UNIVERSIDAD CENTROCCIDENTAL "LISANDRO ALVARADO"
DECANATO DE CIENCIAS Y TECNOLOGÍA
DEPARTAMENTO DE INVESTIGACIÓN DE OPERACIONES Y ESTADÍSTICA

"Regresión Logística: un Ejemplo de Aplicación en Fonoaudiología"

Autora: Zuly Mary Briceño

TRABAJO DE ASCENSO

Presentado ante la Ilustre
Universidad de Centroccidental "Lisandro Alvarado"
como requisito para ascender a la categoría de Profesor Asociado

BARQUISIMETO, VENEZUELA
ABRIL, 2014

CONTENIDO

Resumen

El estudio estadísitico contenido en este trabajo formó parte de las actividades cumplidas en una pasantía de investigación desarrollada entre 2012 y 2013 en la Universidad Estadual de Maringá en el estado de Paraná (Brasil) y esta actividad de pasantía a su vez estuvo enmarcada en el programa de estudios doctorales en estadística, por parte de la autora, en la Universidad Central de Venezuela.

La intención fundamental es dar a conocer alguna de las aplicaciones del Análisis de Regresión Logística, el cual fue utilizado para resolver un problema en el área de la fonoaudiología.

El objetivo principal al construir los modelos de regresión logística fue estimar la prevalencia de factores de riesgo para las variables Ronquido Habitual, Riesgo Alto de Apnea y Riesgo Cardiovascular en la población adulta de Maringá-PR-Brasil.

Palabras Clave: Regresión Logística, Riesgo cardiovascular, Apnea obstructiva, Ronquido habitual.

Agradecimiento

A la Dra. Eniuce Menezes de Souza del Departamento de Estadística de la Universidad Estadual de Maringá (UEM), a la Lcda. Fabiana Southier Romano Avelar de la Maestría en Ciencias de la Salud (UEM), a la DFPA de la UCLA por el apoyo financiero para el desarrollo de los estudios doctorales en Caracas y las pasantías en Brasil, a la Universidad Central de Venezuela por la oportunidad para mi formación.

CAPÍTULO 1

Fundamentación y aspectos formales

En este capítulo se establecen aspectos formales de la investigación desarrollada, brevemente descrita en el resumen, tales como planteamiento del problema, justificación, metodología y objetivos. En el capítulo 2 se establecen elementos teóricos de la regresión logística y en el capítulo 3 se desarrolla la aplicación de la regresión logística en el problema de relacionar variables de naturaleza respiratoria con el riesgo cardiovascular en adultos de Maringá, en Brasil.

1.1. Diagnóstico situacional y planteamiento del problema

Datos reportados por diversos organismos como la Organización Mundial de la Salud en 2011 ([28]) muestran un alarmante índice mundial de ocurrencias fatales derivadas de las enfermedades cardiovasculares, constatando 17,3 millones de decesos por año, victimas de tales enfermedades. De estas victimas 80 % pertenecen a países de ingresos bajos y medios. En Brasil la condición se muestra igualmente grave en proporción, ya que ese número se aproxima a 300 mil por año, según datos de la Organización Panamericana de la Salud en 2007 ([29]).

El padecimiento de enfermedades cardiovasculares está asociado al sedentarismo, la obesidad central, la alimentación inadecuada, diabetes, dislipidemia, hipertensión, uso de alcohol y tabaco, además de características sociodemográficas y de historial médico familiar. Estas complicaciones de naturaleza comportamental y biológica acarrean varias patologías, entre ellas el SAOS (Síndrome de Apena Obstructiva del Sueño), altamente incidente en la reducción de la expectativa de vida, aumento de la morbilidad y mortalidad de la población, como es referido por Noal et. al. en [27] y [26]; así como también incide negativamente en las capacidades productivas de la población afectada, como lo refiere [34]. La necesidad de conocer las relaciones entre diversas de estas enfermedades con

ciertos desordenes y patologías del sueño ha sido resaltada por diversos autores, como lo evidencian [3] y [2] y las citas allí contenidas.

La interacción relevante entre el riesgo cardiovascular y la apnea obstructiva es acentuada en la medida en que la frecuencia de esta última genera, entre otros factores, arritmias cardíacas recurrentes en 58 % de los casos, vea [14], [15] y [41].

Esta complicación respiratoria se caracteriza por paradas del flujo aeronasal en las vías respiratorias superiores durante el sueño, debido a adherencias entre las paredes faríngeas con la consecuente caída en la concentración de oxígeno en la sangre, frecuentemente observada en la población adulta, [13].

Las Sociedades Brasileñas de Cardiología y de Otorrinilaringología informan [36], [5] [6] que esta enfermedad afecta a cerca de 4 en cada 100 hombres obesos y 2 en cada 100 mujeres en al misma condición. Aproximadamente 15 millones de brasileños sufren de esta enfermedad.

Entre otros síntomas que ocurren durante el sueño y que conducen al SAOS están la apnea obstructiva, el ronquido habitual y la somnolencia excesiva [11], [13]. El ronquido afecta al menos al 45 % de los adultos normales y se muestra habitual en el 25 % de los individuos, siendo más frecuente en hombres, personas obesas y empeora con la edad ([6], [31]). Estos síntomas deterioran la calidad de vida del individuo y a largo plazo se somatizan, generando un círculo vicioso en el que no se aprecia al principio la relación dañina con la salud cardiovascular, puesto que no siempre es clara la asociación adecuada entre ellos, pudiendo estar presentes una multiplicidad de factores importantes, vea [30].

Actualmente se reconoce que en Brasil son escasos los estudios de ámbito nacional que intenten esclarecer la relación entre el riesgo cardiovascular con la presencia de ronquido habitual y apnea obstructiva, particularmente en situaciones en las que se vincula el riesgo cardiovascular con la circunferencia cervical (propuesta por [37]), íntimamente relacionada con la obesidad como agravante de las condiciones de riesgo en general. Trabajos pioneros con amplias perspectivas han sido desarrollados apenas recientemente en São Paulo [38].

1.2. Justificación

Los cambios en los perfiles demográfico y epidemiológico de las poblaciones han tenido como consecuencia una mayor exposición de los individuos a los factores de riesgo relacionados con enfermedades crónicas no transmisibles, especialmente las cardiovasculares.

Con esto, la salud de la población brasileña viene presentando una importante transición en las condiciones de vida: la población tiene mayor edad, con nuevos patrones de trabajo y de esparcimiento, nuevos hábitos y estilos de vida [9] y [23]. Así se eleva la prevalencia de la obesidad, que por su parte genera otros daños como los respiratorios, entre varios otros, específicamente durante el sueño, destacándose el ronquido y la apnea obstructiva.

Es relevante entonces investigar acerca de la relación entre variables que representen el riesgo cardiovascular y la presencia de la apnea obstructiva asociada al ronquido habitual, que permita explicar una porción importante del perjuicio causado a la salud global, toda vez que al afectarse la configuración del sueño los individuos pueden hacerse más vulnerables a complicaciones de alcance cardiovascular, vea [10] para una descripción actual sobre el impacto en la salud de los trastornos del sueño. Partiendo de las informaciones obtenidas se podría alertar a los individuos portadores de estos disturbios respiratorios, como son la apnea del sueño y el ronquido habitual, sobre la posibilidad de procurar atención precoz en vista del impacto que pudiera tener para su salud y bienestar general. Tales planteamientos han sido hechos por organismos oficiales del sector salud del estado brasileño ([7]), y se requiere de estudios que caractericen la realidad en torno a estas enfermedades para la orientación toma de decisiones acertadas y oportunas en materia de prevención, diagnóstico y tratamiento tempranos.

1.3. Metodología y Materiales

Se trató de un estudio transversal, descriptivo analítico basado en una muestra representativa de 413 adultos con edades comprendidas entre 20 y 59, residentes en el Municipio de Maringá, Estado de Paraná (PR), Brasil en el 2012.

Fue seleccionada una muestra representativa por conglomerados. La unidad muestral primaria fue representada por los sectores censitarios del municipio, delimitados por el censo realizado en el año 2000 por el Instituto Brasileiro de Geografía e Estatística (IBGE).

La recolección de los datos se realizó mediante la aplicación de una versión complementada del Cuestionario de Berlin [39], ya probado y validado en estudios anteriores realizados en Brasil, relacionados con investigaciones de la condición cardiovascular y de salud respiratoria, ronquido, somnolencia y calidad del sueño. Los datos fueron recolectados en cada domicilio con un periodo de tiempo de duración de aproximadamente 20 minutos. Con el objeto de disminuir errores en la base de datos, durante el periodo de recolección de los datos, los cuestionarios se revisaron periódicamente a fin de detectar

fallas de llenado.

El desarrollo del estudio ocurrió conforme con los preceptos éticos y fue respetada la resolución 196/96 del Conselho Nacional de Saúde, habiendo sido aprobado el proyecto por el Comitê Permanente de Ética em Pesquisa com Seres Humanos (COPEP) de la Universidad Estadual de Maringá (UEM).

El tratamiento de los datos consistió principalmente en la construcción de modelos de regresión logística, siendo este un procedimiento de uso frecuente y de elevada efectividad en estudios e investigaciones de naturaleza médica, como es ampliamente conocido y ha sido reseñado en la literatura desde hace mucho tiempo (vea [4], [22] como referencias clásicas), siendo usado también en la actualidad en estudios de gran alcance, particularmente sobre el síndrome de apnea del sueño como [38]. Desarrollos generales sobre el uso de regresión logística en ciencias médicas y biológicas se encuentran en [40] y [21], entre otros.

El análisis de los datos se efectuó con el software SPSS v19 (IBM SPSS Statistics for Windows, Version 19.0, IBM Corp. Released 2010) con licencia de uso restringido del Laboratório de Estatística Aplicada (LEA) del Departamento de Estadística de la Facultad de Ciencias y Tecnología de la Universidad Estadual Paulista "Júlio de Mesquita Filho"(UNESP), en su Câmpus de Presidente Prudente en el interior de São Paulo. Cabe destacar que, por su cercanía geográfica, los Departamentos de Estadística de la UEM y de la UNESP mantienen estrecha colaboración y tienen miembros comunes en su planta profesoral de postrgrado, lo que permitió el uso de los recursos del Laboratorio de Estadística Aplicada en esta investigación, como se ha hecho de manera regular en muchas otras.

1.4. Objetivos

El propósito de este trabajo es mostrar una aplicación de la regresión logística en el problema de establecer asociaciones entre el ronquido y la apnea del sueño como factores del riesgo cardiovascular, como fue descrito el inicio de este capítulo introductorio. La aplicación estadística se enmarca como parte de una investigación del área de fonoaudiología, desarrollada por un equipo de investigación en la Universidad Estadual de Maringá en Brasil, entre 2013 y 2014.

El estudio se planteó, como objetivo general, la estimación de la prevalencia de factores de riesgo para enfermedades cardiovasculares y la presencia de ronquido habitual

y apnea obstructiva en la población adulta de la ciudad de Maringá, en el estado de Paraná, en Brasil.

El objetivo específico asociado fue la derminación del grado de simultaneidad entre el Ronquido Habitual, la Apnea Obstructiva autoreferida (por los individuos) y el Elevado Riesgo Cardiovascular en la población adulta de Maringá-PR-Brasil.

CAPÍTULO 2

Modelo de Regresión Logística

2.1. Introducción

En este capítulo se presentan de manera breve y a modo de fundamento teórico, los elementos básicos de la teoría de regresión logística. Para mayores especificaciones sobre le tema, se refiere al lector a fuentes tales como [32], [21], [20], [17], [12] y [8], solo para mencionar algunas de las más comunes y de amplia difusión, que fueron empleadas en la elaboración de este trabajo.

En el artículo [18] se introducen y establecen algunos de los procedimientos más importantes de la regresión logística utilizada en el presente trabajo, como la prueba de Hosmer y Lemeshow; por otro lado la en referencia [19] se discute el carácter de adecuación de los métodos de regresión logística como herramienta de la inferencia estadística para datos binarios.

Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre:

- Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).
- Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial, si bien su transformación logarítmica (*logit*) permite su uso como una función lineal.

