Universidad Autónoma de Querétaro

Facultad de Ingeniería

Licenciatura en
Matemáticas Aplicadas

Eficiencia de modelos binarios para datos por conglomerados

TESIS

Que como parte de los requisitos para obtener el grado de
Licenciado en Matemáticas Aplicadas

Presenta:

Ariadna Berenice Juárez Colunga

Dirigido por:

Dr. Elizabeth Juárez-Colunga y Dr. Eduardo Castaño Tostado

SINODALES

Dr. Elizabeth Juárez-Colunga Director	Firma
Dr. Eduardo Castaño Tostado Co-Director	Firma
Dr. Mario Santana Cibrian Sinodal	Firma
M. en C. Wilfrido Jacobo Paredes García Sinodal	Firma
Lic. Yesenia Morales Maciel Sinodal	Firma

Dr. Manuel Toledano Ayala
Director de la Facultad de Ingeniería

Centro Universitario
Querétaro, QRO
México.
Septiembre 2019

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

© 2019 - Ariadna Berenice Juárez Colunga

Todos los derechos reservados.

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

Esta tesis está dedicada a todas esas personas que llegaron a mi vida para permanecer en ella, a quienes se fueron, no porque quisieran, sino porque era su tiempo; a todos ellos gracias por contribuir de manera positiva en mi persona y en mi historia.

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

Agradecimientos

A Dios por otorgarme la paciencia y tranquilidad necesaria a lo largo de estos meses, porque cuando estuve a punto de soltar todo, Él puso en mi camino a las personas correctas para evitarlo.

A mi madre, quien desde siempre ha cuidado de mí, quien me formó, quien me enseñó sobre responsabilidad, disciplina, esfuerzo y se ha preocupado por mí en cada instante de mi vida.

A mi papá que me ha mostrado de sencillez y trabajo duro, quien me mostró que algunas veces la voluntad sola no basta, pero con buena voluntad, las cosas suceden.

A mi hermana Liz, por aceptar la aventura de apoyarme con esta tesis, por apoyarme y guiarme durante este año; por enseñarme un sentido más fuerte de la responsabilidad y la dedicación, porque a pesar de que no crecimos juntas siempre ha estado para mí y me ha dado de los consejos más acertados que he recibido.

A mis hermanos Fernando, Sheila, Candy y Erik, los cuales han estado siempre cuidándome como su bebé, con ellos he compartido diversas etapas de mi vida y en diferentes situaciones, les agradezco su paciencia, por ser todos (incluyendo a Liz) un ejemplo para mí de fortaleza y superación. Gracias por hacerme reír y por estar siempre juntos.

A Óscar por haber sido mi compañero en gran parte de toda esta aventura, por no dejarme desistir cuando quise hacerlo, por ser mi apoyo cuando creí que no podía.

A mis amigos y compañeros: mi amiga (Yesenia), Gyivan's, Vero, Fernando y Pepe, que de alguna forma siempre estuvieron para mí, escuchándome y ayudándome.

Al Dr. Eduardo Castaño, por todo el apoyo brindado en este trabajao, por haber sido uno de los profesores de los que más aprendí durante la licenciatura y quien, con sus clases, me motivó para concluirla.

A todas las personas que han recorrido este camino conmigo y que han aportado de alguna forma a mi vida.

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

Resumen

Los datos de conteo por conglomerados, conocidos por su estructura jerárquica o anidada, surgen en muchos estudios sociales, clínicos y epidemiológicos. Es común en la práctica encontrarse con modelos que no ajustan con los datos reales, por ejemplo, en el modelo de regresión Poisson lo anterior es casi una generalidad; es habitual encontrar que la esperanza de los datos es menor que la varianza $E(Y) < Var(Y)$ cuando los datos siguen una distribución Poisson $Y \sim Poisson(\lambda)$. En estos casos es común emplear el modelo Poisson con *efectos aleatorios*; sin embargo, hay otros métodos utilizados en la práctica con la intención de hacer más amigable el análisis, por ejemplo, haciendo un cambio de la variable de conteo a una variable binomial. En el siguiente trabajo se propondrán diversos escenarios para investigar la pérdida de potencia entre los modelos binarios que los analistas utilizan ocasionalmente, y el modelo de conteos. Esta investigación se realizó usando simulación para comparar la potencia de los diferentes modelos, estudio en el cual se encontró que sí hay una pérdida de potencia al hacer el cambio de variable.

