UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Regresión logística y Poisson y sus diferencias, una aplicación a la base de datos de la Secretaría de Salud de defunciones en México, 2012.

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuaria

P R E S E N T A:

Karen De Lucio Jiménez

DIRECTOR DE TESIS:
Dr. Ricardo Ramírez Aldana

Ciudad de México, 2018

Hoja de datos del jurado

1. Datos del alumno
De Lucio
Jiménez
Karen
55 3053 6109
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
30703925-3

2. Datos del tutor
Dr.
Ricardo
Ramírez
Aldana

3. Datos del sinodal 1
Dra.
Silvia
Ruiz-Velasco
Acosta

4. Datos del sinodal 2
M. en C.
David Chaffrey
Moreno
Fernández

5. Datos del sinodal 3
Act.
Gerardo
Sisniega
Lira

6. Datos del sinodal 4
Act.
Edna Gabriela
López
Estrada

7. Datos del trabajo escrito.
Regresión logística y Poisson
y sus diferencias,
una aplicación a la base de datos
de la Secretaría de Salud
de defunciones en México,
2012.

117 páginas.
Año 2018.

Agradecimientos

A mis padres, por el apoyo incondicional durante toda mi trayectoria en mis estudios y en mi formación como persona. A mi hermano, por ser mi amigo y mi colega. A Vero, por tu confianza, tus consejos y tu fé en cada uno de tus estudiantes. A Gaby y a Ninive, por su bella labor, su compañerismo y su amistad. A mis profesores Edna y Ricardo Aldana por haberme introducido en la teoría de los Modelos Lineales Generalizados, su apoyo en el aspecto teórico, libros y ejemplos. Un especial agradecimiento a Ricardo Aldana por su paciencia. Por que si no hubieras creído en que yo podía hacerlo, más difícil hubiera sido mi camino a conseguir este gran logro. A Karen Lanzguerrero por haberme enseñado a manejar Excel. Gracias a tu ayuda fue posible este trabajo. A mis sinodales por sus comentarios y especialmente a David Chaffrey por el tiempo que dedicó a este trabajo. Por que gracias a ustedes pude saber que equivocaciones cometí y corregirlas. A los jóvenes que he asesorado, por darme la oportunidad de apoyarlos. Y con ello hacerme más humilde. A mis compañeros y amigos de la Facultad de Ciencias que me mostraron con sus ejemplos el valor del trabajo y la persistencia. Y a Héctor Saavedra, por que lo prometido es deuda, gracias por tu ayuda en LaTeX. A ésta institución, en la que en sus instalaciones y jardines tuve alguna vez un segundo hogar.

Al cosmos por darme una gran lección. Pues al final de mis altibajos encontré que no se trata de la meta perseguida, sino del camino. A pesar de mi miedo e inseguridad en el futuro, si conservo la confianza en lo que más anhelo y persisto, podré trascender aquellos sentimientos y convertirlos en oro.

Índice general

Resumen

En ésta tesis se presentan los resultados del modelado estadístico con una base de datos. Para guiar al lector sobre los temas que se abordan en ésta tesis, se plantea la pregunta de investigación de qué procedimiento puede realizarse para obtener modelos explicativos utilizando modelos lineales generalizados logístico y loglineal de Poisson, aplicados a una base de datos real.

Los objetivos son:

1. Presentar un procedimiento para obtener un modelo explicativo.
2. Exponer los modelos lineales generalizados Binomial logístico y loglineal Poisson.
3. Comparar los modelos logístico y loglineal Poisson.
4. Mostrar una alternativa de los modelos logístico y loglineal para corregir la sobredispersión. Es decir, introducir los modelos quasi-verosímiles, en particular los modelos quasibinomial y quasipoisson.
5. Utilizar el programa estadístico R.

Ya que se ha planteado la pregunta y se han mostrados los objetivos, se puede decir brevemente que los modelos explicativos se utilizan para probar hipótesis teóricas. En el caso de ésta tesis se desea probar la hipótesis sobre qué posibles causas pudieron contribuir a la defunción dado que se tenía diabetes mellitus, en la República Mexicana en 2012.

Puede observarse que la base de datos utilizada para la elaboración de la prueba empírica fue de la Secretaría de Salud. Cabe mencionar que la aplicación de los modelos explicativos logístico y de Poisson, hecha en éste trabajo, es del campo de la medicina.

Sin embargo también existen otros campos en los cuales pueden ser aplicados ésta clase de modelos. Por ello vale la pena tener una guía para su elaboración. Así también

debe reconocerse que su obtención no es inmediata. Con éstos se puede explicar un fenómeno bajo ciertas condiciones en diversos campos de la ciencia . Algunos ejemplos de estudios que utilizan modelos explicativos con modelos lineales generalizados logísticos se encuentran en el libro *Categorical data analysis* de Alan Agresti [1], como los estudios de alcohol y su relación con la malformación infantil (pg 345), la contaminación del aire y la respiración (cap. 9.5.5), el ingreso y su relación con el número de tarjetas de crédito (cap. 5.19, pg 206). Por otro lado, un ejemplo de un modelo explicativo de regresión Poisson se tiene en el mismo libro [1], es "Horshoe crabs and satellites" (J. Brackmann, Ethology 1996, Agresti(2002) sec. 4.3).

Por tanto puede decirse que en el campo de la medicina y de las finanzas el modelo explicativo logístico ha sido útil.

Y en el caso del modelo Poisson se ha utilizado para una rama de la biología y psicología experimental llamada etología que estudia el comportamiento de los animales, en éste caso el comportamiento de los cangrejos herradura (horshoe crab).

En los siguientes párrafos se presenta una síntesis de ésta tesis.

La teoría requerida para la elaboración de la presente fue la del modelado explicativo expuesta en el primer capítulo y la de los modelos lineales generalizados (capítulo 5). En especial se expondrán los modelos logístico y loglineal Poisson (capítulo 5).

Se utilizaron conceptos tales como modelos lineales, tablas de contingencia, estimadores verosímiles, funciones de masa y distribución de la familia exponencial y en especial las funciones Binomial y Poisson. Así también se utilizan conceptos utilizados en cursos especializados de regresión Logística y Poisson .

Se utilizaron dos bases de datos de la Secretaría de Salud, con una de ellas (DEFUN12.dbf, registradas en la variable "lista1") se elaboró una tabla en SPSS sobre las enfermedades que causaron defunciones en México en el año 2012 (capítulo 3). En ella se identificaron 85,354 casos de defunciones a causa de Diabetes Mellitus, de un total de 602,354 defunciones. A modo de comparación, se puede decir que otras causas de defunciones identificadas en dicha tabla fueron Enfermedades isquémicas del corazón con 12.3 % de defunciones del total registrado y enfermedades del hígado con 5.5 % de defunciones en tercer lugar. Otras enfermedades tuvieron un porcentaje menor al 5.5 % del total de causas de defunción. En cuanto al primer lugar que fue Diabetes Mellitus, tuvo un porcentaje del 14 %, lo que se traduce a que 14 de cada 100 pacientes fallecieron a causa

de Diabetes Mellitus en el año 2012, en todo el país. Y como la Diabetes Mellitus fue la mayor causa de mortalidad para ese año se seleccionó para el estudio. Después se realizó una investigación documental acerca de la Diabetes Mellitus, consultando diversas fuentes de información. En el capítulo 2 se muestra ésta información.

En el siguiente capítulo se muestra la organización de la estructura de éste trabajo y el resumen de cada capítulo.

Capítulo 1

El modelado Estadístico

En este primer capítulo se mencionan cuales son los pasos que se siguieron para obtener los modelos explicativos.

1.1. Estructura de ésta tesis

La estructura de ésta tesis se elaboró de acuerdo al procedimiento de modelado explicativo del artículo científico *To Explain or to Predict* [7] (Galit Schmueli,2010). Dicho procedimiento consiste en ocho pasos los cuales son "Definir el objetivo (paso 1),Diseño del estudio y recolección de los datos (paso 2),Preparación de los datos(3), Análisis exploratorio de los datos (4), Elección de variables (5), Elección de métodos (6), Evaluar, validar y seleccionar modelo (7) y Uso y del modelo y reporte (8)". Puede decirse que se siguió al pie de la letra dicho procedimiento, sin embargo se juntaron dos pasos en dos capítulos los cuales fueron "Diseño del estudio y recolección de datos y Preparación de los datos", así como "Análisis exploratorio de datos y "Elección de variables". Adicionalmente en el último capítulo se agregó la comparación de los MLG Logístico, quasibinomial, poisson y quasipoisson. En el cuadro 1.1 se muestra cómo se estructuró esta tesis.

A continuación se muestra la síntesis de cada capítulo en cada sección.

Pasos	Nombre del capítulo
1	Definición del objetivo del análisis (capítulo 2)
2 y 3	Diseño del estudio, recolección y preparación de los datos (capítulo 3)
4 y 5	Análisis exploratorio de los datos y elección de variables (capítulo 4)
6	Elección de métodos (capítulo 5)
7	Selección,evaluación y validación de los MLG Logístico, quasibinomial, poisson y quasipoisson (capítulo 6)
8	Uso, reporte y comparación de los MLG Logístico, quasibinomial, poissson y quasipoisson (capítulo 7)

Cuadro 1.1: Pasos del modelado estadístico según Schmueli y Capítulos de ésta tesis

1.1.1. Definición del objetivo del análisis (capítulo 2)

En este paso se pretende explicar el para qué se realizó este estudio. Es decir, la finalidad de su elaboración. Sin embargo, debe mencionarse que parte del objetivo es la elaboración de un modelo explicativo. Esto puede verse con más detalle en la sección 2.1.

1.2. Diseño del estudio, recolección y preparación de los datos (capítulo 3)

En el modelado explicativo el objetivo es estimar un modelo denotado por f, definido como un modelo estadístico que representa adecuadamente un modelo teórico que se denota por F. En este tipo de modelado la potencia estadística es la consideración principal. Eso se resume en reducir el sesgo, lo cual requiere de suficientes datos para la prueba de especificación del modelo según Galit Schmueli. En ésta tesis se utilizan todos los datos de la unión de las bases de datos de defunciones y egresos hospitalarios de la secretaría de salud del año 2012, de toda la república mexicana. En la teoría también se menciona que para la explicación causal, los datos experimentales se prefieren más que los observacionales. Sin embargo este estudio es observacional pues los datos relacionados con la defunción a causa de Diabetes Mellitus obtenidos en la base de datos de la secretaría de salud, "se han observado y registrado sin intervenir en el curso natural de este acontecimiento" (según Scielo.conicyt.cl, 2018 [18]).

El siguiente paso es preparación de los datos. Este paso se refiere al proceso de elimi-

nación de valores que no aportan significado al conjunto de datos. Es decir, es insatisfactorio utilizar valores perdidos para modelos explicativos según Galit Schmueli. Es preciso mencionar que se eliminaron todos los valores perdidos para este estudio. En este capítulo (capítulo 3) se explica cómo se codificaron las variables para identificarlas en el momento de realizar el análisis con los Modelos lineales generalizados (ver sección 3.3).

1.3. Análisis exploratorio de los datos y elección de variables (capítulo 4)

El análisis exploratorio de los datos consiste en resumir los datos numéricamente y gráficamente, reducir su dimensión y prepararlos para un paso de modelado más formal. Los visualizadores son usados para explorar datos para confirmar hipótesis o manipular un artefacto que permita ver imágenes. En la visualización exploratoria, el usuario no sabe necesariamente que está buscando . En un modelo explicativo, un resumen numérico podría enfocarse en las relaciones teóricas (Schmueli,2010). Para el caso de ésta tesis se generaron tablas de contingencia las cuales muestran la relación de las defunciones y afecciones de Diabetes Mellitus con las variables explicativas (mostradas en la sección 3.4.1).

Por otro lado "el paso de la elección de variables se basa en el rol del constructo, la estructura teórica causal y en la operativización" según Schmueli. En el caso del trabajo presente debe resaltarse que como se obtuvieron las variables de unas bases de datos no especializadas en la enfermedad, se limitó la selección de las variables. Por tanto se eligieron las variables Sexo, Región Socioeconómica, Tipo de complicación, Tipo de Diabetes Mellitus, Edad Decenal y Tipo de Diabetes Mellitus que fueron obtenidas de la unión de las bases de datos de la secretaría de salud de defunciones y egresos hospitalarios 2012 (sección 4.2).

1.4. Elección de métodos (capítulo 5)

Schmueli escribió también que "el modelado explicativo requiere modelos estadísticos f fáciles de interpretar y que estén ligados al modelo teórico F." Por éstas razones, en ésta tesis se utilizaron los modelos lineales generalizados. Es decir, porque son fáciles de

interpretar y porque permiten tener más opciones al momento de elegir la distribución de la variable dependiente, como en éste caso la variable de defunción por Diabetes Mellitus. En ese capítulo se muestran las ventajas y desventajas de utilizar los modelos lineales generalizados (MLG). Así como se expone en breve la teoría de los modelos lineales generalizados y en particular los modelos logístico, poisson, quasibinomial y quasipoisson.

1.5. Selección,evaluación y validación de los MLG Logístico, quasibinomial, poisson y quasipoisson (capítulo 6)

En el capítulo 6 se muestra el proceso de evaluación, validación y selección del modelo. Se comenzó a elegir al seleccionar modelos de entre un conjunto de modelos posibles (elaborados con las variables mencionadas en la sección anterior). Para ello se utilizaron criterios como el AIC y se compararon con los cocientes de dispersión. En este caso se mostrará el ajuste con cuatro modelos distintos, los cuales son los modelos lineales generalizados Binomial logístico, Quasibinomial, Poisson y Quasipoisson.

1.5.1. Evaluación del modelo

"La intención de este paso es calcular la potencia explicativa la cual mide la fuerza de asociación" según Galit Schmueli. Para ello se utilizan estadísticos como la pseudo-R^2. Para ésta tesis se utilizaron varias pseudo-R^2 (ver sección 6.2).

1.5.2. Validación

Según Schmueli consiste en dos partes:

- La validación del modelo: valida que f, el modelo estadístico, representa adecuadamente F, el modelo teórico.
- El ajuste del modelo: Valida que $\hat{f}$, el modelo seleccionado ajusta los datos X, Y

En ésta parte del proceso de modelado se usa la inferencia para coeficientes individuales, que también se usa para detectar sobreespecificación o infraespecificación. Así

también implica pruebas de bondad de ajuste (e.g. Análisis de devianza) y diagnóstico del modelo tal como análisis residual.

Para el trabajo presente se utilizaron estadísticos como residuos, así también se utilizaron la devianza residual y la nula, gráficas de interacciones de elogits, parámetros de dispersión y gráficas de varianza contra media estimadas. Todos éstos estimadores y gráficos son definidos y ejemplificados en la sección 6.3.

1.5.3. Uso, reporte y comparación de los MLG Logístico, quasibinomial, poisson y quasipoisson (capítulo 7)

Este último paso lo ideal sería (según Schmueli) "derivar conclusiones estadísticas utilizando inferencias y luego transladarlas a conclusiones científicas considerando el modelo teórico F y las variables explicativas X y la hipótesis causal. Como los modelos explicativos se enfocan en la teoría, en el sesgo, la causalidad y en el análisis retrospectivo, entonces se utilizan para probar o comparar la existencia de teorías causales." Aquí sólo se mostrará la forma en cómo se interpretaron los parámetros estimados de los cuatro modelos. También se muestra la interpretación de sus respectivos intervalos de confianza.

Capítulo 2

Definición del objetivo del análisis

Lo que se muestra a continuación es una serie de definiciones que llevarán al entendimiento de la definición del objetivo para el cual se ha elaborado el modelado explicativo de ésta tesis. Algunas definiciones se extrajeron del artículo científico *To Explain or to Predict* [7] (Galit Schmueli,2010). Mientras que el resto se extrajeron de varios artículos y páginas web mencionados en cada sección. El objetivo del análisis se define al final de éste capítulo.

2.1. El modelado explicativo

El modelado explicativo se refiere a la aplicación de modelos estadísticos a datos, para probar hipótesis causales sobre constructos teóricos. Vale la pena mencionar que "en el modelado explicativo el rol de la teoría es muy fuerte y la dependencia de los datos y el modelado estadístico está hecha estríctamente a través de los lentes del modelado teórico" (Schmueli, 2010).

A continuación se da una serie de definiciones que ayudarán a comprender el proceso del modelado explicativo.

2.1.1. Constructo

"Los constructos teóricos son abstracciones que describen un fenómeno de interés teórico" (Edwards y Bagozzi, 2000)[11]. " También pueden ser llamados variables abstractas.

Algunos ejemplos de constructos son el hambre, la pobreza, el buen comportamiento, etc." (G. Schmueli, 2010). Y en el caso de éste estudio se utilizarán como variables abstractas la defunción de diabetes mellitus, la edad, las regiones socioeconómicas, el sexo, y los tipos de complicación de la diabetes mellitus. En el capítulo siguiente se hondará la operativización de dichas variables, concepto que será aclarado en la subsección 2.1.4.

2.1.2. Características del modelado explicativo

Según Schmueli:

1. Causalidad-Asociación. El modelo estadístico f representa una función subyacente causal, y se asume que X (los factores subyacentes) causan Y (el efecto subyacente).

2. Teoría-Datos. En el modelado explicativo, f está construido cuidadosamente, basado en F, (función por la cual el constructo **x** mediante cierta teoría, se dice que causa **y** de forma que apoye la interpretación de la relación estimada entre X y Y y que pruebe las hipótesis causales.

3. Es retrospectivo. El modelado explicativo es retrospectivo, en el cual f es utilizado para probar un conjunto de hipótesis ya existentes.

4. Sesgo-Varianza. Según Hastie, Tibshirani y Friedman en su libro *The elements of statistical learning* [5] (2017):

   entre más complejo se hace el modelo estimado, menor es el sesgo (cuadrado) pero mayor es la varianza, (para un modelo predictivo). Mientras Galit Schmueli dice que en el modelado explicativo el enfoque es en mimimizar el sesgo para obtener la representación más adecuada para la teoría subyacente. A continuación se detalla la forma en como Hastie, Tibshirani y Friedman exponen la descomposición Sesgo-Varianza.

   Si se asume que $Y = f(x) + \epsilon$ donde $E(\epsilon) = 0$ y $Var(\epsilon) = \sigma_\epsilon^2$, se puede deducir una expresión para el error de predicción esperado para un ajuste de regresión $\hat{f(x)}$ en un punto de entrada $X = x_0$, utilizando un error cuadrático de pérdida como sigue:

$$\begin{aligned} Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 \\ &= \sigma_\epsilon^2 + Sesgo^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \\ &= \text{Error irreducible} + Sesgo^2 + Varianza \end{aligned}$$

El primer término es la varianza objetivo, alrededor de su media verdadera $f(x_0)$, y no puede ser eludida, no importa que tan bien se estimó $f(x_0)$ a no ser que $\sigma_\epsilon^2 = 0$. El segundo término es el sesgo cuadrado, cantidad por la cual el promedio de nuestro estimador difiere para la media verdadera; El último es la varianza; la desviación esperada cuadrada de $\hat{f}(x_0)$ alrededor de su media.

5. Para la explicación causal, se prefieren los datos experimentales a los observacionales. Por lo que para el diseño se elije un escenario experimental, pero esto depende de la disponibilidad, es decir, de la oferta y de los recursos obligados. Por ejemplo en el caso de ésta tesis los datos que se utilizaron fueron observacionales debido a los recursos que requiere montar un estudio especializado en Diabetes Mellitus.

2.1.3. Hipótesis causales

De acuerdo a la página web Academia.edu (Triviño, 2018 [19]),

Las hipótesis causales o bien hipótesis que establecen relaciones de causalidad afirman relaciones entre dos o más variables, cómo se dan dichas relaciones y también proponen un «sentido de entendimiento» entre ellas.

Las hipótesis causales se simbolizan como:

$$X \overset{\text{influye en o causa}}{\longrightarrow} Y$$

La causa X es el conjunto de variables independientes, mientras que el efecto Y es la variable dependiente. Para poder establecerse causalidad antes debe haberse demostrado correlación entre las causas y la variable dependiente, pero además la causa debe ocurrir antes que el efecto.

Por ejemplo, con relación esta tesis, una hipótesis causal es “En 2010 en México, la edad promedio de defunción por DM fue 66.7 años, por lo que se redujo la esperanza de vida 10 años” (Hernández-ávila, Pablo Gutiérrez and Reynoso-Noverón, 2013)[9] . En páginas posteriores se mencionan estas hipótesis causales por cada constructo.

2.1.4. Operativización

Conforme a lo explicado en la página Explorable.com (2018) [20]:

Como operativización se entiende el proceso de definir estrictamente variables en factores medibles. Es decir, es el proceso de definir los constructos de forma que sean medibles empírica y cuantitativamente. La operativización también establece definiciones exactas de cada variable, incrementando la calidad de los resultados y mejorando la robustez del diseño. Otra observación importante de la operativización es que su objetivo es facilitar la replicación exacta del proceso de investigación.

Se debe aclarar que en esta tesis se utiliza la operativización de los constructos sexo, edad, regiones socioeconómicas , tipos de compicación y tipos de diabetes Mellitus como variables de los modelos explicativos y no los constructos que se refieren a éstas variables. En la sección 3.6 se muestra como se operativizaron dichos constructos.

Definición formal del modelado explicativo

Según G. Schmueli:

Considérese una teoría que postula que el constructo **x** causa un constructo **y**, mediante la función subyacente F, tal que **y**=F(x). F se representa comunmente por un conjunto de oraciones o enunciados cualitativos, una gráfica o fórmulas matemáticas.

Sean las variables medibles X y Y operativizaciones de **x** y **y** respectivamente. Y f , un modelo estadístico que operativiza F tal que **E**(Y)=f(X).

Lo que se desea lograr en el modelado explicativo es hacer coincidir f y F tanto como sea posible en la inferencia estadística para aplicar f a las hipótesis teóricas. Los datos X y Y son herramientas para estimar f, la cual a su vez se utiliza para probar hipótesis causales.

2.2. Definición de Diabetes Mellitus (DM)

Según la revista American Diabetes Association (Diagnosis and Classification of Diabetes Mellitus, 2003) [12] la Diabetes Mellitus (ó DM) “es un grupo de padecimientos metabólicos caracterizados por la hiperglucemia, resultado de los defectos de la secreción de la insulina, la acción de la insulina o ambos.”

Ahora bien, de acuerdo con la página web American Diabetes Association, 2018 [21]

la hiperglucemia es el término técnico que utilizamos para referirnos a los altos niveles de azúcar en la sangre. El alto nivel de glucemia aparece cuando el organismo no cuenta con la suficiente cantidad de insulina o cuando la cantidad de insulina es muy escasa. La hiperglucemia también se presenta cuando el organismo no puede utilizar la insulina adecuadamente.

Y en términos numéricos, "cuando la glucosa en sangre en ayunas es superior a 126 mg/dL en por lo menos dos pruebas de sangre tomadas en diferentes momentos, tenemos criterio para el diagnóstico de la diabetes" (Mdsaude.com, 2018) [22].

Para fines de observación de las causas de defunción por diabetes mellitus puede mostrarse según Mediavilla Bravo (2018) [23] que la DM se puede asociar con complicaciones agudas que pueden dar lugar a alteraciones importantes, en caso de no tratamiento urgente, como precipitación de accidentes cardiovasculares o cerebrovasculares, lesiones neurológicas y coma. Igualmente, la hiperglucemia crónica de la diabetes se asocia a largo plazo a lesiones que provocan disfunción y fallo de varios órganos, en especial ojos, riñones, nervios, corazón y vasos sanguíneos.

Algo más que debe añadirse es que la DM fue la principal causa de muerte de las mujeres en nuestro país en el año 2000 y la segunda para los hombres (Estadísticas de mortalidad en México: muertes registradas en el año 2000, 2002).([13])

Según la Asociación Diabetes Madrid, (2018) [24] existen cuatro clases de diabetes mellitus las cuales son de tipo I, en la cual el cuerpo no produce insulina y en la de tipo II, la más común, en la que el cuerpo no produce suficiente o no usa la insulina de manera adecuada, la diabetes Gravídica o diabetes gestacional, de aparición en el embarazo y otros tipos de diabetes; En ésta tesis se utilizan sólo los casos de los afectados y difuntos a causa de diabetes mellitus tipo I y tipo II.

Diabetes Mellitus tipo I y tipo II

Diabetes Mellitus tipo I La DM tipo I es un subtipo de DM caracterizado por deficiencia de insulina. Se manifiesta por el inicio repetido de hiperglucemia severa, progresión rápida hacia cetoacidosis diabética y la muerte, a menos que sea tratada con insulina. La enfermedad puede ocurrir a cualquier edad, pero es más común durante la infancia y la adolescencia. (Dtc.ucsf.edu, 2018) [25].

Diabetes Mellitus tipo II La diabetes tipo II, es el tipo más común de diabetes. Se dice también que algunos grupos tienen mayor riesgo de contraer éste tipo de diabetes como los afroamericanos, los latinos/hispanos, indígenas americanos, estadounidenses de origen asiático, nativos de Hawái y otros isleños del pacífico, como también entre las personas mayores. Con la DM tipo 2 el cuerpo no produce suficiente insulina o las células no hacen uso de la insulina. La insulina es necesaria para que el cuerpo pueda usar la glucosa como fuente de energía. (Dtc.ucsf.edu, 2018) [26].

Tipo de complicaciones por DM

Coma Diabético Es una grave enfermedad que transcurre como una complicación de la diabetes tipo II y que se caracteriza por niveles extremadamente altos de azúcar (glucosa) en la sangre. Se dice que puede ser causada por una enfermedad infecciosa, o por otra enfermedad como infarto al miocardio, o accidente vascular cerebral. Por edad avanzada, insuficiencia renal o suspensión de insulina. Suele verse en personas con DM tipo II o bien DM insulinodependiente (Geosalud.com, 2018) [27].

