
# Un Modelo Logístico para la Evolución de Neonatos Prematuros con Bajo Peso al Nacer, Atendidos en el Hospital Universitario del Valle, durante el período 2002 a 2010

Yessica María Bonilla Castillo

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2019

# Un Modelo Logístico para la Evolución de Neonatos Prematuros con Bajo Peso al Nacer, Atendidos en el Hospital Universitario del Valle, durante el período 2002 a 2010

**Yessica María Bonilla Castillo**

Trabajo de grado presentado como requisito parcial para optar al título de:
**Profesional en Estadística**

Director:
Iván Mauricio Bermudez, M.Sc.

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2019

# *Agradecimientos*

*En primer lugar, quiero brindar mi mayor gratitud a Dios, por guiarme en el camino y fortalecerme espiritualmente para empezar un camino lleno de éxito. Adicional, quiero mostrar mi gratitud a todas aquellas personas que estuvieron presentes en la realización de esta meta que es tan importante para mí, agradecer por todas sus ayudas, palabras motivadoras, conocimientos, consejos y dedicación.*

*A mis compañeros, quienes a través del tiempo fuimos fortaleciendo una amistad, gracias por la colaboración, experiencias, alegrías, frustraciones, llantos, tristezas, peleas y celebraciones.*

*Agradezco también a mi asesor de tesis Iván Mauricio Bermúdez, por haberme brindado la oportunidad de recurrir a sus capacidades y conocimientos estadístico durante el desarrollo de la tesis.*

*Por último, quiero agradecer a la base de todo, a mi familia, en especial a mi madre, por ser el pilar fundamental en todo lo que soy, en toda mi educación, tanto académica como de la vida y por su incondicional apoyo mantenido a través del tiempo.*

*¡Muchas gracias por todo!*

# Resumen

Con el objetivo de construir el modelo más parsimonioso y mejor ajustado para observar la evolución de neonatos prematuros con bajo peso, pertenecientes al programa de seguimiento de alto riesgo y canguro del Hospital Universitario del Valle (HUV), se acude al modelo de regresión logístico para identificar los recién nacidos de alto riesgo. Se estudió una muestra de 145 historias clínicas de neonatos desde febrero de 2002 hasta noviembre de 2010. La metodología estadística utilizada, basada en la modelación permite obtener un modelo que describa la evolución del peso del neonato prematuro y la descripción de los posibles factores de asociación, de tal forma que se puedan realizar inferencias significativas clínicamente a partir de los parámetros del modelo. En primer lugar, se realizó un análisis descriptivo, luego, se obtuvieron las curvas asociadas a los diferentes tiempos de seguimientos sin realizar ningún tipo de corrección ni modelaje y después se obtuvieron los posibles modelos, para al final seleccionar el modelo mejor ajustado estadísticamente para la descripción de la evolución del peso de los neonatos prematuros a través del tiempo.

**Palabras Claves:**

Neonato, datos longitudinales, Bajo Peso al Nacer, Prematuro, Modelo de Regresión Logístico.

# Abstract

With the objective of building the most parsimonious and best adjusted model to observe the evolution of premature infants with low weight, belonging to the high risk and kangaroo follow-up program of the Hospital Universitario del Valle (HUV), the logistic regression model is used to identify high-risk newborns. A sample of 145 clinical records of neonates was studied from February 2002 to November 2010. The statistical methodology used, based on modeling, allows us to obtain a model that describes the evolution of premature neonate weight and the description of possible association factors , so that clinically meaningful inferences can be made from the parameters of the model. First, a descriptive analysis was carried out, then, the curves associated with the different follow-up times were obtained without making any type of correction or modeling and afterwards the possible models were obtained, for the end selecting the best statistically adjusted model for the description of the evolution of the weight of premature infants through time.

**Keywords:**

Neonate, longitudinal data, Low Birth Weight, Premature, Logistic Regression Model.

# Contenido

# Lista de Tablas

# Lista de Figuras

# Capítulo 1

# Introducción

Las curvas de crecimiento son la herramienta para realizar el control de seguimiento del crecimiento y desarrollo de un niño o niña en diferentes instantes del tiempo (Milad et al., 2010), generando información de interés para los pediatras. Uno de los indicadores más importantes para medir el crecimiento y desarrollo es el peso del niño o niña, ya que permite observar el grado de desnutrición y desigualdad social (Paraje, 2009). Mediante estas curvas se pueden generar diferentes modelos matemáticos que permitan predecir el peso a través del tiempo en una población de niños e inferir sobre los parámetros del modelo para generar estrategias de prevención (Hachuel et al., 2006).

El Hospital Universitario del Valle (HUV) de Cali, es el centro de referencia de embarazos de alto riesgo en la parte sur occidental colombiana, presentando una tasa alta de neonatos prematuros con bajo peso (Bermudez et al., 2015). El bajo peso al nacer (BPN) está relacionado con la incidencia de enfermedades y mortalidad infantil y neonatal, siendo causa principal de alto riesgo de padecer problemas cardíacos, infecciones, asfixia perinatal, aspiración de meconio, hipotermia, desnutrición, parálisis infantil, deficiencias mentales y trastornos del aprendizaje (Velázquez Quintana et al., 2004). Adicional, el BPN constituye un verdadero problema de salud pública, al incrementar la morbilidad y mortalidad infantil, el tiempo y costo de hospitalización (Veintimilla Dávila, 2017). Con el objetivo de modelar la evolución de los neonatos que pertenecen al programa de alto riesgo y canguro del HUV, se considera necesario realizar un modelo logístico para modelar la evolución del peso del neonato prematuro a través del tiempo. Todo esto, con el propósito de generar estrategias de prevención y obtener información estadísticamente significativa para las entidades públicas prestadores de servicio de salud.

Se utiliza la regresión logística como técnica para modelar la influencia de un conjunto de variables regresoras en la probabilidad de ocurrencia de tener un peso adecuado o bajo en la edad gestacional. La principal finalidad es modelar la evolución y realizar inferencia de los parámetros. Donde el impacto social esperado es netamente preventivo, siendo una ayuda para el personal de salud que tiene relación directa con el control prenatal y la recepción

de recién nacido, los administradores de salud que elaboran políticas al disponer de datos sobre factores de riesgo relacionados con BPN, promoverán intervenciones preventivas en los distintos niveles de atención pública.

La estructura del documento esta definida de la siguiente forma: En el capítulo 2, se discute el planteamiento del problema y la justificación asociada. En el capítulo 3, se mencionan los objetivos estadísticos para dar solución al planteamiento del problema. En el capítulo 4, se realiza la revisión de literatura correspondiente a los antecedentes usados para la modelación del bajo peso y las distintas metodologías para abordar el problema. En el capítulo 5, se presenta el marco teórico contextual del problema de bajo peso y el marco teórico estadístico sobre el modelo de regresión logístico. En el capítulo 6, se plantea la metodología utilizada para el respectivo análisis de los datos. En el capítulo 7 y 8, se presenta la discusión de los resultados obtenidos, las conclusiones a las que se llegaron y algunas recomendaciones para futuros estudios.

# Capítulo 2

# Planteamiento del problema y Justificación

## 2.1. Planteamiento del problema

La unidad de recién nacidos del Hospital Universitario del Valle (HUV) presenta una alta incidencia de neonatos con bajo peso al nacer (BPN), la cual representa el 75 % de las muertes perinatales, constituyendo este uno de los temas más importantes en el área de la salud materno perinatal (Torres-Muñoz et al., 2016). Según la Organización et al. (2003), se estima que cada año nacen en el mundo unos 15 millones de bebés antes de llegar a término (37 semana de gestación), es decir, más de uno en 10 nacimientos. El bajo peso al nacer se refiere a los bebés que nacen con un peso inferior a 5 libras y 8 onzas (2.500 gramos). Además, definen que un peso al nacer menor a 3 libras y 4 onzas (1.500 gramos) se considera extremadamente bajo.

El bajo peso al nacer está relacionado con la incidencia de enfermedades y mortalidad infantil y neonatal, siendo causa principal de alto riesgo de padecer problemas cardíacos, infecciones, asfixia perinatal, aspiración de meconio, hipotermia, hipoglucemia, hipocalcemia, policitemia, desnutrición, parálisis infantil, deficiencias mentales y trastornos del aprendizaje (Velázquez Quintana et al., 2004). Además, se ha encontrado que en la etapa adulta los niños que han nacido con bajo peso tienen una mayor predisposición a diabetes y enfermedades cardiovasculares. El estudio sobre el bajo peso es un indicador antropométrico que se utiliza para evaluar el estado nutricional y de crecimiento infantil; por lo tanto, es un indicador de desigualdad social.

En el HUV la alta incidencia de bajo peso al nacer ha causado preocupación y ha llamado la atención de realizar diferentes estudios estadísticos que permitan modelar ese comportamiento. El Hospital Universitario del Valle es el centro de referencia para embarazadas de alto riesgo en la ciudad de Cali y de la región suroccidental Colombiana (Bermudez et al., 2015), presenta altas tasas de incidencia en neonatos con bajo peso

al nacer, lo que conlleva a considerar de suma importancia desarrollar un análisis con datos longitudinales, es decir, con mediciones realizadas en diferentes instante del tiempo al mismo neonato y analizar las relaciones de causalidad entre el peso del neonato y diferentes covariables.

El BPN es un problema presente hoy día, con importantes repercusiones para el futuro de nuestra sociedad, por ello es indispensable identificar los factores de riesgos involucrados. En particular, este proyecto busca determinar las variables que describan la evolución de neonatos prematuros con bajo peso al nacer y la construcción de un modelo logístico. El objetivo principal del ajuste, es modelar el peso de los neonatos a través del tiempo y realizar inferencias acerca de los parámetros del modelo estimado. Con el fin de generar información a los pediatras para la orientación de acciones de control y estrategias de prevención para el cuidado del neonato.

## 2.2. Justificación

Según Hachuel et al. (2006), el bajo peso al nacer constituye un problema de Salud Pública en Colombia, no sólo por su alta morbilidad y mortalidad infantil, sino también por las secuelas que puede ocasionar en la edad adulta (hipertensión arterial, diabetes, obesidad entre otros). Menciona que el BPN es desde el punto de vista epidemiológico un trazador para la identificación de determinantes sociales en diferentes condiciones de vida, independientemente de las características biológicas de la madre.

La neonatología del Siglo XX logró disminuir la morbilidad perinatal y a pesar de ello, el BPN no ha dejado de ser un problema de salud pública, presente en 90 % de los nacimientos en los países no desarrollados, con una mortalidad neonatal para América Latina 35 veces mayor que la esperada (Torres et al., 2006). En los Estados Unidos el porcentaje de bajo peso al nacer es 6.8 %, en Colombia la prevalencia es del 11 %, en el Instituto del Seguro Social en Bogotá es del 25 % y en la unidad de recién nacidos del Hospital Universitario del Valle (HUV) en Cali, el 75 % de los neonatos tienen un peso <2,500 gramos (Bermudez et al., 2015).

Se han realizado investigaciones donde se ha encontrado que la formulación de un modelo estadístico apropiado, permite intentar responder como los contextos sociales afectan los resultados y el riesgo en la salud (Hachuel et al., 2006). Es decir, que mediante la modelación podemos encontrar los factores que inciden en el desarrollo del bajo peso al momento del nacimiento del bebe e involucrar aspectos sociales que contribuyen en la incidencia del mismo.

Ahora bien, se define como población objetivo a los neonatos pertenecientes al Programa Madre Canguro de la Unidad de Recién Nacidos del HUV, la cual inicio en Agosto de 2002, como una alternativa segura y de bajo costo para el cuidado de los bebes de bajo peso al nacer. Los beneficiarios de esta investigación, serán los niños(as) en estudio, sus familias, el

HUV y la comunidad en general, puesto que al determinar la influencia de factores de riesgo en los recién nacidos con bajo peso, se permitirá obtener un mejor enfoque y seguimiento de los casos. Además, el mostrar el probable efecto que pueden tener los factores estudiados, se podrá implementar medidas de prevención.

Esta investigación busca determinar las variables que describan la evolución de neonatos prematuros con bajo peso al nacer y la posible construcción de un modelo logístico. Se utiliza la regresión logística como técnica para modelar como influyen un conjunto de variables regresoras en la probabilidad de ocurrencia de tener un peso adecuado o bajo en la edad gestacional. El objetivo principal del ajuste de este modelo, es realizar inferencias acerca de los parámetros. La importancia de este proyecto se base en modelar el peso de los neonatos a través del tiempo, con el fin de generar una gran ventaja para los pediatras que realizan seguimiento en la orientación de acciones de control y estrategias de prevención para el cuidado del neonato.

Finalmente, el impacto social esperado es netamente preventivo. Es una ayuda para el personal de salud que tiene relación directa con el control prenatal y la recepción de recién nacido, las entidades administradoras de salud que elaboran políticas, mediante la disposición de datos concernientes a factores de riesgo relacionados con BPN, y de esta manera promoverán intervenciones preventivas en los diferentes niveles de atención pública (Cruz Montesinos et al., 2013).

## Pregunta de Investigación

El bajo peso al nacer incrementa la morbilidad y mortalidad infantil, además del tiempo y costo de hospitalización, constituyendo un verdadero problema de salud pública. Por eso, es importante intervenir en la prevención y se expone el siguiente interrogante:

1. ¿ Cuáles son las variables que describen y permiten modelar la evolución del neonato prematuro con bajo peso al nacer perteneciente al programa de alto riesgo y canguro del HUV?

# Capítulo 3

# Objetivos

## Objetivo general

- Modelar la evolución de neonatos prematuros con bajo peso al nacer, atendidos en el Hospital Universitario del Valle durante el periodo 2002 al 2010, teniendo en cuenta el peso en la edad gestacional según Ballard.

## Objetivos específicos

- Identificar las posibles relaciones existentes entre las variables que explican la evolución del neonato.
- Determinar las variables que explican el peso del neonato a través del tiempo.
- Modelar la evolución del neonato prematuro con bajo peso al nacer a través del tiempo, mediante un modelo de regresión logístico.

# Capítulo 4

# Antecedentes

El BPN, fue definido por la Organización Mundial de la Salud (OMS) en el año 1960 y en la clasificación internacional de enfermedades como el peso menor de 2.500 gramos, identificándolo como el principal factor determinante de la mortalidad neonatal e infantil (Organización et al., 2003). Hasta el momento se han realizado diferentes estudios para encontrar los principales factores que caracterizan al BPN, debido a la importancia de esté indicador de desigualdad social.

Bermudez et al. (2015), realizó un estudio en el HUV debido a la alta incidencia de neonatos prematuros con bajo peso, y modeló la evolución de neonatos con BPN pertenecientes al Programa Madre Canguro de la Unidad de Recién Nacidos del HUV; Torres et al. (2006) en el articulo de Programa Madre Canguro del HUV, encontraron que el programa es una alternativa segura para el manejo de los niños con bajo peso al nacer, pues les garantiza un egreso temprano, contacto directo con la madre, un crecimiento adecuado y una alimentación inicial basada en la leche materna. El objetivo principal del estudio de (Bermudez et al., 2015), fue encontrar las relaciones entre grupos de neonatos que mostraban un comportamiento similar respecto a las variables de crecimiento físico, morbilidad e intervención y modelar la evolución del peso del neonato a través del tiempo mediante un modelo de coeficientes aleatorios, que permitiera estimar la curva de evolución del peso de cada neonato. Encontrando que la variable que caracterizó los dos grupos de neonatos obtenidos mediante análisis multivariado fue el Control Prenatal. En esta investigación se hace la recomendación de obtener modelos que permitan evaluar las variaciones tanto para la talla como el perímetro cefálico de los neonatos en futuros trabajos.

Berhie and Gebresilassie (2016), investigaron en Etiopía sobre los factores determinantes socioeconómicos, demográficos, médicos, conductales y ambientales de la mortinatalidad fetal, tomando como referencia la Encuesta Demográfica y de Salud para el año 2011 y como marco de muestreo el censo de población y vivienda realizado por la Agencia Central de Estadística (ASC) del año 2000. Para el desarrollo de la investigación se realizó un muestreo estratificado en dos etapas. Este estudio reveló que la tasa de nacidos muertos entre las

mujeres en edad de procrear era aproximadamente 25.5 por 1000 partos en Etiopía. Además, los factores como el nivel de educación, la paridad, el índice de masa corporal (IMC) y el nivel de anemia se asociaron significativamente con la experiencia de muerte fetal en el modelo de regresión logística binaria ajustado, lo que es consistente con la mayoría de los estudios en la literatura.

Navarro Manotas et al. (2015), hicieron una investigación en la ciudad de Barranquilla y su área metropolitana, con el objetivo de ajustar un modelo logístico polinómico que permitiera encontrar los factores que tienen mayor influencia con el bajo peso en neonatos. En el artículo, se explican las posibles causas y consecuencias del bajo peso al nacer encontradas con mayor frecuencia en la literatura médica. Para el desarrollo de la investigación, se utilizó una muestra de 200 registros proporcionados por el Departamento Administrativo de Estadística (DANE) para el año 2008. La variable de interés fue el peso del nacido vivo, la cual fue categorizada en tres niveles: bajo peso al nacer ($<$ 2.500 gr), peso deficiente (2.500 gr a 2.999 gr) y peso normal ($\geq$ 3.000 gr); donde la variable de referencia es el peso normal. Finalmente, se encontró que la talla del nacido, el tiempo de gestación y el sexo del neonato son factores que influyen significativamente en el peso del neonato; por ejemplo, se estima un riesgo similar de bajo peso del nacido cuando se comparan tallas y el tiempo de gestación independientemente del sexo, se determinó que el riesgo de tener un hijo con bajo peso siempre es mayor cuando el tiempo de gestación es menor; de igual manera, se evidencio que en general se tiene mayor riesgo de nacer con peso deficiente que con peso bajo, de acuerdo con el modelo y la muestra tomada.

Cruz Montesinos et al. (2013) realizó una investigación en Ecuador sobre los factores de riesgo perinatales en el bajo peso; para ello, llevo acabo un estudio epidemiológico analítico retrospectivo de casos y controles de la historia Gineco Obstétrica materna, con el fin de conocer la influencia de los factores de riesgo perinatales en los recién nacidos a término de bajo peso y compararlos con recién nacidos a término con peso adecuado, en el Hospital Gíneco-Obstétrico (HGOIA) de la ciudad de Quito en el año 2012. Encontrando, que en Ecuador la prevalencia de bajo peso al nacer fue del 16 % en la zona urbana y el 19 % en la zona rural y para el año 2004 se reportó una prevalencia del 16.1 %. Finalmente, en esta tesis se concluye que la prevalencia de bajo peso en recién nacidos a término en el HGOIA es 8.48 %, existiendo una disminución porcentual en relación a los años previos.

Con los datos utilizados en Bermudez et al. (2015), en este proyecto se trabajará con un modelo de regresión logístico. Esta técnica de la regresión se originó en la década de los 60 con el trabajo de Cornfield, Gordon y Smith. El modelo ha sido utilizado por muchos años, pero no fue hasta que Truett, Cornfiel y Kannel (1967) que aplicaron el modelo utilizando los datos de Framingham, el cual trata de un estudio del corazón, donde se pudo apreciar el poder y la aplicación de estos modelos (Flores, 2002). Su uso se incrementa desde principios de los 80 como consecuencia de los adelantos ocurridos en el campo de la

computación (Domínguez Domínguez, 2010). Está técnica estadística resulta especialmente útil para identificar factores de riesgos y factores de prevención de enfermedades.

Hachuel et al. (2006), mencionan que se han realizado investigaciones donde se ha encontrado que la formulación de un modelo estadístico apropiado, permite intentar responder como los contextos sociales afectan los resultados y el riesgo en la salud. Es decir, que mediante la modelación podemos encontrar los factores que inciden en la evolución del BPN al momento del nacimiento del bebé e involucrar aspectos sociales que contribuyen en la incidencia del mismo.

Por otro lado, en la Universidad Nacional de Rosario en Venezuela se realizó una investigación utilizando un modelo logístico para el estudio del bajo peso al nacer, encontrando que este modelo permite obtener un enfoque apropiado para el análisis de datos longitudinales. Se contó con variables explicativas como la edad, nivel educacional máximo alcanzado, situación de convivencia, cantidad de controles realizados durante el embarazo, condición de primípara y forma de terminación del parto (Hachuel et al., 2006).

# Capítulo 5

# Marco Teórico

En este capítulo se muestran los conceptos más importantes asociados al bajo peso al nacer como: definición, clasificación, factores de riesgos, crecimiento y desarrollo y la importancia de las curvas de crecimiento como herramienta de control del crecimiento de un niño. Adicional, se explica la teoría del tipo de modelo que se utilizará para modelar la evolución del peso del neonato prematuro a través del tiempo.

## 5.1. Marco Teórico Contextual

### 5.1.1. Bajo Peso al Nacer en Recién Nacidos

Según la Organización et al. (2003), el bajo peso al nacer se refiere a los bebés que nacen con un peso inferior a 2.500 g independientemente de la edad gestacional. Según Cruz Montesinos et al. (2013), todo neonato con bajo peso al nacer (BPN), presenta mayor riesgo de morbilidad y mortalidad en los primeros veintiocho días y el primer año de vida en comparación al neonato con peso adecuado al nacer. Además, menciona que el bajo peso al nacer está relacionado con diferentes factores perinatales directos e indirectos.

La prevalencia de bajo peso al nacer debido a la prematurez o al retraso del crecimiento intrauterino (RCIU), en países desarrollados es del 4 % al 8 %, en vía de desarrollo la prevalencia puede ser del 1 % al 15 % y en Colombia según el Ministerio de Salud es del 9 % (Cruz Montesinos et al., 2013). En el Hospital Universitario del Valle en Cali, tiene una la prevalencia del 19.5 % y se encuentra asociada con el 75 % de las muertes perinatales (Herrera et al., 2013); motivo por el cual se considera un problema de salud pública que se debe abordar desde la perspectiva de la promoción y la prevención de forma tal que incida en el mejoramiento en las condiciones de vida de la población. Además, el BPN es un trazador para la identificación de desigualdades en el proceso salud, enfermedad y atención ya que es sensible a diferentes condiciones de vida (Paraje, 2009).