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

Índice general

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

Índice de figuras

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

Índice de Tablas

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

CAPÍTULO 1

Introducción

Actualmente, en la práctica, es común analizar datos donde la variable dependiente toma valores discretos, y en muchos casos, valores discretos no negativos. Para esta tesis, la variable primaria de interés es una variable de conteo, es decir, un variable con valor entero no negativo. Es habitual encontrarse con este tipo de variables en el ámbito de los servicios de salud, por ejemplo, el número de veces que se acudió a los servicios de emergencia, el número de medicamentos prescritos o el número de días de estancia hospitalaria. Otras variables discretas comúnmente utilizadas, son las variables binarias, como su nombre lo dice, son aquellas que toman sólo dos valores, ya sea cero (0) o uno (1); ejemplos de ello puede ser el hecho de que un paciente tenga cierta enfermedad o no, si se le tomó la presión arterial o no, si se le administró algún tratamiento específico o no, entre otros.

En los servicios de salud, la forma estándar para probar si un nuevo tratamiento o intervención es efectivo para atender una enfermedad, son los estudios clínicos aleatorizados (ECA). En ocasiones, los individuos aleatorizados en el estudio se encuentran agrupados en hospitales, clínicas o centros especializados de enfermería (CEE), a estas agrupaciones se les llama *clusters* o conglomerados. El ejemplo que motiva este trabajo es un ECA donde los individuos están siendo monitoreados en algunos CEE, centros de enfermería donde se ayuda a pacientes que necesitan de atención intensiva, por ejemplo, pacientes de la tercera edad con insuficiencia renal o cardiaca que no tienen quién les atienda en casa. Para llevar a cabo un análisis de datos de los experimentos y hacer una inferencia adecuada, así como para tomar decisiones estratégicas basadas en información correcta, es necesario tomar en cuenta algunas diferencias en las características de los clusters; por ejemplo, algunos CEE reciben individuos más enfermos que otros de acuerdo a sus normas de ingreso, o el número de individuos de un clúster a otro puede variar.

El problema particular que ha motivado este trabajo de tesis es un estudio de insuficiencia cardiaca en individuos de la tercera edad que son ingresados en 30 distintos CEE. Una de las variables de mayor importancia, y la variable de interés para esta tesis, es el número de veces que se les toma la presión arterial a los pacientes; cabe mencionar que la estancia de los individuos varían entre $15 - 50$ con una media de 21 días, ya que algunos son dados de alta antes o bien mueren durante su estancia.

En algunas ocasiones se opta por utilizar la variable respuesta de conteo para construir una variable respuesta binaria; la idea principal es marcar un punto de corte delimitado por el analista de acuerdo a su experiencia, así la variable binaria se construye siendo 0 para los individuos donde el conteo sea menor al punto de corte, y siendo 1 para los conteos que sean mayores. Esta tesis

Dirección General de Bibliotecas UAQ

se enfoca en conocer cuáles son los efectos de los resultados del análisis al usar la variable binaria en lugar de los conteos y encontrar la magnitud de pérdida de precisión. También es importante estudiar la potencia, pues ésta puede ayudar a planificar nuevos estudios, ya que los investigadores pueden, en general, articular mejor la probabilidad de resultado que el número promedio de casos para diferentes exposiciones.

1.1. Motivación

El principal ejemplo que motiva este trabajo es un experimento clínico realizado en centros de enfermería especializada (CEE) de Denver, CO, EU. El objetivo del experimento es probar un programa intensivo de atención a pacientes con insuficiencia cardiaca que necesitan un seguimiento especial antes de ser dados de alta a casa. La población de interés en este experimento son los individuos mayores de 40 años que se canalizan a los CEE con diagnóstico de insuficiencia cardiaca después de una hospitalización; nos referimos a esta primera hospitalización como el índice de hospitalización. La variable respuesta de mayor interés en este estudio fue hospitalización por cualquier causa dentro de los siguientes 30 días después del índice de hospitalización. Sin embargo, hay otras variables respuesta secundarias, en particular la que nos interesa en esta tesis es el número de veces que se le revisa la presión arterial a cada individuo, es decir, la variable respuesta es una variable de conteo. Relevante a esta variable respuesta, es importante notar que el tiempo de cada individuo en el CEE puede variar, ya sea porque mueren o porque los dan de alta (cuando ya no necesitan atención de enfermeras) y se pueden ir a su casa. Así que estrictamente la variable de interés es la razón de revisiones; en otras palabras, el número de revisiones dividido por el número de días en el CEE.