Cetoacidosis diabética Afección grave que puede producir coma diabético o incluso la muerte. Cuando las células no reciben la glucosa que necesitan como fuente de energía, el cuerpo comienza a quemar grasa para tener energía, lo que produce cetonas. El cuerpo hace ésto cuando no tiene suficiente insulina para usar glucosa, la fuente normal de su cuerpo. Cuando las cetonas se acumulan en la sangre, ésto hace que la sangre sea más ácida. Es poco común en personas con DM tipo II. La cetoacidosis, una vez que se está enfermo de DM, puede ser causada por no inyectarse suficiente insulina, saltarse comidas (no alimentarse con suficiencia) o como reacción a la insulina.(American Diabetes Association, 2018) [28]

Diabetes Mellitus con complicaciones renales En personas con DM las nefronas (unidades pequeñas del riñon) lentamente se engruesan y con el tiempo cicatrizan. Las nefronas comienzan a dejar pasar proteína (albúmina) a la orina. Este daño puede suceder antes del comienzo de cualquier síntoma. Dichas complicaciones pueden ser provocadas por falta de control de la azúcar en la sangre, la hipertensión arterial o por que se tiene DM tipo I que comenzó antes de los 20 años. También puede influir la presencia de familiares

con DM y problemas renales o también que la persona con DM fuma. (Medlineplus.gov, 2018) [29]

Complicaciones oculares de la Diabetes

- **Glaucoma.** Las personas con diabetes son 40 % más propensas a tener glaucoma que las personas sin diabetes. Cuanto más tiempo la persona haya tenido diabetes, más común el glaucoma. El riesgo también aumenta con la edad. El glaucoma ocurre cuando aumenta la presión en el ojo. En la mayoría de los casos, la presión causa que el humor acuoso drene más lentamente, de manera que se acumula en la cámara anterior. La presión aplasta los vasos sanguíneos que llevan sangre a la retina y el nervio óptico. Se pierde la visión gradualmente porque se daña la retina y el nervio.

- **Cataratas.** Muchas personas sin diabetes tienen cataratas, pero esta afección de los ojos es 60 % más común entre las personas con diabetes. Además, las cataratas tienden a afectar a las personas con diabetes a menor edad y a avanzar más rápido. Con cataratas, el lente claro del ojo se nubla, bloqueando la luz.

- **Retinopatía.** Retinopatía diabética es un término general para todos los trastornos de la retina causados por la diabetes. Hay dos tipos principales de retinopatía: no proliferativa y proliferativa. Retinopatía no proliferativa Con la retinopatía no proliferativa, el tipo más común de retinopatía, los vasos capilares en la parte trasera del ojo se hinchan y forman bolsas. La retinopatía no proliferativa puede tener tres etapas (leve, moderada y severa), a medida que se obstruyen más y más vasos sanguíneos.

- **Edema macular.** A pesar de que la retinopatía por lo general no causa pérdida de visión en esta etapa, las paredes capilares pueden perder la capacidad de controlar el flujo de sustancias entre la sangre y la retina. Puede haber fugas de líquido a la parte del ojo donde ocurre el enfoque, la mácula. Cuando la mácula se hincha con líquido, una afección llamada edema macular, la visión se vuelve borrosa y se puede perder del todo. Si bien la retinopatía no proliferativa generalmente no requiere tratamiento, es necesario tratar el edema macular, pero afortunadamente, el tratamiento generalmente logra detener y a veces revertir la pérdida de la visión.

- **Retinopatía proliferativa.** En ciertas personas, después de varios años, la retinopatía avanza y se convierte en un tipo más serio, llamado retinopatía proliferativa. Con este tipo, hay tanto daño a los vasos sanguíneos que estos se cierran. En respuesta, comienzan a crecer nuevos vasos sanguíneos en la retina. Estos nuevos vasos son débiles y pueden tener fugas de sangre, lo que bloquea la visión y se denomina hemorragia vítrea. Los nuevos vasos sanguíneos también pueden causar cicatrices. Cuando las cicatrices se encogen, pueden distorsionar la retina o jalarla fuera de lugar, un trastorno llamado desprendimiento de retina. Varios factores influyen en la retinopatía:

  - control de la glucosa en la sangre
  - presión arterial
  - tiempo que ha tenido diabetes
  - factores genéticos

  Mientras más tiempo haya tenido diabetes, mayor la probabilidad de tener retinopatía. Casi todas las personas con diabetes tipo 1 a fin de cuentas tienen retinopatía no proliferativa. La mayoría de las personas con diabetes tipo 2 también la padecen. Pero es mucho menos común la retinopatía proliferativa que puede causar ceguera. (American Diabetes Association, 2018) [30]

Neuropatía diabética Las neuropatías diabéticas son un conjunto de trastornos nerviosos causados por la diabetes. Con el tiempo, las personas con diabetes pueden desarrollar daño de los nervios en todo el cuerpo. Algunas personas con daño nervioso no presentan síntomas, mientras que otras pueden presentar síntomas tales como dolor, hormigueo o adormecimiento—pérdida de sensación—en las manos, brazos, piernas y pies. Los problemas de los nervios pueden presentarse en cualquier sistema de órganos, incluidos el tracto digestivo, el corazón y los órganos sexuales.

Cerca de un 60 a 70 por ciento de personas con diabetes sufren algún tipo de neuropatía. Las personas con diabetes pueden desarrollar trastornos nerviosos en cualquier momento, pero el riesgo aumenta con la edad y con una diabetes más prolongada. Las tasas más altas de neuropatía se encuentran en personas que tienen diabetes por al menos durante 25 años. Las neuropatías diabéticas también parecen ser más comunes en personas

que tienen problemas en controlar la glucosa en la sangre, también llamado azúcar en la sangre, así como en aquellas personas con niveles elevados de grasa corporal y presión arterial, y en aquellas que tienen sobrepeso. (National Institute of Diabetes and Digestive and Kidney Diseases, 2018) [31]

Enfermedad Vascular periférica Transtorno de la circulación lento y progresivo que incluye todas las enfermedades en cualquiera de los vasos sanguíneos fuera del corazón y las enfermedades de los vasos linfáticos. Los órganos que reciben suministro de sangre pueden no recibir el flujo sanguíneo adecuado. Sin embargo las partes más afectadas son las piernas y los pies.

Un factor de riesgo es todo aquello que puede aumentar la probabilidad que tiene una persona de desarrollar una enfermedad. Puede ser una actividad, la alimentación, los antecedentes familiares o muchas otras cosas. Ciertos factores de riesgo de la enfermedad vascular periférica pueden modificarse o tratarse, mientras que otros no.

Entre estos últimos podemos mencionar:

- edad (en especial a partir de los 50 años)
- antecedentes de enfermedad cardíaca
- ser varón
- diabetes mellitus (diabetes de tipo I)
- postmenopausia
- antecedentes familiares de dislipidemia (niveles altos de lípidos en la sangre, como colesterol), hipertensión o enfermedad vascular periférica

Entre los factores de riesgo que pueden modificarse o tratarse podemos mencionar:

- enfermedad coronaria
- disminución de la tolerancia a la glucosa
- dislipidemia
- hipertensión (presión sanguínea alta)

- obesidad
- inactividad física
- fumar o consumir productos de tabaco

(Gwheartandvascular.org, 2018)[32]

Letalidad "Es la cantidad de personas que mueren en un lugar y en un período de tiempo determinados en relación con el total de la población." (Oxford Dictionaries | Español, 2018) [33]

Ya establecidas éstas definiciones, a continuación se enlistan un conjunto de enunciados que corresponden a la teoría subyacente del modelo explicativo.

2.3. Constructos y sus hipótesis de dependencia causal

Para elegir los constructos se hizo investigación en libros y en internet y se extrajo lo que se consideró que era de mayor relevancia para explicar las defuciones a causa de DM en el año 2012, en nuestro país. Sin embargo se encontraron publicaciones como "Estadísticas y causa de mortalidad en la diabetes tipo 2"(Salgado Pineda et al., 2001) [14] y se consideraron los constructos que tomaron para dicha publicación de que son los siguientes:

- Datos de la filiación.
- Datos clínicos como datos relacionados al control de la diabetes, control glucémico, perfil lipídico, tratamientos para la diabetes, incidencia de complicaciones crónicas y tiempo medio de evolución de la diabetes.
- Datos clínicos no relacionados directamente con la diabetes como tabaquismo, índice de masa corporal, control de la tensión arterial y otros tratamientos administrados.
- Causa clínica de la muerte.
- Causa de mortalidad que figura (es, decir se obtuvo el dato de los certificados de defunción del registro civil) el registro civil.

Podría decirse que los constructos antes mencionados corresponden a los constructos ideales deseados para el estudio de mortalidad de Diabetes Mellitus. Pero los recursos para esta tesis se limitan a una base de datos de la secretaría de salud donde sólo se encontraron los constructos que a continuación se mencionan. En el capítulo 6 puede observarse el costo que limitados constructos conllevan sobre el modelo explicativo.

2.3.1. Edad

En la base de datos de defunciones y egresos hospitalarios de la Secretaría de Salud en 2012, la edad, fue la edad obtenida del acta o certificado de defunción entregada por el registro civil y la edad registrada en las unidades médicas de la secretaría de salud y los servicios estatales de salud respectivamente (ver sección 3.2). Dicha edad es el dato estadístico que se utilizó en los modelos lineales generalizados mencionados en el capítulo 6. Una hipótesis de dependencia causal para éste constructo podría ser en 2010 en México, la edad promedio de defunción por DM fue de 66.7 años, por lo que se redujo la esperanza de vida 10 años (Hernández-ávila, Pablo Gutiérrez and Reynoso-Noverón, 2013)[9].

2.3.2. Sexo

El constructo sexo, al igual que el constructo edad, para este trabajo (obtenidos de las bases de datos mencionadas en el capítulo 3), se obtuvo del acta o certificado de defunción entregada por el registro civil y fue el sexo registrado en las unidades médicas de la secretaría de salud y los servicios estatales de salud. Una hipótesis causal para el sexo podría ser que “a partir del año 2000, la diabetes mellitus es la primera causa de muerte en mujeres. Y la segunda en hombres” (Sites.google.com, 2018). [34]

2.3.3. Complicaciones

Complicaciones según la página web con título Definiciones y conceptos fundamentales para la calidad en salud (2018)[35] es el “evento que sobreviene durante la prestación de la atención médica, ya sea por la historia natural de la enfermedad, por riesgo inherente, o por iatropatogenia.” “Donde la iatropatogenia es “la lesión generada a un paciente consecuencia de impericia, temeridad, negligencia o dolo del personal de salud.” (Diccionario.leyderecho.org, 2018) [37] Una hipótesis causal para las complicaciones de DM

es que las complicaciones renales tuvieron el mayor porcentaje de letalidad en 1994, en México (Peña y Rico-Verdín, 1996). [36] (cuadro 2.1).

Complicación	**Porcentaje de letalidad (%)**
Coma	36
Renales	11
Cetoasidosis	5
C. periféricas	3
Neurológicas	1

Cuadro 2.1: Cuadro de Complicación contra porcentaje de letalidad. Se estimó la letalidad de cada complicación al dividir el número de defunciones por causa especifica entre el total de egresos por esa causa (Peña y Rico-Verdín, 1996)
.

2.3.4. Tipo de Diabetes Mellitus

El constructo de tipo de diabetes Mellitus se extrajo de la variable tipo de complicación, realizando un procedimiento de filtro en excel. Lo anterior se explica con mayor detalle en la sección 3.6. Según la página Diabetes Mellitus MX (2018) [39] las personas con diabetes Mellitus tipo 2 tuvieron un 15 % más de riesgo de muerte prematura comparado con las personas sanas. Lo anterior puede funcionar como ejemplo de hipótesis causal para éste constructo.

2.3.5. Entidad Federativa

El constructo entidad federativa se operacionalizó con la variable Entidad de residencia (Ent_resid) encontrada en las bases de datos de la secretaría de salud de defunciones y egresos hospitalarios (2012) mencionadas en la sección 3.2. Se eligió la entidad de residencia y no la entidad de registro que también se encontraba en dichas bases de datos, porque se consideró que las condiciones de vida del lugar de residencia podrían tener relación con la defunción a causa de DM.

Según el folleto encontrado en la página Oment.uanl.mx [38], “Más de la mitad de las entidades federativas del país reportan alarmantes tasas de mortalidad por cada 100,000

mil habitantes, superiores al promedio de las naciones de la OCDE y ninguna de ellas se encuentra por debajo del promedio."

Y ésto sirve para mostrar que en otros estudios han utilizado a las entidades federativas como constructo para sus estudios de DM.

En el mismo folleto se mencionó que la mayor parte de los estados que reportan altas tasas de mortalidad por diabetes mellitus están ubicadas en la región centro; entidades como la Ciudad de México (110), Veracruz (103), Tlaxcala (93), Puebla (92), Tabasco (92), Morelos (90), Guanajuato (90), Coahuila (88), Michoacán (87), Colima (81) y el Estado de México (81). Comparativamente, en 2015 la tasa de homicidios dolosos fue de 16 por cada 100,000 mil habitantes.

Por tanto una hipótesis causal de dicho constructo podría ser "La asociación de mayor fuerza con la mortalidad por diabetes se registró en la Ciudad de México, (RR 2,5; IC 2,33-2,68 en 2000; RR 2,06; IC 1,95-2,18 en 2007)"(Sánchez-Barriga, 2018)[15].

2.4. Conclusión del capítulo: Objetivo del modelado explicativo

Por lo tanto el objetivo del modelado explicativo es probar las siguiente hipótesis causales en el contexto de la república mexicana en el año 2012:

a) La edad promedio de defunción por DM fue de 66.7 años, por lo que se redujo la esperanza de vida 10 años.

b) La diabetes Mellitus fue la primer causa de defunción para las mujeres.

c) El coma diabético fue la complicación de mayor letalidad.

d) La diabetes mellitus tipo 2 causó una mayor cantidad de defunciones que la diabetes mellitus tipo 1.

e) La asociación de mayor fuerza con la mortalidad por diabetes se registró en la Ciudad de México.

Capítulo 3

Diseño del estudio, recolección y preparación de los datos

3.1. Diseño de estudio

El diseño que se eligió para este estudio es un modelo explicativo. Dicho diseño tuvo como objetivo principal ilustrar la aplicación de los modelos lineales generalizados logístico, Poisson, quasilogístico y quasipoisson. Como se vió en el capítulo anterior las variables seleccionadas para el modelo se encontraron en varios artículos y páginas web. En parte por ello se eligieron para el modelo, aunque también por el estudio formal realizado con las tablas de contingencia del capítulo 4. Además dichas variables fueron las únicas localizadas en las bases de datos de defunciones y egresos hospitalarios (2012) de la secretaría de salud, como ya se mencionó en el capítulo anterior. Por lo anterior al realizar la unión de las bases de datos sólo se consideraron cuatro constructos los cuales fueron: entidad federativa, que se operacionalizó con las variables de entidad de residencia de las bases de datos mencionadas; la edad que se operacionalizó con la variable edad en ambas bases de datos; el sexo, operacionalizado con la variable sexo de dichas bases de datos; complicaciones y tipo de DM, operacionalizadas con la variable causa de defunción y afección principal CIE de las bases de datos de defunciones y egresos hospitalarios respectivamente.

En la siguiente sección se explica detalladamente la fuente de las bases de datos de defunciones y egresos hospitalarios 2012.

3.2. Recolección de datos

Para éste trabajo se utilizaron dos bases de datos. Una de defunciones y otra de egresos hospitalarios las cuales se detallan a continuación.

Base de datos de las defunciones de 2012 a nivel nacional (DEFUN12) Se utilizó la base de datos de las defunciones de 2012 a nivel nacional (DEFUN12) de la secretaría de salud, de la página (DEFUNCIONES GENERALES (INEGI/SALUD), 2018) [40]. En dicha página se menciona que la base de datos se obtuvo mediante la aplicación llamada Subsistema Epidemiológico y Estadístico de Defunciones (SEED).

Es importante mencionar que la base de datos considera la defunción "como la desaparición permanente de todas las funciones vitales de una persona ocurridas después de ser declarado nacido vivo". En el archivo con extensión zip descargado en el link mencionado se encontraba un archivo que describía la base de datos

`Descripcion_BD_Defunciones_2012.pdf` con formato *.pdf*, normalizada, en la que se distribuía la información obtenida de las actas o certificados de defunción entregados por el registro civil y cuadernos estadísticos que proveen las agencias del ministerio público.

Cabe destacar que la base de datos contiene la información de todos los difuntos registrados a nivel nacional en el año 2012.

Base de datos de egresos hospitalarios de 2012 a nivel nacional También se utilizó la base de datos de egresos hospitalarios extraída de la página (SECRETARÍA DE SALUD 2000-2015, 2018) [41]. De ella se descartaron las defunciones de la parte de egresos hospitalarios para obtener información de sólo los pacientes vivos que fueron internados a causa de estar enfermos de Diabetes Mellitus. Esta página menciona que la base de datos se obtuvo mediante la aplicación tecnológica denominada subsistema Automatizado de Egresos Hospitalarios (SAEH).

Por egreso hospitalario se entiende de ahora en adelate al evento de salida del paciente del servicio de hospitalización que implica la desocupación de una cama censable. Incluye altas por curación, mejoría, traslado a otra unidad hospitalaria, defunción, alta voluntaria o fuga. Excluye movimientos entre diferentes servicios dentro del mismo hospital. Esta base de datos no considera los recién nacidos sanos. Cabe mencionar que la base de datos

contenía el número de egresos hospitalarios registrados en las unidades médicas de la secretaría de salud y los servicios estatales de salud.

3.3. Preparación de los datos

En la sección actual se explica de qué forma se preparó la base de datos antes de la realización del análisis de regresión.

3.3.1. Obtención de la base de datos

Para la obtención de la base de datos se requirieron dos bases de datos mencionadas en los párrafos anteriores. Debe recordarse que las bases de datos contenían datos de los registros de los certificados de defunción y egresos hospitalarios respectivamente. Por lo cual contenían los casos de los individuos difuntos y los que salieron de los hospitales de las unidades médicas de la Secretaría de Salud y los servicios estatales. Antes de todo el proceso de unión de las bases se abrió la base de datos en excel y se realizó un filtro manual en la variables de causa de defunción y afección principal de las bases de datos de defunciones y egresos hospitalarios respectivamente. Con tal filtro manual se seleccionaron sólo los difuntos y enfermos por diabetes mellitus tipo I y tipo II. A continuación se explica paso a paso la obtención de la base elaborada para ejemplificar los modelos lineales generalizados logístico, poisson y quasipoisson.

1) Como se deseaba unir las bases de datos de defunciones y egresos hospitalarios 2012, se debían encontrar casos coincidentes, por lo cual se escogieron variables que se encontraban en ambas bases de datos y estaban codificadas numéricamente y coincidían entre ellas (ver sección 3.6). En seguida se enlistan dichas variables.

    - Para la base de datos de Defunciones 2012:
        a) Entidad de residencia, encontrada como “ent_resid”.
        b) Edad del paciente, que se encontró como “EDAD”.
        c) Sexo del paciente, hallada como “sexo”.
        d) causa de defunción, encontrada como “causa_def”.
    - Para la base de datos de Egresos Hospitalarios 2012:

a) Entidad de residencia, encontrada como “ENTIDAD”.

b) Edad del paciente, que se encontró como “EDAD”.

c) Sexo del paciente, hallada como “SEXO”.

d) afección principal, encontrada como “AFECPRIN”.

2) El siguiente paso fue recodificar la edad en decenas. Tal edad fue renombrada como edad decenal. Lo anterior se realizó con la intención de crear una menor cantidad de categorías para dicha variable, así como para simplificar el modelo explicativo. Pues de lo contrario se ubiera tenido más de cien categorías para dicha variable (que tenía edades desde 0 al 110). En cambio al recodificar sólo se obtuvieron cinco categorías de edad decenal.

3) Con las variables mencionadas se realizaron tablas dinámicas en excel para simplificar la información. A partir de la base de datos se seleccionaron dichas variables y se sumó el total de defunciones y de egresos hospitalarios.

4) Se ordenaron los casos individuales de egresos hospitalarios y defunciones por entidad de registro (Del 1 al 32).

5) Se observó que las variables de sexo, causa de defunción o afección principal debían ordenarse para ambas bases de datos. Por esto se eligió generar treinta y dos hojas para separar las entidades (32), por cada base de datos. Es decir en 64 hojas de excel se contenieron los casos por entidad federativa.

6) Así, el siguiente paso fue ordenar la variable “edad decenal” por lo que por cada entidad se obtuvo 5 hojas. Es decir se desplazaron en 5 hojas distintas las entidades por categoría y por edad. Es decir se elaboraron 160 hojas por cada base de datos.

7) Se ordenaron las hojas resultantes por las categorías de sexo femenino y sexo masculino (es decir, cada fila), mediante filtros manuales en excel.

8) Después se utilizaron las hojas de la Base de datos de Egresos Hospitalarios para desplazar los datos ordenados y separados de la otra base de datos. Luego entonces se unió la cantidad de afecciones y de defunciones obtenidas con las tablas dinámicas en el paso 3). Lo anterior fue elaborado con una función en excel que condicionaba que si se tenía “sexo masculino” entonces se utilizaba la función “BUSCARV”

para obtener el valor de la afección en la matriz donde coincidía con el sexo de los hombres y sino se buscaba el valor de la cantidad de difuntos en la matriz de las mujeres.

9) Por último se eliminaron los casos cuyos egresos fueron nulos y defunciones nulas al mismo tiempo. Dicho proceso se realizó con un filtro manual.

Hasta aquí concluye la explicación de como se colapsaron ambas bases de datos. Las siguientes secciones se basan en el libro "An introduction to categorical Analysis"(Agresti, 2007)([1]). En ellas se muestran definiciones que sirvieron para realizar la operativización de los constructos.

3.4. Datos de respuesta categórica

Una variable categórica tiene una escala de medida consistente en un conjunto de categorías (en el caso de la variable DM que más adelante se define, es difunto a causa de DM o vivo afectado por DM).

3.4.1. Distinción entre variable explicativa y variable respuesta

Variables respuesta o dependientes

Para este estudio se tomó como variable respuesta la variable defunción a causa de Diabetes Mellitus. Esta variable consiste en el número de defunciones a causa de Diabetes Mellitus en todo el país durante el 2012.

Variables explicativas o variables independientes

Las variables explicativas son discretas, pues solo toman un conjunto finito de valores. Por ejemplo, la variable región socioeconómica está codificada en los valores desde el 1 hasta el 6. Así, todas las variables que aquí se utilizan son cualitativas. La mayoría de los análisis distingue entre variables respuesta (o dependientes) y variables explicativas o independientes.

Por consiguiente, los modelos de regresión describen como la media de una variable respuesta, tal como la defunción a causa de DM, cambia de acuerdo con los valores de

las variables explicativas. En este caso son sexo, edad decenal, región socioeconómica y afección principal.

3.5. Distinción entre escala nominal y ordinal

Las variables categóricas tienen dos tipos de escalas primarias. Las variables que tiene categorías sin un orden natural son llamadas nominales. Y en el caso de la variable DM que distingue entre difunto y vivo , puede notarse que no sigue un «orden natural», por tanto es **nominal**. Y con respecto a las variables explicativas sexo, afección principal y región socioeconómica, también son nominales. Nótese que la variable edad decenal es una variable que pasó a ser de intervalo a ordinal.

Distinción entre variables cualitativas y cuantitativas

Las variables nominales son cualitativas ya que la distinción entre las categorías difiere en cualidad y no en cantidad mientras que las variables de intervalo son cuantitativas, pues distinguen niveles que tienen diferencias en las cantidades de una característica de interés.

3.6. Operativización de los constructos

A continuación se muestran las tablas donde se muestra la forma cómo se codificaron las variables explicativas utilizadas para los modelos.

Región socioeconómica Esta variable se obtuvo a partir de la variable entidad de residencia Ent_resid la cual no tenía un orden natural en sus categorías, es decir es **nominal**. Dicha variable estaba compuesta por las 32 entidades de la República Mexicana. Cabe mencionarse que esta variable era común entre las bases de datos de las defunciones de Diabetes Mellitus en 2012 ($DEFUN12$) y de los afectados por Diabetes Mellitus en 2012. Luego, para poder realizar el análisis de los 32 estados, se dividieron en 7 regiones socioeconómicas creadas por INEGI, según indicadores relativos al bienestar como educación, ocupación, salud, vivienda y empleo. Las siete regiones se describen en la tabla 3.1.

Region socioeconomica	Entidades pertenecientes
RS1	Chiapas, Guerrero, Oaxaca.
RS2	Campeche, Hidalgo, Puebla, San Luis Potosí, Tabasco y Veracruz
RS3	Durango, Guanajuato, Micoacán y Tlaxcala
RS4	Colima, México, Morelos, Nayarit, Querétaro, Quintana Roo, Sinaloa y Yucatán.
RS5	Baja California, Baja California Sur, Chihuahua, Sonora y Tamaulipas.
RS6	Aguascalientes, Coahuila, Jalisco y Nuevo León.
CDMX	Ciudad de México.

Cuadro 3.1: Tabla de codificación de la variable regiones socioeconómicas

Sexo La variable sexo era una variable común entre las bases de datos de Defunciones y de Afecciones. Por esta razón fue escogida para éste estudio. También es una variable **nominal**, así como la variable de Región Socioeconómica. En la tabla 3.2 se muestran las codificaciones del sexo que la bases de datos tenían.

codificación	Sexo
1	masculino
2	femenino

Cuadro 3.2: Codificación de la variable Sexo

Edad decenal recodificada Variable la cual fue obtenida a partir de la variable edad, la cual estaba ordenada del 0 al 120. Dicha variable se categorizó para obtener resultados semejantes a un estudio ya realizado por el IMSS por Olaiz-Fernández et al., (2007) [16] acerca de Diabetes Mellitus. Además la recodificación fue con fines de facilitar el estudio. La tabla 3.3 muestra la recodificación.

