos (1-especificidad) y en el eje vertical (ordenada) la tasa de verdaderos positivos (sensibilidad), y permite determinar un punto de corte entre 0 y 1 de tal forma que los valores de probabilidad pronosticada mayores o igual que el punto de corte, indican que la observación se ubica en $Y_i = 1$; de lo contrario la observación se ubica en $Y_i = 0$.

La siguiente gráfica representa la curva ROC para diferentes puntos de corte de un modelo.

Figura 4

Proporción de falso positivos y verdaderos positivos



Si el modelo logístico binario múltiple estimado es perfecto, hay una región en la que cualquier punto de corte tienen sensibilidad y especificidad igual a 1 y la curva tiene solo el punto (0 , 1).

Si el modelo logístico binario múltiple estimado no ayuda en la clasificación, la tasa de verdadera positiva (sensibilidad) es igual a la tasa de falsos positivos (1-especificidad) y la curva corresponde a la diagonal del cuadrado, cuyos puntos son (0 , 0) a (1 , 1).

El modelo logístico estimado tiene mejor desempeño si la curva ROC se aleja lo más posible de la diagonal del cuadrado (línea de no discriminación). El área bajo la curva es igual a la probabilidad de que la prueba clasifique correctamente a dos observaciones, uno de ellos pertenece a la clase para la cual $Y_i = 1$, y el otro a la clase para la cual $Y_i = 0$. El punto de corte adecuado para la curva ROC, es aquel punto que está bien a la izquierda de la abscisa y lo más alto posible de la ordenada. El área bajo la curva ROC proporciona información del poder discriminante del modelo estimado. De acuerdo a Hosmer-Lemeshow (2013), cuando el área bajo la curva ROC es por lo menos 0.7, el modelo logístico estimado tiene capacitad de discriminación aceptable.

2.3.4.9 Criterio de Información de Akaike.

Este criterio de información de Akaike (AIC), proporciona una medida de la calidad relativa del ajuste de varios modelos logístico, para un mismo conjunto de datos $(y_i, x_{i1}, x_{i2}, ..., x_{ik})$, medida que penaliza al modelo que contiene más variable predictoras. Akaike (1974), para varios modelos ajustados, usando un mismo conjunto de datos, el modelo preferido es el que tiene el valor mínimo en el AIC.

$$AIC = -2LnL(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k) + 2(k+1)$$

donde:

$-2LnL(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k)$ = -2 veces el logaritmo neperiano de la función de verosimilitud del modelo en estudio.

k = N° de predictores del modelo en estudio.

2.3.4.10 Interacción entre las variables predictoras.

Para la construcción del modelo logístico, si existe la presencia de variables predictoras que interactúan simultáneamente en forma

significativas, por ejemplo $(x_i.x_j.x_k)$, deben considerarse en el modelo los órdenes jerárquicos inferiores de las variables, esto es $(x_i.x_j, x_i.x_k, x_j.x_k, x_i, x_j, x_k)$. El programa SPSS, proporciona las posibles interacciones significativas entre ellas. Existe interacción entre dos variables cuando la asociación entre ellas varía, según los diferentes niveles de otra u otras variables.

2.3.5 Adecuación del modelo de regresión logístico binario.

2.3.5.1 Análisis de los residuos e influencia

Mediante el análisis de los residuos del modelo de regresión logística binaria estimada, se detecta observaciones discordantes (son aquellas observaciones que no guardan un patrón de comportamiento con respecto a todos los demás), y si existen observaciones que influyan significativamente en la estimación de los parámetros del modelo, es de suma importancia examinar si algunas observaciones de la muestra están alterando o modificando propiedades del modelo (coeficientes estimados, error estándar, etc.).

Residuos del modelo estimado

Diferencia entre el valor que toma la variable respuesta Y_i observada y el valor estimado por el modelo $\hat{\pi}_i$.

$$e_i = y_i - \hat{\pi}_i \qquad i = 1,2,...,n \tag{45}$$

Siendo

$y_i = 0,1 \quad i = 1,2,...,n$

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+...+\hat{\beta}_j x_{ij}+...+\hat{\beta}_k x_{ik}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+...+\hat{\beta}_j x_{ij}+...+\hat{\beta}_k x_{ik}}}$$

Los residuos más utilizados, para detectar si una observación es discordante o no. Se presenta a continuación:

Residuos estandarizados.

Es el residuo no estandarizado dividido por la estimación de su error estándar

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad i = 1,2,...,n \tag{46}$$

Residuos estudiantizados.

Es el cociente de la diferencia entre la respuesta observada y la respuesta estimada, eliminando la observación i-ésima con su error estándar eliminando la observación i-ésima en la muestra.

$$st_i = \frac{y_i - \hat{\pi}_{(i)}}{\sqrt{\hat{\pi}_{(i)}(1-\hat{\pi}_{(i)})}} \qquad i = 1,2,...,n \tag{47}$$

Donde $\hat{\pi}_{(i)}$, es la estimación de π_i obtenida eliminando la observación i-ésima en la muestra

Un residuo estandarizado o estudiantizados superior a 2 en valor absoluto indica una observación discordante.

Residuos desvianza.

Para cada observación de la muestra, la desvianza se calcula por:

$$d_i = signo(y_i - \hat{\pi}_i)(2\left[y_i \ln\frac{y_i}{\hat{\pi}_i} + (1-y_i)\ln(\frac{1-y_i}{1-\hat{\pi}_i}) \right])^{\frac{1}{2}} \tag{48}$$

Un residuo de desvianza superior a 4 en valor absoluto indica que la observación es discordante

Medidas de influencia

Son medidas de gran utilidad, para detectar la presencia de observaciones que se encuentran lejos de la gran mayoría, influyendo

sobre los parámetros estimados o las predicciones, cuanto más grande sea su valor comparado con los demás, mayor es la influencia que ejerce una observación. Existen fórmulas para detectar la presencia observaciones influyentes.

Medida de apalancamiento (Leverage)

Son valores para identificar observaciones que influyen en las predicciones de la variable respuesta. Se calcula usando la matriz sombrero $\hat{H}$, matriz cuadrada de orden nxn.

$$\hat{H} = \sqrt{V}X(X'VX)^{-1}X'\sqrt{V} \qquad (49)$$

$$V = diag\left[\hat{\pi}_i(1-\hat{\pi}_i)\right] \qquad X = Matriz \quad diseño \quad de \quad orden \quad nx(k+1)$$

El elemento $i-esimo$ de la diagonal principal de la matriz $\hat{H}$ es: $h_{ii} \in <0,1>$, con un valor medio de $(k+1)/n$. Las medidas como la distancia de Cook y Dfbeta miden el impacto que tiene una observación en los coeficientes estimados del modelo.

Distancia de Cook

Mide la influencia de cada observación en la estimación de los parámetros del modelo.

$$Cook_i = \frac{r_i^2 h_{ii}}{1-h_{ii}} \qquad (50),$$

dónde: r_i son residuos estandarizados

Un valor de $Cook_i > 1$, indica que la observación i es influyente en la estimación $\hat{\beta}$

Dfbetas

Mide la Influencia de cada observación en la estimación de cada elemento del vector $\hat{\beta}_j$, donde cada

$$\Delta\hat{\beta}_i = \frac{(X'VX)^{-1}x_i'(y_i-\hat{\pi}_i)}{1-h_{ii}} \qquad (51)$$

Siendo: $x_i' = (1, x_{i1}, x_{i2}, ..., x_{ik})$,

2.3.5.2 Multicolinealidad

Se dice que existe multicolinealidad, entre las variables independientes (predictores) de un modelo logístico binario múltiple, cuando entre estas existe una relación de tipo lineal; es decir una variable (predictor) es combinación lineal de las otras. Cuando la colinealidad o multicolinealidad es perfecta no se puede estimar los coeficientes asociados de las variables independientes y los errores estándar serian infinitos, ya que la matriz $(X'VX)$ es singular (carece de inversa). Para evitar la multicolinealidad entre las variables independientes, tratar de no elegir variables correlacionadas, otra seria eliminar la variable que actúa como dependiente de las otras, teniendo cuidado de que esta variable eliminada no cause sesgo de especificación. Ya que la presencia de alta multicolinealidad entre las variables explicativa trae como consecuencia aceptar la hipótesis nula, cuando esta debería ser rechazada (Gujarati et al. 2005, pp. 167-181).

Las consecuencias de la multicolinealidad son:

- Los errores estándar de los estimadores, tiende a ser mayores a medida que aumenta el grado de colinealidad.
- Los límites de confianza para los parámetros, tienden a ser grande.
- Una de las soluciones para la multicolinealidad entre las x_j es eliminar la variable que están relacionada linealmente con el resto de las variables o considerar nuevas variables independientes.

Factor de Inflación de varianzas (FIV)

Es un valor que nos indica en cuanto se infla (crece) la varianza del coeficiente estimado del modelo logístico y la vuelve imprecisa,

trayendo consigo un modelo inestable (en el sentido de que otra muestra pueden dar resultados diferentes). Se calcula siguiendo los siguientes pasos:

Cálculo

Para obtener los factores de inflación de varianza para cada variable independiente para una regresión logística, se hace uso de la matriz de correlación

$$R = D^{-1/2}.COV(\hat{\beta})D^{-1/2}$$

$D^{-1/2}$ Matriz diagonal, cuyos elementos está conformado por las desviaciones estándar de los coeficientes estimados.

$COV(\hat{\beta})$ Matriz de varianza y covarianzas de los coeficientes estimados.

El valor de FIV_J , varía entre 1 hasta infinito, un valor 1 indica que la variable x_j es independiente del resto de variables; un valor superior a 1 indica que la variable x_j no es independiente del resto de las otras, en este caso el coeficiente asociado a la variable se vuelve inestable ya que la varianza queda inflada en $(1 - FIV_j)100\%$, una regla general es que el valor $FIV_j < 3$ es aceptable; para valores intermedios $3 \leq FIV_j < 10$ se tiene problemas de multicolinealidad no tan severos o débiles; un valor $FIV_j > 10$,implica problemas severos de multicolinealidad. También el programa Minitab, proporcionan los FIV.

2.4 Glosario de términos

- **Colesterol alto**

  El exceso de colesterol en nuestro cuerpo provoca enfermedades en el corazón, se mide a través de una análisis en la sangres en ayunas, y se recomienda para bajar el

colesterol elevado hacer ejercicio y dieta sin grasa (Texas Heart Institute, 2022, p.1).

- **Consumo de alcohol**

  Su consumo indiscriminado conduce a muchas patologías y es menor cuando su consumo es moderado. Se mide por onza liquida (ml) (Texas Heart Institute, 2022, p.4).

- **Consumo de tabaco**

  Se refiere al fumar un cigarrillo, el humo (nicotina) es perjudicial para la salud de las personas, además de otros compuestos químicos como el alquitrán y el monóxido de carbono endurecen las arterias y pueden causar mal funcionamiento del ritmo cardiaco elevando la presión arterial (Texas Heart Institute, 2022, p.3).

- **Diabetes mellitus**

  Las personas que padecen de diabetes, debe ser controlado por un médico, porque el buen control de los niveles de glucosa (azúcar) en sangre puede disminuir su riesgo cardiovascular. Si cree que pueda tener diabetes pero no está seguro, consulte al médico quien le realizará los análisis correspondientes (Texas Heart Institute, 2022, p.2).

- **Inactividad física**

  La falta de ejercicio en muestro cuerpo conlleva a muchas enfermedades, sobre todo sufrir un ataque al corazón, ya que el ejercicio regular de 30 a 45 minutos diarios quema calorías para mantener un peso saludable y permite controlar los niveles de colesterol (Texas Heart Institute, 2022, p.3).

- **Obesidad y sobrepeso**

  Una persona tiene sobrepeso si el índice de masa corporal (IMC) es superior a 25 Kg/m2 y es obesa si es mayor a 30 Kg/m2 (Texas Heart Institute, 2022, p.2).

$$IMC = \frac{PESO(kg)}{(ESTATURA \quad mts)^2}$$

- **Presión arterial sistólica (PAS) y Presión arterial diastólica (PAD)**

  Se mide en miligramos de mercurio (mmHg), es la presión que soportan los vasos sanguíneos en nuestro cuerpo, se dice que una persona tienen hipertensión arterial cuando la presión sistólica es mayor o igual a 14 mmHg y la diastólica mayor o igual a 90 mmHg (OMS y OPS, 2020, p.1).

- **Triglicéridos**

  Los triglicéridos son grasas que provienen de alimentos que se consumen: mantequilla, aceites y otras grasas; los triglicéridos también provienen de calorías adicionales, estas son las calorías que usted come pero que su cuerpo no necesita de inmediato, tener un alto nivel de triglicéridos puede aumentar el riesgo de enfermedades del corazón (MedlinePlus, 2021, pp.1-3).

2.5 Hipótesis del estudio

Hipótesis general

Para cumplir el objetivo general del trabajo de investigación, se plantea la siguiente hipótesis general:

Los factores relacionados a las enfermedades cardiovasculares son determinados mediante la regresión logística binaria múltiple.

Hipótesis especificas

Para cumplir los objetivos específicos del trabajo de investigación, se plantea las siguientes hipótesis específicas:

Hipótesis especifica 1:

Los factores demográficos (edad, género, raza, índice de masa corporal), factor estilo de vida (alcoholismo, tabaquismo, actividad física) y factores clínicos (colesterol, triglicéridos, presión sistólica, presión diastólica, diabetes) están relacionados a la enfermedad cardiovascular, pueden ser identificados mediante pruebas estadísticas bivariadas.

Hipótesis especifica 2:

Los factores demográficos (edad, género, raza, índice de masa corporal), factor estilo de vida (alcoholismo, tabaquismo, actividad física) y factores clínicos (colesterol, triglicéridos, presión sistólica, presión diastólica, diabetes), están relacionados a la enfermedad cardiovasculares, son evaluados utilizando la regresión logística binaria múltiple.

CAPITULO 3: METODOLOGIA

3.1 Identificación de variables

3.1.1 Variable dependiente(respuesta)

Enfermedad cardiovascular

3.1.2 Variables independientes (predictores)

A. Género
B. Edad
C. Peso
D. Estatura
E. Raza
F. Alcoholismo
G. Tabaquismo
H. Actividad física
I. Nivel de colesterol
J. Nivel triglicéridos
K. Presión arterial sistólica
L. Presión arterial diastólica
M. Diabetes mellitus

3.2 Tipo y diseño de la investigación.

El presente trabajo de investigación es tipo cuantitativo y el diseño es observacional, correlacional, no experimental, retrospectivo y de corte transversal.

3.3 Población

Todas las personas adultas (policías en actividad, retiro y familiares directos) de 20 años a más, atendidas por consultas externas, en el Hospital PNP Augusto B. Leguía-Rímac. Lima-Perú. Se excluye a menores de 20 años.

3.4 Unidad

Cada persona adulta (policías en actividad, retiro y familiares directos) de 20 años a más, atendidas por consultas externas, en el Hospital PNP Augusto B. Leguía-Rímac. Lima-Perú.

3.5 Muestra seleccionada

Para el presente trabajo de investigación, de la población en estudio, se eligió una muestra al azar que permita estimar la proporción o porcentaje de pacientes con la enfermedad cardiovascular con un margen de error de 2.641% y un nivel de confianza del 95% (significa que se elige una muestra que estime la proporción de paciente con la enfermedad que no se aleje de la proporción de la poblacion en valor absoluto en menor o igual a 0.02641 con una confianza del 95%), por información referencial se sabe que la proporción estimada de adultos atendidos que padecen la enfermedad cardiovascular es $\hat{\pi} = 0.37$,para calcular el tamaño de muestra para un muestreo simple aleatorio se usa la siguiente formula en la cual se asume una poblacion muy grande de atenciones:

$$n = \frac{z^2_{(0.025)}\hat{\pi}(1-\hat{\pi})}{e^2}$$

dónde:

$e =$ Margen de error aceptado por el investigador 2.641% (0.02641), permitió obtener una muestra grande y estimar los parámetros del modelo logístico, con más precisión.

$z =$ Valor de abscisa de la distribución normal estándar con una probabilidad acumulada del 2.5% (0.025) =1.96

"$\hat{\pi}$" =Proporción estimada de personas adultas en la población con problemas cardiovasculares. De trabajos pasados se sabe que el porcentaje de personas adultas atendidas por consultas es 37% (0.37). Se asume que el tamaño de la población (concurrencia de paciente) se asume grande.

$$n = \frac{1.96^2(0.37)(1-0.37)}{(0.02641)^2} = 1284 \quad pacientes$$

El tamaño de la muestra fue de 1284 pacientes.