Como se ve, las covariables pueden ser cuantitativas o cualitativas. Las covariables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia (esta codificación es importante, ya que cualquier otra codificación provocaría modificaciones en la interpretación del modelo). Pero si la covariable cualitativa tuviera

más de dos categorías, para su inclusión en el modelo debería realizarse una transformación de la misma en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas variables *dummy*), de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual. En general, si la covariable cualitativa posee n categorías, habrá que realizar $n-1$ covariables ficticias.

Por sus características, los modelos de regresión logística permiten dos finalidades:

1. Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, conocer la *odds ratio* para cada covariable).

2. Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

No cabe duda que la regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización en esas áreas.

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

Descripción de la regresión logística

La regresión logística es un instrumento estadístico de análisis bivariado o multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de m variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables ficticias o simuladas ("*dummy*").

El propósito del análisis es:

1. Predecir la probabilidad de que a alguien le ocurra cierto evento: por ejemplo, "estar desempleado" = 1 o "no estarlo" = 0; "ser pobre" = 1 o "no ser pobre" = 0; "graduarse como sociólogo" = 1 o "no graduarse" = 0;

2. Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión.

Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos. Por ejemplo, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) los sujetos que están efectivamente desocupados (= 1) y los que no lo están (= 0). En base a ello, predecirá a cada uno de los sujetos – independientemente de su estado real y actual – una determinada probabilidad de ser desocupado (es decir, de tener valor 1 en la variable dependiente). Es decir, si alguien es un joven no amo de casa, con baja educación y de sexo masculino y origen emigrante (aunque esté ocupado) el modelo le predecirá una alta probabilidad de estar desocupado (puesto que la tasa de desempleo del grupo así definido es alta), generando una variable con esas probabilidades estimadas. Y procederá a clasificarlo como desocupado en una nueva variable, que será el resultado de la predicción. Además, analizará cuál es el peso de cada una de estas variables independientes en el aumento o la disminución de esa probabilidad. Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de ser desocupado. En cambio, cuando el sexo pase de 0 = "mujer" a 1 = "varón", aumentará en algo la probabilidad de desempleo porque la tasa de desempleo de los jóvenes de sexo masculino es mayor que la de las mujeres jóvenes. El modelo, obviamente, estima los coeficientes de tales cambios.

Cuanto más coincidan los estados pronosticados con los estados reales de los sujetos, mejor ajustará el modelo. Uno de los primeros indicadores de importancia para apreciar el ajuste del modelo logístico es el doble logaritmo del estadístico de verosimilitud (*likelihood*). Se trata de un estadístico que sigue una distribución similar a χ^2 y compara los valores de la predicción con los valores observados en dos momentos: (a) en el modelo sin variables independientes, sólo con la constante y (b) una vez introducidas las variables predictoras. Por lo tanto, el valor de la verosimilitud debiera disminuir sensiblemente entre ambas instancias e, idealmente, tender a cero cuando el modelo predice bien.

Regresión logística binaria

Los modelos de regresión logística binaria resultan los de mayor interés ya que la mayor parte de las circunstancias analizadas en medicina responden a este modelo (presencia o no de enfermedad, éxito o fracaso, etc.). Como se ha visto, la variable dependiente será una variable dicotómica que se codificará como 0 o 1 (respectivamente, "ausencia" y "presencia"). Este aspecto de la codificación de las variables no es banal (influye en la forma en que se realizan los cálculos matemáticos), y habrá que tenerlo muy en cuenta si se emplean paquetes estadísticos que no recodifican automáticamente las variables cuando éstas se encuentran codificadas de forma diferente (por ejemplo, el uso frecuente de 1 para la presencia y -1 o 2 para la ausencia).

La ecuación de partida, que se denomina distribución logística, en los modelos de regresión logística es:

$$\Pr(y=1|x) = \frac{\exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}{1 + \exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)} \tag{1.3b}$$

donde:

$\Pr(y=1\|X)$	es la probabilidad de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X;
X	es un conjunto de n covariables $\{x_1, x_1, \ldots, x_n\}$ que forman parte del modelo;
b_0	es la constante del modelo o término independiente;
b_i	los coeficientes de las covariables.

Si se divide la expresión (1.3b) por su complementario, es decir, si se construye su *odds* (en el ejemplo de presencia o no de enfermedad, la probabilidad de estar enfermo entre la probabilidad de estar sano), se obtiene una expresión de manejo matemático más fácil:

$$\frac{\Pr(y=1|x)}{1-\Pr(y=1|x)} = \exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right) \tag{2.4}$$

Pero esta expresión aún es difícil de interpretar.

Si ahora se realiza su transformación logaritmo natural, se obtiene una ecuación

lineal que lógicamente es de manejo matemático aún más fácil y de mayor comprensión:

$$\log\left(\frac{\Pr(y=1|x)}{1-\Pr(y=1|x)}\right) = b_0 + \sum_{i=1}^{n} b_i x_i \tag{2.5}$$

o simplificando:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_i x_i \tag{1.6a}$$

En la expresión (2.5) se ve a la izquierda de la igualdad el llamado *logit*, es decir, el logaritmo natural de la *odds* de la variable dependiente (esto es, el logaritmo de la razón de proporciones de enfermar, de fallecer, de éxito, etc.). El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \tag{2.7}$$

La regresión lineal presenta una diferencia fundamental respecto al modelo de regresión logística. En el modelo de regresión lineal se asume que los errores estándar de cada coeficiente siguen una distribución normal de media 0 y varianza constante (homoscedasticidad). En el caso del modelo de regresión logística no pueden realizarse estas asunciones pues la variable dependiente no es continua (sólo puede tomar dos valores, 0 o 1, pero ningún valor intermedio). Llamando ϵ al posible error de predicción para cada covariable x_i se tendrá que el error cometido dependerá del valor que llegue a tomar la variable dependiente, tal como se ve en (2.8).

$$y = \Pr(x) + \epsilon \qquad \begin{cases} y = 1 \implies \epsilon = 1 - \Pr(x) \\ y = 0 \implies \epsilon = -\Pr(x) \end{cases} \tag{2.8}$$

Esto implica que ϵ sigue una distribución binomial, con media y varianza proporcionales al tamaño muestral y a $\Pr(y=1|x_i)$ (la probabilidad de que $y=1$ dada la presencia de x_i).

2.2. Requisitos y etapas de la regresión logística

- Recodificar las variables independientes categóricas u ordinales en variables ficticias o simuladas y de la variable dependientes en 0 y 1;
- Evaluar efectos de confusión y de interacción del modelo explicativo;

- Evaluar la bondad de ajuste de los modelos;
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponenciales y estadísticos de prueba (Wald).

Estimación de los coeficientes del modelo y de sus errores estándar

Para la estimación de los coeficientes del modelo y de sus errores estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que hagan máxima la probabilidad de obtener los valores de la variable dependiente Y proporcionados por los datos de nuestra muestra. Estas estimaciones no son de cálculo directo, como ocurre en el caso de las estimaciones de los coeficientes de regresión de la regresión lineal múltiple por el método de los mínimos cuadrados. Para el cálculo de estimaciones máximo-verosímiles se recurre a métodos iterativos, como el método de Newton-Raphson. Dado que el cálculo es complejo, normalmente hay que recurrir al uso de rutinas de programación o a paquetes estadísticos. De estos métodos surgen no sólo las estimaciones de los coeficientes de regresión, sino también de sus errores estándar y de las covarianzas entre las covariables del modelo.

El siguiente paso será comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo. Para ello se pueden emplear básicamente tres métodos: el estadístico de Wald, el estadístico G de razón de verosimilitud y la prueba Score.

El estadístico de Wald

Contrasta la hipótesis de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza 1. Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. La obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald puede proporcionar falsas ausencias de significación (es decir, se incrementa el error tipo II). Tampoco es recomendable su uso si se están empleando variables de diseño.

El estadístico G de razón de verosimilitud

Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico G sigue una χ^2 con un grado de libertad (no se asume normalidad). La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo. Esta prueba no asume ninguna distribución concreta, por lo que es la más recomendada para estudiar la significación de los coeficientes.

La prueba Score

Su cálculo para el caso de una única variable viene dado por:

$$S = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{2.9}$$

En el caso de múltiples covariables hay que utilizar cálculo matricial, si bien no requiere un cálculo iterativo (precisamente su rapidez de cálculo sería su aspecto más favorable). En contra del mismo dos aspectos:

(a) Se sabe que este estadístico se incrementa conforme aumenta el número de covariables (es decir tiende a dar significación con mayor frecuencia).

(b) Este estadístico también asume una distribución normal con media 0 y varianza 1.

Al igual que en los casos anteriores, si alcanza significación indica que la covariable debería permanecer en el modelo. Su uso en algunos paquetes estadísticos ha quedado relegado a la selección de variables en métodos paso a paso (por la mayor rapidez de cálculo).

Cuando la covariable es cualitativa con n categorías (siendo $n > 2$), en el modelo se analizará la significación de cada una de sus $n-1$ variables ficticias, así como la significación global de la covariable comparando la presencia en bloque frente a la ausencia en bloque de sus $n-1$ covariables ficticias.

Una vez se ha estimado los coeficientes de regresión y sus correspondientes errores estándar se calcularán los correspondientes intervalos de confianza para las estimaciones,

bajo la hipótesis de que dichos coeficientes se distribuyen según respectivas distribuciones normales. Para un determinado coeficiente, su intervalo de confianza al 95 % vendrá dado por:

$$\begin{aligned} IC\,95\,\%(\beta) &= [(\beta - 1,96ee), (\beta + 1,96ee)] \\ IC\,95\,\%(OR) &= [\exp(\beta - 1,96ee), \exp(\beta + 1,96ee)] \end{aligned} \tag{2.10}$$

Junto a la significación del estadístico empleado para contrastar la significación de los coeficientes de regresión, la inclusión de la unidad en el intervalo de confianza es, lógicamente, indicativa de la ausencia de significación.

En ocasiones existirán modelos que llaman la atención por la falta de sentido de sus estimaciones. Esta sorpresa suele venir dada por la presencia de estimaciones de grandes errores estándar, con frecuencia asociadas a estimaciones de coeficientes de regresión también anormalmente elevados. Las posibles causas de este hecho pueden ser:

(a) Presencia de una frecuencia cero en una tabla de contingencia $Y \times X$. Cuando esto ocurre provoca en el cálculo de la correspondiente *odds* la presencia de un 0 en el denominador (y por tanto no es calculable). Si esta covariable se intenta introducir en el modelo de regresión que se está diseñando, el software puede comportarse de forma incorrecta: desde excluirla por entender que predice perfectamente la variable dependiente, a incluirla y comunicar un error (porque la rutina de iteración para el cálculo de estimaciones de máxima verosimilitud o bien no llega a converger o bien llega al máximo de iteraciones prefijadas). Esta circunstancia puede y debe ser detectada durante el análisis univariado. En el caso de tratarse de una variable cualitativa con más de dos categorías, una solución es colapsar dos de esas categorías.

(b) También puede ocurrir que se incluyan interacciones que impliquen una excesiva estratificación para la muestra disponible. El resultado puede ser una estimación elevada del correspondiente coeficiente de regresión y de su error estándar. En el análisis univariado, al realizar efectivamente las dos tablas de contingencia de la estratificación, se observará que alguna de las ocho celdas contiene el cero. Si no puede recurrir al colapso de categorías, puede decidirse diseñar una nueva variable que sea la combinación de las dos covariables con sus correspondientes categorías, e incluirla como tal en el modelo.

Presencia de una o más covariables que discriminan perfectamente las dos categorías de la variable dependiente. Algunos ejemplos servirán para explicar esta circunstancia: Si siempre que se administran antimicrobianos los sujetos con una determinada

enfermedad infecciosa viven y siempre que no se administran mueren, la covariable "antimicrobianos" discrimina perfectamente a la variable "muerte"; o si siempre que se tienen más de 65 años se padece de cardiopatía isquémica y por debajo no, la covariable "edad" discrimina perfectamente a la variable "cardiopatía isquémica". En la práctica esta circunstancia impide que se puedan realizar estimaciones de coeficientes por máxima verosimilitud, lo que no quiere decir que el paquete estadístico necesariamente no de falsas estimaciones, como en el punto anterior.