### 5.1.2. Clasificación del BPN

De acuerdo con Rodríguez et al. (2008) el BPN es clasificada como:

- Bajo peso al nacer (BPN), los recién nacidos pesan menos de 2500 g, ya sea debido a la prematuridad, debido a que son pequeños para su edad gestacional, o ambas cosas.
- De peso muy bajo al nacer (MBPN), los recién nacidos pesan menos de 1500 g (3 lb 5 oz) al nacer.
- De peso extremadamente bajo al nacer (EBPN) los recién nacidos pesan menos de 1000 g (2 libras 3 onzas) al nacer.

### 5.1.3. Factores de Riesgos Perinatales para el BPN

Actualmente, los factores que incrementan la posibilidad de presentar BPN son (Velázquez Quintana et al., 2004):

- **Socio demográficos maternos:** edades cronológicas extremas, relación de pareja, bajo nivel escolar, etnia, condiciones económicas desfavorables, hacinamiento (cuatro personas o más en un dormitorio) y la altura geográfica de residencia.
- **Riesgos médicos anteriores al embarazo:** antecedente de bajo peso al nacer, enfermedades crónicas (hipertensión arterial crónica, cardiopatías, nefropatías), multíparidad y estado nutricional materno.
- **Riesgos médicos del embarazo actual:** preeclampsia, eclampsia, anemia, infección urinaria, hemorragias del primero, segundo y tercer trimestre de la gestación, ganancia de peso insuficiente durante la gestación, primíparidad y período intergenésico corto (menor a 24 meses).
- **Cuidados prenatales inadecuados:** sea porque estos se inicien de forma tardía o porque el número de controles durante la gestación sea insuficiente
- **Riesgos ambientales y hábitos tóxicos:** incluye trabajo materno excesivo, estrés excesivo, tabaquismo, alcoholismo y drogadicción

### 5.1.4. Crecimiento y Desarrollo

El crecimiento y desarrollo es un proceso observacional que permite analizar los cambios físicos mediante medidas antropométricas como mediciones cuantitativas, por ejemplo la talla, el peso, circunferencia craneana, torácica y abdominal. El concepto de desarrollo abarca tanto a la maduración en los aspectos físicos, cognitivos, lingüísticos, socioafectivos y temperamentales como el desarrollo de la motricidad fina y gruesa (Serrano, 2002). Este

proceso observacional en un niño es de gran utilidad para determinar el estado de salud en edad pediátrica, basado en mediciones repetidas en diferentes instantes del tiempo.

El desarrollo es el proceso por el cual los seres vivos logran mayor capacidad funcional de sus sistemas, comprende un aumento de la complejidad y destreza de una persona para adaptarse al medio. La consulta pediátrica en la etapa de la infancia es básica para asegurarse de que el niño se esta desarrollando dentro de los limites establecidos como normales.

### 5.1.5. Curvas de Crecimiento

Las curvas de crecimiento son una herramienta para registrar y evaluar el crecimiento de los niños a través del tiempo. El objetivo principal del seguimiento mediante estas curvas, es que el pediatra y los padres del niño conozcan la evolución del crecimiento, con el fin de desarrollar el máximo potencial ya sea del peso o la talla y detectar o corregir a tiempo posibles alteraciones en el proceso. Según la Organización Mundial Organización et al. (2003), las curvas muestran como debería ser el crecimiento de los niños y niñas menores de cinco años, cuando sus necesidades de alimentación y cuidados de salud son satisfechas en cualquier parte del mundo. Por otro lado, se entiende que el crecimiento no es solo resultado de la nutrición sino de factores heredados.

Mediante estas curvas se pueden ajustar diferentes modelos matemáticos, que permitan realizar estimaciones de los parámetros, con el objetivo de interpretarlos y brindar ayuda para caracterizar la población y predecir algunos parámetros valiosos, como el peso que alcanzaría la población en crecimiento o el peso promedio de la misma.

## 5.2. Marco Teórico Estadístico

### 5.2.1. Modelo Logístico

Con el objetivo de modelar la evolución a través del tiempo del neonato que nace con bajo peso, perteneciente al programa Madre Canguro del HUV, se considera trabajar con un Modelo de Regresión Logístico, que permita encontrar los factores que describan la evolución del peso y las relaciones existentes entre ellos. El modelo regresión logístico o modelo logit, es un modelo clásico de regresión lineal simple o múltiple, donde la variable dependiente es dicótomica. Es decir, adopta solo dos posibles valores (Hosmer Jr et al., 2013).

La regresión logística es un modelo especial que se utiliza para explicar y predecir una variable categórica en función de varias variables independientes que a su vez pueden ser cuantitativas o cualitativas. Este tipo de modelos se usa comúnmente en las ciencias médicas y sociales (Hachuel et al., 2006). Es un modelo con buena capacidad para el análisis de datos en el área de investigación clínica y epidemiologica. Permite encontrar la mejor adaptación,

de tal forma que se pueda interpretar y describir las relaciones existentes entre las variables independientes y la variable dependiente.

Una de las características más importantes de este modelo es que la variable respuesta es binaria o dicótomica (Hosmer Jr et al., 2013). Además, tiene facilidad de uso y la estimación de los parámetros permite obtener información significativa para efectos clínicos (Montgomery et al., 2012).

Hosmer Jr et al. (2013) existen dos importantes diferencias entre el modelo de regresión lineal y logístico:

1. La relación entre el valor esperado y las variables independientes: Para la regresión logística el valor esperado de la variable respuesta dado un valor de la covariable, viene definida como:

$$E(Y/X) = \beta_0 + \beta_1 X \qquad -\infty < X < \infty \qquad 0 \leq E(Y/X) \leq 1 \tag{5.1}$$

   La siguiente expresión hace referencia a la forma especifica del modelo de regresión y la transformación logit asociada (Hosmer Jr et al., 2013):

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{5.2}$$

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 X \tag{5.3}$$

   Al realizar la transformación se tiene que *g(x)* tiene muchas de las propiedades deseables de un modelo de regresión logístico (Hosmer Jr et al., 2013).

2. La distribución condicional de la variable respuesta, es una de las diferencias entre el modelo logístico y lineal (Hosmer Jr et al., 2013) .

   En la regresión lineal se tiene que el error asociado $\varepsilon$ expresa la desviación respecto a la media condicional y a su vez se distribuye normal con media cero y varianza constante. Para el caso de la regresión logística no ocurre esto, dado que la variable respuesta es dicótoma, el error $\varepsilon$ solo puede tomar dos posibles valores:

   - Si Y=1

     El error viene denotado como $\varepsilon = 1 - \pi(x)$ con una probabilidad asociada igual a $\pi(x)$.

- Si Y=0

  El error viene denotado como $\varepsilon = -\pi(x)$ con una probabilidad asociada igual a $[1 - \pi(x)]$.

Finalmente, se obtiene que $\varepsilon$ tiene media cero y varianza igual a $\pi(x)[1 - \pi(x)]$. Ahora bien, la distribución condicional se distribuye binomial con probabilidad dada por la media condicional $\pi(x)$.

### Variables Explicativas Categóricas

En el caso de que las variables explicativas sean categorías, se debe realizar una codificación parcial que será representada mediante variables auxiliares. Dicha codificación exige tomar una categoría de referencia, de tal manera que todas las variables auxiliares toman el valor de 0 para dicha categoría. Es decir, para las categorías restantes, la variable auxiliar toma el valor 1 para la categoría asociada y 0 para el resto, el valor para la variable de diseño $m$-ésima asociada a la categoria $A_m$ sería (Reche, 2013):

$$X_{im}^{A} = X_{im}^{A}|(A = A_i) = \begin{Bmatrix} 1 & i = m \\ 0 & i \neq m \end{Bmatrix} \forall m = 2...., I; i = 1, ...I \tag{5.4}$$

Por otro lado, se debe tener en cuenta que en un modelo de regresión logístico debe haber por lo menos una variable predictora cuantitativa. Si todas las variables predictoras son cualitativas entonces el problema se convierte en uno de diseños experimentales (Hosmer Jr et al., 2013).

### Estimación de los Parámetros

Inicialmente se debe realizar la codificación pertinente a las variables estudiadas, asignando el valor de 1 cuando hay presencia de la característica y 0 representando la ausencia de la misma.
Una vez realizada la codificación, se debe obtener todos los posibles modelos con las variables estudiadas y mirar si existe interacción entre estas. Ahora, para realizar el ajuste previo del modelo se requiere estimar los valores de los parámetros desconocidos $\beta_0$, $\beta_1$, ..., $\beta_i$, para ello se pueden utilizar los siguientes métodos:

- **Método de mínimos cuadrados:** Se minimizan las desviaciones respecto a la suma de cuadrados de los valores observados versus los predichos en el modelo. Hosmer Jr et al. (2013), menciona que en la vida cotidiana este método no es efectivo debido a que la variable respuesta es dicótoma y los estimadores obtenidos no cumplen las propiedades requeridas.

Por otro lado, si se cumplen los supuestos de regresión lineal, se dice que la función de verosimilitud viene denotada como (Montgomery et al., 2012):

$$f(x_i) = \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \tag{5.5}$$

$$l(\beta) = \prod_{n}^{i=1} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \tag{5.6}$$

Donde $y_i$ denota el valor de una variable dicótoma y $x_i$ el valor de la variable independiente de tipo cuantitativo y/o cualitativo. Ahora bien, para maximizar los valores de $\beta$, se utiliza la diferenciación de la función log-verosímil respecto a cada coeficiente.

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} (y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]) \tag{5.7}$$

$$\sum [y_i - \pi(x_i)] = 0$$

$$\sum x_i [y_i - \pi(x_i)] = 0$$

Reche (2013) mencionan los siguientes métodos adicionales para la estimación de los parámetros de un modelo de regresión logístico:

- **Estimación por el método de Newton-Raphson:** Es un algoritmo para encontrar aproximaciones de los ceros o raíces de una función real.
- **Estimación por métodos de optimización generales:** Consiste en una función objetivo y un conjunto de restricciones en la forma de un sistema de ecuaciones o inecuaciones.
- **Estimación por mínimos cuadrados iterativos ponderados:** Estimación de los parámetro de un modelo lineal generalizado cuya variable respuesta tiene distribución multinomial y de poisson.

La función *glm* en el software estadístico R utiliza la estimación por mínimos cuadrados iterativamente ponderados. Está estimación parte de una aproximación local cuadrática a la función de log-verosimilitud, la maximización de esta aproximación se realiza mediante un modelo lineal ponderado. Suponiendo $\beta^{(t)}$ el vector que contiene la estimación de los parámetros del GLM en la iteración $t$ . Entonces $\eta_i^{(t)} = x^{'}\beta^{(t)}$ es el predictor lineal para la observación $i$-ésima y $\mu_i^{(t)} = g^{-1}(\eta_i^{(t)})$ los valores ajustados (probabilidades ajustadas en el caso de regresión logística y usando la función logit como función de enlace $g$, la función de la varianza sería $v_i^{(t)} = Var(\mu_i^{(t)})\phi$, donde $\phi$ es el parámetro de dispersión o escala (Hosmer Jr et al., 2013).
La siguiente expresión hace referencia a unos valores de respuesta intermedios o de trabajo. Se les llama de trabajo porque cambian en cada iteración:

$$z_i^t = \eta_i^t + (y_i - \mu_i^t)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{(t)} \tag{5.8}$$

y se tienen unas ponderaciones de trabajo

$$w_i^{(t)} = \frac{1}{c_i v_i^{(t)}\left[\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{(t)}\right]^2} \tag{5.9}$$

Donde $c_i$ son constantes fijas para la familia binomial $c_i = n_i^{-1}$.
El algoritmo ajusta una regresión por mínimos cuadrados ponderados de $z^{(t)}$ sobre los predictores lineales, minimizando la suma de cuadrados ponderada $\sum_{i=1}^{n} w_i(z_i - x_i^{'}\beta)^2$, donde $x_i^{'}$ es la fila $i$-ésima de la matriz regresora (incluyendo la columna de unos). Finalmente, de esta forma se obtienen unos nuevos parámetros $\beta^{(t+1)}$. Este proceso continua hasta que se estabilizan los parámetros, obteniendo la estimación máximo verosímil $\hat{\beta}$.

### Pruebas para la significancia de los parámetros

Una vez estimado el modelo, se realizan las inferencias necesarias, mediante pruebas estadísticas para determinar si las variables independientes en el modelo son significativas respecto a la variable respuesta (Hosmer Jr et al., 2013). A continuación se mencionan algunos métodos:

- **Suma de Cuadrados del Error:** Es la comparación de las distancias entre los valores observados y predichos:

$$SSE = \sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{5.10}$$

La comparación se realiza mediante la siguiente expresión (Reche, 2013):

$$D = -2ln\left[\frac{\text{Probabilidad del modelo ajustado}}{\text{Probabilidad del modelo saturado}}\right] \tag{5.11}$$

Es equivalente a tener:

$$D = -2ln\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \tag{5.12}$$

$$D = -2ln(\text{Probabilidad del modelo ajustado}) \tag{5.13}$$

El anterior estadístico es conocido como deviación, y para el modelo de regresión representa la suma de cuadrados residual de la regresión lineal.

- **Contraste de Wald:** Se utiliza para poner a prueba el verdadero valor del parámetro basado en la estimación de la muestra. Este contraste está basado en la normalidad asintótica de los estimadores. Se quiere contrastar si un parámetro $\beta_r = 0$, con r = 1, . . . ,R, frente a que no lo sea (Reche, 2013).

$$H_0 : \beta_r = 0 \qquad\qquad H_1 : \beta_r \neq 0$$

Wald, demostró que bajo la hipótesis nula, el estadístico de contraste es:

$$W_r = \frac{\hat{\beta}}{\hat{SE(\beta_r)}} \tag{5.14}$$

Donde $\hat{\beta}_r$ es la estimación del parámetro $\beta_r$ y $\hat{SE(\beta_r)}$ el error estándar asociado para $r = 1, 2, 3....$

- **Contraste condicional de razón de verosimilitud:** Las hipótesis de este contraste son las mismas que en el anterior:

$$H_0 : \beta_r = 0 \qquad\qquad H_1 : \beta_r \neq 0$$

pero el enfoque es distinto. En vez de considerar la distribución de los parámetros se comparan dos modelos, uno dónde se haya estimado el parámetro $\beta_r$, frente a otro modelo que se diferencie del primero en que no esté dicho parámetro, Es decir, comparamos modelos anidados. El test que se utiliza es el test condicional de razón de verosimilitudes.

$$G^2_{modelo1}|G^2_{modelo2} = -2log\frac{V_{modelo1}}{V_{modelo2}} = -2(L_{modelo1} - L_{modelo2}) = G^2_{modelo1} - G^2_{modelo2} \tag{5.15}$$

Donde $V$ es la máxima verosimilitud y $L$ la máxima log-verosimilitud en cada modelo. En otras palabras El estadístico representa, la disminución en la devianza que se produce al pasar de un modelo a otro al añadir una variable.

Se rechaza la hipótesis nula con un nivel de significancia $\alpha$ cuando:

$$\left(G^2_{modelo1} - G^2_{modelo2}\right) \geq \chi^2_{1,\alpha} \tag{5.16}$$

**Intervalos de Confianza**

Es importante resaltar que las estimaciones no son exactas, para ello se emplean los intervalos de confianza que permiten obtener un rango de variación de los posibles valores que pueden tomar los coeficientes del modelo .

- **Estimación mediante test de Wald:** En las siguientes expresiones $\hat{SE}$ representa la estimación del error estándar basado en los parámetros del modelo. $\beta_r$ sigue una distribución asintóticamente normal (Hosmer Jr et al., 2013) :

$$P\left[-z_{\alpha/2} \leq \frac{\hat{\beta}_r - \beta_r}{SE(\hat{\beta}_r)} \leq z_{\alpha/2}\right] = 1 - \alpha \tag{5.17}$$

$$\hat{\beta}_r \pm z_{-\alpha/2}\hat{S}E(\hat{\beta}_r)$$

La siguiente estimación es para la parte logit del modelo:

$$g(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{Var}[\hat{g}(x)] = \hat{Var}[\hat{\beta}_0] + x^2\hat{Var}[\hat{\beta}_1] + 2x\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

$$\hat{g}(x) \pm z_{-\alpha/2}\hat{SE}[\hat{g}(x)]$$

Estimación para los valores puntuales:

$$\frac{e^{\hat{g}(x)\pm z_{-\alpha/2}\hat{SE}[\hat{g}(x)]}}{1 - e^{\hat{g}(x)\pm z_{-\alpha/2}\hat{SE}[\hat{g}(x)]}} \tag{5.18}$$

La interpretación de un intervalo es que si se repite la muestra aleatoriamente un elevado número de veces y en cada una de ellas construyendo un intervalo de confianza para el parámetro $\beta_r$, aproximadamente $(1-\alpha)\,\%$ de los intervalos contendrían al verdadero valor del parámetro.

**Intervalos de Confianza para los $e^{\beta_r}$**

Los siguientes test permiten obtener los intervalos de confianza para los cocientes de ventajas o oportunidad:

$$H_0 : e^{\beta_r} = 1 \qquad\qquad H_1 : e^{\beta_r} \neq 1$$

- **Test condicional de razón de verosimilitud:** La idea consiste en invertir el contraste condicional de razón de verosimilitudes para obtener los I.C, para eso se considera una aproximación a través de la denominada profile-likelihood. Estos intervalos tienen las características de ser más precisos y de que no tienen por qué ser simétricos (Montgomery et al., 2012).

  El objetivo es la maximización de la función de log-verosimilitud con la restricción de que $\theta_j = \beta_0$. Un intervalo de confianza para $\theta_j$ basado en esta función, viene dado por:

$$\bar{L}_j(\beta) = max_{\theta\epsilon\Theta_1} L(\theta) \tag{5.19}$$

$$\left\{\beta| - 2[\bar{L}_j(\beta) - L(\hat{\theta}))] \leq q_1(1-\alpha)\right\} \tag{5.20}$$

Donde $q_1(1-\alpha)$ es el cuantil $(1-\alpha)$ de una distribución chi-cuadrado con un grado de libertad.

### Modelo de Regresión Logístico Múltiple

Según Hosmer Jr et al. (2013), el modelo de regresión múltiple viene denotado como:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{5.21}$$

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \tag{5.22}$$

Recordemos, que para este caso se utilizan las variables auxiliares y solo se necesitan (*k-1*) variables:

$$g(x) = \beta_0 + \beta_1 X_1 + ... + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p X_p \tag{5.23}$$

El proceso de estimación de los coeficientes del modelo y las pruebas de significancias de los coeficientes son equivalentes al caso univariable (Bewick et al., 2005).

## 5.2.2. Ajuste del Modelo

Mediante el ajuste del modelo se desea seleccionar el modelo más parsimonioso y mejor ajustado a los datos.

### Valores ajustados y predicciones

Los valores estimados por el modelo están guardados dentro del objeto glm (función del software estadístico R Studio), y se puede acceder a ellos mediante el operador pesos. A estos mismo valores se puede acceder con la función fitted.values.

### Residuos

El análisis de los residuos es fundamental para evaluar la adecuación del modelo y detectar los valores anómalos e influyentes (Reche, 2013). Se enumera a continuación los distintos tipos de residuos que se pueden obtener para un modelo de regresión logístico:

- Residuos de Respuesta: Diferencia entre el valor observado y el valor estimado por el modelo (Hosmer Jr et al., 2013).

- Residuos de Pearson: Son los $q$-ésimos componentes del estadístico $X^2$ de Pearson de bondad del ajuste global:

$$ep_q = \frac{y_q - n_q\hat{p_q}}{\sqrt{n_q\hat{p_q}(1-\hat{p_q})}} \tag{5.24}$$

  donde los valores estimados por el modelo en cada perfil son $n_q\hat{p_q}$. Se consideran significativos aquellos residuos cuyo valor absoluto sea mayor a 2.

- Residuos de Pearson estandarizados: Se trata de una modificación de lo anteriores, de forma que dichos residuos tengan distribuciones asintóticas normales estándar. Serán significativos aquellos residuos cuyo valor absoluto sea mayor a 2:

$$eps_q = \frac{ep_q}{(1-h_q)^{1/2}} \tag{5.25}$$

### Medidas de bondad del ajuste

En la práctica, no hay garantía de que un modelo de regresión logística se ajuste bien a los datos. Cuando los datos están en forma binaria, una manera de detectar la falta de ajuste es realizando contrastes condicionales de razón de verosimilitudes entre modelos anidados.

### Estadístico $G^2$ de Wilks de razón de verosimilitudes

El estadístico tiene la siguiente expresión:

$$G^2(M) = 2[L_S - L_M]$$

dónde $L_M$ es la log-verosimilitud del modelo ajustado y $L_S$ la log-verosimilitud del modelo saturado. Este estadístico tiene, bajo la hipótesis nula de que el modelo $M$ es adecuado, una distribución asintótica $X^2$ con $Q$ menos ($R$+1) grados de libertad, dónde $Q$ es el número de combinaciones de valores de las $R$ variables explicativas.

### Estadístico $X^2$ de Pearson

Este estadístico tiene, bajo la hipótesis nula de que el modelo es adecuado, la misma distribución asintótica que el estadístico de Wilks. El estadístico $X^2$ de Pearson se puede calcular fácilmente partiendo de los residuos de Pearson. Y el $p$-valor lo obtenemos de nuevo, con la función pchisq.

$$X^2(M) = \sum_{q=1}^{Q} \frac{(y_q - n_q\hat{p_q})^2}{n_q\hat{p_q}(1-\hat{p_q})} \tag{5.26}$$

Donde $\hat{p_q}$ son las probabilidades estimadas en cada combinación de variables predictoras y $n_q$ el número de casos totales en cada combinación.