3.6 Diseño muestral

La selección de la muestra se realizó a través de un muestreo aleatorio simple de 1284 pacientes e historias clínicas. El marco de muestreo estuvo conformado por todos los pacientes, atendidos por consultas con edades de 20 años a más, la fecha de recojo de datos correspondió al período de Octubre a Diciembre 2017 en el Hospital PNP Augusto B Leguía. Lima-Perú. La selección de la muestra se realizó con un nivel de confianza del 95% y un margen de error del 2.641% (el margen de error pequeño permite tener una muestra grande, ya que el modelo logístico funciona bien para muestra grande).

3.7 Técnica de recolección de datos

El presente trabajo de investigación, los datos provienen de fuente primarias (historias clínicas de pacientes con edades de 20 años a más que son atendidas por consultas en Hospital PNP Augusto B Leguía distrito del Rímac. Lima-Perú). El periodo de la recolección de los datos se realizó en los meses de octubre a diciembre del 2017, a través de la revisión de historias clínicas.

3.8 Análisis e interpretación de la información

A. Los datos recopilados, fueron organizados y presentados en tablas de distribución de frecuencias y gráficos.

B. Se exploró los datos mediante, medidas de tendencia central y dispersión, según sea el caso (variables cualitativas y cuantitativas), así como también asociaciones bivariadas (coeficientes de correlación lineal entre los predictores y asociaciones entre cada predictor y la variable en estudio través de las pruebas correspondientes)
C. El estudio de la relación de los predictores con la enfermedad cardiovascular, se realiza haciendo uso del modelo de regresión logístico binaria múltiple, con el objetivo de determinar qué predictores están relacionados o influyen en la enfermedad cardiovascular.

Todos estos, se llevó a cabo usando el programa estadístico SPSS, Minitab, R-Proyect, Excel.

3.9 Definición de variables

Tabla 06

Definición de variables en el estudio

Nombre	Definición	Unidad o categoría
Enfermedad cardiovascular	Es una enfermedad que afecta el corazón o los vasos sanguíneos.	0=no tiene la enfermedad cardiovascular 1=tiene la enfermedad cardiovascular
Factores demográficos		
Genero	Cualidad	0=mujer 1=varón
Edad	Tiempo	Años
Raza	Cualidad	0=blanca 1=negra 2=otras
Índice de masa corporal (IMC)	Es una razón matemática que asocia la masa corporal y la talla de un individuo, se obtiene al dividir el peso en kilogramos por el cuadrado de la estatura en metros (kg/m2), índice utilizado frecuentemente para clasificar el sobrepeso y la obesidad en adultos.	Kg/m^2 <18.5 bajo peso 18.5-24.9 peso normal 25.0-29.9 pre-obesidad o sobre peso 30.0-34.9 obesidad clase I 35.0-39.9 obesidad clase II >40 obesidad clase III
Factores estilo de vida		
Alcoholismo	Si la persona ha declarado tomar una o más veces a la semana al momento de la entrevista.	0=no toma 1=toma
Tabaquismo	Si la persona ha declarado fumar uno a más cigarros al día al momento de la entrevista.	0=no fuma 1=fuma
Actividad física	Si la personas durante la última semana han practicado deporte o realizada actividad física fuera del horario de trabajo, durante 30 minutos o más cada vez, al menos 3 veces a la semana.	0=hace deporte 1=no hace deporte
Factores clínicos		
Colesterol	Sustancia cerosa que se encuentra en la sangre. El cuerpo necesita de esta sustancia pero tener niveles altos puede aumentar el riesgo de sufrir una enfermedad cardíaca. Los resultados del colesterol total se expresan en miligramos por decilitro (mg/dl) de sangre.	0=Normal: menos de 200 mg/dl 1=Normal-alto: entre 200 y 240 mg/dl. Se considera hipercolesterolemia a los niveles de colesterol total superiores a 200 mg/dl. 2=Alto: por encima de 240 mg/dl
Triglicéridos	Un tipo de grasa que se encuentra en la sangre.. El cuerpo los usa para obtener energía. Se necesitan algunos triglicéridos para tener una buena salud. Sin embargo, un nivel alto de triglicéridos podría aumentar el riesgo de enfermedad cardíaca y puede ser una señal del síndrome metabólico. Los resultados del triglicérido total se expresan en miligramos por decilitro (mg/dl) de sangre.	0=normal (triglicéridos menores a 150 mg/dl 1=moderado (triglicéridos entre 150-200 mg/dl) 2=muy alto (triglicéridos superiores a 200 mg/dl)
Presión arterial sistólica (PAS)	Corresponde al valor máximo de la presión arterial en sístole (cuando el corazón se contrae). La presión arterial se mide en mmHg. 1mmHg es la presión ejercida por 1 milímetro (mm) de mercurio (Hg).	90 a 130 mmHg corresponde a una presión sistólica normal, menor a 90 mmHg corresponde a una presión sistólica baja. Presión arterial sistólica superior a 140 mmHg se considera alta y peligrosa para nuestra salud.

Presión arterial diastólica (PAD)	Corresponde al valor mínimo de la presión arterial cuando el corazón está en diástole o entre latidos cardíacos. La presión arterial se mide en mmHg. 1 mmHg es la presión ejercida por 1 milímetro (mm) de mercurio (Hg).	60 a 80 mmHg corresponde a una presión diastólica normal Menor a 60 mmHg corresponde a una presión diastólica baja. Presión arterial diastólica superior a 90 mmHg se considera alta y peligrosa para nuestra salud.
Diabetes	Enfermedad crónica que afecta la forma en que el cuerpo convierte los alimentos en energía.	0=no tiene 1=tiene

Fuente: OMS y artículos de medicina-2020

CAPITULO 4: RESULTADOS Y DISCUSION

4.1 Resultados

4.1.1 Análisis descriptivo univariado

El presente estudio se realizó con una muestra elegida al azar de 1284 pacientes adultos de 20 años a más, atendidos por consultas externas en el Hospital PNP Augusto B. Leguía, ubicado en el distrito del Rímac, entre los meses de octubre a diciembre del año 2017. El objetivo de la investigación es determinar factores de riesgos relacionados a las enfermedades cardiovasculares, a través de un modelo de regresión logístico binario múltiple, siendo la variable dependiente la enfermedad cardiovascular y las variables independientes (explicativas o predictores) 12 elegidos por antecedentes de otros trabajos de investigación: género, raza, edad, índice de masa corporal (IMC), alcoholismo, tabaquismo, actividad física, colesterol, triglicéridos, diabetes, presión arterial sistólica (PAS), presión arterial diastólica (PAD). En un primer intento de conocer el comportamiento de las variables se realiza el análisis descriptivo univariado y bivariado; con el objetivo de conocer la distribución de los datos, las medidas de tendencia central, dispersión y la relación (asociación) de los predictores con la enfermedad cardiovascular. A estos predictores para analizarlo exhaustivamente, se clasificó en : Factor demográfico (género, edad, raza, índice de masa corporal); factor estilo de vida (alcoholismo, tabaquismo, actividad física) y factor clínico (colesterol, triglicéridos, presión arterial sistólica (PAS), presión arterial diastólica (PAD) y diabetes.

Factor demográfico

De una muestra elegida al azar de 1284 historias clínicas de pacientes atendidos por consultas externas: 715(55.69%) son varones; el 1029(80.14%) son de otras razas (*); 314(24.45%) pertenecen al grupo etario que tienen edades entre 60 a 70 años; 691(53.82%) tienen índice

de masa corporal (IMC) entre 25 a 30 Kg/m2 (obesidad) y 477(37.15%) tienen la enfermedad cardiovascular (tabla 7).

Tabla 7

Factor demográfico de pacientes atendidos por consultas externas, hospital PNP Augusto B Leguía, Octubre a Diciembre 2017(n=1284)

Factores demográficos	Número de pacientes	Porcentajes
Género		
Mujer	569	44.31%
Varón	715	55.69%
Raza		
Blanca	125	9.74%
Negra	130	10.12%
Otras*	1029	80.14%
Edad(años)		
20-30(0)	53	4.13%
30-40(1)	132	10.28%
40-50(2)	231	17.99%
50-60(3)	288	22.43%
60-70(4)	314	24.45%
70-80(5)	190	14.80%
80-90(6)	71	5.53%
90-100(7)	5	0.39%
(IMC) (kg/m^2)		
=<25 normal	381	29.67%
25-30 obesidad leve	691	53.82%
30-35 obesidad tipo 1	180	14.02%
35-40 obesidad tipo 2	30	2.34%
>=40 obes. 3 mórbida	2	0.15%
Enfermedad cardiovasc		
Si	477	37.15%
No	807	62.85%

Nota. (*) Mestizo, quechua, Aymara, Nativo, Shipibo, Nikkei, etc.

Con respecto a la edad del paciente: La mínima es de 21 años y la máxima 93 años, la edad promedio 57.45 años con una desviación estándar de 14.93 años; para el índice de masa corporal (IMC): valor mínimo 19.57 kg/m^2 y valor máximo 46.48 kg/m^2, su promedio 26.85 kg/m^2 con desviación estándar de 3.19 kg/m^2 (tabla 10).

Factor estilo de vida

De los 1284 pacientes atendidos por consultas: 770(60.0%) consumen algún tipo de bebida alcohólica; 329(25.6%) fuman y 955(60.8%) no realizan actividad física (tabla 8).

Tabla 8

Factor estilo de vida de pacientes atendidos por consultas externas, hospital PNP Augusto B Leguía, Octubre a Diciembre 2017(n=1284)

Factor estilo de vida	Número de pacientes	Porcentajes
Alcoholismo		
No toma	514	40.03%
Si toma	770	59.97%
Tabaquismo		
No fuma	955	74.38%
Si fuma	329	25.62%
Actividad física		
No practica deporte	781	60.83%
Si practica deporte	503	39.17%

Factor clínico

De los 1284 pacientes atendidos por consultas externas: 952(74.14%), tienen nivel de colesterol normal; 880(68.54%), nivel triglicéridos normal; 1066(83.02%) presión arterial sistólica normal; 1261(98.21%) presión arterial diastólica normal y 842(65.58%) no tienen diabetes (tabla 9).

Tabla 9

Factor clínico en pacientes atendidos por consultas externas, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)

Factor clínico	Número de pacientes	Porcentajes
Colesterol(mg/dl)		
<200 normal	952	74.14%
200-240 normal-alto	110	8.57%
>=240 alto	222	17.29%
Triglicéridos(mg/dl)		
<150 normal	880	68.54%
150-200 moderado	171	13.32%
>=200 alto	233	18.14%
Presión arterial sistólica (PAS)mmHg		
<140 normal	1066	83.02%
>=140 alto	218	16.98%
Presión arterial diastólica (PAD)mmHg		
<90 normal	1261	98.21%
>=90 alto	23	1.79%
Diabetes		
No tiene	842	65.58%
Tiene	442	34.42%

Para el nivel de colesterol del paciente: El nivel mínimo es 114.00 mg/dl y el máximo de 482.00 mg/dl, el nivel promedio es de 189.67 mg/dl con una desviación estándar de 49.59 mg/dl; para el nivel de triglicéridos: su valor mínimo es 48 mg/dl y su máximo 482.00 mg/dl, con un nivel promedio de 140.49 mg/dl y una desviación estándar de 77.79 mg/dl; Para la presión arterial sistólica (PAS): El valor mínimo 80 mmHg y su máximo 190 mmHg, con una presión promedio de 115.65 mmHg y desviación estándar de 22 mmHg y para la presión arterial diastólica (PAD): La presión minina 40 mmHg y la máxima 100 mmHg, con una presión promedio de 64.99 mmHg y desviación estándar de 6.98 mmHg; el índice de masa corporal mínimo es 19.57 kg/m^2 y su máximo 46.48 kg/m^2 , con un índice de masa corporal promedio de 26.85 kg/m^2 y desviación estándar de 3.19 kg/m^2 (tabla 10).

Tabla 10

Estadísticas de variables cuantitativas relacionadas a las enfermedades cardiovasculares, Hospital PNP Augusto B. Leguía, Octubre-Diciembre 2017 (n=1284)

Variables predictores	Mínimo	Máximo	Media	Mediana	Moda	Desviación estándar
Edad(años)	21	93	57.45	58	64	14.93
Colesterol(mg/dl)	114	480	189.67	179	185	49.59
Triglicéridos(mg/dl)	48	482	140.49	103	93	77.79
PAS(mmHg)	80	190	115.65	110	100	22
PAD(mmHg)	40	100	64.99	60	60	6.98
IMC(kg/m^2)	19.57	46.48	26.85	26.32	25	3.19

VARIABLES	GENERO											
	MUJER						VARON					
	Media	Recuento	Máximo	Mínimo	Desviación estándar	Cv	Media	Recuento	Máximo	Mínimo	Desviación estándar	cv
Edad(años)	55.46	569	93.00	21.00	15.78	28.46%	59.03	715	91.00	22.00	14.03	23.77%
Colesterol(mg/dl)	186.60	569	480.00	114.00	46.98	25.18%	192.12	715	420.00	114.00	51.46	26.79%
Triglicéridos(mg/dl)	134.99	569	482.00	48.00	74.33	55.07%	144.88	715	435.00	50.00	80.22	55.37%
IMC(kg/m^2)	26.53	569	40.46	19.65	3.26	12.28%	27.11	715	46.48	19.57	3.12	11.51%
PAS(mmHg)	110.73	569	190.00	80.00	21.11	19.06%	119.56	715	190.00	90.00	21.92	18.34%
PAD(mmHg)	63.93	569	90.00	40.00	6.81	10.65%	65.83	715	100.00	50.00	7.00	10.64%

4.1.2 Análisis bivariado

En el análisis bivariado se estudia la relación de independencia o dependencia entre la enfermedad cardiovascular (variable de estudios) y las variables predictoras, las correlaciones lineales entre las variables predictoras cuantitativas y otras asociaciones que se detallan:

a. Relación entre la enfermedad cardiovascular con posibles variables predictores de riesgo, utilizando la prueba estadística de Chi-cuadrado para un nivel de significación del 5%, con el objetivo de identificar qué variables que se relacionan con la enfermedad cardiovascular.

b. Relación entre los predictores cuantitativos, utilizando la correlación lineal de Pearson, con el objetivo de identificar independencia o dependencia, en caso de independencia entre los predictores favorece en la estimación del modelo logístico binario y todo lo contrario ocurre cuando se presenta alta correlación entre los predictores.

c. Identificar variables predictoras, que son confusoras (variables que alteran el verdadero efecto de asociación de otro factor con la enfermedad cardiovascular), se utiliza la prueba de Mantel Haenszel para identificarlos.

Asociación entre la enfermedad cardiovascular y las variables predictoras

Al igual que en el análisis univariado, también se analizó relaciones de la enfermedad con los predictores clasificados en factores: demográfico, estilo de vida y clínico:

Factor demográfico

Está conformado por los predictores: género, edad, raza, índice de masa corporal (IMC=peso/estatura^2), utilizando la prueba de independencia Chi-cuadrado con un nivel se significación del 5%, se concluye que los predictores: Género (varón, mujer); raza (blanca, negra y otras); edad (grupos etarios) y IMC

(niveles), están relacionados a la enfermedad cardiovascular. Para las tablas de 2 por 2 se calculó los Odds ratio: para el género, los varones tienen 1.741 veces mayor posibilidad que las mujeres de tener la enfermedad cardiovascular; para la edad, el grupo de pacientes con edades mayor o igual a los 60 años tienen 2.765 veces mayor posibilidad de tener la enfermedad cardiovascular que aquellos con edades menores a los 60 años; con respecto al índice de masa corporal IMC) los pacientes con IMC mayor o igual a 25 kg/m^2, tienen 1.894 veces mayor posibilidad de tener la enfermedad cardiovascular que aquellos con IMC menor a 25 Kg/m^2 (tabla 11 y algunas figuras 5,6,7,8 y 9).