Este problema está en estrecha relación con el tamaño muestral y el número de covariables que se desean introducir en el modelo: la probabilidad de discriminación completa es elevada en los modelos con muestras con tamaños muestrales pequeños, sobre todo cuando una de las categorías de la variable dependiente está poco representada, y tanto más cuanto mayor es el número de covariables introducidas en el modelo.

(c) Multicolinealidad. Si bien existen pruebas que permiten comprobar la existencia de colinealidad entre covariables (que se verá más adelante), cabe reseñar aquí que al igual que en los casos anteriores, los modelos con multicolinealidad entre las covariables introducidas llamarán la atención por la presencia de grandes errores estándar, y frecuentemente, estimaciones de coeficientes anormalmente elevadas. Sin embargo la multicolinealidad no afecta al sentido de las estimaciones (la multicolinealidad no hará que aparezca significación donde no la hay, y viceversa).

2.3. Multicolinealidad

Se dice que existe multicolinealidad cuando dos o más de las covariables del modelo mantienen una relación lineal. Cuando la colinealidad es perfecta, es decir, cuando una covariable puede determinarse según una ecuación lineal de una o más de las restantes covariables, es posible estimar un único coeficiente de todas las covariables implicadas. En estos casos debe eliminarse la covariable que actúa como dependiente.

Normalmente lo que se hallará será una multicolinealidad moderada, es decir, una mínima correlación entre covariables. Si esta correlación fuera de mayor importancia, su efecto sería, como ya se vio anteriormente, el incremento exagerado de los errores estándar, y en ocasiones, del valor estimado para los coeficientes de regresión, lo que hace las estimaciones poco creíbles.

Un primer paso para analizar este aspecto puede ser examinar la matriz de coeficien-

tes de correlación entre las covariables. Coeficientes de correlación muy elevados llevarán a investigar con mayor profundidad. Sin embargo, este método, bueno para detectar colinealidad entre dos covariables, puede conducir a no poder detectar multicolinealidad entre más de dos de ellas.

Existen otros procedimientos analíticos para detectar multicolinealidad. Puede desentenderse por el momento de la variable dependiente y realizar sendos modelos en los que una de las covariables actuará como variable dependiente y las restantes covariables como variables independientes de aquella. A cada uno de estos modelos se le puede calcular la R^2 (o dispersión total, medida de ajuste que se verá más adelante). Se denomina tolerancia al complementario de R^2, $(1 - R^2)$, y factor de inflación de la varianza (FIV) al inverso de la tolerancia, $1/(1 - R^2)$. Cuando existe estrecha relación entre covariables la tolerancia tiende a ser 0, y por tanto FIV tiende al infinito. Como regla general debería preocupar tolerancias menores de $0,1$ y FIV mayores de 10. SPSS ofrece la matriz de correlaciones, pero no aporta índices de multicolinealidad para la regresión logística.

La solución a la multicolinealidad no es fácil:

- Puede intentarse eliminar la variable menos necesaria implicada en la colinealidad, a riesgo de obtener un modelo menos válido;
- Puede intentar cambiarse la escala de medida de la variable en conflicto (es decir, transformarla), para evitar sacarla del modelo, si bien no siempre se encontrará una transformación de forma directa. Algunas transformaciones frecuentes son el centrado respecto de la media, la estandarización o la creación de variables sintéticas mediante un análisis previo de componentes principales (que es otro tipo de análisis multivariado). Estas transformaciones por el contrario hacen al modelo muy dependiente de los datos actuales, invalidando su capacidad predictiva;
- También se puede recurrir a aumentar la muestra para así aumentar la información en el modelo, lo que no siempre será posible.

Las variables simuladas (*dummy*)

A veces se necesita incorporar al modelo de regresión logística variables independientes que no son numéricas sino categóricas. Supóngase, por ejemplo, que se quiere predecir la probabilidad de una persona de “ser pobre”.

Tal vez resulte importante incorporar variables que no son cuantitativas: por ejemplo, la categoría ocupacional ("empleador", "trabajador por cuenta propia", "asalariado", "trabajador sin remuneración"). En este caso, esta variable podría ser incorporada a la ecuación si se la transforma en una variable simulada. Ello consiste en generar $n-1$ variables dicotómicas con valores cero y uno, siendo n el número de categorías de la variable original.

Se crearían tres variables dicotómicas: la primera de ellas sería "empleador". Quien lo sea tendrá valor 1 en esa variable y valor cero en las variables "cuenta propia" y "asalariado". Los "por cuenta propia" tendrán valor 1 en la segunda variable y cero en las otras, etc. No se necesita crear, en cambio, una variable llamada "trabajador sin remuneración": lo será quien tenga valores cero en las tres anteriores. Esta última es la categoría "base" de las variables simuladas.

Una vez realizada esta transformación, estas variables pueden ser incorporadas en una ecuación de regresión: sus valores sólo pueden variar entre cero y uno y sus coeficientes b indicarán, en cada caso, cuanto aumentan o disminuyen los "*odds*" de probabilidad del evento que se procura predecir cuando una de estas variables pasa de cero a uno (por ejemplo, cuando alguien es un empleador, seguramente la probabilidad de que sea pobre disminuirá, lo que se expresará en un coeficiente b negativo en la ecuación logística).

Función de verosimilitud

Se sabe que cualquier variable dependiente de otra u otras variables, toma valores según los valores de las variables de las que depende. Por otra parte, esa variable dependiente irá tomando valores siguiendo o describiendo una determinada distribución de frecuencias; es decir, tomen los valores que tomen las variables independientes, si el experimento se repite múltiples veces, la variable dependiente tomará para esos valores de las independientes un determinado valor, y la probabilidad de ocurrencia de dicho valor vendrá dado por una distribución de frecuencias concreta: una distribución normal, una distribución binomial, una distribución hipergeométrica, etc. En el caso de una variable dependiente dicotómica (como el caso que nos ocupa), la distribución de frecuencias que seguirá será la binomial, que depende de la tasa de éxitos (X sujetos de un total de N, que sería el elemento variable), para un determinado tamaño muestral N y probabilidad $\Pr(\cdot)$ de ocurrencia del evento valorado por la variable dependiente (parámetros constantes). La función de densidad de esta distribución de frecuencias vendrá dada por la siguiente

expresión:

$$\Pr(y) \approx f(x) = \binom{N}{x} p^x (1-p)^{N-x} \tag{2.11}$$

Si en la expresión anterior introducimos los datos concretos de nuestra muestra de N sujetos (es decir, se convierte el elemento variable X en parámetro), y se hace depender el resultado de la función de densidad del parámetro "probabilidad de ocurrencia" (p, que de esta forma se convierte en variable), se está generando su función de verosimilitud, $f(p|x)$ (función dependiente de p dado el valor muestral de x) o $L(P)$ (L de *likelihood*), que ofrece como resultados las probabilidades de la función de densidad ajustada a los datos:

$$f(p|x) = \binom{N}{x} p^x (1-p)^{N-x} \tag{2.12}$$

Se deduce que, para una muestra concreta, esa probabilidad será diferente según qué valores tome el parámetro "probabilidad de ocurrencia":

Se demuestra que la mejor estimación del parámetro $\widehat{e}$ es aquel valor que maximiza esta función de verosimilitud, ya que son estimadores consistentes (conforme crece el tamaño muestral, la estimación se aproxima al parámetro desconocido), suficientes (aprovechan la información de toda la muestra), asintóticamente normales y asintóticamente eficientes (con mínima varianza), si bien no siempre son insesgados (no siempre la media de las estimaciones para diferentes muestras tenderá hacia el parámetro desconocido).

2.4. Método de Newton-Raphson

Se trata de un método iterativo, empleado en diversos problemas matemáticos, como en la determinación de las raíces de ecuaciones, y en el presente caso, en la estimación de los coeficientes de regresión β por el procedimiento de máxima verosimilitud.

Por facilidad de cálculo toda la formulación se expresará en forma de matrices. Las particularidades del cálculo matricial escapan del ámbito de este documento. Téngase presente la base de datos (una tabla con filas y columnas). Se dispondrá de:

- Una variable Y, que es la variable dependiente. Expresada como matriz será una matriz de N filas y una columna, cuyo contenido será de ceros y unos (ya que se

trata de una variable dicotómica).

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} \tag{2.13}$$

- Un conjunto de M covariables, que pueden expresarse como una matriz de N filas y M columnas. Sin embargo, dado que el modelo contiene una constante, ésta se puede expresar como una columna adicional en la que todos sus elementos son "1". Por tanto la matriz X queda como una matriz con N filas y $(M+1)$ columnas, de la forma:

$$X = \begin{pmatrix} 1 & x_{1,2} & \cdots & x_{1,m+1} \\ 1 & x_{2,2} & \cdots & x_{2,m+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,2} & \cdots & x_{n,m+1} \end{pmatrix} \tag{2.14}$$

- Y por último un conjunto de coeficientes de regresión β, uno para cada covariable, incluida la covariable creada para la constante, con una fila y $(M+1)$ columnas

$$\beta = (\beta_1, \beta_2, \ldots, \beta_{m+1}) \tag{2.15}$$

El proceso se inicia construyendo la función de verosimilitud (*likelihood function*) de la ecuación de regresión logística:

$$L(\beta) = p_i^{\sum y_i}(1-p_i)^{(N-\sum y_i)} \tag{2.16}$$

o mejor, su transformación logarítmica (*log likelihood*):

$$LL(\beta) = \sum y_i \ln(p_i) + (N - \sum y_i)\ln(1-p_i) \tag{2.17}$$

donde p_i es la probabilidad de ocurrencia de $y = 1$ con los valores muestrales de las covariables $X = \{x_1, x_2, \ldots, x_{m+1}\}$, para el sujeto $i = \{1, 2, \ldots, N\}$. El valor $2LL(\beta)$ se llama devianza y mide en qué grado el modelo se ajusta a los datos: cuanto menor sea su valor, mejor es el ajuste.

Se trata de conocer aquellos valores de β que hacen máxima la función de verosimilitud (o su logaritmo). Se sabe que si se iguala a cero la derivada parcial de una función

respecto a un parámetro, el resultado es unos valores de dicho parámetro que hacen llevar a la función a un valor máximo o un valor mínimo (un punto de inflexión de la curva). Para confirmar que se trata de un máximo y no de un mínimo, la segunda derivada de la función respecto a dicho parámetro debe ser menor de cero.

La primera derivada de $LL(\beta)$ respecto de β (llamada función *score*) en su forma matricial es:

$$U(\beta) = \frac{\partial LL(\beta)}{\partial \beta} = X'(Y - p) \tag{2.18}$$

donde:

p es una matriz de N filas y una columna que contiene las probabilidades de cada individuo de que tengan su correspondiente evento y_i

La segunda derivada, llamada matriz informativa o hessiana, es:

$$H(\beta) = \frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta} = -\mathbf{X'WX} \tag{2.19}$$

donde:

$\mathbf{W}$ es una matriz diagonal de n filas y n columnas, en la que los elementos de su diagonal vienen dados por los respectivos productos $p_i(1 - p_i)$, de manera que $\mathbf{W}$ queda de la forma siguiente:

$$\mathbf{W} = \begin{pmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n(1-p_n) \end{pmatrix} \tag{2.20}$$

y para cada fila su p_i es:

$$p_i = \frac{1}{1 + \exp\left(-\sum_{j=1}^{m+1} \beta_j x_{ij}\right)} \tag{2.21}$$

Una vez se dispone todos los elementos necesarios, se procede a explicar como tal el método iterativo para la determinación de los coeficientes de regresión.

Se le asigna un valor inicial empírico a los coeficientes de regresión, en general cero a todos ellos.

En cada iteración t la matriz de nuevos coeficientes de regresión experimentales resulta de sumar matricialmente un gradiente a la matriz de coeficientes experimentales

del paso anterior. Este gradiente es el resultado del cociente entre la primera derivada y la segunda derivada de la función de verosimilitud de la ecuación de regresión.

$$\widehat{\beta}_t = \widehat{\beta}_{t-1} + (\mathbf{X}'\mathbf{W}_{t-1}\mathbf{X})'\mathbf{X}'(\mathbf{Y} - p_{t-1}) \tag{2.22}$$

El segundo paso se repite tantas veces como sea necesario hasta que la diferencia entre la matriz de coeficientes de regresión en dicha iteración y la matriz de la iteración previa, sea 0 o prácticamente 0 (por ejemplo $< 10^{-6}$). Los paquetes estadísticos suelen tener un límite de iteraciones que pueden modificarse si no se obtiene convergencia inicialmente. SPSS® tiene además otras condiciones de parada:

- $LL(\beta)$ muy cercana a cero;
- Diferencia entre $LL(\beta)$ de dos iteraciones consecutivas muy cercana a cero.