Tipo de Complicación y Tipo de Diabetes Mellitus Las variables Tipo de Complicación y Tipo de Diabetes Mellitus fueron extraídas de las variables Afección principal y causa de defunción encontradas en las bases de datos de Egresos hospitalarios 2012 y Defunciones 2012. Vale la pena recordar que dichas variables eran la misma para la base de datos unida, que se formó con la intersección de los casos de ambas variables (ver sección3.3.1). Debe mencionarse que la variable Afección principal estaba codificada según

codificación	Edad Decenal
1	30a39
2	40a49
3	50a59
4	60a69
5	70a79

Cuadro 3.3: Codificación de la variable Edad decenal

la Organización mundial de Salud OMS. Por ejemplo si un caso agrupado decía E110 eso significaba que Tenía Diabetes mellitus tipo I y Tenía DM con la complicación de Coma. Por tanto la clave fue dividida en dos. Los primeros tres dígitos significan la enfermedad DM y su Tipo (Tipo I o Tipo II) y con ellos se formó la variable tipo de Diabetes Mellitus. El cuarto dígito sirvió para formar la variable Tipo de Complicación la cual se explica a continuación.

Tipo de Complicación Esta variable fue considerada crucial. Pues, según la teoría, la defunción depende del tipo de complicación. Por ello fue relevante que se encontrara en ambas bases de datos ya que su eliminación del estudio tiene un efecto estadístico importante, el cual podrá observarse en el capítulo 5 en los efectos del cociente de dispersión 5.4.4 que se aborda en el mismo capítulo. Esta variable es **nominal**, discreta y cualitativa. Se puede observar en el cuadro 3.4 que la variable tipo de complicación tiene 9 categorías. Sin embargo, se utilizarán para el modelado sólo 5 de estas: de la 0 a la 5. Lo anterior debido a que la categoría 6 y 7 son complicaciones múltiples y otras complicaciones especificadas, las cuales se mencionan poco en la teoría de defunción por DM. Y las categorías 8 y 9 no proveen información para el modelado explicativo, por lo que quedaron omitidas. Vale la pena recordar que en el modelado explicativo es poco deseable manejar valores perdidos o no especificados.

Tipo de Diabetes Mellitus Esta variable fue probada en el modelo por separado en lugar de Tipo de Complicación. Sin embargo al probar ambas variables la dispersión del modelo mejoró y también se hizo una prueba de cocientes de verosimilitudes (devianza) para observar si era o no significativa, como se verá en la sección 6.1.2. Dicha variable se codificó según el cuadro 3.5.

Después de haber mostrado cómo se codifican las variables, puede mostrarse si se

codificación	**Nombre de la complicación**
0	DM con coma
1	DM con cetoacidosis
2	DM con complicaciones renales
3	DM con complicaciones oftálmicas
4	DM con complicaciones neurológicas
5	DM con complicaciones circulatorias periféricas
6	DM con con otras complicaciones especificadas
7	DM con complicaciones múltiples
8	DM con complicaciones No especificadas
9	DM sin mención de complicación

Cuadro 3.4: Codificación de la variable tipo de complicaciones de la Diabetes Mellitus

codificación	**Tipo de Diabetes Mellitus**
E10	Tipo I
E11	Tipo II

Cuadro 3.5: Codificación de la variable Tipo de Diabetes Mellitus

relacionan con la variable respuesta. El siguiente capítulo trata de como seleccionar las variables que formaron parte del modelo.

Capítulo 4

Análisis exploratorio de los datos

Para ésta sección se retomó información del libro " An introduction to categorical data analysis" (Agresti, 2007) [1, p. 322], así como la página de internet llamada "Contingency Tables and the Chi Square Statistic Interpreting" (Albrecht, 2018) [42]. El propósito de éste capítulo es mostrar las tablas de contingencia que se elaboraron en el programa estadístico R y su respectiva interpretación con la prueba chi-cuadrada. A continuación se muestran los elementos teóricos requeridos para dicho análisis.

4.1. Tablas de contingencia

Una tabla de contingencia es una tabla que muestra la relación entre dos o más variables categóricas.

Según Agresti, [1, p. 78] para una muestra multinomial con probabilidades $\widehat{\pi}_{ij}$ en una tabla de contingencia IxJ, la hipótesis nula de independencia estadística es

$$H_0 = \pi_{ij} = \pi_{i+}\pi_{+j} \forall i, j \tag{4.1}$$

Donde

$$\pi_{i+} = \sum_j \pi_{ij}$$

y

$$\pi_{+j} = \sum_i \pi_{ij}$$

Esta prueba utiliza el estadístico χ^2 con el número de casos totales $\sum_i \sum_j n_{ij} = n$ y

$$\widehat{\mu}_{ij} = n\widehat{\pi}_{i+}\widehat{\pi}_{+j} = n_{i+}n_{+j}/n$$

pues

$$\widehat{\pi}_{i+} = n_{i+}/n \tag{4.2}$$

y

$$\widehat{\pi}_{+j} = n_{j+}/n \tag{4.3}$$

Supuestos para la prueba chi-cuadrada

1. Ambas variables, dependientes e independientes requieren ser categóricas. Y en este caso, en el capítulo anterior se mostró que lo son pues la mayoría de las variables son nominales y una es intervalar.

2. Se requiere un tamaño de la muestra adecuado. Generalmente la muestra debe ser de al menos 100 casos. En el caso de la base de datos que se utiliza para éste estudio se tienen 30,884 casos en total.

3. El número de respuestas en cada celda debe ser de al menos 5. Si no se puede utilizar la prueba exacta de Fisher u otras pruebas. En el caso de éstas tablas de contingencia, la cantidad mínima de casos que una celda contiene es de 8.

Por lo anterior esta prueba es adecuada para mostrar asociación. En la sección siguiente se muestran las tablas de contingencia entre cada variable seleccionada para formar parte del modelo. Para cada tabla, Difunto corresponde a la cantidad de defunciones a causa de DM mientras que no difunto corresponde a los Afectados por esta causa.

Antes de exponer las tablas de contingencia se plantean las hipótesis nula y alternativa:

H_0: La variable (sexo, edad decenal, región socioeconómica, tipo de complicación o tipo de DM) es independiente de la defunción a causa de DM del paciente.

H_a: La variable (sexo, edad decenal, región socioeconómica, tipo de complicación o tipo de DM) está asociada a la defunción a causa de DM del paciente.

Dichas hipótesis tienen un nivel de significación de α =0.05. Y el criterio de decisión es : H_0 se rechaza si $P - value \leq \alpha$ (Ji Cuadrado, Scribd, 2018)[43].

SEXO	**Difunto**	**No difuntos**
hombre	51 % (9156)	49 % (8758)
mujer	40 % (5177)	60 % (7793)

Cuadro 4.1: χ^2= 629.21, df = 1, $p - value$ <2.2e-16

Variable sexo contra la defunción a causa de DM en el año 2012

Interpretación En la tabla de contingencia que se muestra en el cuadro 4.1 puede observarse que más de la mitad de los hombres que enfermaron de DM, fallecieron por ésta causa y en cambio aproximadamente 10 % menos mujeres fallecieron por ésta causa. Por lo cual parece existir una mayor incidencia de defunciones a causa de DM en los hombres que en las mujeres.

Prueba chi-cuadrada Como la probabilidad asociada (es decir el p-value) al estadístico χ^2=629.21 es menor que 0.05, entonces se rechaza la hipótesis nula de independencia entre el sexo y la defunción a causa de diabetes Mellitus y no se rechaza que hay una fuerte relación entre la variable defunción a causa de Diabetes Mellitus y sexo.

Variable Región socioeconómica contra la defunción a causa de DM en el año 2012

Interpretación Se puede observar en el cuadro 4.2 que la región en la cual hay una menor incidencia de muertes a causa de DM es en la región 3 y en las demás regiones socioeconómicas las defunciones y afecciones fueron casi la mitad y la mitad del total de difuntos y afectados.

Prueba chi-cuadrada Como la probabilidad asociada al estadístico χ^2 =1231.4 es menor que 0.05, entonces no se rechaza la hipótesis alternativa de que hay relación entre la variable defunción a causa de DM y la variable región socioeconómica.

Región socioeconómica	Difunto	No difunto
Región1	46 % (1374)	54 % (1631)
Región2	42 % (2858)	58 % (3924)
Región3	37 % (1961)	63 % (3325)
Región4	58 % (3631)	42 % (2650)
Región5	54 % (1150)	46 % (985)
Región6	57 % (1210)	43 % (897)
Región7	41 % (2149)	59 % (3139)

Cuadro 4.2: Tabla de contingencia. Región socioeconómica contra supervivencia de enfermos de DM en el año 2012 (χ^2 = 1231.4,df = 6, $p - value$ <2.2e-16)

Tipo de Diabetes Mellitus	Difunto	No difunto
DM tipo I	18.8 % (150)	81.1 % (645)
DM tipo II	47.1 %(14,183)	52.8 (15,906)

Cuadro 4.3: Tabla de contingencia. Tipo de Diabetes Mellitus contra la defunción a causa de DM en el año 2012. $\chi^2 = 21471$, df = 1 , $p - value <$ 2.2e-16

Variable Tipo de Diabetes Mellitus contra la defunción a causa de DM en el año 2012

Interpretación Se puede apreciar en el cuadro 4.3 que la población adquirida para éste estudio tuvo una cantidad de casos de vivos afectados por DM mayor que los difuntos. Sin embargo el porcentaje de difuntos de diabetes mellitus tipo II fue más del doble, que el porcentaje de difuntos de diabéticos con el tipo I.

Prueba chi-cuadrada Dado que la probabilidad asociada al estadístico $\chi^2 = 21471$ es menor a 0.05, entonces se rechaza que hay relación entre la variable afección principal y la variable de defunciones a causa de DM.

Variable Afección principal contra la defunción a causa de DM en el año 2012

Interpretación En el cuadro 4.4 se observa que la complicación renal encabeza el mayor porcentaje de defunciones ya que poco más del 60 % del total de afectados y difuntos por DM fallecieron por esta causa, seguido la complicación cetoacidosis con poco

Afección principal	Difunto	No difunto
DM con coma	27 % (396)	73 % (1075)
DM con cetoacidosis	48 % (1065)	52 % (1173)
DM con complicaciones renales	61 % (12277)	39 % (7805)
DM con complicaciones oftalmicas	10 % (13)	90 % (112)
DM con complicaciones neurológicas	3 % (8)	97 % (247)
DM con complicaciones circulatorias periféricas	9 % (574)	91 % (6139)

Cuadro 4.4: Tabla de contingencia. Afección principal Mellitus contra la defunción a causa de DM en el año 2012. $\chi^2 = 21471, df = 9, p - value < 2{,}2e - 16$

menos del 50 % y la complicación de la DM que menos defunción causó fue la neurológica.

Prueba chi-cuadrada Dado que la probabilidad asociada al estadístico $\chi^2 = 21471$ es menor a 0.05, entonces indica que no se rechaza la hipótesis alternativa de que hay relación entre la variable afección principal y la variable de defunciones a causa de DM.

Variable Edad Decenal contra la defunción a causa de DM en el año 2012

Edad Decenal	Difunto	No difunto
30a39	16 % (200)	84 % (1058)
40a49	33 % (1491)	67 % (3050)
50a59	44 % (4323)	56 % (5408)
60a69	51 % (4666)	49 % (4549)
70a79	60 % (3653)	40 % (2486)

Cuadro 4.5: Tabla de contingencia. Edad Decenal contra la defunción a causa de DM en el año 2012. $\chi^2 = 3618{,}5, df = 4, p - value <$2.2e-16

Interpretación En este caso, en el cuadro 4.5 puede observarse que las defunciones van en incremento conforme se avanza en la edad y pasa de un 16 % en el grupo de edad de 30 a 39 años a más de 30 % en el grupo siguiente, lo cual corresponde a un crecimiento en defunciones del doble. En el siguiente grupo se tiene un incremento de aproximadamente 11 % y en los últimos grupos es decir los de 60 a 69 y de 70 a 79 hay un incremento

del 10 % con respecto al grupo anterior que le corresponde. Así pasa a ser un 15 % las defunciones al 60 % para el último grupo de 70 a 79. Por lo cual se ve una asociación entre defunciones y el incremento en la edad.

Prueba chi-cuadrada La probabilidad asociada al estadístico χ^2 =3618.5 es menor a 0.05 por lo que no se rechaza la hipótesis alternativa de que hay asociación entre la variables edad decenal y defunción a causa de DM.

4.2. Elección de variables

Según la teoría subyacente del capítulo 2 de esta tesis, se tienen los siguientes factores asociados a las defunciones a causa de diabetes Mellitus tipo I y tipo II:

- Edad
- Sexo
- Tipo de complicación
- Tipo de Diabetes Mellitus.
- Región socioeconómica.

Las variables anteriores se eligieron por la disponibilidad en las bases de datos y por relevancia en la parte teórica que describe las causas de defunción por diabetes Mellitus.

El capítulo que sigue trata sobre los métodos que se utilizaron para modelar la variable respuesta Defunción a causa de Diabetes Mellitus.

Capítulo 5

Elección de métodos

5.1. Introducción

En el capítulo anterior se pudo observar que la variable respuesta de supervivencia por DM tenía asociacion con cada una de las variables explicativas. Sin embargo, según la página web *STAT 504 Analysis of discrete data* (2018) [44], para analizar simultáneamente los efectos de las variables, deben usarse los Modelos lineales generalizados (MLG). Lo anterior es un ejemplo de una aplicacion de los MLG, también se usan para:

1. Incluir mezclas de variables categoricas y variables continuas.
2. Para trabajar con más de dos o tres variables categoricas.
3. Extender los modelos de regresion ordinaria al incluir un gran rango de variables respuesta cuya distribucion es no normal y al modelar funciónes de la media.

Esta herramienta estadística se enfoca en estimar los parámetros de los modelos. Dichos parámetros proveen medidas de fuerza de las asociaciones entre las variables. Además la forma estructural del modelo pueden mostrar las interacciones de las variables.

5.2. Modelos Lineales Generalizados (MLG)

El término modelo lineal generalizado se refiere a una clase de modelos extensamente popularizada por McCullagh y Nelder (1982,2a edicion 1989) [6]. En estos modelos, se

asume que la variable respuesta Y sigue una distribucion en la familia exponencial con media μ_i. Dicha media asume ser alguna función de una combinacion lineal $X'\beta$. Algunos llaman a esta función no lineal porque μ_i o $E(Y)$ es una función de covariables no lineal, pero McCullagh y Nelder los consideraron lineales, porque las covariables afectan la distribucion de y_i solo mediante la combinacion lineal $X'\beta$.

El primer paquete informático utilizado ampliamente para ajustar estos MLG fue llamado "GLIM", pues GLIM fue bien aceptado para abreviar el concepto de Generalized lineal model. Hoy en día, los modelos lineales generalizados son ajustados por muchas paqueterías como SAS proc Genmod y R function glm(). Notese que Agresti (2002)[1] utiliza GLM en lugar de GLIM como abreviacion y en esta tesis se utiliza MLG.

Hay tres componentes para formar cualquier MLG:

1. El *componente aleatorio*: identifica la variable respuesta Y, y su distribucion de probabilidad.

2. Un *componente sistemático*: especifica las variables explicativas utilizando una función llamada predictor lineal $X'\beta$.

3. Y una *función liga* g(): la cual especifica la función de $E(Y)$ que el modelo iguala al componente sistemático.

(*STAT 504 Analysis of discrete data*, 2018)

5.2.1. Componentes de los modelos lineales generalizados

Según Agresti (2002)[1, pgs. 116,117] el *componente aleatorio* de un Modelo lineal generalizado consiste en una variable respuesta Y con observaciones independientes $(y_1, \ldots, y_N)$ cuya distribucion forma parte de la familia exponencial natural. Esta familia tiene función de densidad de probabilidad o función de masa de la forma:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)] \tag{5.1}$$

.

La razon para restringir los MLG a la familia exponencial de distribuciones de Y es que el mismo algoritmo aplica para estimar sus parámetros.

Algunas distribuciones importantes como las de Poisson y binomial son casos especiales de esta familia de distribuciones. El valor del parámetro θ_i puede variar para $i = 1, \ldots, N$, dependiendo de los valores de las variables explicativas. El término $Q(\theta)$ es llamado el *parámetro natural*.

El segundo componente es el *componente sistemático* de un MLG el cual relaciona un vector $(\eta_1, \ldots, \eta_N)$ con las variables explicativas mediante un modelo lineal. Sea x_{ij} el valor del predictor tal que $j = 1, 2, \ldots, p$ para un sujeto u observacion i. Entonces

$$\eta_i = \sum_j \beta_j x_{ij} i = 1, \ldots, N \tag{5.2}$$

esta combinacion lineal de variables explicativas es llamada el *predictor lineal*. Comúnmente se toma para una de las j's, usualmente la primera, $x_{ij} = 1$ para todo i, y de esta forma se obtiene el intercepto (comúnmente denotado por α) en el modelo.

El tercer componente de un MLG es una *función liga* que conecta los componentes sistemático y aleatorio. Sea $\mu_i = E(y_i)$, $i = 1, \ldots, N$. El modelo liga μ_i a $\eta_i = g(\mu_i)$, donde la función liga g es una función monotona y diferenciable. Entonces, g enlaza $E(Y_i)$ a las variables explicativas mediante la formula:

$$g(\mu_i) = \sum_j \beta_j x_{ij}, i = 1, \ldots, N. \tag{5.3}$$

La *función liga* $g(\mu) = \mu$, se llama liga identidad, tiene $\eta_i = \mu_i$. La función liga identidad especifica un modelo lineal para la media de sí misma. Esto es la función liga para una regresion ordinaria con Y respuesta cuya distribucion es normal. La función liga que transforma la media a un parámetro natural es llamado la liga canonica. Para esto, $g(\mu_i) = Q(\phi_i)$, y $Q(\phi_i) = \sum_j \beta_j x_{ij}$.

También es de ayuda extender la notacion para un MLG de tal forma que se puedan manejar distribuciones que tengan un segundo parámetro. El componente aleatorio de un MLG especifica que las N observaciones $(y_1, \ldots, y_N)$ sobre Y son independientes con función masa o densidad para y_i de la forma:

$$f(y_i, \theta_i, \phi) = exp[\frac{y_i - b(\theta_i)}{a(\phi)} + c((y_i), \phi)] \tag{5.4}$$

Donde, ϕ es el llamado parámetro de dispersion, y a, b, c son funciónes que definen varios miembros de la familia.

Cabe añadir que cuando ϕ es desconocida, se simplifica a la forma 5.1.

Las distribuciones de la familia exponencial tienen medias y varianzas $E(Y) = \mu = b'(\theta)$, $var(Y) = b''(\theta)a(\phi)$ respectivamente.

5.2.2. Supuestos de los MLG

- Los datos $y_1, y_2 \dots y_n$ son observaciones independientes con distribucion en la familia exponencial.

- La variable dependiente y no necesita ser distribuída normal, pero asume típicamente una distribucion de la familia exponencial (es decir, binomial, Poisson, multinomial, normal...).

- Los MLG no asumen una relacion lineal entre la variable dependiente y las variables independientes, pero asume una relacion lineal entre la respuesta transformada en términos de la función liga y las variables explicativas.

- La homogeneidad de la varianza no necesita ser satisfecha. De hecho, en muchos casos no es posible dar la estructura del modelo y la sobredispersion (cuando la varianza observada es mayor que la que el modelo adopta) podría estar presente.

- Los errores necesitan ser independientes pero no normalmente distribuídos.

- Utiliza la estimacion de máxima verosimilitud más que mínimos cuadrados ordinarios para estimar los parámetros, por lo que depende de las aproximaciones de muestras grandes.

- Las medidas de bondad de ajuste dependen de las muestras suficientemente grandes. Donde una regla heurística [1] es que no más de 20 % de los conteos de las celdas esperadas sean menores a 5.

[1] f. En algunas ciencias, manera de buscar la solución de un problema mediante métodos no rigurosos, como por tanteo, reglas empíricas, etc. (Dle.rae.es, 2018) [45]

5.2.3. Ventajas y desventajas de la formulacion de los MLG

Los MLG proveen una teoría unificada de modelacion que abarcan los más importantes modelos para variables continuas y discretas. Los modelos estudiados en este texto son MLG con componente aleatorio Poisson o binomial. La razon para restringir los MLG a la familia de distribucion exponencial para Y es que el mismo algoritmo se aplica para esta familia entera, para cualquier elección de la función liga. Adicionalmente tiene ventajas como:

- No se requiere transformar la variable respuesta Y para tener una distribucion normal.
- La elección de la liga está separada de la elección del componente aleatorio. Por lo que se tiene mayor flexibilidad en el modelado.
- Si la liga produce efectos aditivos, entonces no se necesita una varianza constante.
- Los modelos son estimados mediante máxima verosimilitud; por ello los estimadores tienen propiedades optimas.
- Todas las herramientas de inferencia y chequeo del modelado como las pruebas de Wald, razon de verosimilitud, devianza, residuales, intervalos de confianza y sobredispersion aplican para todos los MLG's.
- Hay un solo procedimiento en los diversos paquetes informáticos para capturar todos los modelos, por ejemplo: PROC GENMOD en SAS o glm() en R, etc. con opciones de variar los tres componentes.

Aunque también hay algunas limitaciones para los MLG como:

- Se puede tener un solo predictor lineal en el componente sistemático.
- Las respuestas deben ser independientes.

STAT 504 Analysis of discrete data (2018)

En las siguientes secciones se muestran los dos tipos de distribuciones que la variable de defunciónes por DM puede asumir por tratarse de conteos.

5.3. Distribuciones de la Variable respuesta DM

Las siguientes subsecciones se basan en las notas dadas en el sitio web Data.princeton.edu por Germán Rodríguez, (2018) [46, caps. 3 y 4].

5.3.1. Distribucion Binomial

Se considera primero el caso donde la respuesta y_i es binomial se asume solo dos valores que por conveniencia se codifican como cero o uno. Por ejemplo, se define

$$y_i := \begin{cases} 1 & \text{si el i-ésimo paciente fallecio por DM ,} \\ 0 & \text{en otro caso} \end{cases}$$

Se ve a y_i como una realizacion de una variable aleatoria Y, que puede tomar valores uno o cero con probabilidades π_i y 1-π_i, respectivamente. La distribucion de Y es llamada Bernoulli con parámetro π_i, y puede ser escrito de forma compacta como:

$$Pr(Y = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{5.5}$$

,

para $y_i = 0, 1$. Note que si $y_i = 1$ se obtiene π_i, y si y_i=0 se obtiene $1 - \pi_i$. Es fácil verificar por calculo directo que el valor esperado y la varianza de Y_i son:

$$E(Y) = \mu_i = \pi_i, Var(Y) = {\sigma_i}^2 = \pi_i(1 - \pi_i) \tag{5.6}$$

Notese que la media y la varianza dependen de la probabilidad subyacente π_i. Cualquier factor que afecte la probabilidad va a alterar no solo la media sino también la varianza de las observaciones. Esto sugiere que un modelo lineal que permite a las variables explicativas en este caso afectar la media, en el cual se asume que la varianza constante podría no ser adecuada para el análisis de datos binarios. suponga que las unidades bajo estudio pueden ser clasificadas de acuerdo a los factores de interés en K grupos tales que la forma que todos los individuos en un grupo tienen valores idénticos de todas las covariables.

Por ejemplo, sea

y_i= número de defunciónes a causa de DM en la entidad i, en el año 2012.

Se observa a y_i como una realizacion de la variable aleatoria y_i que toma los valores $0, 1, \ldots, n_i$. Si las n_i observaciones en cada grupo son independientes y todas estas tiene la misma probabilidad π_i de tener el atributo de interés, entonces la distribucion de y_i es binomial con parámetros π_i y n_i, lo cual se escribe $y_i \sim B(n_i, \pi_i)$

La función de distribucion de masa de y_i es dada por

$$Pr(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \tag{5.7}$$

Para $y_i = 0$, $i = 1, \ldots, n_i$. Aquí $\pi_i^{y_i}(1 - \pi_i)^{n_i - y_i}$ es la probabilidad de obtener y_i éxitos y $n_i - y_i$ fallas en algún orden específico, número de formas de obtener y_i éxitos en n_i ensayos. La media y varianza de y_i puede ser mostrado como:

$$E(y_i) = \mu_i = n_i \pi_i, Var(y_i) = \sigma^2 = n_i \pi_i (1 - \pi_i) \tag{5.8}$$

El camino más fácil de obtener este resultado es como sigue: sea y_{ij} la variable indicadora que toma valores uno o cero si la j-ésima unidad en el grupo i es un éxito o falla, respectivamente. Note que y_{ij} es una variable aleatoria Bernoulli con media y varianza dadas en la ecuación 5.6. Se puede escribir el número de éxitos y_i en el grupo i como la suma de variables indicadoras individuales, entonces $y_i = \sum_j y_{ij}$. La media de y_i es entonces la suma de las medias individuales, y por independencia, su varianza es la suma de varianzas individuales, destacando el resultado en la ecuación 5.8 depende de la probabilidad subyacente π_i. Desde un punto de vista práctico es importante notar que si las variables explicativas son factores discretos y las salidas son independientes, se pueden usar la distribucion Bernoulli para los datos individuales cero-uno o la distribucion binomial para datos agrupados consistentes en conteos de éxitos en cada grupo. Las dos aproximaciones son equivalentes, en el sentido de que estas llevan exactamente a la misma función de verosimilitud y por tanto a las mismas estimaciones y errores estándares.

La distribucion Poisson puede ser derivada como una forma del límite de la distribucion binomial si se considera la distribucion del número de éxitos en un muy grande número de ensayos Bernoulli con una probabilidad pequeña de éxitos en cada intento.

Específicamente, si $X \sim B(n, \pi)$ entonces la distribucion de Y como $n \to \infty$ y $\pi \to 0$

con $\mu = n\pi$ se aproxima a una distribucion Poisson con media μ. Entonces la distribucion Poisson proveé una aproximacion a la distribucion binomial para el análisis de eventos raros donde π es pequeño y n es grande.

5.3.2. Distribucion Poisson

Una derivacion alternativa de la distribucion Poisson es en términos de un proceso estocástico que describe algo informalmente como sigue:

Debe recordarse que para la media y la varianza, cualquier factor que afecte uno podría también afectar al otro.

Por lo anterior, el supuesto usual de homocedasticidad podría no ser apropiado para los datos Poisson.