Tabla 11

Factor demográfico y las enfermedades cardiovasculares, en pacientes atendidos por consultas externas, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)

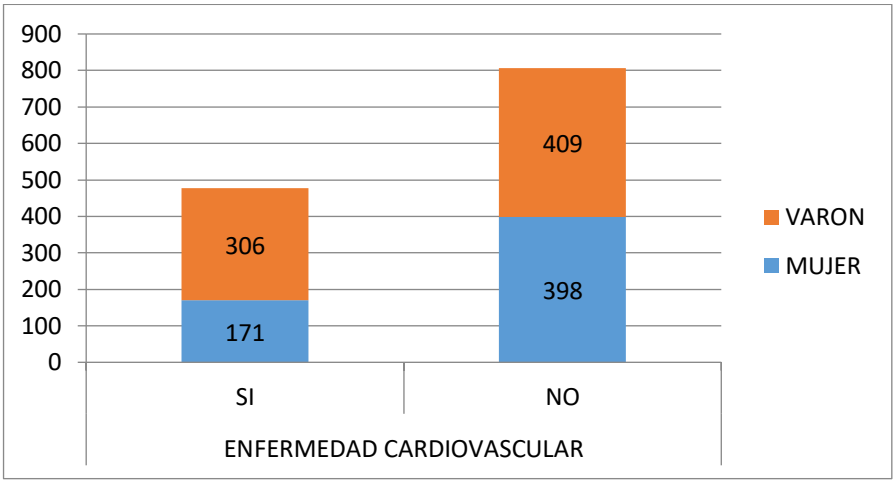
Factor demográfico	Enfermedad cardiovascular				Prueba de Chi cuadrado	Riesgo(ventaja)	
	Tiene		No tiene			OR	IC(95%)
	n	%	n	%			
Género							
Varón(y=1)	306	42.8	409	57.2	0.000(**)	1.741	1.380-2.197
Mujer(y=0)	171	30.1	398	69.9			
Razas							
Blanca	33	26.4	92	73.6	0.007(**)		
Negra	59	45.4	71	54.6			
Otras(1)	385	37.4	644	62.6			
Edad(años)							
20-30	4	7.5	49	92.5	0.000(**)		
30-40	16	12.1	116	87.9			
40-50	57	24.7	174	75.3			
50-60	110	38.2	178	61.8			
60-70	145	46.2	169	53.8			
70-80	102	53.7	88	46.3			
>=80	43	56.6	33	43.4			
Edad>= 60 años	290	50.0	290	50.0	0.000(**)	2.765	2.189-3.491
Edad<60 años	187	26.6	517	73.4			
IMC(kg/m^2)							
<25 normal	114	27.5	301	72.5	0.000(**)		
25-30 obesidad leve	246	37.4	412	62.6			
30-35 obesidad tipo1	98	54.7	81	45.3			
>=35 obesidad tipo 1 y 2	19	59.4	13	40.6			
IMC>=25 kg/m^2	363	41.8	506	58.2	0.000(**)	1.894	1.469-2.442
IMC<25 kg/m^2	114	27.5	301	72.5			

(**) significación 1%

(1) Mestizo, quechua, aimara, nativo, indígena, shipibo, Nikkei, etc.

Figura 5

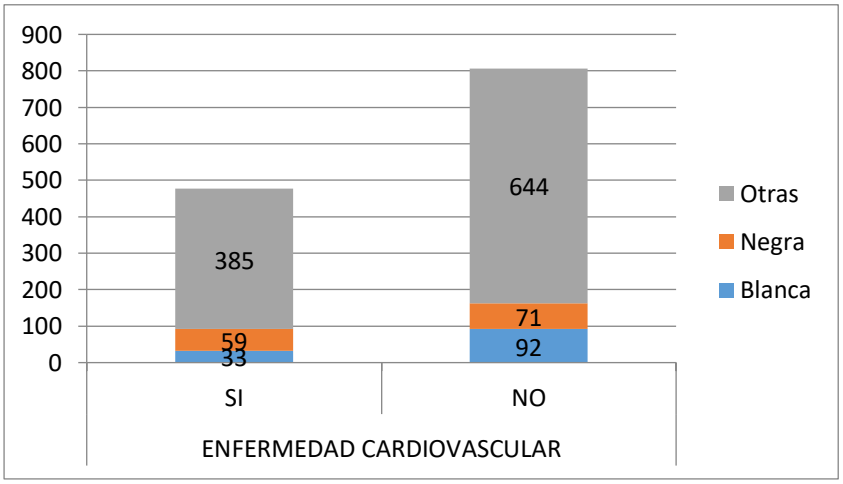
Género de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa que los pacientes con la enfermedad cardiovascular, en varones (306) es 1.79 veces mayor a las mujeres (171).

Figura 6

Razas de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)

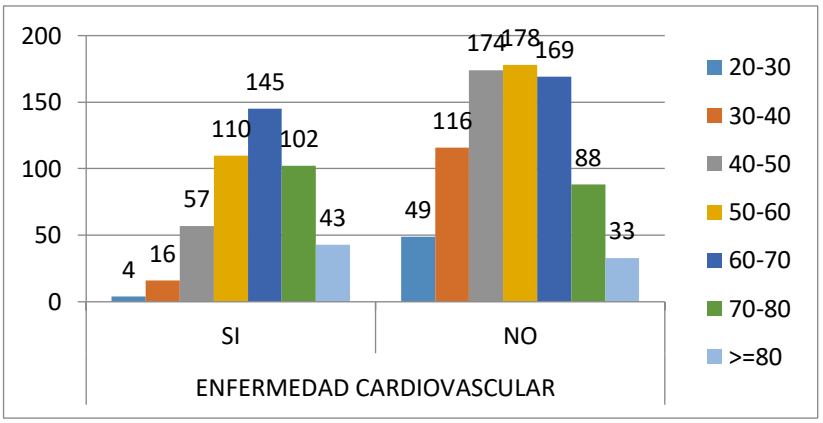


Nota. Otras razas (Mestizo, quechua, aimara, nativo, indígena, shipibo, Nikkei, etc.)

Se observa que con respecto a la raza de los pacientes con enfermedad cardiovascular, la raza negra es 1.8 veces mayor a la raza blanca y otras razas es aproximadamente: 6.5 veces mayor a la raza negra y 11.7 veces mayor a la raza blanca.

Figura 7

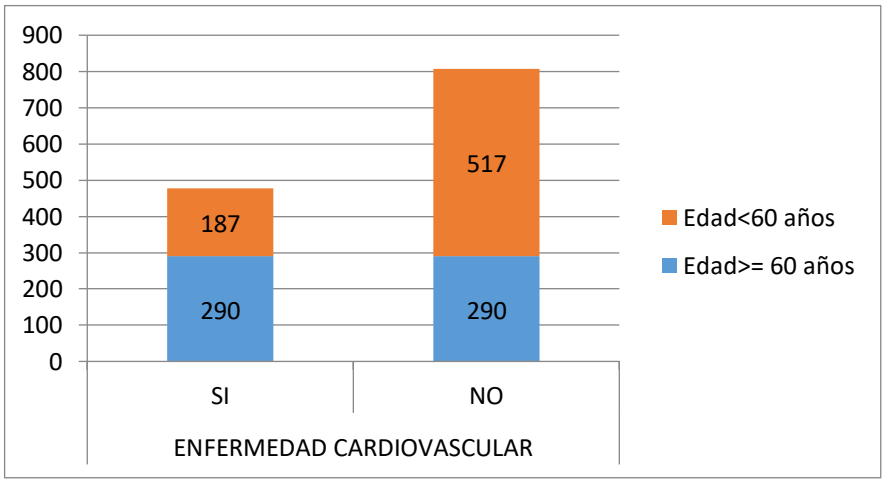
Edades de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa que con respecto a las edades (grupos etarios): los pacientes con enfermedad cardiovascular, se presenta con mayor frecuencia en el grupo etario de 60 a 70 años.

Figura 8

Grupos etarios de pacientes con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)

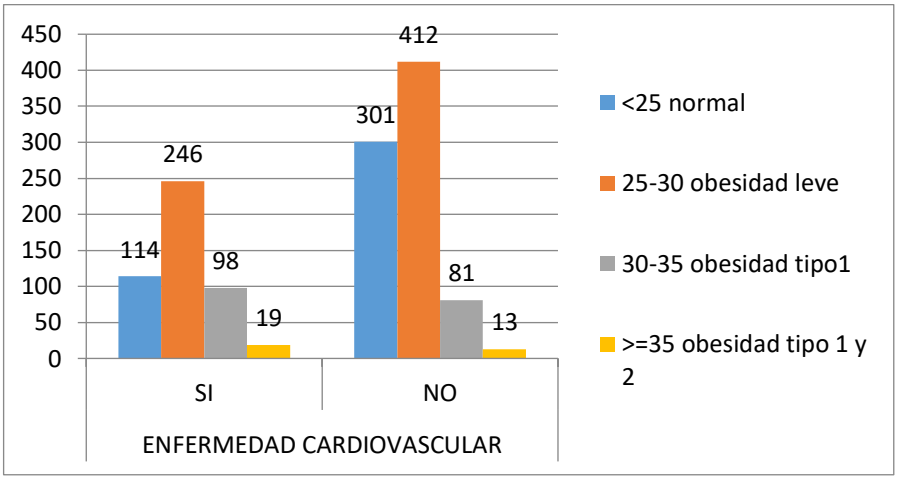


Se observa que con respecto a la cantidad de pacientes con enfermedad cardiovascular, el grupo de pacientes con edades

mayor o igual a los 60 años es 1.56 veces mayor al grupo de pacientes con edades inferior a los 60 años.

Figura 9

Índice de masa corporal (IMC) con o sin la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa que la cantidad de pacientes con un índice de masa corporal superior a los 30 kg/m^2., es 1.24 veces mayor para los que tienen la enfermedad cardiovascular (117) en comparación para los que no la tienen (94).

Factor estilo de vida

Está conformado por los predictores: Alcoholismo, tabaquismo, actividad física, utilizando la prueba de independencia Chi-cuadrado, se concluye que existe una asociación estadísticamente significativa al 1% para las variables alcoholismo y actividad física con la enfermedad cardiovascular, no resultando la prueba significativa para el tabaquismo. Para las tablas de doble entrada se calcula los Odds ratio, resultando el alcoholismo protector para la enfermedad cardiovascular, los pacientes que no toman algún tipo de bebidas alcohólica tienen

1.8 veces mayor posibilidad de tener la enfermedad cardiovascular, en comparación de aquellos que si toman en forma moderada; con respecto a la actividad física resultó de riesgo para la enfermedad, es decir, las personas que no hacer algún tipo de deporte tienen 2.561 veces mayor posibilidad de tener la enfermedad cardiovascular que aquellos que si lo practican (tabla 12).

Tabla 12

Factor de estilo de vida y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía-Rímac, Octubre a Diciembre 2017 (n=1284)

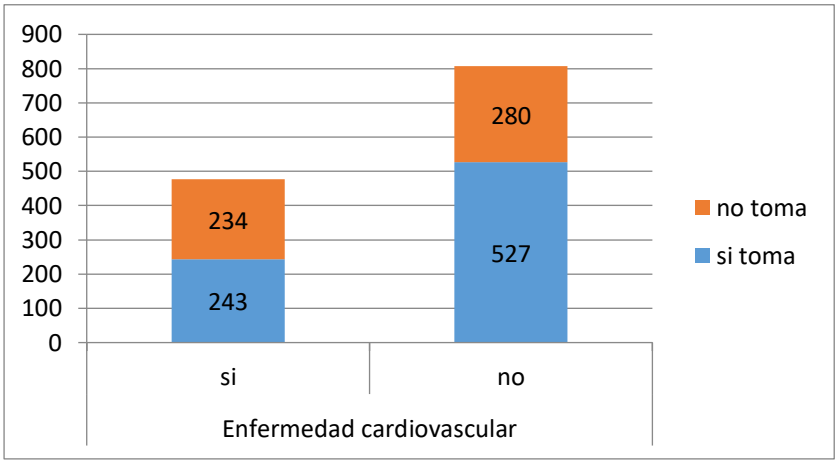
Factor estilo de vida	Enfermedad cardiovascular				Prueba de Chi cuadrado	Riesgo(ventaja)	
	Tiene		No tiene			OR	Intervalo de confianza 95% o 99%
	n	%	n	%			
Alcoholismo							
Si toma	243	31.6	527	68.4	0.000(**)	0.552	0.438-0.695
No toma	234	45.5	280	54.5			
Tabaquismo							
Si fuma	105	31.9	224	68.1	0.023(*)	0.735	0.563-0.958
No fuma	372	39	583	61.0			
Actividad física							
No deporte	354	45.3	427	54.7	0.000(**)	2.561	2.000-3.280
Si deporte	123	24.5	380	75.5			

(**) significación 1%

(*) significativo 5%

Figura 10

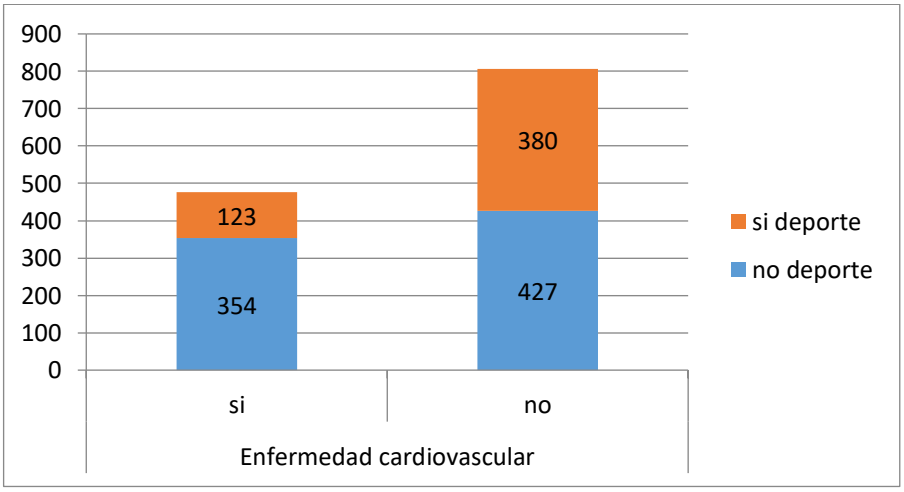
Alcoholismo y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa que la cantidad de pacientes que no tienen la enfermedad y toman, es 2.25 veces mayor con respecto a los que no toman y están enfermo. Se puede inferir a priori para esta muestra de paciente de que el alcoholismo no es de riesgo para la enfermedad cardiovascular.

Figura 11

Actividad física y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa que la cantidad de pacientes que no tienen la enfermedad y hacen deporte (380), es 1.07 veces mayor a la cantidad de pacientes con la enfermedad y no hacen deporte (354).

Factor clínico

Está conformado por los predictores: Colesterol, triglicéridos, diabetes, presión arterial sistólica (PAS) y presión arterial diastólica (PAD), utilizando la prueba de independencia Chi-cuadrado para un nivel de significación del 1%, se concluye que los predictores mencionados están relacionados a la enfermedad cardiovascular. Además, para las tablas de 2 por 2 se calcula los Odds ratio (Riesgos) se tiene:

Los pacientes con un nivel de colesterol mayor o igual a 200 mg/dl, tienen un riesgo de 4.497(CI:3.452; 5.859) veces más posibilidad de padecer la enfermedad cardiovascular que aquellos con nivel de colesterol menor a 200 mg/dl.

Los pacientes con un nivel de triglicéridos mayor o igual a 150 mg/dl, tienen 6.327 (LC: 4.887; 8.191) veces más posibilidad de padecer la enfermedad cardiovascular que aquellos con nivel de triglicéridos menor a 150 mg/dl.

Los pacientes con diabetes tienen 1.978 (LC: 1.561-2.506) veces más posibilidad de padecer la enfermedad cardiovascular que aquellos que no tienen diabetes.

Los pacientes con una presión arterial diastólica superior a 140 mmHg tienen 46.0 (LC: 25.845-81.988) veces más posibilidad de padecer la enfermedad cardiovascular que aquellos con nivel inferior a 140 mmHg (tabla 13).