Una vez finalizadas las iteraciones, la inversa de la matriz informativa de la última iteración ofrece los valores de varianzas y covarianzas de las estimaciones de los coeficientes de regresión estimados. En concreto, el error estándar de cada coeficiente de regresión coincide con la raíz cuadrada del elemento respectivo de la diagonal principal (es decir el elemento $(1,1)$ sería el cuadrado del error estándar del coeficiente β_1, el elemento $(2,2)$ el cuadrado del error estándar del coeficiente β_2, y así sucesivamente). Por debajo de esta diagonal quedan las covarianzas de cada pareja de covariables (es decir, el elemento $(2,1)$ es la covarianza de β_1 y β_2, el elemento $(3,2)$ es la covarianza de β_2 y β_3, etc.). Hay programas estadísticos que ofrecen esta matriz de varianzas y covarianzas; SPSS® no lo hace, sino que ofrece la matriz de correlaciones. En ese caso se podrá calcular la matriz de varianzas y covarianzas sabiendo que la covarianza de dos variables es igual al producto del coeficiente de correlación de ambas (r) y los dos respectivos errores estándar:

$$\text{cov}(\beta_1, \beta_2) = r(\beta_1, \beta_2)ee(\beta_1)ee(\beta_2) \tag{2.23}$$

Entender esta formulación y el algoritmo de las iteraciones puede ser de gran utilidad, pues con conocimientos básicos de programación facilita el desarrollo de rutinas propias, por ejemplo en $VisualBasic$® dentro de una base de datos de $Access$®, que pueden liberar de la dependencia de costosos paquetes estadísticos.

Odds ratio: Es un cociente de proporciones de enfermos por cada sano entre el grupo con un factor de riesgo y el grupo sin dicho factor de riesgo.

En este caso, entre los que tienen el factor de riesgo hay 20 enfermos por cada 80 sanos $(0,25)$, y entre los que no tienen el factor de riesgo hay 30 enfermos por cada 270

sanos $(0, 11)$, por lo que las personas con el factor de riesgo tienen un riesgo de enfermar $2, 25$ veces superior $(0, 25/0, 11)$ que las personas sin el factor de riesgo.

Principio jerárquico: siempre que se incluya en el modelo un término de interacción, el modelo debe incluir también todos los términos de orden inferior, y si el término de interacción resultase significativo y permaneciese en el modelo, también deberían permanecer los términos de orden inferior, aunque no se lograra demostrar significación para ellos.

Modelo con interacción de primer orden:

$$y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$$

Modelo con interacción de segundo orden:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_1x_3 + b_6x_2x_3 + b_7x_1x_2x_3$$

Principio de parsimonia: En igualdad de condiciones la solución más sencilla que explique completamente un problema es probablemente la correcta (Guillermo de Ockham). Según este principio, cuando más de un modelo se ajuste a las observaciones, se retendrá el modelo más simple que explique dichas observaciones con un grado adecuado de precisión.

2.5. Modelos de regresión logística

Modelo logístico univariante simple

$$y = \frac{\exp(a + bx)}{1 + \exp(a + bx)}$$

Modelo logístico univariante múltiple

$$y = \frac{\exp(\beta_{i,0} + \beta_1\, x_1 + \beta_{i,2}\, x_2 + \cdots + \beta_{i,k}\, x_k)}{1 + \exp(\beta_{i,0} + \beta_1\, x_1 + \beta_{i,2}\, x_2 + \cdots + \beta_{i,k}\, x_k)}$$

Modelo logístico multivariante simple

$$y_i = \frac{\exp(a + bx)}{1 + \exp(a + bx)}$$

Modelo logístico multivariante múltiple

$$y_i = \frac{\exp(\beta_{i,0} + \beta_1\, x_1 + \beta_{i,2}\, x_2 + \cdots + \beta_{i,k}\, x_k)}{1 + \exp(\beta_{i,0} + \beta_1\, x_1 + \beta_{i,2}\, x_2 + \cdots + \beta_{i,k}\, x_k)}$$

Interpretación del modelo logístico

Los parámetros del modelo son: β_0, la ordenada en el origen, y $\boldsymbol{\beta_i} = \{\beta_1, \beta_2, \ldots, \beta_k\}$. A veces, se utilizan también como parámetros $\exp(\beta_0)$ y $\exp(\boldsymbol{\beta_i})$, que se denominan *odds ratios* o razón de probabilidades. Estos valores indican cuánto se modifican las probabilidades por unidad de cambio en las variables x. De (1.6a) se deduce que:

$$O_i = \frac{p_i}{1-p_i} = \exp(\beta_0) \prod_{j=1}^{k} \exp(\beta_j)^{x_j}$$

Supóngase que dos elementos tienen valores iguales en todas las variables menos en una.

Sean $(x_{i,1}, x_{i,2}, \ldots, x_{i,h}, \ldots, x_{i,k})$ los valores de las variables para el primer elemento y $(x_{j,1}, x_{j,2}, \ldots, x_{j,h}, \ldots, x_{j,k})$ para el segundo, y todas las variables son las mismas en ambos elementos menos en la variable h donde $x_{i,h} = x_{j,h} + 1$. Entonces, el *odds ratio* para estas dos observaciones es:

$$\frac{O_i}{O_j} = \exp(\beta_h)$$

e indica cuánto se modifica el *ratio* de probabilidades cuando la variable x_j aumenta en una unidad.

Si se considera $p_i = 0,5$ en el modelo *logit*, entonces

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} = 0$$

es decir,

$$x_{1,i} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{k} \frac{\beta_j x_{i,j}}{\beta_1}$$

donde $x_{1,i}$ representa el valor de x_1 que hace igualmente probable que un elemento cuyas restantes variables son $(x_{2,i}, \ldots, x_{k,i})$ pertenezca a la primera o la segunda población.

Medidas de confiabilidad del modelo

Devianza Es similar a la suma de cuadrados del error de la regresión lineal y se define como:

$$D = -2 \sum_{i=1}^{n} \left(y_i \log \left(\frac{\widehat{p}}{y_i} \right) + (1 - y_i) \log \frac{1-\widehat{p}}{1-y_i} \right)$$

Si D es mayor que una χ^2 con $n-p$ grados de libertad para un nivel de significación dado entonces el modelo logístico es confiable.

Criterio AIC de Akaike Se define como

$$AIC = D + 2(p + 1)$$

donde p es el número de variables predictoras.

Prueba de bondad de ajuste de Hosmer-Lemeshov Se define como

$$c = \sum_{i=1}^{g} \frac{(O_i - n_i' \overline{p}_i)^2}{n_i' \overline{p}_i (1 - \overline{p}_i)}$$

donde

g	es el número de grupos;
n_i'	es el número de observaciones en el i-ésimo grupo;
O_i	es la suma de las Y en el i-ésimo grupo; y
$\overline{p}_i$	es el promedio de las p_i en el i-ésimo grupo.

2.6. Estadísticos influenciales para regresión logística

Existen varios tipos de residuales que permiten cotejar si una observación es influencial o no.

Residuales de Pearson

Definidos como:

$$r_i = \frac{y_i - m_i \widehat{p}_i}{\sqrt{m_i \widehat{p}_i (1 - \widehat{p}_i)}}$$

donde y_i representa el número de veces que $y = 1$ entre las m_i repeticiones de X_i si los valores de la variable de respuesta están agrupadas.

El residual de Pearson es similar al residual estudentizado usado en regresión lineal. Así, un residual de Pearson mayor que 2 indica un dato anormal.

Si el modelo es correcto, los residuales de Pearson serán variables de media cero y varianza unidad que pueden servir para hacer el diagnóstico de dicho modelo. El estadístico $\chi_0^2 = \sum_{i=1}^{k} e_1^2$ permite realizar un contraste global de la bondad del ajuste. Se distribuye asintóticamente como una χ^2 con $(n - k - 1)$ grados de libertad, donde $k + 1$ es el número de parámetros en el modelo.

En lugar de los residuos de Pearson se pueden utilizar, también, las desviaciones o pseudoresiduos definidos por:

$$d_i = -2(y_i \log \widehat{p}_i + (1 - y_i) \log(1 - \widehat{p}_i))$$

Residuales de devianza

Definidos como:

$$D_i = -2sign(y_i - m_i\, p_i)\sqrt{y_i \log \frac{m_i\, p_i}{y_i} + (m_i - y_i) \log \frac{m_i\,(1 - \widehat{p}_i)}{m_i - y_i}}$$

Si la devianza es mayor que 4 entonces la observación correspondiente es anormal.

Uso de la regresión logística en clasificación

Para efectos de clasificación la manera más fácil de discriminar es considerar que si $p > 0,5$ entonces la observación pertenece a la clase que interesa. Pero algunas veces esto puede resultar injusto sobre todo si se conoce si una de las clases es menos frecuente que la otra.

Métodos alternativos son:

(a) Representar gráficamente el porcentaje de observaciones que poseen el evento (o sean que pertenecen al grupo (1) y que han sido correctamente clasificadas (sensibilidad) frente a distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (especificidad) frente a los mismos niveles de probabilidad anteriormente usados, en la misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtienen cortando las dos curvas.

(b) Usar la curva ROC (*receiver operating characteristic*). En este caso se representa gráficamente la sensibilidad frente a (1-especifidad)100 %, y se escoge como el p ideal aquel que está más cerca a la esquina superior izquierda, o sea al punto $(100, 0)$.

Diagnóstico en regresión logística

Verificar que el modelo es adecuado, (bondad de ajuste):

- Con datos agrupados: deviancia residual;

- Con datos individuales hace falta una referencia, que puede obtenerse a partir del modelo saturado, siempre que se trabaje con pocas variables y éste sea estimable;
- Otros estadísticos:
  - $\frac{\sum(O-E)^2}{E}$ sobre cada observación;
  - Hosmer y Lemeshow: $\frac{\sum(O-E)^2}{E}$ sobre 10 categorías de p.

Modelos predictivos

El objetivo del modelo puede ser:

- Generar una ecuación con capacidad predictiva, como una clasificación (análisis discriminante);
- Buscar qué factores tienen capacidad predictiva.

Si la respuesta es la aparición de un evento, pueden llamarse modelos pronósticos. En este tipo de estudios es típico contar con un gran número de variables a explorar.

2.7. Métodos de selección automática

Hacia adelante

1. Se inicia con un modelo vacío (sólo α);
2. Se ajusta un modelo y se calcula el p valor de incluir cada variable por separado;
3. Se selecciona el modelo con la más significativa;
4. Se ajusta un modelo con la(s) variable(s) seleccionada(s) y se calcula el p valor de añadir cada variable no seleccionada por separado;
5. Se selecciona el modelo con la más significativa;
6. Se repite 4 – 5 hasta que no queden variables significativas para incluir.

Hacia atrás

1. Se inicia con un modelo con TODAS las variables candidatas;
2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar;
3. Se selecciona para eliminar la menos significativa;
4. Se repite 2 – 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

Stepwise

- Se combinan los métodos adelante y atrás;
- Puede empezarse por el modelo vacío o por el completo, pero en cada paso se exploran las variables incluidas, por si deben salir y las no seleccionadas, por si deben entrar;
- No todos los métodos llegan a la misma solución necesariamente.

Consideraciones

- Criterio exclusivamente estadístico: no se tienen en cuenta otros "conocimientos" sobre las variables más interesantes a incluir (aunque se puede forzar a que algunas variables siempre estén en el modelo);
- Si hay un conjunto de variables muy correlacionadas, sólo una será seleccionada;
- No es fácil tener en cuenta interacciones entre variables (los modelos deben ser jerárquicos).