La probabilidad de distribucion del número de ocurrencias del evento en un intervalo de tiempo fijo es Poisson con media $\mu = \lambda t$ donde λ es la tasa de ocurrencia del evento por unidad de tiempo y t es la longitud del intervalo de tiempo.

La motivacion más importante la distribucion Poisson para el punto de vista de la estimacion estadística, sin embargo, es la relacion entre la media y la varianza.

Se va a remarcar este punto cuando se discuta nuestro ejemplo, donde los supuestos de limitar un Proceso binomial y Poisson es particularmente realista, pero el modelo Poisson captura muy bien el hecho de que, como es común en el caso de datos de conteo, la varianza tiende a incrementar con la media.

La función de distribucion Poisson que se escribe como $y_i \sim P(\mu_i)$, cuya media y varianza son μ_i está dada por

$$Pr(Y = y) = \frac{\exp^{-\mu} \mu^y}{y!} \tag{5.9}$$

Una propiedad útil de la distribucion Poisson es que la suma de variables aleatorias independientes Poisson es también Poisson. Específicamentente, si y_1 y y_2 son independientes $y_1 + y_2 \sim P(\mu_1 + \mu_2)$ este resultado se generaliza a la suma de más de dos observaciones Poisson lo cual se utiliza en este estudio pues se suman los casos por entidad y por grupo según grupo de edad, sexo y tipo de complicacion. Por lo que si cada caso

$y_i \sim P(\mu_i)$ entonces la $\sum y_i \sim P(\sum_i \mu_i)$.

Una importante consecuencia práctica de este resultado es que podemos analizar datos individuales. Especificamente, suponga que tenemos un grupo de individuos n, con valores covariables idénticos. Sea y_i, la variable que denota el número de eventos esperados por la unidad j-ésima en el i-esimo grupo, y sea y_i la variable que denota el número total de eventos en el grupo i, entonces bajo los supuestos usuales de independencia, si $y_{ij} \sim P(\mu_i)$

Para $j = 1, 2, \ldots, n_i$, entonces

$y_{ij} \sim P(n_i\mu_i)$.

En palabras, si los conteos individuales y_{ij} son Poisson con media μ_i el grupo total y_i es Poisson con media $n_i\mu_i$. En términos de estimacion, se obtiene exactamente la misma función de verosimilitud si se trabaja con los conteos individuales y_{ij} o el grupo de conteos y_i.

5.4. Modelos Lineales Generalizados para conteos y tasas

El más conocido MLG para datos de conteo asume un distribucion Poisson para Y. Dicho modelo es útil también para tasas, de lo cual se hablará en las secciones siguientes. Para el ajuste hecho en este trabajo se requieren los modelos logísticos (que modelan conteos de éxitos) y los modelos loglineales de Poisson que modelan tasas.

5.4.1. Modelo de Regresion Binomial

Esta subsección se basa en el libro "Extending the linear model with R: Generalized linear, Mixed effects and nonparametric regression Models"(Faraway J.Julian, 2016) [2, caps. 2 y 3].

Suponga que la variable respuesta y_i para $i = 1, \ldots, n_i$ es distribuida binomialmente $B(n_i, p_i)$.

Se asume que y_i son independientes. Los ensayos individuales que componen la respuesta y_i son todos sujetos a las mismas variables explicativas q $(x_{i1}, \ldots, x_{iq})$. El grupo de ensayos es conocido como una clase covariable. Se necesita un modelo que describe la relacion de $x_i, \ldots, x_q$ con p.

Siguiendo el modelo lineal aproximado, se construye un predictor lineal:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}$$

Como en el predictor lineal se pueden acomodar predictores cuantitativos y cualitativos con el uso de variables dummy y también permite las transformaciones y combinaciones de predictores originales, es muy flexible y de fácil interpretacion.

Para la función liga, elegir $\eta_i = p_i = \pi_i$ no es apropiado porque se requiere que $0 \leq \pi_i = p_i \leq 1$. En su lugar hay que usar la función liga g tal que $\eta_i = g(p_i) = g(\pi_i)$, para esto se necesita una función g monotona y tal que $0 \leq g^{-1}(\eta) \leq 1$ para cualquier η.

Hay tres elecciones comunes:

1. Logit: $\eta = log(p/(1-p))$
2. Probit: $\eta = \phi^{-1}(p)$ donde ϕ^{-1} es la distribucion inversa normal.
3. Log-log complementario: $\eta = log(-log(1-p))$

La idea de la función liga es también una de las ideas centrales de los MLG. Esta es usada para enlazar el predictor lineal a la media de la respuesta en clases de modelos amplios. En el caso del trabajo presente se utiliza la función liga: Logit, formando el modelo logit: $Log(p/(1-p)) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_q X_{iq}$

Otros elementos requerido para realizar el ajuste para datos de respuesta binomial son y y n. Se verá en el capítulo 6 que al ajustar en R, una forma de hacerlo es formando una matriz de dos columnas, con la primer columna representando el número de "éxitos"y y la segunda el número de "fallas"$n - y$. Y para los datos aquí ajustados la primer columna es defunción a causa de DM y la segunda es afeccion de DM. Y n correspondería a los casos totales de afecciones más defunciónes por DM.

La transformacion Logit

El siguiente paso en definir un modelo para nuestros datos concierne a la función liga. Como se dijo en la seccion anterior, se desea tener las probabilidades π_i que dependen de un vector de covariables observadas x_i. La idea más simple podría ser que π_i sean una función lineal de covariables, es decir,

$$\pi_i = x_i' \beta \tag{5.10}$$

donde β es un vector de coeficientes de regresion. El modelo 5.10 es a veces llamado el modelo lineal de probabilidad. Este modelo es estimado comúnmente para datos individuales utilizando mínimos cuadrados ordinarios (OLS).

Un problema con este modelo es que la probabilidad π_i sobre el vector del lado izquierdo tiene que estar entre cero y uno, pero el predictor lineal $x_i'\beta$ sobre el lado derecho puede tomar cualquier valor real, entonces no hay garantía de que los valores predichos puedan estar en el rango correcto, excepto cuando se imponen restricciones complejas sobre los coeficientes.

Una solucion simple a este problema es transformar la probabilidad para remover las restricciones del rango, y modelar la transformacion como una función lineal de las covariables. Se hace lo anterior en dos pasos. El primero es pasar de la probabilidad π_i a los momios (*odds*).

$$odds_i = \frac{\pi_i}{1 - \pi_i}$$

El momio asociado a cierto suceso es: "La razon entre la probabilidad de que ocurra tal suceso y la probabilidad de que no ocurra".

Los momios funcionan de la forma siguiente: si la probabilidad de un evento es un medio, los momios (*odds*) son uno a uno. Si la probabilidad es 1/3, los odds son uno a dos.

El segundo paso es tener los logaritmos, calculando el logit o los log-odds.

$$\eta_i = logit(\pi_i) = log\frac{\pi_i}{1 - \pi_i} \tag{5.11}$$

que tiene el efecto de remover la restriccion del "piso". Para ver este punto destáquese que cuando la probabilidad va hacia cero, los logit se aproximan a $-\infty$.

En el otro extremo, a medida que la probabilidad se aproxima a uno los odds se aproximan a $+\infty$ y el logit también.

Por lo tanto, los logits "escanean" probabilidades del rango $(0, 1)$ a la línea recta real. Note que si la probabilidad es $1/2$, los odds son entonces uno y el logit es cero.

Los logits negativos representan las probabilidades debajo de un medio. Y los logits positivos corresponden a probabilidades arriba de la mitad.

Los logits también pueden definirse en términos de la media binomial $\mu_i = n_i\pi_i$ como el logaritmo de la tasa esperada de éxitos μ_i sobre las fallas esperadas $n_i - \mu_i$. El resultado es exactamente el mismo porque el denominador binomial n_i se cancela cuando se calculan los momios.

La transformacion logit es biyectiva. La inversa de la transformacion se llama antilogit y nos permite regresar de los logits a las probabilidades.

Resolviendo de la ecuación 5.11

$$\pi_i = logit^{-1}(\eta_i) = \frac{\exp \eta_i}{1+\exp \eta_i}$$

(Rodríguez, 2018) [46].

5.4.2. El modelo de regresion logística

supongase que se tienen k observaciones independientes $y_1, \ldots, y_k$ y que las i-ésimas observaciones pueden ser tratadas como una realizacion de una variable aleatoria y_i. Se asume que y_i tiene una distribucion binomial

$$y_i \sim B(n_i, \pi_i) \tag{5.12}$$

Con probabilidad π_i. Con datos individuales $n_i = 1 \; \forall i$. Lo anterior define la estructura estocástica del modelo (o componente aleatorio). supongase que el logit de la probabilidad subyacente π_i es una función lineal de los predictores

$$logit(\pi_i) = x_i'\beta \tag{5.13}$$

donde x_i es un vector de covariables y β es un vector de coeficientes de regresion. Esto es definido como estructura sistemática o componente sistemático del modelo.

El modelo definido en las ecuaciones 5.12 y 5.13 es un modelo lineal generalizado con respuesta binomial y liga logit. Note que es más natural considerar la distribucion de la respuesta y_i que la distribucion del error implicado $y_i - \mu_i$.

Los coeficientes de regresion β pueden ser interpretados teniendo en mente que el lado izquierdo es un logit más que una media.

Por tanto β_j representa el cambio en el logit de la probabilidad asociada con una unidad de cambio en el j-ésimo predictor manteniendo todos los otros predictores constantes.

Al exponenciar la ecuación 5.13 se pueden encontrar los momios para la i-ésima unidad dada por:

$$\frac{\pi_i}{1-\pi_i} = \exp x_i'\beta \tag{5.14}$$

La expresion 5.14 define un modelo multiplicativo para los momios. Por ejemplo si se cambia el j-ésimo predictor por una uidad mientras mantenemos todas las demás variables constantes, se multiplicaran los odds por $exp\beta_j$.

Para ver este punto suponga que el predictor lineal es $x_i'\beta$ y se incrementa x_i por uno, se obtiene $x_i'\beta + \beta_j$. Aplicando la función exponencial a la ecuación anterior se obtiene $\exp x_i'\beta$ veces $\exp \beta_j$. Por tanto, el coeficiente exponenciado $\exp \beta_j$ representa un *odds ratio*, es decir una razon de momios, también son llamados razón de posibilidades según TAPIA-GRANADOS, (1997) [17] . Traduciendo los resultados en efectos multiplicativos sobre los momios, o razon de momios, es comúnmente de ayuda, porque se pueden tratar con una escala más familiar mientras que se retiene un modelo relativamente más simple.

Resolviendo para la probabilidad π_i en el modelo logit en la ecuación 5.13 se obtiene un modelo más complicado

$$\pi_i = \frac{\exp x_i'\beta}{1+\exp x_i'\beta}$$

mientras que del lado izquierdo es la escala de probabilidad familiar, el lado derecho es una función no lineal de predictores,y no es un camino simple expresar el efecto de la probabilidad si un predictor incrementa por una unidad mientras que se mantienen las otras variables constantes.

Estimacion de máxima verosimilitud para el modelo logístico

La función de verosimilitud de n observaciones binomiales independientes es un producto de las densidades dadas por la función de distribucion de probabilidad binomial dada en la sección 5.3.1 en la ecuación 5.7. Obteniendo el logaritmo se tiene que, excepto por una constante involucrada con los términos combinatorios, la función de log-verosimilitud

es

$$logL(\beta) = \sum y_i log(\pi_i + (n_i - y_i) log(1 - \pi_i) \tag{5.15}$$

donde π_i depende de las covariables x_i y el vector de p parámetros β a través de la transformacion logit de la ecuación 5.13.

Ahora deben calcularse las segundas derivadas esperadas para obtener el promedio y la matriz de informacion y el desarrollo del procedimiento Fisher scoring para maximizar la log-verosimilitud. El procedimiento equivale a los mínimos cuadrados iterativamente reponderados (IRLS). Dado un estimador actual estimado $\widehat{\beta}$ de los parámetros, se calcula el predictor lineal $\widehat{\eta} = x_i'\widehat{\beta}$ y los valores ajustados $\widehat{\mu} = logit^{-1}(\eta)$.

Bondad de Ajuste para el modelo logístico

Suponga que se tiene un modelo ajustado y se busca evaluar qué tan bien ajusta a los datos. Para ello se requiere una medida de discrepancia entre los valores ajustados y los observados. Dicha medida es la devianza, dada por:

$$D = 2 \sum y_i log(\frac{y_i}{\widehat{\mu}_i} + (n_i - y_i) log(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}) \tag{5.16}$$

,

donde y_i es el valor observado y $\widehat{\mu}_i$ es el valor ajustado para cada i-ésima observacion. Note que este estadístico es dos veces la suma de "los valores observados por el logaritmo de los valores esperados sobre los valores esperados", donde la suma es de las fallas y éxitos (es decir se comparan y_i y $n_i - y_i$ con sus valores esperados).

En un ajuste perfecto la razon de los valores observados sobre los esperados es uno. Y su logaritmo es cero, entonces la devianza es cero.

Con los datos agrupados (como en el caso de esta tesis) la distribucion del estadístico devianza como el grupo de tamaños $n_i \to \infty$ para todo i, converge a una distribucion chi-cuadrada con $n - p$ grados de libertad, donde "n" es el número de grupos y "p" es el número de parámetros en el modelo, incluyendo la constante.

Así para grupos razonablemente grandes, la devianza proveé una prueba de bondad de ajuste para el modelo. Con datos individuales la distribucion de la devianza, converge a la distribucion chi-cuadrada (o a otra conocida) , y no puede ser usada como una prueba

de bondad de ajuste. Sin embargo puede considerarse otras herramientas de diagnostico para datos individuales. Una medida alternativa de bondad de ajuste es el estadístico chi-cuadrado de Pearson, que para datos binomiales puede ser escrito como

$$X_p^2 = \sum \frac{n_i(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i(n_i - \widehat{\mu}_i)}$$

Note que cada término en la suma es la diferencia cuadrada entre los valores observados y los ajustados y_i y $\widehat{\mu}_i$ respectivamente, divididos entre la varianza la cual es $\frac{\mu_i(n_i-\mu_i))}{n_i}$, estimada utilizando $\widehat{\mu}_i$ para μ_i. Con los estadísticos agrupados Pearson, se tiene para muestras grandes aproximadamente una distribucion chi-cuadrada con $n-p$ grados de libertad, y es asintoticamente equivalente a la devianza o al cociente verosímil del estadístico chi-cuadrado. El estadístico no puede ser usado como una prueba de bondad de ajuste con datos individuales, pero provee una base para calcular residuales que se verán posteriormente.

5.4.3. Modelos Loglineales de Poisson para Datos de conteos

Suponga que tenemos una muestra de n observaciones $y_1, y_2, \ldots y_n$ las cuales pueden ser tratadas como realizaciones de variables aleatorias independientes poisson, con $y_i \sim P(\mu_i)$, y suponga que se busca obtener la media μ_i (y por tanto la varianza) depende de un vector de variables explicativas x_i. Se obtiene un modelo de la forma $\mu_i = x_i'\beta$, pero este modelo tiene una desventaja de que el predictor lineal de la derecha puede asumir cualquier valor real, mientras que la media Poisson del lado izquierdo, que representa un conteo esperado, tiene que ser no negativo. La solucion es el logaritmo de la media usada en lugar de un modelo lineal. Por tanto, se obtienen los logs, calculando $\eta_i = log(\mu_i)$ y se asume que las medias transformadas siguen un modelo lineal $\eta_i = x_i'\beta$. Por tanto, se considera un modelo lineal generalizado con una liga log. Combinando estos dos pasos en uno, se puede obtener lo que a continuacion se escribe:

$$log(\mu_i) = x_i'\beta \tag{5.17}$$

En este modelo de regresion el coeficiente β_j representa el cambio esperado en el log de la media por unidad de cambio en el predictor x_j. En otras palabras, el incremento de x_j por una unidad está asociado con un incremento de β_j en el log de la media. Aplicando

la exponencia a la ecuación 5.17 se obtiene un modelo multiplicativo de la media misma:

$$\mu_i = \exp x_i'\beta \tag{5.18}$$

En el modelo 5.18 un coeficiente de regresion exponenciado $\exp \beta_j$ representa un efecto multipicativo en el j-ésimo predictor sobre la media. El incremento de x_j por unidad multiplica la media por un factor $\exp \beta_j$. Otra ventaja de usar el enlace log viene de la observacion empírica de que con datos de conteo los efectos de los predictores son a menudo multiplicativos en lugar de aditivos. Lo anterior significa que se observa típicamente efectos pequeños para recuentos pequeños y efectos grandes para recuentos grandes. Si de hecho, el efecto es proporcional al recuento, trabajar en la escala típicamente logarítmica conduce a un modelo más sencillo.

Regresion poisson para tasas

Según Agresti (2002)[1, secs. 4.3.5 y 9.7], cuando las salidas ocurren sobre el tiempo, el espacio u otro indice de tamaño, es más relevante modelar su tasa de ocurrencia más que su número de columna.

Cuando el conteo de la respuesta n_i tiene un índice igual a t_i, la tasa muestral es n_i/t_i y su valor esperado es μ_i/t_i. Con una variable explicativa x, un modelo loglineal para valores esperados tiene la forma

$$log(\mu_i/t_i) = \alpha + \beta x_i \tag{5.19}$$

Este modelo tiene una representacion equivalente:

$$log(\mu_i) - log(t_i) = \alpha + \beta x_i$$

El término ajustado $log(t_i)$ es el logaritmo de la liga de la media y se llama offset. El ajuste corresponde a usar $log(t_i)$ como un predictor sobre la mano derecha y forzando su coeficiente a ser igual a 1. Para el modelo 5.4.3, la respuesta esperada de conteos satisface

$$\mu_i = t_i exp(\alpha + \beta x_i)$$

La media es proporcional al índice, con constante proporcional que depende del valor de

x. La liga identidad es también útil. El modelo es entonces

$$\mu_i/t_i = \alpha + \beta x_i$$

o

$$\mu_i = \alpha t_i + \beta x_i t_i$$

Esto no requiere un offset. Esto corresponde a un MLG Poisson que utiliza la liga identidad con t_i y $x_i t_i$ como variables explicativas y sin intercepto. También provee efectos predictivos aditivos, más que multiplicativos. Y es menos útil con algunos predictores, como al ajustar procesos, ya que podría fallar si los conteos fueran negativos.

Para este estudio el conteo de la respuesta n_i = *la cantidad de difuntos a causa de DM*, tiene un índice igual a t_i = *la cantidad de difuntos por DM+la cantidad de afectados por DM* , la tasa muestral es n_i/t_i = *la cantidad de afectados por DM/(la cantidad de difuntos por DM+la cantidad de afectados por DM*).

Por lo tanto el modelo de tasas que se emplea es:

$$log(n_i/t_i) = \alpha + \beta x_i$$

que es equivalente a

$$log(n_i) = log(t_i) + \alpha + \beta x_i$$

Por lo tanto el offset es $log(t_i)$ es decir el logaritmo de la suma de la cantidad de difuntos por DM más la cantidad de afectados por DM.

Devianza de un MLG Poisson o Binomial

Según A. Agresti [1, sec. 4.1.5] devianza de un MLG Poisson o Binomial se define como

$$-2[L(\hat{\mu}; y)] - L[y; y]$$

.

Este es el estadístico de razon de verosimilitud para probar la hipotesis nula de que el modelo se mantiene contra la alternativa general. (Es decir el modelo saturado).

Para algunos MLG's Poisson y Binomial el número de observaciones "N" se mantiene fijo, mientras que los conteos individuales aumentan de tamaño. Entonces la devianza tiene una distribucion nula asintotica Chi-cuadrada.

Los grados de libertad (d.f. ó g.l.) $g.l. = N - p$ donde "p" es el número de parámetros en los modelos saturados e insaturados. La devianza por lo tanto provee una prueba de ajuste del modelo.

Estimacion de máxima verosimilitud para el modelo loglineal de Poisson

Esta subsección y la siguiente se extrajeron de la página web Data.princeton.edu. (Rodríguez, 2018)[46, secs. 4.2.1 y 4.2.2]

La función de versomilitud para n observaciones independientes Poisson es el producto de probabilidades dadas por la ecuación 5.9. Obteniendo los logaritmos e ignorando la constante involucrada con el logaritmo de $y!$ se tiene que la función de log-verosimilitud es:

$$logL(\beta) = \sum y_i log(\mu_i) - \mu_i \tag{5.20}$$

donde μ_i depende de las covariables x_i y un vector de p parámetros β mediante la liga log de la ecuación 5.18.

Tomando las segundas derivadas de la función de verosimilitud con respecto a los elementos de β e igualándolas a cero se puede demostrar que la máxima verosimilitud en los modelos log-lineares de Poisson satisfacen las ecuaciones estimadas:

$$X'y = X'\widehat{\mu} \tag{5.21}$$

Aquí, X es la matriz del modelo, con una fila para cada observacion y una columna para cada predictor, incluyendo la constante (si hay alguna), "y" es el vector respuesta y $\widehat{\mu}$ es un vector de valores ajustados, calculado de los estimadores máximo verosímiles $\widehat{\beta}$ al exponenciar el predictor lineal $\eta = X'\widehat{\beta}$.

Tal ecuación estimada surge no solo en los modelos log-lineares de Poisson, sino más generalmente en cualquier modelo lineal generalizado con liga canonica, incluyendo los modelos lineales para datos normales y modelos de regresion logísticos para conteos binomiales.

Bondad de ajuste para modelos loglineales de Poisson

Una medida de discrepancia entre los valores y ajustados es la devianza. Para las respuestas Poisson la devianza toma la forma:

$$D = 2\sum(y_i \log(y_i/\widehat{\mu}_i) + (n_i - y_i)log(\frac{n_i - y_i}{n_i - \hat{\mu}_i}) \tag{5.22}$$

El primer término es idéntico a la devianza binomial, representando "dos veces la suma de los valores observados por el logaritmo de los valores sobreajustados". El segundo término, una suma de las diferencias entre los valores observados y ajustados, es comúnmente cero porque los estimadores máximo verosímiles en los modelos Poisson tienen la propiedad de reproducir los totales marginales.

Para muestras grandes la distribucion de la devianza es aproximadamente una chi-cuadrada con $n - p$ grados de libertad, donde n es el número de observaciones y p el de parámetros. Así la devianza puede ser usada directamente para probar bondad de ajuste para el modelo.

Una medida alternativa para bondad de ajuste es el estadístico chi-cuadrado de Pearson definido como

$$X_p^2 = \sum \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

El numerador es la diferencia cuadrada entre los valores observados y ajustados y el denominador es la varianza del valor observado. El estadístico de Pearson tiene la misma forma para datos Poisson y binomiales. En muestras grandes el estadístico de Pearson es aproximadamente chi-cuadrado con $n - p$ grados de libertad. Una ventaja de la devianza sobre el chi-cuadrado de pearson es que puede ser usado para comparar modelos anidados.

5.4.4. Sobredispersion o infradispersion para MLG de Poisson

Suponga que la respuesta Poisson tiene una tasa λ que es por sí misma una variable aleatoria. La tendencia de fallo de una máquina puede variar por unidad en unidad a pesar de que se modela con un mismo modelo. Es más adecuado en este caso modelar con

λ distribuida gamma y $E(\lambda) = \mu Y$ y $var(\lambda) = \mu/\phi$. Ahora Y se distribuye binomial negativa con media $E(Y) = \mu$. La media es la misma que la Poisson, pero la varianza $var(Y) = \frac{\mu(1+\phi)}{\phi}$ la cual es distinta que μ. En este caso puede ocurrir sobredispersion.

Si se conoce el mecanísmo específico como en el ejemplo de arriba, se podría modelar la respuesta como una binomial negativa u otra distribucion más flexible. Si el mecanismo no se conoce, se puede introducir el parámetro de dispersion ϕ tal que $var(Y) = \phi E(Y) = \phi\mu$ Si $\phi = 1$, se trata del caso regular de la regresion Poisson, mientras que si $\phi > 1$ es sobredispersion y $\phi < 1$ es infradispersion. El parámetro de dispersion puede ser estimado utilizando:

$$\widehat{\phi} = \frac{X^2}{n-p} = \frac{\sum_i (y_i - \widehat{\mu}_i)^2/\widehat{\mu}_i}{n-p} \tag{5.23}$$

Este parámetro será utilizado en el siguiente capítulo, estimado en R dividiendo el estadístico de la devianza residual entre sus grados de libertad. (Faraway J.Julian,2016)[2, sec.3.1]

5.4.5. Bondad de ajuste para modelos lineales Generalizados

Las siguientes dos subsecciones son traducciones del libro "Extending the linear model with R: Generalized linear, Mixed effects and nonparametric regression Models" [2, cap. 6]. Para un MLG en particular con observaciones $y = (y_1, ..., y_N)$, sea $L(\mu; y)$. La función de verosimilitud que se expresa en términos de las medias

$$\mu = (\mu_1, ...\mu_N)$$

Sea $L(\hat{\mu}; y)$ la máxima log verosimilitud para el modelo. Considérese que para todos los modelos posibles la máxima log verosimilitud alcanzable es L(y;y).

Esto ocurre para el modelo más general, teniendo un parámetro separado para cada observacion y el ajuste perfecto $\hat{\mu} = y$. Dicho modelo es llamado el modelo saturado. Este modelo no es útil, pues no proveé reduccion de datos. Sin embargo sirve como base de comparacion con otros ajustes del modelo.

A continuacion se detalla más al respecto de los modelos saturado y nulo.

5.4.6. Modelo nulo y modelo Saturado

El modelo nulo es el modelo más pequeño que se podría obtener, mientras que el modelo completo o saturado es el más complejo. El modelo nulo representa la situacion donde no hay relacion entre los predictores y la respuesta. Usualmente esto significa que se ajusta una media común μ para toda y, esto es, solo un parámetro.

Por otro lado en el modelo saturado, los datos se explican exactamente. Por lo general, se utilizan n parámetros para n datos puntuales. Esto puede ser logrado al ajustar un polinomio de un suficiente alto orden o mediante tratar los valores numéricos como predictores cuantitativos como codigos. Si suficientes interacciones son incluídas, el modelo podría ser saturado. Este modelo no nos dice más que los datos por sí mismos y comunmente son no informativos.