Tabla 13

Factor clínico y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)

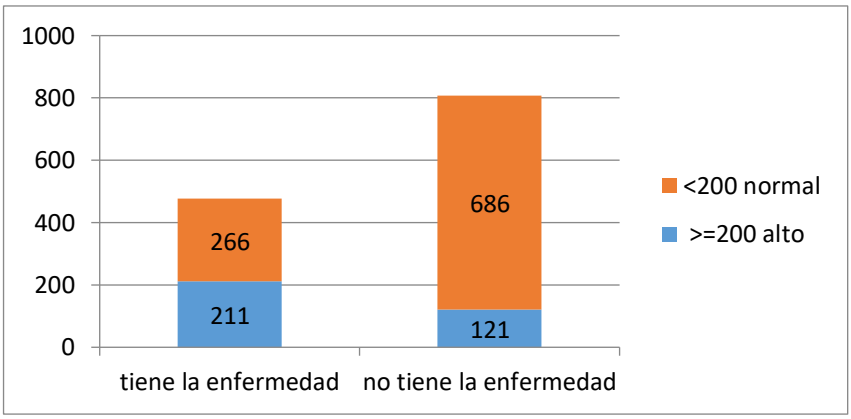
Factor clínico	Enfermedad cardiovascular				Prueba de Chi-cuadrado p-valor	Riesgo(ventaja)	
	Tiene		No tiene			OR	IC (95%)
	n	%	N	%			
Colesterol (mg/dl)							
>=200 alto	211	63.6	121	36.4	0.000(**)	4.497	3.452-5.859
<200 normal	266	27.9	686	72.1			
Triglicéridos (mg/dl)							
>=150 alto	268	66.3	136	33.7	0.000(**)	6.327	4.887-8.191
<150 normal	209	23.8	671	76.2			
Diabetes							
Tiene	211	47.7	231	52.3	0.000(**)	1.978	1.561-2.506
No tiene	266	31.6	576	68.4			
PAS(mmHg)							
>=140 alto	205	94	13	6	0.000(**)	46.032	25.845-81.988
<140 normal	272	25.5	794	74.5			
PAD(mmHg)							
>=90 alto	23	100	0	0	0.000(**)	nsc	nsc
<90 normal	454	36	807	64			

(**) Significación al 1%

(nsc) no se puede calcular

Figura 12

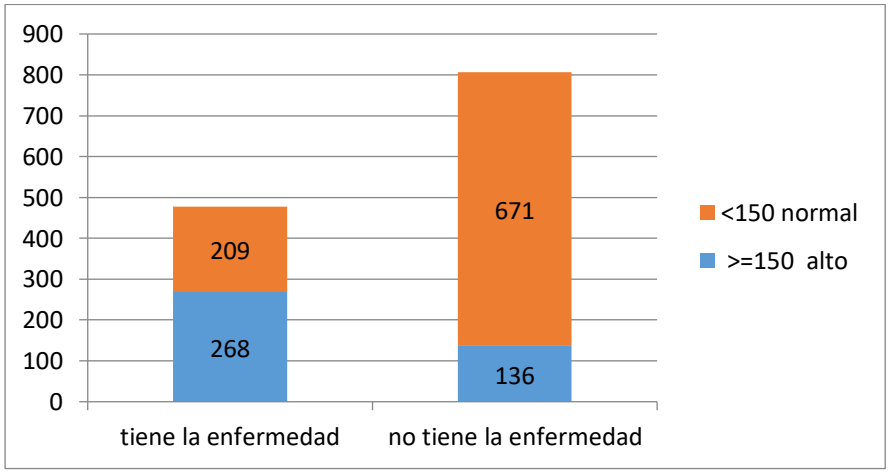
Nivel de colesterol y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



De la figura, se observa que la cantidad de pacientes que tienen un nivel de colesterol superior a 200 mg/dl, es 1.74 veces mayor para los que tienen la enfermedad (211) comparado para los que no la tienen (121).

Figura 13

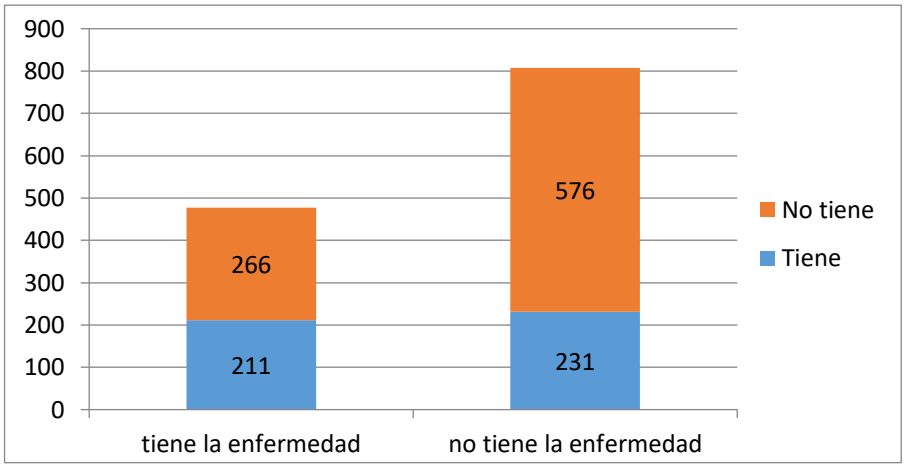
Nivel de triglicéridos y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa que la cantidad de pacientes que tienen un nivel de triglicéridos superior a 150 mg/dl, es 1.97 veces mayor para los que tienen la enfermedad (268) comparado para los que no la tienen (136).

Figura 14

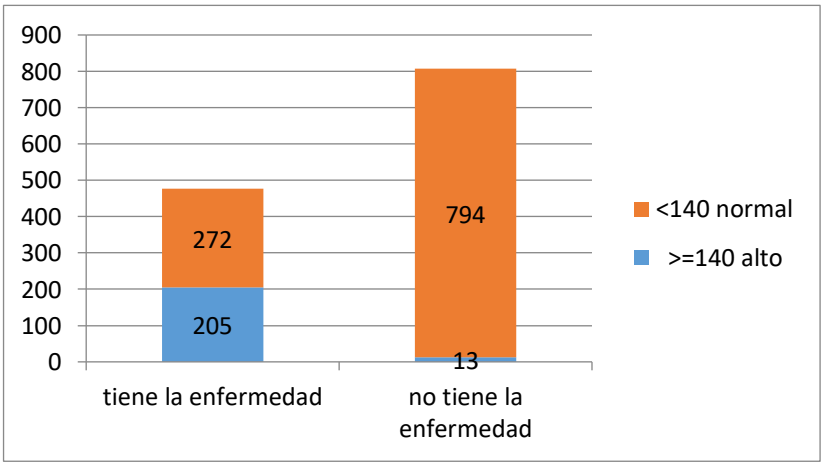
La diabetes en pacientes y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



De la figura, se observa que la cantidad de pacientes que tienen diabetes, son 1.1 veces mayor para los que no tienen la enfermedad cardiovascular comparado con lo que si tienen la enfermedad.

Figura 15

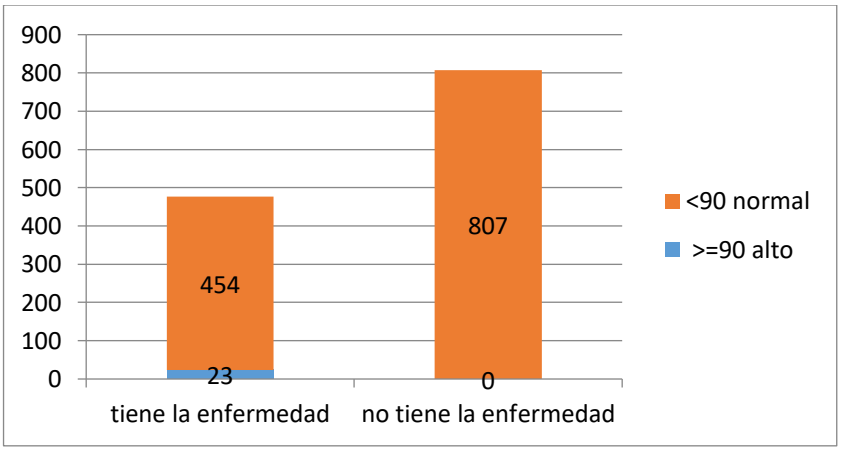
Presión arterial sistólica y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



De la figura, se observa que la cantidad de pacientes que tienen un nivel de presión sistólica superior a 140 mmHg, es 15.8 veces mayor para los que tienen la enfermedad comparada para los que no la tienen.

Figura 16

Presión arterial diastólica y la enfermedad cardiovascular, hospital PNP Augusto B. Leguía, Octubre a Diciembre 2017 (n=1284)



Se observa para los pacientes con nivel de presión diastólica por debajo de 90 mmHg es 1.8 veces mayor ´para los que no tienen la enfermedad cardiovascular, en comparación para los que sí tienen la enfermedad.

Relación lineal de las variables predictoras.

Las correlaciones lineales de las variables: edad, nivel colesterol, nivel triglicéridos, presión arterial sistólica (PAS), presión arterial diastólica (PAD) e índice de masa corporal (IMC), los cuales hacen un total de15 coeficientes de correlación, solo uno de ellos supera el 0.5 (coeficiente de correlación lineal entre la presión arterial sistólica y la presión arterial diastólica 0.649), se puede decir que de la totalidad de

correlaciones no están asociadas linealmente (tabla 14 y Anexo 6).

Identificación de variables confusoras

Una variable confusora es aquella que su presencia altera o modifica el efecto de otra variable que se encuentra asociada a la variable en estudio; para el presente trabajo de investigación se identificó al alcoholismo como variable confusora entre el tabaquismo y la enfermedad cardiovascular con un OR ajustado 0.962 (LC: 0.717-1.289), prueba estadística de (Mantel y Haenszel, 1959) resultando no significativo la prueba (anexo 04).

Tabla 14

Correlaciones entre las variables predictoras cuantitativas

Variables predictoras	Edad	Colesterol	Triglicéridos	PAS	PAD	IMC
Edad	1	0.14	0.124	0.301	0.227	0.152
Colesterol	0.14	1	0.391	0.251	0.196	0.253
Triglicérido	0.124	0.391	1	0.353	0.259	0.316
PAS	0.301	0.251	0.353	1	0.649	0.13
PAD	0.227	0.116	0.259	0.649	1	0.095
IMC	0.152	0.253	0.316	0.13	0.095	1

4.1.3 Modelo de regresión logístico binario múltiple.

Con el objetivo de determinar factores (predictores) relacionados a las enfermedades cardiovasculares en personas adultas de 20 años a más, se procedió a ajustar el modelo de regresión logístico binario múltiple:

$$P(Y=1/x_1;x_2,...x_{12})=\frac{e^{\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+...+\beta_{12}x_{12i}}}{1+e^{\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+...+\beta_{12}x_{12i}}}$$

Siendo:

$P(Y=1/x_1;x_2,...x_{12})=$ La probabilidad de que un paciente tenga la enfermedad cardiovascular $(y=1)$ dado un conjunto de predictores $x_1,x_2,...x_{12}$

La variable dependiente:

$Y=1$, paciente con la enfermedad cardiovascular

$Y=0$, paciente sin la enfermedad cardiovascular

Las variables predictoras

$X_1=$ Género del paciente $\begin{cases} x_1=0 & mujer \\ x_1=1 & \text{var}on \end{cases}$

$X_2=$ Edad (años)

$X_3=$ Índice de masa corporal del paciente (Kg/m^2)

$X_4=$ Raza en el paciente $\begin{cases} x_4=1 & blanca \\ x_4=2 & negra \\ x_4=3 & otras \end{cases}$

$X_5=$ Alcoholismo en el paciente $\begin{cases} x_5=0 & no\ \ toma \\ x_5=1 & toma \end{cases}$

$$X_6 = \text{Tabaquismo en el paciente} \begin{cases} x_6 = 0 \quad no \quad fuma \\ x_6 = 1 \quad fuma \end{cases}$$

$$X_7 = \text{Actividad física del paciente} \begin{cases} x_7 = 0 \quad si \quad actividad \quad fisica \\ x_7 = 1 \quad no \quad actividad \quad fisica \end{cases}$$

$$X_8 = \text{Colesterol (mg/dl)} \begin{cases} x_8 = 0 \quad (< 200 \quad normal) \\ x_8 = 1 \quad (200 - 240 \quad normal \quad alto) \\ x_8 = 2 \quad (>= 240 \quad alto) \end{cases}$$

$$X_9 = \text{Triglicéridos (mg/dl)} \begin{cases} x_9 = 0 \quad (< 150 \quad normal) \\ x_8 = 1 \quad (150 - 200 \quad \text{mod}\, erado) \\ x_8 = 2 \quad (>= 200 \quad alto) \end{cases}$$

$$X_{10} = \text{PAS (mmHg)} \begin{cases} x_{10} = 0 \quad (< 140 \quad normal) \\ x_{10} = 1 \quad (>= 140 \quad alto) \end{cases}$$

$$X_{11} = \text{PAD (mmHg)} \begin{cases} x_{11} = 0 \quad (< 90 \quad normal) \\ x_{11} = 1 \quad (>= 90 \quad alto) \end{cases}$$

$$X_{12} = \text{Diabetes} \begin{cases} x_{12} = 0 \quad no \quad tiene \quad diabetes \\ x_{12} = 1 \quad tiene \quad diabetes \end{cases}$$

4.1.3.1 Estimación del modelo logístico

El modelo logístico, fue ajustado para una muestra de 1284 pacientes con edades 20 años a más, considerando como variable dependiente a la enfermedad cardiovascular (tiene la enfermedad=1 o no la tiene=0), y un total de 12 variables predictoras como se indicó. La estimación de los parámetros del modelo logístico, lo proporciona el programa estadístico SPSS versión 18. La evaluación del modelo usando la prueba de razón de verosimilitud (estadístico G) al 5% de significación, se rechaza la hipótesis nula, que indica que los coeficientes del modelo logístico son nulo, esto implica que por lo menos algún predictor del modelo se relaciona (influye) en la probabilidad de ocurrencia de la enfermedad cardiovascular $P(Y = 1)$; la prueba

estadística Wald para probar la significación de cada parámetro del modelo, resultaron significativos al 5% los siguientes predictores: Edad,, alcoholismo, colesterol, triglicéridos, presión arterial sistólica, presión arterial diastólica, lo que implica que los predictores asociados a los parámetros se relacionan (influyen) en la probabilidad de ocurrencia de la enfermedad cardiovascular $P(Y=1)$. Asimismo es posible evaluar el porcentaje de explicación del modelo estimado a través de los pseudo coeficiente determinación $R^2\ Cox\ y\ Snell$ es igual a 47.5%, indica que 47.5% de la variación en los resultados en la enfermedad cardiovascular, es explicada en forma conjunta por los predictores que resultaron significativos en el modelo estimado; $R^2\ Negel\ker ke$ es igual a 64.8% y $R^2\ McFadden$ 48.8%; la prueba de bondad de ajuste. Hosmer-Lemeshow, resulto no significativa (p=0.206), lo cual indica que el modelo estimado se ajusta (acomoda) bien a las observaciones Y_i (no hay diferencia significativa entre los resultados observados y los predichos por el modelo estimado).

4.1.3.2 Interpretación de los coeficientes estimados

La Interpretación de los resultados se facilita si el modelo estimado se presenta en término de OR=razón de Odds:

La edad (años): Su coeficiente estimado fue de 0.025 y resultó significativo al 5% a través de la prueba Wad, lo cual indica que la variable edad se relaciona con la probabilidad de ocurrencia de la enfermedad cardiovascular con un Odds-ratio, $OR=e^{0.025}=1.025(LC:1.012-1.039)$, valor que indica que la edad es variable de riesgo para la enfermedad, se interpreta de la siguiente forma, si la edad un paciente se incrementa en 1 año, manteniéndose fijos las otras variables del modelo estimado, la posibilidad de tener la enfermedad cardiovascular se incrementa en 1.025 veces mayor.

Se hace de su conocimiento que si la edad se incrementa en 10 años, es decir por ejemplo, una persona de tiene hoy 40 años después de 10 años tiene 50 años, la probabilidad de padecer la enfermedad cardiovascular cuando tenga los 50 años, será $OR = e^{(10*0.025)} = 1.284$ veces mayor.

Alcoholismo: Su coeficiente estimado fue de -0.441 resultando significativo al 5% a través de la prueba Wald, lo cual indica que la variable alcoholismo influye en la probabilidad de ocurrencia de enfermedad cardiovascular, con un $OR = e^{-0.441} = 0.644(LC: 0.447 - 0.928)$, valor que es menor a la unidad, lo que significa que alcoholismo es protector para la enfermedad cardiovascular, calculando su inversa resulta 1.55, lo cual significa que las personas que no consumen alcohol, tienen 1.55 veces mayor posibilidad de tener la enfermedad cardiovascular en comparación de aquellas que si consumen alcohol en forma moderada, manteniéndose fijos las otras variables.