Valoración de la capacidad predictiva del modelo

- El modelo permite calcular una predicción del resultado en escala de probabilidad;
- Puede decidirse clasificar un individuo en el grupo de sucesos si su probabilidad supera un valor π:

$$clasificación = \begin{cases} \Pr > \pi \implies y_e = 1 \\ \Pr \leqslant \pi \implies y_e = 0 \end{cases}$$

Clasificación

- Sensibilidad $= VP/(VP + FN)$
- Especificidad $= VN/(VN + FP)$
- Área bajo la curva ROC construida para todos loa posibles puntos de corte de π para clasificar individuos:

Cálculo del área bajo la curva ROC

- Guardar los valores que predice el modelo (esperados)
- Calcular la U de Mann-Whitney respecto a los esperados:

$$AUC = 1 - \frac{U}{n_1\, n_0}$$

donde n_1 y n_0 son respectivamente el número esperado de "1" y "0".

$$AUC = 1 - \frac{U}{n_1\, n_0} = 1 - \frac{26273}{295 \times 286} = 0,69$$

Un $AUC = 0,5$ corresponde a una capacidad predictiva nula. El máximo es 1.

Elección del punto de corte óptimo

- Debe optimizarse la sensibilidad y la especificidad, y elegir un punto según la naturaleza del modelo predictivo
- El cambio en el punto de corte corresponde a emplear diferentes constantes en el modelo logístico
- Con frecuencia la constante estimada, α, consigue una sensibilidad y especificidad máxima, pero puede no ser el caso.

Validación del modelo

- El cálculo de la capacidad predictiva (CP) del modelo sobre la misma muestra que lo generó siempre es optimista, y debe validarse;

- Diferentes estrategias:
  - Probar el modelo en otra muestra diferente;
  - Elaborar el modelo con un 75 % de la muestra y calcular la CP en el 25 % restante;
  - Usar la misma muestra, pero calcular los indicadores de CP mediante técnicas de bootstrap o validación cruzada, que corrigen el "optimismo".

Regresión logística condicional

- Estudios con datos apareados;
- Generalización (modelo) del test de McNemar;
- Las observaciones no son independientes. Si se ignora, la correlación (positiva) entre las observaciones genera un sesgo hacia la hipótesis nula;
- Se pierde poder estadístico.

Estudios de casos y controles apareados

- Para cada caso se elige uno o varios controles que son idénticos (o muy parecidos) al caso en variables que se quiere controlar forzosamente: sexo, edad, residencia, fecha de diagnóstico;
- Estas variables quedarán igualadas entre casos y controles: no se podrá estimar un efecto;
- Debe controlarse la correlación que generan: modelo condicional;
- Cada caso se compara exclusivamente con sus controles;
- Las parejas (caso-control) que sean iguales en el factor de interés no son informativas;
- Es un método menos eficiente

$$I(\beta, x) = \sum_{sets} \frac{\exp(\alpha + \beta\, x_0)}{\exp(\alpha + \beta\, x_0) + \exp(\alpha + \beta\, x_1) + \cdots + \exp(\alpha + \beta\, x_r)}$$

- La constante del modelo se anula (es igual para casos y para controles).

Ajuste del modelo

- Software:
  - Se puede usar un programa para analizar modelos de Cox de supervivencia, pues la función de verosimilitud es la misma.
- Es un modelo lineal generalizado, por lo que pueden emplearse los mismos métodos para valorar efectos;
- Los coeficientes β son $\log(OR)$: $OR = e^{\beta}$.

Regresión multinomial

- La variable dependiente es categórica con más de dos grupos;
- Puede analizarse con regresión logística politómica (modelo multinomial);
- Se elige una categoría como referencia y se modelan varios *logits* simultáneamente, uno para cada una de las restantes categorías respecto a la de referencia.

  Ejemplo: hábito tabáquico

- La variable resultado tiene tres categorías:
  - Fumador;
  - Exfumador;
  - No fumador (referencia).
- Se modelan dos *logits* simultáneamente:
  - $logit(fumador \,/\, no\, fumador \,|\, z) = \alpha_1 + \beta_1\, z$;
  - $logit(ex - fumador \,/\, no\, fumador \,|\, z) = \alpha_2 + \beta_2\, z$
- Las covariables z son comunes pero se estiman coeficientes diferentes para cada *logit* (incluso diferente constante).

2.8. Regresión ordinal

- La variable respuesta tiene más de dos categorías ordenadas;
- Se modela un único *logit* que recoge la relación (de tendencia) entre la respuesta y las covariables;
- Hay varios modelos posibles según interese modelar la tendencia:
  - *odds* proporcionales (acumulado);
  - categorías adyacentes (parejas).

Odds-proporcionales

- Se compara un promedio de los posibles logia acumulados (respecto a la primera categoría).
- Cada *logit* tiene una constante diferente pero comparten el coeficiente de las covariables.
- Modelo de *odds* proporcionales:

$$logit_k(y > y_k \,|\, z) = \alpha_k + \beta\, z$$

donde

$$y = 1, 2, \ldots, C;$$
$$k = 2, 3, \ldots, C.$$

- Supone que el cambio entre diferentes puntos de corte de la respuesta es constante (β), pero parte de diferentes niveles (α_k).

Categorías adyacentes

- Compara cada categoría con la siguiente.
- Cada *logit* tiene una constante diferente pero comparten el coeficiente de las covariables.

Modelo de categorías adyacentes:

$$logit_k(y_k > y_{k-1} \mid z) = \alpha_k + \beta\, z$$

donde

$y = 1, 2, \ldots, C$;
$k = 2, 3, \ldots, C$.

- Supone que el cambio entre categorías adyacentes de la respuesta es constante (β), pero parte de diferentes niveles (α_k).

CAPÍTULO 3

Aplicación de la Regresión Logística

A continuación se muestra un ejemplo de la aplicación de la regresión logística en una investigación en el campo de la fonoaudiología. Existe una gran variedad de aplicaciones efectivas de los métodos de regresión logística en ciencias de la salud, medicina y biología; para citar algunas de relevancia con respecto el presente estudio se mencionan [35], [4], [22], [40] y [24]. La importancia de la regresión logística en aplicaciones de gran alcance y en el área de enfermedades del sueño y apnea ostructiva fue reportada en [38].

Los datos fueron procesados usando el software SPSS v19 (IBM SPSS Statistics for Windows, Version 19.0, IBM Corp. Released 2010) perteneciente a y con licencia de uso restringido del Laboratório de Estatística Aplicada (LEA) del Departamento de Estadística de la Facultad de Ciencias y Tecnología de la Universidad Estadual Paulista "Júlio de Mesquita Filho"(UNESP), en su Câmpus de Presidente Prudente en el interior de São Paulo. Para referencias sobre esta herramienta consulte, por ejemplo, [33].

Con el objeto de lograr los objetivos planteados se construyeron modelos estadísticos. Debido a la naturaleza de las variables de interés en la investigación se realizó un análisis de regresión logística en el cual fueron consideradas como variables dependientes las variables dicotómicas Ronquido Habitual, Riesgo Alto de Apnea y Riesgo Cardiovascular.

En la tabla 3.1 se muestran las variables incluidas como posibles predictoras y su operacionalización.

Variable	Operacionalización
Edad	**1. De 19 a 44** **2. De 45 a 44** **3. De 55 o más**
Sexo	**1. Masculino** **0. Femenino**
Estado Civil	**1. Soltero** **2. Casado** **3. Otros**
Color de la piel	**1. Negra** **2. Parda** **3. Amarilla** **4. Blanca**
Condición Socioeconómica	**1. A1-A2** **2. B1-B2** **3. C1-C2** **4. D** **5. E**
Largas Pausas al Respirar	**1. Presenta Largas Pausas** **0. No Presenta Largas Pausas**
Contracciones Movimientos de Piernas	**1. Presenta La Patología** **0. No Presenta La Patología**
Dificultad de Atención	**1. Tiene Dificultad** **0. No Tiene Dificultad**
Ronco Habitual	**1. Sí Ronca** **0. No Ronca**
Sudor Nocturno	**1. Presenta sudor nocturno** **0. No Presenta Sudor nocturno**
Fumante	**1. Fuma** **0. No Fuma**
Ex Fumador	**1. Fue Fumador** **0. No Fue fumador**
Uso abusivo del Alcohol	**1. Sí Ronca** **0. No Ronca**
Baba durante el Sueño	**1. Baba Durante el sueño** **0. No Baba Durante el sueño**
Riesgo de Apnea	**1. Presenta Riesgo de Apnea** **0. No Presenta Riesgo de Apnea**
Descanso Intencional	**1. Dormita Intencionalmente** **0. No Dormita Intencionalmente**

Tabla 3.1: Operacionalización de las variables incluidas como posibles predictoras en los modelos.

3.1. Análisis de Datos y Resultados

A continuación será explicada con detalles una salida de resultados de una Regresión Logística Binaria, realizada con el programa estadístico SPSS 19.

Se utilizará la base de datos generada para el proyecto en discusión. Se explicara con detalle las salidas obtenidas al construir un modelo de regresión logística binaria con siguientes variables: "Ronca durante el Sueño", "Sexo", "Sudor Nocturna" y "Dificultad en la Atención". La variable dependiente será la variable "Ronca durante el Sueño".

Al ejecutar el procedimiento Regresión Logística, el programa SPSS produce un archivo de salida conteniendo las siguientes tablas:

Resumen de Procesamiento de Casos

Casos sin Pesos[a]		N	Porcentaje
Casos Seleccionados	Incluidos en el Análisis	412	99,8
	Casos Faltantes	1	,2
	Total	413	100,0
Casos no Seleccionados		0	,0
Total		413	100,0

a. Si un peso es efectivo, vea la tabla de clasificación para el número total de casos.

Tabla 3.2: Resumen de los casos

La tabla 3.2 nos muestra un cuadro resumen con el número de casos (n) introducidos, en este caso 412, los seleccionados para el análisis y los excluidos en este caso 1 (casos perdidos, por tener algún valor faltante).

Codificación de Variables

Dependientes

Valor Original	Valor Interno
No Ronca	0
Si Ronca	1

Tabla 3.3: Codificación de la variable dependiente

La tabla 3.3 especifica la codificación de la variable dependiente (que debe ser dicotómica). Internamente el programa asigna el valor 0 al menor de los dos códigos, y el valor 1 al mayor. En este caso coincide con la codificación empleada en la base de datos, 1 para el grupo de individuos que ronca y 0 para el grupo de pacientes que no ronca. Es importante que el valor 1 identifique la categoría de la variable dependiente que resulte ser el resultado evaluado (en nuestro caso "Si Ronca"), ya que ello permite comprender mejor el coeficiente b_i de las variables independientes y de control.

El modelo se estima en dos bloques. En el primer bloque se introduce únicamente el término constante.

Bloque 0: Bloque inicial

Historia de Iteración[a,b,c]

Iteración		-2 Log Verosimilitud	Coeficientes Constante
Paso 0	1	569,754	,117
	2	569,754	,117

a. La Constante está incluida en el modelo.

b. -2 Log Verosimilitud inicial: 569,754

c. La estimación terminó a la iteración 2 porque los parámetros estimados variaron menos de ,001.

Tabla 3.4: Historial de las Iteraciones

La tabla 3.4 muestra información sobre el proceso de iteración, que se utilizará para construir el modelo con un solo coeficiente, la constante. En este bloque inicial se calcula la verosimilitud de un modelo que sólo tiene el término constante (a ó b_0). Puesto que la verosimilitud L es un número muy pequeño (comprendido entre 0 y 1), se suele ofrecer el logaritmo neperiano de la verosimilitud (LL), el cual en este caso es un número negativo, o el menos dos veces el logaritmo neperiano de la verosimilitud ($-2LL$), que es un número positivo.

El estadístico $-2LL$ mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de "desviación". Cuanto más pequeño sea el valor, mejor será el ajuste. En este primer paso sólo se ha introducido el término constante en el modelo. Un modelo sin poder predictivo alguno asigna a cualquier sujeto la probabilidad 0,5.

