



Comparación de Métodos para Estimar Razones de Prevalencia

Dallys Nicol Sinisterra Gutierrez

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Cali, Colombia
2024

Comparación de Métodos para Estimar Razones de Prevalencia

Dallys Nicol Sinisterra Gutierrez

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

Director: Ph.D. Jose Rafael Tovar Cuevas

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Cali, Colombia
2024

Agradecimientos

A mi amada Universidad del Valle, por permitirme alcanzar un escalafón más en mi vida. A los profesores quienes me dictaron clases por cada una de sus enseñanzas durante mi proceso académico. De manera especial, al profesor José Rafael Tovar, Director de este proyecto, quien me ha guiado con su paciencia y conocimientos hacia este objetivo. Agradezco enormemente su forma de enseñar y el apoyo que me brindó en diferentes situaciones que me permitieron formarme a nivel académico e integral. Agradezco a todos y todos los que confiaron en mí y aportaron de una u otra manera al desarrollo de este proyecto.

Dedicatoria

A Dios por darme salud y sabiduría para culminar mi carrera de manera satisfactoria.

A mis padres Juvencio y Maria, mis hermanas Stefanny y Julieth, a mis amigos y allegados por su amor y apoyo incondicional a lo largo de este proceso de formación.

Gracias por alentarme, animarme, limpiar mis lágrimas y estar conmigo. A mis grandes amigos y a quienes estuvieron dándome el apoyo cada día en el que creí que no sería posible. A todos estos, les dedico este logro, gracias por ser parte de el.

*Y sabemos que para los que aman a Dios, todas las cosas cooperan para bien, esto es, para los que son llamados conforme a su propósito.
Romanos 8:28*

Resumen

Este estudio abordó la comparación de métodos para estimar las razones de prevalencia desde una perspectiva analítica y mediante un ejercicio de simulación para diferentes tamaños de muestra y momentos de prevalencia, considerando casos de presencia media, alta y baja de la enfermedad, con el objetivo de facilitar los procesos para seleccionar modelos y reducir la sobreestimación, identificando la naturaleza de los datos. Se realizaron comparaciones con y sin modelos estadísticos, como los modelos Logístico, Log-binomial, Poisson y Poisson Robusto, enfatizando el tipo de estudio epidemiológico asociado con cada uno, desde un enfoque clásico y bayesiano. También se desarrolló una comparación a través de un caso aplicado de pacientes con eventos adversos, visto longitudinalmente considerando los tiempos de admisión y transversalmente con el número total de eventos adversos en el período estudiado.

Palabras clave: Regresión logística, regresión log - binomial, regresión poisson, regresión poisson robusta, razón de prevalencia, sobreestimación.

Abstract

This study addressed the comparison of methods for estimating prevalence ratios from an analytical perspective and through a simulation exercise for different sample sizes and prevalence moments, considering cases of medium, high and low presence of the disease, with the aim of facilitating the processes for selecting models and reducing overestimation, identifying the nature of the data. Comparisons were made for the Logistic, Log-binomial, Poisson and Robust Poisson models, emphasizing the type of epidemiological study associated with each one. The comparison was also developed through an applied case of patients with adverse event, seen longitudinally considering the admission times and transversally with the total number of adverse events in the period studied.

Keywords: Logistic regression, log-binomial regression, poisson regression, robust poisson regression, prevalence ratio, overestimation

Contenido

Resumen	v
1 Introducción y planteamiento del problema	4
1.1 Introducción	4
1.2 Planteamiento del problema	5
1.2.1 Pregunta de investigación	6
1.3 Justificación	6
1.4 Objetivos	7
1.4.1 Objetivo General	7
1.4.2 Objetivos Específicos	7
2 Antecedentes	9
2.1 Antecedentes	9
3 Marco teórico	12
3.1 Inferencia Bayesiana	12
3.1.1 Teorema de Bayes	12
3.2 Diseños epidemiológicos	13
3.3 Medidas de frecuencia	14
3.4 Medidas de asociación	16
3.5 Modelos de Regresión para Variable de respuesta con estructura Binaria . . .	17
3.5.1 Modelo de Regresión Logística	17
3.5.2 Modelo de Regresión Log-Binomial	19
3.5.3 Modelo de Regresión Poisson	20
3.5.4 Modelo de Regresión Poisson Robusto	22
4 Metodología	25
4.1 Estructuras analíticas	27
4.2 Modelo Estadístico	27
4.2.1 Estimación del parámetro, asumiendo dos grupos de comparación: . .	27
4.2.2 Estimaciones Bayesianas asumiendo un experimento Poisson en la recolección de los datos	31

4.2.3	Modelo de Regresión para datos bajo experimento Poisson (Diseño prospectivo)	32
4.2.4	Modelo de regresión cuando los datos son obtenidos usando experimentos binomiales	38
4.3	Estudio de Simulación	43
4.4	Estudio con datos reales	44
5	Resultados	45
5.1	Estudio de Simulación	45
5.1.1	Caso $n = 369$ y $\lambda = 5$ bajo supuesto de homocedástico	45
5.1.2	Caso $n = 384$ y $\lambda = 2$ bajo supuesto de heterocedasticidad	47
5.1.3	Caso $n = 1536$ y $\lambda = 4$ bajo supuesto de heterocedasticidad	49
5.1.4	Caso $n = 6144$ y $\lambda = 8$ bajo supuesto de heterocedasticidad	50
5.2	Caso Aplicado: Unidad de Hemato - Oncología Pediátrica en una clínica nivel IV en Cali	51
6	Conclusiones y recomendaciones	54
6.1	Conclusiones	54
6.1.1	Estudio de simulación	54
6.1.2	Estudio de Datos Reales en la Unidad de Hemato-Oncología Pediátrica	56
6.1.3	Consistencia de los resultados	56
6.2	Recomendaciones	58
7	Anexos	59
7.1	Anexos	59
7.2	Códificación en Rstudio	59

Índice de Tablas

4-1	Esquema de simulación de datos	44
5-1	Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra($n=369$ y $\lambda = 5$).	45
5-2	Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra ($n=384$ y $\lambda = 2$).	47
5-3	Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra ($n=1536$ y $\lambda = 4$).	49
5-4	Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra ($n=6144$ y $\lambda = 8$).	50
5-5	Tabla de contingencia para caso aplicado (Evento adverso Vs Exposición a TMO)	51
5-6	Comparación de modelos ajustados en el caso aplicado, metodología clásica .	52
5-7	Comparación de modelos ajustados en caso aplicado, metodología Bayesiana	52

Índice de figuras

3-1	Esquemas conceptuales de diseño y análisis para un estudio transversal Fajardo-Gutiérrez (2017)	16
4-1	Esquema conceptual para los escenarios metodológicos	26
4-2	Esquema conceptual de los diseños para obtener las observaciones en campo y los modelos estadísticos asociados al análisis de los datos	27

1 Introducción y planteamiento del problema

1.1. Introducción

Para la comprensión de la asociación entre la exposición y los resultados de interés en epidemiología, se requiere una estimación precisa de las razones de prevalencia (RP). Esta medida epidemiológica expresa la proporción de individuos en una población que presentan una condición o enfermedad en un momento específico en función de una exposición de interés. Esta métrica es comúnmente usada en estudios transversales y longitudinales, ya que permite evaluar la fuerza de la asociación entre una exposición (como un factor de riesgo) y un resultado de salud (como una enfermedad o condición). Una Razón de prevalencia mayor que 1 indica que la exposición está asociada con una mayor prevalencia del resultado, mientras que una RP menor que 1 sugiere una asociación protectora.

Tradicionalmente, se ha utilizado la regresión logística para ajustar el posible efecto de variables de confusión y estimar odds ratios (OR), aproximaciones a las RP. Sin embargo, su utilidad ha sido cuestionada en situaciones de alta prevalencia o riesgo, donde las OR pueden sobreestimar la verdadera asociación, generando la necesidad de explorar métodos alternativos. Este estudio investiga y compara métodos clásicos y bayesianos, como la Regresión Logística, Log-binomial, Poisson y Poisson robusta, para ofrecer estimaciones precisas de la RP en estudios longitudinales y de corte transversal, además de la obtención de las estimaciones a través de un ejercicio de conteo, en distintos momentos del tiempo.

A través de un desarrollo metodológico segmentado en tres fases, la primera fase, integrada por la construcción y búsqueda de las formas analíticas de cada modelo, desde su enfoque clásico y bayesiano. La segunda, un esquema de simulaciones ajustado a cuatro tamaños de muestra y distintos valores de prevalencias. Finalmente, una tercera fase integrada por el análisis de un caso aplicado en una Unidad de Hemato - Oncología pediátrica en la ciudad de Santiago de Cali.

Los resultados revelaron que tanto el modelo log- binomial como el Poisson proporcionan resultados similares en términos de RP desde un enfoque clásico. El modelo Poisson robusto tiende a sobrestimar más que el Poisson estándar. A pesar de ser el modelo con mayor confiabilidad al acercarse a los datos reales, el modelo Log-binomial muestra limitaciones en su aplicabilidad en ciertos contextos epidemiológicos. Sin embargo, estos se pueden solventar a través del enfoque bayesiano.

Se destaca la importancia de elegir el modelo adecuado según el diseño del estudio, la prevalencia de la enfermedad y el tamaño de la muestra, así como considerar el enfoque bayesiano para abordar las limitaciones de los modelos clásicos.

Los resultados obtenidos detallan diferencias en la precisión y confiabilidad de los diferentes modelos, tanto en el enfoque clásico como en el bayesiano, resaltando la importancia de una evaluación cuidadosa en la elección del método apropiado.

1.2. Planteamiento del problema

Desde una perspectiva epidemiológica, la precisa estimación de las Razones de Prevalencia (RP) desempeña un papel fundamental en la comprensión de las asociaciones entre la exposición y los resultados de interés. Una RP bien estimada es crucial al permitir cuantificar de manera directa la proporción de la población que presenta una enfermedad o condición en función de la exposición, lo que facilita la identificación de factores de riesgo y la planificación de intervenciones. Además, permite evitar interpretaciones erróneas sobre la relación entre exposición y resultado, particularmente en la evaluación de la efectividad de sucesos de exposición.

Aunque la regresión logística ha sido tradicionalmente utilizada para ajustarse a posibles variables de confusión y estimar odds ratios (OR), que posteriormente se interpretan como aproximaciones a las Razones de Prevalencia (RP), se ha cuestionado esta aproximación, especialmente en situaciones de alta prevalencia o riesgo. En estos casos, las OR pueden sobrestimar la verdadera asociación, lo que subraya la necesidad de métodos alternativos que proporcionen estimaciones más precisas y directamente interpretables de la RP.

Como señalan Petersen and Deddens (2008), esta problemática ha generado extensas discusiones sobre el uso de las odds ratios y el riesgo relativo cuando se ajusta un modelo estadístico diferente al logístico. Este mismo tema ha sido identificado por otros autores en

casos donde se refieren a otros tipos de modelos que pueden carecer de validez externa al generar estimaciones que no consideran la naturaleza de los datos, lo que ocasiona problemas en la interpretación de los resultados.

En este contexto, distintos autores como Coutinho et al. (2008) y Blizzard and Hosmer (2006) han desarrollado las técnicas para abordar este problema haciendo uso de modelos de regresión, los cuales a partir del contexto pueden presentar limitaciones en su formulación.

El objetivo de este trabajo de grado es abordar el problema metodológico a través de un enfoque bayesiano, comparando los resultados obtenidos en este ejercicio con los publicados por otros autores en el contexto de la comparación de estos modelos desde el enfoque clásico. Estos métodos permitirán estimar con precisión la razón de prevalencia en escenarios desafiantes de alta y baja prevalencia, considerando la influencia de covariables y contrastando los resultados con los datos observados. Así, se busca identificar cuál de estos métodos ofrece estimaciones más precisas y menos sesgadas de la razón de prevalencia en diversos contextos, especialmente en estudios transversales donde las enfermedades o lesiones no son raras, mejorando la precisión y validez de las conclusiones en investigaciones epidemiológicas y estableciendo estándares más consistentes y adecuados para la estimación y reporte de medidas de asociación.

1.2.1. Pregunta de investigación

¿Qué tan comparable es el estimador de Razón de Prevalencias Bayesiano, con los estimadores clásicos comúnmente usados?

1.3. Justificación

En numerosas ocasiones, resulta más conveniente modelar y estimar razones de prevalencia en lugar de razones de odds en estudios transversales, especialmente cuando las enfermedades o lesiones no son poco frecuentes. Sin embargo, los problemas asociados con los métodos existentes para modelar razones de prevalencia, como la falta de convergencia, la sobreestimación de errores estándar y la extrapolación de fórmulas univariadas simples a modelos multivariados, han generado la necesidad de explorar y comparar métodos alternativos para obtener estimaciones estadísticas más precisas. Esto es crucial para la toma de decisiones y criterios para el seguimiento y la interpretación de la información, garantizando una lectura adecuada del momento epidemiológico estudiado, el cual está intrínsecamente vinculado con el diseño del estudio a realizar. Petersen and Daddens (2008)

En este sentido, se hace necesaria la búsqueda de alternativas que permitan optimizar la forma de obtención de los resultados, de modo que si se trabaja con los datos, sin modelo de regresión (en grupos separados), el enfoque bayesiano permitirá obtener los resultados de manera ágil, para cualquier tamaño de muestra, haciendo uso de la información externa, incluyendo distribuciones a priori, entregando estimaciones robustas y regiones de credibilidad estrechas dentro de los límites del espacio paramétrico adecuado, aun cuando la información externa sea insuficiente o vaga.

En este mismo sentido, al usar modelos de regresión, el enlace logbinomial presenta problemas cuando la muestra es pequeña o, cuando la cantidad de éxitos tiende a cero o, cuando se tiene ambas características en el archivo de datos. Los modelos Poisson por su parte, pueden presentar sobredispersión y, aunque el modelo Poisson robusto ajusta el problema de la varianza, las estimaciones pueden ser insesgadas. En este aspecto, el enfoque Bayesiano mejora la robustez de los resultados, permitiendo ajustar los modelos con información externa (a priori) sobre los parámetros beta, lo cual implica un ajuste en la estimación de la media posterior del modelo. Martinez et al. (2017)

En consecuencia, este estudio conducirá a resultados orientados en la comparación y selección del método con el ajuste más apropiado según las características de los datos de estudio, comparando los resultados obtenidos con y sin modelos. Se utilizará un soporte matemático estadístico que permita la clasificación e identificación de posibles escenarios para superar los problemas presentes en las estimaciones clásicas.

1.4. Objetivos

1.4.1. Objetivo General

Comparar el desempeño estadístico de cuatro métodos utilizados comúnmente para estimar razones de prevalencia con los resultados obtenidos al utilizar métodos de inferencia bayesiana.

1.4.2. Objetivos Específicos

- Establecer las formas analíticas de los modelos utilizados comúnmente y obtener la estructura de las estimaciones de Razón de Prevalencia.

- Desarrollar un estudio de simulación, considerando escenarios de alta y baja prevalencia con diferentes tamaños de muestra.
- Comparar el desempeño estadístico de las estimaciones obtenidas a partir de un caso aplicado.

2 Antecedentes

2.1. Antecedentes

Coutinho et al. (2008), realizaron una revisión literaria sobre estudios de corte transversal, cuyo objetivo fue comparar empíricamente las regresiones de Cox, log-binomial, Poisson y logística para estimar razones de prevalencia en estudios de corte transversal para una base poblacional de 2.072 adultos mayores diagnosticados con demencia, trastornos mentales comunes y autopercepción de salud deficiente en la ciudad de São Paulo, incluyendo así, desenlaces de prevalencia baja, intermedia y alta.

Se utilizaron variables de confusión con dos o más categorías o valores continuos. Los puntos de referencia para las estimaciones de punto e intervalo de las razones de prevalencia (RP) se determinaron a través del método de estratificación de Mantel-Haenszel. Se realizaron cálculos de estimaciones ajustadas empleando regresiones de Cox y Poisson con varianza robusta, así como mediante regresión log-binomial. Los odds ratios (OR) brutos y ajustados se derivaron mediante el uso de regresión logística.

Las estimaciones de punto e intervalo derivadas de las regresiones de Cox y Poisson fueron similares a las obtenidas a través de la estratificación de Mantel-Haenszel, sin importar la prevalencia del resultado y las covariables presentes en el modelo.

El modelo log-binomial experimentó dificultades de convergencia cuando el resultado tenía una prevalencia elevada y existía una covariable continua en el modelo. La regresión logística generó estimaciones de punto e intervalo superiores a las obtenidas con los demás métodos, especialmente para los resultados con prevalencias iniciales más altas. En caso de interpretarse como estimaciones de razón de prevalencia (RP), los odds ratios (OR) exagerarían las asociaciones para resultados con prevalencia inicial baja, intermedia y alta.

Así pues, las estimaciones puntuales arrojadas por el modelo log-binomial, pueden considerarse más adecuadas a las generadas por otros modelos, no obstante, puede enfrentar

problemas de convergencia al utilizar covariables de forma continua. Para este caso, se obtuvieron mejores resultados en los modelos de Cox y Poisson Robusto, al ofrecer estimaciones imparcia les.

Petersen and Deddens (2008) realizaron una comparación de dos métodos usados para estimar razones de prevalencia, el método Poisson Robusto y el método log-binomial, a través de datos simulados y datos reales. Se confronta el método Poisson Robusto, el cual emplea la distribución de Poisson y un estimador de varianza tipo sándwich, con el método log-binomial, que utiliza la distribución binomial para generar estimaciones basadas en la máxima verosimilitud. Las simulaciones replicaron algunas de las realizadas por otros autores.