### Medidas tipo $R^2$

En la regresión lineal por mínimos cuadrados tenemos una medida $R^2$, que nos ofrece una medida de la bondad del ajuste comparando sumas de cuadrados.

- **Pseudo $R^2$ de McFadden (McFadden et al., 1973):** Esta medida compara la log-verosimilitud del modelo ajustado con la log-verosimilitud del modelo que sólo tiene el término constante.

$$R^2 = 1 - \frac{\hat{L}_{Modelo}}{\hat{L}_{Intercepto}} \tag{5.27}$$

  Asemejándose a la expresión del $R^2$ en la estimación por mínimos cuadrados, dónde $\hat{L}_{Intercepto}$ sería la suma de cuadrados totales y $\hat{L}_{Modelo}$ haría el papel de la suma de cuadrado de los errores. Este $R^2$ da una idea de en cuánto se reduce la devianza de los datos al ajustar el modelo.

- **Pseudo $R^2$ de Cox y Snell (Cox and Snell, 1989):**

$$R^2_{CN} = 1 - exp\left(-\frac{\widehat{LR}}{N}\right); \quad \text{donde} \quad \widehat{LR} = \hat{L}_{intercept} - \hat{L}_{modelo} \tag{5.28}$$

  Donde $\hat{L}_{Intercepto}$ sería la suma de cuadrados totales y $\hat{L}_{Modelo}$ haría el papel de la suma de cuadrado de los errores.

- **Pseudo $R^2$ de Nagelkerke (Nagelkerke et al., 1991):** Se trata de una modificación del $R^2_{CN}$ que alcanza su máximo en:

$$maxR^2_{CN} = 1 - exp\left(-\frac{\hat{L}_{intercept}}{N}\right) \qquad R^2_N = \frac{R^2_{CN}}{maxR^2_{CN}} \tag{5.29}$$

  dónde $\hat{L}_{Intercepto}$ sería la suma de cuadrados totales y $N$ el número de perfiles ajustados.

#### Medidas basadas en la tabla de clasificación. Curvas ROC

Otra forma de evaluar el desempeño de un modelo de clasificación es a través de medidas relacionadas con la tabla de clasificación. Para determinar cómo clasifica el modelo a cada observación, se elige un punto de corte, si la probabilidad predicha por el modelo es mayor que el punto de corte se clasifica como éxito y si es menor como fracaso.

Se considera que un modelo es mejor que otro si la curva ROC se acerca al borde superior izquierdo, o lo que es lo mismo, que el área bajo la curva sea mayor (Reche, 2013).

**Tabla 5.1**: Clasificación del Modelo. Fuente: Libro de Reche (2013)

| | Clasificación: éxito | Clasificación: fracaso |
|---|---|---|
| Éxtio | Verdaderos Positivos (VP) | Falsos Negativos (FN) |
| Fracaso | Falsos Positivos (FP) | Verdaderos Negativos (VN) |

Este tipo de análisis permite observar varias medidas relacionadas con la tabla de clasificación, y representarlas gráficamente. Nos puede ayudar a elegir el punto de corte que maximiza la tasa de clasificaciones correctas, y ver como varía esta tasa según los diferentes puntos de corte.

### 5.2.3. Métodos de selección de variables

La selección automática de variables se basa en la comparación de la devianza entre modelos. El procedimiento habitual es el llamado stepwise (paso a paso) en el que, mediante contrastes condicionales de razón de verosimilitudes se comparan modelos con diferentes variables (Hosmer Jr et al., 2013). La selección forward (hacia adelante) empieza con el modelo más simple y va añadiendo en cada paso la variable más significativa según el contraste condicional de razón de verosimilitudes (Hosmer Jr et al., 2013).
Existen diferentes métodos de selección de variables, los más utilizados son: forward, backward o stepwise, pero se tiene que los modelos pueden ser totalmente diferentes con cada uno de estos métodos (Reche, 2013). En esta investigación se utilizará el método Backward, el cuál es uno de los más usados en el área clínica. Además, se utilizará el paquete glmulti del software estadístico R Studio, para hacer una elección de variables de acuerdo al modelo de regresión logístico.

#### Selección Backward

En este caso se inicia el proceso a partir del modelo que contiene todas las posibles variables predictoras. En cada iteración se generan modelos a los que se les elimina un único predictor a la vez y se selecciona el que tiene menor suma de cuadrados de error asociado al modelo. Este proceso se repite hasta llegar al modelo nulo, sin ningún predictor. Al final, entre los

mejores modelos seleccionados se identifica mediante el criterio de información Akaike (AIC), que permite evaluar cada variable en presencia de las otras, el modelo escogido como aquel que tenga menor valor de AIC.

### Criterio de Información Akaike (AIC)

Es una medida que permite analizar la calidad de un modelo estadístico para un conjunto de datos (Martınez et al., 2009).

$$AIC = 2K - 2ln(L) \tag{5.30}$$

Donde $K$ es el número de parámetros en el modelo, y $L$ es el valor máximo de la función de verosimilitud del modelo estimado.

### Método glmulti

Este paquete del software estadístico R, utiliza la función *glm* (Nelder and Baker, 1972), para comparar automáticamente todos los posibles modelos teniendo en cuenta el modelo que se desea ajustar. Para este proyecto es el modelo de regresión logístico, por ende se utiliza la familia binomial. Esta selección genera todos los posibles modelos y encuentra los mejores bajo el criterio de información (AIC) (Calcagno et al., 2010). Se utiliza el método de detección exhaustivo, obteniendo todas las posibles combinaciones. Se especifica la función y la familia, en este caso, se especifica que es *gml* de familia *Binomial*, que representan un modelo de regresión logístico.

En comparación con las demás técnicas de selección, no es un proceso iterativo, ya que todos los modelos son comparados y es completamente automático. Se utiliza con más de 10 variables predictoras y los cálculos son bastantes largos. Se requiere la ejecución de Java y el paquete rJava, para completar el proceso.

### Modelo de Regresión Logística para el estudio de Casos y Controles

Este tipo de modelo es muy utilizado en el área de investigación clínica, debido a la facilidad de obtener odds ratios ajustados de los coeficientes. Para un estudio de casos y controles la variable respuesta se fija por estratificación y hace referencia a las variable de exposición (Hosmer Jr et al., 2013).

Según Hosmer Jr et al. (2013), la función de verosimilitud es el producto de las funciones de verosimilitud específicas del estrato y depende de la probabilidad de seleccionar al sujeto y la distribución de probabilidad de las covariables.

### 5.2.4. Diagnóstico y Validación

En esta sección se estudia la falta de ajuste a nivel de cada observación y cómo se afecta el modelo en general. En los modelos lineales generalizados, es importante realizar el análisis de los residuos ya que pueden indicar si se hizo una mala elección de la función de enlace o si el efecto del conjunto de variables explicativas no es lineal (Reche, 2013).

#### Análisis de los residuos

Para este tipo de análisis se asume que para efectos de diagnostico un residuo es significativo distinto de 0 si su valor absoluto es mayor que 2. En la subsección 5.2.2 se especifica los tipos de residuos que se pueden calcular para un modelo de regresión logístico.

#### Medidas de influencias

Las medidas de influencias son importantes de estudiar, ya que nos permiten detectar los valores influyentes que puedan generar efectos en los parámetros del modelo. Para los modelos lineales generalizados, se obtienen mediante la aproximación de medidas de influencias en (Williams, 1987) y en (Cook and Weisberg, 1982).

Las medidas de influencias más usadas son *hat values* y las *distancias de cook*, la siguiente expresión se utiliza para los modelos glm:

$$D_q = \frac{e^2_{PSq}}{k+1} * \frac{h_{qq}}{1-h_{qq}} \tag{5.31}$$

Donde $e_{PSq}$ es el error cuadrático medio del modelo de regresión, $k$ es la cantidad de variables y $h_{qq}$ son los *hat values*.

#### Colinealidad y factores de inflación de la varianza generalizada (GVIF)

En esta sección se estudia la relación lineal entre los predictores del modelo. Debido a que la presencia de colinealidad genera mayor variabilidad estimada para los coeficientes. Tal y como se describe en Fox and Weisberg (2011) no es adecuado utilizar el factor de inflación de la varianza (VIF) para los modelos con variables categóricas como predictoras. Para ello Fox and Monette (1992), generaliza la noción de la inflación de la varianza asociada, esta media se conoce como *factor de inflación de la varianza generalizado* (GVIF).

En este caso si la variable predictora tiene $p$ regresores, el valor de $GVIF^{1/2p}$ es una medida dé como disminuye la precisión de la estimación de los coeficientes debido a la existencia

de colinealidad (Reche, 2013). Para este caso, un valor próximo a 1 de $GVIF^{1/2p}$ indica ausencia de colinealidad.

$$GVIF = \frac{detR_{11}detR_{22}}{detR} \tag{5.32}$$

Donde $R_{11}$ es la matriz de correlaciones entre el conjunto de regresores en cuestión (como los $p$ regresores de una variable con $p$+1 categorías). $R_{22}$ la matriz de correlaciones entre los otros regresores del modelo y $R$ la matriz de correlaciones entre todos los regresores del modelo.

### Validación Cruzada

La validación cruzada es utilizada para analizar el sobreajuste del modelo. Con el objetivo de comprobar si el modelo predice correctamente un nuevo conjunto de datos. Para ello, se considera la técnica *K-Fold cross-validation*, consiste en dividir la muestra en $k$ submuestras, de tal forma que se utilicen *k-1* para estimar el modelo y el restante como submuestra de evaluación. Este proceso se repite $K$ veces, donde cada submuestra es utilizada una vez para evaluar el modelo y *K-1* veces el ajuste. Para concluir respecto a la validación, se utiliza la tasa de clasificaciones correctas o su complemento (Canty et al., 2012).

# Capítulo 6

# Metodología

En este capítulo se explicará la metodología implementada para dar cumplimiento a los objetivos planteados en el proyecto, se propone la aplicación de un modelo de regresión logístico, que permita modelar y distinguir aspectos importantes en la evolución de los neonatos prematuros con bajo peso al nacer, con la información disponible del neonato y la madre.

## 6.1. Población Objeto de Estudio

Son los neonatos atendidos en la Unidad de Recién Nacido del HUV entre febrero de 2002 y noviembre de 2010, pertenecientes al Programa de Seguimiento de Alto Riesgo y Canguro.

## 6.2. Unidad Experimental

Se definen las unidades experimentales como “el neonato prematuro con bajo peso al nacer"; estos deben cumplir con los siguientes criterios de inclusión para la adaptación intrahospitalaria definidos por la Unidad de Recién Nacidos:

1. Recién nacidos con menos de 2001 gr.
2. Regulación térmica y del patrón respiratorio
3. Saturación de oxígeno normal
4. Procedentes de la ciudad de Cali y que no planeen establecerse fuera de la ciudad.
5. Madre y/o familia dispuesta a colaborar con el programa de seguimiento, cumplir con las recomendaciones y asistir a los controles.

## 6.3. Consolidación de la Información

Se contó con una muestra de 145 historias clínicas, las cuales contienen 49 variables que brindan información de la madre y el neonato. Se realizó un primer filtro, donde se eliminaron las variables que no brindaban información significativa, como por ejemplo: el numeral de la historia clínica, nombre del doctor asignado, fecha de los controles de seguimiento, fecha de la próxima cita, identificador del individuo, entre otras. Luego, se efectuaron pruebas chi-cuadrado con una significancia estadística del 5 % (**Ver Apéndice .1**, con el fin de encontrar relaciones significativas entre variables cualitativas. Adicional, se realizaron consultas con pediatras del HUV, para seleccionar las variables que se consideran relevantes en el estudio desde un punto de vista medico. Finalmente, se obtuvieron 16 variables para el estudio del modelo de la evolución del peso del neonatos a través del tiempo (**Ver Figura 6.1**).

## 6.4. Análisis Exploratorio

El objetivo de realizar el análisis exploratorio es resumir y visualizar los datos de los neonatos de manera que se facilite la identificación de tendencias o patrones que los subyacen y que son relevantes para responder a la pregunta de investigación.

Inicialmente, se realizó un análisis exploratorio para el conjunto de variables de los grupos maternos y neonatales estudiados. Para conocer la distribución de frecuencias y porcentajes de neonatos asociados a cada una de las categorías de las variables estudiadas, se realizó un análisis univariado. Adicional, se llevo a cabo un análisis bivariado de peso del neonato a través del tiempo, utilizando medidas de tendencia central como el promedio e indicadores de variabilidad como la desviación estándar, el coeficiente de variación y el rango de variación asociado.

Adicional, se observo la distribución del peso del neonato por mes, mediante las curvas sin ninguna transformación en los datos. Finalmente, se analizó mediante gráfico de barras diferentes variables en relación con el peso en la edad gestacional del neonato, si es adecuado o bajo.

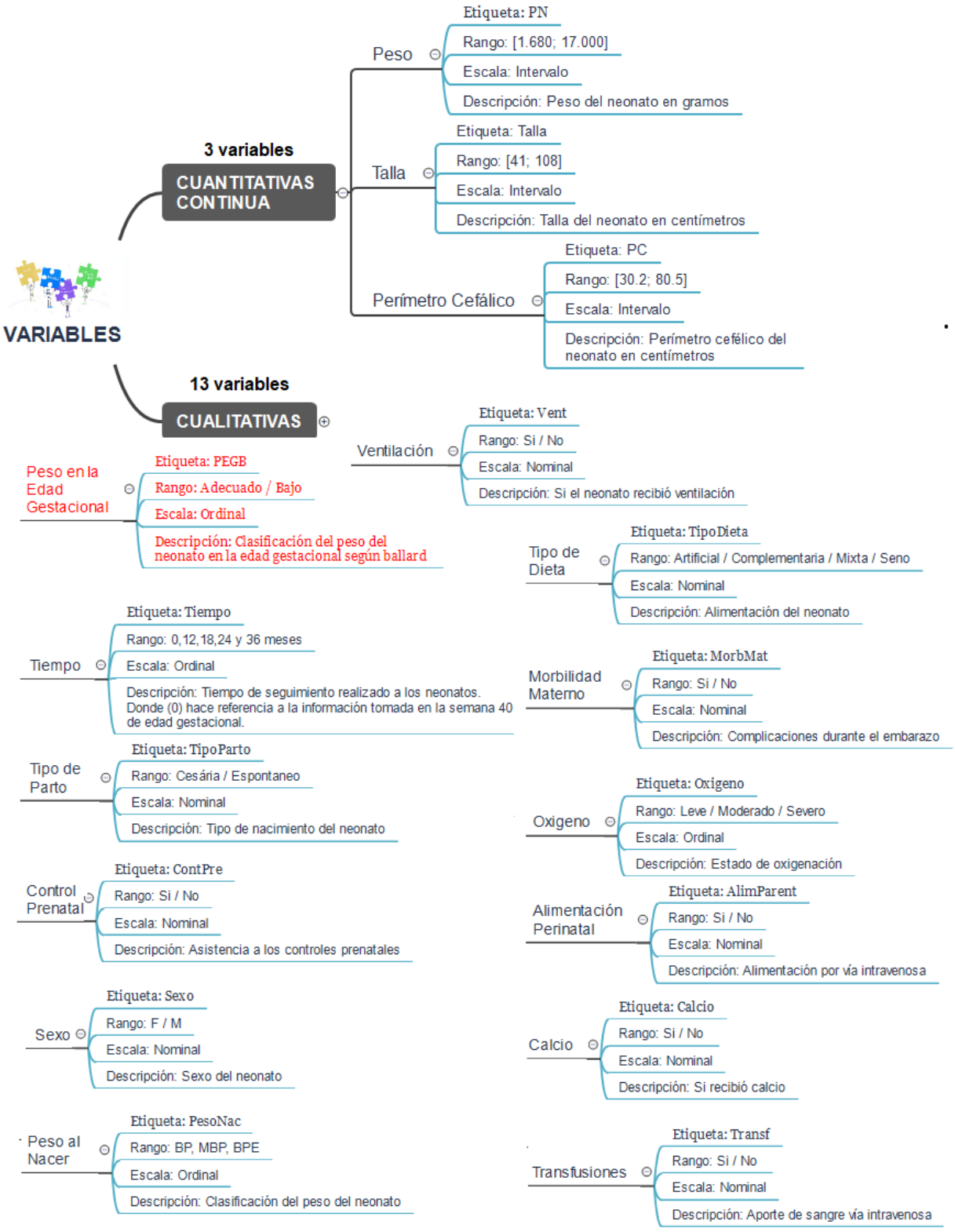


**Figura 6.1**: Conjunto de variables para la estimación del modelo.

## 6.5. Modelo Regresión Logístico

Con base en los objetivos planteados, como descripción estadística de los datos de la evolución de los neonatos prematuros, se utilizó el modelo de regresión logístico, con el fin de obtener un posible modelo que permita modelar e identificar la dinámica de la evolución del neonato prematuro.

Primeramente, se realizó una codificación parcial de las variables cualitativas, tomadas como factor de un total de 16 variables a estudiar, donde 3 son variables cuantitativas y 13 variables cualitativas. Se asigno el valor de (1) a las categorías como BNP Adecuado, si presentó control prenatal, si presentó morbilidad materna, entre otras. Variables como el sexo, tiempo de seguimiento, tipo de dieta, clasificación del peso al nacer, sexo y oxigeno se tomaron como factor (**Ver Tabla 6.1**), donde la variable respuesta es el peso en la edad gestacional según Ballard (PEGB). Luego, se observó la presencia de valores faltantes en la base de datos de los neonatos y se gráficó la frecuencia, con el objetivo de analizar cuál(es) variable(s) presentaban frecuencias altas de datos faltantes, para ser excluidas por su alta frecuencia y realizar la modelación. Para ello, se utilizó el paquete *AMELIA* del software estadístico R Studio (Honaker et al., 2010), con el objetivo de graficar la frecuencia de datos:

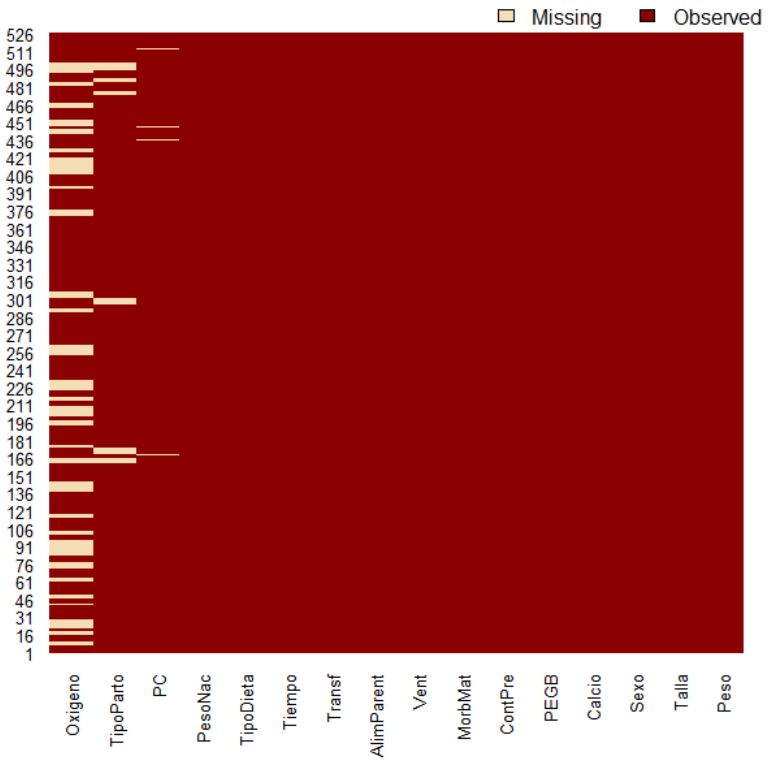


**Figura 6.2**: Frecuencia de Datos Faltantes

**Tabla 6.1**: Base de datos para los primero 10 neonatos estudiados

| | Peso | Talla | PC | Sexo | Calcio | PEGB | ContPre | MorbMat | Vent | AlimParent | Transf | Tiempo | TipoParto | TipoDieta | PesoNac | Oxigeno |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2020.00 | 44.00 | 33.00 | Masculino | 0 | Adecuado | 0 | 1 | 0 | 0 | 1 | 0 | Cesarea | Seno | BPE | Moderado |
| 2 | 7280.00 | 71.50 | 43.50 | Masculino | 0 | Adecuado | 0 | 1 | 0 | 0 | 1 | 12 | Cesarea | Complementaria | BPE | Moderado |
| 3 | 8580.00 | 74.50 | 41.50 | Masculino | 1 | Adecuado | 0 | 1 | 0 | 0 | 1 | 18 | Cesarea | Complementaria | BPE | Moderado |
| 4 | 9360.00 | 83.00 | 44.50 | Masculino | 0 | Adecuado | 0 | 1 | 0 | 0 | 1 | 24 | Cesarea | Complementaria | BPE | Moderado |
| 5 | 11570.00 | 93.00 | 46.00 | Masculino | 0 | Adecuado | 0 | 1 | 0 | 0 | 1 | 36 | Cesarea | Complementaria | BPE | Moderado |
| 6 | 2380.00 | 46.00 | 33.10 | Masculino | 0 | Adecuado | 0 | 0 | 0 | 0 | 1 | 0 | Cesarea | Mixta | MBP | Moderado |
| 7 | 8620.00 | 75.00 | 45.50 | Masculino | 0 | Adecuado | 0 | 0 | 0 | 0 | 1 | 12 | Cesarea | Complementaria | MBP | Moderado |
| 8 | 10240.00 | 85.00 | 47.00 | Masculino | 0 | Adecuado | 0 | 0 | 0 | 0 | 1 | 18 | Cesarea | Complementaria | MBP | Moderado |
| 9 | 11100.00 | 87.00 | 47.40 | Masculino | 1 | Adecuado | 0 | 0 | 0 | 0 | 1 | 24 | Cesarea | Complementaria | MBP | Moderado |
| 10 | 13200.00 | 96.50 | 48.50 | Masculino | 0 | Adecuado | 0 | 0 | 0 | 0 | 1 | 36 | Cesarea | Complementaria | MBP | Moderado |

En la Figura **6.1**, se visualiza la cantidad de observaciones realizadas en los tiempos de seguimiento frente a cada una de las variables. Los cuadros de color amarillo muestran la cantidad de información faltante por variable estudiada y la zona café o más oscura del gráfico, representa la información completa. Se concluye que las variables: oxigeno(157), tipo de parto(29) y perímetro cefálico(4), presentan datos faltantes, siendo la variable oxigeno la de mayor frecuencia, por ende, se excluye del estudio para evitar un mal ajuste del modelo y estimación de los parámetros.