Tabla de Clasificación[a,b]

			Predicho		
			V44 Si el individuo tiene RONQUIDO HABITUAL		Porcentaje
	Observado		No Ronca	Si Ronca	
Paso 0	V44 Si le individuo tiene RONQUIDO HABITUAL	No Ronca	0	194	,0
		Si Roca	0	218	100,0
	Porcentaje Global				52,9

a. La constante está incluida en el modelo.

b. El valor de corte es ,500

Tabla 3.5: Tabla de Clasificación

La tabla 3.5 contiene información muy parecida a la empleada para valorar una prueba diagnóstica, permite evaluar el ajuste del modelo de regresión (hasta este momento, con un solo parámetro en la ecuación), comparando los valores predichos con los valores observados. Por defecto se ha empleado un punto de corte de 0,5 para clasificar a los individuos en cada grupo: esto significa que aquellos sujetos para los que la ecuación (con éste único término) calcula una probabilidad $< 0,5$ se clasifican como Ronquido Habitual $= 0$ (No Ronca), mientras que si la probabilidad resultante es mayor que $0,5$ se clasifican como Ronquido Habitual $= 1$ (Si Ronca). En este primer paso el modelo ha clasificado correctamente a un $52,9\,\%$ de los casos, y ningún sujeto que "no ronca" ha sido clasificado correctamente.

Variables en la Ecuación

		B	S.E.	Wald	df	Sig.	Exp(B)
Paso 0	Constante	,117	,099	1,396	1	,237	1,124

Tabla 3.6: Variables en la ecuación

La tabla 3.6 nos muestra las variables incluidas en la ecuación de regresión, recordemos que en este primer paso la ecuación solo incluye el término constante habiendo quedado fuera el resto de las variables. Esta tabla también nos muestra los valores del parámetro estimado (B), su error estándar ($E.T.$) y su significación estadística con la prueba de Wald (estadístico que sigue una ley Chi cuadrado con 1 grado de libertad) y la estimación de la OR ($\exp(B)$).

En el segundo bloque se incluyen las variables "sexo", "sudor nocturno", "dificultad en la atención".

Bloque 1

Historia de Iteración[a,b,c,d]

Iteración		-2 Log verosimilitud	Coeficientes			
			Constante	Sexo	Sudoresenoturna	Dificultad de Atención
Paso 1	1	545,870	-,303	,582	,885	,559
	2	545,701	-,317	,614	1,015	,591
	3	545,700	-,317	,614	1,020	,591
	4	545,700	-,317	,614	1,020	,591

a. Método: Inserción

b. La constante está incluida en el modelo.

c. -2 Log Verosimilitud inicial: 569,754

d. La estimación terminó a la iteración 4 porque los parámetros estimados variaron menos de ,001.

Tabla 3.7: Historial de las Iteraciones

La tabla 3.7 muestra información sobre el proceso de iteración, que se utiliza para construir el modelo, el cual ahora incluye cuatro coeficientes, los correspondientes a la constante (ya incluida en el anterior paso) y las variables SEXO, SUDOR_NOCTURNO y DIFICULTADATENCIÓN. Vemos como disminuye el $-2LL$ respecto al paso anterior (el modelo sólo con la constante tenía un valor de este estadístico de $569,754$, mientras que ahora se reduce a $545,700$), y el proceso termina con cuatro bucles. Los coeficientes calculados son los siguientes: para la constante $b_0 = -0,317$, y para las variable SEXO $b_1 = 0,614$, SUDOR_NOCTURNO, $b_2 = 1,020$ y DIFICULTADATENCIÓN $b_3 = 0,591$.

Prueba de Omnibus de los Coeficientes del Modelo

		Chi-square	df	Sig.
Paso 1	Paso	24,054	3	,000
	Bloque	24,054	3	,000
	Modelo	24,054	3	,000

Tabla 3.8: Prueba Omnibus sobre los coeficientes del modelo

La tabla 3.8 muestra una prueba Chi Cuadrado que evalúa la hipótesis nula de que los coeficientes (P) de todos los términos (excepto la constante) incluidos en el modelo son cero (Esta prueba de bondad de ajuste es comparable al test F global que en la Tabla ANOVA se realiza para evaluar el modelo de Regresión Lineal). El estadístico Chi Cuadrado para este contraste es la diferencia entre el valor de $-2LL$ para el modelo sólo con la constante y el valor de $-2LL$ para el modelo actual:

Chi cuadrado = $(-2LL\text{MODELO 0}) - (-2LL\text{MODELO 1}) = 569,754 - 545,700 = 24,054$

Esta tabla está formada por tres entradas: Paso, Bloque y Modelo.

- La primera fila (PASO) es la correspondiente al cambio de verosimilitud (de $-2LL$) entre pasos sucesivos en la construcción del modelo, contrastando la H_0 de que los coeficientes de las variables añadidas en el último paso son cero.
- La segunda fila (BLOQUE) es el cambio en $-2LL$ entre bloques de entrada sucesivos durante la construcción del modelo. Si como es habitual en la práctica se introducen las variables en un solo bloque, el Chi Cuadrado del Bloque es el mismo que el Chi Cuadrado del Modelo.
- La tercera fila (MODELO) es la diferencia entre el valor de $-2LL$ para el modelo sólo con la constante y el valor de $-2LL$ para el modelo actual.

La significación estadística $(0,000)$ nos indica que el modelo con la introducción de las nuevas variables mejora el ajuste de forma significativa con respecto al que se tenía en el paso anterior.

Resumen del Modelo

Paso	-2 Log Verosimilitud	Cox & Snell R Square	Nagelkerke R Square
1	545,700[a]	,057	,076

a. La estimación terminó a la iteración 4 porque los parámetros estimados variaron menos de ,001.

Tabla 3.9: Tabla Resumen de los Modelos

En la tabla 3.9 se reportan tres medidas complementarias a la presentada en la tabla anterior, para evaluar de forma global su validez: la primera es el valor del $-2LL$

y las otras dos son Coeficientes de Determinación (R^2), parecidos al que se obtiene en Regresión Lineal, que expresan la proporción (en tanto por uno) de la variación explicada por el modelo. Un modelo perfecto tendría un valor de $-2LL$ muy pequeño (idealmente cero) y un R^2 cercano a uno (idealmente uno). Recordemos que:

- $-2\log$ de la verosimilitud ($-2LL$) mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de "desviación". Recordemos que cuanto más pequeño sea el valor, mejor será el ajuste.

- La R cuadrado de Cox y Snell es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes). La R cuadrado de Cox y Snell se basa en la comparación del log de la verosimilitud (LL) para el modelo respecto al log de la verosimilitud (LL) para un modelo de línea base. Sus valores oscilan entre 0 y 1. En nuestro caso es un valor muy discreto ($0,039$) que indica que sólo el $3,9\,\%$ de la variación de la variable dependiente es explicada por la variable incluida en el modelo.

- La R cuadrado de Nagelkerke es una versión corregida de la R cuadrado de Cox y Snell. La R cuadrado de Cox y Snell tiene un valor máximo inferior a 1, incluso para un modelo "perfecto". La R cuadrado de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1.

La tabla 3.10 nos muestra una prueba de ajuste global del modelo que se conoce como **Prueba de Hosmer y Lemeshow**. Esta es otra prueba para evaluar la bondad del ajuste de un modelo de regresión logística [16].

Si el ajuste es bueno, un valor alto de la probabilidad predicha (p) se asociará con el resultado 1 de la variable binomial dependiente, mientras que un valor bajo de p (próximo a cero) corresponderá —en la mayoría de las ocasiones— con el resultado $Y = 0$. Se trata de calcular, para cada observación del conjunto de datos, las probabilidades de la variable dependiente que predice el modelo, ordenarlas, agruparlas y calcular, a partir de ellas, las frecuencias esperadas, y compararlas con las observadas mediante una prueba χ^2.

El estadígrafo de Hosmer y Lemeshow no se computa cuando, para algunos grupos, E_i (valores esperados) ó $E_i \cdot (n_i - E_i)$ son nulos o muy pequeños (menores que 5). Por otra parte, lo que se desea en esta prueba es que no haya significación (lo contrario a lo que suele ser habitual). Por eso, muchos autores proponen simplemente cotejar valores

Prueba de Hosmer y Lemeshow

Paso	Chi-square	df	Sig.
1	1,278	3	,734

Tabla de Contingencia para la Prueba de Hosmer y Lemeshow

		V44 Si el individuo tiene RONQUIDO HABITUAL = No Ronca		V44 Si el individuo tiene RONQUIDO HABITUAL = Si Ronca		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	113	113,406	83	82,594	196
	2	29	29,801	40	39,199	69
	3	30	30,262	41	40,738	71
	4	19	16,174	35	37,826	54
	5	3	4,358	19	17,642	22

Tabla 3.10: Prueba de Hosmer y Lemeshow

observados y esperados mediante simple inspección y evaluar el grado de concordancia entre unos y otros a partir del sentido común.

Sobre este razonamiento, una forma de evaluar la ecuación de regresión y el modelo obtenido es construir una tabla 2×2 (tabla 3.11) clasificando a todos los individuos de la muestra según la concordancia de los valores observados con los predichos o estimados por el modelo, de forma similar a como se evalúan las pruebas diagnósticas. Una ecuación sin poder de clasificación alguno tendría una especificidad, sensibilidad y total de clasificación correctas igual al 50 % (por el simple azar). Un modelo puede considerarse aceptable si tanto la especificidad como la sensibilidad tienen un nivel alto, de al menos el 75 %. Con nuestro modelo la tabla de clasificación obtenida es la siguiente:

En la tabla de clasificación podemos comprobar que nuestro modelo tiene una especificidad de 58 % y una sensibilidad de $61,9$ %. Con la constante y tres variables predictoras, no clasifica tan mal a los individuos que roncan durante el sueño cuando el punto de corte de la probabilidad de Y calculada se establece (por defecto) en 50 % $(0,5)$.

La tabla 3.12 nos ofrece información sobre las variables que formaran el modelo, sus coeficientes de regresión con sus correspondientes errores estándar, el valor del estadístico de Wald para evaluar la hipótesis nula ($P_i = 0$), la significación estadística asociada, y el valor de la OR ($\exp(B)$) con sus intervalos de confianza.

Tabla de Clasificación[a]

			Predicho		
			V44 Si el individuo tiene RONQUIDO HABITUAL		Porcentaje
	Observado		No Ronca	Si Ronca	
Paso 1	V44 Si el individuo tiene RONQUIDO HABITUAL	No Ronca	113	81	58,2
		Si Ronca	83	135	61,9
	Porcentaje Global				60,2

a. El valor de corte es ,500

Tabla 3.11: Tabla de clasificación del modelo

Variables en la Ecuación

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Inferior	Superior
Paso 1[a]	Sexo	,614	,225	7,487	1	,006	1,848	1,190	2,870
	Sudoresenoturna	1,020	,425	5,763	1	,016	2,774	1,206	6,379
	Dificultad de Atención	,591	,225	6,910	1	,009	1,806	1,162	2,806
	Constante	-,317	,137	5,362	1	,021	,728		

a. Las variable(s) ingresada(s) en el paso 1: Sexo, Sudoresenoturna, Dificultad de Atención.

Tabla 3.12: Variables incluidas en el modelo estimado

Con estos datos podemos construir la ecuación de regresión logística, que en nuestro ejemplo sería:

$$P(\text{Individuo Roncar Dormido}) = \frac{1}{1+\exp(-0,317+0,614\times\text{Sexo}+1,020\times\text{Sudoresenoturna}+0,591\times\text{DificultaNaAtencao})}$$

Esta fórmula nos sirve para estimar la probabilidad de un individuo roncar dormido en función del sexo, de si presenta o no sudor nocturno o si presenta dificultad en la atención. Así, un individuo de sexo masculino tendría, según esta ecuación logística, la siguiente probabilidad de roncar dormido.

$$P(\text{Individuo Roncar Dormido}) = \frac{1}{1+\exp(-0,317+0,614+1,020+0,591)}$$
$$= \frac{1}{1+2,718(1,908)} = 1/6,18 = 0,1618$$

Puesto que la probabilidad predicha es menor que $0,50$ el individuo se clasifica como un individuo que ronca.

Con la información obtenida a través del procedimiento Regresión logística del programa SPSS, resumida en la tabla 3.12 se construye la siguiente tabla resumen:

Variable	B	Error Estándar	Sig.	Odds Ratio	I.C de 95 % Odds ratio	
					Inferior	Superior
Intercepto	-,317	,137	,021	,728		
Sexo	.614	.225	,006	1,848	1,190	2,870
Sudor Nocturno	1,020	,425	,016	2,774	1,206	6,379
Dificultad en la Atención	,591	,225	,009	1,806	1,162	2,806

Tabla 3.13: Resultados del Análisis de Regresión Logística utilizando Como Variable Dependiente: Presencia de Ronquido

En la tabla 3.13 se observa que todas las variables incluidas en la construcción del modelo resultaron estadísticamente significativas. Esta significación indica que las variables Sexo, Sudor Nocturno y Dificultad en la Atención, tienen influencia en la probabilidad de un individuo presentar Ronquido Habitual.