Específicamente, se llevaron a cabo para la situación de una covariable continua, X , distribuida uniforme de 0 a 10. En cada valor de X , se generó aleatoriamente un valor de Y a partir de una distribución binomial con un tamaño de muestra de 1 y una prevalencia de $\exp^{(\beta_0 + \beta_1 x)}$. Se eligieron tres valores para β_1 , se fijó un tamaño de muestra $n = 100$ para cada simulación porque se consideró lo suficientemente grande para que se cumplieran las propiedades de muestras grandes, pero no tan grande como para que ambos métodos tuvieran una potencia demasiado alta para la comparación. Los datos se replicaron 1000 veces para cada simulación. Para todas las simulaciones, además de las estimaciones de los métodos log-binomial y Poisson Robusto, se obtuvieron estimaciones exactas de máxima verosimilitud (incluso cuando no tenían convergencia en los datos originales).

Para prevalencias grandes y un tamaño de muestra moderado, el método Poisson Robusto genera estimaciones de las razones de prevalencia con menor sesgo que el método log-binomial. No obstante, en situaciones de prevalencias bajas y un tamaño de muestra reducido, el método log-binomial ofrece estimaciones ligeramente menos sesgadas que el método Poisson Robusto. En la mayoría de los casos, el método log-binomial muestra una potencia superior y errores estándar más reducidos en comparación con el método Poisson Robusto.

Aunque el método Poisson Robusto a menudo brinda estimaciones razonables de la razón de prevalencia y es muy fácil de usar, el método log-binomial resulta en menos sesgo en la mayoría de las situaciones comunes, errores estándar más pequeños y, a diferencia del Poisson Robusto, siempre produce prevalencias entre cero y uno.

Martinez et al. (2017), realizaron una revisión literaria de artículos con diseños transversales

para evaluar el método estadístico utilizado y la interpretación de la medida estimada de asociación; posteriormente, ilustraron el uso de métodos estadísticos alternativos que estiman directamente la Razón de Prevalencia (RP) a través de una revisión de los métodos estadísticos informados en estudios transversales con resultados binarios en medicina veterinaria, evaluando la adecuación de la interpretación de la medida de asociación estimada. Se realizaron pruebas con cuatro modelos alternativos para la regresión logística que estiman directamente la RP, utilizando un conjunto de datos propio de un estudio transversal sobre el virus de la diarrea viral bovina.

Se utilizó un conjunto de datos de un estudio en el sur de Brasil para ejemplificar métodos que estiman directamente la RP. El estudio buscaba determinar la prevalencia de anticuerpos contra el virus de la diarrea viral bovina (BVDV) en la leche de tanque a granel. Se tomaron muestras al azar de una población de 81307 hatos lecheros en el estado de Río Grande do Sul, Brasil, recopilando datos de 388 hatos.

Se ajustó un modelo de Poisson robusto multivariable empleando las variables cuyo resultado fue significativo con el BVDV ($p < 0,05$). También se incorporaron en otros cuatro modelos: regresión log-binomial, regresión logística, regresión de Poisson y un enfoque bayesiano para la regresión log-binomial. Todos los modelos se ajustaron al mismo conjunto de datos utilizando el software estadístico SAS, excepto para la regresión log-binomial con enfoque bayesiano.

La prevalencia estimada de BVDV en los datos utilizados fue del 24 %. En la última regresión de Poisson robusta se identificaron tres variables significativamente asociadas con la seropositividad al BVDV ($p < 0,05$). En relación con los cambios observados entre los métodos en comparación con la regresión log-binomial, se observa que las estimaciones puntuales en la regresión logística mostraron las mayores diferencias, oscilando entre 25.7 y 69.9 %, mientras que la regresión de Poisson y la regresión de Poisson robusta, que produjeron la misma estimación puntual esperada, tuvieron una diferencia que osciló entre 1.4 y 16.2 %. El enfoque bayesiano para la regresión log-binomial tuvo la diferencia más baja, que osciló entre 3.5 y 6.8 %.

En todos los métodos, la regresión logística presentó intervalos más amplios, el rango de los intervalos de la regresión de Poisson robusta fue menor que los calculados por la regresión de Poisson, la regresión log-binomial tuvo los intervalos más estrechos entre todos los métodos.

3 Marco teórico

El presente trabajo de grado aborda un problema metodológico relacionado con la selección del modelo más adecuado para estimar la razón de prevalencia y comparar la prevalencia de una enfermedad o condición entre dos grupos distintos, típicamente un grupo expuesto y un grupo no expuesto a un factor de interés. Este estudio incluye un análisis de las diferentes formas analíticas utilizadas en el proceso de estimación, así como una revisión de los conceptos clave necesarios para comprender cada modelo empleado.

3.1. Inferencia Bayesiana

La inferencia bayesiana, de acuerdo con Páez et al. (2011), es el proceso de razonamiento para ajustar un modelo de probabilidad como medida de creencia que permita simular las diferentes condiciones de los datos estudiados, cuando no se conoce la veracidad de las hipótesis. Observando la evidencia y calculando el valor estimado de acuerdo con la creencia planteada en la hipótesis.

En el caso de la inferencia bayesiana, a diferencia de la clásica, todas las conclusiones están determinadas bajo el uso del Teorema de Bayes, el cual permite el ajuste de probabilidades a priori, sobre eventos no necesariamente de naturaleza aleatoria, permitiendo la incorporación de información externa de manera subjetiva. Dichos fundamentos, están condicionados sobre los valores observados de Y , a partir de probabilidades condicionadas definidas como $P(\theta|y)$, los que implícitamente, también se encuentran condicionados por la covariable X .

3.1.1. Teorema de Bayes

Sea, $y = (y_1, y_2, \dots, y_n)$, un vector de n observaciones con distribución de probabilidad asociada, $f(y|\theta)$, donde y depende de los k parámetros θ , donde, $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, (un vector de parámetros desconocido, pero con información previa), por tanto, se puede asumir

que tiene una distribución de probabilidad a priori denominada $\pi(\theta)$ Box and Tiao (2011).

Considerando los datos observados de y , la distribución condicional de θ , se puede definir como:

$$f(\theta | y) = \frac{L(\theta | y) \pi(\theta)}{\int_{\Theta} L(\theta | y) \pi(\theta) d\theta} \quad (3-1)$$

Donde Θ es el espacio paramétrico y,

$$E(f(y | \theta)) = c^{-1} = \int_{\Theta} L(\theta | y) \pi(\theta) d\theta \quad (3-2)$$

Donde la integral está definida sobre todo el espacio paramétrico θ , y c , es la constante de normalización para asegurar que la distribución posterior se integre de forma completa y resulte en uno.

3.2. Diseños epidemiológicos

De acuerdo con Hernández-Avila et al. (2000), se han establecido criterios de clasificación que orientan la construcción y la diferenciación de enfoques en la investigación. Bajo este contexto, los estudios epidemiológicos se han dividido en tres categorías:

1. Experimentales: En este tipo de estudios, se ejerce control sobre la exposición y se llevan a cabo asignaciones de manera aleatoria.
2. Pseudo-experimentales: Se controla la exposición, pero no se realizan asignaciones de manera aleatoria.
3. No experimentales: No hay intervención ni control por parte del investigador en el momento de la ocurrencia del fenómeno.

Dentro de estos enfoques, se incluyen las mediciones de la ocurrencia del evento, que permiten distinguir entre diseños longitudinales (con dos o más mediciones, una basal y otra subsecuente) y diseños transversales (con una sola determinación y evaluación concurrente después del evento de interés).

Los estudios transversales, en particular, posibilitan la selección de la muestra después de la exposición, explorando la presencia o ausencia de los sujetos tras el evento determinístico. Este enfoque escapa al control directo del investigador, brindando estimaciones valiosas para estudiar la prevalencia de enfermedades en la población.

3.3. Medidas de frecuencia

Prevalencia

De acuerdo con la definición planteada con Ferrer and del Prado González (2013), la prevalencia se define como la proporción del grupo de individuos que presentaron el evento o característica clínica en determinado momento del tiempo (t), representado a través de una población previamente definida, con individuos con y sin el problema bajo estudio. En otras palabras, representan la medida de frecuencia de la enfermedad, en estudios de corte transversal. En este sentido, existen dos tipos de prevalencia.

- **Prevalencia puntual** Se determina en el momento del muestreo para cada individuo, no necesariamente el mismo momento para todos.
- **Prevalencia de periodo** Hace referencia a los casos presentes en cualquier momento en un periodo específico.

$$\text{Prevalencia (Punto)} = \frac{\text{Nº de casos presentes de enfermedad}}{\text{Total de la población estudiada}} \quad (3-3)$$

$$\text{Prevalencia (Periodo)} = \frac{\text{Nº de casos nuevos} + \text{Nº de casos presentes de enfermedad}}{\text{Total de la población estudiada}} \quad (3-4)$$

Incidencia

La incidencia en estudios epidemiológicos, centrados en la investigación causal o la evaluación de medidas preventivas, se enfoca en medir el flujo entre la salud y la enfermedad, es decir, la aparición de casos nuevos. Esta medida expresa el cambio de estado y se manifiesta como la frecuencia de nuevos eventos. A diferencia de los estudios de prevalencia, la incidencia parte de poblaciones susceptibles libres del evento y observa la presentación de casos nuevos a lo largo de un periodo de seguimiento. Además de indicar el volumen final de casos nuevos, los resultados permiten establecer relaciones de causa-efecto entre características poblacionales

específicas y enfermedades. La incidencia de una enfermedad se mide de dos formas: mediante la tasa de incidencia, que se basa en el tiempo-persona y mide la velocidad de ocurrencia de la enfermedad, y mediante la incidencia acumulada, que se basa en el número de personas en riesgo y mide la probabilidad de que un individuo desarrolle el evento en estudio, también conocida como riesgo (Moreno-Altamirano et al., 2000).

Riesgo relativo

El Riesgo Relativo se define como la razón entre las tasas de incidencia de una enfermedad en individuos expuestos y no expuestos. Esta medida cuantifica la fuerza de la asociación entre la exposición a un factor de riesgo y la enfermedad, indicando cuántas veces es más probable que ocurra el evento en los expuestos en comparación con los no expuestos.

En términos más simples, el Riesgo Relativo es un cociente de probabilidades, representando la probabilidad de desarrollar la enfermedad en los expuestos en relación con la probabilidad en los no expuestos. Se expresa comúnmente como la tasa de incidencia en expuestos dividida por la tasa de incidencia en no expuestos ($RR = \text{Tasa incidencia en expuestos} / \text{Tasa de incidencia en no expuestos}$). Este cálculo revela cuántas veces es más probable que la enfermedad se desarrolle en el grupo expuesto en comparación con el grupo no expuesto Colimon (1990).

$$RR = \frac{\text{Incidencia en expuestos}}{\text{Incidencia en No expuestos}} = \frac{P(Y = 1 | X = 1)}{P(Y = 1 | X = 0)} \quad (3-5)$$

Este cociente sólo puede tomar valores positivos y, si la exposición es un factor de riesgo de la enfermedad, el valor que arroje será superior a 1, mientras que si la exposición es un factor protector, su valor será inferior a 1. En casos donde sea estrictamente igual a 1, se entiende que no existe asociación entre la exposición y la enfermedad, ya que la incidencia de expuestos y no expuestos es la misma.

Relación entre incidencia y prevalencia

Sea P la proporción de personas enfermas en una población en un momento determinado, interpretada como la prevalencia instantánea. Esta prevalencia depende de la velocidad de contagio en la población, expresada como la incidencia dividida por los casos por unidad de observación, y del período medio de permanencia de la enfermedad (D). Cuando se tienen valores estables para la incidencia y la duración de la enfermedad, se puede expresar la relación mediante la ecuación:

$$P = \frac{Ix D}{1 + Ix D} \quad (3-6)$$

Dado que el producto de una proporción (menor a 1) y una tasa (mayor a 1) generalmente es menor a uno, el denominador puede considerarse igual a uno, simplificando la ecuación a:

$$P = I \times D \quad (3-7)$$

Esto significa que la prevalencia es aproximadamente igual al producto de la incidencia y la duración media de la enfermedad, teniendo en cuenta la dimensionalidad y las unidades de cada variable Granados (1995).

3.4. Medidas de asociación

Las medidas de asociación dependen del estudio con el que se esté trabajando, y se obtienen al llevar a cabo un estudio transversal analítico, donde su principal característica es obtener una muestra aleatoria y probabilística cuyo objetivo es conocer la prevalencia de una enfermedad y sus factores asociados.

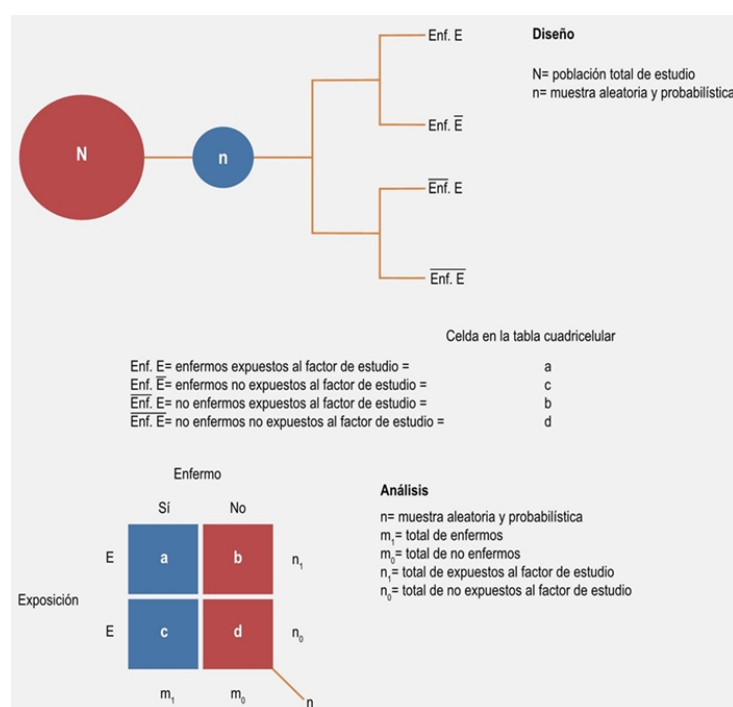


Figura 3-1: Esquemas conceptuales de diseño y análisis para un estudio transversal Fajardo-Gutiérrez (2017)

Las principales son:

1. Razón de prevalencia: La razón de prevalencia es una medida epidemiológica que describe la fuerza de asociación entre una exposición y una enfermedad en estudios transversales. Se calcula como el cociente entre la prevalencia de la enfermedad en el grupo expuesto y la prevalencia de la enfermedad en el grupo no expuesto en un momento específico. Matemáticamente, se expresa como:

$$\text{Razón de prevalencia} = \frac{\text{Prevalencia en el grupo expuesto}}{\text{Prevalencia en el grupo no expuesto}} \quad (3-8)$$

Este indicador proporciona información sobre cuántas veces es más probable que ocurra la enfermedad en el grupo expuesto en comparación con el grupo no expuesto en un punto particular en el tiempo. La razón de prevalencia es especialmente útil en situaciones donde la enfermedad estudiada no es rara y se utiliza comúnmente en investigaciones epidemiológicas de enfermedades crónicas Fajardo-Gutiérrez (2017).

2. Odds Ratio: Es un parámetro estadístico utilizado en estudios epidemiológicos para medir la fuerza de asociación entre una exposición y un resultado, especialmente en estudios de casos y controles. Se define como la razón de las odds de un evento entre los expuestos en comparación con los no expuestos, se expresa de la siguiente manera:

$$\text{Odds Ratio} = \frac{\text{Odds en el grupo expuesto}}{\text{Odds en el grupo no expuesto}} \quad (3-9)$$

El Odds Ratio representa la probabilidad de que ocurra un evento dividida por la probabilidad de que no ocurra. En el contexto del OR, se comparan las odds de que ocurra un evento entre los individuos expuestos con las odds entre los individuos no expuestos. Un OR de 1 indica que no hay asociación entre la exposición y el resultado, mientras que un OR mayor que 1 sugiere una asociación positiva, y un OR menor que 1 sugiere una asociación negativa. El OR es especialmente útil cuando la prevalencia del resultado es baja y no se puede calcular la razón de prevalencia de manera directa Tenny and Hoffman (2017).

3.5. Modelos de Regresión para Variable de respuesta con estructura Binaria

3.5.1. Modelo de Regresión Logística

La Regresión Logística, según Bonilla Castillo (2019), constituye un modelo específico diseñado para explicar y predecir la probabilidad de ocurrencia de un resultado categórico (variable dependiente) en función de otras variables independientes o covariables. Este modelo encuentra una aplicación más frecuente en los campos de las ciencias humanas y médicas

debido a su eficaz capacidad para analizar datos en el ámbito de la epidemiología clínica. Su utilidad radica en la capacidad para identificar la mejor adaptación del modelo, permitiendo así la explicación de las relaciones existentes entre las variables dependientes e independientes. Una característica distintiva y relevante de este modelo es que la variable de respuesta o dependiente es binaria, simplificando de esta manera el uso del modelo como la estimación de los parámetros.

En este contexto, se hace posible definir el modelo de regresión logística simple y múltiple, considerando el número de variables explicativas que tendrá el modelo.

1. Modelo de Regresión Logística Simple:

Formulación del modelo: Sea Y , una variable de respuesta binaria codificada 0,1 y X como variable explicativa, entonces:

$$E[Y|X = x] = P[Y = 1|X = x] = \pi(x) \quad (3-10)$$

Se obtiene la forma específica del modelo planteado por Hosmer Jr et al. (2013),

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (3-11)$$

Se realiza la transformación logit de $\pi(x)$, lo que permite obtener propiedades deseables del modelo de regresión lineal, al ser lineal en sus parámetros, los cuales pueden ser continuos y tener un rango de $(-\infty, +\infty)$ que depende del rango de valores de X , lo que permite expresar la ecuación anterior como:

$$\text{logit}(\pi(x)) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x \quad (3-12)$$

De acuerdo con esta definición, este modelo hace parte de los modelos lineales generalizados, donde la función logit hace el papel de función de enlace.