Posteriormente se realizó la selección de variables como se mencionó en la sección 5.2.3. Con las variables seleccionadas por los algoritmos implementados, se procedió a realizar la estimación del modelo mediante la función *gml* del software estadístico R studio.

### 6.5.1. Selección de Variables

El modelo lineal generalizado para este proyecto es el modelo de regresión logístico, que permite modelar una variable dependiente dicótomica. Con el objetivo de obtener el modelo más parsimonioso y mejor ajustado se utiliza el paquete **glmulti** para realizar la selección del modelo, utilizando un enfoque de teoría de la información (Calcagno et al., 2010). Se utiliza el método de detección exhaustivo, obteniendo todas las posibles combinaciones de variables. Para ello, se utilizó el 70 % (368/526) de las observaciones como es recomendado en diferentes estudios,es decir, 102 neonatos prematuros con BPN, que serían los datos de entrenamiento para estimar los modelos (validación cruzada) (Reche, 2013).

Con los resultados obtenidos en el análisis exploratorio, se analiza un panorama más descriptivo de las variables que mejor discriminan el peso del neonato en la edad gestacional. Ahora, se realiza la selección de las variables que se ajusten a un modelo de regresión logístico, verificando el ajuste y la verosimilitud de varios modelos. Para ello, se obtendrán modelos que no contienen ninguno, uno y todas las variables estudiadas. Se estimaron 100 diferentes modelos y se utilizó como criterio de decisión *AIC* para cada modelo, ya que esté intenta seleccionar el modelo que mejor describe una realidad desconocida y de alta dimensión, en comparación con el criterio *BIC* que intenta encontrar el modelo verdadero entre el conjunto de candidatos (Reche, 2013).

La Figura (**6.3**), muestra los valores de AIC para los 100 modelos estimados. La línea roja horizontal diferencia entre los modelos cuyo AIC es menor en comparación con más de 2 unidades de distancia del posible mejor modelo (es decir, el modelo con AIC más bajo). El resultado anterior muestra que hay 9 modelos cuyo AIC está a menos de 2 unidades del posible mejor modelo, sin embargo, no debemos colgarnos demasiado de tales divisiones (algo arbitrarias), este criterio diferencial se hace entre los modelos para observar la importancia de cada uno. Se podría considerar que los modelos con valores de más de 2 unidades son menos plausibles que aquellos con valores más cercanos a los del posible mejor modelo.

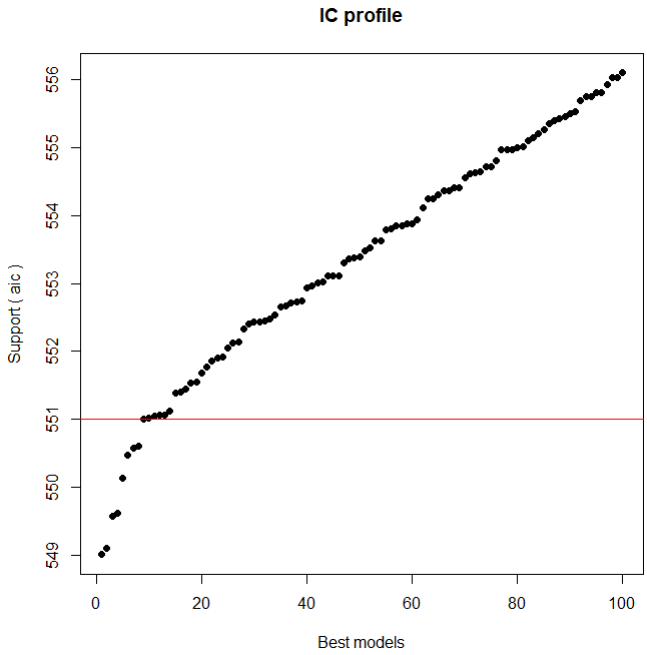


**Figura 6.3**: Gráfica de los valores de AIC para los 100 modelos

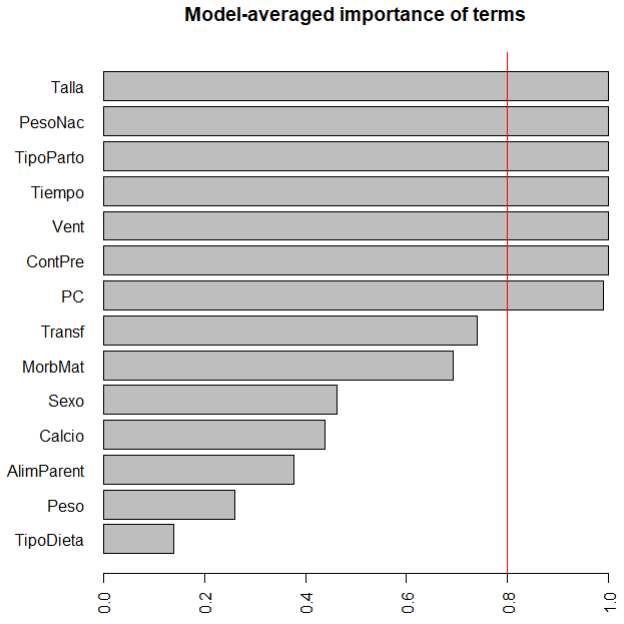


**Figura 6.4**: Variables Importantes para la modelación

La Tabla (**6.2**), muestra los 9 modelos que se mencionaron anteriormente. Donde el mejor modelo es aquel que incluye las variables sexo, control prenatal, morbilidad, ventilación, transfusiones, tiempo, tipo de parto, peso al nacer, talla y perímetro cefálico. El segundo mejor modelo incluye las variables predictoras mencionadas en el primer modelo sin tener en cuenta el sexo del neonato. Se obtienen los mismo pesos para ambos modelos (weights), llamados también *pesos Akaike*, el cual se considera como la probabilidad de que el modelo sea el mejor modelo, debido a que minimiza la perdida de información. Entonces, aunque el mejor modelo tiene peso alto, esté no es sustancialmente mayor que el segundo. Por lo tanto, se podría estudiar el comportamiento de ambos modelos y seleccionar el que mejor se ajuste al comportamiento del peso del neonato.

Se grafican las variables consideradas con mayor importancia, teniendo en cuenta todos los posibles modelos. La Figura (**6.4**), muestra la importancia relativa de los diversos términos del modelo. Donde el valor de importancia para un predictor particular es igual a la suma de los pesos para los modelos en los que aparece la variable.

La línea roja vertical se dibuja en 0.8, como punto de corte para diferenciar entre variables importantes y no tan importantes, tomando esta división algo arbitraria. Siendo así, se tiene que las variables perímetro cefálico, control prenatal, ventilación, tiempo, tipo de parto, peso al nacer y la talla, son las que presentan la mayor importancia y se deben tener en cuenta en el modelo de regresión logístico.

**Tabla 6.2**: Los 9 mejores modelos de 100 estudiados

| | model | aic | weights |
|---|---|---|---|
| 1 | PEGB ~ 1 + Sexo + ContPre + MorbMat + Vent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 549.01 | 0.06 |
| 2 | PEGB ~ 1 + ContPre + MorbMat + Vent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 549.10 | 0.06 |
| 3 | PEGB ~ 1 + Sexo + Calcio + ContPre + MorbMat + Vent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 549.57 | 0.04 |
| 4 | PEGB ~ 1 + Calcio + ContPre + MorbMat + Vent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 549.61 | 0.04 |
| 5 | PEGB ~ 1 + ContPre + MorbMat + Vent + AlimParent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 550.13 | 0.03 |
| 6 | PEGB ~ 1 + Sexo + ContPre + MorbMat + Vent + AlimParent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 550.47 | 0.03 |
| 7 | PEGB ~ 1 + Sexo + ContPre + Vent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 550.57 | 0.03 |
| 8 | PEGB ~ 1 + Calcio + ContPre + MorbMat + Vent + AlimParent + Transf + Tiempo + TipoParto + PesoNac + Talla + PC | 550.60 | 0.03 |
| 9 | PEGB ~ 1 + Sexo + ContPre + MorbMat + Vent + Transf + Tiempo + TipoParto + PesoNac + Peso + Talla + PC | 551.00 | 0.02 |

### 6.5.2. Elección del Modelo Logístico

Con los resultados obtenidos en la Tabla (**6.2**), se realizó el análisis de los dos principales modelos encontrados para modelar si el peso del neonato en la edad gestacional es *ADECUADO* (variable auxiliar del modelo). El objetivo, fue analizar cuáles de los modelos se ajusta mejor a los datos, mediante un conjunto de variables predictoras.

**Tabla 6.3**: Resumen Estadístico del Modelos No.1

| Coeficientes | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.3086 | 2.1316 | -3.90 | 0.0001 |
| SexoMasculino | 0.3509 | 0.2434 | 1.44 | **0.1494** |
| ContPre1 | -0.8184 | 0.2615 | -3.13 | 0.0017 |
| MorbMat1 | 0.4761 | 0.2549 | 1.87 | **0.0618** |
| Vent1 | 0.8928 | 0.2689 | 3.32 | 0.0009 |
| Transf1 | -0.6258 | 0.2789 | -2.24 | 0.0248 |
| Tiempo12 | -5.2413 | 1.0683 | -4.91 | 0.0000 |
| Tiempo18 | -6.4465 | 1.3074 | -4.93 | 0.0000 |
| Tiempo24 | -7.2977 | 1.5158 | -4.81 | 0.0000 |
| Tiempo36 | -9.2889 | 1.8252 | -5.09 | 0.0000 |
| TipoPartoEspontaneo | 0.2524 | 0.2531 | 1.00 | **0.3186** |
| PesoNacBPE | -2.0724 | 0.4521 | -4.58 | 0.0000 |
| PesoNacMBP | -0.8926 | 0.2982 | -2.99 | 0.0028 |
| Talla | 0.1901 | 0.0406 | 4.68 | 0.0000 |
| PC | 0.0312 | 0.0429 | 0.73 | **0.4664** |

En los resultados del resumen obtenido para los dos principales modelos a estudiar, se tiene que: Para el modelo 1 (Ver Tabla (**6.3**)), las variables predictoras asociadas al modelo son el sexo masculino como referencia, presentar control prenatal, morbilidad materna, ventilación, transfusiones, el tiempo de evaluación del peso (12, 18, 24 y 36), el tipo de parto espontáneo, clasificación del peso como extremo y muy extremo, la talla y el perímetro cefálico. Este resumen permite analizar que las variables sexo masculino, presentar morbilidad, tipo de parto espontáneo y perímetro cefálico, a una significancia $\alpha = 0.05$ los parámetros asociados a las variables no son significativos o distintos de 0 según el contrates de Wald (Pr(>|z|)). Por lo tanto, se recomienda excluir estas variables del modelo debido al ajuste del mismo.

Para el modelo 2 (Ver Tabla (**6.4**)), las variables predictoras asociadas son las mismas que el anterior modelo sin tener en cuenta el sexo marculino. Para este modelo, el tipo de parto espontáneo y el perímetro cefálico a una significancia $\alpha = 0.05$ los parámetros asociados a las variables no son significativos o distintos de 0 según el contrates de Wald (Pr(>|z|)). Por lo tanto, se recomienda excluir estas variables del modelo.

**Tabla 6.4**: Resumen Estadístico del Modelos No.2

| Coeficientes | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.1842 | 2.1230 | -3.86 | 0.0001 |
| ContPre1 | -0.8837 | 0.2590 | -3.41 | 0.0006 |
| MorbMat1 | 0.4989 | 0.2543 | 1.96 | 0.0498 |
| Vent1 | 0.8895 | 0.2682 | 3.32 | 0.0009 |
| Transf1 | -0.5399 | 0.2708 | -1.99 | 0.0462 |
| Tiempo12 | -5.2721 | 1.0646 | -4.95 | 0.0000 |
| Tiempo18 | -6.4726 | 1.3028 | -4.97 | 0.0000 |
| Tiempo24 | -7.3264 | 1.5095 | -4.85 | 0.0000 |
| Tiempo36 | -9.3135 | 1.8172 | -5.13 | 0.0000 |
| TipoPartoEspontaneo | 0.2176 | 0.2504 | 0.87 | **0.3848** |
| PesoNacBPE | -2.2759 | 0.4351 | -5.23 | 0.0000 |
| PesoNacMBP | -0.9926 | 0.2917 | -3.40 | 0.0007 |
| Talla | 0.1896 | 0.0405 | 4.68 | 0.0000 |
| PC | 0.0346 | 0.0435 | 0.79 | **0.4267** |

Con el análisis previo de ambos modelos, se excluyen las variables sexo masculino, morbilidad materna, tipo de parto espontáneo, perímetro cefálico y transfusiones, donde esta ultima variable para ambos modelos era significativa, pero al realizar la eliminación de las variables mencionadas, está variable se volvió no significativa a una significancia $\alpha = 0.05$. Finalmente, se obtiene el modelo final de regresión logístico que mejor se ajusta a los datos y sus coeficientes son significativos (Ver Tabla (**7.3**)).

## 6.6. Inferencias del modelo final de regresión logístico

Una vez estimado el modelo, se realiza las inferencias necesarias del modelo, con el fin de extrapolar los resultados al total de la población. Se realiza el cálculo de los contrastes de bondad del ajuste global del modelo y los intervalos asociados.

### 6.6.1. Contrastes de los parámetros

El objetivo es contrastar si los coeficientes estimados son significativos distintos de 0. En otras palabras, analizar si una determinada variable explicativa tiene un efecto significativo sobre el peso adecuado en la edad gestacional o no.

**Contraste de Wald**

Este contraste está basado en la normalidad asintótica de los estimadores. Se quiere contrastar si un parámetro $\beta_r = 0$, con r = 1, . . . ,10, frente a que no lo sea.

$$H_0 : \beta_r = 0 \qquad \forall r = 1, ..., 10 \qquad H_1 : \beta_r \neq 0 \qquad \forall r = 1, ..., 10$$

$$W_r = \frac{\hat{\beta}}{\hat{SE(\beta_r)}} \tag{6.1}$$

Donde $\hat{\beta}$ y $\hat{SE(\beta_r)}$ son las estimaciones del modelo para $\beta_r$ y el error estándar de $\beta_r$.

**Contraste condicional de razón de verosimilitud**

Las hipótesis de este contraste son las mismas que en el anterior. Se utilizó la librería *car* del software estadístico R Studio, para obtener la función *Anova*, la cual realiza los contrastes condicionales de razón de verosimilitud sobre los parámetros asociados a cada una de las variables del modelo, sin necesidad de especificar los distintos modelos.

### 6.6.2. Intervalos de Confianza para los parámetros

Los intervalos de confianza a nivel $(1-\alpha)$, están directamente relacionados con un contraste a un nivel de significancia $\alpha$. Se construyen intervalos de confianza basados en el test de Wald y test condicional de razón de verosimilitud. Adicional, se utiliza la técnica conocida como bootstrap (Efron, 1992). Estos métodos dan como resultados intervalos de valores plausibles $\hat{\beta}_r \pm z_{-\alpha/2}\hat{SE}(\hat{\beta}_r)$ (Ver subsección (5.2.1)).

Para el modelo final, se obtienen los intervalos de confianza a un nivel del 95 %. Si el intervalo de confianza para los coeficientes $(\beta_r)$ incluye el 0, significa que al nivel $\alpha = 0.05$ no se podría rechazar la hipótesis nula de que $\beta_r = 0$. Para la razón de oportunidad $(e^{\beta_r})$ si incluye el 1, no se podría rechazar la hipótesis nula de que $e^{\beta_r} = 1$.

Luego, se obtienen los intervalos de confianza mediante la técnica Bootstrap. El procedimiento consiste en tratar la muestra original como si fuera la población (Reche, 2013). Se extrae mediante un muestreo aleatorio con reemplazo, un número elevado de muestras (1.000 muestras), por ende se obtiene la distribución en el muestreo de los parámetros. Los intervalos se calculan tomando el 95 % de los valores centrales en la distribuciones empíricas obtenidas. Se utilizó la función *bootCase* del paquete *car* (Fox and Weisberg, 2011); para el calculo de los intervalos de confianza, se utilizó la función *quantile*.

### 6.6.3. Valores ajustados, predicciones del modelo y residuos

Con el objetivo de estudiar el ajuste del modelo, se analizan los valores ajustados y predicciones, para un valor individual de la variable dependiente asociado a los valores dado las variables independientes. Además, se examinan los residuos como la parte de la observación no explicada por el modelo ajustado.

#### Valores ajustados y predicciones

Se obtienen los valores ajustados y las predicciones para las primeras 10 observaciones, correspondientes al seguimiento de 4 neonatos en diferentes instantes del tiempo (Ver Tabla (**7.8**)). El resultado obtenido son probabilidades de ocurrencias. Adicional, se calcula el error estándar y los intervalos de confianza asociados a cada estimación del predictor lineal $(\hat{p}_i \pm 2 * \widehat{se(p_i)})$ con un nivel de confianza aproximado del 95 %.

#### Residuos

En la regresión lineal se tiene que el error asociado $\varepsilon$ expresa la desviación respecto a la media condicional y a su vez se distribuye normal con media cero y varianza constante. Para el caso de la regresión logística no ocurre esto, dado que la variable respuesta es dicótoma, el error $\varepsilon$ solo puede tomar dos posibles valores (Ver sección (5.2.1)).

El análisis de los residuos es fundamental para evaluar la adecuación del modelo y detectar los valores anómalos e influyentes. Teniendo en cuenta, que un residuo es la diferencia entre un valor observado y un valor ajustado. Es la parte de la observación no explicada por el modelo ajustado.

Se realizan los cálculos correspondientes para obtener los residuos de pearson, residuos de pearson estandarizados y los residuos estudentizados; los cuales se consideran significativos o influyentes si el valor absoluto es mayor a 2. Por otro lado, se obtiene el residuo de devianza y el residuo de devianza estandarizado, para completar el análisis. Donde este ultimo, se considera influye si el valor absoluto es mayor a 4.

### 6.6.4. Medidas de bondad del ajuste

Permite describir el ajuste del modelo al conjunto de observaciones. Con el objetivo de observar la evolución del peso del neonato a través el tiempo, mediante la información suministrada por las variables regresoras del modelo.

### Contrastes Estadísticos $G^2$ y $X^2$

La hipótesis que se contrasta para el caso de la regresión logística es:

$$H_0 : p_q = \frac{exp(\sum_{r=0}^{R} \beta_r x_{qr})}{1 + exp(\sum_{r=0}^{R} \beta_r x_{qr})} \ \forall_q = 1, ..., Q \tag{6.2}$$

$$H_1 : p_q \neq \frac{exp(\sum_{r=0}^{R} \beta_r x_{qr})}{1 + exp(\sum_{r=0}^{R} \beta_r x_{qr})} \qquad \forall_q = 1, ..., Q \tag{6.3}$$

Donde $Q$ es el número de perfiles o combinaciones posibles de lar variables predictoras y $p_q$ las probabilidades a estimar. El resultado deseado es que no se pueda rechazar la hipótesis nula de que el modelo se ajusta bien a los datos (Hosmer Jr et al., 2013).

Para los contrastes clásicos como los estadísticos $G^2$ de Wilks de razón de verosimilitud (Ver subsección(5.2.2)) y el estadístico $X^2$ de Pearson (Ver subsección(5.2.2)), se utilizan las mismas hipótesis a contrastar. El primer contrastes se calcula partiendo de la suma de los cuadrados de los residuos de la devianza y el segundo estadístico se calcula partiendo de los residuos de pearson. Estos contrastes sirven para comparar la diferencia entre los valores predichos y observados.

### Contrastes basado en el estadísticos de Hosmer Lemeshow

Hosmer y Lemeshow proponen crear 10 grupos de la variable respuesta en base a las probabilidades estimadas por el modelo, y comparar las frecuencias del éxito observado con las estimadas, mediante el estadístico $X^2$ de Pearson con 8 grados de libertad (Reche, 2013).

Inicialmente, se dividen las probabilidad estimadas en intervalos de igual amplitud, a partir de estos puntos de corte, se construyen las tablas de los valores observados y esperados. Con las tablas de frecuencias obtenidas, se calcula el estadístico de contraste y el valor-p asociado. Se utilizó la función *HLgof.test* del paquete *MKmisc*, para la bondad de ajuste global con 10 grupos.

### Medidas tipo $R^2$

La medida $R^2$ nos ofrece una medida de la bondad del ajuste del modelo. Para el caso del modelo de regresión logística, se presentan varios inconvenientes con esta medida, ya que los estimadores máximo verosimiles no son los estimadores que maximizan esta medida o que no tiene en cuenta la dependencia de la varianza de $Y$ respecto a $p$. Por lo siguiente, se propone trabajar con los pseudo $R^2$ de McFadden, Cox-Snell y Nagelkerke (Reche, 2013) (Ver subsección (5.2.2)).