Colesterol (mg/dl): Su coeficiente estimado fue de 0.012, resultando significativo (0.05>p), lo cual indica que la variable colesterol influye en la probabilidad de ocurrencia enfermedad cardiovascular, con un OR=1.012(LC: 1.009-1.016), valor que indica que la variable colesterol es de riesgo para la enfermedad, se interpreta se la siguiente forma, si el nivel de colesterol de un paciente se incrementa en 1 mg/dl, manteniéndose fijos las otras variables, la posibilidad de tener la enfermedad cardiovascular es 1.012 veces mayor.

Triglicéridos (mg/dl): Su coeficiente estimado fue de 0.008, resultando significativo (0.05>p), lo cual indica que la variable triglicérido influye en la probabilidad de ocurrencia de la enfermedad cardiovascular con un OR=1.011(LC:1.008-1.014), valor que indica que la variable de riesgo para la enfermedad, se interpreta de la siguiente forma, si el nivel de triglicéridos de un paciente, se incrementa en 1 mg/dl, manteniéndose fijos las

otras variables, la posibilidad de tener la enfermedad cardiovascular es 1.01 veces mayor.

Presión arterial sistólica (mmHg) : Su coeficiente estimado fue de 0.081, resultando con la prueba Wald significativo (0.05>p), lo cual indica que la variable presión arterial sistólica se relaciona o influye en la probabilidad de ocurrencia de la enfermedad cardiovascular con un OR=1.085(LC:1.070-1.099), valor que indica que es un factor de riesgo, cuya interpretación es, si la presión arterial sistólica de un paciente se incrementa en 1 mmHg, manteniendo las otras variables, la posibilidad de tener la enfermedad cardiovascular es 1.085 veces mayor.

Presión arterial diastólica (mmHg): Su coeficiente estimado fue de 0.05, resultando con la prueba Wald significativo (0.05>p), lo cual indica que la variable presión arterial diastólica se relaciona o influye en la probabilidad de ocurrencia de la enfermedad cardiovascular con un OR=1.051(LC: 1.018-1.085), valor que indica, si la presión arterial diastólica es factor de riesgo, su valor se interpreta de la siguiente forma, si la presión arterial sistólica de un paciente se incrementa en 1 mmHg, manteniéndose fijo a las otras variables, la posibilidad de tener la enfermedad cardiovascular es 1.051 veces mayor. Se concluye que para el factor demográfico la variable edad se relaciona con la enfermedad cardiovascular, para el factor estilo de vida, la variable alcoholismo se relaciona con la enfermedad cardiovascular y para el factor clínico, las variables colesterol, triglicéridos, presión arterial sistólica y presión arterial sistólica se relacionan con la enfermedad cardiovascular (taba 15).

Tabla 15

Variables relacionados a las enfermedades cardiovasculares, en pacientes adultos atendidos en el hospital PNP Augusto B Leguía-Rímac, utilizando el modelo logístico (n=1284)

variables	Coeficientes estimados	significación estadística	$OR = e^{\hat{\beta}_j}$	Intervalo de confianza 95%
Edad	0.025	0.000(**)	1.025	1.012-1.039
Alcoholismo	-0.441	0.018(*)	0.644	0.447-0.928
Colesterol	0.012	0.000(**)	1.012	1.009-1.016
Triglicéridos	0.008	0.000(**)	1.008	1.005-1.010
PAS	0.081	0.000(**)	1.085	1.070-1.099
PAD	0.050	0.002(**)	1.051	1.018-1.085
Intercepto	-18.042	0.000(**)		
Bondad de ajuste:				
Estadística G	827.131	0.000(**)		
$-2LL(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4,\hat{\beta}_5,\hat{\beta}_6)$	867.099			
Hosmer-Lemeshow	10.930	0.206(ns)		
R^2 McFadden	48.8%			
R^2 Cox y Snell	47.5%			
R^2 Nagelkerke	64.8%			
% global correcto	85.6%			
Curva ROC	0.920	0.000(**)		0.905-0.935

(**) Significación 1%
(*) Significación 5%
(ns) no significativa

4.1.3.3 Evaluación de los supuestos del modelo

- Matriz de correlación, multicolinealidad:

La matriz de correlaciones entre las estimaciones de los parámetros del modelo, nos permite hacer un análisis de la existencia de colinealidad entre las variables predictora. En este caso se puede concluir que todas las correlaciones son bajas (tabla 16).

El factor de inflación de varianza (FIV) para cada factor, son menores a 3, lo cual indica que las variables predictoras presentan baja multicolinealidad (tabla 17).

Tabla 16

Matriz de correlaciones entre las estimaciones de los parámetros del modelo estimado.

Predictores	Const.	Edad	Alcohol(1)	Colest.	Trigli.	PAS	PAD
Const.	1.000	-.342	-.140	-.323	-.076	-.302	-.640
Edad	-.342	1.000	.344	.005	.070	-.089	.033
Alcohol	-.140	.344	1.000	-.083	-.002	-.079	.019
Colest.	-.323	.005	-.083	1.000	-.272	.081	.001
Trigli.	-.076	.070	-.002	-.272	1.000	-.040	.025
PAS	-.302	-.089	-.079	.081	-.040	1.000	-.385
PAD	-.640	.033	.019	.001	.025	-.385	1.000

Tabla 17

Factor de inflación de varianzas de los predictores del modelo estimado

Predictores	Coeficiente estimado	Error estándar del coeficiente estimado.	FIV
Constante	-18.04	1.22	
Edad	0.02506	0.00669	1.15
Colesterol	0.01234	0.00195	1.10
Triglicéridos	0.00756	0.00120	1.09
Presión arterial sistólica	0.08131	0.00677	1.19
Presión arterial diastólica	0.0501	0.0163	1.18
Alcoholismo	-0.441	0.187	1.15

Nota. Obtenido por Minitab versión 14

- **Análisis de los residuos e influencia**

Por el análisis de residuos permite detectar observación discordante (observaciones que no siguen un patrón de comportamiento regular o típico) y observaciones influyentes (observaciones que influyen en la estimación de los parámetros del modelo), el programa SPSS

detecta 44 observaciones discordantes, observaciones que superan ± 2 desviaciones estándar ver tabla 27 de anexo 05 que representa el 3.43% de la muestra (1284). Este análisis se realiza con el fin de mejorar el modelo ajustado. De las figuras (17, 18, 19, 20) se observa que no existen valores atípicos ni tampoco valores influyentes en la estimación de los parámetros estimados.

Figura 17

Residuos estandarizados y observaciones

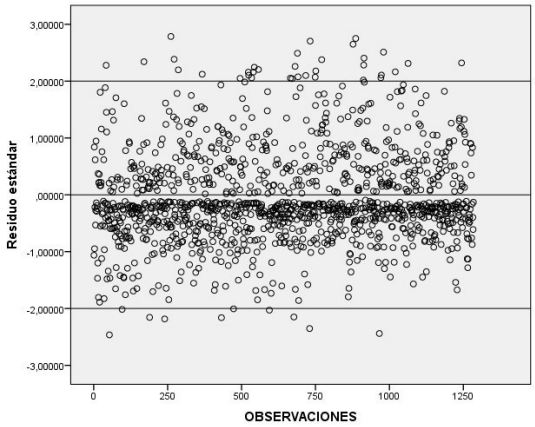
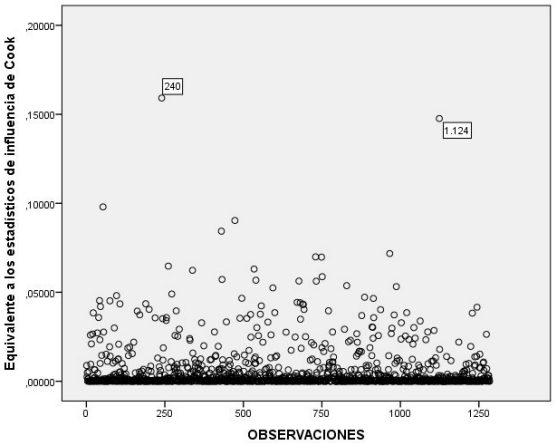


Figura 18

Distancia de Cook y observaciones.



Se observa de que no existe datos influyentes ya que los datos 240 y 1124 no superan el valor de la distancia de $Cook > 1$.

Figura 19

Probabilidad estimada y distancia de Cook (n=1284)

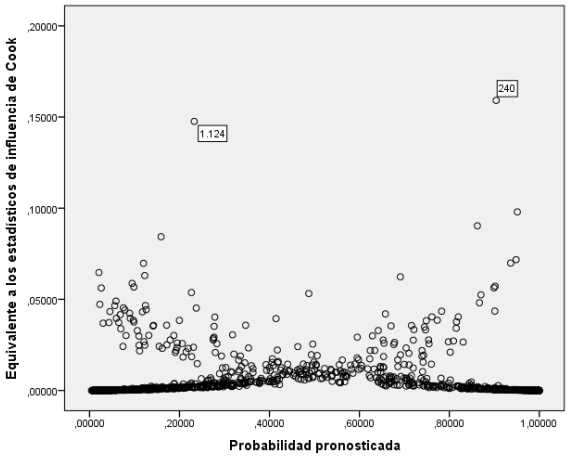
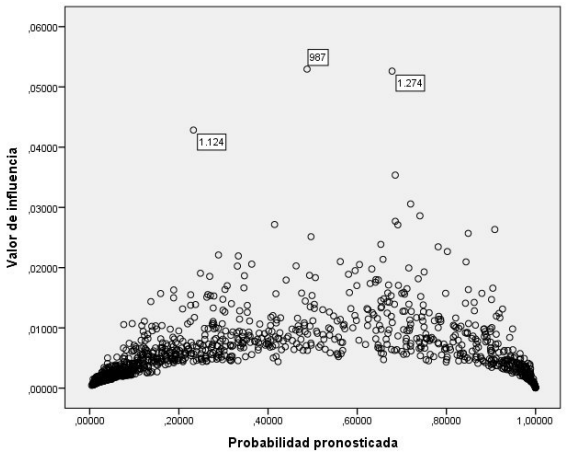


Figura 20

Probabilidad estimada y valores de influencia (n=1284)



- **Interacciones de predictores**

No se logró determinar interacciones significativas entre los predictores para la enfermedad cardiovascular.

4.1.3.4 Capacidad predictiva del modelo estimado

El modelo de regresión estimado con punto de corte 0.5, permite clasificar a los pacientes de la siguiente forma: Del total de pacientes que no tienen la enfermedad el 91.1% fueron

correctamente clasificados, mientras que el 8.9% no lo fueron, y del total de paciente que si tienen la enfermedad el 76.3% fueron correctamente clasificados correctamente, mientras que solo el 23.7% no lo fueron. La tabla de clasificación presenta además que con este modelo estimado, fueron clasificados correctamente (sin y con la enfermedad) el 85.6% de los pacientes y 14.4% fueron clasificados incorrectamente llamado error aparente (tabla 18).

Tabla 18

Clasificación de pacientes, mediante el modelo logístico estimado, con punto de corte c=0.50 (n=1284)

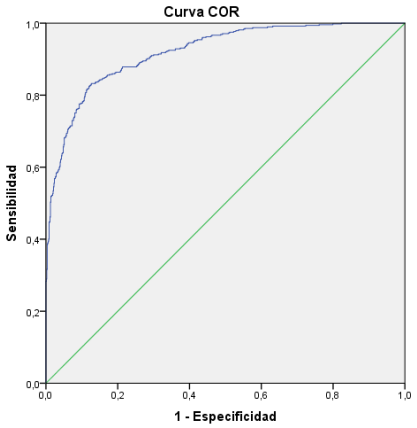
Observados		Pronosticado Enfermedad		Porcentaje correcto
		No=0	Si=1	
Enfermedad cardiovascular	**No=0**	735	72	91.1%
	Si=1	113	364	76.3%
Porcentaje global				85.6%

4.1.3.5 Prueba ROC, modelo logístico estimado.

El área bajo la curva ROC es 0.92, resultado significativo al 5% IC (0.905; 0.935), valor que indica que el 92% es la probabilidad de que el modelo logístico binario múltiple estimado clasifique correctamente a dos pacientes, el sano en sano y al enfermo en enfermo de la enfermedad cardiovascular (figura 21).

Figura 21

Curva ROC del modelo, tasa de falsos positivos (1-especificidad) y tasa de verdaderos positivos (sensibilidad)



Con la curca ROC, se estima un nuevo punto de corte de c=0.426, valor que permite obtener una alta sensibilidad (0.805) y alta especificidad (0.89), los resultados son: del total de pacientes (807) que no tienen la enfermedad el 89.0% de los pacientes fueron correctamente clasificados, mientras que el 11.0% no fueron correctamente clasificados; del total de pacientes (477) que tienen la enfermedad, el 80.5% fueron clasificados correctamente, mientras que el 19.5% no fueron clasificados correctamente. La tabla de clasificación presenta además que con este modelo estimado, fueron clasificados correctamente el 85.8% y 14.2%, fueron clasificado incorrectamente (tasa de error) tabla 19.

Tabla 19

Clasificación de pacientes mediante el modelo logístico estimado con punto de corte c=0.426 (n=1284)

Observados		Pronosticado		
		Enfermedad		Porcentaje correcto
		No=0	Si=1	
Enfermedad cardiovascular	No=0	718	89	89.0%
	Si=1	93	384	80.5%
Porcentaje global				85.8%

Otra forma de realizar la clasificación de las observaciones, es por medio de la validación cruzada, consiste en eliminar la $i-$ ésima observación y con las restantes observaciones $(n-1)$, se estima el modelo logístico y luego se estima $\hat{\pi}_{(i)}$ y así sucesivamente (Berrendero, J., Cárcamo, J. ,2019). Con punto de corte de c=0.5, los resultados son: del total de pacientes (807) que no tienen la enfermedad el 90.6% de los pacientes fueron correctamente clasificados, mientras que el 9.4% no fueron correctamente clasificados; del total de pacientes (477) que tienen la enfermedad, el 76.1% fueron clasificados correctamente, mientras que el 23.9% no fueron clasificados correctamente. La tabla de clasificación

presenta además que con este modelo estimado, fueron clasificados correctamente el 85.2% y 14.8%, no fueron clasificado correctamente (tasa de error por validación cruzada) tabla 20.

Tabla 20

Clasificación de pacientes, a través de la validación cruzada con punto de corte c=0.5, (n=1284),

Observados		Pronosticado		
		Enfermedad		Porcentaje correcto
		No=0	Si=1	
Enfermedad cardiovascular	**No=0**	731	76	90.6%
	Si=1	114	363	76.1%
Porcentaje global				85.2%

Tabla 21

Estadísticas resumen del modelo estimado de regresión logística binaria múltiple.

Tamaño de la muestra	-2 log Likelihood, para el modelo estimado	Estadística G	R^2 Cox-Snells	R^2 McFadden	R^2 Nagelkerke	% observaciones correctamente clasificados con la enfermedad (sensibilidad)	% observaciones correctamente clasificados sin la enfermedad (especificidad)	% de observaciones correctamente clasificados con y sin la enfermedad	Prueba de Hosmer Lemeshow P-valor	Área de la curva ROC	Factores significativos en el modelo estimado
1284	867.099	827.131	47.5%	48.8%	64.8%	76.3%	91.1%	85.6%	0,206	0,92	6/12

Modelo logístico binario múltiple estimado con todos las variables predictoras: Género, edad, índice de masa corporal (IMC), raza, alcoholismo, tabaquismo, actividad física, colesterol, triglicéridos, presión arterial sistólica (PAS), presión arterial diastólica (PAD) y diabetes, para una muestra elegida al azar de 1284 pacientes, resultaron significativo al 5% para predecir la enfermedad cardiovascular. Las variables: Edad corresponde al factor demográfico; alcoholismo corresponde al factor estilo de vida y las variables como colesterol, triglicéridos, presión arterial sistólica y presión arterial diastólica todos ellos corresponden al factor clínico, otras característica del modelo estimado con solo 6 predictores se presentan en la tabla, además si este modelo se compara con los otros modelo estimados para solo factores demográficos, estilo de vida y clínicos que se presenta en el anexo 05, el modelo elegido tiene el menor criterio de información de Akaike (AIC=881.099).