Medidas de asociación

- **Odds Ratio:** El término de ventajas o de odds para este modelo significa el incremento β_1 veces de la respuesta 1 por cada unidad de X , es decir:

$$\frac{\pi(x)}{1 - \pi(x)} = \frac{e^{(\beta_0 + \beta_1 x)}}{1 - e^{(\beta_0 + \beta_1 x)}} \quad (3-13)$$

3.5.2. Modelo de Regresión Log-Binomial

Blizzard and Hosmer (2006), en “Parameter Estimation and Goodness-of-Fit in Log Binomial Regression” investigaron acerca del modelo de Regresión Log-binomial, e identificaron su uso para la estimación de riesgos a partir de variables de confusión, especialmente en casos de enfermedades comunes en la población. Este modelo, siendo un tipo de modelo lineal generalizado, vincula la variable de respuesta a una distribución binomial, lo que implica una respuesta binaria, mediante una función de enlace logarítmico.

Suponga que se tienen n observaciones independientes asociadas a una variable de respuesta Y , codificadas como variables binarias (1,0), y p covariables no constantes. Se definen los datos observados como, $(y_i, \mathbf{x}_i); i = 1, 2, \dots, n$, para $\mathbf{x}_i' = (x_0, x_{i1}, x_{i2}, \dots, x_{ip})$, con $x_0 = 1$.

Con covariables que pueden estar asociadas o resultantes de otras combinaciones. Este modelo de regresión usará la transformación del logaritmo de una proporción log con rango $\pi_x \in [0, 1]$.

Sin embargo, puede ocurrir que los valores de $\exp(\beta'X)$, estén por fuera de los límites al cumplirse $\exp(\beta'X) > 0$. En consecuencia, las probabilidades predictivas podrán encontrarse por fuera de los intervalos $[0, 1]$ imposibilitando la estimación apropiada de los parámetros (Melis and Cadarso-Suárez, 2017).

Formulación del modelo: Sea, $P(Y = 1|x) = e^{x'\beta}$, la probabilidad condicional del modelo asociado, se define el predictor lineal como:

$$x'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (3-14)$$

Lo que de forma completa permite expresar el modelo como:

$$\log(\pi_x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (3-15)$$

Medidas de asociación

- **Risk Ratio:** Sea la razón de riesgo, la exponenciación del producto de los coeficientes por la diferencia de los valores de las covariables. Entonces, en un modelo simple

de única variable codificada (0,1), se expresa como:

$$\frac{P(Y = 1|x_1)}{P(Y = 1|x_0)} = e^{(x_1 - x_0)' \beta} \quad (3-16)$$

Aplicando la propiedad para el modelo de regresión log binomial sobre la no simetría de sus coeficientes con respecto a la razón de riesgo, se tiene que:

$$\frac{P(Y = 1|x = 1)}{P(Y = 1|x = 0)} = e^{\beta} \quad (3-17)$$

Este modelo, a diferencia del modelo Logístico donde se tiene que el odds ratio para $Y = 0$ es la inversa del odds ratio para $Y = 1$, no es simétrico con respecto a la probabilidad de 0.5 lo que impide invertir el calculo:

$$\frac{P(Y = 0|x = 1)}{P(Y = 0|x = 0)} \neq e^{-\beta} \quad (3-18)$$

En otras palabras, cuando $X_i = 1$ es ajustado para las demás covariables, su cálculo se realiza de la siguiente manera:

$$RR_i, PR_i = \frac{P(Y = 1|X_1, \dots, X_i = 1, \dots, X_p)}{P(Y = 1|X_1, \dots, X_i = 0, \dots, X_p)} \quad (3-19)$$

3.5.3. Modelo de Regresión Poisson

La distribución de Poisson se emplea para modelar la cantidad de eventos que ocurren en un período de tiempo específico, describiendo la función de masa de probabilidad:

$$P(Y = y) = \frac{e^{-\mu} \mu^y}{y!} \quad (3-20)$$

Para $y = 0, 1, 2, \dots$, y $\mu > 0$, con media y varianza provenientes de la propiedad de equidispersión: $E(Y) = \mu = Var(Y)$.

Siendo una distribución perteneciente a la familia exponencial, el modelo de regresión asociado, pertenece a la familia de los modelos lineales generalizados (Gómez Gallego, 2019).

Formulación del modelo: El modelo de Regresión Poisson, introduce las covariables

en la ecuación de la media de forma log lineal, como:

$$E[y_i] = \mu_i = e^{x_i'\beta} \quad (3-21)$$

Para $i = 1, 2, \dots, n$. De modo que:

$$\log(\mu_i) = \eta_i = x_i'\beta = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \quad (3-22)$$

Medidas de asociación

Este modelo permite predecir una tasa a partir de la presencia o ausencia de la exposición. En su forma simple se tiene un modelo con una variable independiente definida como:

$$\log(\mu_i) = \beta_0 + \beta_1 X \quad (3-23)$$

Donde si, X adopta el valor de cero en el grupo no expuestos y uno en el grupo de expuestos, es posible estimar la tasa predicha para cada uno de los grupos de exposición como:

■ **Razón de tasas:**

- Tasa de no expuestos:

$$X = 0 \rightarrow \beta_0 + \beta_1 0 = e^{\beta_0} \quad (3-24)$$

- Tasa de expuestos:

$$X = 1 \rightarrow \beta_0 + \beta_1 1 = e^{(\beta_0 + \beta_1)} \quad (3-25)$$

En consecuencia, la razón de tasas puede estimarse como una razón de funciones, donde en el numerador se tiene la tasa del grupo en exposición y en el denominador la tasa del grupo de referencia o no expuesto. Diaz-Quijano (n.d.)

$$RT_{x_1} = \frac{T_{expuestos}(X_1 = 1)}{T_{noexpuestos}(X_1 = 0)} = \frac{e^{\beta_0} e^{\beta_1(X_1=1)}}{e^{\beta_0} e^{\beta_1(X_1=0)}} \quad (3-26)$$

Simplificando la ecuación anterior se tiene:

$$RT_{x_1} = e^{\beta_1} \quad (3-27)$$

3.5.4. Modelo de Regresión Poisson Robusto

El modelo Poisson de varianza robusta es una extensión del modelo de Regresión Poisson que aborda la posible heterocedasticidad, es decir, la variabilidad no constante de los errores en el modelo original. En un modelo de regresión Poisson estándar, se asume que la varianza de los errores es igual a la media, lo cual es característico de la distribución Poisson .

En situaciones donde la igualdad de varianza no se cumple, el modelo Poisson de varianza robusta se utiliza un enfoque que permite a la varianza ser una función más general de la media, mediante el uso de un estimador de varianza robusto que es menos sensible a los valores atípicos y a la heterocedasticidad. El método de varianza robusta se basa en la matriz de información observada y utiliza técnicas robustas para calcular estimaciones de los parámetros y errores estándar que son menos propensos a ser influenciados por valores atípicos o incumplimientos de supuestos. Haciendo que el modelo sea más flexible y capaz de manejar de manera efectiva datos que pueden tener variabilidad no constante.

Formulación del modelo:

Consideremos el caso en que las observaciones (y_i, x_i) , $i = 1, \dots, n$, son pares de observaciones independientes, idénticamente distribuidas.

El modelo estándar Poisson asumiría que, $y_i \sim P(\lambda_i)$, con $\lambda_i = \exp(x_i\beta)$, bajo el supuesto de equidispersión donde, $E(Y_i|x_i; \beta) = \text{Var}(Y_i|x_i; \beta) = \exp(x_i; \beta)$, bajo una especificación errónea donde la familia Poisson $\{P_\beta, \beta \in B\}$, no contiene el modelo estimado, indicando que la verdadera función de probabilidad puede no ser Poisson , o la media puede estar mal especificada debido a regresores omitidos o un enlace erróneo por la no linealidad del predictor, o inclusive, una mala especificación de la varianza.

Por lo anterior, el planteamiento de este modelo se realiza a través de la maximización de la función de log verosimilitud Poisson :

$$\text{Log}L = \sum_{i=1}^n [y_i x_i \beta - \exp(x_i \beta) - \log(y_i!)] \quad (3-28)$$

Resolviendo el problema de equidispersión del modelo clásico, el cual tendrá ahora una función de cuasi log-verosimilitud y $\hat{\beta}$, un cuasi estimador de máxima verosimilitud.

Para casos de muestras independientes idénticamente distribuidas se considera:

$$\sqrt{n} \left(\hat{\beta} - \beta^* \right) \sim N \left(0, A^{-1} B A^{-1} \right) \quad (3-29)$$

Donde, β^* es el vector de parámetros que minimiza la diferencia esperada entre el logaritmo de la distribución verdadera y el logaritmo de la cuasiverosimilitud.

Gourieroux et al. (1984) demostraron que $\hat{\beta}$ es consistente para β_0 , si la función de cuasi log verosimilitud pertenece a la familia exponencial y la media está bien especificada.

Las matrices $A(\beta)$ y $B(\beta)$, asumen el valor esperado del hessiano y del producto exterior del vector con respecto a la densidad verdadera de los datos (Winkelmann and Zimmermann, 1992).

$$A(\beta) = \left\{ -E_0 \left(\partial^2 \log f(y, x, \beta) / \partial \beta \partial \beta' \right) \right\} \quad (3-30)$$

$$B(\beta) = \left\{ E_0 \left(\partial \log f(x, y, \beta) / \partial \beta \cdot \partial \log f(y, x, \beta) / \partial \beta' \right) \right\} \quad (3-31)$$

Ambas matrices deben ser idénticas si se está trabajando con la función de probabilidad verdadera y resultarán en la matriz de información de Fisher.

En caso que no se esté trabajando con la probabilidad verdadera, A y B serán divergentes. Por la ley débil de los grandes números, $n^{-1} E_y \sum_{i=1}^n \log f(y|x, \beta) \rightarrow E_0 \log f(y, x, \beta)$, E_y , es la probabilidad condicional dado X . Tsou (2006)

Finalmente, usando la probabilidad condicional de Y , se establecen las matrices correspondientes A y B. Esto, aplicado en un modelo de regresión simple $\log(\mu_i) = \beta_0 + \beta_1 x$.

$$A = \lim_{n \rightarrow \infty} \sum_i^n \mu_{i,0} (x_i - \tilde{x})^2 / n \quad (3-32)$$

$$B = \lim_{n \rightarrow \infty} \sum_i^n \text{Var}_h(Y_i) (x_i - \tilde{x})^2 / n \quad (3-33)$$

Donde, $\tilde{x} = (\sum_i^n x_i \mu_{i,0}) / (\sum_i^n \mu_{i,0})$

Medidas de asociación

Dado que la generalidad y forma del modelo Poisson no cambian con este método sino que más bien se ajusta la verosimilitud antes de obtener los coeficientes del modelo,

este modelo trabaja con las mismas medidas de asociación que el caso tradicional:

■ **Razón de tasas:**

- Tasa de no expuestos:

$$X = 0 \rightarrow \beta_0 + \beta_1 0 = e^{\beta_0} \quad (3-34)$$

- Tasa de expuestos:

$$X = 1 \rightarrow \beta_0 + \beta_1 1 = e^{(\beta_0 + \beta_1)} \quad (3-35)$$

$$RTx_1 = \frac{T_{expuestos}(X_1 = 1)}{T_{noexpuestos}(X_1 = 0)} = \frac{e^{\beta_0} e^{\beta_1(X_1=1)}}{e^{\beta_0} e^{\beta_1(X_1=0)}} = e^{\beta_0} e^{\beta_1} \quad (3-36)$$

2. Generalidades

En un modelo log-lineal las estimaciones de los coeficientes se realizan así:

- $e^{\hat{\beta}_0}$, es el valor esperado de la variable dependiente cuando las variables explicativas son nulas o iguales a 0.
- Los valores de β_j , se interpretan en función de si es mayor o menor a 0, de modo que si $\beta_j > 0$, se produjo un incremento porcentual de la variable de respuesta cuando se produce un incremento unitario en la variable explicativa correspondiente, permaneciendo constante en el resto de variables predictoras. De la misma manera, cuando $\beta_j < 0$, se produce un decrecimiento porcentual de la variable explicativa cuando se produce un incremento en las variables predictoras. Gómez Gallego (2019)

4 Metodología

La metodología empleada para comparar las estimaciones, tanto con como sin el uso de modelos estadísticos, para los modelos logístico, log-binomial, Poisson y Poisson robusto, se detalla a continuación. El enfoque se centra en la estimación de la razón de prevalencia, teniendo en cuenta el diseño epidemiológico correspondiente a cada modelo. Este proceso se divide en las siguientes fases:

1. *Simulación*: Se llevó a cabo un ejercicio de simulación con dos variables aleatorias: una considerada como la variable de respuesta o evento, y la otra como la covariable o factor de exposición. Este ejercicio se ajustó para cuatro tamaños de muestra (369, 384, 1536, 6144) y dos escenarios de prevalencia:
 - E_1 : 0.40 para los expuestos, 0.10 para los no expuestos.
 - E_2 : 0.60 para los expuestos, 0.25 para los no expuestos.

Una vez definido el tamaño de muestra y la prevalencia, se estableció una semilla para asegurar la reproducibilidad de los resultados. Se crea un dataframe con una columna de identificación para cada individuo y se asigna una variable binaria de exposición 0 a 1 que contiene un identificador único para cada individuo.

Para los individuos expuestos ($Exposición = 1$), se asignaron eventos con una probabilidad de 0.6 o 0.4, mientras que para los no expuestos ($Exposición = 0$), los eventos se asignaron con una probabilidad de 0.25 o 0.10.

Para los modelos Logístico y Log-Binomial, se quiere estimar la relación entre la exposición y el evento, tal que se obtengan las probabilidades de ocurrencia para cada individuo (RP) y se obtienen intervalos de confianza para el 95 %.

Para los modelos Poisson y Poisson Robusto se realiza el ajuste de los modelos para conocer la relación entre la exposición y el evento, más, se obtienen la prevalencia, a

través de la aproximación de la incidencia y el Riesgo relativo.

Tras ajustar los modelos desde el enfoque clásico, se procedió a realizar el ajuste utilizando un enfoque bayesiano. Los modelos se ajustaron nuevamente, agregando una distribución a priori a los parámetros del modelo y empleando cadenas de Markov con 1000 iteraciones en cada una. Se generó un resumen del modelo ajustado, incluyendo estadísticas descriptivas de las distribuciones posteriores de los parámetros. Luego, se extrajeron muestras de la distribución posterior para estimar los coeficientes del nuevo modelo, con los cuales se calculó la razón de prevalencia, utilizando estos resultados para determinar la región de credibilidad.

2. *Estudio de Caso Aplicado:* Se desarrolló un estudio aplicado en una unidad de hematología pediátrica en una clínica de nivel IV en Santiago de Cali, enfocado en analizar la presencia de eventos adversos (EA), definido como cualquier incidente inesperado, documentado en la historia clínica, que causa lesión, discapacidad, aumento de la estancia hospitalaria, o incluso la muerte, como consecuencia de la atención hospitalaria y no de la enfermedad subyacente.

Se aplicaron modelos estadísticos para validar su capacidad de estimar la relación entre la exposición a un factor de riesgo, en este caso el trasplante de médula ósea (TMO), y la ocurrencia de eventos adversos en los pacientes, visto como experimento Poisson, donde los niños pueden hacer evento adverso en distintos momentos del tiempo durante el año de observación y, Binomial al analizar el final de todos los periodos de hospitalización tal que el niño hizo al menos un evento adverso. (Figura 4-1)

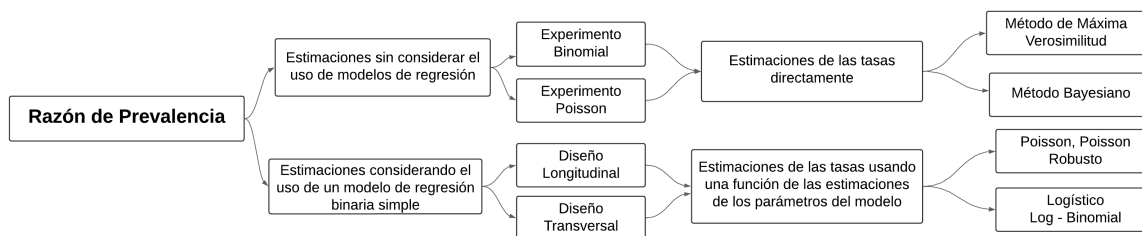


Figura 4-1: Esquema conceptual para los escenarios metodológicos

4.1. Estructuras analíticas

A continuación, se detalla la elaboración de las formas de estimación y cada uno de los modelos estadísticos a usarse. Para ello, se abordan dos tipos de estudios (Transversales y Longitudinales) de acuerdo a las observaciones en campo. (Figura 4-2)

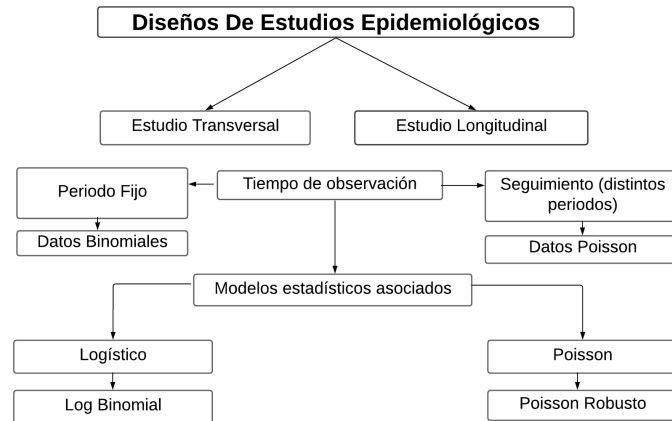


Figura 4-2: Esquema conceptual de los diseños para obtener las observaciones en campo y los modelos estadísticos asociados al análisis de los datos

4.2. Modelo Estadístico

4.2.1. Estimación del parámetro, asumiendo dos grupos de comparación:

Se tienen dos poblaciones identificadas por la presencia del evento:

B = Factor de exposición presente

B' = Factor de exposición ausente

Al interior de cada población se puede realizar la siguiente estrategia para recolección de los datos de estudio.