**Tabla 6.5**: Medidas de Pseudo $R^2$ para el análisis de bondad del ajuste del modelo

| **Medida Tipo $R^2$** | **Formula ($R^2$)** |
|---|---|
| Pseudo de McFadden | $R^2 = 1 - \frac{\hat{L}_{Modelo}}{\hat{L}_{Intercepto}}$ |
| Pseudo de Cox y Snell | $R^2_{CN} = 1 - exp\left(-\frac{\widehat{LR}}{N}\right)$ |
| Pseudo de Nagelkerke | $R^2_N = \frac{R^2_{CN}}{maxR^2_{CN}}$ |

**Medidas basadas en la tabla de clasificación - Curva ROC**

Otra forma de evaluar el desempeño del modelo, es mediante la tabla de clasificación. Se elige un punto de corte igual a 0.5, si la probabilidad predicha por el modelo es mayor que el punto de corte se clasifica como éxito y si es menor como fracaso. Adicional, se evalúa el ajuste del modelo, mediante la curva ROC representada en un gráfico mediante la fracción de falsos positivos definida como FP/(FP+VN), frente a la fracción de verdaderos positivos definida como VP/(VP+FN). Se utilizó los paquete *pROC* y *ROCR*.

## 6.7. Diagnóstico y Validación

Después de comprobar el ajuste global del modelo de regresión logística, se estudia la falta de ajuste a nivel de cada observación y cómo afecta al modelo general.

El objetivo principal es realizar el diagnostico y validación del modelo ajustado, ya que indica si hay falta de ajuste por los efectos de las variables o las observaciones.

### 6.7.1. Análisis de los residuos

En la subsección (7.3.2) se describieron los distintos tipos de residuos bajo la hipotesis nula de que el modelo se ajusta bien a los datos. Como se ha mencionado anteriormente el modelo de regresión logística tiene una diferencia significativa con los modelos de regresión lineal, y es que el error asociado $\varepsilon$ que expresa la desviación respecto a la media condicional, no sigue una distribución normal con media cero y varianza constante. Para el caso de la regresión logística no ocurre esto, dado que la variable respuesta es dicótoma, el error $\varepsilon$ solo puede tomar dos posibles valores (Ver sección (5.2.1)).

Se considera a efectos de diagnósticos que un residuos es significativamente distinto de 0 si su valor absoluto es mayor que 2. Para realizar el cálculos de la frecuencia de residuos

significativos se utilizaron las funciones *residuals, rstandard* y *rstudent.*

Se utilizó la función *residualPlots*, del paquete *car*, para representar los residuos de Pearson frente a las variables predictoras y la transformación logit (predictor lineal).

### 6.7.2. Medidas de Influencias

Las medidas de influencia detectan los valores influyentes analizando el efecto que tienen en los parámetros del modelo, se trata de valores atípicos. Se utilizó las funciones *hatvalues*, *influence.measures* y *cooks.distance* para obtener los valores. Es recomendable detectar aquellos valores que tengan una distancia de Cook superior a 1, y comprobar su influencia en la estimación de los parámetros del modelo. Se muestran las medidas de influencia para las 6 primeras observaciones, donde se especifican las medidas DFBETAS (útiles para ver sobre qué variable del modelo es influyente cada observación), los DFFITS (cómo cambia el valor ajustado para la observación, cuando esta no se ha utilizado en el ajuste del modelo), el COVRATIO (mide el cambio en el determinante de la matriz de covarianzas de las variables predictoras, al eliminar la observación).

$$D_q = \frac{e^2_{PSq}}{k+1} * \frac{h_{qq}}{1-h_{qq}} \tag{6.4}$$

Donde $e_{PSq}$ es el error cuadrático medio del modelo de regresión, $k$ es la cantidad de variables y $h_{qq}$ son los *hat values.*

### 6.7.3. Colinealidad y Factores de Inflación de la Varianza Generalizado (GVIF)

El principal objetivo es estudiar si existe una relación lineal fuerte entre los predictores del modelo, ya que reduce la precisión de los coeficientes estimados, es decir, se incrementa su varianza. De este modo, se utiliza el factor de inflación de la varianza generalizado (GVIF), haciendo uso de la función *vif* del paquete *car*, para examinar la multicolinealidad en el modelo ajustado.

$$GVIF = \frac{detR_{11}detR_{22}}{detR} \tag{6.5}$$

Donde $R_{11}$ es la matriz de correlaciones entre el conjunto de regresores en cuestión (como los $p$ regresores de una variable con $p$+1 categorías). $R_{22}$ la matriz de correlaciones entre los otros regresores del modelo y $R$ la matriz de correlaciones entre todos los regresores del modelo.

### 6.7.4. Validación Cruzada

Se estudia el sobreajuste, que implica que el modelo es menos generalizable en el ajuste de otros datos. Una forma de analizar este problema, es verificando que el modelo predice correctamente. Para ello, se utiliza la validación cruzada mediante el método K-Fold, se utiliza la función *CVbinary* del paquete *DAAG* con un número de submuestras a considerar K=10.

Finalmente, con la propuesta metodológica mencionada anteriormente, se desea encontrar el modelo más parsimonioso y mejor ajustado que permita modelar la evolución del peso del neonato perteneciente a la Unidad de Cuidados Intensivos Neonatal del Hospital Universitario del Valle (HUV) y aportar información importante para la toma de decisiones que ayuden a orientar acciones de control y estrategias de prevención que minimicen el problema.

# Capítulo 7

# Resultados y Discusión

En este capitulo se muestran los resultados obtenidos mediante un análisis bivariado para encontrar las posibles relaciones de las variables estudiadas respecto a la evolución del peso del neonato prematuro del HUV. Se obtienen las inferencias del mejor modelo ajustado, contrastes de los parámetros, intervalos de confianzas, valores ajustados, predicciones, residuos, medidas de bondad de ajuste, diagnostico y validación del modelo de regresión logístico ajustado. Las mediciones se realizaron en diferentes instantes del tiempo (meses), llevando un seguimiento de la evolución del peso del neonato en los tiempos: 0, 12, 18, 24 y 36 meses, donde "0" hace referencia a la semana 40 de edad gestacional.

## 7.1. Estadísticas Descriptivas

Se contó con un total de 145 recién nacidos que presentaron pesos menores a 2.500 gramos. De estos 145 neonatos, el 61 % son de sexo femenino y el 39 % masculino. La edad promedio de las gestantes fue de 25 años, , el 75 % tienen una edad menor o igual a 30 años, la edad mínima fue de 14 años y la máxima de 44 años.

Más de la mitad de los neonatos nacen con Muy Bajo Peso (MBP) correspondiente a un 54,5 %, pesos entre 1.000 y 1.500 gramos, seguidos por un 33,1 % con Bajo Peso (BP), pesos entre 1.500 y 2.001 gramos y un 12,4 % con Bajo Peso Extremo (BPE), pesos inferiores a los 1.000 gramos. El 66,9 % de los pesos son adecuados a la edad gestacional, el 68.3 % de los neonatos necesitaron ventilación asistida y 63.4 % no se les realizó transfusiones. Además, el 68.3 % de las madre no asistieron a controles prenatales y el 63.8 % presentaron un parto por cesárea y la dieta que mayor predomino en un 49 % de los neonatos fue la mixta compuesta por leche maternizada junto con la leche materna (Ver Tabla (**7.1**)).

**Tabla 7.1**: Distribución de Frecuencias en Variables Cualitativas Observadas

| **Variables** | **Etiqueta** | **Frecuencias** | |
|---|---|---|---|
| Peso Nacimiento | | **Frec.** | % |
| Bajo Peso | BP | 48 | 33,1 |
| Muy Bajo Peso | MBP | 79 | 54,5 |
| Extremo Bajo Peso | BPE | 18 | 12,4 |
| Peso en la Edad Gestacional | | | |
| Adecuado | Ade | 97 | 66,9 |
| Bajo | Bajo | 48 | 33,1 |
| Alimentación Parenteral | | | |
| Si | Si | 89 | 61,4 |
| No | No | 56 | 38,6 |
| Ventilación | | | |
| Si | Si | 99 | 68,3 |
| No | No | 46 | 31,7 |
| Transfusiones | | | |
| Si | Si | 53 | 36,6 |
| No | No | 92 | 63,4 |
| Morbilidad Materna | | | |
| Si | Si | 100 | 69 |
| No | No | 45 | 31 |
| Control Prenatal | | | |
| Si | Si | 46 | 31,7 |
| No | No | 99 | 68,3 |
| Tipo de Parto | | | |
| Cesárea | Ces | 88 | 63,8 |
| Espontáneo | Esp | 50 | 36,2 |
| Necesidad de Oxigeno | | | |
| Leve | Leve | 40 | 38,8 |
| Moderado | Mode | 55 | 53,4 |
| Severo | Seve | 8 | 7,8 |
| Tipo de Dieta | | | |
| Artificial | Arti | 6 | 4,1 |
| Complementaria | Comp | 2 | 1,4 |
| Mixta | Mix | 71 | 49 |
| Seno | Seno | 66 | 45,5 |
| Sexo | | | |
| Femenino | F | 88 | 60,7 |
| Masculino | M | 57 | 39,3 |

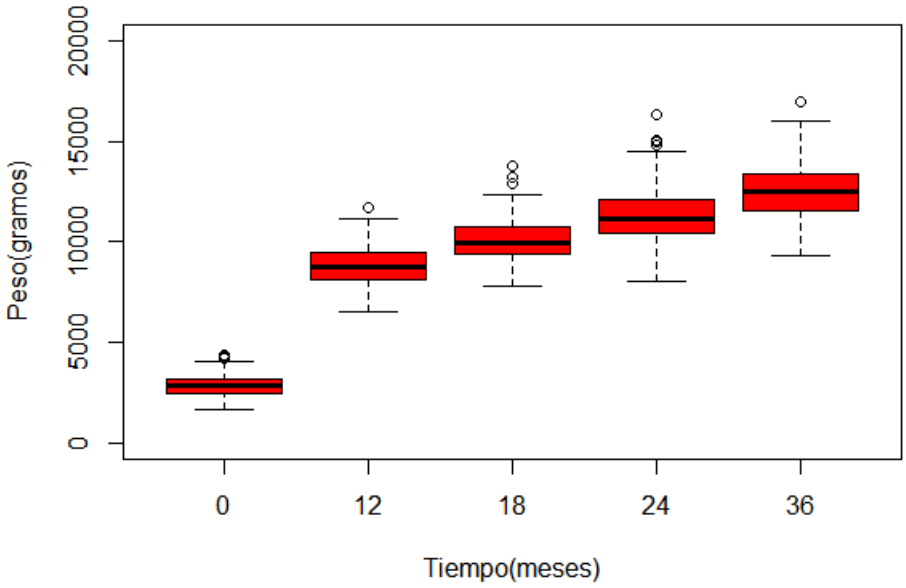


**Figura 7.1**: Distribución Peso Neonatos por Mes

Para observar la distribución del peso a través del tiempo, se gráfica dicho comportamiento, en la **Figura (7.1)**, se muestra que a medida que aumenta la edad aumenta el peso de neonato de una forma no lineal positiva, esté comportamiento creciente tiende a ser estable a través del tiempo, con algunas variaciones en el peso medio de los neonatos a medida que aumenta la edad (**Tabla (7.2)**). Se tiene que los pesos en las edades 12 y 18 presentan la variación más baja en sus pesos, presentando un coeficiente de variación (CV) del 11 %, mientras que al tiempo cero, correspondiente a la semana 40 de edad gestacional, los pesos presentan mayor variación (CV=19 %). El peso máximo de 17.000 gramos, lo presento una niña en los 36 meses de edad gestacional, siendo clasificado como un peso adecuado.

**Tabla 7.2**: Estadísticas Descriptivas de Peso en el Tiempo

| ***Edad (Meses)*** | ***Frecuencia*** | ***Media*** | ***Des*** | ***CV( %)*** | ***Mín*** | ***Máx*** |
|---|---|---|---|---|---|---|
| **0** | 143 | 2830 | 539 | 19 % | 1680 | 4340 |
| **12** | 126 | 8756 | 998 | 11 % | 6480 | 11980 |
| **18** | 111 | 10067 | 1121 | 11 % | 7780 | 13780 |
| **24** | 101 | 11368 | 1463 | 13 % | 8040 | 16360 |
| **36** | 45 | 12730 | 1647 | 13 % | 9340 | 17000 |
| **Total** | 145 | 8263 | 3683 | 45 % | 1680 | 17000 |

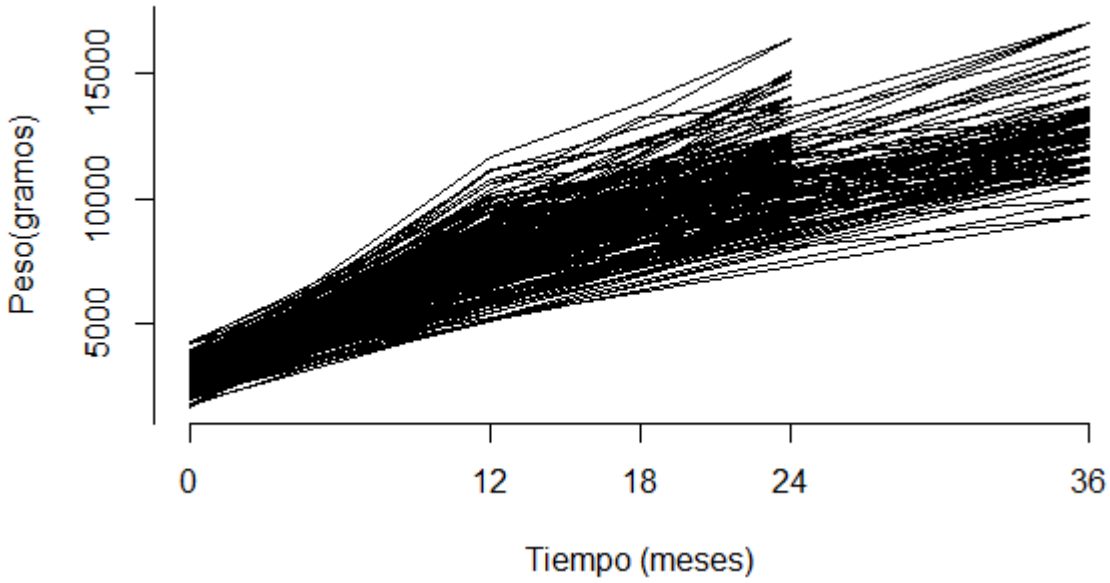


**Figura 7.2**: Curvas del peso por Neonato-Datos observados

Se graficaron las curvas asociadas a la evolución del peso a través del tiempo para cada uno de los neonatos en estudio. En la **Figura (7.2)**, se analiza que no todos los neonatos presentaron mediciones en los mismos tiempos, ya que con el pasar de éste algunos presentaron deserción o controles en tiempos diferentes a los establecidos.

Esta información obtenida para los neonatos, se caracteriza por:

1. Ser de medidas repetidas, ya que se realiza un seguimiento al mismo neonato en diferentes instante de tiempo, generando información dependiente.
2. El comportamiento general de las curvas presenta un crecimiento a través del tiempo.
3. Existen variaciones del peso de individuo a individuo.
4. Hay información faltante en algunos neonatos, causado por la deserción de los mismo.

Adicional, se tienen como indicadores del crecimiento y desarrollo de los neonatos la talla y el perímetro cefálico, en la **Figura (7.3)**, se visualiza un crecimiento positivo en la talla del neonato a través del tiempo, presentando un comportamiento similar a las curvas observadas originales del peso del neonato. Por otro lado, el perímetro cefálico presenta un aumento a través del tiempo pero a partir de los 24 meses de edad gestacional tiende a permanecer constante en 45 cm.

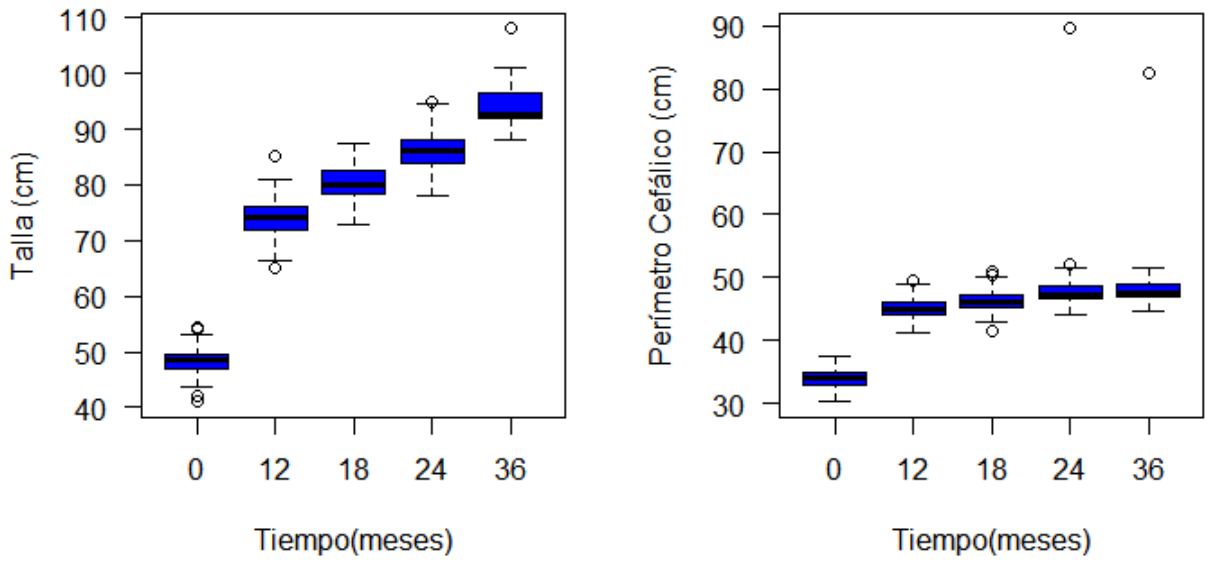


**Figura 7.3**: Talla y Perímetro Cefálico por Mes

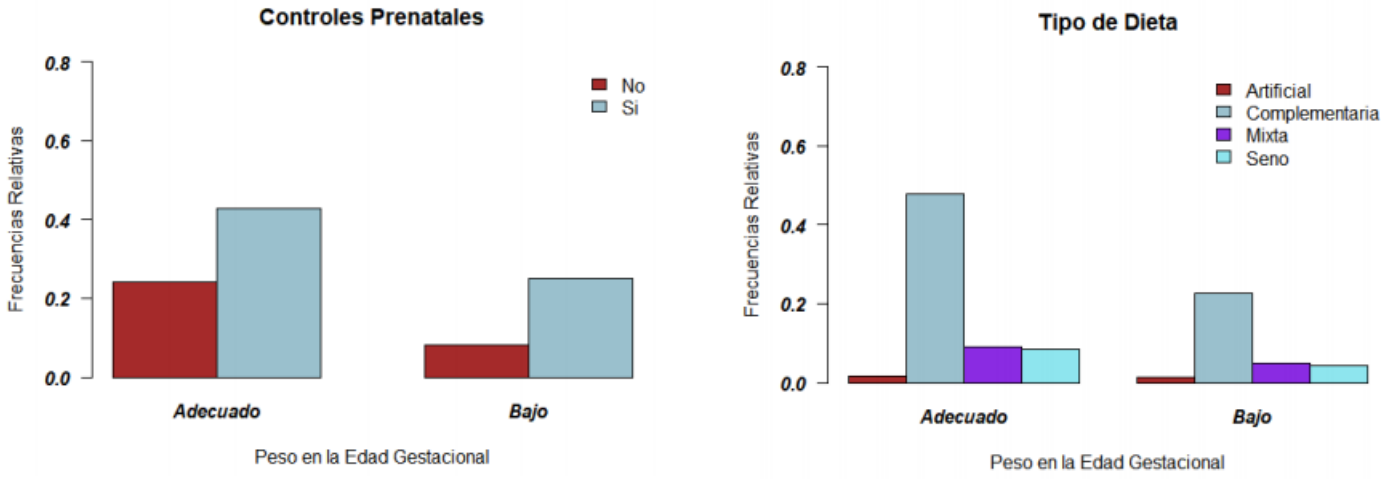


**Figura 7.4**: Controles Prenatales y el Tipo de Dieta, según el Peso del Neonato

En la investigación de Bermudez et al. (2015), se obtuvo un modelo de coeficientes aleatorio, donde el control prenatal fue la única variable significativa para el peso del neonato. En este proyecto se tiene que el presentar controles prenatales durante el periodo de embarazo permite que el peso del neonato sea adecuado para la edad gestacional. Por lo tanto, se podría considerar como una variable significativa para la modelación del evolución del peso (**Ver Figura (7.4)**).

Las madres de los neontatos presentan una dieta complementaria o reciben otros alimentos adicionales a la leche materna. En general, no discrimina el comportamiento respecto al peso en la edad gestacional, solo permite conocer que las madres tienden a utilizar una dieta complementaria.