4.1.3.6 Predicciones

Una vez estimado los coeficientes el modelo logístico binario múltiple, es posible realizar estimaciones de la probabilidad de ocurrencia de la enfermedad, dado un conjunto de predictores, $\hat{P}(Y=1/x_1,x_2,...x_k)$

Utilizando los coeficientes del modelo estimado

$$\hat{P}(Y=1/X)=\frac{e^{-18.042+0.025*edad-0.441*alcoholismo+0.012*colesterol+0.008*trigliceridos+0.081*PAS+0.050*PAD}}{1+e^{-18.042+0.025*edad-0.441*alcoholismo+0.012*colesterol+0.008*trigliceridos+0.081*PAS+0.050*PAD}}$$

Para las predicciones se consideran dos pacientes: paciente-1, con valores normales y paciente-2 con valores no normales, las probabilidades estimadas para cada uno de ellos se presentan en la tabla 22.

Paciente1: Con niveles normales.

$$P(Y=1/Edad,alcoholismo,colesterol,trigliceridos,PAS,PAD)=0.2586$$

Si se elige al azar a un paciente-1, con los valores normales (tabla 22), la probabilidad de que tenga la enfermedad cardiovascular es de 25.86%.

Paciente-2: Con niveles por encima de lo normal.

$$P(Y=1/Edad,alcoholismo,colesterol,trigliceridos,PAS;PAD)=0.9568$$

Si se elige al azar a un paciente2, con los valores no normales (tabla 22), la probabilidad de que tenga la enfermedad cardiovascular es de 95.68%.

Tabla 22

Variables que se relacionan con las enfermedades cardiovasculares, a través del modelo logístico binario

Predictores	Coeficientes del modelo estimado	Paciente-1 con valores normales	Paciente-2 con valores por encima de lo normales
Edad(años) riesgo ECV	0.025	60	60
Alcoholismo protector ECV	-0.441	1	0
Colesterol(mg/dl) riesgo ECV	0.012	140	210
Triglicéridos(mg/dl) riesgo ECV	0.008	90	160
PAS(mmHg) riesgo ECV	0.081	130	140
PAD(mmHg) riesgo ECV	0.050	60	90
Constante	-18.042		

4.2 Discusión

El objetivo de la investigación fue determinar los factores relacionados a las enfermedades cardiovasculares, basado en una muestra simple aleatoria de 1284 pacientes adultos con edades de 20 años a más, atendidos por consulta externa en el Hospital PNP Augusto B. Leguía-Rímac. Lima-Perú, mediante el análisis de regresión logística binaria múltiple, usando el programa SPSS, por el método de selección de variables predictoras hacia atrás; fueron identificados: Edad, alcoholismo, colesterol, triglicéridos, presión arterial sistólica, presión arterial diastólica. Esto indica que están relacionados o influyen en la probabilidad de ocurrencia de la enfermedad cardiovascular $P(Y=1)$.

Los resultados se corroboran con otras investigaciones tales como:
La edad como factor de riesgo: Luis (2021), Álvarez et al. (2020), Hernández-Martínez et al. (2019), Cisneros & Carrazana (2013), Rojas et al. (2016) y Zuni-Chávez et al. (2019); alcoholismo como factor protector: Cedeño-Zambrano et al. (2016) y Sacanella et al. (2019); colesterol factor de riesgo: Luis (2021), Zuni-Chávez et al (2019), Patiño-Villada et al. (2011), Cisneros & Carrazana (2013); triglicéridos factor de riesgo: Patiño-Villada et al. (2011), Cisneros & Carrazana (2013), Rojas et al. (2016); presión arterial sistólica factor de riesgo: Berdú et al. (2020), Hernández -Martínez et al. (2019), Luis (2021), Zuni-Chávez et al. (2019), Patiño-Villada et al. (2011), Cisneros & Carrazana (2013), Rojas et al (2016); presión arterial diastólica factor de riesgo: Berdú et al. (2020), Hernández -Martínez et al. (2019), Zuni-Chávez et al (2019), Patiño-Villada et al. (2011), Cisneros & Carrazana (2013).

Sin embargo, las otras variables: Género, raza, índice de masa corporal, tabaquismo, actividad física, diabetes, para mi estudio (multivariado) resultaron no significativos. Para los otros autores mencionados en los antecedentes resultaron significativos dichas variables para la enfermedad cardiovascular, ya que ellos solo

realizaron el análisis bivariado (variable de riesgo y la enfermedad cardiovascular).

Conclusiones

El objetivo fue determinar factores relacionados a las enfermedades cardiovasculares en personas adultas, en una muestra al azar de 1284 pacientes adultos de 20 años a más, atendidos por consultas externas en el Hospital PNP Augusto B Leguía–Rímac, Lima-Perú, entre los meses de Octubre a Diciembre 2017.

Se han identificado mediante la regresión logística binaria múltiple a los siguientes factores: Edad (factor de riesgo); alcoholismo (factor protector); colesterol (factor de riesgo), triglicéridos (factor de riesgo), presión arterial sistólica (factor de riesgo) y presión arterial diastólica (factor de riesgo).

Los resultados de esta investigación son corroborados con los antecedentes del estudio.

Recomendaciones

Si bien es cierto que los gasto en salud en casi todos los países del mundo, genera una fuerte inversión de miles de dólares para el estado, tomando en cuenta la discusión y conclusión expuesta con respecto a los resultados encontrados sobre los factores que se relacionan o influyen en la enfermedad cardiovascular, en particular a la población de pacientes que concurren al Hospital PNP Augusto B Leguía-Rímac-Lima-Perú, recomendamos:

1. Hacer de conocimiento a todos los pacientes que concurren al Hospital Nacional PNP Augusto B Leguía-Rímac y también en sus centros de trabajos, sobre estos factores de riesgo cardiovasculares (Edad, alcoholismo, colesterol, triglicéridos, presión arterial sistólica (PAS), presión arterial diastólica (PD)), sobre todo el control de estos factores en forma semestral o anual, a fin de disminuir la aparición de estas enfermedades y difusión de charlas y panfletos de prevención.

2. Los resultados de este estudio de investigación, que sea una llamada de atención de promover la prevención, a través de la alimentación saludable, cambio de estilo de vida y control de su salud por lo menos una vez al año.

3. Si bien es cierto que la edad es un factor de riesgo cardiovascular no modificable, se pueden controlar otros factores modificables: como colesterol, triglicéridos, presión arterial sistólica y presión arterial diastólica, a través de práctica de algún deporte acompañados de buenos hábitos alimenticios.

4. Con los resultados de la tesis hacer de conocimientos al personal médico sobre los factores de riegos que se asocian

a la enfermedad cardiovascular, a afín de que ellos puedan instruir a los pacientes sobre la prevención. Generalizar estos resultados a otros centros de Salud.

5. Para próximos estudios sobre enfermedades cardiovasculares, debería tomarse en cuenta otros factores que no fueron estudiados, tales: Hábitos en la alimentación, el estrés y la herencia genética. No se disponía de esos datos y otras literaturas lo mencionan.

6. Que el colegio de estadísticos, gestione ante los distintos ministerios del país, que los profesionales en estadística tengamos acceso a datos estadísticos (salud, educación, economía, agricultura, pesquería, minería, etc.), para luego revertir en decisiones acertadas en mejora de la calidad de vida de los compatriotas.

Referencias Bibliográficas

Agresti, A. (2013). *Categorical Data Analysis.* John Wiley & Sons.

Akaike, H. (1974). *Una nueva mirada a la identificación del modelo estadístico*. https //doi.org /10.1109/TAC. 1974.1100705

Álvarez, L., Frías, J., Fernández de Bobadilla, J., Díaz, M. (2020). Prevalencia de factores de riesgo cardiovascular en trabajadores de un hospital terciario de Madrid. *Rev Asoc Esp Med Trab • Diciembre 2020 • vol 29 • Núm. 4 • 257-392* .https://scielo.isciii.es/pdf/medtra/v29n4/1132-6255-medtra-29-04-274.pdf

Barrera, D. (2015). *Factores de riesgo para enfermedades cardiovasculares según los determinantes de la salud presentes en los choferes de transporte público 2014*. [Tesis de licenciatura, Universidad Nacional Mayor de San Marcos].

Berdú, J., Chacón, T., Fonseca, A. & Pérez, R. (2020). Factores de riesgo de las enfermedades cardiovasculares en pacientes de 50 años y menos. *Revista Médica. Granma RPNS-1853 773*. http://scielo.sld.cu/pdf/mmed/v24n4/1028-4818-mmed-24-04-772.pdf

Berrendero, J., Cárcamo, J. (2019). Las componentes lineales de una regla de clasificación cuadrática.

https://caminosaleatorios.wordpress.com/2018/09/04/las-componentes-lineales-de-una-regla-de-clasificacion-cuadratica.

Chambergo-Michilot, D., Velit-Rios, B., Cueva-Parra, A. (2020). Prevalencia de enfermedades cardiovasculares en el Hospital Nacional Dos de Mayo de Perú. *Revista Mexicana Angiología*. 48(3):84-89, https://www.rmangiologia.com/frame_esp.php?id=33

Cedeño-Zambrano, J., Vásquez-Jaramillo, P. y Roca-Lino, V. (2016). Riesgo cardiovascular relacionado con el consumo de alcohol. REVISTA CIENTÍFICA DOMINIO DE LA CIENCIA, 2(4), 17-27. file:///C:/Users/Lenovo/Downloads/Dialnet-RiesgoCardiovascularRelacionadoConElConsumoDeAlcoh-5761632.pdf

Chapra, S., Canale, R. (2006). *Métodos Numérico para ingenieros*, 5ta. Ed., Mc Graw Hill.

Cisneros, L., Carrazanas, E. (2013). Factores de riesgo de la cardiopatía Isquémica. *Revista Cubana de Medicina General Integral, 29(4),369-378.*

Cook, R..& Weisberg, S. (1982). *Residuos e influencia en la regresion* Chapman and Hall.

Cox, D. & Snell, E. (1989). *Analysis of Binary Data*. Chapman & Hall.

De la Garza, J., Morales, B., Gonzales, B. (2013). *Análisis estadístico Multivariante*. Editorial: Mc Graw Hill. (pp.60-107).

Evans, M., Rosenthal, J. (2013). *Probabilidad y estadística*. Editorial Reverte, S.A. Barcelona.

FUNDACIÓN ESPAÑOLA DEL CORAZÓN, (2022). La raza como factor de riesgo cardiovascular [Archivo PDF]. https://fundaciondelcorazon.com/prevencion/marcadores-de-riesgo/raza-etnia-linaje.html

Gujarati, N., Garmendia, D. y Arango, G. (2005). *Econometría* (pp. 167-181) *(*4° ed.) México: McGraw-Hill.

Hernández-Martínez, J., Varona-Uribe, M., Hernández, G (2019). Prevalencia de factores asociados a la enfermedad cardiovascular y su relación con el ausentismo laboral de los trabajadores de una entidad oficial, Revista Colombiana de Cardiología 2020; 27(2):109-116. https://www.rccardiologia.com/previos/RCC%202020%20Vol.%2027/RCC_2020_27_2_MAR_ABR/RCC_2020_27_2_109-116.pdf

Hosmer, D., Lemeshow, S (2000). *Applied logistic regression*. John Wiley & Sons.

INEI. (2012). *Condiciones de Riesgo Cardiovasculares*. Perú: Situación de Salud de la Población Adulta Mayor. [Archivo PDF]. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1146/cap02.pdf

Luis, A. (2021). Riesgo cardiovascular en adultos que acuden a la consulta médica en un policlínico privado en Carabayllo. Revista Científica Cuidado y Salud Pública. 2021; 1(1):18-23. https://www.cuidadoysaludpublica.org.pe/index.php/cuidadoysaludpublica/article/view/11/135.

Mantel y Haenszel (1959). Aspectos estadísticos del análisis de datos de estudios retrospectivos de enfermedades. *Revista del Instituto Nacional del Cáncer, 22 (4), 719–748.*

MedlinePlus en español [Internet]. Bethesda (MD): Biblioteca Nacional de Medicina (EE. UU.) [actualizado 22 nov. 2021]. Enfermedad Cardiovascular; [actualizado 22 nov. 2021; consulta 30 ago. 2022]; [aprox. 2 p.]. Disponible en:

https://medlineplus.gov/spanish/ency/patientinstructions/000759.htm

McCullagh, P. & Nelder, J. (1983). *Generalized Linear Models*. Chapman and Hall.

McFadden, D. (1974). *Análisis logit condicional del comportamiento de elección cualitativa, en P. Zarembka*, editor, Frontiers in Econometrics Nueva York, 1974.

Montgomery, D., Peck, E. & Vining, G. (2004). *Introducción al análisis de regresión Lineal* (pp. 399-414) (3° Ed.). Compañía Editorial Continental.

Nagelkerke, N.J.D. (1991). *A note on a general definition of the coefficient of determination*. Biometrika, 78, 691-692.

Nelder, J. and Wedderburn. R (1972). *Generalized Linear Models*. Chapman and Hall. London.

OMS. (2020). Temas de salud. Enfermedades cardiovasculares. Obtenido de https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1

OPS. (29 de setiembre de 2021). Las enfermedades del corazón siguen siendo la principal causa de muerte en las Américas.

https://www.paho.org/es/noticias/29-9-2021-enfermedades-corazon-siguen-siendo-principal-causa-muerte-americas

Patiño-Villada, F., Arango-Vélez, E., Quintero-Velásquez, M. & Arenas-Sosa, M. (2011). Factores de riesgo cardiovascular en una población urbana de Colombia. *Revista de Salud pública*, 13(3), 433-445.

Pérez, C. (2014). *Técnicas Estadísticas Predictivas Con IBM SPSS Modelos* (pp.126-138). Editorial Garceta.

Rivas, E. (2016). Enfermedades cardiovasculares y actividad física: Recomendaciones para la Atención Primaria de Salud en Cuba. CorSalud, 8(3),139-143.

Rojas, S., Querales, M. Villarino, A. (2016). Evaluación de los factores de riesgos que predisponen a la hipertensión arterial a través de un modelo de regresion logística. *Salus.*, 20(2),18-23.

Sacanella, I., Casas, R., Viñas, E., Castro, S. y Sacanella, E. (2019). Prevención de la enfermedad cardiovascular y bebidas alcohólicas fermentadas. ¿Realidad o ficción?. Servicio de Medicina Interna. Hospital Clínic. Barcelona. Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS). Universidad de Barcelona. Barcelona. CIBER de Obesidad y Nutrición. Instituto de Salud Carlos III. Madrid.https://dialnet.unirioja.es/servlet/articulo?codigo=7048944

TEXAS HEART INSTITUTE, (2022) Factores de riesgo cardiovascular, consultado Julio 2022, disponible en: https://www.texasheart.org/heart-health/heart-information-center/topics/factores-de-riesgo-cardiovascular.

Uriel, E. & Aldas, J. (2005). *Análisis Multivariante Aplicado(5ª ed.)*. Thomson

Wald, A (1943) *Tests of statistical hypotheses concerning several parameters when the number of observations is large*. Trans.Amer. Math. Soc. 54 426-482.