A partir de los valores mostrados bajo la columna **odds ratios** se puede concluir lo siguiente:

- ✓ Los individuos del sexo masculino tienen $1,85$ veces más riesgos de presentar Ronquido nocturno en comparación con los individuos del sexo femenino.
- ✓ Los individuos que padecen de Sudor Nocturna tienen 2,77 veces más riesgos de presentar Ronquido nocturno en comparación con los individuos que no padecen Sudor Nocturno.
- ✓ Los individuos que presentan Dificultad en la Atención tienen 1,80 veces más riesgos de presentar Ronquido nocturno en comparación con los individuos que no presentan Dificultad en la Atención.

Utilizando el procedimiento Regresión Logística de manera similar con el resto de las variables, se obtienen las siguientes tablas que resumen de manera organizada la infor-

mación que corresponde a cada modelo estimado para las variables "riesgo alto de apnea" y "riesgo cardiovascular".

Variable	B	Error Estándar	Sig.	Odds Ratio	I.C de 95 % Odds ratio	
					Inferior	Superior
Intercepto						
Circunferencia Cervical	.149	.040	.000	1.160	1.073	1.255
Largas Pausas al Respirar	1.241	.422	.003	3.461	1.513	7.915
Contracciones movimientos de piernas	1.476	.621	.017	4.375	1.296	14.769
Dificultad en Atención	.867	.306	.005	2.379	1.307	4.332
Ronco Habitual	2.999	.495	.000	20.067	7.602	52.973

Tabla 3.14: Resultados del Análisis de Regresión Logística utilizando como variable Dependiente: Riesgo de Apnea

En la tabla 3.14 se observa que todas las variables resultaron estadísticamente significativas, esta significación indica que las variables Circunferencia Cervical, Largas Pausas en la Respiración, Contracciones y Movimientos en las Piernas, Dificultad en la Atención y Ronquido Habitual tienen influencia en la probabilidad de un individuo tener Riesgo de Apnea.

El valor positivo del coeficiente correspondiente a la variable Circunferencia Cervical nos indica que a medida que aumenta la circunferencia cervical aumentan las posibilidades de un individuo padecer riesgos de apnea. Un aumento en una unidad de los valores de la circunferencia cervical aumenta el riesgo de apnea en un factor $1,149$.

A partir de los valores mostrados bajo la columna **odds ratios** se puede concluir lo siguiente:

- ✓ Las personas que presentan Largas Pausas de Respiración tienen 3,46 veces más riesgos de pertenecer al grupo de individuos con riesgo de apnea en comparación con las personas que no presentan esta ocurrencia del sueño.
- ✓ Las personas que presentan Dificultad en la Atención tienen 2,37 veces más riesgos

de pertenecer al grupo de individuos con riesgo de apnea en comparación con las personas que no presentan Dificultad en la Atención.

✓ Las personas que presentan Ronquido Habitual tienen 20,67 veces más riesgos de pertenecer al grupo de individuos con riesgo de apnea en comparación con las personas que no presentan esta patología del sueño.

La Variable Riesgo Cardiovascular se codificó en dos posibles valores, el primer grupo estuvo constituido por los individuos que tenían un riesgo cardiovascular bajo, el segundo estuvo constituido por los individuos con un riesgo cardiovascular Moderado, Elevado y Muy Elevado. Se calculó la asociación de dicha variable con el resto de variables ya nombradas anteriormente obteniéndose los siguientes resultados:

Variable	**Chi - Cuadrado**	**Sig.**	**Odds Ratio**	**I.C de 95 % Odss ratio**	
				Inferior	**Superior**
Presión Arterial	28,855	$P < 0{,}01$	3,503	2,188	5,609
Circunferencia Abdominal	27,682	$P < 0{,}01$	3,039	1,995	4,629
Circunferencia Cervical	37,876	$P < 0{,}01$	4,406	2,692	7,209
Fumante	49,338	$P < 0{,}01$	6,826	3,807	12,238
Ex Fumador	3,250	0.075	1,728	0,949	3,148
Uso abusivo del Alcohol	28,843	$P < 0{,}01$	3,239	2,090	5,019
Baba durante el sueño	1,419	0,234	1,336	0,829	2,154
Riesgo de Apnea	12,213	$P < 0{,}01$	2,241	1,418	3,543
Ronco Habitual	0,463	0,496	1,149	0,770	1,715
Descanso Intencional	5,554	0,018	0,53	0,310	0,903

Tabla 3.15: Resultados del Análisis de Regresión Logística utilizando Como Variable Dependiente: Riesgo cardiovascular

De los valores listados bajo la columna **sig** de la tabla 3.15 se observa que no existe dependencia entre las variables Ex-Fumante, Baba Durante el sueño, Ronquido Habitual y Reposo Intencional con la variable Riesgo Cardiovascular.

A partir de los valores mostrados bajo la columna **odds ratios** se puede concluir lo siguiente:

✓ Los individuos con una presión arterial alta tienen $3,5$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas con presión arterial baja.

- ✓ Los individuos con una Circunferencia Abdominal alterada tienen $3,09$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas con una Circunferencia Abdominal normal.
- ✓ Los individuos con una Circunferencia Cervical alterada tienen $4,40$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas con una Circunferencia Cervical normal.
- ✓ Los individuos Fumadores tienen $3,80$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas que no Fuman.
- ✓ Los individuos que hacen Uso abusivo del Alcohol tienen $2,090$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas con una Circunferencia Abdominal normal.
- ✓ Los individuos con Riesgo de Apnea tienen $1,41$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas que no tienen Riesgo de Apnea.

3.2. Conclusiones

En el nmarco de los objetivos propuestos y para la población estudiada, las conclusiones más relvantes del estudio son las siguientes.

- ✓ Las personas que presentan Ronquido Habitual tienen 20,67 veces más riesgos de pertenecer al grupo de individuos con Resgo de Apnea en comparación con las personas que no presentan esta patología del sueño.
- ✓ Los individuos con Riesgo de Apnea tienen $1,41$ más probabilidades de tener riesgo cardiovascular moderado, alto o muy alto comparado con las personas que no tienen Riesgo de Apnea.

Anexo

Cuestionario de Berlin

RESMED

Questionário de Berlim
AVALIAÇÃO DO SONO

Nome _____

Cidade, Estado e CEP _____

1. Complete o seguinte:

Altura _____ Idade _____

Peso _____ Masculino/Feminino _____

O seu peso mudou?
- ☐ Aumentou
- ☐ Diminuí
- ☐ Permaneceu inalterado

2. Você ronca?
- ☐ Sim ☐ Não ☐ Não sei

Se roncar:

3. Seu ronco é . . .
- ☐ Ligeiramente mais alto que a respiração
- ☐ Tão alto quanto a fala
- ☐ Mais alto que a fala
- ☐ Muito alto

4. Com que freqüência você ronca?
- ☐ Quase todos os dias
- ☐ 3-4 vezes por semana
- ☐ 1-2 vezes por semana
- ☐ 1-2 vezes por mês
- ☐ Nunca ou quase nunca

5. Seu ronco incomoda outras pessoas?
- ☐ Sim ☐ Não

6. Alguém já notou que você pára de respirar duranteo sono?
- ☐ Quase todos os dias
- ☐ 3-4 vezes por semana
- ☐ 1-2 vezes por semana
- ☐ 1-2 vezes por mês
- ☐ Nunca ou quase nunca

7. Você acorda cansado?
- ☐ Quase todos os dias
- ☐ 3-4 vezes por semana
- ☐ 1-2 vezes por semana
- ☐ 1-2 vezes por mês
- ☐ Nunca ou quase nunca

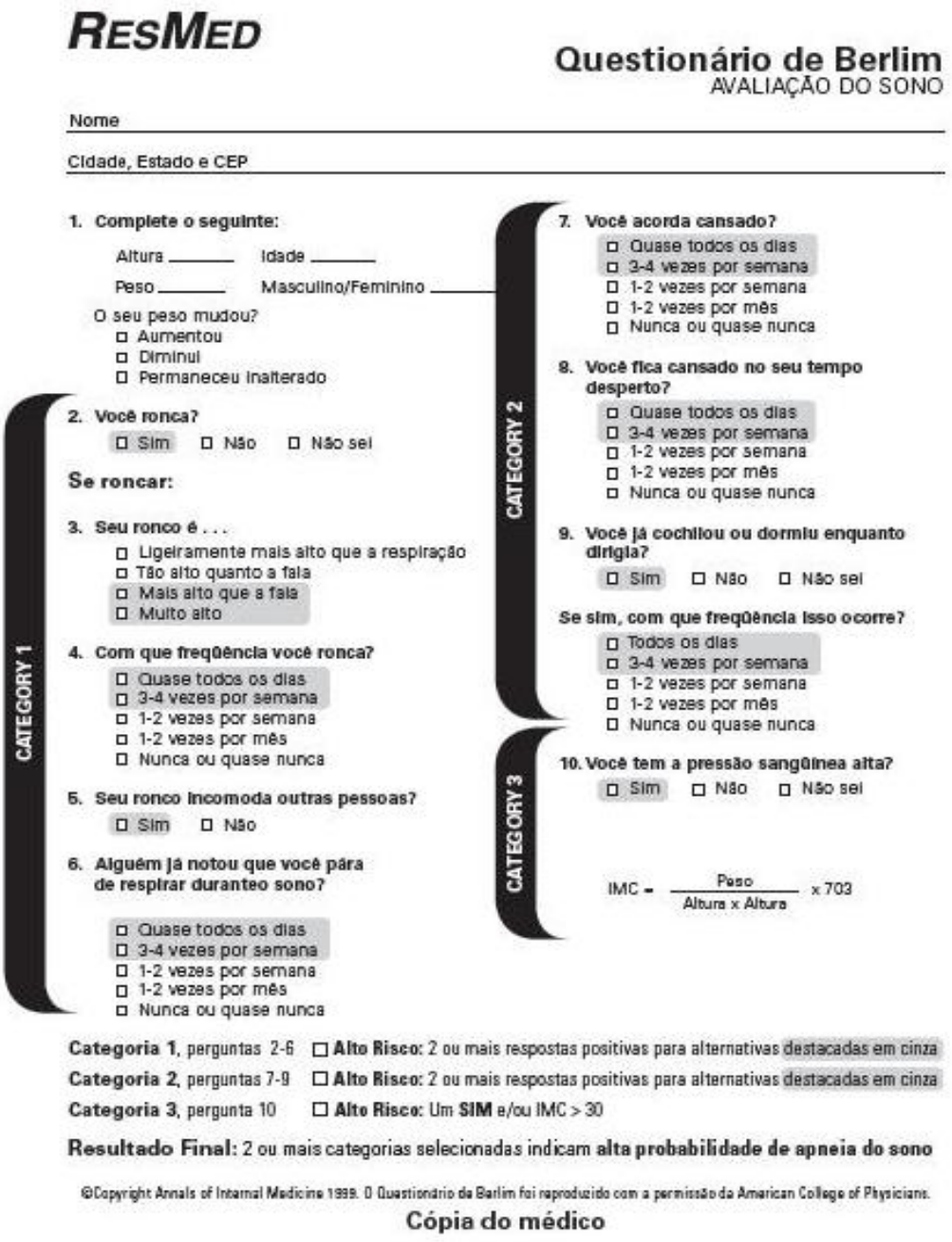
8. Você fica cansado no seu tempo desperto?
- ☐ Quase todos os dias
- ☐ 3-4 vezes por semana
- ☐ 1-2 vezes por semana
- ☐ 1-2 vezes por mês
- ☐ Nunca ou quase nunca

9. Você já cochilou ou dormiu enquanto dirigia?
- ☐ Sim ☐ Não ☐ Não sei

Se sim, com que freqüência isso ocorre?
- ☐ Todos os dias
- ☐ 3-4 vezes por semana
- ☐ 1-2 vezes por semana
- ☐ 1-2 vezes por mês
- ☐ Nunca ou quase nunca