Un modelo estadístico describe como se particionan los datos en una estructura sistemática y una variacion aleatoria. El modelo nulo representa un extremo donde los datos son representados enteramente como la variacion aleatoria. Mientras que el modelo saturado o completo representa a los datos como ser enteramente sistemático. El modelo completo proveé una medida de qué tan bien cualquier modelo podría ajustar posiblemente.

5.5. Métodos de Quasi-Verosimilitud

La información de esta sección proviene del libro "Foundations of Linear and Generalized Linear Models" de Agresti (2015)[3, cap. 8] Para un MLG $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$ las ecuaciónes de verosimilitud son:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{V(\mu_i)}(\frac{\partial \mu_i}{\partial \eta_i}) = 0, j = 1, \ldots, p \tag{5.24}$$

dependiendo de la distribucion de probabilidad supuesta para y_i solo mediate μ_i y la función varianza, $V(\mu_i) = var(y_i)$. La elección de la distribucion para y_i determina la relacion $V(\mu_i)$ entre la varianza y la media. Momentos más altos tales como el sesgo pueden afectar las propiedades del modelo, como que tan rápido $\hat{\beta}$ convergen a la normalidad, pero ello no tiene impacto sobre el valor de $\hat{\beta}$ y su matriz de covarianza de una muestra

grande.

Una estimacion alternativa, la estimacion quasi-verosímil, especifica una función liga y un predictor lineal $g(\mu_i) = \sum_j \beta_j x_{ij}$ como un modelo lineal generalizado, pero esto no supone una distribucion para y_i. Esto aproxima los estimadores β_j al resolver las ecuaciones que se parecen a las ecuaciones 5.5 para los MLG. Sin embargo se asume solo una relacion media-varianza para la distribucion de y_i. Las estimaciones son la solucion de las ecuaciones 5.5, pero $V(\mu_i)$ son reemplazadas por cualquier función de varianza apropiada. En una situacion particular con un ajuste correspondiente para los errores estandarizados. Para ilustrar una aproximacion del modelado para los conteos se supone que y_i son variables Poisson independientes, para las cuales $V(\mu_i) = \mu_i$. Sin embargo la sobredispersion ocurre comúnmente tal vez por que hay heterogeneidad no modelable entre los datos de estudio.Para permitir el modelado con esta situacion se puede seleccionar $V(\mu_i) = \phi\mu_i$ para alguna constante desconocida ϕ. La quasi-verosimilitud simple para el modelo poisson o binomial simplemente supone una inflacion de la varianza para un modelo estandarizado. El método quasi-verosimilitud utiliza la misma estimacion en los MLG ordinarios pero infla los errores estandarizados al tomar en cuenta la variabilidad empírica.

5.5.1. Enfoque quasiverosímil de la variacion inflada.

Suponga que un modelo estándar especifica la función $V^*(\mu_i)$ para la varianza como una función de la media. Pero se cree que la varianza puede diferir de $V^*(\mu_i)$. Para permitir esto, se debe asumir que:

$$Var(y_i) = \phi V^*(\mu_i)$$

Para alguna constante ϕ. El valor $\phi > 1$ representa la sobredispersion. Cuando se sustituye $V(\mu_i) = \phi V^*(\mu_i)$ en la ecuación 5.5, ϕ queda libre.

Las ecuaciones son idénticas a las ecuaciones verosímiles para el MLG con función varianza $V^*(\mu_i)$, y los estimadores de los parámetros del modelo son idénicos también. Con la función varianza generalizada.

$$W_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{var(y_i)} = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\phi V^*(\mu_i)} \tag{5.25}$$

Donde la varianza asintotica $var(\hat{\beta}) = (X^{\perp}WX)^{-1}$ es ϕ veces la varianza de los MLG ordinarios. Cuando la función varianza tiene la forma $V(\mu_i) = \phi V^*(\mu_i)$. Comúnmente ϕ es desconocida. Sea

$$\chi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\hat{V^*(\mu_i)}}$$

el estadístico de Pearson generalizado 5.5.1 para el modelo simple con $\phi = 1$. Cuando χ^2/ϕ es aproximadamente chi-cuadrada, entonces tiene p en el predictor lineal y $E(\chi^2/\phi) \approx n-p$. Por lo tanto $E(\chi^2/n-p) \approx \phi$. Utilizando la estimacion por momentos coincidentes, $\hat{\phi} = \chi^2/(n-p)$ es el múltiplo estimado al aplicar a la matriz de covarianza estimada. En resumen, esta aproximacion de la quasi-verosimilitud es simple: ajusta el MLG y utiliza sus parámetros estimados por máxima verosimilitud $\hat{\beta}$. Multiplica los errores estandarizados estimados por $\sqrt{\chi^2/(n-p)}$. este método es apropiado; sin embargo, solo si el modelo elegido describe bien la relacion estructural entre $E(y_i)$ y las variables explicativas. Si un estadístico χ^2 es grande es debido a algún otro tipo de falta de ajuste, tal como fallar al incluir un término relevante de interaccion, ajustar para sobredispersion podría no abordar la insuficiencia.

5.5.2. MLG Poisson y binomial sobredispersos

Se ilustra la quasi-verosimilitud aproximacion de la inflacion-varianza con la alternativa de un MLG Poisson cuya media y varianza tienen la forma

$$V(\mu_i) = \phi\mu_i$$

Los parámetros quasi-verosímil estimados son idénticos a los maximo-verosímiles bajo supuesto de MLG Poisson. Con la liga canonica log, la matriz de covarianza ajustada es $(X'WX)^{-1}$ con $w_i = (\partial\mu_i/\partial\eta_i)^2/var(y_i) = (\mu_i)^2/\phi\mu_i = \mu_i/\phi$. Debido a la función liga, el estadístico de Pearson es

$$\chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

y $\hat{\phi} = \chi^2/(n-p)$ es la inflacion de la varianza estimada. Una aproximacion alternativa

utiliza un modelo paramétrico que permite variabilidad extra, tal como el MLG binomial negativo [3](seccion 7.3). Una ventaja de esa aproximacion es que este modelo es de hecho un modelo con función de máxima verosimilitud. La sobredispersion podría ocurrir también para conteos de datos binarios agrupados. Suponga que y_i es la proporcion de éxitos en n_i ensayos con parámetro π_i para cada intento,$i = 1, \ldots, n$. El y_i podría exhibir mayor variabilidad que lo que el binomial permite. Esto puede ocurrir en dos formas distintas. Una manera involucra heterogeneidad, con observaciones sobre un conjunto particular de variables explicativas teniendo probabilidades de éxito que varían de acuerdo a los valores de variables no observables. Para tratar con esto, se podría utilizar un modelo con mezcla jerárquico que permita que π_i tenga por si mismo una distribucion, tal como una distribucion beta. Alternativamente, variabilidad extra puede ocurrir cuando los ensayos Bernoulli a cada i son correlacionados positivamente. Se presentan modelos que reflejan estas posibilidades en el libro referenciado al principio de la sección actual [3, sec.8.2] Para ajustar una muestra supuesta binomial (es decir, independiente, ensayos bernouli idénticos), la quasiverosimilitud de la inflacion de la varianza utiliza la función varianza

$$V(\pi_i) = \phi\pi_i(1 - \pi_i)/n_i$$

para la proporcion y_i. Los estimadores quasiverosímiles son los mismos como los estimadores maximoverosímiles para el modelo binomial, y la matriz de covarianza asintotica se multiplica por ϕ. El estimador de $\phi = \chi^2/(n - p)$ utiliza el estadístico χ^2 para el modelo ordinario binomial con p parámetros,cuya ecuación es

$$\chi^2 = \sum_i \frac{(y_i - \hat{\pi})^2}{[\hat{\pi}_i(1 - \hat{\pi}_i)]/n_i}$$

A pesar de que esta aproximacion quasiverosímil con $v(\pi_i) = \phi\pi_i(1 - \pi_i)/n_i$ tiene la ventaja de ser simple, es inapropiada cuando $n_i = 1$: entonces $P(y_i = 1) = \pi_i = 1 - P(y_i = 0)$, y necesariamente $E(y_i^2) = E(y_i) = \pi_i$ y $var(y_i = \pi_i(1 - \pi_i)$. Para datos binarios desagrupados, necesariamente $var(y_i = \pi_i(1-\pi_i)$, y solo $\phi = 1$ hace sentido. este problema estructural no ocurre para la mezcla de modelos o para un modelo quasiverosímil con una función varianza que corresponde a una mezcla de modelo.

Capítulo 6

Selección, evaluación y validación de los MLG Logístico,quasibinomial,poisson y quasipoisson

En este capítulo se realiza la aplicación de los conceptos definidos en el capítulo 5. Así, en la primera sección se selecciona entre algunos modelos con funciones liga Binomial y Poisson. El cálculo de los estadísticos y el ajuste se elaboró utilizando el programa estadístico R. Aquí se explican detalladamente los códigos requeridos en R. A continuación se escribe cómo se manda a llamar la base de datos. En el siguiente código se manda a llamar a la base de datos, seleccionando y copiando esta previamente, desde una hoja de excel.

```
>DM<-read.delim("clipboard",header=T)
>names(DM)
>summary(DM)
```

En el código siguiente se muestra como se utilizó la función “factor”. Así cada variable se le asigna un número por categoría, excepto una, la categoría de referencia. estos números, en R, son llamados “levels". Se podrían utilizar otras funciones en lugar de esta para realizar la asignación. Si no se hiciera esta asignación, se podría utilizar la función “glm"para correr las funciones lineales generales, pero sería incorrecto, ya que R las tomaría como variables continuas. Para más información sobre funciones en R para codificar

variables categóricas ver la página web IDRE Stats, (2018) [47].

```
>DM$SEXO<-factor(DM$SEXO)
>DM$Edad_Dec_recod<-factor(DM$Edad_Dec_recod)
>DM$Reg_socioec<-factor(DM$Reg_socioec)
>DM$AFECPRIN3<-factor(DM$AFECPRIN3)
>attach(DM)
```

Una vez que se han asignado niveles a cada categoría por variable, se utiliza la función “relevel"para reordenar los niveles de cada factor. Para este análisis se utilizó la última categoría como nivel de referencia. Por ello se seleccionó en la función relevel el último nivel. Pues de esta forma se coloca en primer lugar al último nivel. R toma por defecto el primel nivel, el cual se indicó con “relevel"que es el último.

```
relevel(SEXO,2)->S
relevel(Reg_socioec,7)->RS
relevel(AFECPRIN3,5)->AF2
relevel(Edad_Dec_recod,5)->ED
```

El propósito de la recodificación fue para simplificar su selección.

A partir de ahora los siguientes símbolos del cuadro 6 representan las variables categóricas sexo, región socioeconómica, afección principal o tipo de complicación y edad decenal, así como la variable respuesta defunción por diabetes Mellitus:

6.1. Selección de los modelos

A continuación se explica la forma en cómo se seleccionaron los “mejores” modelos de un conjunto de modelos probables fue utilizando criterios como del de Akaike. Esta sección se basó en el libro “Categorical Data Analysis” de Agresti (2002)[1, sec. 6.1.4].

AIC, selección del modelo y modelo correcto AIC significa criterio de información Akaike, por sus siglas en inglés (Akaike information criterion) y se define de la siguiente forma:

```
AIC=-2(máxima verosimilitud - número de parámetros en el modelo)
```

Sus funciones son juzgar un modelo mediante qué tan cercanos están sus valores ajustados con sus valores reales, en términos de un cierto valor esperado. El AIC también nos ayuda

código	**variables**	**significado**
S	S1	sexo masculino
	S2	sexo femenino (variable de referencia (v.r.))
RS	RS1	Región socioeconómica 1
	RS2	Región socioeconómica 2
	RS3	Región socioeconómica 3
	RS4	Región socioeconómica 4
	RS5	Región socioeconómica 5
	RS6	Región socioeconómica 6 (v.r.)
AF	AF20	DM con complicación coma
	AF21	DM con complicación cetoacidosis
	AF22	DM con complicaciones renales
	AF23	DM con complicaciones oftálmicas
	AF24	DM con compicaciones neurológicas
	AF25	DM con compicaciones circulatorias periféricas (v.r.)
ED	ED1	De 30 a 39 años de edad
	ED2	De 40 a 49 años de edad
	ED3	De 50 a 59 años de edad
	ED4	De 60 a 69 años de edad
	ED5	De 70 a 79 años de edad (v.r.)
TDM	TDMTipo I	Diabetes Mellitus tipo I
	TDMTipo II	Diabetes Mellitus tipo II (v.r.)

Cuadro 6.1: Recodificación en R de las variables utilizadas en todos los MLG . Nótese que para cada modelo la variable de referencia (v.r.) es omitida. Véase la sección 3.6 para mayor información acerca de las regiones socioeconómicas.

a seleccionar un buen modelo en términos de estimar cantidades de interés. Además, el AIC penaliza un modelo por tener muchos parámetros.

Es preciso mencionar que al seleccionar un modelo, se está equivocado si se piensa que se ha encontrado el único verdadero. Cualquier modelo es una simplificación de la realidad. Por ello se requieren ciertas estrategias o criterios que ayuden a decidir entre el conjunto de modelos que es posible elaborar, sin tener que comparar entre todos y cada uno de estos.

Según Agresti [1, pgs. 211 y 212], las estrategias para la selección del modelo explicativo son, primero que el modelo debería ser complejo lo suficiente para ajustar bien los datos. Por otro lado, debería ser simple de interpretar. La mayoría de los estudios son diseñados para responder ciertas preguntas. Estas preguntas guían la elección de los términos del modelo.

En nuestro caso algunas preguntas podrían ser si hay relación entre el lugar de fallecimiento y el número de defunciones por DM y análogamente si hay relación entre el sexo, Tipo de Complicación, Tipo de Diabetes Mellitus, Edad decenal y el fallecimiento a causa de DM.

Para la selección del modelo se corrieron una gran cantidad de modelos entre todos los posibles que pudieran formarse con estas variables. Sin embargo, el procesador de la máquina utilizada fue incapaz de correr los modelos saturados. Por eso se eligió otro método de selección del modelo y a continuación se explica.

6.1.1. Tablas de comparación de modelos binomial y poisson

Para elaborar estas tablas se tomó en cuenta que realizar todos los modelos posibles hubiese sido costoso. Ya que requería de realizar sólo contando las interacciones de 2 en 2 posibles, las cuales eran 10, las siguientes combinaciones posibles:

$$C_{15}^{1} * C_{15}^{2} * C_{15}^{3} * C_{15}^{4} * C_{15}^{5} * \ldots * C_{15}^{15}$$

Lo cual es una cifra enorme. Aunque, más que calcular todos los modelos posibles, es necesario tener una estrategia, es decr, empezar con el modelo más simple y agregar después los términos principales, interacciones de segundo orden, etc. O bien, empezar

con un modelo con "muchas" variables y luego ir quitando variables. Por tanto, se llegó a la conclusión de que se deseaba un modelo simple, explicativo. Entonces se dió preferencia a lo que en la teoría se dice y lo que estaba disponible en las bases de datos. Es decir, a las cinco variables principales. A saber sexo, región socioeconómica, Afección pricipal , Edad decenal y Tipo de Diabetes Mellitus. Y se observaron los resultados con las interacciones entre 2 variables. Para ello se compararon, entre algunos otros, los siguientes modelos que se muestran en los cuadros 6.2 y 6.3. Así se fueron agregando interacciones, a partir del modelo simple de cinco variables principales.

componente sistemático	**estimador del parámetro de dispersión**	**AIC**
S:RS	9.8	16302.9
S:AF2	5.6	10187.2
S:ED	9.2	15559.3
RS:AF2	4.8	9049.2
RS:ED	9.3	15422.7
AF2:ED	5.2	9590.3
S	10.4	17319.7
RS	10.2	16923.9
AF2	6.0	10774.9
ED	9.8	16331.6
TDM	10.4	17429.0
S+RS+AF2+ED	4.0	7920.7
S+RS+AF2	4.8	9046.3
S+RS	9.9	16552.3
S+RS+AF2+ED+S:RS	3.9	7787.9
S+RS+AF2+ED+S:AF2	4.0	7901.2
S+RS+AF2+ED+S:ED	3.8	7614.7
S+RS+AF2+ED+RS:AF2	3.6	7323.4
S+RS+AF2+ED+RS:ED	4.0	7817.1
S+RS+AF2+ED+AF2:ED	4.0	7879.2
S+RS+AF2+ED+RS:ED	4.0	7817.1
S+RS+AF2+ED+AF2:ED	4.0	7879.2
S+RS+AF2+ED+TDM	3.9	7859.464
S+RS+AF2+ED+TDM+S:RS	3.89	7728.811

Cuadro 6.2: Tabla estimadores de dispersión y AIC para cada modelo con función liga logit

Del cuadro 6.2 se pueden observar los estimadores del parámetro de dispersión (ver definición en sección 5.5.2) que sirvieron para seleccionar un modelo que mejor describiera los datos. Para ello se utilizó el criterio de seleccionar el modelo si el estimador es un número cercano a 1. Entre más cercano a 1 es el estimador del parámetro menor dispersión.
Para la interpretación del modelo véase el cuadro 6 al inicio del capítulo. Nótese que los símbolos ":" significan que las variables colocadas una después de los dos puntos tienen interacción. Por lo cual $S : RS$ simboliza que S y RS tienen interacción y el símbolo "+" se agrega entre categoría y categoría, por ejemplo $S + RS$ significa que a las variables de la categoría sexo se les agregó las variables de las categorías de sexo y de región socioeconómica (ver cuadro 6). En resumen puede decirse de esta tabla que los modelos con mayor sobredispersión son los que tienen sólo una variable para describir la defunción por DM. Sin embargo resaltó el modelo con componente sistemático AF2 que tuvo un estimador del parámetro de dispersión de 5.9. Lo cual indica que por si misma esta variable que es Afección principal jugó un papel fundamental en el modelo que explica la defunción por DM para el año 2012, en la república Mexicana.

El cuadro 6.3 muestra estimadores del parámetro de dispersión menores a los del modelo con liga logística, de la tabla anterior. También los criterios de Akaike tienen valores menores que en la cuadro 6.2.

6.1.2. Buscando el mejor modelo con la función de "análisis de devianza".

Otra herramienta útil para buscar un mejor modelo, una vez seleccionando los mejores posibles, es utilizar la función de "Tabla de análisis de devianza". Esta es una prueba formal para ver si cada término agregado es significativo. Consiste en comparar los modelos entre sí, 2 a 2, introduciendo primero un modelo con más términos. Esto se realiza a través de los cocientes de verosimilitud correspondientes. La hipótesis nula es que no hay efecto sobre el modelo. Entonces se quiere rechazar esta prueba (que el valor del p-valor sea menor a α). La forma en que lo haremos será comparar el p-valor con el nivel de significancia (α), el cual se asume como $\alpha = 5\,\% =$0.05

componente sistemático	estimador del parámetro de dispersión	AIC
S:RS	5.7	10968.8
S:AF2	3.0	7081.9
S:ED	5.4	10508.2
RS:AF2	2.7	6590.8
RS:ED	5.4	10488.6
AF2:ED	2.8	6753.9
S	6.1	11508.0
RS	5.9	11312.3
AF2	3.2	7357.6
ED	5.7	10929.3
TDM	10.4	17429.0
S+RS+AF2+ED	2.4	6152.9
S+RS+AF2	2.7	6641.6
S+RS	5.8	11116.9
S+RS+AF2+ED+S:RS	2.3	6069.4
S+RS+AF2+ED+S:AF2	2.4	6121.1
S+RS+AF2+ED+S:ED	2.3	6014.4
S+RS+AF2+ED+RS:AF2	2.2	5868.2
S+RS+AF2+ED+RS:ED	2.4	6098.7
S+RS+AF2+ED+AF2:ED	2.4	6074.7
S+RS+AF2+ED+TDM	2.3	6122.6
S+RS+AF2+ED+TDM+S:RS	2.3	6040.1

Cuadro 6.3: Estimadores del parámetro de dispersión y AIC para cada modelo con función liga loglineal

Los siguientes modelos obtenidos en R son MLG binomiales liga logística y son comparados entre sí, usando de “Tabla de análisis de devianza".

```
model00<-glm(cbind(Cant_def,Cant_afec)~
               S+RS+AF2+ED
             ,binomial)
model01<-glm(cbind(Cant_def,Cant_afec)~
               S+RS+AF2+ED+S:RS
             ,binomial)
model02<-glm(cbind(Cant_def,Cant_afec)~
               S+RS+AF2+ED+TDM
            ,binomial)
```

```
model1b<-glm(cbind(Cant_def,Cant_afec)~
               S+RS+AF2+ED+TDM+S:RS
             ,binomial)
```

Los anteriores modelos son comparados en la siguiente salida:

```
> anova(model00,model01,test="Chi")
Analysis of Deviance Table

Model 1: cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED
Model 2: cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED + S:RS
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1451     5832.0
2      1445     5687.3  6   144.72 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> anova(model01,model02,test="Chi")
Analysis of Deviance Table
Model 1: cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED + S:RS
Model 2: cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED + TDM
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1445     5687.3
2      1450     5768.8 -5  -81.526 4.023e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(model01b,model02,test="Chi")
Analysis of Deviance Table

Model 1: cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED + S:RS
Model 2: cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED + TDM
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1445     5687.3
2      1450     5768.8 -5  -81.526 4.023e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se pudo observar que los parámetros agregados fueron significativos, pues en cada prueba de “Tabla de análisis de devianza"se rechazó la hipótesis nula. Por tanto no se rechaza que cada término agregado tiene efecto sobre el modelo para el MLG binomial liga logística. El siguiente es el código en R del modelo Poisson:

```
> modelo01<-glm(Cant_def~
+               S+RS+AF2+ED+TDM
+ ,offset=LTotal,family=poisson(link=log))
> modelo1p<-glm(Cant_def~
+               S+RS+AF2+ED+TDM+S:RS
+             ,offset=LTotal,family=poisson(link=log))
> anova(modelo01,modelo1p,test="Chi")
Analysis of Deviance Table

Model 1: Cant_def ~ S + RS + AF2 + ED + TDM
Model 2: Cant_def ~ S + RS + AF2 + ED + TDM + S:RS
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1450     3452.9
2      1444     3358.5  6   94.429 < 2.2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ocurrió lo mismo que con el modelo1b y no se rechaza la interacción S:RS. Sin embargo puede observarse que con otras interacciones el modelo tenía una devianza y AIC menor. Pero la cantidad de parámetros aumentaba considerablemente. Por ello se descartaron esos modelos, pues perdía la simplicidad del modelo.

De esta manera se llegó al siguiente predictor lineal, para las funciones liga logit (modelo1b), loglineal (modelo1p) y los MLG quasibinomial (modelo2b) y quasipoisson (modelo2p) (cuyos códigos se mencionan en la sección 6.3.2):

$$\eta = S + RS + AF2 + ED + TDM + S:RS$$

6.2. Evaluación de los modelos (modelo1b, modelo2b, modelo1p y modelo2p)

```
> library(BaylorEdPsych)
> PseudoR2(modelo1b)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
      0.64030973       0.63711317       0.99891112       0.99893467
McKelvey.Zavoina           Effron            Count        Adj.Count
      0.46780943       0.41796724       0.45912807      -0.08174387
             AIC    Corrected.AIC
```

```
   5674.19293282    5675.02453365
> PseudoR2(modelo2b)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
      0.64030973       0.63711317       0.99891112       0.99893467
McKelvey.Zavoina           Effron            Count        Adj.Count
      0.46780943       0.41796724       0.45912807      -0.08174387
             AIC    Corrected.AIC
   5674.19293282    5675.02453365
```

Debe observarse que se utilizó la librería "BaylorEdPsych"para poder obtener las diferentes Pseudo R^2 existentes, entre las cuales está la de Nagelkerke, la cual se utiliza en el libro "Extending the linear model with R: Generalized linear, Mixed effects and nonparametric regression Models"[2, sec. 2.9.]. La cual se dice debe estar entre 0 y 1, lo cual es cierto y mientras más cercana a 0, mejor ajuste se tiene. Puede observarse que con estos modelos se tiene un mal ajuste a los datos ya que ambas pseudo R^2 son iguales a 0.99.

```
> PseudoR2(modelo1p)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
       0.6299849        0.6244762        0.9796604        0.9816868
McKelvey.Zavoina           Effron            Count        Adj.Count
             NA        0.9501662        0.4455041       -0.1089918
            AIC    Corrected.AIC
    3406.4852718     3407.3168727
> PseudoR2(modelo2p)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
       0.6299849        0.6244762        0.9796604        0.9816868
McKelvey.Zavoina           Effron            Count        Adj.Count
             NA        0.9501662        0.4455041       -0.1089918
            AIC    Corrected.AIC
    3406.4852718     3407.3168727
```

Puede observarse que para la pseudo R^2 de Nagelkerke en estos dos modelos loglineales Poisson y quasipoisson, las pseudo R^2's son iguales entre sí y son ligeramente menores a las de los modelos logístico y quasibinomial, por una décima. Por lo cual puede decirse que, aunque es un mal ajuste, se tiene uno mejor con los modelos poisson y quasipoisson.

6.3. Validación de los modelos (modelo1b, modelo2b, modelo1p y modelo2p)

En esta sección se explican estadísticos tales como residuos y pruebas estadísticas que ayudan a verificar la validez de los modelos seleccionados. Según Agresti [3](pgs. 122 y 123), es importante probar para los MLG los supuestos que los soportan, como con los modelos lineales estándar. Los métodos de diagnóstico de los MLG son análogos a los de los modelos lineales estándar. Sin embargo, es necesario que se hagan algunas adaptaciones y, dependiendo del tipo de MLG, no todos los métodos diagnósticos pueden ser aplicados.

6.3.1. Diagnósticos del modelo

Los métodos de diagnóstico de modelos se pueden dividir en dos tipos. Unos que detectan casos simples o pequeños grupos de casos que no ajustan el patrón del resto de los datos. La detección de valores atípicos es un ejemplo de estos. Otros métodos son diseñados para probar los supuestos del modelo. Estos últimos se subdividen en aquellos que pueden probar la forma estructural del modelo, tal como la elección y transformación de los predictores, y aquellos que pueden probar la parte estocástica del modelo, tal como la naturaleza de la varianza acerca de la respuesta media.