- Fijar un periodo de tiempo de observación (seguimiento) dividido en intervalos de tiempo fijos (p intervalos) del tipo $t_i - t_{i-1}$ igualmente espaciados; $[t_i - t_{i-1}] = [t_j - t_{j-1}]$
- Al interior de cada intervalo se realizan observaciones del tipo éxito - fracaso y se guarda el número de éxitos descontándolo del total de los elementos expuestos al momento t_0 .

Al final del periodo de seguimiento se tendrá un total de éxitos descontados de la cantidad N .

Para la construcción del modelo estadístico, se deben considerar dos posibilidades dependiendo del objetivo establecido para obtener los datos:

Posibilidad No.1: Trabajar con los conteos de éxitos observados en los diferentes periodos de observación (Estudio prospectivo para estimar incidencia y riesgo relativo).

■ **Proceso Generador de Datos (PGD)**

Condiciones en las que se encuentra presente el factor de exposición. A mayor presencia, mayor será la cantidad esperada de éxitos en el periodo de observación.

■ **Experimento Estadístico**

Tomar un elemento y observa si es un éxito. Si es así, se separa del conjunto total de elementos.

■ **Evento de interés**

$A = \{\text{El evento observado es un éxito}\}$

Sea X la variables aleatoria tal que:

$$X = \begin{cases} 1 & \text{Si ocurre } A \\ 0 & \text{Si no ocurre } A \end{cases}$$

Dado que se está repitiendo el experimento estadístico en muestras de tamaño $N \mapsto \infty$ de dos poblaciones definidas por la presencia del evento B , entonces, la variable X se redefinirá como:

$$X_j = \begin{cases} 1 & \text{Si ocurre } A \text{ en la población } J \\ 0 & \text{Si no ocurre } A \text{ en la población } J \end{cases}$$

Con, $J = 0, 1$

$$J = \begin{cases} 1 & \text{Si es la poblacion donde está presente } B \\ 0 & \text{Si es la poblacion donde no está presente } B \end{cases}$$

Entonces, $X_j \sim \text{Bernoulli}(\theta_j)$ y $P(X_j = x_j | \theta_j) = \theta_j^{x_j} (1 - \theta_j)^{1-x_j}$

Para $x_j \in \{0, 1\}$ y, $\theta_j \in (0, 1)$

Dado que al interior de cada periodo de tiempo se ejecutan los conteos, entonces se debe crear una variable aleatoria y al interior de cada periodo de observación, de la siguiente forma: $Y_{Jk} = \sum_{i=1}^{\infty} x_{ijk}$, de modo que: $Y_{Jk} \sim Poisson(\lambda_j)$

$$P(Y_{Jk} = y_{jk} \mid \lambda_j) = \frac{\lambda_j^{y_{jk}} e^{-\lambda_j}}{y_{jk}!};$$

Con $\lambda_j > 0$, $Y_{Jk} \in \{0, 1, 2, \dots\}$; $J = 0, 1$; $k = 1, 2, 3, \dots, p$.

Se asume entonces que el Proceso Generador de Datos (PGD), permanece en condiciones iguales durante los p periodos de tiempo, así por ejemplo, las observaciones x_{111} y x_{211} , son observaciones iid en el periodo 1 y la población 1, lo que lleva a que las observaciones $y_{j1}, y_{j2}, \dots, y_{jp}$ sean independientes e idénticamente distribuidas (Poisson con el mismo λ_j en todos los periodos).

La función masa de probabilidad conjunta será:

$$P(Y_{j1}, Y_{j2}, \dots, Y_{jp} \mid \lambda_j) = \prod_{k=1}^p \frac{\lambda_j^{y_{jk}} e^{-\lambda_j}}{y_{jk}!} = c * \lambda_j^u e^{-p\lambda_j} \quad (4-1)$$

donde, $u_j = \sum_{k=1}^p y_{jk}$ y $c = 1 / \prod_{k=1}^p y_{jk}!$.

La función de verosimilitud será proporcional a:

$$L(\lambda_j \mid y_{j1}, \dots, y_{jp}) \propto \lambda_j^u e^{-p\lambda_j} \quad (4-2)$$

Y el estimador λ_j ; $j = 0, 1$ obtenido usando el método de máxima verosimilitud será:

$$\hat{\lambda}_j = \frac{u_j}{p} = \sum_{k=1}^p y_{jk} / p \quad (4-3)$$

$$E(\hat{\lambda}_j) = \frac{E(u_j)}{p} = \frac{p\lambda_j}{p} = \lambda_j \quad (4-4)$$

$$Var(\hat{\lambda}_j) = \frac{1}{p^2} Var\left(\sum_{k=1}^p y_{jk}\right) = \frac{\lambda_j}{p} \quad (4-5)$$

Cuando no es posible asumir igual varianza en todos los periodos, $\lambda_j \neq \lambda_{j'}$, se debe recurrir a estrategias como la mezcla $Y_j \sim Poisson(\lambda_{jk})$ y $\lambda_j \sim Gamma(\alpha_j, \beta_j)$. Para realizar la estimación, otra alternativa viable, y más usada, es emplear un modelo de regresión Poisson que no considere igualdad de varianza.

Posibilidad No.2: Trabajar con el acumulado de éxitos al final del estudio y su complemento, *los no éxitos*, (Estudio transversal para estimar prevalencias y razón de prevalencia). Esta prevalencia intragrupo podrá considerarse una incidencia acumulada.

Se tiene $t_0 \rightarrow N_1$ que corresponden a los *expuestos*, N_2 a los *No expuestos* y N es el total de individuos.

Durante el tiempo t_1 se observan y_{j1} éxitos, entonces, al final del periodo se tiene $y_{j1} + y_{j2} + \dots y_{jk}$ éxitos acumulados y $N_j - y_{j1} - y_{j2} - \dots - y_{kj}$ fracasos.

El proceso continúa hasta el periodo p y al final del seguimiento se tienen $u_j = \sum_{k=1}^p y_{jk}$ éxitos y $N_j^* = N_j - u_j$ fracasos.

El Proceso Generador Datos se mantiene igual al anterior. Sin embargo, la estrategia con la que se obtienen los datos cambia. El experimento estadístico no cambia, al igual que el evento de interés y la variable aleatoria binaria que representa la ocurrencia del evento de interés.

Se crean las variables aleatorias: $u_j = \sum_{k=1}^p y_{jk} = \sum_{k=1}^p \sum_{i=1}^N x_{ijk}$

Entonces: $U_j \sim Binomial(N_j, \theta_j)$ y,

$$P(U_j = u_j | N_j, \theta_j) = \binom{N_j}{u_j} \theta_j^{u_j} (1 - \theta_j)^{N_j - u_j} \quad (4-6)$$

Para, $\theta_{j \in \{0,1\}} \in \{0,1\}$; $u_j \in \{0,1,\dots,N_j\}$; $j = 0,1$

La función de verosimilitud será proporcional a:

$$L(\theta_j | u_j, N_j) \propto \theta_j^{u_j} (1 - \theta_j)^{N_j - u_j} \quad (4-7)$$

El estimador por máxima verosimilitud, su esperanza y varianza serán:

$$\hat{\theta}_{jMV} = \frac{U_j}{N_j}; E(\hat{\theta}_{jMV}) = \theta_j; V(\hat{\theta}_{jMV}) = \frac{\theta_j(1 - \theta_j)}{N_j} \quad (4-8)$$

4.2.2. Estimaciones Bayesianas asumiendo un experimento Poisson en la recolección de los datos

■ **Estimaciones bayesianas para datos bajo experimento Poisson (Diseño prospectivo)**

Siendo λ_j , el parámetro de interés, se le puede asociar una distribución a prior Gamma, con hiperparámetros (α_j, β_j) :

$$\lambda_0 \sim \text{Gamma}(\alpha_0, \beta_0) \rightarrow f(\lambda_0) = \frac{\beta_0}{\Gamma(\alpha_0)} (\beta_0 \lambda_0)^{\alpha_0 - 1} e^{-\beta_0 \lambda_0}$$

$$\lambda_1 \sim \text{Gamma}(\alpha_1, \beta_1) \rightarrow f(\lambda_1) = \frac{\beta_1}{\Gamma(\alpha_1)} (\beta_1 \lambda_1)^{\alpha_1 - 1} e^{-\beta_1 \lambda_1}$$

Se observa que las distribuciones a priori planteadas para λ_j no tienen que tener exactamente los mismos valores de los hiperparámetros.

La distribución a posteriori conjunta se obtiene de la siguiente manera:

Dado que $Y_{jk} \sim \text{Poisson}(\lambda_j)$, la función de verosimilitud para p tiempos es:

$$L(\lambda_j | Y_{jk}) = \prod_{k=1}^p \frac{\lambda_j^{Y_{jk}} e^{-\lambda_j}}{Y_{jk}!}$$

La distribución a posteriori proporcional se obtiene como:

$$P(\lambda_j | Y_{jk}) \propto \lambda_j^{\sum_{k=1}^p Y_{jk} + \alpha_j - 1} e^{-(p + \beta_j)\lambda_j}$$

La forma de la distribución a posteriori conjunta indica que λ_j sigue una distribución

Gamma con parámetros actualizados:

$$\lambda_j | Y_{jk} \sim \text{Gamma} \left(\alpha_j + \sum_{k=1}^p Y_{jk}, \beta_j + p \right)$$

Por lo tanto, la esperanza de λ_j con $j = 0, 1$ es:

$$E[\lambda_j | Y_{jk}] = \frac{\alpha_j + \sum_{k=1}^p Y_{jk}}{\beta_j + p}$$

■ **Estimaciones bayesianas para datos bajo experimento Binomial (Usando modelos de regresión binaria con una covariable)**

En esta aproximación es posible usar modelos de regresión Binaria o regresión Poisson para modelar, de forma lineal, la relación entre el factor de exposición (asociando su presencia a una variable explicativa o covariable binaria) y la probabilidad de ocurrencia del evento de éxito. La respuesta de interés puede ser modelada a través de la probabilidad de éxito $P(Y = 1 | Z = j) = \theta_j$ o a través de $\lambda_j = E(Y_j) = N_j \theta_j$ la cantidad media de éxitos por periodo.

Con el parámetro θ_j se modela la prevalencia y para ello se puede usar un modelo de regresión Binaria con alguno de los enlaces conocidos (logístico, probit, log-log complementario) o un modelo de regresión Log- Binomial.

Cuando la aproximación se realiza usando datos de conteo por periodo (datos Poisson), se puede ajustar un modelo de regresión Poisson en alguna de sus variantes (Poisson homogéneo, Poisson -Gamma, Poisson con varianza robusta). Se desglosan los siguientes casos:

4.2.3. Modelo de Regresión para datos bajo experimento Poisson (Diseño prospectivo)

- **Modelo Poisson** Se expresa el modelo Poisson como:

$$\log(\lambda_j) = \beta_0 + \beta_1 x_{1i} \tag{4-9}$$

$\lambda_j; j = 0, 1$ representando la tasa promedio por población y x la covariable.

Entonces:

$$\begin{cases} \lambda_0 = E(Y|X=1) & \lambda_0 = N_1\theta_0 \\ \lambda_1 = E(Y|X=0) & \lambda_1 = N_2\theta_1 \end{cases}$$

Se despejan de la expresión los valores de θ_1, θ_2 , los cuales serán equivalentes al cálculo del riesgo relativo y, posteriormente, τ como la razón de tasas.

$$\begin{aligned} \theta_1 &= \frac{\lambda_1}{N_1}; \theta_2 = \frac{\lambda_2}{N_2} \\ \tau &= \frac{\theta_1}{\theta_2} = \frac{\frac{\lambda_1}{N_1}}{\frac{\lambda_2}{N_2}} = \frac{\lambda_1 N_2}{\lambda_2 N_1} \end{aligned}$$

Estimación Bayesiana

Bajo la estructura anterior, desde el enfoque bayesiano se realizan las estimaciones de los valores de e^{β_0}, e^{β_1} .

$$\log(\lambda_j) = \beta_0 + \beta_1 x_{1i}$$

Entre los expuestos, el modelo tomará la forma de:

$$\begin{aligned} \lambda_1 &= E(Y|X=1) = \log(\lambda_1) = \beta_0 + \beta_1 \\ \lambda_2 &= E(Y|X=0) = \log(\lambda_2) = \beta_0 \end{aligned}$$

Aplicando exponencial a ambos lados;

$$\begin{aligned} \lambda_0 &= e^{\beta_0 + \beta_1} = N_0\theta_0 \rightarrow \theta_0 = \frac{\lambda_0}{N_0} = \frac{e^{\beta_0 + \beta_1}}{N_1} \\ \lambda_1 &= e^{\beta_0} = N_1\theta_1 \rightarrow \theta_1 = \frac{\lambda_1}{N_1} = \frac{e^{\beta_0}}{N_1} \end{aligned}$$

Reemplazando en la forma del τ :

$$\tau = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} = \frac{N_0\theta_0}{N_1\theta_1} = \frac{e^{\beta_0 + \beta_1} N_1}{e^{\beta_0} N_2} = \frac{N_1 e^{\beta_1}}{N_0}$$

Entonces $\hat{\tau}$, será:

$$\hat{\tau} = \frac{N_1 e^{\hat{\beta}_1}}{N_0}$$

Si se asumen distribuciones normales con media μ y desviación σ como distribuciones

a priori para los parámetros beta del modelo de regresión, se tiene que:

$$\text{Sea, } Y_i | \lambda_j \sim \text{Poisson}(\lambda_j)$$

Donde $\log(\lambda_j) = \beta_0 + \beta_1 x_{1i}$ entonces: $\lambda_j = e^{\beta_0 + \beta_1 x_i}$ con $i = 1, \dots, p$.

Partiendo de que, $\beta_0 || \beta_1$, las distribuciones a priori tendrían la siguiente forma:

$$\begin{aligned} \beta_0 &\sim \text{Normal}(a, b) \rightarrow f(\beta_0) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{1}{2b^2}(\beta_0 - a)^2} \\ \beta_1 &\sim \text{Normal}(c, d) \rightarrow f(\beta_1) = \frac{1}{\sqrt{2\pi d}} e^{-\frac{1}{2d^2}(\beta_1 - c)^2} \end{aligned}$$

Con a, b, c y d como los hiperparámetros.

Podemos obtener la función de verosimilitud como:

$$L(\beta_0, \beta_1 | Y, X) = \prod_{i=1}^p \frac{e^{-(\beta_0 + \beta_1 X_i)} (\beta_0 + \beta_1 X_i)^{Y_i}}{Y_i!} \quad (4-10)$$

Tomando el logaritmo de la verosimilitud:

$$\log L(\beta_0, \beta_1 | Y, X) = \sum_{i=1}^p [-e^{\beta_0 + \beta_1 X_i} + Y_i(\beta_0 + \beta_1 X_i) - \log(Y_i!)]$$

La distribución a posteriori conjunta es proporcional al producto de la verosimilitud y las distribuciones a priori:

$$P(\beta_0, \beta_1 | Y, X) \propto \exp \left(\sum_{i=1}^n [-e^{\beta_0 + \beta_1 X_i} + Y_i(\beta_0 + \beta_1 X_i)] - \frac{(\beta_0 - \mu_0)^2}{2\sigma_0^2} - \frac{(\beta_1 - \mu_1)^2}{2\sigma_1^2} \right)$$

Como se puede observar, en este caso la distribución posterior no presenta una forma conocida, por lo que se hace necesario usar métodos de Montecarlo basados en cadenas de Markov (MCMC) implementados a través del software estadístico de uso libre OpenBugs. Raftery and Lewis (1996)

■ Modelo Poisson Robusto

Se expresa el modelo Poisson como:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} \quad (4-11)$$

$\lambda_j; j = 0, 1$ representando la tasa promedio por población y x la covariable.

Entonces:

$$\begin{cases} \lambda_0 = E(Y|X = 1) & \lambda_0 = N_1 \theta_0 \\ \lambda_1 = E(Y|X = 0) & \lambda_1 = N_2 \theta_1 \end{cases}$$

Definiendo la expresión de θ_0, θ_1 como el despeje de la igualdad anterior, ajustándose al cálculo del Riesgo relativo o razón de tasas como:

$$\begin{aligned} \theta_0 &= \frac{\lambda_0}{N_0}; \theta_1 = \frac{\lambda_1}{N_1} \\ \tau = \frac{\theta_0}{\theta_1} &= \frac{\frac{\lambda_0}{N_0}}{\frac{\lambda_1}{N_1}} = \frac{\lambda_0 N_1}{\lambda_1 N_0} \end{aligned}$$

Bajo la estructura anterior, al ser casos de conteo, se puede realizar la estimación del Riesgo relativo con la estimación robusta de los coeficientes β_i .

$$\hat{\beta}_{iPPR} = \left(X' \hat{W} X \right)^{-1} X' \hat{W} \hat{Z}$$

Donde $\hat{W} = \text{diag}(\lambda_i)$, $\hat{Z}_i = \log \left[\hat{\mu}_i \left(\hat{\beta}_{iPPR} \right) \right] + \left[\frac{Y_i - \hat{\mu}_i(\hat{\beta}_{iPPR})}{\hat{\mu}_i(\hat{\beta}_{iPPR})} \right]$, X como la matriz de covariables.