Zuni-Chávez, K., More-Sandoval, B., Fernández-Vargas, C., García-Fuentes, B., Ruiz-Olano, J. & Pérez-Rodríguez, V. (2019). Prevalencia de factores de riesgo cardiovascular en pacientes hospitalizados en un hospital de Lima. Revista de la Facultad de Medicina Humana, Lima-Perú, 19(4), 68-73.

http://revistas.urp.edu.pe/index.php/RFMH/article/view/2343

ANEXOS

ANEXO: 01 MATRIZ DE CONSISTENCIA

FACTORES DE RIESGO RELACIONADOS CON LAS ENFERMEDADES CARDIOVASCULARES, EN PERSONAS ADULTAS, DE 20 AÑOS A MAS , ATENDIDOS EN EL HOSPITAL PNP AUGUSTO B. LEGUÍA -LIMA-PERÚ 2017

Problema	Objetivo	Marco Teórico	Hipótesis	Diseño Metodológico
Problema general ¿Cuáles son los factores de riesgos de las enfermedades cardiovasculares en personas adultas, de 20 años a mas atendidos en el Hospital PNP Augusto B Leguía -Lima-Perú 2017?	Objetivo general Determinar los factores de riesgo de las enfermedades cardiovasculares, en personas adultas, atendidos en el Hospital PNP Augusto B. Leguía-Rimac-2017, mediante la regresión logística binaria múltiple. a. Identificar factores relacionados a las enfermedades cardiovasculares más frecuentes, en personas adultas, atendidas en el Hospital PNP Augusto B. Leguía-Rímac-2017. b. Evaluar factores relacionados a las enfermedades cardiovasculares, en personas adultas, atendidos en el Hospital PNP Augusto B. Leguía-Rímac-2017.	Marín, A. (2003). Factores de riesgo asociados a enfermedad cardiovascular en el municipio de Zacapa, departamento de Zacapa. El objetivo fue Describir los factores de riesgo asociados a enfermedad cardiovascular, en personas adultas en la zona urbana del municipio de Zacapa, departamento de Zacapa. Se estudió una muestra de 457 personas, a las cuales se entrevistaron, realizando mediciones de peso, talla y presión arterial. Conclusión se identificó que el factor de riesgo más frecuente asociado a enfermedad cardiovascular fue el sedentarismo con una tasa de prevalencia de 90.6%, seguida por de obesidad con 50.1% y antecedentes familiares patológicos con 39.4%. Patiño, V.(2009). Factores de riesgo cardiovascular en una población urbana de Colombia. El objetivo Evaluar la prevalencia de los factores de riesgo cardiovascular en una población urbana entre 25 y 50 años del municipio de Santa Rosa de Osos, Antioquia 2009. Se realizó un estudio de prevalencia, con muestreo aleatorio bietápico en 357 personas sin	Hipótesis general Los factores relacionados a las enfermedades cardiovasculares son determinados mediante la regresión logística binaria múltiple. Hipótesis específicas: Hipótesis especifica 1: Los factores demográficos (edad, género, raza, índice de masa corporal), factor estilo de vida (alcoholismo, tabaquismo, actividad física) y factores clínicos (colesterol, triglicéridos, presión sistólica, presión diastólica, diabetes) están relacionados a la enfermedad cardiovascular, pueden ser identificados mediante pruebas estadísticas bivariadas. Hipótesis especifica 2: : Los factores demográficos (edad, género, raza, índice de masa corporal), factor estilo de vida (alcoholismo, tabaquismo, actividad física) y factores clínicos (colesterol, triglicéridos, presión sistólica, presión diastólica, diabetes), están relacionados a la enfermedad cardiovasculares, son	Tipo: aplicada Nivel: correlacional y explicativa Diseño: no experimental Instrumentos de medición: historia clínica Población: todos los adultos de ambos sexo, con edades mayor o igual a los 20 años, atendidos en el Hospital PNP Augusto B Leguía -Lima-Perú 2017? Muestra: pacientes atendidos en el Hospital PNP Augusto B Leguía -Lima-Perú 2017? Modelo estadístico: Regresión logística binaria múltiple.

		enfermedad cardiovascular conocida. Conclusión son: Los factores de riesgo cardiovascular más importantes fueron baja actividad física, obesidad, tabaquismo.	evaluados utilizando la regresión logística binaria. Múltiple.	

ANEXO 02

OPERACIONALIZACIÓN DE VARIABLES

Operacionalización de Variables				
Variables	Definición conceptual	Definición operacional	Escala de medición	Valores finales
1.Género	Característica de las personas en varón o mujer	Cualidad	Cualitativa nominal	0=mujer; 1=varón
2.Edad	Característica del tiempo de vida de una persona, se mide en años, meses, días, horas, minutos y segundos.	Tiempo	Cuantitativa continua	Años
3.Indice de masa corporal	El índice de masa corporal (IMC), es un número que se calcula tomando en cuenta el peso (Kg) y la estatura (Mt.) de una persona. Su valor es un indicador confiable del peso. Se usa para identificar las categorías de peso que puede llevar a problemas de salud.	Para una persona, valor que se obtiene al dividir el peso (kg) y la estatura (mts) elevada al cuadrado.	Cuantitativa continua	IMC(1) ESTADO <18.5 bajo peso 18.5-24.9 peso Normal 25.0-29.9Pre-obes Sobrepeso 30.0--34.9 obesidad 1 35.0-39.9 obesidad 2 >40 obesidad 3
4.Raza	Término que se usa para describir a un grupo de personas que comparten características físicas, como el color de la piel o los rasgos	Cualidad	Cualitativa nominal	0=blanca 1=negra 2=otros

	faciales. Quizás también compartan identidades sociales o culturales y orígenes ancestrales similares			
5.Consumo de alcohol	El alcohol es un depresor del Sistema Nervioso Central, que pertenece al grupo de sedantes junto con los barbitúricos y las benzodiacepinas. El alcohol etílico se obtiene de dos maneras: 1. Por fermentación de frutas, vegetales o granos. Ej.: vinos, cervezas. 2. Por destilación: medio artificial para aumentar la concentración del alcohol de una bebida. Ej.: coñac, ginebra, whisky y vodka	Se considera como consumidora de alcohol a las personas que declararon tomar una o más veces a la semana al momento de la entrevista.	Cualitativa nominal	0=no toma alcohol durante la semana 1=si toma alcohol durante la semana
6.Consumo de tabaco	Se denominan productos del tabaco los que están hechos total o parcialmente con tabaco, sean para fumar, chupar, masticar o esnifar. Todos contienen nicotina, un ingrediente psicoactivo muy adictivo.	Se considera como fumadora a las personas que declararon fumar uno a más cigarros al día al momento de la entrevista.	Cualitativa nominal	0=no fuma durante el día 1=si fuma durante el día
7.Actividad física	Se define como actividad física, aquella personas que hace ejercicio físico no menos de 30 minutos 3 veces a la semana	Se definió como falta de actividad física a las personas que durante el último mes no habían practicado deporte o realizado actividad física fuera del horario de trabajo, durante 30 minutos o más, al menos 3 veces a la semana.	Cualitativa nominal	0=si practica algún deporte durante la semana 1=no practica deporte durante la semana

8.Nivel de colesterol	El **colesterol** es una sustancia grasa natural presente en todas las células del cuerpo humano necesaria para el normal funcionamiento del organismo. La mayor parte del colesterol se produce en el hígado, aunque también se obtiene a través de algunos alimentos.	El colesterol total se mide usando un análisis de sangre llamado panel de lipoproteínas, que también mide los triglicéridos. Para esta prueba, usted proporciona una muestra de sangre que se toma después de haber ayunado durante 9 a 12 horas antes de su prueba. Su puntuación total de colesterol se calcula usando la ecuación: nivel de HDL + nivel de LDL + 20 por ciento de su nivel de triglicéridos. Los resultados del colesterol total se expresan en miligramos por decilitro (mg/dl) de sangre y se clasifican de la siguiente manera: nivel deseable – menos de 200 mg/dl límite alto – 200-239 mg/dl alto – 240 mg/dl y superior	Cuantitativa continua	0=Normal: menos de 200 mg/dl 1=Normal-alto: entre 200 y 240 mg/dl. Se considera hipercolesterolemia a los niveles de colesterol total superiores a 200 mg/dl. 2=Alto: por encima de 240 mg/dl
9.Nivel de triglicéridos	Los triglicéridos son un tipo de grasa que se encuentra en la sangre. El exceso de este tipo de grasa puede aumentar el riesgo de enfermedad de las arterias coronarias, especialmente en mujeres.	Un análisis de sangre mide su nivel de triglicéridos, junto con el colesterol. Los niveles de triglicéridos normales están por debajo de 150. Los niveles superiores a 200 se consideran altos.	Cuantitativa continua	0=normal (triglicéridos menor a 150 mg/dl, 1=moderado (triglicéridos entre 150-199 mg/dl), 2=muy alto(triglicéridos superior a 200)
Prensión arterial. 10. Presión arterial sistólica(PAS) 11. Presión arterial diastólica(PAD)	La presión arterial es necesaria para aportar oxígeno y nutrientes a los órganos corporales. La presión arterial es la fuerza que ejerce la sangre contra las paredes de las arterias. Cada vez que el corazón late, bombea sangre hacia las arterias, que es	La presión arterial se mide en mmHg. Primero se registra siempre el valor sistólico y después el diastólico. Por ejemplo: 120/80 mmHg significa que la presión arterial sistólica es de 120 mmHg y la presión arterial diastólica de 80 mmHg. 1 mmHg es la presión ejercida por 1 milímetro (mm) de mercurio (Hg). Instrumento: tensiómetro de brazo	Cuantitativa continua	1= tiene hipertensión alta si PAS>=140 mmHg y PAD>=90 mmHg 0=no tiene hipertensión alta otro caso

	cuando su presión es más alta. A esto se le llama presión sistólica (PAS). Cuando su corazón está en reposo entre un latido y otro, la presión sanguínea disminuye. A esto se le llama la presión diastólica (PAS).	• 119/79 o menos es considerada presión arterial normal • 140/90 o más se considera hipertensión arterial Entre 120 y 139 para el número más elevado, o entre 80 y 89 para el número más bajo es pre hipertensión. La pre hipertensión significa que puede desarrollar presión arterial alta, a menos que tome medidas.		
12.Diabetes	Es una enfermedad que se produce cuando el páncreas no puede fabricare insulina suficiente o cuando esta no logra actuar en el organismo porque las células no responden a su estimulo	Usted puede medir el nivel de azúcar en la sangre utilizando un pequeño instrumento llamado medidor de azúcar o glucómetro.	Cualitativa nominal	0=no tiene diabetes 1=tiene diabetes
Variable dependiente				
Enfermedad cardiovascular	La enfermedad cardiovascular es un término amplio para problemas con el corazón y los vasos sanguíneos. Estos problemas a menudo se deben a la aterosclerosis. Esta afección ocurre cuando la grasa y el colesterol se acumulan en las paredes del vaso sanguíneo (arteria)	La cardiopatía coronaria La insuficiencia cardíaca Las arritmias Las enfermedades de las válvulas cardíacas La arteriopatía periférica .La presión arterial alta (hipertensión) Un accidente cerebrovascular La cardiopatía congénita	Cualitativa nominal	0=no tiene enfermedad cardiovascular 1=tiene enfermedad cardiovascular

ANEXO 03

FICHA DE RECOJO DE DATOS

Cuestionario, realizado a pacientes del Hospital PNP Augusto B Leguía, con edades de 20 años a más. Marcar con aspa (X) la opción correspondiente.

==========

1. Género Varón ( ) Mujer ( )
2. Edad (años) :……
3. Peso (Kg) :……
4. Estatura (metros) :……
5. Razas Blanca( ) negra( ) Otras( )
6. Consumo de alcohol durante la semana Si( ) No( )
7. Consumo de tabaco durante la semana Si( ) No( )
8. Practica deporte durante la semana Si( ) No( )
9. Nivel de colesterol (mg/dl) :……..
10. Nivel de triglicéridos (mg/d) :……..
11. Presión arterial sistólica (mmHg) :……
12. Presión arterial diastólica (mmHg) :……..
13. Tiene diabetes Si( ) No( )
14. Sufre de enfermedad cardiovascular Si( ) No( )

Gracias.

Anexo 04

VARIABLE CONFUSORA

Es aquella variable predictora que su presencia altera o modifica al verdadero efecto de otra variable predictora asociada con la variable en estudio (variable dependiente). La variable confusora se asocia a otra variable predictora que puede considerarse de riesgo a la variable en estudio. Para que una variable predictora sea de confusión en otra variable predictora la cual es de riesgo para la variable en estudio (enfermedad), esta debe tener las siguientes características:

- La variable confusora (predictor) debe ser una variable de riesgo para la variable en estudio(enfermedad)
- La variable confusora (predictor) debe estar asociada a la variable predictora de riesgo (exposición).
- La variable confusora no debe ser un paso intermedio entre la exposición y la enfermedad.

Para controlar el efecto de la variable confusora, se forman grupos o estratos con respecto a esta variable, en estadística la estratificación se utiliza con el objetivo de formar grupos de modo que las observaciones sean homogéneas dentro de los estrados formados y heterogéneos entre ellos. Para detectar la verdadera asociación dentro de cada grupo (exposición y la enfermedad). Un procedimiento de Mantel y Haenszel (1959), es calcular los OR para cada grupo y probar su homogeneidad estadística (prueba de Breslow-Day o la de Tarone), si se logra probar la homogeneidad, entonces se controló el efecto de la variable confusora es decir la fuerza de asociación es la misma en cada grupo y se procede a calcular el OR generalizado o ajustado, indicador verdadero de la asociación entre la exposición y la enfermedad controlando el efecto de la variable

confusora, si la prueba estadística para el OR ajustado o generalizado es no significativo, indica que la variable que nos permitió estratificar es una variable confusora, de lo contrario no lo es.

Supongamos que la variable X_2 confusora tiene k estratos, y para cada estrato, se tiene la tabla de 2x2 con n_i elementos (clasificación de los elementos de la variable respuesta con la supuesta variable de riesgo), es decir,

El estrato i-ésimo de la posible variable confusora, por la variable respuesta y la variable de riesgo

a_i	b_i	
c_i	d_i	
		n_i

$$OR_{ajustado} = \frac{\sum_{i=1}^{k} \frac{a_i.d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i.c_i}{n_i}}$$

$$\sum_{i=1}^{k} n_i = n \qquad n_i = a_i + b_i + c_i + d_i$$

El programa SPSS versión 23, proporciona la prueba estadística de Mantel y Haenszel (1959), que permite identificar si una variable es confusora o no.

Para nuestro trabajo de investigación, se analizaron los datos para las 12 variables predictores y se encontró al alcoholismo como variable confusora para la relación entre el tabaquismo y enfermedad cardiovascular, con un OR ajustado 0.962 (IC: 0.717-1.289), resultando no significativo.

De los resultados obtenidos de las 12 variables predictoras: género, edad, raza, IMC, alcoholismo, tabaquismo, actividad física, colesterol, triglicéridos, presión arterial sistólica, presión arterial diastólica y diabetes, están asociados en forma bivariada a la enfermedad cardiovascular con un nivel de significación del 5% y son de riesgos, a excepción la variable alcoholismo resultó protectora a la enfermedad.

TABAQUISMO → *ENFERMEDAD*
↑ ↑
ALCOHOLISMO

Pruebas de homogeneidad de razón de ventajas

	Chi cuadrado	Gl	Significación asintótica (bilateral)
Breslow-Day	,605	1	,437
De Tarone	,605	1	,437

Pruebas de independencia condicional

	Chi cuadrado	gl	Significación asintótica (bilateral)
Cochran	,069	1	,793
Mantel-Haenszel	,035	1	,851

Estimación de razón de ventajas común de Mantel-Haenszel

Estimación			,962
ln(Estimación)			-,039
Error estándar de ln(Estimación)			,149
Significación asintótica (bilateral)			,793
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	,717
		Límite superior	1,289
	ln(razón de ventajas común)	Límite inferior	-,332
		Límite superior	,254

La estimación de razón de ventajas común de Mantel-Haenszel se ha distribuido normalmente de forma asintótica bajo la razón de ventajas común de 1,000 supuesto. Así pues, es el logaritmo natural de la estimación.

Se concluye que el alcoholismo es un factor de confusión entre el tabaquismo y la enfermedad cardiovascular.

Fórmulas para calcular el riesgo e intervalos de confianza para una tabla dos por dos.

Variable	Enfermedad(evento)		Total
	Tiene (Y=1)	No tiene (Y=0)	
Edad >=60(1)	A	b	
Edad <60(0)	C	d	
			n

$$\hat{OR} = \frac{ad}{bc} \begin{cases} > 1 & factor\ \ de\ \ riesgo\ \ para\ \ la\ \ enfermedad \\ < 1 & factor\ \ protector\ \ para\ \ la\ \ enfermedad \\ = 1 & ni\ \ riesgo\ \ ni\ \ protector\ \ para\ \ la\ \ enfermedad \end{cases}$$

Intervalo de confianza del $(1-\alpha)100\%, \quad \alpha = 0.01, 0.05, 0.10$

$$LC(OR) = \hat{OR}.e^{\pm z_{(\frac{\alpha}{2}}.\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

El OR es significativo, cuando el intervalo de confianza no contiene a la unidad.

Anexo 05

SALIDAS REGRESIONES LOGÍSTICAS PARA LOS FACTORES DEMOGRÁFICOS, ESTILO DE VIDA Y CLÍNICOS RELACIONADOS CON LA ENFERMEDAD CARDIOVASCULAR.