6.3.2. Prueba de significancia de los parámetros

Para esta prueba las hipótesis a contrastar son $H_0 : \beta_i = 0$ contra $H_a : \beta_i \neq 0$ para cada $i = 1, \ldots, 23$, ya que nuestro modelo tiene 23 parámetros. Se desea rechazar la hipótesis nula. Y se rechaza si el p-valor del estadístico de Z, es decir el estadístico de Wald. Es menor que α.

Prueba de significancia de los parámetros de los modelos Binomial y quasibinomial

A continuación se muestran las salidas de R de las pruebas de significancia de los parámetros mencionados en el subtítulo, y se unieron las salidas para realizar una comparación entre ambos.

```
> modelo1b<-glm(cbind(Cant_def,Cant_afec)~
+                 S+RS+AF2+ED+TDM+S:RS
+               ,binomial)


> summary(modelo1b)
Call:
glm(formula = cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED +
    TDM + S:RS, family = binomial)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14.3262   -0.9662   -0.3781    0.4796    8.6780


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.31122    0.08112 -40.820  < 2e-16 ***
S1            1.15425    0.07061  16.346  < 2e-16 ***
RS1           1.28418    0.08427  15.240  < 2e-16 ***
RS2           0.85096    0.07342  11.590  < 2e-16 ***
RS3           0.69884    0.07679   9.101  < 2e-16 ***
RS4           1.38662    0.07681  18.053  < 2e-16 ***
RS5           1.62369    0.09609  16.898  < 2e-16 ***
RS6           1.46276    0.09957  14.691  < 2e-16 ***
AF20          1.55431    0.07680  20.240  < 2e-16 ***
AF21          2.73303    0.06559  41.667  < 2e-16 ***
AF22          3.04193    0.04828  63.003  < 2e-16 ***
AF23          0.50805    0.30442   1.669 0.095142 .
AF24         -0.74521    0.36539  -2.040 0.041399 *
ED1          -2.13490    0.09155 -23.320  < 2e-16 ***
ED2          -1.15886    0.04785 -24.218  < 2e-16 ***
ED3          -0.66420    0.03943 -16.846  < 2e-16 ***
ED4          -0.49272    0.03981 -12.376  < 2e-16 ***
TDMDM tipo I -0.78385    0.10416  -7.526 5.25e-14 ***
S1:RS1       -0.86968    0.11240  -7.737 1.02e-14 ***
S1:RS2       -0.31476    0.09133  -3.446 0.000568 ***
S1:RS3       -0.78739    0.09558  -8.238  < 2e-16 ***
S1:RS4       -0.20503    0.09428  -2.175 0.029646 *
S1:RS5       -1.02102    0.12544  -8.140 3.96e-16 ***
```

```
S1:RS6       -0.43830    0.12876  -3.404 0.000664 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 15641.8  on 1467  degrees of freedom
Residual deviance:  5626.2  on 1444  degrees of freedom
AIC: 7728.8
Number of Fisher Scoring iterations: 5
```

Puede observarse que los coeficientes estimados corresponden a estimate en R, cuyos nombres están al inicio, en la columna "coefficients". Sus pruebas de significancia están al final del cuadro y puede observarse que el parámetro β_{AF23}, tiene un $p - valor$ mayor a $\alpha = 5\,\%$ y para este no se rechaza la hipótesis nula de que los parámetros son cero, por lo cual no se rechaza que no es significante para el modelo.

Código del "summary" o resumen del modelo quasibinomial.

```
glm(formula = cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED +
    S + TDM + S:RS, family = quasibinomial)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14.3262   -0.9662   -0.3781    0.4796    8.6780
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.31122    0.15469 -21.406  < 2e-16 ***
S1            1.15425    0.13465   8.572  < 2e-16 ***
RS1           1.28418    0.16069   7.992 2.70e-15 ***
RS2           0.85096    0.14002   6.078 1.56e-09 ***
RS3           0.69884    0.14644   4.772 2.01e-06 ***
RS4           1.38662    0.14647   9.467  < 2e-16 ***
RS5           1.62369    0.18323   8.862  < 2e-16 ***
RS6           1.46276    0.18987   7.704 2.44e-14 ***
AF20          1.55431    0.14644  10.614  < 2e-16 ***
AF21          2.73303    0.12508  21.850  < 2e-16 ***
AF22          3.04193    0.09207  33.039  < 2e-16 ***
AF23          0.50805    0.58052   0.875   0.3816
AF24         -0.74521    0.69677  -1.070   0.2850
ED1          -2.13490    0.17458 -12.229  < 2e-16 ***
ED2          -1.15886    0.09125 -12.700  < 2e-16 ***
ED3          -0.66420    0.07519  -8.834  < 2e-16 ***
```

```
ED4          -0.49272    0.07592  -6.490 1.18e-10 ***
TDMDM tipo I -0.78385    0.19862  -3.946 8.31e-05 ***
S1:RS1       -0.86968    0.21434  -4.057 5.23e-05 ***
S1:RS2       -0.31476    0.17416  -1.807   0.0709 .
S1:RS3       -0.78739    0.18227  -4.320 1.67e-05 ***
S1:RS4       -0.20503    0.17978  -1.140   0.2543
S1:RS5       -1.02102    0.23920  -4.269 2.10e-05 ***
S1:RS6       -0.43830    0.24554  -1.785   0.0745 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasibinomial family taken to be 3.636429)
    Null deviance: 15641.8  on 1467  degrees of freedom
Residual deviance:  5626.2  on 1444  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5
```

Puede apreciarse que la única diferencia entre los modelos son los errores estandarizados, por lo cual al realizar la pruebas de significancia de cada β_i, puede observarse que también difieren al efectuar dicha prueba. Parámetros como β_{AF23} , β_{AF24}, $\beta_{S1:RS2}$, $\beta_{S1:RS4}$, $\beta_{S1:RS6}$, resultaron no significativos para el modelo quasibinomial, cuyo parámetro de dispersión se encuentra entre paréntesis como "Dispersion parameter", cuyo valor es de aproximadamente $3{,}6$, lo cual ya se había estimado al realizar el cociente de el estadístico de la devianza del modelo binomial (modelo1b), χ^2 entre sus grados de libertad.

prueba de significancia de los parámetros de los modelos Poisson y quasipoisson

Resumen del modelo Poisson.

```
> summary(modelo1p)
Call:
glm(formula = Cant_def ~ S + RS + AF2 + ED + TDM + S:RS,
family = poisson(link = log), offset = LTotal)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-12.1226   -0.8837   -0.3864    0.3822    6.7138
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.01046    0.06860 -43.883  < 2e-16 ***
S1            0.69055    0.05664  12.191  < 2e-16 ***
```

```
RS1            0.74385    0.06339  11.735  < 2e-16 ***
RS2            0.53211    0.05946   8.949  < 2e-16 ***
RS3            0.47256    0.06175   7.652 1.97e-14 ***
RS4            0.78214    0.05941  13.165  < 2e-16 ***
RS5            0.86759    0.06688  12.973  < 2e-16 ***
RS6            0.81103    0.06949  11.672  < 2e-16 ***
AF20           1.19458    0.06549  18.240  < 2e-16 ***
AF21           1.83611    0.05212  35.226  < 2e-16 ***
AF22           1.94575    0.04288  45.382  < 2e-16 ***
AF23           0.34045    0.28059   1.213   0.2250
AF24          -0.83118    0.35602  -2.335   0.0196 *
ED1           -1.07306    0.07395 -14.510  < 2e-16 ***
ED2           -0.45978    0.03112 -14.775  < 2e-16 ***
ED3           -0.21967    0.02271  -9.673  < 2e-16 ***
ED4           -0.15427    0.02224  -6.935 4.05e-12 ***
TDMDM tipo I -0.43539    0.08348  -5.215 1.84e-07 ***
S1:RS1        -0.58530    0.07839  -7.466 8.24e-14 ***
S1:RS2        -0.32958    0.06834  -4.823 1.42e-06 ***
S1:RS3        -0.51772    0.07262  -7.129 1.01e-12 ***
S1:RS4        -0.39909    0.06701  -5.955 2.60e-09 ***
S1:RS5        -0.64602    0.08189  -7.889 3.05e-15 ***
S1:RS6        -0.47774    0.08192  -5.832 5.49e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 9076.6  on 1467  degrees of freedom
Residual deviance: 3358.5  on 1444  degrees of freedom
AIC: 6040.2
Number of Fisher Scoring iterations: 5
```

El siguiente código es del "summary"del Modelo quasipoisson.

```
> modelo2p<-glm(Cant_def~S+RS+AF2+ED+S+TDM+S:RS,
offset=LTotal,family=quasipoisson(link=log))
> summary(modelo2p)
Call:
glm(formula = Cant_def ~ S + RS + AF2 + ED + S + TDM + S:RS,
    family = quasipoisson(link = log), offset = LTotal)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-12.1226   -0.8837   -0.3864    0.3822    6.7138
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.01046    0.09573 -31.448  < 2e-16 ***
S1            0.69055    0.07904   8.736  < 2e-16 ***
RS1           0.74385    0.08845   8.410  < 2e-16 ***
RS2           0.53211    0.08297   6.413 1.93e-10 ***
RS3           0.47256    0.08617   5.484 4.90e-08 ***
RS4           0.78214    0.08290   9.435  < 2e-16 ***
RS5           0.86759    0.09332   9.297  < 2e-16 ***
RS6           0.81103    0.09696   8.364  < 2e-16 ***
AF20          1.19458    0.09139  13.071  < 2e-16 ***
AF21          1.83611    0.07274  25.244  < 2e-16 ***
AF22          1.94575    0.05983  32.522  < 2e-16 ***
AF23          0.34045    0.39155   0.870 0.384717
AF24         -0.83118    0.49680  -1.673 0.094534 .
ED1          -1.07306    0.10320 -10.398  < 2e-16 ***
ED2          -0.45978    0.04342 -10.588  < 2e-16 ***
ED3          -0.21967    0.03169  -6.932 6.25e-12 ***
ED4          -0.15427    0.03104  -4.970 7.49e-07 ***
TDMDM tipo I -0.43539    0.11650  -3.737 0.000193 ***
S1:RS1       -0.58530    0.10939  -5.351 1.02e-07 ***
S1:RS2       -0.32958    0.09537  -3.456 0.000564 ***
S1:RS3       -0.51772    0.10134  -5.109 3.68e-07 ***
S1:RS4       -0.39909    0.09351  -4.268 2.10e-05 ***
S1:RS5       -0.64602    0.11427  -5.653 1.89e-08 ***
S1:RS6       -0.47774    0.11432  -4.179 3.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for quasipoisson family taken to be 1.947217)
    Null deviance: 9076.6  on 1467  degrees of freedom
Residual deviance: 3358.5  on 1444  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5
```

Es preciso observar entre estos dos modelos que del modelo Poisson al quasipoisson no se difiere en la selección de parámetros significativos, a comparación de los modelos binomial y quasibinomial.

También es importante mencionar que esta diferencia radica en que el cociente de dis-

persión en este modelo quasipoisson es menor que en el modelo quasibinomial. Mientras más cercano a 1, menos sobredispersión existe. Podría concluirse entonces que un modelo más apropiado para nuestros datos es el Poisson. Sin embargo deben realizarse análisis de residuos para validar el modelo.

6.3.3. Análisis del MLG con función liga logit,(modelo1b)

Hasta aquí es importante aclarar que, como se vió en el capítulo 5, los modelos quasi-verosímiles sólo modifican los errores estandarizados de los modelos poisson y binomial. Por eso sus estadísticos dan exactamente lo mismo que los de los modelos poisson y binomial. Por ello no se mostrarán algunos de ellos en los siguientes párrafos. Otro objetivo de las siguientes secciones es mostrar el proceso mediante el cuál se eligió utilizar los modelos quasibinomial y quasipoisson.

6.3.4. Bondad de Ajuste (modelo1b)

La devianza residual es la devianza para el modelo actual, mientras que la devianza nula es la devianza para un modelo sin predictores y sólo un término intercepto.

Devianza residual (modelo1b)

Para esta prueba la hipótesis nula (H_0) es que el modelo es un ajuste adecuado. Y se rechaza si el p-valor es menor que α =0.05. La salida en R fue la siguiente:

```
pchisq(deviance(modelo1b),df.residual(modelo1b), lower=FALSE)
```

y lo que se obtuvo fue que $p-valor$ <0.05= α, por lo que se rechaza la hipótesis nula de que el modelo es un ajuste adecuado. Y aún si no se rechazara, éso no significaría que el modelo fuera correcto, o que un modelo más simple no ajusta adecuadamente.

Devianza nula(modelo1b)

Aquí H_0: El modelo con un sólo término constante y el modelo saturado son similares. Se desea rechazar este modelo. Y ocurre si $p - valor < \alpha$. Salida en R:

```
1-pchisq(5626,1444)
```

Se obtuvo $p - valor = 0$ <0.05 por lo tanto se rechaza la H_0, es decir no se rechaza la hipótesis alternativa de que los modelos con un sólo término constante y el saturado son distintos.

Puede observarse que para el modelo actual (modelo1b) tiene un devianza de 5,626 lo cual es grande para 1,444 grados de libertad, por tanto podría sospecharse que existe sobredispersión.

Pero antes de concluír que existe sobredispersión se requiere eliminar otras explicaciones potenciales.

Debe observarse que cada combinación de casos, lo cual corresponde a cada fila de la Base de datos tiene mínimo un afectado diabético, la media es 21.04 y el máximo es 757 casos totales de diabéticos, con una combinación específica. Lo anterior muestra que la distribución de los conteos está sesgada a la derecha, por ello la varianza es grande. Esta podría ser una de las razones de la sobredispersión. También se deben checar los valores atípicos y observar la fórmula del modelo. Se verán en la subsección 6.3.13.

La gráficas de logits empíricos se utilizan para comprobar si los predictores están correctamente expresados. Los logits empíricos se definen como:

$log(\frac{y+\frac{1}{2}}{m-y+\frac{1}{2}})$

Donde m: es la cantidad de casos totales de la suma de cantidad de afectados y difuntos. y y: la cantidad de difuntos por DM. Los logits empíricos son los mejores estimadores de los logits según Agresti [1, sec. 5.1.2]. Fueron nombrados como logits empíricos por Faraway [2, sec. 2.11]

Los siguientes son los códigos en R de las gráficas de interacción encontradas en la figura 6.1.

```
#logits emp\'iricos
elogits<-log((DM$Cant_def+0.5)/(DM$Cant_afec+0.5))
with(DM,interaction.plot(Reg_socioec,SEXO,elogits))
with(DM,interaction.plot(Reg_socioec,AFECPRIN3,elogits))
with(DM,interaction.plot(Reg_socioec,Edad_Dec_recod,elogits))
with(DM,interaction.plot(Edad_Dec_recod,SEXO,elogits))
with(DM,interaction.plot(Edad_Dec_recod,AFECPRIN3,elogits))
with(DM,interaction.plot(AFECPRIN3,SEXO,elogits))
with(DM,interaction.plot(SEXO,AFECPRIN2,elogits))
```

```
with(DM,interaction.plot(Reg_socioec,AFECPRIN2,elogits))
with(DM,interaction.plot(AFECPRIN3,AFECPRIN2,elogits))
with(DM,interaction.plot(Edad_Dec_recod,AFECPRIN2,elogits))
```

Los gráficos que estos códigos produjeron (figura 6.1) están en orden de derecha a izquierda, y de arriba hacia abajo y coinciden con los códigos. A cada código le corresponde un gráfico.

Se dice que las gráficas de interacciones son siempre difíciles de interpretar conclusivamente. Según la página web Support.minitab.com. [48] si las líneas son paralelas entre sí entonces no existen interacciones entre las variables. Si, por el contrario no son paralelas entonces ocurren interacciones. Sin embargo, mientras más paralelas son las líneas, mayor es la fuerza de la interacción. Puede observarse que en la mayoría de las gráficas de interacciones son no paralelas. Sin embargo, "parece no haber señales obvias de grandes interacciones. Entonces no hay evidencia de que el modelo lineal sea inadecuado" según Faraway [2, pg. 46]. Y en nuestro caso parece no ser inadecuado el modelo lineal.

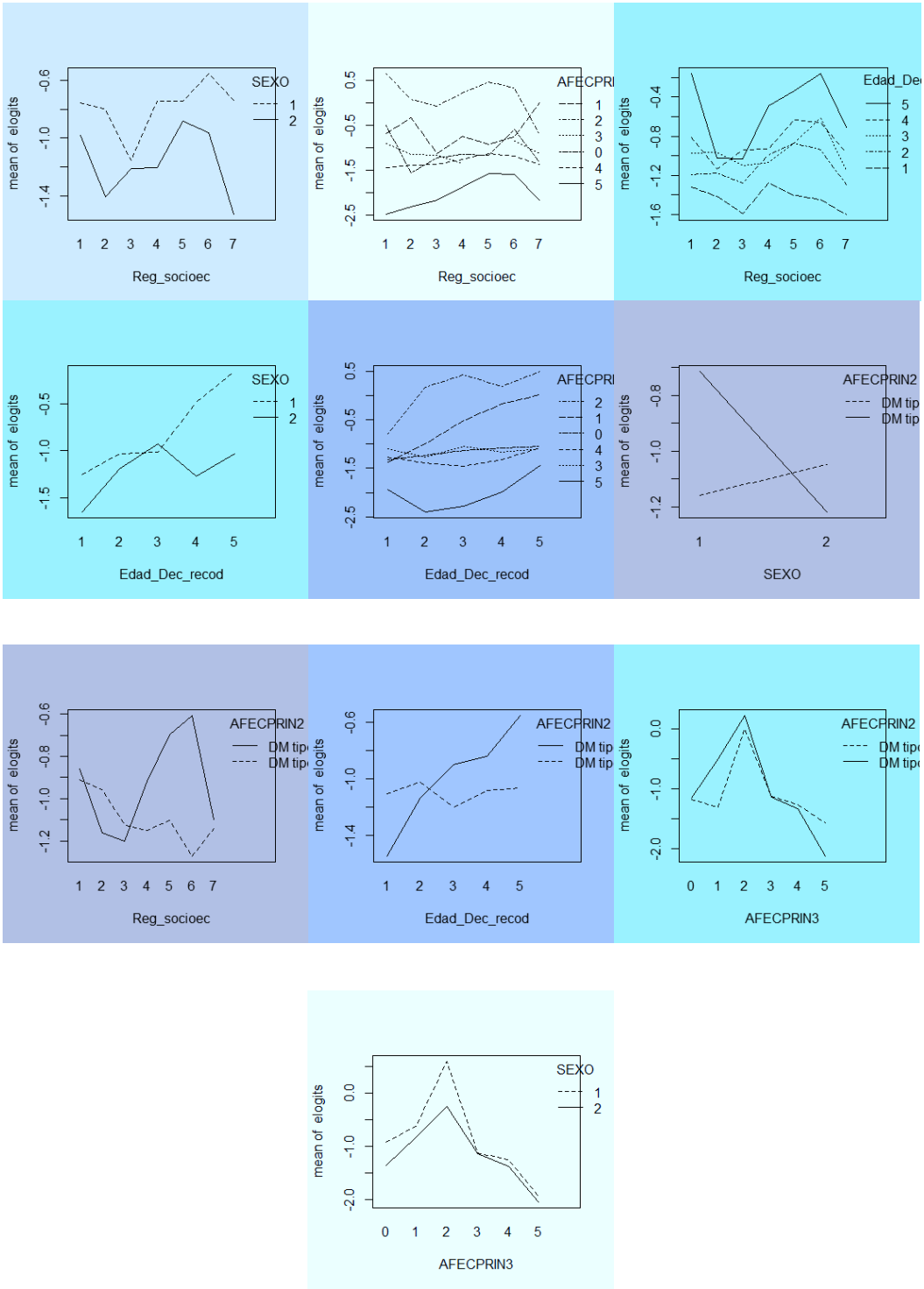
6.3.5. Dispersión(modelo1b)

Si no se tiene ningún valor atípico y la forma funcional del modelo parece adecuado, pero la devianza es todavía grande a la esperada, entonces se podría culpar a la sobredispersión.

Lo siguiente es el código R donde se calcula el parámetro de dispersión.

```
> sres2<-sum(residuals(modelo1b,type="pearson")^2)
> sigma2<-sres2/df.residual(modelo1b)
> sigma2
[1] 3.636278
```

Figura 6.1: Elogits del modelo 1b para visualizar si se requiere añadir interacciones o no



Como dicho parámetro “sigma2” es mayor a 1 entonces parece que existe sobredispersión. Y para corregir los parámetros del modelo con sobredispersión se puede utilizar un modelo quasibinomial.

6.3.6. MLG Quasibinomial(modelo2b)

A continuación se hace una prueba F sobre los predictores.

```
> drop1(modelo1b,scale=sigma2,test="F")
Single term deletions
Model:
cbind(Cant_def, Cant_afec) ~ S + RS + AF2 + ED + TDM + S:RS
scale:  3.636278
       Df Deviance    AIC  F value    Pr(>F)
<none>      5626.2 7728.8
AF2     5  12680.0 9658.7 362.0815 < 2.2e-16 ***
ED      4   6644.4 8000.8  65.3351 < 2.2e-16 ***
TDM     1   5687.3 7743.6  15.6886 7.831e-05 ***
S:RS    6   5768.8 7756.0   6.1021 2.513e-06 ***
--
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In drop1.glm(modelo1b, scale = sigma2, test = "F") :
  F test assumes ’quasibinomial’ family
```

Puede observarse que todos los parámetros fueron significativos para esta prueba F. Entonces no se rechazan todos estos para el modelo quasibinomial. Y la escala está dada por el parámetro de dispersión estimado con sigma2, obtenido en la sección anterior. El mensaje “warning” nos recuerda que el uso de un parámetro de dispersión libre, resulta en un MLG quasibinomial.

6.3.7. Análisis del MLG con función liga loglineal de Poisson,(modelo1p)

6.3.8. Bondad de Ajuste(modelo1p)

```
> pchisq(deviance(modelo1p),df.residual(modelo1p), lower=FALSE)
[1] 9.798843e-154
 Devianza Nula
> 1-pchisq(deviance(modelo1p),df.residual(modelo1p))
[1] 0
```

Para la devianza residual definida en la subsección 6.3.4 se obtuvo que $9{,}798843e-154$ <0.05, por lo que se rechaza la hipótesis nula de que el modelo es un ajuste adecuado.

Mientras que para la devianza nula se obtuvo $p-valor = 0$ <0.05 por lo tanto se rechaza la H_0, es decir no se rechaza la hipótesis alternativa de que los modelos con un sólo término constante y el saturado son distintos.

6.3.9. Dispersión (modelo1p)

```
> deviance(modelo1p)/df.residual(modelo1p)
[1] 2.325821
```

Puede observarse que el cociente de dispersión estimado para el modelo loglineal tipo Poisson fue de “2.32”, lo cual muestra que existe sobredispersión. Sin embargo puede observarse que es menor a la sobredispersión existente que para el modelo logit. Para solucionar ésto se procede como en el modelo anterior al buscar un modelo más flexible, que pueda utilizar un parámetro adicional para corregir la sobredispersión. Ésto puede realizarse utilizando un modelo quasipoisson, con un parámetro de dispersión aproximadamente 2.3. Sin embargo existe otra forma de estimar el parámetro de dispersión, la cual es la siguiente:

```
#estimando el parámetro de dispersión
> (dp<-sum(residuals(modelo1p, type="pearson")^2)
/modelo1p$df.res)
[1] 1.94718
```

Se puede investigar más al respecto de la sobredispersión con la gráfica de varianza estimada contra la media . Se desea que la varianza sea igual a la media. Investigemos esta relación para este modelo. Es difícil estimar la varianza para un valor dado de la media pero$(y - (\hat{\mu}))^2$ sirve como aproximación cruda.
Lo que sigue es el gráfico de esta varianza estimada contra la media en R:

```
plot(log(fitted(modelo1p)),
log((DM$Cant_def-fitted(modelo1p))^2),
xlab=expression(hat(mu)),
ylab=expression((y-hat(mu))^2))
abline(0,1)
```

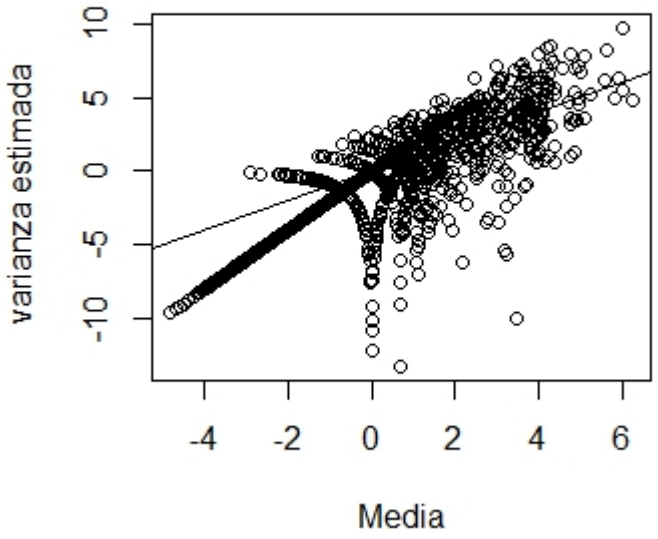


Figura 6.2: Relación entre la media y varianza. La linea mostrada representa la media igual a la varianza.

De la figura 6.2 se puede decir que hay tendencias con distintas formas. Unas son curveadas y otros están lejos de tendencia más marcada . esta tendencia muestra que la varianza es proporcional, pero más grande que la media.

Cuando el supuesto del modelo de regresión Poisson se rompe pero la función liga y la elección de predictores es correcta, los estimadores de los parámetros β son consistentes, pero los errores estandarizados podrían ser incorrectos. La distribución Poisson tiene un sólo parámetro y por tanto no es muy flexible para propósitos de ajuste empírico. Se puede generalizar permitiendo un parámetro de sobredispersión. La sobredispersión o infrardispersion, puede ocurrir en los modelos Poisson. Si no se conoce el parámetro de sobredispersión se puede ajustar un modelo quasipoisson. Lo que sigue es este ajuste.