## 7.2. Modelo final de regresión logístico

El modelo de regresión logístico mejor ajustado, esta determinado por las variables control prenatal, ventilación, tiempo de evaluación, clasificación del peso al nacer y la talla:

**Tabla 7.3**: Resumen Estadístico del Modelos No.3

| Coeficientes | Estimate $\beta$ | $e^{\beta}$ | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|---|
| (Intercept) | -6.7873 | 0.0011 | 1.7744 | -3.83 | 0.0001 |
| ContPre1 | -0.6743 | 0.5095 | 0.2248 | -3.00 | 0.0027 |
| Vent1 | 0.6264 | 1.8708 | 0.2345 | 2.67 | 0.0076 |
| Tiempo12 | -4.5742 | 0.0103 | 0.9786 | -4.67 | 0.0000 |
| Tiempo18 | -5.6894 | 0.0034 | 1.2047 | -4.72 | 0.0000 |
| Tiempo24 | -6.5620 | 0.0014 | 1.4067 | -4.66 | 0.0000 |
| Tiempo36 | -8.2947 | 2e-04 | 1.6999 | -4.88 | 0.0000 |
| PesoNacBPE | -1.6784 | 0.1867 | 0.3417 | -4.91 | 0.0000 |
| PesoNacMBP | -1.0072 | 0.3652 | 0.2470 | -4.08 | 0.0000 |
| Talla | 0.1786 | 1.1956 | 0.0365 | 4.89 | 0.0000 |

$$\begin{aligned} Peso_{ti} = logit[p(x)] = ln\left[\frac{p(x)}{1-p(x)}\right] = -6.79 - 0.67 * ContPre1_i \\ +0.63 * Vent1_i - 4.57 * Tiempo12_i - 5.69 * Tiempo18_i \\ -6.56 * Tiemp24_i - 8.29 * Tiempo36_i - 1.68 * PesoNacBPE_i \\ -1.01 * PesoNacMBP_i + 0.18 * Talla_{ti} \end{aligned} \tag{7.1}$$

Para t=1,2,3,4,5 (Tiempos de medición: 0,12,18,24,36 meses)
Para i=1,2,...,145 (Representa al neonato)

El coeficiente del intercepto (-6.79), es el logit de peso adecuado del neonato en la edad gestacional (Tiempo0) que hace referencia a las 40 semanas de edad gestacional. Además, de que la madre no presente controles, el neonato no recibe ventilación, la clasificación del peso al nacer es bajo y no se tiene información de la talla.

Los coeficientes del modelo se interpretan con base en la categoría de referencia elegida. Con los resultados obtenidos en la Tabla (**7.3**), se observa que la razón de oportunidad ($e^{\beta}$) asociadas a las variables: presentar control prenatal, los diferentes tiempos de evaluación del peso y la clasificación del peso al nacer del neonato, son factores protectores que reducen o atenúan la probabilidad de presentar un peso adecuado en la edad gestacional. Por el contrario, el presentar ventilación y la talla del neonato, son factores de riesgo que aumentan la probabilidad de presentar un peso adecuado en la edad gestacional.

## 7.3. Inferencias del Modelo Final de Regresión Logístico

### Contrastes de los parámetros: Contraste de Wald

Con los resultados obtenidos en la Tabla (**7.3**), a una significancia del $\alpha = 0.05$ los parámetros asociados a las variables del modelo son significativos o distintos de 0, ya que el valor absoluto del estadístico de Wald para cada una de las variables es mayor a $Z_{\alpha/2} = 1.96$. (Ver columna z value).

### Contrastes de los parámetros: Contraste condicional de razón de verosimilitud

La siguiente Tabla (**7.4**) muestra los contrastes condicionales de razón de verosimilitud sobre los parámetros asociados a cada una de las variables del modelo, sin necesidad de especificar los distintos modelos:

**Tabla 7.4**: Tabla de Análisis de Devianza

| | LR Chisq | Df | Pr(>Chisq) |
|---|---|---|---|
| ContPre | 9.41 | 1 | 0.0022 |
| Vent | 7.46 | 1 | 0.0063 |
| Tiempo | 25.83 | 4 | 0.0000 |
| PesoNac | 29.65 | 2 | 0.0000 |
| Talla | 26.28 | 1 | 0.0000 |

Ahora, se tiene en cuenta en el contraste condicional de razón de verosimilitudes si un subconjunto de parámetros son iguales a 0, frente a que alguno de ellos no lo sea:

$$\begin{aligned} Model1 : PEGB \sim 1 \\ Model2 : PEGB \sim ContPre + Vent + Tiempo + PesoNac + Talla \end{aligned} \tag{7.2}$$

De los resultados obtenidos en las Tablas (**7.4**) y (**7.5**), se concluye que los coeficientes asociados al modelo final, son significativos y distintos de 0. Teniendo en cuenta las Pr(>Chi) asociadas a cada variable, ya que son menores a la significancia $\alpha = 0.05$.

**Tabla 7.5**: Tabla de Análisis de Devianza, teniendo en cuenta el modelo nulo

| | Resid. Df | Resid. Dev | Df | Deviance | Pr(>Chi) |
|---|---|---|---|---|---|
| modelo 1 | 525 | 667.75 | | | |
| modelo 2 | 516 | 591.19 | 9 | 76.56 | 0.0000 |

### 7.3.1. Intervalos de Confianza para los parámetros

**Tabla 7.6**: Intervalos de Confianza asociados a los coeficientes del modelo estimado

| | **Test de Wald** | | | | **Test Condicional de RV** | | | |
|---|---|---|---|---|---|---|---|---|
| | $\beta_r$ | | $e^{\beta_r}$ | | $\beta_r$ | | $e^{\beta_r}$ | |
| | 2.5 % | 97.5 % | 2.5 % | 97.5 % | 2.5 % | 97.5 % | 2.5 % | 97.5 % |
| (Intercept) | -10.27 | -3.31 | 0.00 | 0.04 | -10.34 | -3.37 | 0.00 | 0.03 |
| ContPre1 | -1.12 | -0.23 | 0.33 | 0.79 | -1.12 | -0.24 | 0.33 | 0.79 |
| Vent1 | 0.17 | 1.09 | 1.18 | 2.96 | 0.17 | 1.10 | 1.19 | 2.99 |
| Tiempo12 | -6.49 | -2.66 | 0.00 | 0.07 | -6.53 | -2.69 | 0.00 | 0.07 |
| Tiempo18 | -8.05 | -3.33 | 0.00 | 0.04 | -8.10 | -3.37 | 0.00 | 0.03 |
| Tiempo24 | -9.32 | -3.80 | 0.00 | 0.02 | -9.38 | -3.86 | 0.00 | 0.02 |
| Tiempo36 | -11.63 | -4.96 | 0.00 | 0.01 | -11.70 | -5.02 | 0.00 | 0.01 |
| PesoNacBPE | -2.35 | -1.01 | 0.10 | 0.36 | -2.36 | -1.02 | 0.09 | 0.36 |
| PesoNacMBP | -1.49 | -0.52 | 0.23 | 0.59 | -1.50 | -0.53 | 0.22 | 0.59 |
| Talla | 0.11 | 0.25 | 1.11 | 1.28 | 0.11 | 0.25 | 1.11 | 1.29 |

Los resultados de la Tabla (**7.6**) muestran que los intervalos de confianza asociados a los coeficientes ($\beta_r$) del modelo, si se repite la muestra aleatoria a un elevado número de veces, el 95 % de las veces los intervalos contendrían el valor verdadero del parámetro. Para los coeficientes ($\beta_r$) los intervalos no contienen el 0, significa que al nivel de significancia $\alpha = 0.05$ se rechaza la hipótesis nula de que ($\beta_r = 0$). Del mismo modo, se analiza para la razón de oportunidad ($e^{\beta_r}$), concluyendo que los intervalos asociados no contienen el 1, por lo tanto, se rechaza la hipótesis nula ($e^{\beta_r} = 1$).

En general, el rango de variabilidad de los coeficientes en los intervalos obtenidos es pequeño y contiene el valor del parámetro estimado.

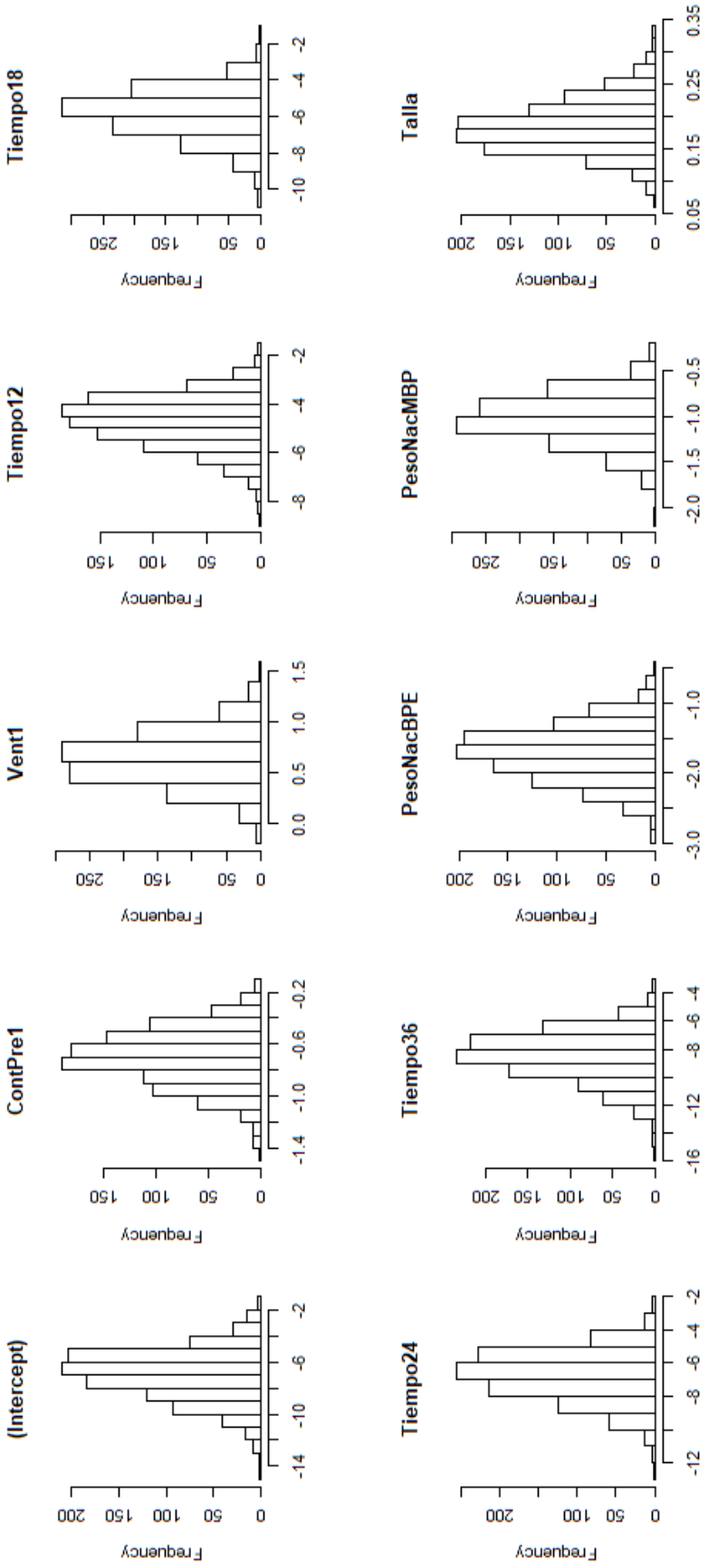


**Figura 7.5:** Histogramas de los coeficientes estimados mediante Bootstrap

Ahora, se obtendrán los intervalos de confianza mediante la técnica Bootstrap. En la Figura (**7.5**) se muestran los histogramas de los coeficientes estimados mediante bootstrap con 1.000 muestras. Mostrando que tienden a centrarse en su valor medio estimado.

**Tabla 7.7**: Intervalos de Confianza asociados a los coeficientes del modelo estimado, mediante la técnica Bootstrap

| | Bootstrap SD | 2.5 % | 97.5 % |
|---|---|---|---|
| (Intercept) | 1.92 | -11.07 | -3.56 |
| ContPre1 | 0.22 | -1.13 | -0.31 |
| Vent1 | 0.25 | 0.16 | 1.14 |
| Tiempo12 | 1.03 | -6.84 | -2.90 |
| Tiempo18 | 1.29 | -8.62 | -3.61 |
| Tiempo24 | 1.49 | -9.79 | -4.11 |
| Tiempo36 | 1.74 | -12.23 | -5.47 |
| PesoNacBPE | 0.39 | -2.48 | -0.99 |
| PesoNacMBP | 0.26 | -1.57 | -0.54 |
| Talla | 0.04 | 0.12 | 0.27 |

Al igual que los resultados obtenidos con el test de Wald y test condicional de razón de verosimilitud para los intervalos, se obtiene en la Tabla (**7.7**) que los intervalos de los coeficientes ($\beta_r$) no contienen el 0, significa que al nivel de significancia $\alpha = 0.05$ se rechaza la hipótesis nula de que ($\beta_r = 0$).

Finalmente, se concluye que las tres técnicas utilizadas para el cálculos de los intervalos de confianza asociados a los coeficientes del modelo estimado, son muy similares en sus rangos de variabilidad. Tanto los coeficientes como la razón de oportunidad son significativos para el modelo.

### 7.3.2. Valores ajustados, predicciones del modelo y residuos

#### Valores ajustados y predicciones

Se obtienen los valores ajustados y las predicciones para las primeras 10 observaciones, correspondientes al seguimiento de 4 neonatos en diferentes instantes del tiempo (Ver Tabla (**7.8**)). El resultado obtenido son probabilidades de ocurrencias. Adicional, se calcula el error estándar y los intervalos de confianza asociados a cada estimación del predictor lineal, con un nivel de confianza aproximado del 95 %.

#### Residuos

En la Tabla (**7.9**), las observaciones señaladas son los residuos influyentes, debido al aumento significativo del neonato de un periodo de evaluación a otro, en comparación con el progreso

**Tabla 7.8**: Valores estimados del predictor lineal para las primeras 10 observaciones estudiadas

| Obs. | Valor Ajustado | IC (2.5 %) | IC (97.5 %) | Error |
|---|---|---|---|---|
| 1 | 0.35 | 0.18 | 0.52 | 0.08 |
| 2 | 0.43 | 0.26 | 0.61 | 0.09 |
| 3 | 0.30 | 0.13 | 0.47 | 0.09 |
| 4 | 0.45 | 0.26 | 0.64 | 0.09 |
| 5 | 0.46 | 0.24 | 0.68 | 0.11 |
| 6 | 0.60 | 0.47 | 0.74 | 0.07 |
| 7 | 0.74 | 0.63 | 0.85 | 0.06 |
| 8 | 0.85 | 0.76 | 0.94 | 0.04 |
| 9 | 0.77 | 0.66 | 0.87 | 0.05 |
| 10 | 0.76 | 0.62 | 0.90 | 0.07 |
| 11 | 0.83 | 0.74 | 0.92 | 0.04 |
| 12 | 0.80 | 0.70 | 0.91 | 0.05 |
| 13 | 0.77 | 0.64 | 0.89 | 0.06 |
| 14 | 0.84 | 0.75 | 0.93 | 0.05 |
| 15 | 0.73 | 0.56 | 0.90 | 0.08 |
| 16 | 0.83 | 0.74 | 0.92 | 0.05 |
| 17 | 0.70 | 0.55 | 0.85 | 0.07 |
| 18 | 0.75 | 0.61 | 0.88 | 0.07 |
| 19 | 0.70 | 0.53 | 0.86 | 0.08 |
| 20 | 0.72 | 0.55 | 0.89 | 0.09 |

de los demás neonatos. Se analizan los residuos de pearson y los residuos de pearson estandarizados, ya que son lo únicos que indican influencia en las observaciones 11, 12 y 14; dado que el valor absoluto es mayor a 2. El análisis de devianza no detecta esta influencia, en la subsección (7.4.2) se realizará un análisis a fondo de este tema.

### 7.3.3. Medidas de bondad del ajuste

#### Contrastes Estadísticos $G^2$ y $X^2$

Para el modelo estimado del peso del neonato en la edad gestacional, se obtiene en la Tabla (**7.10**) que el contraste del estadístico $G^2$ de razón de verosimilitud rechaza la hipótesis nula con una significancia del $\alpha = 0.05$, por lo tanto, sugiere que el modelo no se ajusta bien a los datos. Sin embargo, el contraste del estadístico $X^2$ de Pearson con la misma significancia no rechaza la hipótesis nula de que el modelo se ajusta globalmente bien a los datos. Con este resultado es difícil concluir mediante estos contraste si el modelo esta bien ajustado o no. Para ello, se utiliza un contraste alternativo propuesto por (Hosmer Jr et al., 2013), ya que

**Tabla 7.9**: Tipos de residuos para las primeras 10 observaciones estudiadas

| Obs. | Res.Pearson | Res.Devianza | Res.P.Estandarizados | Res.Dev.Estandarizada | Res.Estudentizados |
|---|---|---|---|---|---|
| 1 | 1.35 | 1.44 | 1.38 | 1.47 | 1.46 |
| 2 | 1.14 | 1.29 | 1.16 | 1.31 | 1.31 |
| 3 | 1.53 | 1.55 | 1.55 | 1.58 | 1.58 |
| 4 | 1.11 | 1.26 | 1.13 | 1.29 | 1.28 |
| 5 | 1.08 | 1.24 | 1.10 | 1.27 | 1.26 |
| 6 | 0.81 | 1.00 | 0.82 | 1.01 | 1.01 |
| 7 | 0.60 | 0.78 | 0.60 | 0.79 | 0.79 |
| 8 | 0.43 | 0.58 | 0.43 | 0.58 | 0.58 |
| 9 | 0.55 | 0.73 | 0.56 | 0.74 | 0.73 |
| 10 | 0.56 | 0.74 | 0.57 | 0.75 | 0.75 |
| 11 | **-2.24** | -1.89 | **-2.25** | -1.91 | -1.91 |
| 12 | **-2.02** | -1.81 | **-2.04** | -1.82 | -1.83 |
| 13 | -1.81 | -1.71 | -1.83 | -1.73 | -1.73 |
| 14 | **-2.29** | -1.91 | **-2.31** | -1.93 | -1.94 |
| 15 | -1.64 | -1.62 | -1.68 | -1.65 | -1.65 |
| 16 | 0.45 | 0.61 | 0.46 | 0.62 | 0.61 |
| 17 | 0.65 | 0.84 | 0.66 | 0.85 | 0.85 |
| 18 | 0.58 | 0.77 | 0.59 | 0.77 | 0.77 |
| 19 | 0.66 | 0.85 | 0.67 | 0.86 | 0.86 |
| 20 | 0.63 | 0.81 | 0.64 | 0.83 | 0.82 |

**Tabla 7.10**: Contrastes clásicos de medidas de bondad de ajuste del modelo estimado

| CONTRASTE | ESTADISTICO | DF | P-VALOR |
|---|---|---|---|
| Estadístico $G^2$ de RV | 591,19 | 516 | 0,0120 |
| Estadístico $X^2$ de Pearson | 517,87 | 516 | 0,4685 |

los contrastes utilizados no llegan a la misma conclusión para el caso de datos no agrupados.

### Contrastes basado en el estadísticos de Hosmer Lemeshow

Las siguientes tablas representan los valores observados y esperados, dónde la columna *V1* representa el número de observaciones en cada grupo que tienen valor PEGB=0 y *y* el número de observaciones en cada grupo que tienen valor PEGB=1.

Con las tablas (**7.11**) y (**7.12**) de frecuencias anteriores, se obtiene el estadístico de contraste y el valor-p asociado.

**Tabla 7.11**: Tabla de frecuencia de los valores observados

| Intervalos | V1 | y |
|---|---|---|
| [0.199,0.277] | 5.00 | 2.00 |
| (0.277,0.355] | 13.00 | 4.00 |
| (0.355,0.433] | 22.00 | 7.00 |
| (0.433,0.51] | 26.00 | 26.00 |
| (0.51,0.588] | 31.00 | 45.00 |
| (0.588,0.666] | 19.00 | 44.00 |
| (0.666,0.743] | 18.00 | 60.00 |
| (0.743,0.821] | 25.00 | 55.00 |
| (0.821,0.899] | 13.00 | 65.00 |
| (0.899,0.977] | 2.00 | 44.00 |

**Tabla 7.12**: Tabla de frecuencia de los valores esperados

| Intervalos | V1 | yhat |
|---|---|---|
| [0.199,0.277] | 5.24 | 1.76 |
| (0.277,0.355] | 11.55 | 5.45 |
| (0.355,0.433] | 17.46 | 11.54 |
| (0.433,0.51] | 27.35 | 24.65 |
| (0.51,0.588] | 34.19 | 41.81 |
| (0.588,0.666] | 23.68 | 39.32 |
| (0.666,0.743] | 23.07 | 54.93 |
| (0.743,0.821] | 17.40 | 62.60 |
| (0.821,0.899] | 10.92 | 67.08 |
| (0.899,0.977] | 3.14 | 42.86 |

```
        Hosmer-Lemeshow H statistic

data:  fitted(modelo.final) and modelo.final$y
X-squared = 12.47, df = 8, p-value = 0.1314
```

**Figura 7.6**: Contraste de Hosmer Lemeshow

Con una significancia $\alpha = 0.05$ no se rechaza la hipótesis nula, según el estadístico de Hosmer Lemeshow. Es decir, que el modelo estimado se ajusta globalmente bien a los datos.

### Medidas tipo $R^2$

Con los resultados obtenidos en la Tabla (**7.13**), donde el $R^2$ da una idea de cuánto se reduce la devianza de los datos al ajustar el modelo, se concluye que la reducción obtenida es poca

**Tabla 7.13**: Medidas de Pseudo $R^2$ para el análisis de bondad del ajuste del modelo

| **Medida Tipo $R^2$** | **Formula ($R^2$)** | **Resultado ($R^2$)** |
|---|---|---|
| Pseudo de McFadden | $R^2 = 1 - \frac{\hat{L}_{Modelo}}{\hat{L}_{Intercepto}}$ | 0,1147 |
| Pseudo de Cox y Snell | $R^2_{CN} = 1 - exp\left(-\frac{\widehat{LR}}{N}\right)$ | 0,1355 |
| Pseudo de Nagelkerke | $R^2_N = \frac{R^2_{CN}}{max R^2_{CN}}$ | 0,1884 |

en los tres pseudos estudiados. En general, la reducción de devianza generada por el modelo estimado es aproximadamente del 19 %.