10. Você tem a pressão sangüínea alta?
- ☐ Sim ☐ Não ☐ Não sei

Cópia do Paciente



RESMED

Questionário de Berlim
AVALIAÇÃO DO SONO

Nome

Cidade, Estado e CEP

CATEGORY 1

1. **Complete o seguinte:**

   Altura _____ Idade _____

   Peso _____ Masculino/Feminino _____

   O seu peso mudou?
   - ☐ Aumentou
   - ☐ Diminuiu
   - ☐ Permaneceu inalterado

2. **Você ronca?**
   - ☐ Sim ☐ Não ☐ Não sei

Se roncar:

3. **Seu ronco é . . .**
   - ☐ Ligeiramente mais alto que a respiração
   - ☐ Tão alto quanto a fala
   - ☐ Mais alto que a fala
   - ☐ Muito alto

4. **Com que freqüência você ronca?**
   - ☐ Quase todos os dias
   - ☐ 3-4 vezes por semana
   - ☐ 1-2 vezes por semana
   - ☐ 1-2 vezes por mês
   - ☐ Nunca ou quase nunca

5. **Seu ronco incomoda outras pessoas?**
   - ☐ Sim ☐ Não

6. **Alguém já notou que você pára de respirar durante o sono?**
   - ☐ Quase todos os dias
   - ☐ 3-4 vezes por semana
   - ☐ 1-2 vezes por semana
   - ☐ 1-2 vezes por mês
   - ☐ Nunca ou quase nunca

CATEGORY 2

7. **Você acorda cansado?**
   - ☐ Quase todos os dias
   - ☐ 3-4 vezes por semana
   - ☐ 1-2 vezes por semana
   - ☐ 1-2 vezes por mês
   - ☐ Nunca ou quase nunca

8. **Você fica cansado no seu tempo desperto?**
   - ☐ Quase todos os dias
   - ☐ 3-4 vezes por semana
   - ☐ 1-2 vezes por semana
   - ☐ 1-2 vezes por mês
   - ☐ Nunca ou quase nunca

9. **Você já cochilou ou dormiu enquanto dirigia?**
   - ☐ Sim ☐ Não ☐ Não sei

Se sim, com que freqüência isso ocorre?
   - ☐ Todos os dias
   - ☐ 3-4 vezes por semana
   - ☐ 1-2 vezes por semana
   - ☐ 1-2 vezes por mês
   - ☐ Nunca ou quase nunca

CATEGORY 3

10. **Você tem a pressão sangüínea alta?**
    - ☐ Sim ☐ Não ☐ Não sei

$$\text{IMC} = \frac{\text{Peso}}{\text{Altura} \times \text{Altura}} \times 703$$

Categoria 1, perguntas 2-6 ☐ **Alto Risco:** 2 ou mais respostas positivas para alternativas destacadas em cinza

Categoria 2, perguntas 7-9 ☐ **Alto Risco:** 2 ou mais respostas positivas para alternativas destacadas em cinza

Categoria 3, pergunta 10 ☐ **Alto Risco:** Um **SIM** e/ou IMC > 30

Resultado Final: 2 ou mais categorias selecionadas indicam **alta probabilidade de apneia do sono**

Cópia do médico

RESMED

Tabela do Índice de Massa Corporal

Altura (metros) \ Peso (kilos)	54.43	58.96	63.50	68.03	72.57	77.11	81.64	86.18	90.71	95.25	99.79	104.3	108.8	113.4
1.524	23	25	27	29	31	33	35	37	39	41	43	45	47	49
1.574	22	24	26	27	29	31	33	35	37	38	40	42	44	46
1.625	21	22	24	26	28	29	31	33	34	36	38	40	41	43
1.676	19	21	23	24	26	27	29	31	32	34	36	37	39	40
1.727	18	20	21	23	24	26	27	29	30	32	34	35	37	38
1.778	17	19	20	22	23	24	26	27	29	30	32	33	35	36
1.828	16	18	19	20	22	23	24	26	27	29	30	31	33	34
1.879	15	17	18	19	21	22	23	24	26	27	28	30	31	32

Referencias Bibliográficas

[1] Allison, Paul D. (1999). *Logistic Regression Using The SAS System, Theory and Application.* Willey.

[2] Alves da Silva, G., Haueisen Sander, H., Eckeli, A., França, R. M., Barbosa Coelho, E., Nobre, F. (2009). *Conceitos básicos sobre síndrome da apneia obstrutiva do sono.* Rev Bras Hipertens vol.16(3):150-157.

[3] Barros Soares, E. Barbalho Pires, J., Menezes, M., Silva de Santana, S., K., Fraga, J. (2010). *Fonoaudiologia X Ronco/Apneia do sono.* Rev. CEFAC. 12(2):317-325.

[4] Bender, R. & Grouven, U. (1997) . *Ordinal logistic regression in medical research.* Journal of the Royal College of Physicians of London. Vol. 31 No. 5, p. 546-551.

[5] Cintra F., Poyares D., Guilleminault C., Carvalho A. C., Tufik S., de Paola A. (2006) *Alterações Cardiovasculares na Síndrome da Apnéia Obstrutiva do Sono.* Arquivos Brasileiros de Cardiologia, Volume 86, No. 6.

[6] Cintra, F., Tufik, S., de Paola, A., Feres, M., Mello-Fujita, L., Oliveira, W., Rizzi, C., Poyares, D. (2011). *Perfil Cardiovascular em Pacientes com Apneia Obstrutiva do Sono.* Arq Bras Cardiol 96(4):293-299.

[7] Conselho Federal de Medicina. (2012) *O Projeto Diretrizes, Iniciativa conjunta da Associação Médica Brasileira e Conselho Federal de Medicina. Apneia Obstrutiva do Sono e Ronco Primário.* Diagnóstico 2012.

[8] Cox, D., Snell, E. (1989). *The Analysis of Binary Data.* 2 ed. London, Chapman and Hall. 231 p.

[9] Gondim F.J., Lessa I. (2005) *Prevalência e fatores associados ao sedentarismo no lazer em adultos.* Cadernos de Saude Publica 21(3):870.

[10] Grunstein, R. R. (2012). *Global perspectives on sleep and health issues.* J. Natl. Inst. Public Health, 61 (1) Japan.

[11] Grunstein, R. (2006). *Syndrome Zzzzzzzzz: The overlap between snoring, sleep apnea, and metabolic syndrome.* Drug Development Research, 67(7), 616-618.

[12] Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (1995). *Multivariate Data Analysis.* 4th edition. Prentice-Hall, Inc.

[13] Hedner, J., Berend, N., Grunstein, R., Phillips, C. (2005). *Diurnal and obstructive sleep apnea influences on arterial stiffness and central blood pressure in men.* Sleep, 28(5), 604-609.

[14] Hoffstein, V., Mateika, S. (1994). *Cardiac Arrhythmias, Snoring, and Sleep Apnea.* Chest. 106(2):466-471.

[15] Hoffstein V, Viner S, Mateika S, Conway J. (1992). *Treatment of obstructive sleep apnea with nasal continuous positive airway pressure. Patient compliance, perception of benefits, and side effects.* Am Rev Respir Dis. Apr;145(4 Pt 1):841-5.

[16] Hosmer D.W. and Lemeshow S. (1980) *A goodness-of-fit test for the multiple logistic regression model.* Communications in Statistics A10:1043-1069.

[17] Hosmer, D., Lemeshow, S. (1989). *Applied Logistic Regression.* New York, John Wiley.

[18] Hosmer, D., Lemeshow, S., Klar, P. (1988). *Goodness of Fit testing for Multiple Logistic Regression Analysis when the estimated probabilities are small.* Biometrical Journal 30: 911-924.

[19] Jennings, D. (1986). *Judging inference adequacy in logistic regression.* Journal of American Statistical Association 81: 471-476.

[20] Johnson, R. A.; Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* 5th edition. Prentice Hall.

[21] Kleinbaum D., Klein M. (2002) *Logistic regression: a self-learning text -2nd ed.* Springer Verlag, Statistics for Biology and Health.

[22] Lee, J. (1986) *An insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence.* International Journal of Epidemiology,15 (1), 22-29.

[23] Lessa, I. (2004) *Doenças crônicas não-transmissíveis no Brasil: um desafio para a complexa tarefa da vigilância.* Ciênc. saúde coletiva v.9 n.4 Rio de Janeiro.

[24] Long, WJ, Grith JL, Selker, HP, D'Agostino RB. (1993). *A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain.* Biomedical Research,26: 74-97.

[25] Lorenzi Filho, G. (2008) *Apnéia obstrutiva do sono: um grave problema de saúde pública.* Pneumologia Paulista. São Paulo, v.21, n.3, p.5.

[26] Marshall, N., Wong, K., Cullen, S., Knuiman, M., Grunstein, R. R. (2014). *Sleep Apnea and 20-Year Follow-Up for All-Cause Mortality, Stroke, and Cancer Incidence and Mortality in the Busselton Health Study Cohort.* Journal of Clinical Sleep Medicine. 15;10(4):355-62.

[27] Noal, R. B. (2008). *Ronco habitual e apnéia obstrutiva observada em adultos: estudo de base populacional, Pelotas, RS.* Revista de Saúde Pública, v. 42, n. 2, p. 224-233.

[28] *Organización Mundial de la Salud (2011). Estadísticas sanitarias mundiales 2011.* Ediciones de la OMS, Organización Mundial de la Salud, 20 Avenue Appia, 1211 Ginebra 27, Suiza.

[29] Organización Panamericana de la Salud (2007). *Situación de las Estadísticas Vitales, de Morbilidad y de Recursos y Servicios en Salud de los países de las Américas (Informe Regional) 11/2007.*

[30] Pedrosa, PR., Montenegro, M., Pedrosa, L., Celestino, D., Filho, S., Lorenzi-Filho, G. (2009). *Apneia do sono e hipertensão arterial sistêmica.* Rev Bras Hipertens vol.16(3):174-177.

[31] Pedrosa RP, Krieger EM, Lorenzi-Filho G, Dragar LF. (2011). *Avanços Recentes do Impacto da Apneia Obstrutiva do Sono na Hipertensão Arterial Sistêmica.* Arq. bras. cardiol;97(2).

[32] Peña, D. (2002). *Análisis de Datos Multivariantes.* McGraw Hill.

[33] Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos, Aplicaciones con SPSS.* Pearson.

[34] Richard, R. F., Gay, P. C., Farrell, P. C. (2006). *The Economics of Sleep Disordered Breathing.* RT Magazine. com,

[35] Silva, L. C. (1995). *Excursión a la Regresión Logística en Ciencias de la Salud*. Ediciones Díaz de Santos, S.A.

[36] Sociedade Brasileira de Sono, Sociedade Brasileira de Rinologia, Sociedade Brasileira de Otorrinolaringologia (2000). *I Consenso em ronco e apnéia do sono.* Coordenação Sérgio Tu, Perboyre Lacerda Sampaio, Luc Louis Maurice Weckx. São Paulo: Sociedade Brasileira de Sono. 67p.

[37] Stirbulov R. (2007). *Respiratory repercussions of obesity.* J Bras Pneumol. 33(1):vii-viii.

[38] Tufik, S., Santos-Silva, R., Taddei, J. A., Azeredo, L. R. (2010). *Obstructive Sleep Apnea Syndrome in the Sao Paulo Epidemiologic Sleep Study.* Sleep Medicine. Volume 11, Issue 5 , Pages 441-446,

[39] Vaz A.P., Drummond M., Mota P.C., Severo M., Almeida J., Carlos Winck J. (2011) *Tradução do Questionário de Berlim para língua Portuguesa e sua aplicação na identificação da SAOS numa consulta de patologia respiratória do sono.* Rev Port Pneumol. 2011;17(2):59-65.

[40] Vittinghoff E., Shiboski S., Glidden D., McCulloch C. (2005) *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models.* Springer Verlag, Statistics for Biology and Health.

[41] Yaggi, H. K,, Concato, J., Kernan, W. N., Lichtman, J. H., Brass, L. M., Mohsenin, V. (2005). *Obstructive sleep apnea as a risk factor for stroke and death.* N. Engl. J. Med. 353(19):2034-41.