6.3.10. MLG QuasiPoisson(modelo2p)

El código en R para el ajuste quasipoisson es:

```
###modelo quasi-poisson
modelo2p<-glm(Cant_def~S+RS+AF2+ED+S+TDM+S:RS,
offset=LTotal,family=quasipoisson(link=log))
```

este mismo se utilizó para la prueba de significancia de los parámetros.

6.3.11. Residuos de Pearson, de Devianza y residuos estandarizados para MLG.

Para un modelo particular con función varianza $V(\mu)$, los residuos de Pearson para observaciones y_i y su valor ajustado $\hat{\mu}_i$ es:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

La suma de los valores al cuadrado para el estadístico de Pearson generalizado es

$$\chi^2 = \frac{\sum_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

y cuando μ_i son grandes y el modelo se mantiene, e_i tiene una distribución aproximadamente normal y $\chi^2 = \sum_i e_i^2$ tiene una distribución chi-cuadrada aproximada.

Para un MLG binomial en el cual $n_i y_i$ tiene una distribución $bin(n_i, \pi_i)$, los residuos

de Pearson son:

$$e_i = (y_i - \hat{\pi}_i)/\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}$$

y cuando n_i son grandes $\chi^2 = \sum_i e_i^2$ tienen una distribución aproximada chi-cuadrada. En estos casos, tales estadísticos son utilizados en la prueba de bondad de ajuste del modelo.

De la expresión 4.15[3] para la devianza, sea $D(y;\hat{\mu}) = \sum_i d_i$, donde

$$d_i = 2\omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

El residuo de devianza es:

$$\sqrt{d_i} \times signo(y_i - \hat{\mu}_i)$$

Para juzgar cuando un residuo es "grande" es útil ver si tienen valores de sus medias de 0 y varianzas de 1. Sin embargo, los residuos de Pearson y de Devianza tienden a tener varianza menor a 1 porque estos comparan y_i con la media ajustada $\hat{\mu}_i$ más que la media verdadera μ_i.

Análisis de residuos (modelo1b, modelo2b, modelo1p y modelo2p)

A continuación se muestran los residuos de Pearson y de devianza para los modelos binomial liga logística (modelo1b) y quasibinomial (modelo2b).

```
> summary(residuals(modelo1b,"pearson"))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-12.4500  -0.7862  -0.2847  -0.1060   0.5077   8.7400
> var(residuals(modelo1b,"pearson"))
[1] 3.568019
> ##residuales de devianza
> summary(residuals(modelo1b))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-14.3300  -0.9662  -0.3781  -0.2393   0.4796   8.6780
> var(residuals(modelo1b))
[1] 3.777863
##residuales de pearson
>12.4500  -0.7862  -0.2847  -0.1060   0.5077   8.7400

> var(residuals(modelo2b,"pearson"))
[1] 3.568019
```

```
> ##residuales de devianza
> summary(residuals(modelo2b))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-14.3300  -0.9662  -0.3781  -0.2393   0.4796   8.6780
> var(residuals(modelo2b))
[1] 3.777863
```

Del código anterior puede notarse que los residuos tanto de Pearson como de devianza son iguales para los MLG binomial y quasibinomial cuyos componentes sistemáticos son los mismos.

```
 ##residuales de devianza
> summary(residuals(modelo1p,"pearson"))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-8.57200 -0.66510 -0.28250 -0.07675  0.40260  8.23600
 var(residuals(modelo1p,"pearson"))
[1] 1.910757
##residuales de pearson
> summary(residuals(modelo1p))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-12.1200  -0.8837  -0.3864  -0.2991   0.3822   6.7140
> var(residuals(modelo1p))
[1] 2.199847
> ##residuales de pearson
> summary(residuals(modelo2p,"pearson"))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-8.57200 -0.66510 -0.28250 -0.07675  0.40260  8.23600
>  var(residuals(modelo2p,"pearson"))
[1] 1.910757
>  ##residuales de devianza
>  summary(residuals(modelo2p))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-12.1200  -0.8837  -0.3864  -0.2991   0.3822   6.7140
>  var(residuals(modelo2p))
[1] 2.199847
```

Si se comparan las varianzas de los residuos de Pearson del modelo binomial con los del modelo Poisson hay una diferencia de más de una unidad. Lo cual sugiere que el modelo Poisson es un mejor ajuste para los datos, pues su varianza se acerca más a uno

y las medias del modelo Poisson y quasipoisson se acercan más a 0 que las del modelo binomial y quasibinomial. Observese que los modelos quasipoisson y poisson tienen los mismos valores residuales.

6.3.12. Gráficos de residuos contra valores ajustados(modelo1b, modelo2b, modelo1p y modelo2p).

Para los modelos lineales,el gráfico de residuos contra valores ajustados es probablemente el gráfico más valioso. Mientras que para los MLG se debe elegir en la escala más apropiada para los valores ajustados. Comúnmente es mejor graficar los predictores lineales $\hat{\eta}$, mejor que las respuestas predichas.

Del gráfico de residuos de devianza contra $\hat{\mu}$ se buscan dos características. La primera es si hay alguna relación no lineal entre los valores predichos ($\hat{\eta}$ y $\hat{\mu}$) y los residuos. Si sí, esto indica ausencia de ajuste que podría ser rectificado con un cambio en el modelo.

Se puede considerar un cambio en la función liga, pero ésto es con frecuencia indeseable ya que sólo algunas opciones de funciones liga conducen a modelos fácilmente interpretables. Sin embargo es mejor si se puede hacer un cambio en la elección de variables explicativas del predictor lineal ó transformaciones en estos mismos, ya que ésto implica la menor interrupción para el MLG.

A continuación se mostrarán los códigos de tres tipos distintos de gráficos de residuos contra valores ajustados. Cabe destacar que se excluyeron los códigos y gráficos de los modelos quasiverosímiles. Esto fue debido a que el del modelo quasibinomial coincidió con el del modelo binomial y análogamente para el modelo Poisson. El código siguiente es de los gráficos de $\hat{\mu}$ contra los residuos de devianza.

```
plot(residuals(modelo1b)~predict(modelo1b,type="response"),
main="Residuos vs ajustados (1b)",xlab=expression(hat(mu))
,ylab="residuos de devianza")
plot(residuals(modelo1p)~predict(modelo1p,type="response"),
main="Residuos vs ajustados (1)",xlab=expression(hat(mu)),
```

Del primer gráfico correspondiente al modelo binomial se puede notar que los valores de los momios ajustados de la respuesta defunción por DM paracen ser más frecuentes

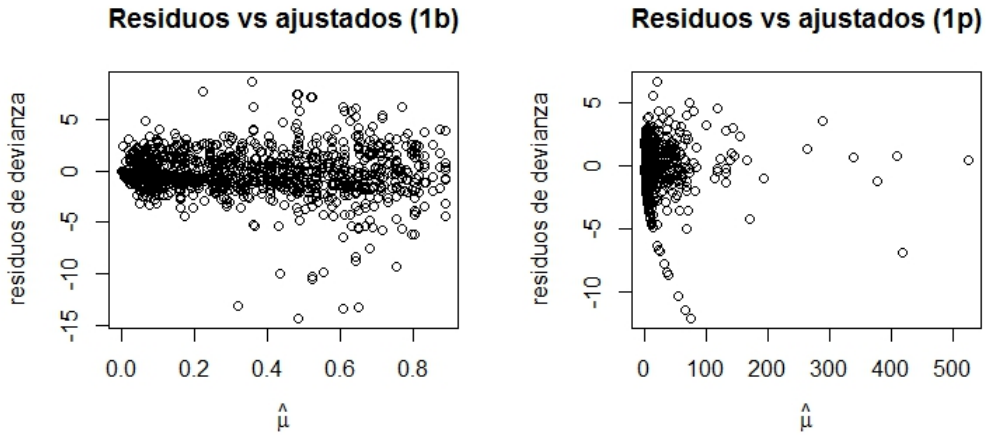


Figura 6.3: Valores ajustados en escala de la respuesta de los modelos 1b y 1p

los que están entre 0 y 0.2. También puede observarse que va disminuyendo su frecuencia conforme el valor de los momios de la defunción por DM aumenta. Lo que también puede notarse es que es uniforme el el crecimiento de los valores ajustados contra los residuos de devianza. Es decir, no tienen relación.

En el segundo gráfico puede observarse que existen muy pocos casos con una media de defunciones por DM muy grandes. Y que la mayoría de los casos se concentra entre 0 y 100. Por esta razón es difícil ver la relación que existe entre los residuales y los valores ajustados por que la mayoría de los puntos se concentran a la izquierda.

A continuación se muestran los códigos y gráficos del segundo tipo de gráfico de residuales contra valores ajustados.

```
plot(residuals(modelo1b)~predict(modelo1b,type="link"),
main="Residuos vs ajustados (1b)",xlab=expression(hat(eta)),
ylab="residuos de devianza")
plot(residuals(modelo1p)~predict(modelo1p,type="link"),
main="Residuos vs ajustados (1)",xlab=expression(hat(eta)),
ylab="residuos de devianza")
```

En este conjunto de gráficos se busca evidencia de no linearidad. En el primer modelo

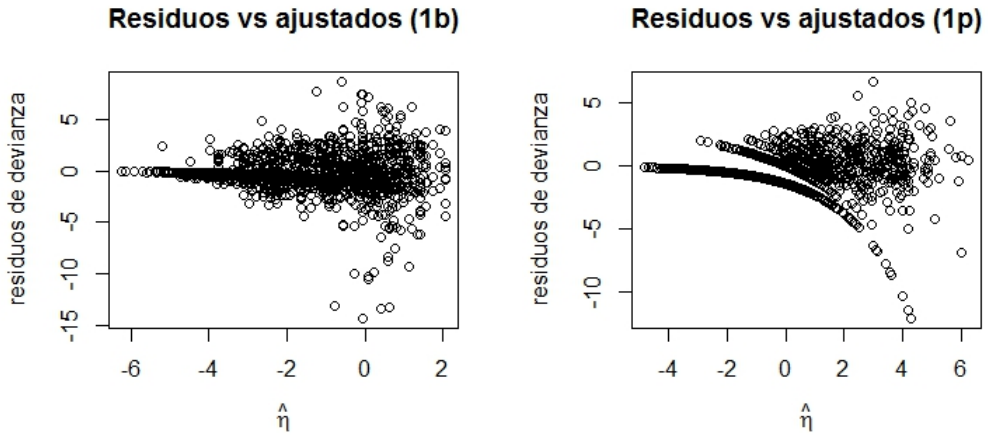


Figura 6.4: Residuos de devianza vs ajustados. Modelo 1b y Modelo1p

parece no existir evidencia de no linearidad. Y el segundo muestra entre los valores ajustados de -4 a 4 una tendencia curva descendente de los valores de los residuos de devianza.

Con ésto podría decirse que en el modelo poisson se presenta una indicación de ausencia de ajuste que podría ser retificado mediante un cambio en el modelo. En este caso se podría considerar un cambio en la elección de las variables explicativas o bien realizar transformaciones de dichas variables con el propósito de conservar la "interpretabilidad" del modelo Poisson.

A continuación son mostrados los códigos y gráficos de los residuos de respuesta contra valores ajustados de $\hat{\eta}$ (i.e. predictores lineales).

```
plot(residuals(modelo1b)~predict(modelo1b,type="response"),
main="Residuos vs ajustados (1b)",xlab=expression(hat(eta)),
ylab="residuos de respuesta")
plot(residuals(modelo1p)~predict(modelo1p,type="response"),
main="Residuos vs ajustados (1)",xlab=expression(hat(eta)),
ylab="residuos de respuesta")
```

La varianza de los residuos con respecto a los valores ajustados es la segunda carac-

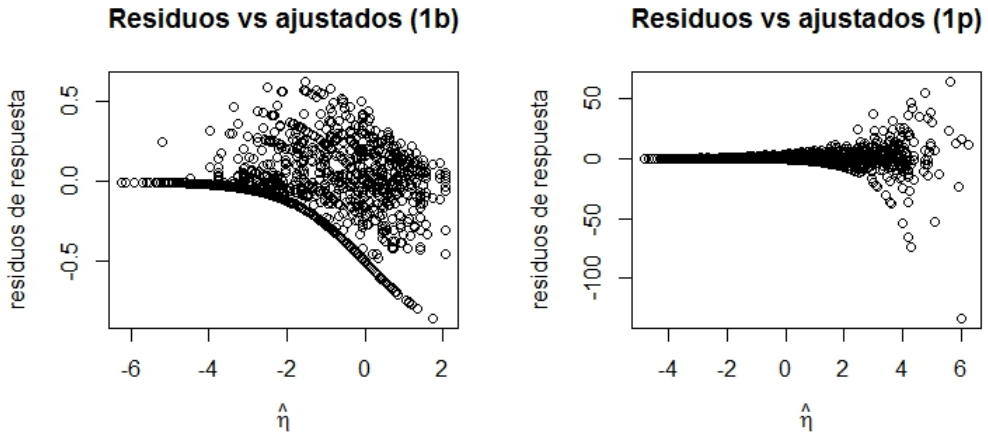


Figura 6.5: Residuos de respuesta vs ajustados. Modelo1b y Modelo1p

terística que se verifica. Los supuestos de los MLG requieren varianza constante en el gráfico. La violación de este supuesto muestra que se debe realizar un cambio en el modelo. Por ésto, se podría considerar un cambio en la función varianza, lo cual implica abandonar el MLG Poisson.

Los gráficos 1 y 2 muestran que existe un patrón de incremento en la variación consistente con Poisson.

Debe considerarse que en algunos casos, esta clase de gráficos no son particularmente de ayuda. Para un modelo con respuesta binaria, los residuos pueden tomar sólo dos valores posibles para la respuesta predicha dada. esta es la situación más extrema, pero puede ocurrir una "discreción"similar para respuestas binomiales con grupos de tamaños pequeños y con respuestas Poisson que son pequeñas. Los gráficos de residuos en estos casos tienden a mostrar lineas curvas de puntos correspondientes al número limitado de respuestas observadas. Tales artefactos pueden oscurecer el propósito principal de los gáficos.

6.3.13. Puntos discrepantes (outliers)

La siguiente subsección se basa en el libro “Extending the linear model with R: Generalized linear, Mixed effects and nonparametric regression Models” Faraway [2, sec. 6.4] Los residuos, leverages y medidas de influencia pueden ser utilizados para comprobar

puntos que no ajustan el modelo o influyen el ajuste excesivamente. Los "Q-Q plot"de los residuales son los gráficos estándares para comprobar los supuestos de normalidad sobre los errores hechos para un modelo lineal. Mientras tanto, para un MLG, no se espera que los residuos sean normalmente distribuidos. Pero se sigue interesado en detectar puntos inusuales. Por ésto es mejor utilizar una gráfica half-normal que compara los residuos ordenados y los cuantiles de la distribucioón “half-normal":

$$\phi^{-1}\left(\frac{n+i}{2n+1}\right)$$

$i = 1, \ldots, n$ Como no se espera que los residuos sean normalmente distribuídos, entonces no se está buscando una línea recta aproximada. Se está buscando puntos inusuales los cuales podrían ser identificados como puntos fuera de la tendencia. Un gráfico half-normal es mejor para este propósito en el sentido de que la resolución del gráfico es duplicada, al tener todos los puntos en una cola. Para checar los puntos inusuales, se deben graficar los residuos jacknife (o también llamados "studentized residuals"). Los gráficos half-normal son útiles también para valuar positivamente diagnósticos tales como los leverages y los estadísticos de cook.

A continuación se muestran los códigos y gráficos de los gráficos halfnormal de los estadísticos de cook y residuos jacknife. Para estos gráficos es requerida la librería “faraway” mencionada en el código.

```
library(faraway) ##librería
halfnorm(cooks.distance(modelo1b),main="modelo1b")
halfnorm(cooks.distance(modelo2b),main="modelo2b")
halfnorm(cooks.distance(modelo1p),main="modelo1p")
halfnorm(cooks.distance(modelo2p),main="modelo2p")
halfnorm(rstudent(modelo1b),main="modelo1b")
halfnorm(rstudent(modelo2b),main="modelo2b")
halfnorm(rstudent(modelo1p),main="modelo1p")
halfnorm(rstudent(modelo2p),main="modelo2p")
```

Los gráficos de arriba parecen mostrar, todos, que el caso 126 es uno atípico.

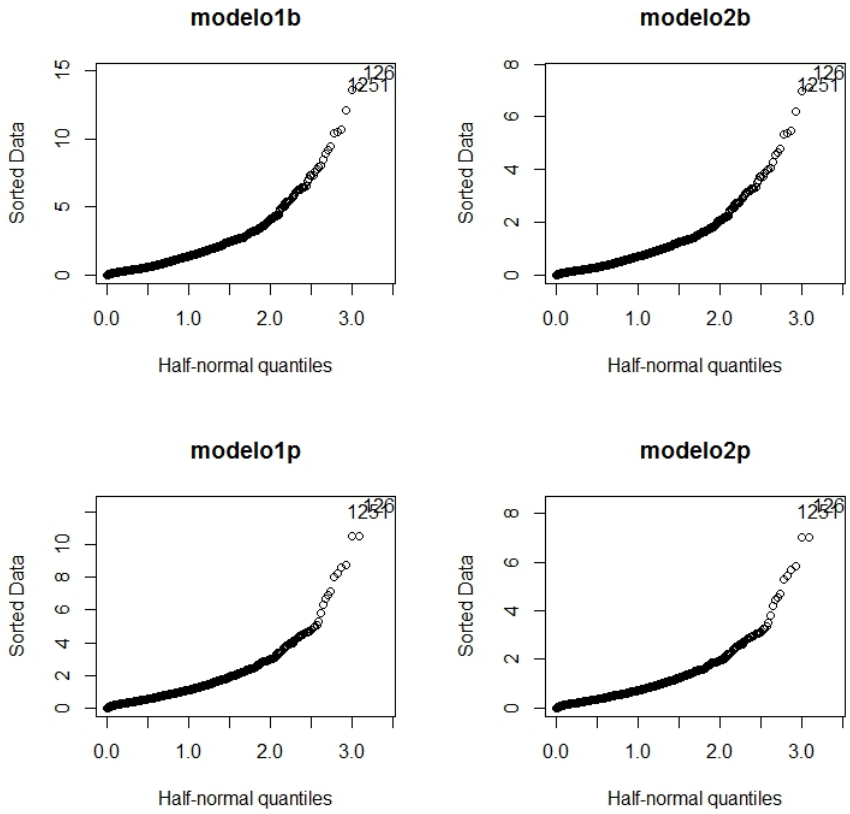


Figura 6.6: Gráficos halfnormal de los residuos jacknife

Puede observarse en los gráficos de la parte superior existe un valor influyente, el cual es el caso 248. Al revisar dicho caso en la base de datos se encontró que tenía 472 conteos

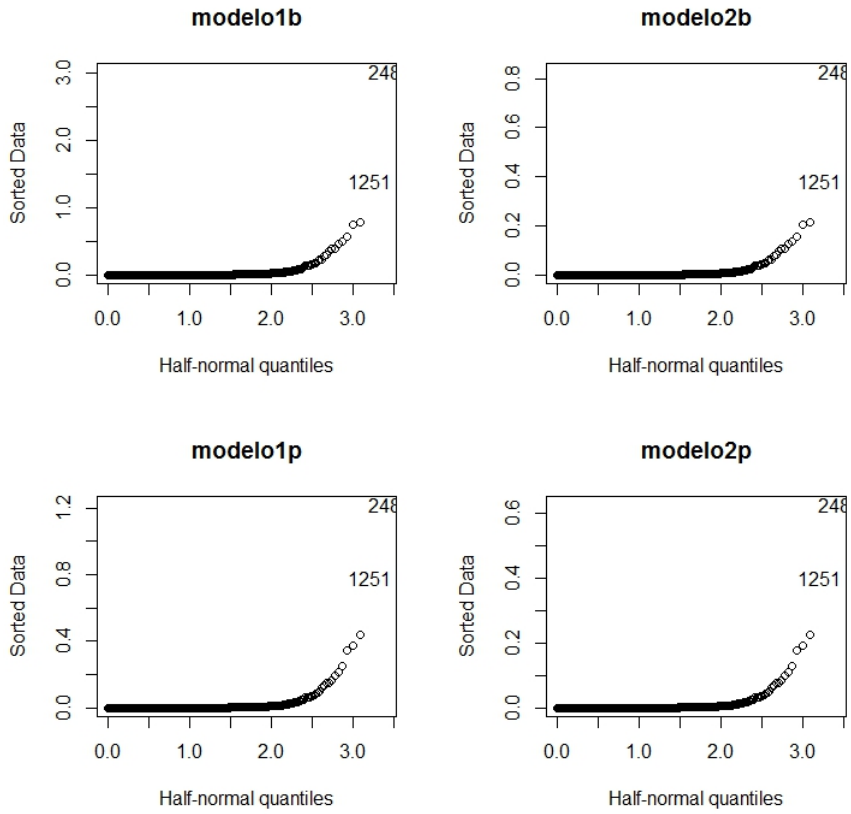


Figura 6.7: Gráficos halfnormal de los estadísticos de Cook

de afecciones y 285 conteos de defunciones, por lo que en total contaba con 757. Este caso se localizó en la región socioeconómica de la CDMX (7), de sexo masculino (1), Edad decenal de 50 a 59 años (3), y pertenecían a los tipos de complicaciones renales (2) y Diabetes Mellitus tipo II.

En el capítulo que sigue se realiza la interpretación de los coeficientes estimados ($\hat{\beta}$) para estos modelos, así como sus respectivos intervalos de confianza.

Capítulo 7

Uso , reporte y comparación de los MLG Logístico,quasibinomial,poisson y quasipoisson

7.1. Uso y reporte de los modelos

En esta sección se muestra la forma en como se interpretaron los coeficientes de los MLG usados en el capítulo 6. Para realizar una interpretación correcta de los parámetros del modelo Poisson se requiere el concepto de **letalidad**. Y se estudia la letalidad porque se está realizando la interpretación sobre las tasas de difuntos con diabetes mellitus sobre un total de enfermos y difuntos de diabetes mellitus en la población Mexicana en el año 2012, es decir, esta es la media estimada con el modelo Poisson. Y así como fue estimada por el modelo poisson también lo fue para el modelo Quasipoisson.

7.1.1. Interpretación de los coeficientes

Para interpretar los coeficientes de este modelo se utilizan los $\exp \beta_{ij}$. Esto es porque se utiliza el resultado 5.11 del capítulo 5 que dice que los “odds ratio"(cocientes de momios ó cocientes de posibilidades pg. 46) son iguales a la función exponencial de los coeficientes del modelo.

Se realizará la interpretación para el modelo binomial y como este modelo sólo corrige

la dispersión y tiene efecto sobre la desviación estándar, entonces la interpretación es la misma para el modelo Quasibinomial.

En R se realizó de la siguiente forma:

```
exp(coefficients(modelo1b))
exp(coefficients(modelo1p))
```

Variable	**modelo1b**	**modelo1p**	**(1)si modelo2b**	**(1)si modelo2p**
(Intercept)	0.04	0.05	1	1
S1	3.17	1.99	1	1
RS1	3.61	2.10	1	1
RS2	2.34	1.70	1	1
RS3	2.01	1.60	1	1
RS4	4.00	2.19	1	1
RS5	5.07	2.38	1	1
RS6	4.32	2.25	1	1
AF20	4.73	3.30	1	1
AF21	15.38	6.27	1	1
AF22	20.95	7.00	1	1
AF23	1.66	1.41	0	0
AF24	0.47	0.44	0	0
ED1	0.12	0.34	1	1
ED2	0.31	0.63	1	1
ED3	0.51	0.80	1	1
ED4	0.61	0.86	1	1
TDME10	0.46	0.65	1	1
S1:RS1	0.42	0.56	1	1
S1:RS2	0.73	0.72	0	1
S1:RS3	0.46	0.60	1	1
S1:RS4	0.81	0.67	0	1
S1:RS5	0.36	0.52	1	1
S1:RS6	0.65	0.62	0	1

Cuadro 7.1: Tabla de parámetros en los modelos modelo1b y modelo1p. El código es 1 si el parámetro fue significativo para el modelo de la columna correspondiente y 0 en caso contrario.

En la tabla 7.1 puede apreciarse cada uno de los exponentes de los parámetros de los modelo binomial (modelo1b) $logit(\pi_i) = S + RS + AF2 + ED + S + TDM : RS$ y poisson $log(\pi_i) = S + RS + AF2 + ED + S + TDM : RS$(modelo1p).

Sería redundante incluir la interpretación de los modelos quasibinomial y quasipoisson, ya que los parámetros son exactamente los mismos sólo que con los errores estandarizados distintos, como ya se vió en 6.3.2.

A continuación se hará la interpretación de algunos cocientes de posibilidades (odds ratio) considerados vitales para el modelo, en cada uno de ellos la defunción es de DM y el contexto es en México, en el año 2012. Así también se realiza la interpretación a modo de comparación para el modelo Poisson.

La constante, o intercepto, para el modelo Poisson representa el logaritmo de la tasa media de defunciones por DM (dado que se padecía esa enfermedad), para la celda de referencia. En este caso la celda de referencia fue Sexo femenino (S2), CDMX (RS7), (AF25) DM con complicaciones circulatorias periféricas, (ED5) Edad de entre 70 a 79 años, (TDME11) DM tipo II y (S1:RS7) sexo masculino y CDMX.

Dado que el exponente de la constante fue 0.05, entonces puede decirse que la tasa media de defunciones por DM fue esta cantidad. O bien, hubo 5 defunciones por cada 100 egresos hospitalarios por esta enfermedad (afectados y difuntos de DM), del sexo femenino, de la CDMX, con complicaciones circulatorias periféricas, de entre 70 a 79 años, con DM tipo II y con la interacción de sexo masculino y CDMX.