Estimación Bayesiana:

Sea, $\lambda_j; j = 0, 1$, el parámetro de interés, se le puede asociar una distribución a prior Gamma, con hiperparámetros (α_j, β_j) :

$$\begin{aligned} \lambda_0 &\sim \text{Gamma}(\alpha_0, \beta_0) \rightarrow f(\lambda_0) = \frac{\beta_0}{\Gamma(\alpha_0)} (\beta_0 \lambda_0)^{\alpha_0-1} e^{-\beta_0 \lambda_0} \\ \lambda_1 &\sim \text{Gamma}(\alpha_1, \beta_1) \rightarrow f(\lambda_1) = \frac{\beta_1}{\Gamma(\alpha_1)} (\beta_1 \lambda_1)^{\alpha_1-1} e^{-\beta_1 \lambda_1} \end{aligned}$$

Se observa que las distribuciones a priori planteadas para λ_j no tienen que tener exactamente

los mismos valores de los hiperparámetros.

Entonces,

$$\lambda_1, \lambda_2 | Y \sim \text{Gamma}(\alpha + n, \beta + t)$$

Donde, $t = Y$ ó, $t = \sum_{i=1}^n y_i$ y, $n = 1$. Si, $\hat{\lambda}_B = E(\lambda | Y)$, $\rightarrow \hat{\tau} = \frac{\hat{\lambda}_{1B} N_2}{\hat{\lambda}_{2B} N_1}$

Considerando que en la estimación del RR sin modelo de regresión no se consideran las estimaciones de los β , estos se realizan al igual que en la manera clásica.

Entonces,

$$\ln(\lambda) = \hat{\beta}_{0PPR} + \hat{\beta}_{1PPR} X$$

Entre los expuestos el modelo tomará la forma de:

$$\begin{aligned} \lambda_1 &= E(Y | X = 1) = \ln(\lambda_1) = \hat{\beta}_{0PPR} + \hat{\beta}_{1PPR} \\ \lambda_2 &= E(Y | X = 0) = \ln(\lambda_2) = \hat{\beta}_{0PPR} \end{aligned}$$

Aplicando exponencial a ambos lados;

$$\begin{aligned} \lambda_1 &= e^{\hat{\beta}_{0PPR} + \hat{\beta}_{1PPR}} = N_2 \theta_1 \rightarrow \theta_1 = \frac{\lambda_1}{N_1} = \frac{e^{\hat{\beta}_{0PPR} + \hat{\beta}_{1PPR}}}{N_1} \\ \lambda_2 &= e^{\hat{\beta}_{0PPR}} = N_1 \theta_2 \rightarrow \theta_2 = \frac{\lambda_2}{N_2} = \frac{e^{\hat{\beta}_{0PPR}}}{N_2} \end{aligned}$$

Reemplazando en la forma del τ :

$$\tau = \frac{e^{\hat{\beta}_{0PPR} + \hat{\beta}_{1PPR}}}{e^{\hat{\beta}_{0PPR}}} = e^{\hat{\beta}_{1PPR}} = \frac{N_1 \theta_1}{N_2 \theta_2} = \frac{e^{\hat{\beta}_{0PPR} + \hat{\beta}_{1PPR}} N_2}{e^{\hat{\beta}_{0PPR}} N_1} = \frac{N_2 e^{\hat{\beta}_{1PPR}}}{N_1}$$

Entonces $\hat{\tau}$, será:

$$\hat{\tau} = \frac{N_2 e^{\hat{\beta}_{1PPR}}}{N_1}$$

Ahora bien, si consideramos la asociación de distribuciones a priori Normal, entonces:

Partiendo de que, $\beta_{0PPR} || \beta_{1PPR}$, las distribuciones a priori tendrían la siguiente forma:

$$\beta_{0PPR} \sim Normal(a, b) \rightarrow f(\beta_{0PPR}) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{1}{2b^2}(\beta_{0PPR}-a)^2}$$

$$\beta_{1PPR} \sim Normal(c, d) \rightarrow f(\beta_{1PPR}) = \frac{1}{\sqrt{2\pi d}} e^{-\frac{1}{2d^2}(\beta_{1PPR}-c)^2}$$

Con a, b, c y d como los hiperparámetros.

Aplicando el teorema de Bayes establecemos la función de Verosimilitud:

$$L(\beta|Y) = \prod_{i=1}^n \frac{e^{-\theta_i} \theta_i^{Y_i}}{Y_i!} = e^{-\sum_{i=1}^n \theta_i} \prod_{i=1}^n \frac{\theta_i^{Y_i}}{Y_i!} \quad (4-12)$$

Permitiendo la expresión de la distribución posterior como (se omite la constante al usar el proporcional):

$$L(\beta_{0PPR}, \beta_{1PPR}|Y) \propto e^{-\sum_{i=1}^n \theta_i} \prod_{i=1}^n \frac{\theta_i^{Y_i}}{Y_i!} \cdot f(\beta_{0PPR}) \cdot f(\beta_{1PPR})$$

$$L(\beta_{0PPR}, \beta_{1PPR}|Y) \propto e^{-\sum_{i=1}^n \theta_i} \prod_{i=1}^n \frac{\theta_i^{Y_i}}{Y_i!} \cdot e^{-\frac{1}{2b^2}(\beta_{0PPR}-a)^2} \cdot e^{-\frac{1}{2d^2}(\beta_{1PPR}-c)^2}$$

$$L(\beta_{0PPR}, \beta_{1PPR}|Y) \propto e^{-\sum_{i=1}^n \theta_i - \frac{1}{2b^2}(\beta_{0PPR}-a)^2 - \frac{1}{2d^2}(\beta_{1PPR}-c)^2} \prod_{i=1}^n \theta_i^{Y_i}$$

Recordamos que si, $\log(\theta_i) = \beta_{0PPR} + \beta_{1PPR}X$, entonces, $\theta_i = e^{\beta_{0PPR} + \beta_{1PPR}X}$

$$L(\beta_{0PPR}, \beta_{1PPR}|Y) \propto e^{-\sum_{i=1}^n e^{\beta_{0PPR} + \beta_{1PPR}X} - \frac{1}{2b^2}(\beta_{0PPR}-a)^2 - \frac{1}{2d^2}(\beta_{1PPR}-c)^2} \cdot \prod_{i=1}^n (e^{\beta_{0PPR} + \beta_{1PPR}X})^{Y_i}$$

Como se puede observar, en este caso la distribución posterior no presenta una forma conocida, por lo que se hace necesario usar métodos MCMC implementados a través de OpenBugs.

4.2.4. Modelo de regresión cuando los datos son obtenidos usando experimentos binomiales

- **Modelo Logístico:** Sea $\text{logit}(\theta_i)$, la forma que tomará el modelo logístico cuando es ajustado para una covariable, con $i = 0, 1$.

$$\text{logit}(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x \quad (4-13)$$

$$\theta_i = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

De manera extendida, la forma que tomarían θ_0, θ_1 :

$$\begin{aligned} \ln\left(\frac{\theta_i}{1-\theta_i}\right) &= \beta_0 + \beta_1 x \\ \left(\frac{\theta_i}{1-\theta_i}\right) &= e^{\beta_0 + \beta_1 x} \\ \theta_i &= e^{\beta_0 + \beta_1 x} (1 - \theta_i) \\ \theta_i &= e^{\beta_0 + \beta_1 x} - \theta_i (e^{\beta_0 + \beta_1 x}) \\ \theta_i + \theta_i (e^{\beta_0 + \beta_1 x}) &= e^{\beta_0 + \beta_1 x} \\ \theta_i (1 + e^{\beta_0 + \beta_1 x}) &= e^{\beta_0 + \beta_1 x} \\ \theta_i &= \frac{e^{\beta_0 + \beta_1 x}}{(1 + e^{\beta_0 + \beta_1 x})} \end{aligned}$$

Cuando $X = 0$, entonces:

$$\hat{\theta}_0 = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} \quad (4-14)$$

Cuando $x = 1$,

$$\hat{\theta}_1 = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}} \quad (4-15)$$

Se define la estructura para la Razón de prevalencia como el cociente entre la prevalencia

de expuestos y no expuestos $\hat{\tau}$, así:

$$RP = \tau = \frac{\theta_1}{\theta_0}$$

$$\theta_1 = P(Y = 1|X = 1)$$

$$\theta_0 = P(Y = 1|X = 0)$$

$$PR = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}$$

Resultando en la estructura de la razón de prevalencia $\hat{\tau}$:

$$\hat{\tau} = \frac{\hat{\theta}_1}{\hat{\theta}_0} = \frac{\frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}}{\frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1} (1 + e^{\hat{\beta}_0})}{e^{\hat{\beta}_0} (1 + e^{\hat{\beta}_0 + \hat{\beta}_1})} = \frac{e^{\hat{\beta}_1} (1 + e^{\hat{\beta}_0})}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}$$

$$\hat{\tau} = \frac{e^{\hat{\beta}_1} + e^{\hat{\beta}_1} e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}} = \frac{e^{\hat{\beta}_1} + e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}$$

Estimación Bayesiana:

Se realizan las estimaciones de los valores de $e^{\hat{\beta}_0}, e^{\hat{\beta}_1}$. Considerando las complicaciones que se presentan al especificar directamente una distribución a priori sobre los coeficientes de regresión Bedrick et al. (1996), se asumen distribuciones a priori normales para el modelo logístico al usar una función de enlace canónico; $Normal(\mu, \sigma^2)$, donde $\mu_j = E(\beta_j); j = 0, 1$.

Sea, $Y_i|\theta_i \sim Bernoulli(\theta_i)$

Donde $\theta_i = P(Y_i = 1)$, con valor esperado $E(Y_i|\theta_i) = \theta_i$.

Para poder explicar θ_i usaremos X como predictor y una función $g()$, para garantizar que θ_i se mueva entre 0 y 1, tal que $g(\theta_i) = \beta_0 + \beta_1 X_1$.

Entonces, $E(Y_i|\theta_i) = \theta_i$ usando la función de enlace logit se puede expresar como $g(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right)$ lo que indica que:

$$Y_i | \beta_0, \beta_1 \sim \text{Bernoulli}(\theta_i)$$

$$\theta_i = \frac{e^{\beta_0 + \beta_1 x}}{(1 + e^{\beta_0 + \beta_1 x})}$$

Se definen las siguientes distribuciones a priori para los coeficientes β_0, β_1 :

$$\beta_0 \sim \text{Normal}(a, b)$$

$$\beta_1 \sim \text{Normal}(a, b)$$

Con a, b como los hiperparámetros.

Así pues, la función de Máxima Verosimilitud se puede expresar como:

$$L(\beta_0, \beta_1) = \prod_{i=1}^n P(Y_i = 1)^{Y_i} [1 - P(Y_i = 1)]^{1-Y_i} \quad (4-16)$$

Asumiendo que, $t_i = \beta_0 + \beta_1 x$ reemplazamos en la ecuación anterior despejando el valor de la probabilidad de éxito.

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \left(\frac{e^{t_i}}{1 + e^{t_i}} \right)^{Y_i} \left[1 - \frac{e^{t_i}}{1 + e^{t_i}} \right]^{1-Y_i}$$

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{e^{t_i Y_i}}{(1 + e^{t_i})^{Y_i}} \left(\frac{1 + e^{t_i} - e^{t_i}}{1 + e^{t_i}} \right)^{1-Y_i}$$

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{e^{t_i Y_i}}{(1 + e^{t_i})^{Y_i}} \left(\frac{1}{1 + e^{t_i}} \right)^{1-Y_i}$$

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{e^{t_i Y_i}}{(1 + e^{t_i})^{Y_i}} \prod_{i=1}^n \left(\frac{1}{1 + e^{t_i}} \right)^{1-Y_i}$$

$$L(\beta_0, \beta_1) = \frac{\prod_{i=1}^n e^{t_i Y_i}}{\prod_{i=1}^n (1 + e^{t_i})^{Y_i}} \cdot \frac{\prod_{i=1}^n (1)^{1-Y_i}}{\prod_{i=1}^n (1 + e^{t_i})^{1-Y_i}}$$

Reemplazando el valor $t_i = \beta_0 + \beta_1 x$ nuevamente,

$$L(\beta_0, \beta_1) = \frac{e^{\sum_{i=1}^n \beta_0 y_i + \sum_{i=1}^n \beta_1 x_i y_i}}{\prod_{i=1}^n (1 + e^{t_i})^{y_i}} \cdot \frac{1}{\prod_{i=1}^n (1 + e^{t_i})^{1-y_i}}$$

Entonces, la forma de la verosimilitud sería:

$$L(\beta_0, \beta_1) = \frac{e^{\beta_0 y_1 + \beta_1 x_1 y_1} \cdot e^{\beta_0 y_2 + \beta_1 x_2 y_2} \cdot \dots \cdot e^{\beta_0 y_n + \beta_1 x_n y_n}}{\prod_{i=1}^n (1 + e^{t_i})^{1-y_i}}$$

$$L(\beta_0, \beta_1) = \frac{e^{\beta_0 \sum_{i=1}^n y_i + \beta_1 \sum_{i=1}^n x_i y_i}}{\prod_{i=1}^n (1 + e^{t_i})^{1-y_i}}$$

Ajustando las distribuciones a priori y considerando que se realiza el ajuste sobre un modelo simple con $(a, c) = \mu$ y $(b, d) = \sigma^2$

$$\beta_0 \sim Normal(a, b) \rightarrow f(\beta_0) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\beta_0 - a}{b}\right)^2} \quad (4-17)$$

$$\beta_1 \sim Normal(c, d) \rightarrow f(\beta_1) = \frac{1}{d\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\beta_1 - c}{d}\right)^2} \quad (4-18)$$

$$f(\beta_0, \beta_1 | Y) \propto \frac{e^{\beta_0 y_i + \beta_1 x_i y_i}}{\prod_{i=1}^n (1 + e^{t_i})} \cdot e^{-\frac{1}{2b^2}(\beta_0 - a)^2} \cdot e^{-\frac{1}{2d^2}(\beta_1 - c)^2} \quad (4-19)$$

Sea, $w_i = \beta_0 y_i + \beta_1 x_i y_i$ entonces, tendremos la forma proporcional de la distribución posterior

$$f(\beta_0, \beta_1 | Y) \propto \frac{e^{w_i}}{\prod_{i=1}^n (1 + e^{t_i})} \cdot e^{-\frac{1}{2b^2}(\beta_0 - a)^2} \cdot e^{-\frac{1}{2d^2}(\beta_1 - c)^2}$$

$$\propto \frac{1}{\prod_{i=1}^n (1 + e^{t_i})} e^{w_i - \frac{1}{2b^2}(\beta_0 - a)^2 - \frac{1}{2d^2}(\beta_1 - c)^2}$$

Como se puede observar, en este caso la distribución posterior no presenta una forma conocida por lo que se hace necesario usar métodos MCMC implementados a través de OpenBugs. Bazán and Bayes (2010)

Dicho esto, la razón de prevalencia $\hat{\tau}_B$, indicará la estimación de la RP usando el método de Bayes:

$$\hat{\tau}_B = \frac{e^{\hat{\beta}_{1B}} + e^{\hat{\beta}_{0B} + \hat{\beta}_{1B}}}{1 + e^{\hat{\beta}_{0B} + \hat{\beta}_{1B}}} \quad (4-20)$$

■ Modelo Log Binomial

$$\ln(\theta(x=0)) = \beta_0$$

$$\ln(\theta(x=1)) = \beta_0 + \beta_1$$

De forma extendida,

$$\begin{aligned} \ln(\theta_0) &= \beta_0 + \beta_1(0) \rightarrow \ln(\theta_0) = \beta_0 \rightarrow \theta_0 = e^{\beta_0} \\ \ln(\theta_1) &= \beta_0 + \beta_1(1) \rightarrow \ln(\theta_1) = \beta_0 + \beta_1 \rightarrow \theta_1 = e^{\beta_0 + \beta_1} \end{aligned}$$

Al ser un estudio transversal se define entonces:

Y_1 = Número de éxitos en n_1

Y_2 = Número de éxitos en n_2

Entonces, $\hat{\theta}_1 = \frac{Y_1}{n_1}$ y, $\hat{\theta}_2 = \frac{Y_2}{n_2}$, así pues, la estimación de τ se expresa como:

$$\hat{\tau} = \frac{\hat{\theta}_1}{\hat{\theta}_2} = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}} = \frac{n_2 Y_1}{n_1 Y_2}$$

Y, usando el modelo de regresión:

$$\hat{\tau} = \frac{\hat{\theta}_1}{\hat{\theta}_0} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1}$$

Estimación Bayesiana:

Se realiza la estimación para los coeficientes del modelo, partiendo del supuesto de que $\beta_0 || \beta_1$, se asumen distribuciones a priori *Uniforme* (a, b). Sea, $Y_i \sim \text{Binomial}(n_i, \theta_i)$

Donde $\text{logit}(\theta_i) = \beta_0 + \beta_1 X$; $i = 1, \dots, n$

Partiendo de que, $P(Y = 1|X) = e^{\beta_0 + \beta_1 X}$

$$\beta_0 \sim \text{Uniforme}(a, b)$$

$$\beta_1 \sim \text{Uniforme}(c, d)$$

Con a, b como los hiperparámetros.