### Medidas basadas en la tabla de clasificación - Curva ROC

**Tabla 7.14**: Tabla de Clasificación del peso del neonato en la edad gestacional, para un punto de corte igual a 0.5

| Observado | **Predicción** | | |
|---|---|---|---|
| | PEGB=1 | PEGB=0 | Total |
| PEGB=1 | 314 | 38 | 352 |
| PEGB=0 | 111 | 63 | 174 |
| Total | 425 | 101 | 526 |

Con la Tabla (**7.14**), se obtiene la sensibilidad y especificidad del modelo. Donde el modelo clasifico al 89.2 % de las observaciones o mediciones del peso del neonato prematuro como *Adecuado* en la edad gestacional, porcentaje que representa la sensibilidad del modelo. Adicional, clasifica al 36.21 % de las observaciones como neonatos con peso *Bajo* en la edad gestacional, haciendo referencia a la especificidad del modelo (Ver Tabla (**7.15**)).
Las Tablas (**7.14**) y (**7.15**), representan las tablas de clasificación del modelo ajustado con un punto de corte de 0.5. Indicando globalmente, que el modelo clasificaría al 71,67 % (314+63/526) de las observaciones o mediciones del pesos del neonato en la edad gestacional a través del modelo ajustado.

Otra forma de evaluar el ajuste del modelo, es mediante la curva ROC, se obtuvo la siguiente curva roc para varios puntos de corte:

**Tabla 7.15**: Tabla de Clasificación en porcentajes por fila sobre el peso del neonato en la edad gestacional, para un punto de corte igual a 0.5

| Observado | **Predicción** | | |
|---|---|---|---|
| | PEGB=1 | PEGB=0 | Porcentaje Correcto (%) |
| PEGB=1 | 89,20 % | 10,80 % | 66,92 % |
| PEGB=0 | 63,79 % | 36,21 % | 33,08 % |
| Porcentaje global (%) | 80,80 % | 19,20 % | 100 % |

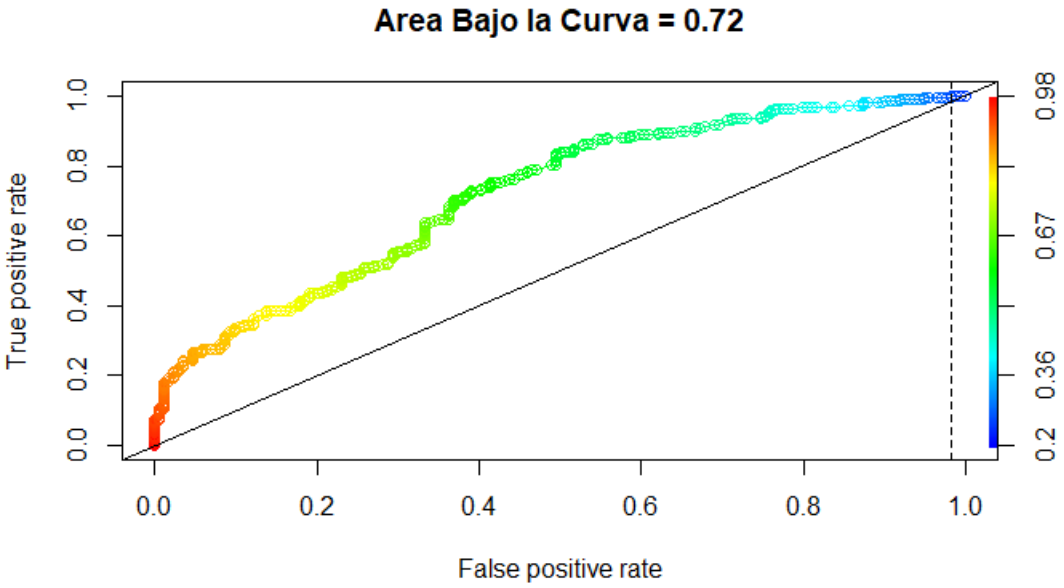


**Figura 7.7**: Curva ROC del modelo estimado para peso del neonato prematuro en la edad gestacional

La Figura (**7.7**), muestra la tasa de clasificaciones correctas para los distintos puntos de corte. Se tiene que para un punto de corte de 0.98 el porcentaje de clasificación correcta es aproximadamente del 72 % de las observaciones. Según las reglas que proporcionan Hosmer Jr et al. (2013) para interpretar el ara bajo la curva (AUC), se concluye que el modelo ajustado tiene una discriminación aceptable (72 %) para clasificar correctamente el peso del neonato prematuro en la edad gestacional.

Finalmente, con las medidas de ajustes realizadas al modelo se concluye que el modelo estimado se ajusta bien a los datos y clasifica correctamente al 72 % de las observaciones o mediciones del peso del neonato en la edad gestacional.

## 7.4. Diagnóstico y Validación

### 7.4.1. Análisis de los residuos

**Tabla 7.16**: Observaciones significativas según el tipo de residuo estudiado

| Tipo de Residuo | Residuos Significativos (Frecuencia) | Residuos Significativos (Porcentaje) |
|---|---|---|
| Residuo de Pearson | 21 | 3,99 % |
| Residuo de Pearson Estandarizado | 21 | 3,99 % |
| Residuo de la Devianza | 5 | 0,95 % |
| Residuo de la Devianza Estandarizado | 5 | 0,95 % |
| Residuo Studentizados | 6 | 0,01 % |

La Tabla (**7.16**) muestra que menos del 5 % de los residuos son significativos. Los residuos de pearson y los residuos de pearson estandanrizado tiene la misma frecuencia de residuos significativos siendo aproximadamente el 4 % de las observaciones de los neonatos prematuros con bajo peso al nacer. Los otros tipos de residuos tienen menor frecuencia de residuos significativos siendo menos del 1 % de las observaciones del seguimiento de los neonatos. En general, se tienen porcentajes bajos de residuos significativos, por lo tanto se puede pensar que no hay falta de ajuste a nivel de cada observación. Ahora, se analizan los residuos más significativos y las variables asociadas en el estudio, ordenados de mayor a menor:

**Tabla 7.17**: Resumen de los primeros residuos más significativos

| Observación | Peso | **Talla** | PC | Sexo | Calcio | **PEGB** | **ContPre** | MorbMat | **Vent** | AlimParent | Transf | **Tiempo** | TipoParto | TipoDieta | **PesoNac** | **Predicho** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 376 | 16980.00 | 108.00 | 82.50 | Masculino | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 36 | Cesarea | Complementaria | MBP | 0.93 |
| 430 | 11820.00 | 88.00 | 46.50 | Masculino | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 24 | Cesarea | Complementaria | BP | 0.91 |
| 428 | 9100.00 | 75.00 | 44.50 | Masculino | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 12 | Cesarea | Complementaria | BP | 0.88 |
| 140 | 15300.00 | 99.00 | 49.00 | Femenino | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 36 | Cesarea | Complementaria | BP | 0.87 |
| 429 | 10500.00 | 80.50 | 43.50 | Masculino | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 18 | Cesarea | Complementaria | BP | 0.86 |
| 305 | 8760.00 | 74.00 | 48.00 | Masculino | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 12 | Cesarea | Artificial | BP | 0.86 |

En la Tabla (**7.17**), vemos que los residuos más altos corresponden a los neonatos que, tanto por control prenatal, ventilación, tiempo de evaluación, clasificación del peso al nacer y talla, el modelo le asigna la pertenencia a la categoría opuesta que es tener un peso adecuado en la edad gestacional, analizando la columna de predicho con probabilidades altas de clasificación.

Para los diferentes tipos de residuos estudiados, se obtuvo que las observaciones 140, 305, 376, 428, 429 y 430, son las que tienen los residuos más significativos.

El residuo más alto corresponde a un neonato masculino, el cual presento los controles prenatales, no recibió ventilación, con una talla de 108 cm en los 36 meses de seguimiento y con una clasificación de peso al nacer muy extremadamente bajo. El modelo asigna a ese perfil una probabilidad prácticamente 1, de que el peso del neonato es adecuado en la edad gestacional.

Los mismo ocurre con el resto de neonatos con residuos altos. Su configuración de variables predictoras, hace que el modelo los clasifique con una alta probabilidad en la categoría contraria. Adicional, se representan los residuos mediante las siguientes gráficas:

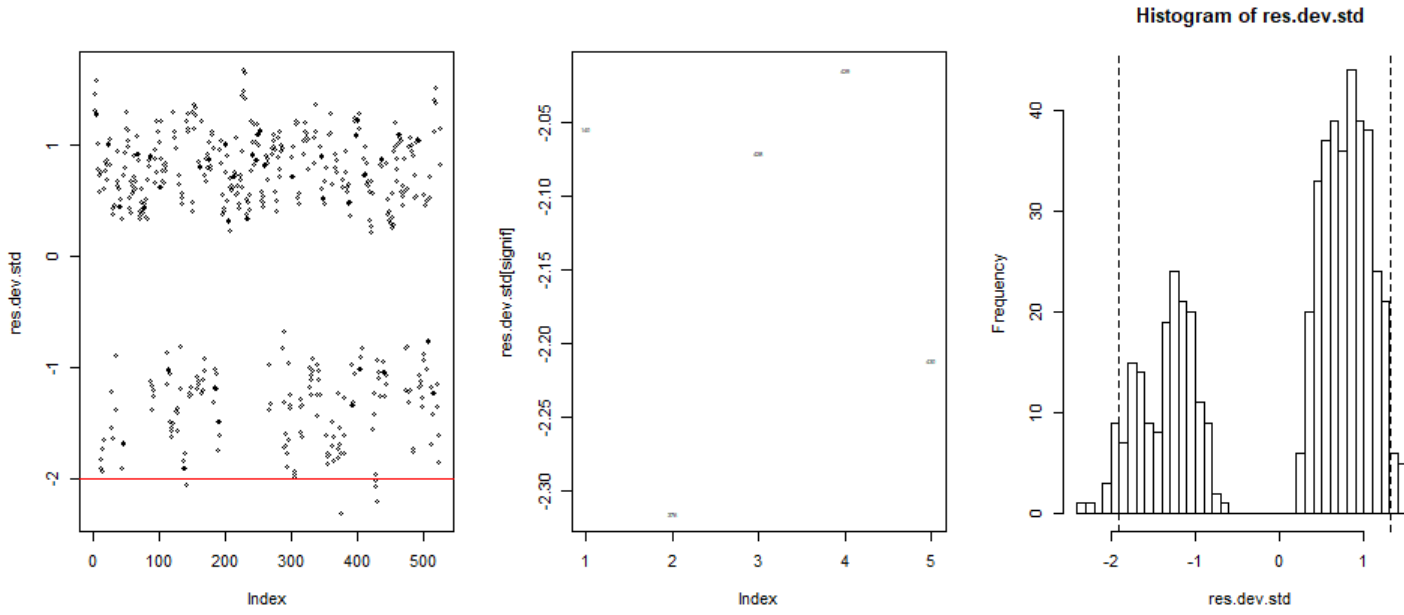


**Figura 7.8**: Residuos de la devianza estandarizados

La Figura (**7.8**) muestra los diferentes gráficos para analizar los residuos de la devianza para cada neonato. En el primer gráfico de izquierda a derecha, se representan los residuos mayores que 2 en valor absoluto, notando que no hay residuos mayores que 2 y hay una frecuencia pequeña de 5 residuos con valor negativo mayor que 2. El segundo gráfico, etiqueta los 5 residuos más significativos que corresponden a las observaciones (376, 430, 428, 140 y 429). Finalmente, el ultimo gráfico es un histograma con los cuantiles 0.025 y 0.975, con lo que entre esas dos líneas estaría el 95 % de los residuos, no se observa un comportamiento normal ya que estamos bajo un modelo de regresión logístico.

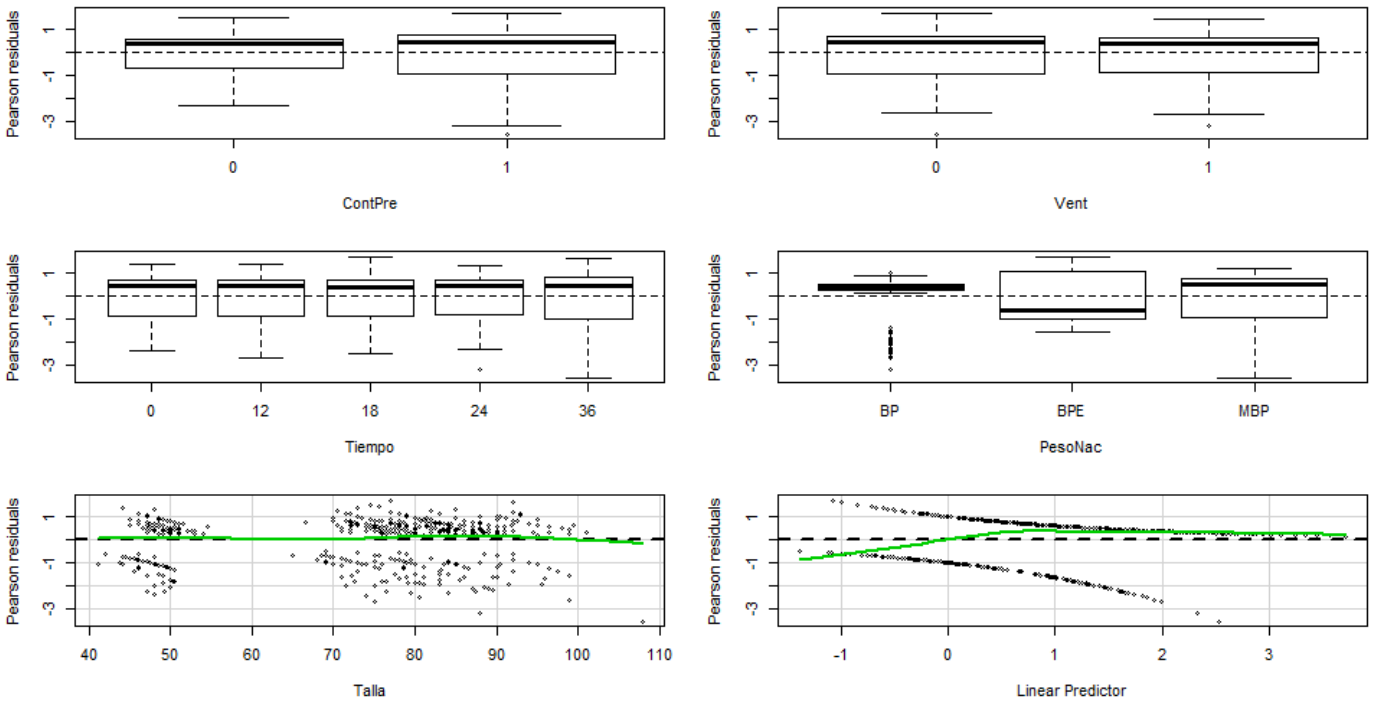


**Figura 7.9**: Residuos de Pearson

La Figura (**7.9**) gráfica los boxplots para observar la relación entre los residuos y cada una de las variables cualitativas predictoras y las categorías asociadas. Estos gráficos no muestran ningún patrón y los boxplots están centrados aproximadamente alrededor de la recta y=0. En el caso de la variable cuantitativa Talla, la línea verde no identifica un efecto cuadrático, por lo cual no es necesario realizar ningún ajuste añadiendo el cuadrado a la variable, por lo contrario la linea verde es aproximadamente un línea recta horizontal, permitiendo concluir que la variable talla es significativa.

El ultimo gráfico, representa los residuos frente al predictor lineal: Si PEGB=0 y la clasificación del modelo predice mal, el residuo debe ser negativos, y si el valor verdadero es PEGB=1, entonces subestimados y los residuos deben ser positivos, generando una relación monótona, donde la línea verde esta cerca de la linea punteada y=0. Como se ha mostrado no hay residuos mayores que +2, lo que indica que el modelo esta clasificando mal a neonatos con peso bajo en la edad gestacional, de hay la relación con los residuos significativos mayores que 2 negativo.

En general, el número de residuos significativos es muy pequeño comparado con el total de residuos. Al final de la sección (7.3), con los test de bondad de ajuste globales y el análisis de la curva ROC para este modelo se indica un ajuste adecuado.

### 7.4.2. Medidas de Influencias

En la Tabla (**7.18**) se muestran las diferentes medidas de influencias, donde el $|DFBETAS| < 2/\sqrt{n}$ para cada una de las variables del modelo, los $|DFFITS| < 2\sqrt{p/n}$, los valores hat $h_{ii} < 2p/n$ y las distancia de cook $D_i < 1$, por lo tanto, se concluye que no hay valores influyentes o atípicos que afecten la estimación de los parámetros.

En la Figura (**7.10**), al lado izquierdo se muestra la distancia de Cook frente al número de observaciones. No hay distancias de Cook mayores que 1, es decir, no hay valores influyentes, y se etiquetan los valores más altos, correspondientes a las observaciones 140, 317 y 376, con una distancia máxima de 0.04. El gráfico del lado derecho, muestra la distancia de Cook frente a los valores hat, se muestran que ninguna de las observaciones con mayor distancia de Cook tiene además un valor alto de $h_{qq}$.

**Tabla 7.18**: Medidas para detectar valores influyentes

| | dfb.1_ | dfb.CnP1 | dfb.Vnt1 | dfb.Tm12 | dfb.Tm18 | dfb.Tm24 | dfb.Tm36 | dfb.PNBP | dfb.PNMB | dfb.Tall | dffit | cov.r | cook.d | hat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.10 | -0.08 | -0.08 | 0.05 | 0.05 | 0.06 | 0.06 | 0.13 | -0.00 | -0.08 | 0.25 | 1.01 | 0.01 | 0.03 |
| 2 | 0.04 | -0.08 | -0.07 | 0.05 | 0.03 | 0.03 | 0.03 | 0.12 | 0.00 | -0.03 | 0.22 | 1.02 | 0.00 | 0.03 |
| 3 | 0.12 | -0.08 | -0.08 | 0.11 | 0.14 | 0.11 | 0.11 | 0.13 | 0.00 | -0.12 | 0.28 | 1.01 | 0.01 | 0.03 |
| 4 | 0.04 | -0.08 | -0.06 | 0.04 | 0.04 | 0.06 | 0.04 | 0.12 | -0.00 | -0.04 | 0.23 | 1.03 | 0.00 | 0.03 |
| 5 | -0.01 | -0.07 | -0.07 | -0.01 | -0.01 | -0.01 | 0.03 | 0.13 | 0.00 | 0.01 | 0.27 | 1.04 | 0.01 | 0.05 |
| 6 | 0.05 | -0.07 | -0.02 | 0.01 | 0.02 | 0.02 | 0.02 | -0.00 | 0.03 | -0.03 | 0.13 | 1.02 | 0.00 | 0.02 |

Otro gráfico que permite ver los valores influyentes, es la Figura (**7.11**) que muestra las distancia de cook, los residuos estudentizados, los p-valores de los contrastes de los residuos si son distintos de 0 y los valores hat:

Se siguen teniendo las mismas observaciones con las distancia de Cook más altas, pero se mantiene la conclusión de que no hay valores influyentes al igual que la medida de valores hat. Los residuos estudentizados solo el 0.01 % (Ver Tabla (**7.16**)) son mayores que 2 en valor absoluto.