Interpretación de los parámetros de la variable SEXO

Para la variable sexo denotada con S1, 1 significa que es sexo masculino, por lo que el cociente de posibilidades del sexo masculino es

$$\hat{\theta}_M = \frac{Momio_{hombres}}{Momio_{mujeres}} = exp(\hat{\beta}_{S1}) =$$

3.17

significa que la posibilidad de defunción de DM de los hombres fue aproximadamente 3.2 veces mayor que de las mujeres.

En el caso de la regresión poisson se tiene $exp(\hat{\beta}_{S1})$ =1.99, por lo cual se dice que, el coeficiente estimado del sexo masculino (1) indica que en el sexo femenino la tasa de

letalidad de diabetes Mellitus fue aproximadamente el doble respecto a la tasa de letalidad que se tenía en el sexo femenino.

Interpretación de los parámetros de la variable Región Socioeconómica

$$\hat{\theta}_{RS(5)} = \frac{Momio_{RS(5)}}{Momio_{RS(CDMX)}} = exp\hat{\beta}_{RS(5)} = 5$$

Se interpreta como que la posibilidad de defunción de los afectados por DM fue 5 veces mayor para los pacientes que vivían en Baja California, Baja California Sur, Chihuahua, Sonora y Tamaulipas que para los pacientes de la CDMX.

Como $\hat{\theta}_{RS_6} \approx \hat{\theta}_{RS_4} = 4$, se puede decir que la defunción de los afectados por DM fue aproximadamente 4 veces mayor para los pacientes que vivían en las regiones socioeconómicas 4 y 6 que en la CDMX. Es decir en los estados de Colima, México, Morelos, Nayarit, Quintana Roo, Sinaloa y Yucatán para la RS4 y Aguascalientes, Coahuila y Nuevo León para la RS6.

Mientras tanto para el modelo Poisson se puede decir que el coeficiente estimado en la RS5 fue 2.3 veces mayor con respecto a la tasa de letalidad que se tuvo en la CDMX.

Y con respecto a la RS4 se tuvo que la tasa de letalidad de DM fue aproximadamente dos veces mayor con respecto a la tasa de letalidad que se tuvo en la CDMX.

Interpretación de los parámetros de los tipos de complicaciones

$$\hat{\theta}_{AF22} = exp\beta_{AF_{22}} = 21$$

Puede decirse que la posibilidad de defunción de los afectados fue aproximadamente 21 veces mayor para los que tenían complicaciones renales que para los afectados con complicaciones circulatorias periféricas.

Puesto que $\hat{\theta}_{AF21} \approx 15$, la posibilidad de defunción de los afectados fue 15 veces mayor para los afectados con cetoacidosis que para los afectados con complicaciones circulatorias periféricas.

En el modelo poisson se tiene para estos parámetros que el coeficiente estimado de los afectados con complicaciones renales la tasa de letalidad fue 7 veces mayor a comparación

de la tasa de letalidad que se tuvo para los afectados con complicaciones circulatorias periféricas.

Se puede mencionar también que en todos los modelos los parámetros de las complicaciones AF23 y AF24 no fueron significativos, con un nivel de significancia del $5\,\%$. Por tanto podría decirse que no hubo diferencia entre tener DM con complicaciones oftálmicas y neurológicas y tener DM con complicaciones circulatorias periféricas. O bien puede decirse que los parámetros fueron prácticamente cero.

Interpretación de los parámetros de las edades decenales

Se tiene que $\hat{\theta_{ED1}} \approx \frac{Momio_{ED1}}{Momio_{ED5}} = exp\beta_{ED1}$ =0.12 , entonces la posibilidad de defunción de los afectados con edades entre los 30 y 39 años fue 88 % menor que los afectados de entre 70 a 79.

$\hat{\theta}_{ED4} \approx 0.61$, por tanto los afectados de edades de entre 60 a 69 años tuvieron una posibilidad de defunción aproximadamente 40 % menor que los afectados de edades de entre 70 a79 años.

Con respecto al modelo Poisson, se puede mencionar que el coeficiente estimado de las edades entre 30 a 39 años la tasa de letalidad fue $64\,\%$ menor que la tasa de letalidad que se tuvo para las edades de entre 70 a 79 años. Y puede observarse que va aumentando el porcentaje del incremento de la letalidad conforme las edades aumentan, para todos los modelos. Lo siguiente es una tabla que muestra el incremento para el modelo binomial.

$\hat{\theta}_{ED1} \approx 0{,}10$	90 % menor
$\hat{\theta}_{ED2} \approx 0{,}30$	70 % menor
$\hat{\theta}_{ED3} \approx 0{,}51$	49 % menor
$\hat{\theta}_{ED4} \approx 0{,}60$	40 % menor

Cuadro 7.2: Interpretación de los coeficientes de variable Edad

Es decir a mayor edad tuvieran los afectados mayor la posibilidad de defunción. Y de los afectados de entre 40 a 49 a los afectados de 50 a 59 tuvieron un incremento del casi 21 % de defunciones.

Interpretación de los parámetros de la variable Tipo de Diabetes Mellitus

Se dice, según la OMS que E10 significa DM tipo I. Como $\hat{\theta}_{TDME10} \approx$=0.46, por tanto los afectados con DM tipo I tuvieron una posibilidad de defunción 64 % menor que los afectados con DM Tipo II.

Interpretación de los parámetros de las interacciones de sexo masculino y regiones socioeconómicas

Debe observarse que todas las razones de momios del sexo masculino, dado que pertenecían a diferentes regiones socioeconómicas, comparados con el sexo femenino y la última región socioeconómica (CDMX), son menores a 1. Esto implica que la defunción de DM en todas las demás regiones socioeconómicas tuvo una posibilidad menor que en la CDMX.

7.1.2. Interpretación de los intervalos de confianza de los coeficientes de los parámetros

El cálculo de los intervalos de confianza en R se realizó de la siguiente forma:

```
%%##intervalos de confianza de los exponentes de los parámetros
%%##estimados para los modelos model1b y model2b
%%limiteinf<-exp(coefficients(modelo1b)-1.96*summary(modelo1b)
%%$coefficients[,2])
%%limitesup<-exp(coefficients(modelo1b)+1.96*summary(modelo1b)
%%$coefficients[,2])
%%limiteinf2b<-exp(coefficients(modelo2b)-1.96*summary(modelo2b)
%%$coefficients[,2])
%%limitesup2<-exp(coefficients(modelo2b)+1.96*summary(modelo2b)
%%$coefficients[,2])
%%cbind(limiteinf2b,limitesup2)
%%## alternativa para calcular intervalos para los par\'ametros
%%exp(confint(modelo1b))
```

```
%%exp(confint(modelo2b))
%%
```

Los intervalos de los modelos 1p y 2p se calculan de forma análoga.

Variable	**modelo1b**	**modelo1p**	**modelo2b**	**modelo2p**
(Intercept)	(0.03 , 0.04)	(0.03 , 0.05)	(0.04 , 0.06)	(0.04 , 0.06)
S1	(2.76 , 3.64)	(2.44 , 4.13)	(1.79 , 2.23)	(1.71 , 2.33)
RS1	(3.06 , 4.26)	(2.64 , 4.95)	(1.86 , 2.38)	(1.77 , 2.50)
RS2	(2.03 , 2.70)	(1.78 , 3.08)	(1.52 , 1.91)	(1.45 , 2.00)
RS3	(1.73 , 2.34)	(1.51 , 2.68)	(1.42 , 1.81)	(1.35 , 1.90)
RS4	(3.44 , 4.65)	(3.00 , 5.33)	(1.95 , 2.46)	(1.86 , 2.57)
RS5	(4.20 , 6.12)	(3.54 , 7.26)	(2.09 , 2.71)	(1.98 , 2.86)
RS6	(3.55 , 5.25)	(2.98 , 6.26)	(1.96 , 2.58)	(1.86 , 2.72)
AF20	(4.07 , 5.50)	(3.55 , 6.30)	(2.90 , 3.75)	(2.76 , 3.95)
AF21	(3.52 , 17.49)	(12.04 , 19.65)	(5.66 , 6.95)	(5.44 , 7.23)
AF22	(9.05 , 23.02)	(17.49 , 25.09)	(6.43 , 7.61)	(6.22 , 7.87)
AF23	(0.92 , 3.02)	(0.53 , 5.19)	(0.81 , 2.44)	(0.65 , 3.03)
AF24	(0.23 , 0.97)	(0.12 , 1.86)	(0.22 , 0.88)	(0.16 , 1.15)
ED1	(0.10 , 0.14)	(0.08 , 0.17)	(0.30 , 0.40)	(0.28 , 0.42)
ED2	(0.29 , 0.34)	(0.26 , 0.38)	(0.59 , 0.67)	(0.58 , 0.69)
ED3	(0.48 , 0.56)	(0.44 , 0.60)	(0.77 , 0.84)	(0.75 , 0.85)
ED4	(0.57 , 0.66)	(0.53 , 0.71)	(0.82 , 0.90)	(0.81 , 0.91)
TDME10	(0.37 , 0.56)	(0.31 , 0.67)	(0.55 , 0.76)	(0.51 , 0.81)
S1:RS1	(0.34 , 0.52)	(0.28 , 0.64)	(0.48 , 0.65)	(0.45 , 0.69)
S1:RS2	(0.61 , 0.87)	(0.52 , 1.03)	(0.63 , 0.82)	(0.60 , 0.87)
S1:RS3	(0.38 , 0.55)	(0.32 , 0.65)	(0.52 , 0.69)	(0.49 , 0.73)
S1:RS4	(0.68 , 0.98)	(0.57 , 1.16)	(0.59 , 0.77)	(0.56 , 0.81)
S1:RS5	(0.28 , 0.46)	(0.23 , 0.58)	(0.45 , 0.62)	(0.42 , 0.66)
S1:RS6	(0.50 , 0.83)	(0.40 , 1.04)	(0.53 , 0.73)	(0.50 , 0.78)

Cuadro 7.3: Intervalos de confianza de los coeficientes de los parámetros de los modelos logístico, quasibinomial, poisson y quasipoisson respectivamente.

Al comparar los intervalos de confianza de los coeficientes de los parámetros entre los modelos binomial y quasibinomial, modelo1b y modelo2b respectivamente, se aprecia una diferencia de centésimos entre limites inferiores y limites superiores. Esta diferencia puede explicarse al momento de hacer el cálculo de los intervalos, ya que para ello se requieren los errores estandarizados. Y como ya se vio en el capítulo 6, estos cambian entre los modelos binomial y quasibinomial. De hecho, esta es la única diferencia entre

las estimaciones asociadas a los parámetros. Lo mismo ocurre con los modelos poisson y quasipoisson.

Por otro lado, los intervalos de confianza para los modelos binomiales sirven para ver si los parámetros son significativos. La prueba consiste en que si el 1 se encuentra en el intervalo de confianza, entonces se dice que el parámetro no es significativo.

Puede observarse en el cuadro 7.3 que el intervalo de confianza de AF23 contiene el 1. Y en el caso del modelo quasibinomial, las variables cuyos intervalos de confianza contenían el 1 fueron AF23, AF24, S1:RS2, S1:RS4 y S1:RS6. Lo cual es consistente con la prueba de significancia de los parámetros de 6.3.2.

Capítulo 8

Conclusiones

- Realizar el análisis con los modelos lineales generalizados fue un proceso complejo. Hubo que realizar investigación respecto a que significa un modelo explicativo y sus diferencias con un modelo predictivo. Por lo que hubo que distinguir entre cuando se realizaba análisis entre uno y el otro, pues buscan objetivos distintos.

- Se aprendió también que es muy poco probable que un conjunto de datos de conteos reales no presenten sobredispersión. Por tanto fue requerido el uso de modelos quasi-verosímiles que permitían un parámetro extra para corregir este problema, modificándose así los errores estandarizados a los parámetros asociados.

- También se aprendió que la sobredispersión puede ser causada por falta de variables que pudieran describir los datos. Como en el caso de nuestra base de datos. O también porque la función liga especificada es inadecuada para la variable independiente.

- Con respecto a la enfermedad de Diabetes Mellitus se aprendió que fue la mayor causa de defunciones en 2012 y las complicaciones renales fueron las más relacionadas con las defunciones según el modelo seleccionado.

- Se pudo observar que el análisis de residuos utilizado para modelos lineales puede ser inadecuado para este tipo de modelos. Y podría ser una mejor idea no utilizarlos para evitar confusiones.

- Se aprendió que para el estudio de la defunción por Diabetes Mellitus se requieren datos clínicos especializados para la enfermedad. Por ejemplo datos del contenido de azúcar en la sangre, índice de masa corporal, entre otros.

- También fue difícil encontrar una base de datos para aplicar a estos modelos ya que los supuestos de ausencia de sobredispersión son poco probables en datos reales.

- Se aprendió que pueden modelarse con liga logit datos de conteos agrupados o desagrupados. En el caso de ser desagrupados, el modelo sólo puede ser logístico y no Poisson. Por ello se agruparon los datos utilizados para esta tesis, para realizar análisis Poisson.

- Al comparar los parámetros estimados entre los modelos binomial y Poisson (ver cuadro 7.1), se pudo observar que sus diferencias oscilaban entre décimos y algunas unidades, para la mayoría de las variables. Sin embargo, se obtuvo que, mientras mayor el valor de los parámetros para el modelo binomial, mayor la diferencia entre estos y los parámetros estimados del modelo Poisson correspondientes. Por ejemplo, para el caso de Afección de tipo renal la diferencia fue de aproximadamente 14 unidades. Esa fue la diferencia más grande entre los parámetros estimados de ambos modelos. Cabe mencionar que no tiene sentido hablar de diferencias entre los parámetros estimados de los modelos quasibinomial y quasipoisson ya que son los mismos que para los modelos binomial y Poisson respectivamente. En la sección 6.3.2. se encuentran los códigos correspondientes a dichos modelos.

- Se pudo observar que para el modelo binomial el intervalo de los errores estándares fue (0.03,0.3) aproximadamente. Mientras que para el modelo quasibinomial fue de (0.07,0.69). Así como para el modelo poisson fue de (0.03,0.35) y (0.03,0.4) para el modelo quasipoisson. Si se comparan los intervalos entre sí, puede observarse una diferencia mayor entre el modelo binomial y el quasibinomial. Esto debido a que el modelo binomial tuvo una mayor sobredispersión que el modelo Poisson.

Bibliografía

[1] AGRESTI ALAN , *Categorical data analysis* 2nd ed. New York: John Wiley and Sons, 2002.

[2] FARAWAY J.JULIAN *Extending the linear model with R: Generalized linear, Mixed effects and nonparametric regression Models*, second Edition, CHAPMAN & HALL BOOK, 2016.

[3] AGRESTI ALAN,*Foundations of Linear and Generalized Linear Models* Wiley Series in Probability and Statistics, 2015.

[4] CHRISTENSEN, R. *Log-linear models and logistic regression*, 2nd ed. New York: Springer-Verlag,1997.

[5] HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. *The elements of statistical learning.* New York: Springer, p.223, 2017.

[6] MCCULLAGH, P. AND NELDER, J.*Generalized Linear Models.*(1989) Chapman and Hall/CRC, Washington, DC. http://dx.doi.org/10.1007/978-1-4899-3242-6

Revistas

[7] GALIT SHMUELI,*To explain or to predict,* Institute of Mahematical Statistics, Vol 25, No.3 289-310, 2010.

[8] JORGE ESCOBEDO DE LA PEÑA,M.C. MS.P.M EN C. BEATRIZ RICO-VERDÍN MC.,*Incidencia y letalidad de las complicaciones agudas y crónicas de la diabetes mellitus en México,* Salud Pública Mex 1996, 38: 236-242.

[9] HERNÁNDEZ-ÁVILA, M., PABLO GUTIÉRREZ, J. AND REYNOSO -NOVERÓN, N.,*Diabetes mellitus en México. El estado de la epidemia.* Salud Pública de México, 55(Supl.2), p.129.,2013.

[10] SÁNCHEZ- BARRIGA, JUAN JESÚS*Mortality trends from diabetes mellitus in the seven socioeconomic regions of Mexico,* 2000-2007, Rev Panam Salud Pública; 28(5) 368-367 nov. 2010.

[11] EDWARDS R. JEFFREY, RICHARD P. BAGOZZI*On the Nature and Direction of Relationships Between constructs and Mesures,* Psycological Methods, 2000, Vol. 5. No.2, 155-174.

[12] *Diagnosis and Classification of Diabetes Mellitus.* Diabetes Care, [en línea] 27(Supplement 1), pp.S5-S10. (2003). Disponible en: https://doi.org/10.2337/diacare.27.2007.S5 [último acceso 21 Mar. 2018].

[13] *Estadísticas de mortalidad en México: muertes registradas en el año 2000.* Salud Pública de México, 44(3) (2002).

[14] SALGADO PINEDA, M., FRANCH NADAL, J., PALLAS ELLACURIA, M., ORIOL ZERBE, C., GRAU BARTOMEU, J. AND CASTELLÀ GARCÍA, J. (2001). *Estadísticas y causas de mortalidad en la diabetes tipo 2.* Atención Primaria, [en linea] 27(9), pp.654-657. Disponible en : http://www.elsevier.es/es-revista-atencion- primaria-27-pdf-S0212656701788750-S300 [último acceso 24 Mar. 2018].

[15] SÁNCHEZ-BARRIGA, J.*Mortality trends from diabetes mellitus in the seven socio-economic regions of Mexico, 2000-2007* . [en linea] PAHO/WHO Institutional Repository. Disponible en : http://iris.paho.org/xmlui/handle/123456789/9611 (2018). [último acceso 27 Mar. 2018].

[16] OLAIZ-FERNÁNDEZ, G., ROJAS, R., AGUILAR-SALINAS, C., RAUDA, J. AND VILLALPANDO, S.*Diabetes mellitus en adultos mexicanos: resultados de la Encuesta Nacional de Salud 2000.* Salud Pública de México, 49, pp.s331-s337. (2007).

[17] TAPIA-GRANADOS, J.*Posibilidades, oportunidades, momios: un comentario sobre la traducción del término odds.* Salud Pública de México, 39(1), pp.69-71. (1997).

Páginas electrónicas

[18] Scielo.conicyt.cl. (2018). SciELO - Scientific electronic library en linea. [en línea] Disponible en: https://scielo.conicyt.cl [último acceso 14 Mar. 2018].

[19] Triviño, L. (2018). FORMULACION DE HIPOTESIS. [en linea] Academia.edu. Disponible en: http://www.academia.edu/6982789/FORMULACION_DE_HIPOTESIS [último acceso 14 Mar. 2018]

[20] Explorable.com. (2018). Operationalization - Defining Variables Into Measurable Factors. [en linea] Disponible en: https://explorable.com/operationalization [último acceso 16 Mar. 2018].

[21] American Diabetes Association.(2018). Hiperglucemia. [en linea] Disponible en: http://www.diabetes.org/es/vivir-con-diabetes/tratamiento-y-cuidado/el-control-de-la-glucosa-en-la-sangre/hiperglucemia.html [último acceso 22 Mar. 2018].

[22] Mdsaude.com. (2018). DIAGNÓSTICO DE LA DIABETES MELLITUS MD.Saúde. [en linea] Disponible en :

https://www.mdsaude.com/es/2015/11/diagnostico-de-la-diabetes-mellitus.html [último ingreso 22 Mar. 2018].

[23] Mediavilla Bravo, J. (2018). la diabetes mellitus tipo 2. [en linea] Elsevier.es. Disponible en : http://www.elsevier.es/es-revista-medicina-integral-63-articulo-la-diabetes-mellitus-tipo-2-13025480 [último acceso 22 Mar. 2018].

[24] Asociación Diabetes Madrid. (2018). Diabetes tipo 1 y tipo 2, definición y diferencias. - Asociación Diabetes Madrid. [en linea] Disponible en : https://diabetesmadrid.org/diabetes-tipo-1-tipo-2-definicion-diferencias/ [último acceso 22 Mar. 2018].

[25] Dtc.ucsf.edu. (2018). ¿Qué es la Diabetes tipo 1? :: Diabetes Education en linea. [en linea] Disponible en : https://dtc.ucsf.edu/es/tipos-de-diabetes /diabetes-tipo-1/comprension-de-la-diabetes-tipo-1/que-es-la-diabetes-tipo-1/ [último acceso 22 Mar. 2018].

[26] Dtc.ucsf.edu. (2018). ¿Qué es la Diabetes tipo 2? ::Diabetes Education en linea. [en linea] Disponible en : https:// dtc.ucsf.edu/es/tipos-de-diabetes/diabetes-tipo-2 /comprension-de-la-diabetes-tipo-2/que-es-la-diabetes-tipo-2/ [último acceso 22 Mar. 2018]

[27] Geosalud.com. (2018). Coma diabetico. Qué es? Síntomas y tratamiento. [en linea] Disponible en : https://www.geosalud.com /diabetesmellitus/coma-diabetico.html [último acceso 23 Mar. 2018].

[28] American Diabetes Association. (2018). Cetoacidosis. [en linea] Disponible en : http://www.diabetes.org/es/ vivir-con-diabetes/complicaciones/cetoacidosis.html [último acceso 23 Mar. 2018].

[29] Medlineplus.gov. (2018). Diabetes y enfermedad renal: MedlinePlus enciclopedia médica. [en linea] Disponible en : https://medlineplus.gov/spanish/ency/article/000494.htm [último acceso 23 Mar. 2018].

[30] American Diabetes Association. (2018). Complicaciones de los ojos. [en linea] Disponible en : http://www.diabetes.org/es/vivir-con-diabetes /complicaciones-en-la-vista.html [último acceso 23 Mar. 2018].

[31] National Institute of Diabetes and Digestive and Kidney Diseases. (2018). Neuropatías diabéticas: el daño de los nervios | NIDDK. [en linea] Disponible en : https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/prevenir-problemas/neuropatias-diabeticas [último acceso 24 Mar. 2018].

[32] Gwheartandvascular.org. (2018). Enfermedad Vascular Periférica. [en linea] Disponible en : http://www.gwheartandvascular.org/education/en-espanol/enfermedades/enfermedades_condiciones/enfermedades_condiciones_periferica/ [último acceso 24 Mar. 2018].

[33] Oxford Dictionaries | Español. (2018). letalidad Definición de letalidad en español de Oxford Dictionaries. [en linea] Disponible en : https://es.oxforddictionaries.com/definicion/letalidad [último acceso 24 Mar. 2018].

[34] Sites.google.com. (2018). Estadísticas en México -Diabetes Mellitus. [en linea] Disponible en : https://sites.google.com/a/uabc.edu.mx/infodiabetes/fio/estadisticas-en-mexico [último acceso 26 Mar. 2018].

[35] Anon, (2018). Definiciones y conceptos fundamentales para la calidad en salud. [en linea] Disponible en : http://www.calidad.salud.gob.mx/site/editorial/docs/dgr-editorial_00E.pdf [último acceso 26 Mar. 2018].

[36] Peña, J. and Rico-Verdín, B. (2018). Incidencia y letalidad de las complicaciones agudas y crónicas de la diabetes mellitus en México. [en linea] Saludpublica.mx. Disponible en : http://saludpublica.mx/index.php/spm/article/ view/5930/6720 [último acceso 26 Mar. 2018].

[37] Diccionario.leyderecho.org. (2018). Iatropatogenia. [en linea] Disponible en : http:// diccionario.leyderecho.org/iatropatogenia/ [último acceso 26 Mar. 2018].

[38] Oment.uanl.mx. (2018). [en linea] Disponible en : http://oment.uanl.mx/wp-content/uploads/2016/11/FMidete_Asumiendo-Control-Diabetes-2016.pdf [último acceso 27 Mar. 2018].

[39] Diabetes Mellitus MX. (2018). Los diabéticos tipo 2 todavía se enfrentan a elevados riesgos de muerte - Diabetes Mellitus MX. [en linea] Disponible en : https://www.diabetesmellitus.mx/los-diabeticos-tipo-2-todavia-se-enfrentan-a-elevados-riesgos-muerte/ [último acceso 27 Mar. 2018].

[40] DEFUNCIONES GENERALES (INEGI/SALUD). (2018). BASES DE DATOS SOBRE DEFUNCIONES. [en linea] Disponible en : http://www.dgis.salud.gob.mx/contenidos/basesdedatos/std_defunciones.html [último acceso 28 Mar. 2018].

[41] SECRETARÍA DE SALUD 2000-2015. (2018). BASES DE DATOS SOBRE EGRESOS HOSPITALARIOS. [en linea] Disponible en : http:// www.dgis.salud.gob.mx/contenidos/basesdedatos/std_egresoshospitalarios_gobmx.html [último acceso 28 Mar. 2018].

[42] Albrecht, C.(2018). Contingency Tables and the Chi Square Statistic Interpreting. [en linea] Studylib.net. Disponible en: http://studylib.net/doc/18300816/contingency-tables-and-the-chi-square-statistic-interpreting [último acceso 30 Mar. 2018].

[43] Scribd. (2018). Ji Cuadrado. [en línea] Disponible en: https://es.scribd.com/document/6703611/Ji-Cuadrado [último acceso 30 Mar. 2018].

[44] STAT 504 Analysis of discrete data. (2018). Introduction to Generalized Linear Models. [en línea] Disponible en: https://en líneacourses.science.psu.edu/stat504/node/216 [último acceso 31 Mar. 2018].

[45] Dle.rae.es. (2018). [en línea] Disponible en: http:// dle.rae.es/?id=KHdGTfC [último acceso 31 Mar. 2018].

[46] Rodríguez, G. (2018). WWS 509. [en línea] Data.princeton.edu. Disponible en: http://data.princeton.edu/wws509/ notes/ [último acceso 31 Mar. 2018].

[47] IDRE Stats. (2018). Coding for Categorical Variables in Regression Models | R Learning Modules - IDRE Stats. [en línea] Available at: https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/ [último acceso 1 Apr. 2018].

[48] Support.minitab.com. (2018). Interpret the key results for Interaction Plot - Minitab Express. [en línea] Disponible en: http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-sta [último acceso 2 Apr. 2018].

Notas de clase.

[49] Ramírez, Ricardo."Modelos loglineales de Poisson". Análisis de Regresión. UNAM. Septiembre de 2015 .