La función de Verosimilitud estaría dada por:

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^n P(Y_i = 1)^{Y_i} [1 - P(Y_i = 1)]^{1-Y_i} \\ L(\beta_0, \beta_1) &= \prod_{i=1}^n (e^{\beta_0 + \beta_1 X})^{Y_i} (1 - e^{\beta_0 + \beta_1 X})^{1-Y_i} I_{\Theta}(\beta) \end{aligned}$$

Donde, $\Theta = \{(\beta_0, \beta_1) : \beta_0 + \beta_1 X \leq 0\}$

$$L(\beta_0, \beta_1) = \prod_{i=1}^n e^{\beta_0 Y_i + \beta_1 X Y_i} (1 - e^{\beta_0 + \beta_1 X})^{1-Y_i} I_{\Theta}(\beta)$$

$$L(\beta_0, \beta_1) = \left(e^{\beta_0 \sum_{i=1}^n Y_i + \beta_1 \sum_{i=1}^n X Y_i} \right) \prod_{i=1}^n (1 - e^{\beta_0 + \beta_1 X})^{1-Y_i} I_{\Theta}(\beta)$$

Considerando que se realiza el ajuste sobre un modelo simple:

$$\beta_0 \sim Uniforme(a, b) \rightarrow f(\beta_0) = \frac{1}{b-a}$$

$$\beta_1 \sim Uniforme(c, d) \rightarrow f(\beta_1) = \frac{1}{d-c}$$

Entonces la forma proporcional de la distribución posterior se puede expresar como:

$$L(\beta_0, \beta_1 | Y) \propto \left(e^{\beta_0 Y_1 + \beta_1 X Y_1} \right) \prod_{i=1}^n (1 - e^{\beta_0 + \beta_1 X})^{1-Y_i} I_{\Theta}(\beta) \cdot \frac{1}{b-a} \cdot \frac{1}{d-c}$$

Como se puede observar, en este caso la distribución posterior no presenta una forma conocida, por lo que se hace necesario usar métodos MCMC implementados a través de OpenBugs.

4.3. Estudio de Simulación

El proceso de modelamiento se llevó a cabo utilizando el software estadístico R, empleando únicamente las librerías incluidas en la instalación base del software. Para asegurar la reproducibilidad de las simulaciones, se utilizó una semilla. Se construyó un dataframe para almacenar los vectores de muestras y probabilidades generados, realizando el ajuste de los modelos mediante la función `glm`, especificando la familia correspondiente para cada modelo. Además, se empleó un ciclo para almacenar cada una de las listas de resultados.

Este ejercicio se realizó para cada modelo bajo escenarios de baja y alta prevalencia. Los conteos realizados permitieron obtener la prevalencia en expuestos, no expuestos, y la razón de prevalencia, siguiendo el esquema detallado a continuación:

Supuestos	λ	P Expuestos ¹	P No Expuestos ²	n^3	R_1^4	R_2^4	R_3^4
Homocedasticidad	5	0.40	0.10	369	22	44	53
		0.60	0.25		16	27	49
Heterocedasticidad	2	0.40	0.10	384	105	96	80
		0.60	0.30		119	93	65
Heterocedasticidad	4	0.40	0.10	1536	104	234	302
		0.60	0.30		108	220	293
Heterocedasticidad	8	0.40	0.10	6144	9	61	179
		0.60	0.30		10	70	156

¹ P expuestos: Prevalencia de expuestos

² P No expuestos: Prevalencia de los no expuestos

³ n: tamaño de muestra del ejercicio de simulación

⁴ $R_i; i = 1, 2, 3$: Número de éxitos en el momento i

Tabla 4-1: Esquema de simulación de datos

Los datos se generaron mediante simulación de variables aleatorias Poisson para dos tamaños de muestra derivados de un nivel de confianza del 95 % y un error estándar de 0.5. Estos escenarios incluyeron prevalencias en expuestos de 0.40 y 0.60, y en no expuestos de 0.10 y 0.30. Se determinaron los tamaños de muestra asociados a la prevalencia y se realizaron conteos de éxitos en tres momentos de tiempo (R_1, R_2, R_3) para casos de homocedasticidad y heterocedasticidad de λ . Esta información permitió realizar la transformación $n\theta$ según el modelo aplicado.

4.4. Estudio con datos reales

Se tienen los datos de 114 pacientes que ingresos a la Unidad de Hemato- Oncología Pediátrica en un centro de salud nivel IV, de lo que se define el evento estadístico como:

- Evento B = El niño presenta trasplante de Medula Espinal
- Evento A = El niño hace al menos un evento adverso durante el p-ésimo periodo de observación

Proceso Generador de Datos: Nivel de cumplimiento de los protocolos de manejo y cuidado de los niños por parte del personal de salud en la Unidad de Oncología pediátrica de un centro de atención en salud nivel IV de la ciudad. **Experimento estadístico:** Tomar la ficha del niño y mirar si hizo al menos un evento adverso en el k-ésimo mes de estudio ($J= 12$ meses de observación)

5 Resultados

Este capítulo presenta los resultados del ejercicio de simulación llevado a cabo para ajustar los modelos estudiados, considerando los tamaños de muestra y niveles de prevalencia definidos anteriormente. Se evaluó el grado de sobreestimación en cada estimación, tanto en los escenarios simulados como en el caso aplicado. Los resultados incluyen la prevalencia en pacientes expuestos y no expuestos, así como la razón de prevalencia para cada modelo, proporcionando una visión integral del desempeño de los modelos bajo diferentes condiciones

5.1. Estudio de Simulación

5.1.1. Caso $n = 369$ y $\lambda = 5$ bajo supuesto de homocedástico

Modelo	P Exp ¹	P No Exp ²	RP Clásica ³	IC (95 %) ⁴	A priori ⁵	RP Bayes ⁶	RC ⁷
Caso No.1 ($n = 369$): $\lambda = 5$ homocedástico (prevalencia expuestos = 0.40 y no expuestos 0.10)							
Logístico	0.46	0.09	4.93	(3.34, 7.29)	Normal (0,1)	3.80	(2.03, 6.57)
Log - Binomial	0.46	0.09	4.93	(3.34, 7.29)	Normal (0,1)	3.81	(2.03, 6.66)
Poisson	0.46	0.10	4.93	(2.67, 9.11)	Normal (0,1)	4.01	(2.20, 7.12)
Poisson Robusto	0.46	0.10	4.93	(2.77, 8.76)	Normal (0,1)	4.05	(2.13, 7.13)
Caso No.2 ($n = 369$): $\lambda = 5$ homocedástico (prevalencia expuestos = 0.60 y no expuestos 0.25)							
Logístico	0.60	0.27	2.19	(1.27, 3.27)	Normal (0,1)	2.09	(1.52, 2.94)
Log - Binomial	0.60	0.27	2.19	(1.27, 3.27)	Normal (0,1)	2.09	(1.53, 2.99)
Poisson	0.60	1.0	2.19	(1.56, 3.09)	Normal (0,1)	2.11	(1.53, 2.95)
Poisson Robusto	0.60	1.0	2.19	(1.67, 2.89)	Normal (0,1)	2.13	(1.47, 3.05)

¹ P Exp: Prevalencia estimada de expuestos.

² P No Exp: Prevalencia estimada de los no expuestos.

³ RP Clásica: Razón de prevalencia calculada por el método clásico.

⁴ IC (95 %): Intervalo de confianza del 95 % para la razón de prevalencia calculada por el método clásico.

⁵ A priori: Distribución a priori asociada al parámetro.

⁶ RP Bayes: Razón de prevalencia estimada a través de métodos bayesianos.

⁷ RC: Región de credibilidad.

Tabla 5-1: Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra($n=369$ y $\lambda = 5$).

Los resultados presentados en la Tabla 5-4 reflejan el desempeño de los modelos ajustados en dos escenarios diferentes de prevalencia.

En el Caso No.1 (prevalencia de expuestos = 0.40 y no expuestos = 0.10), la razón de prevalencia (RP) clásica se mantiene consistentemente alrededor de 4.93 en todos los modelos, con una sobreestimación de 0.93 unidades en comparación con el valor esperado de 4. Esta sobreestimación puede ser atribuida a la variabilidad inherente en los modelos y la estructura de los datos simulados. Las estimaciones bayesianas, en contraste, muestran una ligera reducción en la RP, con valores cercanos a 3.80 para el modelo logístico, en comparación con el valor clásico de 4.93. Esta diferencia se debe a la influencia de la distribución a priori, que tiende a moderar las estimaciones. Las regiones de credibilidad bayesianas, que son más estrechas, indican una mayor precisión en las estimaciones al incorporar información previa.

En cuanto a los intervalos de confianza (IC) para la RP clásica en el Caso No.1, se observan amplios rangos de variabilidad, como el IC de 2.67 a 9.11 para el modelo Poisson. Estos intervalos amplios reflejan una considerable incertidumbre en las estimaciones, aunque todos los IC incluyen el verdadero valor de la razón de prevalencia, sugiriendo que las estimaciones, a pesar de su variabilidad, son consistentes entre los diferentes modelos.

En el Caso No.2 (prevalencia de expuestos = 0.60 y no expuestos = 0.25), la razón de prevalencia estimada con el método clásico es aproximadamente 2.19, cercana al valor de simulación de 2.4, lo que evidencia una ligera subestimación de 0.21 unidades. Las estimaciones bayesianas para el modelo logístico y log-binomial son consistentes en 2.09, mientras que los modelos Poisson y Poisson Robusto muestran una pequeña diferencia entre ellos, de 0.02 unidades. En este caso, todas las razones de prevalencia presentan una subestimación de entre 0.37 y 0.40 unidades.

Los intervalos de confianza para el Caso No.2 muestran amplitudes de 2 unidades para los modelos logístico y log-binomial, mientras que los modelos Poisson y Poisson Robusto tienen amplitudes menores de 1.53 y 1.22, respectivamente. Estos resultados sugieren una mayor precisión en los modelos Poisson y Poisson Robusto. Además, las regiones de credibilidad bayesianas son más precisas, reflejando intervalos más ajustados que pueden ser beneficiosos en estudios epidemiológicos.

En resumen, tanto en el Caso No.1 como en el Caso No.2, se observa una consistencia en la razón de prevalencia clásica entre los diferentes modelos. Sin embargo, las estimaciones bayesianas tienden a ser más conservadoras, posiblemente debido a la inclusión de información

previa a través de distribuciones a priori. Esta diferencia se traduce en estimaciones de RP más bajas en los modelos bayesianos en comparación con los clásicos. Además, los amplios intervalos de confianza para las estimaciones clásicas reflejan una variabilidad considerable y una mayor incertidumbre en la precisión de los resultados, mientras que las estimaciones bayesianas ofrecen una mayor precisión debido a la inclusión de información previa adecuada.

5.1.2. Caso $n = 384$ y $\lambda = 2$ bajo supuesto de heterocedasticidad

Modelo	P Exp ¹	P No Exp ²	RP Clásica ³	IC (95 %) ⁴	A priori ⁵	RP Bayes ⁶	RC ⁷
Caso No.3 (n = 384): $\lambda = 2$ heterocedástico (prevalencia expuestos = 0.40 y no expuestos 0.10)							
Logístico	0.45	0.10	4.19	(2.32, 7.57)	Normal (0,1)	3.41	(1.87, 5.95)
Log - Binomial	0.45	0.10	4.19	(2.32, 7.57)	Normal (0,1)	3.35	(1.91, 5.64)
Poisson	4.19	1.0	4.19	(2.30, 7.63)	Normal (0,1)	3.52	(1.98, 6.15)
Poisson Robusto	4.19	1.0	4.19	(2.40, 7.32)	Normal (0,1)	3.52	(1.98, 6.15)
Caso No.4 (n = 384): $\lambda = 2$ heterocedástico (prevalencia expuestos = 0.60 y no expuestos 0.25)							
Logístico	0.56	0.28	2.01	(1.11, 3.65)	Normal (0,1)	1.93	(1.36, 2.76)
Log - Binomial	0.56	0.28	2.01	(1.11, 3.65)	Normal (0,1)	1.93	(1.38, 2.67)
Poisson	2.01	1.0	2.01	(1.43, 2.83)	Normal (0,1)	1.97	(1.39, 2.72)
Poisson Robusto	2.01	1.0	2.01	(1.53, 2.65)	Normal (0,1)	1.95	(1.39, 2.72)

¹ P exp: Prevalencia estimada de expuestos

² P No exp: Prevalencia estimada de los no expuestos

³ RP Clásica: Razón de prevalencia calculada por el método clásico

⁴ IC (95 %): Intervalo de confianza del 95 % para la RP Clásica

⁵ A priori: Distribución a priori asociada al parámetro

⁶ RP Bayes: Razón de prevalencia estimada a través de métodos bayesianos

⁷ RC: Región de credibilidad

Tabla 5-2: Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra ($n=384$ y $\lambda = 2$).

En la Tabla 5-4 se presentan los resultados de los casos No. 3 y No. 4, ambos con un tamaño de muestra de 384 y un supuesto de heterocedasticidad, para diferentes prevalencias de exposición y no exposición.

Caso No. 3: La prevalencia de exposición es del 40 % y la de no exposición es del 10 %, lo que implica una razón de prevalencia de 4. Los resultados indican que la razón de prevalencia (RP) estimada utilizando los métodos clásicos es 4.19, con un intervalo de confianza del 95 % que varía entre 2.32 y 7.63. Esto muestra una alta amplitud en los intervalos de confianza, reflejando una considerable variabilidad en la estimación. A pesar de que todos los intervalos de confianza contienen el verdadero valor, los métodos bayesianos ofrecen estimaciones ligeramente más cercanas al valor real, con una diferencia que oscila

entre 0.48 y 0.59, lo que es menor que la diferencia observada en los métodos clásicos (0.19). La región de credibilidad bayesiana tiene una menor amplitud, de 0.73 a 4.17 unidades, lo que sugiere una estimación más precisa en comparación con la clásica.

Para el Caso No.4, la prevalencia de exposición es del 60 % y la de no exposición es del 25 %, resultando en una razón de prevalencia de 2.4.

La razón de prevalencia estimada mediante el método clásico es aproximadamente 2.01, con una subestimación de 0.39 unidades del valor real. Todos los modelos muestran una estimación consistente para este caso, aunque con ligeras variaciones en precisión. En los métodos bayesianos, los modelos logístico y log-binomial presentan estimaciones consistentes de la razón de prevalencia, al igual que los modelos Poisson y Poisson Robusto, con una diferencia mínima de 0.02 unidades entre ellos. Las estimaciones bayesianas se encuentran por debajo del valor real entre 0.43 a 0.47 unidades, y las regiones de credibilidad son más precisas que los intervalos de confianza clásicos, aunque ambos contienen el valor verdadero de la razón de prevalencia.

En resumen, en ambos casos (Tabla 5-4), los modelos bayesianos muestran una consistencia en las estimaciones de la razón de prevalencia, aunque las estimaciones bayesianas tienden a estar ligeramente más alejadas del valor real en comparación con las estimaciones clásicas. La menor amplitud de las regiones de credibilidad bayesianas en comparación con los intervalos de confianza clásicos sugiere una mayor precisión en las estimaciones bayesianas, especialmente en el Caso No. 3.

5.1.3. Caso $n = 1536$ y $\lambda = 4$ bajo supuesto de heterocedasticidad

Modelo	P Exp ¹	P No Exp ²	RP Clásica ³	IC (95 %) ⁴	A priori ⁵	RP Bayes ⁶	RC ⁷
Logístico	0.40	0.10	3.93	(2.46, 6.26)	Normal (0,1)	3.74	(2.79, 4.99)
Log - Binomial	0.40	0.10	3.93	(2.46, 6.26)	Normal (0,1)	3.70	(2.74, 5.00)
Poisson	3.93	1	3.93	(2.88, 5.27)	Normal (0,1)	3.74	(1.91, 5.64)
Poisson Robusto	3.93	1	3.93	(2.94, 5.24)	Normal (0,1)	3.74	(2.73, 5.05)
Logístico	0.61	0.23	2.56	(2.16, 3.03)	Normal (0,1)	2.56	(2.14, 3.01)
Log - Binomial	0.61	0.23	2.56	(2.16, 3.03)	Normal (0,1)	2.53	(2.14, 3.02)
Poisson	2.01	1	2.56	(2.15, 3.05)	Normal (0,1)	2.53	(2.14, 3.02)
Poisson Robusto	2.01	1	2.56	(2.22, 2.95)	Normal (0,1)	2.53	(2.12, 2.99)

¹ P exp: Prevalencia estimada de expuestos

² P No exp: Prevalencia estimada de los no expuestos

³ RP Clásica: Razón de prevalencia calculada por el método clásico

⁴ IC (95 %): Intervalo de confianza del 95 % para la RP clásica

⁵ A priori: Distribución a priori asociada al parámetro

⁶ RP Bayes: Razón de prevalencia estimada a través de métodos bayesianos

⁷ RC: Región de credibilidad

Tabla 5-3: Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra ($n=1536$ y $\lambda = 4$).

En la Tabla 5-4 se presentan los resultados de los Casos No. 5 y No. 6, ambos con una muestra de 1536 individuos y bajo la condición de heterocedasticidad.

En el Caso No. 5, la prevalencia de exposición es del 40 % y la de no exposición es del 10 %. La estimación de la razón de prevalencia por métodos clásicos es 3.93, con una diferencia mínima de 0.07 unidades respecto al valor real. Los modelos bayesianos presentan estimaciones de la razón de prevalencia que varían entre 3.70 y 3.74, situándose por debajo del valor real en 0.26 unidades. Los intervalos de confianza para estos modelos oscilan entre 2.3 y 3.8 unidades, mientras que las regiones de credibilidad están en un rango similar de 2.2 a 3.73, mostrando variaciones muy similares. En ambos casos, tanto las regiones de credibilidad como los intervalos de confianza incluyen el valor real de la razón de prevalencia.

En el Caso No. 6, con una prevalencia de exposición del 60 % y una de no exposición del 25 %, la razón de prevalencia estimada por métodos clásicos es 2.56, lo que coincide con la estimación bayesiana del modelo logístico. Los modelos bayesianos restantes muestran una razón de prevalencia ligeramente inferior de 2.53, cercana al valor real de 2.4. Los intervalos de confianza en este caso tienen amplitudes que varían entre 0.73 y 0.9 unidades, mientras que las regiones de credibilidad son más precisas, oscilando entre 0.87 y 0.88 unidades.