Tabla 24

Factor demográfico, relacionados a las enfermedades cardiovasculares, hospital PNP Augusto B Leguía-Rímac, utilizando el modelo de regresión logístico (n=1284)

Predictores	Coeficientes estimados	significación estadística	$OR = e^{\hat{\beta}_j}$	Intervalo de confianza 95%
Genero(V/M)	0.418	0.001(**)	1.519	1.186-1.947
Edad (años)	0.044	0.000(**)	1.045	1.036-1.055
IMC(Kg/m^2)	0.118	0.000(**)	1.125	1.082-1.169
Constante	-6.557	0.000(**)		
Bondad de ajuste				
Estadística G	177.278	0.000(**)		
$-2LL(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$	1516.952			
Hosmer-Lemeshow	13.15	0.107(ns)		
R^2 *Cox Snells*	12.90%			
R^2 *Nagel* ker *ke*	17.60%			
R^2 *McFadden*	10.46%			
Clasificación global correcto %	66.20%			
Curva ROC	0.712	0.000(**)		0.683-0.740
AIC	1524.952			

(*) Significación al 5%
(**) Significación 1%
(ns) no significativo

Tabla 25

Factor estilo de vida, relacionados a las enfermedades cardiovasculares, hospital PNP Augusto B Leguía-Rímac, utilizando el modelo de regresión logístico (n=1284)

Predictores	Coeficientes estimados	significación Estadística	$OR = e^{\hat{\beta}_j}$	Intervalo de confianza 95%
Alcoholismo	-0.388	0.002(**)	0.678	0.533-0.864
Actividad física	0.838	0.000(**)	2.313	1.791-2.986
Intercepto	-0.837	0.000(**)		
Bondad de ajuste				
Estadística G	68.677	0.000(**)		
$-2LL(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$	1625.553			
Hosmer-Lemeshow	0.422	0.810(ns)		
R^2 *Cox Snells*	5.20%			
R^2 *Nagel* ker *ke*	7.10%			
R^2 *McFadden*	4.05%			
Clasificación global correcto %	63.20%			
Curva ROC	0.631			0.600-0.662
AIC	1631.553			

(*) Significación al 5%
(**) Significación 1%
(ns) no significativo

Tabla 26

Factor clínico relacionados a las enfermedades cardiovasculares, en pacientes adultos atendidos en el hospital PNP Augusto B Leguía-Rímac, utilizando el modelo logístico binario (n=1284).

Predictores	Coeficientes estimados	significación estadística	$OR = e^{\hat{\beta}_j}$	Intervalo de confianza 95%
Colesterol	0.012	0.000(**)	1.012	1.008-1.016
Triglicéridos	0.007	0.000(**)	1.007	1.005-1.010
PAS	0.085	0.000(**)	1.089	1.075-1.103
PAD	0.051	0.002(*)	1.052	1.019-1.086
Intercepto	-17.224	0.0000(**)		
Bondad de ajuste del modelo estimado:				
Estadística G	796.822	0.000(**)		
$-2LL(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$	897.408			
Hosmer-Lemeshow	12.063	0.148(ns)		
R^2 Cox y Snell	46.20%			
R^2 Nagelkerke	63.10%			
R^2 McFadden	47.03%			
% global correcto	85.60%			
Curva ROC	0.914	0.000(**)		0.897-0.930
AIC	907.408			

(*) Significación 5%
(**) Significación 1%
(ns) no significativa

Tabla 27

Observaciones que superan ± 2 desviaciones estándar, para el modelo estimado con las variables: edad, alcoholismo, colesterol, triglicéridos, presión arterial sistólica y presión arterial diastólica (n=1284)

Lista por casos

Caso	Estado seleccionado	Observado ENFERMEDAD_CARDIOVASCULAR	Pronosticado	Grupo pronosticado	Variable temporal Resid	Variable temporal ZResid
42	S	S**	,075	N	,925	3,507
53	S	N**	,951	S	-,951	-4,414
96	S	N**	,867	S	-,867	-2,552
170	S	S**	,065	N	,935	3,795
189	S	N**	,901	S	-,901	-3,019
240	S	N**	,904	S	-,904	-3,071
261	S	S**	,021	N	,979	6,867
272	S	S**	,059	N	,941	4,003
286	S	S**	,090	N	,910	3,176
367	S	S**	,106	N	,894	2,902
432	S	N**	,902	S	-,902	-3,030
473	S	N**	,862	S	-,862	-2,502
495	S	S**	,124	N	,876	2,654
523	S	S**	,098	N	,902	3,027
524	S	S**	,110	N	,890	2,847
534	S	S**	,123	N	,877	2,675
540	S	S**	,098	N	,902	3,026
543	S	S**	,081	N	,919	3,365
557	S	S**	,089	N	,911	3,198
594	S	N**	,870	S	-,870	-2,590
665	S	S**	,123	N	,877	2,673
671	S	S**	,126	N	,874	2,638
677	S	N**	,899	S	-,899	-2,987
681	S	S**	,079	N	,921	3,420
689	S	S**	,045	N	,955	4,592
691	S	S**	,117	N	,883	2,741
717	S	S**	,111	N	,889	2,827
730	S	N**	,936	S	-,936	-3,835
732	S	S**	,026	N	,974	6,124
749	S	S**	,120	N	,880	2,709
751	S	S**	,095	N	,905	3,094

771	S	S**	,060	N	,940	3,969
877	S	S**	,030	N	,970	5,681
886	S	S**	,023	N	,977	6,540
912	S	S**	,132	N	,868	2,568
913	S	S**	,074	N	,926	3,531
914	S	S**	,056	N	,944	4,107
966	S	N**	,948	S	-,948	-4,288
971	S	S**	,121	N	,879	2,695
973	S	S**	,110	N	,890	2,850
980	S	S**	,043	N	,957	4,715
1020	S	S**	,097	N	,903	3,048
1063	S	S**	,070	N	,930	3,658
1244	S	S**	,068	N	,932	3,693

a. S = Seleccionado, U = casos sin seleccionar, y ** = casos clasificados incorrectamente.

b. Se listan los casos con residuos estudentizados mayores que 2,000.

Anexo 06

Tabla 28

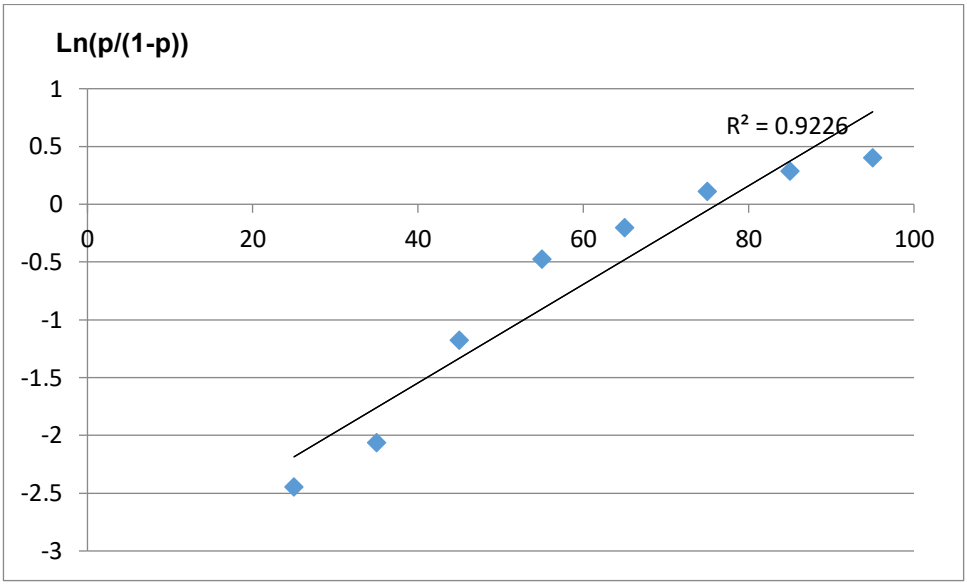
Relación lineal entre cada predictor y logaritmo neperiano del Odds de la prevalencia (n=1284)

Para la variable edad

Edad (años)	Enfermedad cardiovascular		Total	Prevalencia=p	$Ln\left[\frac{p}{1-p}\right]$	Marca de clase
	no	si	total			
20-30	46	4	50	0.08	-2.44234704	25
30-40	110	14	124	0.11290323	-2.06142304	35
40-50	178	55	233	0.2360515	-1.17445037	45
50-60	170	106	276	0.38405797	-0.47235934	55
60-70	167	137	304	0.45065789	-0.19801289	65
70-80	99	111	210	0.52857143	0.11441035	75
80-90	33	44	77	0.57142857	0.28768207	85
90-100	4	6	10	0.6	0.40546511	95
Total	807	477	1284			

Figura 22

Edad y logaritmo neperiano del Odds de la prevalencia de la enfermedad cardiovascular (n=1284)



Existe una relación lineal entre la variable edad y el logaritmo neperiano de Odds de la prevalencia de la enfermedad cardiovascular, para las otras variable cuantitativas se realiza en mismo procedimiento y se llega a las mismas conclusiones.

Calculo del factor de Inflación de varianzas (FIV)

Para obtener los factores de inflación de varianzas para cada variable independiente para una regresion logística, se hace uso de la matriz de correlación.

$$R = D^{-1/2}.COV(\hat{\beta})D^{-1/2}$$

$D^{-1/2}$ Matriz diagonal, cuyos elementos está conformado por las desviaciones estándar de los coeficientes estimados.

$COV(\hat{\beta})$ Matriz de varianzas y covarianzas de los coeficientes estimados

1. Calculo de la matriz de información de Fisher $I(\beta) = (X^{'}VX)$

	p(1-p)	edad	alcoh	colest	trig	ps	pd
p(1-p)	134.842	8089.09	77.014	26349.1	19554.5	15871	8827.9
edad	8089.09	510914	4303.38	1575240	1160907	953693	529582
alcoh	77.014	4303.38	77.014	15350.5	11444.4	9106.64	5045.31
colest	26349.1	1575240	15350.5	5437261	3949882	3094430	1723653
trig	19554.5	1160907	11444.4	3949882	3595361	2302433	1279028
Ps	15871	953693	9106.64	3094430	2302433	1894049	1043193
Pd	8827.9	529582	5045.31	1723653	1279028	1043193	582395

2. La inversa de la matriz de Información de Fisher, es igual a la matriz de varianzas y covarianzas de los coeficientes estimados.

$$COV(\hat{\beta}) = I^{-1}(\hat{\beta}) = (X'VX)^{-1} = S$$

	p(1-p)	edad	alcoh	Colest	trig	ps	pd
p(1-p)	1.47972	-0.00279	-0.03166	-0.00077	-0.00011	-0.00249	-0.01265
edad	-0.00279	4.5E-05	0.00043	6.6E-08	5.6E-07	-4E-06	3.6E-06
alcoh	-0.03166	0.00043	0.03479	-3E-05	-4.6E-07	-1E-04	5.7E-05
colest	-0.00077	6.6E-08	-3E-05	3.8E-06	-6.3E-07	1.1E-06	4.7E-08
Trig	-0.00011	5.6E-07	-4.6E-07	-6.3E-07	1.4E-06	-3.3E-07	4.8E-07
Ps	-0.00249	-4E-06	-1E-04	1.1E-06	-3.3E-07	4.6E-05	-4.2E-05
Pd	-0.01265	3.6E-06	5.7E-05	4.7E-08	4.8E-07	-4.2E-05	0.00026

3. Errores estándar de los coeficientes estimados

Coeficientes	Varianzas	Error estándar
Constante	1.4797185	1.2164368
Edad	0.0000448	0.0066936
Alcoholismo	0.0347905	0.1865222
Colesterol	0.0000038	0.0019516
Triglicéridos	0.0000014	0.0011976
Presión sistólica	0.0000459	0.0067742
Presión diastólica	0.0002646	0.0162663

4. Matriz diagonal D formado por las inversas de las deviaciones estándar de la matriz de varianzas y covarianzas.

0.82207	0	0	0	0	0	0
0	149.397	0	0	0	0	0
0	0	5.36129	0	0	0	0
0	0	0	512.405	0	0	0
0	0	0	0	834.977	0	0
0	0	0	0	0	147.619	0
0	0	0	0	0	0	61.4768

5. Producto de D*S*D=matriz de correlaciones de las estimaciones.

1	-0.34238	-0.13953	-0.32325	-0.07593	-0.30163	-0.63951
-0.34238	1	0.34362	0.00505	0.06983	-0.08866	0.03284
-0.13953	0.34362	1	-0.08289	-0.00206	-0.07888	0.01886
-0.32325	0.00505	-0.08289	1	-0.27164	0.08057	0.0015
-0.07593	0.06983	-0.00206	-0.27164	1	-0.04019	0.02488
-0.30163	-0.08866	-0.07888	0.08057	-0.04019	1	-0.38515
-0.63951	0.03284	0.01886	0.0015	0.02488	-0.38515	1

6. De la matriz de correlación, se calcula la matriz inversa solo para las variables (predictores), los elementos de la diagonal de la matriz inversa, corresponde a los FIV (factor de inflación de varianzas) para cada variable (predictor) del modelo.

	edad	alcoh	colest	trig	ps	pd
edad	1.14943	-0.39533	-0.07089	-0.09743	0.07259	0.0002
alcoh	-0.39533	1.14906	0.1097	0.06166	0.05311	0.01006
colest	-0.07089	0.1097	1.09862	0.30104	-0.09096	-0.04391
trig	-0.09743	0.06166	0.30104	1.08955	0.00696	-0.02284
ps	0.07259	0.05311	-0.09096	0.00696	1.19404	0.45646
pd	0.0002	0.01006	-0.04391	-0.02284	0.45646	1.17624

7. Factor de inflación de varianza (FIV).

FACTORES	**FIV**
EDAD	1.14943401
ALCOHOLISMO	1.14906320
COLESTEROL	1.09861806
TRIGLICERIDOS	1.08955147
PAS	1.19403767
PAD	1.17624151

Existe una baja multicolinealidad entre las variable lo cual es aceptable.

$(FIV < 3 \quad y \quad k > 1/3)$

Salida de Minitab

Variable	Valor	Conteo	
Enef	1	477	(Evento)
	0	807	
	Total	1284	

Tabla de desviaciones

Fuente	GL	Desv.ajust	Media.ajust	Chi-cuadrada	Valor p
Regresión	6	827.13	137.855	827.13	0.000
Edad	1	14.52	14.521	14.52	0.000
Colesterol	1	45.89	45.894	45.89	0.000
Triglicéridos	1	43.05	43.046	43.05	0.000
Presión arterial sistólica	1	213.37	213.368	213.37	0.000
Presión arterial diastólica	1	9.48	9.482	9.48	0.002
Alcoholismo	1	5.58	5.580	5.58	0.018
Error	1277	867.10	0.679		
Total	1283	1694.23			

Resumen del modelo

R-cuad. de desviación	R-cuad. (ajust) de desviación	AIC
48.82%	48.47%	881.10

Coeficientes

Término	Coef	EE del coef.	FIV
Constante	-18.04	1.22	
Edad	0.02506	0.00669	1.15
Colesterol	0.01234	0.00195	1.10
Triglicéridos	0.00756	0.00120	1.09
Presión arterial sistólica	0.08131	0.00677	1.19
Presión arterial diastólica	0.0501	0.0163	1.18
Alcoholismo	-0.441	0.187	1.15

Programa en Rstudio, que permite clasificar las observaciones por validación cruzada.

```
cv1GLM=function(dat)
+ {n=dim(dat)[1]
+ q=dim(dat)[2]
+ y=dat[,q]
+ dat=data.frame(dat[,-q],y)
+ pre=vector()
+
+ for(i in 1:n)
+ {dtest=dat[i,]
+ dtrain=dat[-i,]
+ mod=glm(y~.,dtrain,family=binomial)
+ pre[i]=predict(mod,dtest,type="response")
+ }
+ pre1=(pre>=0.5)*1
+ err=sum(pre1!=y)
+ te=err/n*100
+ ta=table(pre1,y)
+ list(Tabla=ta,Errores=err,Casos=n,TE=te)
+ }
> cv1GLM(matosu)
$Tabla
    y
pre1   0   1
   0 731 114
   1  76 363
$Errores
[1] 190
$Casos
[1] 1284
$TE
[1] 14.79751
```

===///===