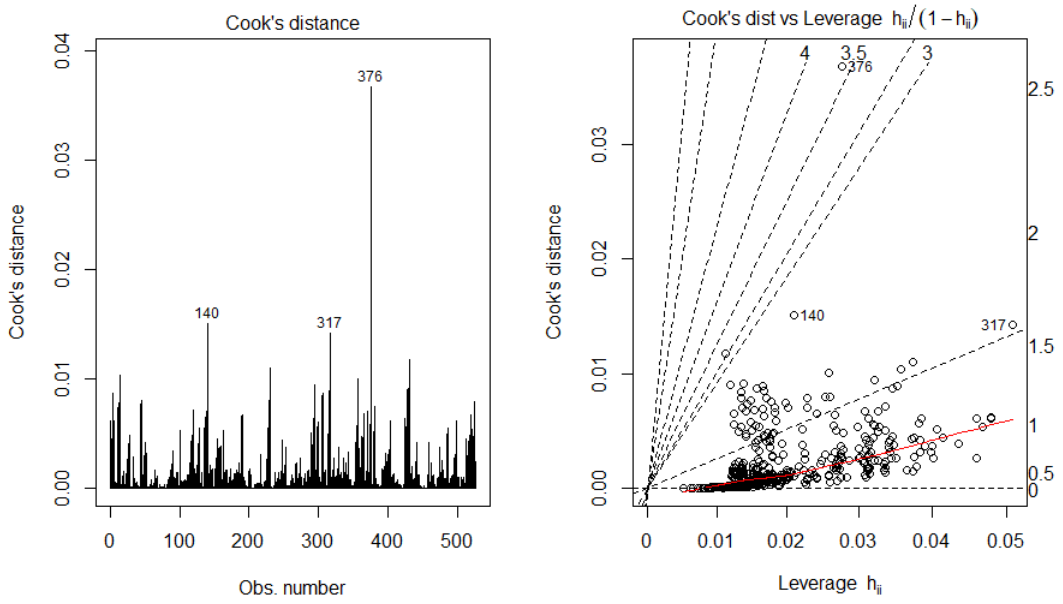


**Figura 7.10**: Medidas de Influencia: Distancia de Cook y Distancia de Cook frente a los valores hat

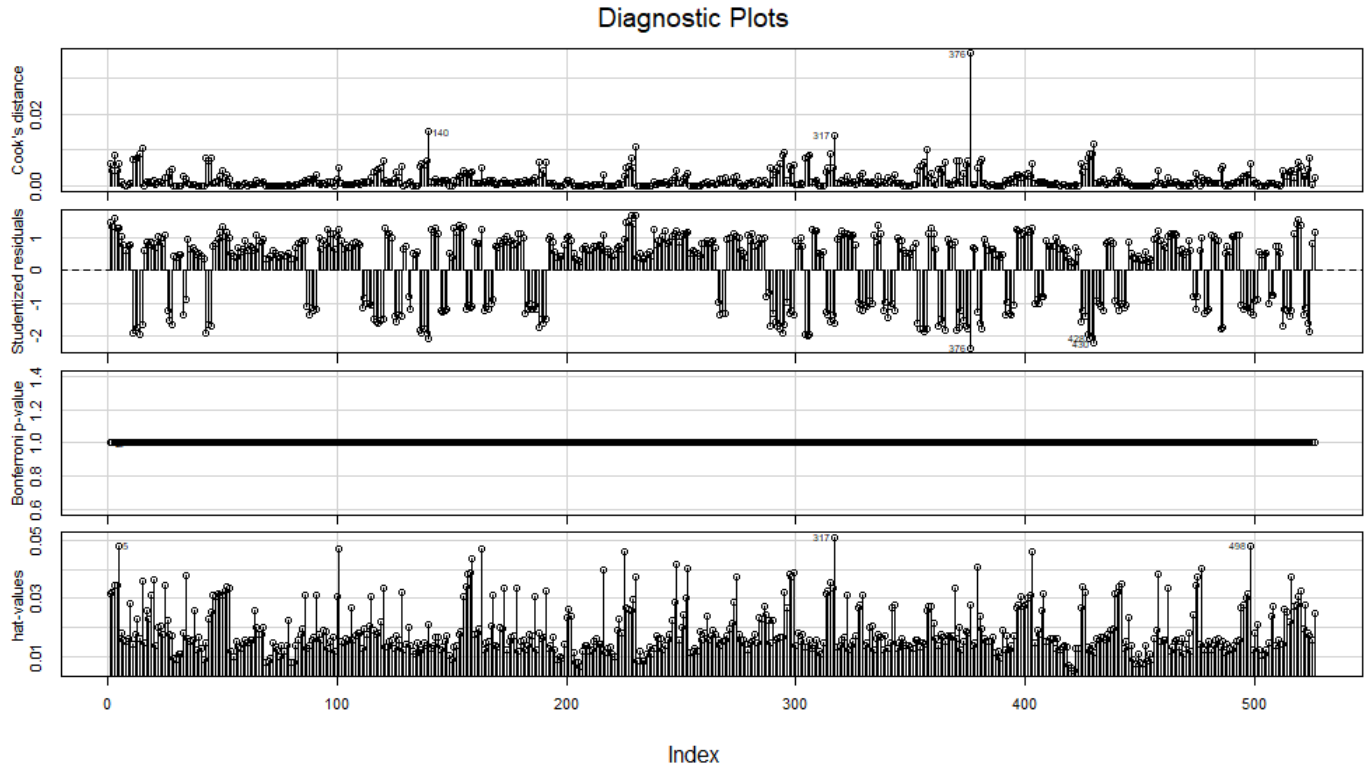


**Figura 7.11**: Gráfico de influencia con influenciaIndexPlot

### 7.4.3. Colinealidad y Factores de Inflación de la Varianza Generalizado (GVIF)

**Tabla 7.19**: Resultados Colinealidad y Factor de Inflación de la Varianza Generalizado

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| ContPre | 1.02 | 1.00 | 1.01 |
| Vent | 1.08 | 1.00 | 1.04 |
| Tiempo | 34.52 | 4.00 | 1.56 |
| PesoNac | 1.10 | 2.00 | 1.02 |
| Talla | 34.36 | 1.00 | 5.86 |

En la Tabla (**7.19**) se obtienen los resultados del GVIF para cada uno de los predictores que componen el modelo ajustado. Se analizan los resultados obtenidos para $GVIF^{1/2p}$, donde p es la cantidad de regresores de la variable predictora, ya que el modelo tiene variables cualitativas. Para las variables ContPre, Vent, Tiempo y PesoNac, se obtiene $GVIF^{1/2p}$ muy cercanos a 1, lo que indica ausencia de colinealidad. Por el contrario, la variable Talla tiene un $GVIF^{1/2p}$ cercano a 5, indicando colinealidad aproximada. La cual, se puede reducir, aumentando el tamaño de la muestra de los neonatos prematuros con bajo peso al nacer. En general, el modelo sigue conservando un buen ajuste y la precisión de la estimación de los coeficientes no se reduce por la colinealidad o relación entre los predictores del modelo.

### 7.4.4. Validación Cruzada

```
Fold:  9 1 2 10 3 5 6 4 7 8
Internal estimate of accuracy = 0.717
Cross-validation estimate of accuracy = 0.698
```

**Figura 7.12**: Validación cruzada método K-Fold con K=10

La tasa de clasificaciones correctas en la estimación en un conjunto de entrenamiento es aproximadamente del 72 %, y por otro lado, la estimación mediante validación cruzada es del 70 %. En general, el valor indica, que en medida, el modelo clasifica correctamente a más del 72 % de los neonatos prematuros con bajo peso al nacer, cuando estos no han sido usados en el ajuste del modelo (Ver resultados (**7.12**)). Este resultado es coherente con el área bajo la curva ROC, obtenida en la subsección (7.3.3).

Finalmente, a la vista de la tasa de clasificaciones correctas calculada mediante validación cruzada, de la no existencia de valores influyentes y del escaso porcentaje de residuos significativos, concluimos que el modelo no presenta falta de ajuste ni tampoco problemas de sobreajuste, en general es un modelo aceptable.

# Capítulo 8

# Conclusiones y Recomendaciones

## 8.1. Conclusiones

El estudio del bajo peso al nacer (BPN) es importante, debido a que constituye un problema de Salud Pública en Colombia, no sólo por su alta morbilidad y mortalidad infantil, sino también por las secuelas que puede ocasionar en la edad adulta. Además, es un trazador para la identificación de desigualdades en el proceso de salud, enfermedad y atención, ya que es sensible a diferentes condiciones de vida (Hachuel et al., 2006).

Una de las principales herramientas para registrar y evaluar el crecimiento de los niños a través del tiempo, son las curvas de crecimiento. El objetivo principal del seguimiento mediante estas curvas, es que el pediatra y los padres del niño conozcan la evolución del crecimiento, con el fin de desarrollar el máximo potencial del peso y detectar o corregir a tiempo posibles alteraciones en el proceso (Serrano, 2002). La formulación de un modelo estadístico, permite intentar responder como los contextos sociales afectan los resultados y el riesgo en la salud (Hachuel et al., 2006).

De acuerdo con las estadísticas descriptivas, la distribución del peso a través del tiempo es creciente a medida que aumenta la edad gestacional del neonato, y tiene una forma no lineal positiva y una variación relativamente constante. En general, el comportamiento de las curvas de seguimiento del peso, presentan un crecimiento con el tiempo. Adicional, la talla y el perímetro cefálico presentan un crecimiento positivo, similar a las curvas observadas del peso del neonato prematuro del HUV.

Considerando el crecimiento de las curvas del peso a través del tiempo para cada uno de los neonatos prematuros con bajo peso al nacer, se obtuvo un modelo de regresión logístico que permite modelar la evolución del peso del neonato en la edad gestacional según Ballard, como peso adecuado o bajo. Se encontró que las posibles variables o patrones que describen la evolución del peso son: la asistencia a los controles prenatales, la clasificación del peso al nacer

(BP,BPE y MBP), la ventilación asistida, la talla del neonato y el tiempo de seguimiento (0, 12, 18, 24 y 36 meses).

Teniendo en cuenta como objetivo principal el ajuste del modelo estimado, se obtuvo las inferencias acerca de los parámetros mediante diferentes técnicas estadísticas, encontrando que los coeficientes asociados son significativos y distintos de 0. Donde la asistencia a los controles prenatales, la clasificación del peso al nacer y los diferentes tiempos de evaluación, son factores protectores que reducen o atenúan la probabilidad de presentar un peso adecuado en la edad gestacional. Por el contrario, la ventilación asistida y la talla del neonato son factores de riesgos, que aumentan la posibilidad de presentar un peso adecuado en la edad gestacional. Esta conclusión ha sido ampliamente divulgada en diferentes investigaciones (Vanegas (2015), Daza et al. (2009), Bermudez et al. (2015)), siendo importante resaltar que el peso esta asociado a la talla y la asistencia a los controles prenatales.

Las medidas de bondad de ajuste realizadas, mostraron que el modelo estimado se ajusta globalmente bien a los datos y clasifica correctamente al 72 % de las observaciones o mediciones del peso del neonato en la edad gestacional, por ende, el modelo tiene una discriminación aceptable según Hosmer Jr et al. (2013). En cuanto, al diagnostico y validación del modelo, el porcentaje de residuos significativos es muy pequeño comparado con el total de las observaciones y esta posiblemente asociado al aumento significativo del peso del neonato de un periodo de evaluación a otro en comparación con el progreso de los demás neonatos. Sin embargo, no hay valores influyentes o atípicos que afecten la estimación de los parámetros. En general, el modelo conserva un buen ajuste y la precisión de la estimación de los coeficientes no se reduce por la colinealidad o relación entre los predictores del modelo.

Finalmente, a la vista de la tasa de clasificaciones correctas calculada mediante validación cruzada, de la no existencia de valores influyentes y del escaso porcentaje de residuos significativos, se concluye que el modelo no presenta falta de ajuste ni tampoco problemas de sobreajuste, en general, es un modelo aceptable. Esta información es soporte de ayuda para los pediatras que realizan seguimiento en la orientación de acciones de control y las entidades públicas prestadores de servicio de salud como el Ministerio de Salud y Protección social, que se encargan de regular las normas y directrices en materia de temas de salud pública (promoción de la salud y prevención de la enfermedad), asistencia social, población en riesgo y pobreza, mediante estrategias de prevención para el cuidado del neonato.

## 8.2. Recomendaciones

Se pueden considerar otros tipos de técnicas estadísticas para llevar a cabo el estudio del bajo peso al nacer del neonato prematuro del Hospital Universitario del Valle (HUV):

- Realizar un estudio epidemiológico observacional y analítico mediante casos y controles anidados, donde los casos como los controles son tomados de la población que participa

en un estudio de cohorte, de tal forma, que permita comparar la proporción de neonatos prematuros con bajo peso al nacer tomados como casos y los neonatos de la misma cohorte pero que ya tienen un peso adecuado para la edad gestacional como controles (Sánchez-Nuncio et al., 2005).

- Otra alternativa, es ejecutar un estudio que permita estimar las curvas de la talla y el perímetro cefálico del neonato, las cuales complementen y proporcionen un adecuado control del crecimiento de los neonatos. Se puede realizar mediante regresión cuantílica, donde las curvas de percentiles se utilizan comúnmente para detectar un crecimiento anormal (Fuentes, 2017).

- Por último, se puede realizar un estudio de análisis de datos espaciales o geoestadística, con el objetivo de inferir aspectos que no han sido medidos u observados (Bivand et al., 2008). Para ello, se debe contar con la información de la ubicación de las madres de los neonatos y de esta manera, generar estrategias de prevención para las comunas más vulnerables.

# Bibliografía


Berhie, K. A. and Gebresilassie, H. G. (2016), 'Logistic regression analysis on the determinants of stillbirth in ethiopia', *Maternal health, neonatology and perinatology* **2**(1), 10.

Bermudez, V. I., Andrade, B. M. and Torres, M. J. (2015), 'Modelación de la evolución de neonatos con bajo peso al nacer, atendidos en el hospital universitario del valle, durante el período 2002 a 2010: Estudio de cohorte', *Archivos de Medicina (Manizales)* **15**(2), 191–199.

Bewick, V., Cheek, L. and Ball, J. (2005), 'Statistics review 14: Logistic regression', *Critical care* **9**(1), 112.

Bivand, R. S., Pebesma, E. J. and Gómez-Rubio, V. (2008), *Applied spatial data analysis with R*, Vol. 747248717, Springer. 405p, New York, USA.

Calcagno, V., de Mazancourt, C. et al. (2010), 'glmulti: an r package for easy automated model selection with (generalized) linear models', *Journal of statistical software* **34**(12), 1–29.

Canty, A., Ripley, B. et al. (2012), 'boot: Bootstrap r (s-plus) functions', *R package version* **1**(7).

Cook, R. D. and Weisberg, S. (1982), *Residuals and influence in regression*, New York: Chapman and Hall.

Cox, D. R. and Snell, E. J. (1989), *Analysis of binary data*, Vol. 32, CRC Press.

Cruz Montesinos, D. L., Llivicura Molina, M. M. et al. (2013), 'Factores de riesgo perinatales para peso bajo en recién nacidos a término del Hospital Gineco Obstétrico Isidro Ayora, Quito 2012'.

Daza, V., Jurado, W., Duarte, D., Gich, I., Sierra-Torres, C. H. and Delgado-Noguera, M. (2009), 'Bajo peso al nacer: exploración de algunos factores de riesgo en el Hospital Universitario San José en Popayán (colombia)', *Revista Colombiana de Obstetricia y Ginecología* **60**(2), 124–134.

Domínguez Domínguez, I. (2010), 'Estudio del bajo peso al nacer en cayo hueso', *Revista Habanera de Ciencias Médicas* **9**(4), 588–594.

Efron, B. (1992), Bootstrap methods: another look at the jackknife, *in* 'Breakthroughs in statistics', Springer, pp. 569–593.

Flores, L. (2002), 'Análisis estadístico de los factores de riesgo que influyen en la enfermedad angina de pecho', *Oficina General del Sistema de Bibliotecas y Biblioteca Central UNMSM* .

Fox, J. and Monette, G. (1992), 'Generalized collinearity diagnostics', *Journal of the American Statistical Association* **87**(417), 178–183.

Fox, J. and Weisberg, S. (2011), *An R companion to applied regression*, Sage Publications.

Fuentes, N. A. (2017), 'Desigualdades de crecimiento municipal en méxico: un análisis mediante regresión cuantílica', *Ensayos Revista de Economía (Ensayos Journal of Economics)* **26**(2).

Hachuel, L. S., Boggio, G. S. and Borra, V. L. (2006), 'Uso de modelos logit mixtos para el estudio del bajo peso al nacer en Rosario'.

Herrera, A. I., Jaramillo, M. R. and Restrepo de Rovetto, C. (2013), 'Bajo peso al nacer y enfermedad renal cronica: reporte de casos en el hospital universitario del valle y propuesta de seguimiento'.

Honaker, J., King, G., Blackwell, M. and Blackwell, M. M. (2010), 'Package 'amelia''.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied logistic regression*, Vol. 398, John Wiley & Sons.

Martınez, D. R., Julio, L. A., Cabaleiro, J. C., Pena, T. F., Rivera, F. F. and Blanco, V. (2009), 'El criterio de información de akaike en la obtención de modelos estadısticos de rendimiento', *XX Jornadas de Paralelismo en Coruña* .

McFadden, D. et al. (1973), 'Conditional logit analysis of qualitative choice behavior', *Institute of Urban and Regional Development, University of California Oakland* .

Milad, M., Fabres, J., Asíllaga, C. and Others (2010), 'Recomendación sobre curvas de crecimiento intrauterino', *Revista chilena de pediatría* **81**(3), 264–274.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012), *Introduction to linear regression analysis*, Vol. 821, John Wiley & Sons.

Nagelkerke, N. J. et al. (1991), 'A note on a general definition of the coefficient of determination', *Biometrika* **78**(3), 691–692.

Navarro Manotas, E., Rodríguez Cuadro, D., Nieves Vanegas, S., Hurtado Ibarra, K. and Camacho Castro, C. (2015), 'Análisis de los factores de riesgo de bajo peso al nacer a partir de un modelo logístico polinómico'.

Nelder, J. A. and Baker, R. J. (1972), *Generalized linear models*, Wiley Online Library.

Organización, W. H., UNICEF et al. (2003), 'Estrategia mundial para la alimentación del lactante y del niño pequeño'.

Paraje, G. (2009), 'Desnutrición crónica infantil y desigualdad socioeconómica en américa latina y el caribe', *Revista Cepal* .

Reche, J. L. C. (2013), 'Regresión logística. tratamiento computacional con r.'.

Rodríguez, S. R., de Ribera, C. G. and Garcia, M. P. A. (2008), 'El recién nacido prematuro', *Asociación Española de Pediatría [libro electrónico]. España* .

Sánchez-Nuncio, H. R., Pérez-Toga, G., Pérez-Rodríguez, P. and Vázquez-Nava, F. (2005), 'Impacto del control prenatal en la morbilidad y mortalidad neonatal', *Revista médica del instituto mexicano del seguro social* **43**(5), 377–380.

Serrano, A. T. (2002), 'Crecimiento y desarrollo', *Revista Mexicana de Medicina Física y Rehabilitación* **14**, 54–57.

Torres, J., Palencia, D., Sánchez, D. M., García, J., Rey, H. and Echandía, C. A. (2006), 'Programa madre canguro: primeros resultados de una cohorte de niños seguidos desde la unidad neonatal hasta la semana 40 de edad postconcepcional', *Colombia Médica* **37**(2).

Torres-Muñoz, J., Rojas, C., Mendoza-Urbano, D., Marín-Cuero, D., Orobio, S. and Echandía, C. (2016), 'Factores de riesgo asociados con el desarrollo de asfixia perinatal en neonatos del Hospital Universitario del Valle, Cali, Colombia, 2010-2011', *Biomédica* **37**.

Vanegas, S. N. (2015), 'Análisis de los factores de riesgo de bajo peso al nacer a partir de un modelo logístico polinómico', *Prospectiva* **13**(1), 76–85.

Veintimilla Dávila, M. G. (2017), 'Comparación de los resultados antes y después de la implementación del programa madre canguro en recién nacidos menores de 2000 gramos en el hospital general luis gabriel dávila, durante los años 2013 a 2016'.

Velázquez Quintana, N. I., Zárraga, M. Y., Luis, J. and Ávila Reyes, R. (2004), 'Recién nacidos con bajo peso; causas, problemas y perspectivas a futuro', *Boletín Médico del Hospital Infantil de México* **61**(1), 73–86.

Williams, D. (1987), 'Generalized linear model diagnostics using the deviance and single case deletions', *Applied Statistics* pp. 181–191.

# A. Resultados Prueba Chi Cuadrado con una significancia del 5 %

**Tabla .1**: Resultados Prueba Chi Cuadrado con una significancia del 5 %

| | Variable1 | Estadistico | DF | Valor.P | Criterio |
|---|---|---|---|---|---|
| 1 | Tiempo | 1.14 | 8 | 1.00 | No Rechazo Ho. |
| 2 | Sexo | 13.42 | 2 | 0.00 | Rechazo Ho. |
| 3 | PN | 841.79 | 168 | 0.00 | Rechazo Ho. |
| 4 | Peso | 591.52 | 632 | 0.87 | No Rechazo Ho. |
| 5 | Talla | 248.77 | 212 | 0.04 | Rechazo Ho. |
| 6 | PC | 189.37 | 230 | 0.98 | No Rechazo Ho. |
| 7 | TipoParto | 9.90 | 2 | 0.01 | Rechazo Ho. |
| 8 | TipoDieta | 2.52 | 6 | 0.87 | No Rechazo Ho. |
| 9 | HA | 2.34 | 2 | 0.31 | No Rechazo Ho. |
| 10 | ContPre | 9.29 | 2 | 0.01 | Rechazo Ho. |
| 11 | MorbMat | 2.34 | 2 | 0.31 | No Rechazo Ho. |
| 12 | PesoNac | 38.62 | 4 | 0.00 | Rechazo Ho. |
| 13 | PEGB | 1052.00 | 4 | 0.00 | Rechazo Ho. |
| 14 | Vent | 12.84 | 2 | 0.00 | Rechazo Ho. |
| 15 | Oxigeno | 11.09 | 4 | 0.03 | Rechazo Ho. |
| 16 | AlimParent | 10.34 | 2 | 0.01 | Rechazo Ho. |
| 17 | Transf | 12.51 | 2 | 0.00 | Rechazo Ho. |
| 18 | Ex Respiratorio | 2.46 | 2 | 0.29 | No Rechazo Ho. |
| 19 | Ex Neurológico | 3.20 | 2 | 0.20 | No Rechazo Ho. |
| 20 | Ex Genitourinario | 0.35 | 2 | 0.84 | No Rechazo Ho. |
| 21 | Ex Cardiovascular | 2.06 | 2 | 0.36 | No Rechazo Ho. |
| 22 | Ex Gastrointestinal | 2.10 | 2 | 0.35 | No Rechazo Ho. |
| 23 | Ex Osteomuscular | 0.58 | 2 | 0.75 | No Rechazo Ho. |
| 24 | Ex Clínica Sist Visual | 3.63 | 2 | 0.16 | No Rechazo Ho. |
| 25 | Ex Clínica Sist Auditivo | 0.51 | 2 | 0.77 | No Rechazo Ho. |
| 26 | Ex Audiológico | 0.12 | 2 | 0.94 | No Rechazo Ho. |
| 27 | Ex Oftamológico | 2.62 | 6 | 0.85 | No Rechazo Ho. |
| 28 | Requiere Oxigeno | 4.13 | 4 | 0.39 | No Rechazo Ho. |
| 29 | Ex Hematopoyerico | 3.15 | 2 | 0.21 | No Rechazo Ho. |
| 30 | Vacunación | 2.07 | 2 | 0.36 | No Rechazo Ho. |
| 31 | Interconsulta | 0.20 | 2 | 0.91 | No Rechazo Ho. |
| 32 | Examenes | 0.62 | 2 | 0.73 | No Rechazo Ho. |
| 33 | Hierro | 0.54 | 2 | 0.76 | No Rechazo Ho. |
| 34 | Calcio | 6.34 | 2 | 0.04 | Rechazo Ho. |
| 35 | Vitaminas | 2.18 | 2 | 0.34 | No Rechazo Ho. |
| 36 | Fe | 1.32 | 2 | 0.52 | No Rechazo Ho. |
| 37 | Suplemento | 3.73 | 2 | 0.15 | No Rechazo Ho. |