Comparando ambos casos, se observa que los modelos bayesianos ofrecen regiones de credibilidad más precisas en comparación con los intervalos de confianza clásicos. Esto sugiere que los modelos bayesianos proporcionan estimaciones más precisas, especialmente en el Caso No. 6. Además, entre los modelos bayesianos, el modelo Poisson Robusto se destaca por ofrecer una región de credibilidad ligeramente más ajustada en comparación con otros modelos.

5.1.4. Caso $n = 6144$ y $\lambda = 8$ bajo supuesto de heterocedasticidad

Modelo	P Exp ¹	P No Exp ²	RP Clásica ³	IC (95 %) ⁴	A priori ⁵	RP Bayes ⁶	RC ⁷
Caso No. 7							
Logístico	0.40	0.10	4.19	(1.08, 1.61)	Normal (0,1)	4.17	(3.60, 4.83)
Log - Binomial	0.40	0.10	4.19	(1.08, 1.61)	Normal (0,1)	4.13	(3.55, 4.77)
Poisson	0.40	0.10	4.19	(3.61, 4.86)	Normal (0,1)	4.17	(3.57, 4.82)
Poisson Robusto	0.40	0.10	4.19	(3.65, 4.86)	Normal (0,1)	4.13	(3.57, 4.81)
Caso No. 8							
Logístico	0.60	0.25	2.27	(1.61, 3.19)	Normal (0,1)	2.29	(2.10, 2.50)
Log - Binomial	0.60	0.25	2.27	(1.61, 3.19)	Normal (0,1)	2.27	(2.07, 2.45)
Poisson	0.60	0.25	2.27	(2.08, 2.47)	Normal (0,1)	2.27	(2.07, 2.47)
Poisson Robusto	0.60	0.25	2.27	(2.12, 2.43)	Normal (0,1)	2.27	(2.08, 2.47)

¹ P exp: Prevalencia estimada de expuestos

² P No exp: Prevalencia estimada de los no expuestos

³ RP Clásica: Razón de prevalencia calculada por el método clásico

⁴ IC (95 %): Intervalo de confianza del 95 % para la razón de prevalencia calculada por el método clásico

⁵ A priori: Distribución a priori asociada al parámetro

⁶ RP Bayes: Razón de prevalencia estimada a través de métodos bayesianos

⁷ RC: Región de credibilidad

Tabla 5-4: Resultados de los ajustes de modelos para diferentes escenarios de prevalencia y tamaños de muestra ($n=6144$ y $\lambda = 8$).

En la tabla 5-4 se muestran los casos No. 7 y No. 8, ambos analizados con una muestra de 6144 individuos bajo la condición de heterocedasticidad.

En el Caso No. 7 (prevalencia de expuestos = 0.40 y no expuestos = 0.10), la razón de prevalencia (RP) se estima en 4.19, sin importar el modelo utilizado. Para la estimación bayesiana, los modelos Logístico y Poisson presentan una estimación de 4.17, mientras que los modelos Log-Binomial y Poisson Robusto estiman 4.13. Los intervalos de confianza para esta razón oscilan entre 0.53 y 1.25 unidades, mientras que las regiones de credibilidad varían entre 1.22 y 1.25. Esto indica que las estimaciones bayesianas ofrecen una mayor precisión en comparación con los intervalos de confianza.

En el Caso No. 8 (prevalencia de expuestos = 0.60 y no expuestos = 0.25), la razón de prevalencia se estima en 2.27 para todos los modelos, excepto para el modelo Logístico, que ofrece una estimación ligeramente superior de 2.29. Los intervalos de confianza para la razón de prevalencia están entre 0.31 y 1.58, mientras que la estimación bayesiana muestra una amplitud entre 0.38 y 0.40. Esto demuestra que las estimaciones bayesianas proporcionan una precisión más alta en este caso.

En ambos casos, las estimaciones bayesianas tienden a ofrecer mayor precisión en comparación con las estimaciones clásicas y los intervalos de confianza.

5.2. Caso Aplicado: Unidad de Hemato - Oncología Pediátrica en una clínica nivel IV en Cali

En esta sección se presentan los resultados obtenidos a partir del análisis estadístico realizado sobre los datos de estudio. Se compararon diferentes modelos para evaluar su desempeño y ajustar adecuadamente las Razones de Prevalencia (RP).

Claro, aquí tienes una versión revisada:

Los resultados del análisis del caso aplicado se basan en una muestra de 114 niños. De estos, 44 niños experimentaron un evento adverso, definido como cualquier incidente inesperado documentado en la historia clínica que haya causado una lesión, incapacidad, incremento en los días de hospitalización o incluso la muerte, y que no esté relacionado con la enfermedad subyacente, sino con la atención hospitalaria recibida. Amaya (2009)

Al 22 % (25) se les realizó Transplante de Médula Ósea. Factor que se entenderá como exposición.

	Evento adverso	Sin Evento adverso	Total
Expuestos	20	5	25
No expuestos	24	65	89

Tabla 5-5: Tabla de contingencia para caso aplicado (Evento adverso Vs Exposición a TMO)

- Prevalencia de expuestos: 0.80
- Prevalencia de No expuestos: 0.26

- Razón de Prevalencia: 2.96

Modelo	P Exp ¹	P No Exp ²	RP Clásica ³	IC (95 %) ⁴	Sobreestimación ⁵
Logístico	0.80	0.26	2.96	(2.90, 3.02)	0.79
Log - Binomial	0.80	0.26	2.96	(2.90, 3.02)	0.79
Poisson	2.96	1	2.96	(1.63, 5.37)	0.36
Poisson Robusto	2.96	1	2.96	(1.63, 5.37)	1.55

¹ P exp: Prevalencia estimada de expuestos

² P No exp: Prevalencia estimada de los no expuestos

³ RP Clásica: Razón de prevalencia calculada por el método clásico

⁴ IC (95 %): Razón de prevalencia calculada por el método clásico

⁵ Sobreestimación: La sobreestimación de la OR con respecto a la RP se calculó con la fórmula: [Sobreestimación = (OR-RP)/(OR-1)] Espelt et al. (2013)

Tabla 5-6: Comparación de modelos ajustados en el caso aplicado, metodología clásica

Modelo	Distribución a Priori	β_0	β_1	RP ¹	Región de Credibilidad
Logístico	Normal (0,1)	-4.37	0.93	2.62	(2.25, 4.12)
Log - Binomial	Normal (0,1)	-4.39	0.92	2.62	(1.67, 3.88)
Poisson	Normal (0,1)	-1.26	0.97	2.75	(1.99, 3.45)
Poisson Robusto	Gamma (1,1)	-0.07	0.37	1.47	(0.00, 2.47)

¹ RP: Estimación de la razón de prevalencia

Tabla 5-7: Comparación de modelos ajustados en caso aplicado, metodología Bayesiana

Acorde con los resultados anteriores se puede mencionar que:

Los modelos considerados proporcionan resultados iguales en términos de la razón de prevalencia desde un enfoque clásico. Sin embargo, el intervalo de confianza asociado a la razón de prevalencia es mucho más amplio cuando se trata del modelo Poisson y el modelo Poisson Robusto. En términos de estimaciones, estos resultados indican que tener una razón de prevalencia (RP) de 2.96 en el análisis clásico significa que los niños que recibieron el trasplante de médula ósea (expuestos) tienen casi tres veces más probabilidades de presentar eventos adversos en comparación con aquellos que no recibieron el trasplante (no expuestos), este valor sugiere una asociación fuerte y significativa entre el trasplante de médula ósea y la ocurrencia de eventos adversos.

En cuanto a las estimaciones considerando estimación bayesiana, se encuentra que el modelo Poisson Robusto se mantiene alejado de los modelos propuestos, teniendo una

estimación de 1.47 mientras que los modelos restantes se mantienen entre 2.62 a 2.75, indicando con ello que los niños con trasplante médula tiene mayor prevalencia de sufrir eventos adversos que el grupo que no fue trasplantado, esto sugiere una asociación positiva entre el trasplante y el evento adverso.

Exceptuando el caso del modelo Poisson Robusto, el resultado obtenido con las estimaciones clásicas y bayesianas son similares.

Ahora bien, evaluando los intervalos de confianza obtenidos, se observa que la amplitud mínima es de 0.12 y se presenta en el modelo Logístico y el modelo Log-Binomial. En términos de estimaciones bayesianas, el modelo Poisson y Logístico logran amplitud mínima de 1.46 y 1.87, respectivamente, mientras que el modelo Poisson y Poisson Robusto presentan la mayor amplitud con 3.74.

Es importante señalar que los modelos Logístico, Log-binomial y Poisson tienden a comportarse de manera similar cuando la prevalencia es alta y la frecuencia de eventos no es extremadamente baja. En situaciones donde la prevalencia de eventos adversos es muy alta o muy baja, modelos como el Logístico y el Log-binomial pueden producir resultados similares. Por ejemplo, si la prevalencia de eventos adversos en expuestos y no expuestos es significativamente diferente (como 0.40 y 0.10), la razón de prevalencia calculada puede ser idéntica en diversos modelos.

6 Conclusiones y recomendaciones

6.1. Conclusiones

6.1.1. Estudio de simulación

Los resultados de los estudios de simulación indican una consistencia notable en las razones de prevalencia (RP) estimadas tanto por métodos clásicos como por métodos bayesianos, con ligeras variaciones dependiendo del tamaño de la muestra y las condiciones de homocedasticidad o heterocedasticidad. En general, los modelos clásicos tienden a proporcionar estimaciones con intervalos de confianza más amplios, lo que refleja una mayor incertidumbre en comparación con las regiones de credibilidad de los modelos bayesianos.

En el caso de un tamaño de muestra de 369 y un parámetro λ de 5 bajo condiciones de homocedasticidad, la RP estimada por los métodos clásicos se mantuvo constante alrededor de 2.19 a 4.93 en todos los modelos, con intervalos de confianza amplios. Por otro lado, las estimaciones bayesianas fueron ligeramente más conservadoras, con RP alrededor de 2.09 a 4.05 y regiones de credibilidad más estrechas, sugiriendo una mayor precisión. En términos de seleccionar un modelo desde la estimación clásica sería el modelo Logístico y Log-Binomial, ya que la amplitud de su intervalo es de 3.95, en el caso de prevalencia de expuestos de 0.4 y no expuestos de 0.10; en el segundo caso prevalencia de expuestos de 0.60 y no expuestos de 0.25 el mejor modelo es el Poisson Robusto con una amplitud de intervalo de 1.22.

En el escenario con un tamaño de muestra de 384 y un parámetro λ de 2 bajo condiciones de heterocedasticidad, las RP estimadas 2.01 a 4.19 en todos los modelos, con intervalos de confianza amplios, mientras que las estimaciones bayesianas se mantuvieron entre 1.93 a 3.52 y regiones de credibilidad más estrechas, sugiriendo una mayor precisión. En términos de seleccionar un modelo desde la estimación clásica sería el modelo Poisson Robusto, ya que la amplitud de su intervalo es de 4.92; sin embargo, de manera general la estimación bayesiana se presenta como mejor alternativa de estimación, ya que la amplitud de las estimaciones se mantienen entre 3.73 a 4.17, en el caso de prevalencia de expuestos de 0.4 y no expuestos de 0.10; en el segundo caso prevalencia de expuestos de 0.60 y no expuestos de 0.25 el mejor modelo es el Poisson Robusto con una amplitud de intervalo de

1.12.

En el escenario con un tamaño de muestra de 1.536 y un parámetro λ de 4 bajo condiciones de heterocedasticidad, las RP estimadas por los métodos clásicos se mantuvieron entre 2.56 a 3.93 en todos los modelos, teniendo en cuenta las estimaciones clásicas. Las estimaciones bayesianas presentaron ligeras discrepancias, con RP alrededor de 2.53 hasta 3.74 y regiones de credibilidad más precisas, lo que nuevamente indica una mayor precisión en las estimaciones bayesianas. En términos de seleccionar un modelo desde la estimación clásica sería el modelo Poisson Robusto, ya que la amplitud de su intervalo es de 2,3, seguido de cerca por el modelo Poisson con una amplitud de 2.39; sin embargo, de manera general la estimación bayesiana con los modelos Logístico y Log-Binomial se presenta como mejor alternativa de estimación, ya que la amplitud de las estimaciones se mantienen entre 2.2 a 2.26, en el caso de prevalencia de expuestos de 0.4 y no expuestos de 0.10; en el segundo caso prevalencia de expuestos de 0.60 y no expuestos de 0.25 el mejor modelo es el Poisson Robusto con una amplitud de intervalo de 0.73.

Finalmente, para un tamaño de muestra de 6.144 y un parámetro λ de 8 bajo condiciones de heterocedasticidad, tanto las RP clásicas como las bayesianas se mantuvieron constantes en 2.27 a 4.19, con intervalos de confianza más estrechos para los modelos bayesianos (por ejemplo, de 3.57 a 4.82) en comparación con los modelos clásicos. En términos de seleccionar un modelo para el caso de prevalencia de expuestos de 0.4 y no expuestos de 0.10 se podría seleccionar el modelo Logístico y Log-Binomial, pues la amplitud del intervalo se mantiene en 0.53; en el segundo caso prevalencia de expuestos de 0.60 y no expuestos de 0.25 el mejor modelo es el Poisson Robusto con una amplitud de intervalo de 0.31.

Ahora en términos de los modelos propuestos, se encuentra que el modelo logístico resulta ser el método de mayor aplicabilidad, más no el de mayor confiabilidad para la realización de estimaciones de la razón de prevalencia, independientemente del diseño por el que hayan sido recolectados los datos, dado que este presenta significativos problemas de sobreestimación. Los métodos bayesianos son una buena alternativa de aproximación a los resultados de la razón de prevalencia cuando los datos se encuentran altamente sobreestimados o, por el contrario, subestimados.

En resumen, para tamaños de muestras grandes, el modelo Poisson Robusto es una alternativa más recomendable debido a su capacidad para producir intervalos de confianza más estrechos y estimaciones más cercanas a los resultados reales. Estos modelos han demostrado ser efectivos en minimizar la variabilidad y proporcionar estimaciones precisas de la razón de prevalencia. Además, el uso de métodos de conteo para llegar a las estimaciones permite

obtener el valor real del parámetro, asegurando una mayor exactitud en los resultados. La combinación de estos enfoques proporciona una metodología robusta para la estimación precisa de la prevalencia en estudios epidemiológicos.

6.1.2. Estudio de Datos Reales en la Unidad de Hemato-Oncología Pediátrica

El análisis de los datos reales de la Unidad de Hemato-Oncología Pediátrica en una clínica de nivel IV en Cali reveló varias conclusiones importantes sobre la prevalencia de eventos adversos y la efectividad de diferentes modelos estadísticos para la estimación de la razón de prevalencia (RP). La base de datos incluyó 114 niños, de los cuales 44 presentaron eventos adversos y 25 recibieron un trasplante de médula ósea, considerado como factor de exposición.

Los métodos bayesianos permiten la incorporación de información previa en el análisis, lo cual es especialmente útil en contextos clínicos donde existe conocimiento acumulado sobre los riesgos y eventos adversos. Esto ayuda a obtener estimaciones más precisas y realistas, especialmente en situaciones con tamaños de muestra limitados. Las estimaciones bayesianas tienden a ser más conservadoras comparadas con las estimaciones clásicas, lo cual puede resultar en intervalos de confianza más estrechos y precisos.

A pesar de la diferencia en las magnitudes de las estimaciones, todos los modelos sugieren una asociación positiva entre el trasplante de médula ósea y los eventos adversos, indicando que los niños que recibieron el trasplante tienen una mayor prevalencia de sufrir eventos adversos en comparación con el grupo no trasplantado. Los hallazgos destacan la importancia de utilizar enfoques estadísticos robustos para obtener estimaciones precisas y confiables, esenciales para la toma de decisiones informadas en la práctica clínica y la formulación de políticas de salud.

6.1.3. Consistencia de los resultados

Para concluir sobre la consistencia de los resultados obtenidos entre el caso aplicado y las investigaciones realizadas por otros autores, nos enfocaremos entre los aspectos relacionados con la estimación de la Razón de Prevalencia y los métodos estadísticos utilizados:

1. Métodos estadísticos utilizados:

- **Marco teórico:** Los estudios revisados exploraron diferentes métodos para estimar la RP, incluyendo regresiones de Cox, log-binomial, Poisson robusto, y métodos bayesianos. Se identificaron ventajas y desventajas de cada método dependiendo del tamaño de muestra, prevalencia, y presencia de covariables continuas.
- **Caso aplicado:** Se emplearon métodos bayesianos, resaltando su utilidad para incorporar información previa y obtener estimaciones precisas en contextos clínicos con tamaños de muestra limitados.

2. Desempeño de los métodos en estimación de la RP:

- **Marco teórico:** Se destacó que el método log-binomial puede ser más adecuado en ciertas condiciones, pero enfrenta problemas de convergencia. El método Poisson robusto ofrece estimaciones más precisas para muestras grandes. Los métodos bayesianos fueron recomendados para obtener estimaciones más precisas, especialmente cuando los datos están sobreestimados o subestimados.
- **Caso aplicado:** Se encontró que las estimaciones bayesianas tienden a ser más conservadoras y precisas, lo que se alinea con la observación de que los métodos bayesianos ofrecen intervalos de confianza más estrechos y precisos.

3. Resultados específicos:

- **Marco teórico:** Los estudios concluyeron que diferentes métodos pueden proporcionar estimaciones puntuales y de intervalo con variabilidad según las condiciones de los datos (prevalencia, tamaño de muestra, etc.). La regresión logística tiende a sobreestimar la RP.
- **Caso aplicado:** Se observó una asociación positiva entre el trasplante de médula ósea y los eventos adversos, consistente con las estimaciones de RP más precisas obtenidas a través de métodos bayesianos y sin el uso de modelos estadísticos.

En resumen, las conclusiones del estudio de datos reales son consistentes con las observaciones del estado del arte. Ambos destacan la importancia de seleccionar métodos estadísticos adecuados para la estimación de la RP, considerando las características específicas del conjunto de datos y el contexto clínico, y subrayan la ventaja de los métodos bayesianos en proporcionar estimaciones más precisas.

6.2. Recomendaciones

Para mejorar la eficacia y la precisión de los modelos utilizados en el estudio de simulación, es fundamental optimizar los parámetros iniciales y asegurar una correcta especificación de los mismos. Los problemas de convergencia observados al fijar la prevalencia inicial de no expuestos en 0.10 sugieren la necesidad de ajustes más precisos. Se recomienda realizar pruebas preliminares de sensibilidad para determinar los parámetros óptimos que faciliten la convergencia de los modelos y garanticen resultados robustos y fiables.

Se recomienda explorar el uso de diferentes distribuciones de probabilidad como a prior en futuros análisis y adicional establecer un proceso para la selección de hiperparámetros en los modelos utilizados en el estudio de simulación. Esto incluye la realización de pruebas de validación cruzada y el uso de técnicas de búsqueda automatizada como la optimización bayesiana o la búsqueda en malla. Estas técnicas pueden ayudar a identificar los conjuntos de hiperparámetros que proporcionan el mejor desempeño del modelo en términos de precisión y estabilidad.

7 Anexos

7.1. Anexos

7.2. Códificación en Rstudio

```
1 ##Objetivo No.3
2 ##Con toma de datos binomial
3 data_binomial <-Base_pacientes_Poisson _y_binomial
4
5 ##Modelo Logistico
6 modelo <- glm(data_binomial$eventoadversodef ~
7 data_binomial$tmo, family = "binomial")
8 summary(modelo)
9
10 # Calcular probabilidades ajustadas
11 probabilidades_ajustadas <- predict(modelo,
12 type = "response")
13
14 # Crear un nuevo dataframe con las probabilidades
15 #y la variable de exposici n
16 nuevo_data <- data.frame(prob = probabilidades_ajustadas,
17 Exposicion = data_binomial$tmo)
18
19 # Calcular prevalencia para expuestos y no expuestos
20 prevalencia_expuestos <- mean(nuevo_data$prob
21 [nuevo_data$Exposicion == 1])
22 prevalencia_no_expuestos <- mean(nuevo_data$prob
23 [nuevo_data$Exposicion == 0])
24 Razon_de_prevalencia <- prevalencia_expuestos/
25 prevalencia_no_expuestos
26
27 # Calcular intervalo de confianza para
28 #la raz n de prevalencia
```

```
29 # Aumentar el número máximo de iteraciones
30
31 modelo <- glm(eventoadversodef ~ tmo,
32 data = data_binomial, family = "binomial", maxit = 100)
33 intervalo_confianza <- confint(modelo, "tmo", level = 0.95)
34 intervalo_confianza
35
36 cat("Prevalencia para expuestos:",
37 prevalencia_expuestos, "\n")
38 cat("Prevalencia para no expuestos:",
39 prevalencia_no_expuestos, "\n")
40 cat("Razón de prevalencias:",
41 Razon_de_prevalencia, "\n")
42 cat("Intervalo de confianza
43 (95%) para la razón de prevalencia:",
44 intervalo_confianza, "\n")
45
46 # Calcular probabilidades ajustadas
47 probabilidades_ajustadas <- predict(modelo,
48 type = "response")
49
50 # Calcular el índice de Brier
51 brier_score <- mean((probabilidades_ajustadas -
52 data_binomial$eventoadversodef)^2)
53
54 # Calcular la sobreestimación
55 sobreestimacion <- brier_score /
56 (mean(data_binomial$eventoadversodef) *
57 (1 - mean(data_binomial$eventoadversodef)))
58
59 cat("Índice de Brier:", brier_score, "\n")
60 cat("Sobreestimación:", sobreestimacion, "\n")
61
62
63
64
65 ##Modelo Logbinomial
66 modelo_logbinomial <- glm(data_binomial$eventoadversodef ~
67 data_binomial$tmo, family = binomial(link = "logit"))
68 summary(modelo_logbinomial)
69
```

```
70 # Calcular probabilidades ajustadas
71 probabilidades_ajustadas_logbinomial <-
72 predict(modelo_logbinomial, type = "response")
73
74 # Crear un nuevo dataframe con las
75 #probabilidades y la variable de exposici n
76 nuevo_data_logbinomial <- data.frame(prob =
77 probabilidades_ajustadas_logbinomial,
78 Exposicion = data_binomial$tmo)
79
80 # Calcular prevalencia para expuestos y no expuestos
81 prevalencia_expuestos_logbinomial <-
82 mean(nuevo_data_logbinomial$prob
83 [nuevo_data_logbinomial$Exposicion == 1])
84 prevalencia_no_expuestos_logbinomial <-
85 mean(nuevo_data_logbinomial$prob
86 [nuevo_data_logbinomial$Exposicion == 0])
87 Razon_de_prevalencia_logbinomial <-
88 prevalencia_expuestos_logbinomial /
89 prevalencia_no_expuestos_logbinomial
90
91 # Mostrar resultados
92 cat("Prevalencia para expuestos (Log-binomial):",
93 prevalencia_expuestos_logbinomial, "\n")
94 cat("Prevalencia para no expuestos (Log-binomial):",
95 prevalencia_no_expuestos_logbinomial, "\n")
96 cat("Raz n de prevalencias (Log-binomial):",
97
98 Razon_de_prevalencia_logbinomial, "\n")
99
100 # Aumentar el n mero m ximo de iteraciones
101
102 modelo <- glm(eventoadversodef ~ tmo,
103 data = data_binomial, family = "binomial", maxit = 100)
104 intervalo_confianza <- confint(modelo_logbinomial, "tmo",
105 level = 0.95)
106 intervalo_confianza
107
108 # Calcular probabilidades ajustadas
109 probabilidades_ajustadas <- predict(modelo_logbinomial,
110 type = "response")
```

```

111
112 # Calcular el ndice de Brier
113 brier_score <- mean((probabilidades_ajustadas -
114 data_binomial$eventoadversodef)^2)
115
116 # Calcular la sobreestimaci n
117 sobreestimacion <- brier_score /
118 (mean(data_binomial$eventoadversodef) *
119 (1 - mean(data_binomial$eventoadversodef)))
120
121 cat(" ndice de Brier:", brier_score, "\n")
122 cat("Sobreestimaci n:", sobreestimacion, "\n")
123
124
125 ##Modelo Poisson
126
127 data_Poisson <-Base_Poisson
128 ##Se carga la base que incluye los meses de ingreso
129 # Crear un modelo Poisson
130 modelo_Poisson <- glm(data_Poisson $
131 eventoadversodef ~ data_Poisson $tmo +
132 data_Poisson $Ene + data_Poisson $Feb +
133 data_Poisson $Mar + data_Poisson $Abr +
134 data_Poisson $May + data_Poisson $Jun +
135 data_Poisson $Jul + data_Poisson $Ago +
136 data_Poisson $Sep + data_Poisson $Oct +
137 data_Poisson $Nov + data_Poisson $Dic, family = "Poisson ")
138
139 # Resumen del modelo
140 summary(modelo_Poisson )
141
142 # Suponiendo que 'exposicion' es tu variable de exposici n
143 nuevo_data_frame <- data.frame(data_Poisson $tmo)
144
145 # Realiza las predicciones
146 exposicion_prediccion <- predict(modelo_Poisson ,
147 newdata = nuevo_data_frame, type = "response")
148
149 # Muestra las predicciones
150 exposicion_prediccion
151

```

```

152 # Asumiendo que 'n' es el tamaño de la muestra
153 tasa_no_expuestos <- exposicion_prediccion[1] /
154 sum(data_Poisson $tmo == 0)
155 tasa_expuestos <- exposicion_prediccion[2] /
156 sum(data_Poisson $tmo == 1)
157 n<-114
158 # Muestra las tasas
159 cat("Tasa de eventos para no expuestos:",
160 tasa_no_expuestos, "\n")
161 cat("Tasa de eventos para expuestos:",
162 tasa_expuestos, "\n")
163
164 tasa_expuestos/tasa_no_expuestos
165
166 modelo_Poisson <- glm(eventoadversodef ~
167 data_Poisson $tmo + data_Poisson $Ene +
168 data_Poisson $Feb + data_Poisson $Mar +
169 data_Poisson $Abr + data_Poisson $May +
170 data_Poisson $Jun + data_Poisson $Jul +
171 data_Poisson $Ago + data_Poisson $Sep +
172 data_Poisson $Oct + data_Poisson $Nov +
173 data_Poisson $Dic, data = data_Poisson ,
174 family = "Poisson ", maxit = 100)
175 intervalo_confianza <- confint(modelo_Poisson ,
176 "tmo", level = 0.95)
177 intervalo_confianza
178
179 # Calcular probabilidades ajustadas
180 probabilidades_ajustadas <- predict(modelo_Poisson,
181 type = "response")
182
183 # Calcular el índice de Brier
184 brier_score <- mean((probabilidades_ajustadas -
185 data_Poisson $eventoadversodef)^2)
186
187 # Calcular la sobreestimación
188 sobreestimacion <- brier_score /
189 (mean(data_Poisson $eventoadversodef) *
190 (1 - mean(data_Poisson $eventoadversodef)))
191
192 cat(" Índice de Brier:", brier_score, "\n")

```

```

193 cat("Sobreestimaci n:", sobreestimacion, "\n")
194
195
196
197 ##Poisson Robusto
198
199 # Instalar y cargar el paquete sandwich
200 #para errores est ndar robustos
201 library(sandwich)
202
203 # Ajustar el modelo Poisson
204 modelo_Poisson_robusto <- glm(data_Poisson $
205 eventoadversodef ~
206 data_Poisson $tmo + data_Poisson $Ene +
207 data_Poisson $Feb +
208 data_Poisson $Mar + data_Poisson $Abr +
209 data_Poisson $May +
210 data_Poisson $Jun + data_Poisson $Jul +
211 data_Poisson $Ago +
212 data_Poisson $Sep + data_Poisson $Oct +
213 data_Poisson $Nov +
214 data_Poisson $Dic , family = "Poisson ")
215
216 # Calcular errores est ndar robustos
217 errores_robustos <- sqrt(diag(vcovHC(
218 modelo_Poisson_robusto)))
219
220 # Resumen del modelo con errores est ndar robustos
221 summary(modelo_Poisson_robusto)
222
223 # Suponiendo que 'exposicion' es tu variable de exposici n
224 nuevo_data_frame <- data.frame(data_Poisson $tmo)
225
226 # Realiza las predicciones
227 exposicion_prediccion_robusto <-
228 predict(modelo_Poisson_robusto,
229 newdata = nuevo_data_frame, type = "response")
230
231 # Asumiendo que 'n' es el tama o de la muestra
232 tasa_no_expuestos_robusto <-
233 exposicion_prediccion_robusto[1] /

```

```

234 sum(data_Poisson $tmo == 0)
235 tasa_expuestos_robusto <-
236 exposicion_prediccion_robusto[2]
237 / sum(data_Poisson $tmo == 1)
238
239 # Muestra las tasas
240 cat("Tasa de eventos para no expuestos (robusto):",
241 tasa_no_expuestos_robusto, "\n")
242 cat("Tasa de eventos para expuestos (robusto):",
243 tasa_expuestos_robusto, "\n")
244
245 # Calcula la raz n de prevalencia
246 razon_prevalencia_robusto <- tasa_expuestos_robusto /
247 tasa_no_expuestos_robusto
248 cat("Raz n de Prevalencia (robusto):",
249 razon_prevalencia_robusto, "\n")
250
251 modelo_Poisson_robusto1 <-
252 glm(eventoadversodef ~ data_Poisson $tmo +
253 data_Poisson $Ene +
254 data_Poisson $Feb + data_Poisson $Mar +
255 data_Poisson $Abr +
256 data_Poisson $May + data_Poisson $Jun +
257 data_Poisson $Jul +
258 data_Poisson $Ago + data_Poisson $Sep +
259 data_Poisson $Oct +
260 data_Poisson $Nov + data_Poisson $Dic,
261 data = data_Poisson ,
262 family = "Poisson ", maxit = 100)
263 intervalo_confianza1 <- confint(modelo_Poisson_robusto1,
264 "tmo", level = 0.95)
265 intervalo_confianza1
266
267 # Calcular probabilidades ajustadas
268 probabilidades_ajustadas <- predict(modelo_Poisson_robusto,
269 type = "response")
270
271 # Calcular el ndice de Brier
272 brier_score <- mean((probabilidades_ajustadas -
273 data_Poisson$eventoadversodef)^2)
274

```

```
275 # Calcular la sobreestimaci n
276 sobreestimacion <- brier_score /
277 (mean(data_Poisson $eventoadversodef) *
278 (1 - mean(data_Poisson $eventoadversodef)))
279
280 cat(" ndice de Brier:", brier_score, "\n")
281 cat("Sobreestimaci n:", sobreestimacion, "\n")
```

Bibliografía

- Amaya, S. L. (2009), 'Conceptos y análisis de eventos adversos', *Cent Gest Hosp* **48**, 6–21.
- Bazán, J. and Bayes, C. (2010), 'Inferencia bayesiana en modelos de regresion binaria usando brmuw', *Reporte de Investigacion. Serie B. Nro* **25**.
- Bedrick, E. J., Christensen, R. and Johnson, W. (1996), 'A new perspective on priors for generalized linear models', *Journal of the American Statistical Association* **91**(436), 1450–1460.
- Blizzard, L. and Hosmer, W. (2006), 'Parameter estimation and goodness-of-fit in log binomial regression', *Biometrical Journal* **48**(1), 5–22.
- Bonilla Castillo, Y. M. Y. M. (2019), 'Un modelo logístico para la evolución de neonatos prematuros con bajo peso al nacer, atendidos en el hospital universitario del valle, durante el periodo 2002 a 2010'.
- Box, G. E. and Tiao, G. C. (2011), *Bayesian inference in statistical analysis*, John Wiley & Sons.
- Colimon, K.-M. (1990), *Fundamentos de epidemiología*, Ediciones Díaz de Santos.
- Coutinho, L., Scazufca, M. and Menezes, P. R. (2008), 'Métodos para estimar razão de prevalência em estudos de corte transversal', *Revista de Saúde Pública* **42**, 992–998.
- Diaz-Quijano, F. A. (n.d.), 'Artículo científico de revisión'.
- Espelt, A., Borrell, C., Palència, L., Goday, A., Spadea, T., Gnavi, R., Font-Ribera, L. and Kunst, A. E. (2013), 'Socioeconomic inequalities in the incidence and prevalence of type 2 diabetes mellitus in europe', *Gaceta Sanitaria* **27**, 494–501.
- Fajardo-Gutiérrez, A. (2017), 'Medición en epidemiología: prevalencia, incidencia, riesgo, medidas de impacto', *Revista Alergia México* **64**(1), 109–120.
- Ferrer, M. E. F. and del Prado González, N. (2013), 'Medidas de frecuencia y de asociación en epidemiología clínica', *Anales de Pediatría Continuada* **11**(6), 346–349.

- Gourieroux, C., Monfort, A. and Trognon, A. (1984), 'Pseudo maximum likelihood methods: Theory', *Econometrica: journal of the Econometric Society* pp. 681–700.
- Granados, J. A. T. (1995), 'Medidas de prevalencia y relación incidencia-prevalencia', *Med Clin (Barc)* **105**, 216–218.
- Gómez Gallego, A. (2019), 'Análisis de datos mediante regresión de poisson'.
- Hernández-Avila, M., Garrido-Latorre, F. and López-Moreno, S. (2000), 'Diseño de estudios epidemiológicos', *salud pública de méxico* **42**, 144–154.
- Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied logistic regression*, Vol. 398, John Wiley & Sons.
- Martinez, B. A. F., Leotti, V. B., Silva, G. d. S. e., Nunes, L. N., Machado, G. and Corbellini, L. G. (2017), 'Odds ratio or prevalence ratio? an overview of reported statistical methods and appropriateness of interpretations in cross-sectional studies with dichotomous outcomes in veterinary medicine', *Frontiers in veterinary science* **4**, 193.
- Melis, G. G. and Cadarso-Suárez, C. (2017), 'El modelo de riesgos proporcionales de cox y sus extensiones. impacto en estadística y biomedicina', *La Gaceta de la RSME* **20**(3), 513–538.
- Moreno-Altamirano, A., López-Moreno, S. and Corcho-Berdugo, A. (2000), 'Principales medidas en epidemiología', *Salud pública de México* **42**(4), 337–348.
- Páez, L. O. M., Lozano, M. R. and Davila, J. A. R. (2011), 'Descripción general de la inferencia bayesiana y sus aplicaciones en los procesos de gestión', *La simulación al Servicio de la Academia* **2**, 1–28.
- Petersen, M. R. and Deddens, J. A. (2008), 'A comparison of two methods for estimating prevalence ratios', *BMC medical research methodology* **8**(1), 1–9.
- Raftery, A. E. and Lewis, S. M. (1996), 'Implementing mcmc', *Markov chain Monte Carlo in practice* pp. 115–130.
- Tenny, S. and Hoffman, M. R. (2017), 'Odds ratio'.
- Tsou, T.-S. (2006), 'Robust poisson regression', *Journal of Statistical Planning and Inference* **136**(9), 3173–3186.
- Winkelmann, R. and Zimmermann, K. F. (1992), Robust poisson regression, in 'Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13–17 July 1992', Springer, pp. 201–206.