La intervención consiste en un programa intensivo de 7 componentes: documentación de la fracción de eyección, evaluación de síntomas y actividad, vigilancia diaria de peso y dieta, valoración médica, educación del paciente y el cuidador, instrucciones de rehabilitación y un seguimiento de visitas dentro de los 7 días posteriores de darle de alta del CEE. El grupo control consiste en el cuidado que generalmente se les proporciona a los pacientes de insuficiencia cardiaca. En esta tesis nos enfocamos en estimar el efecto del programa intensivo respecto al cuidado usual, i.e. efecto del tratamiento o intervención.

Aunque este trabajo ha sido desarrollado tomando como motivación este estudio de CEEs en Denver, CO, EU, los resultados de esta tesis pueden ser útiles en otros contextos donde el estudio sea realizado por conglomerados y donde la variable respuesta sea un conteo. Por ejemplo, en México se realiza cada 6 años la Encuesta Nacional de Salud y Nutrición (ENSANUT), la cual consiste en un conjunto de preguntas que permiten conocer el estado de salud y nutrición de los mexicanos. En esta encuesta se les toman medidas a los individuos, una muestra de sangre del dedo y se les cuestiona sobre el consumo de algunos alimentos y bebidas, enfermedades crónicas, actividad física, vacunación, servicios de salud y programas sociales, entre otras cosas.

La ENSANUT está interesada en todos los hombres y mujeres mexicanos de las 32 entidades federativas del país, a los cuales divide en 4 diferentes rangos de edad: menores de 5 años, niños de 5-11 años, adolescentes de 12-19 años y adultos de 19 años en adelante.

En el contexto de esta tesis, los conglomerados de esta tesis corresponderían a los conglomerados de la encuesta. La variable respuesta podría ser por ejemplo las horas que el individuo pasa a la semana realizando actividad física o el número de horas que pasa frente al televisor. Y el efecto de interés podría ser la diferencia en actividad física u horas frente al televisor entre distintos

Dirección General de Bibliotecas UAQ

grupos, por ejemplo, grupos de edad o niveles de escolaridad.

1.2. Formulación del problema

Se tiene un ensayo clínico aleatorizado por conglomerados, donde la variable respuesta está dada por el número de veces que se les revisó la presión arterial a los individuos durante su estancia en la clínica; además, la estadía de los individuos en el lugar varía. Dicho esto, la variable respuesta debería ser la razón del número de eventos entre el número de días. Sin embargo, en ocasiones analistas optan por simplificar la variable respuesta a una variable binaria, la cual se obtiene por medio de la variable de conteo; esto podría reducir la precisión de la estimación del efecto del tratamiento. Entonces las preguntas de interés son, ¿hay pérdida de precisión en la estimación del efecto del tratamiento?, si es así, ¿cuánta precisión y potencia de detectar un efecto en la intervención se pierde al dicotomizar la variable de conteo?

Hay varios modelos de estimación para el análisis de variables por conglomerados; éstos incluyen específicamente modelos con base en efectos aleatorios y métodos de Ecuaciones Estimadas Generalizadas (GEE). Para propósitos de esta tesis, hemos elegido concentrarnos en el GEE porque es muy robusto y puede proteger contra suposiciones respecto al tipo de covarianza que tienen los datos.

1.3. Objetivos

Investigar la pérdida de potencia cuando una variable de conteo se simplifica/transforma a una variable binaria: un análisis basado en una variable respuesta de conteo contra un análisis de una variable binaria obtenida por medio de la variable de conteo antes mencionada.

1.3.1. Objetivos Específicos

El objetivo específico de esta tesis es:

Comparar la potencia por simulación del análisis basado en una variable de conteo con respecto a una variable binaria construida a partir de la variable de conteo dentro de un contexto de un ensayo clínico aleatorizado (ECA). Específicamente estamos interesados en la potencia del estimador del efecto de la intervención.

1. Comparar la potencia cuando la variable binaria se construye con base en un punto de corte delimitado por la media.

2. Comparar la potencia cuando la variable binaria se construye con base en un punto de corte delimitado por la mediana.

Compararemos la potencia en varios escenarios motivados por ejemplos reales que encontramos en la práctica de ECA. Estos escenarios variarán respecto a el número de clusters en el ECA, el número de individuos por cada clúster, la variabilidad entre clusters y el efecto de la intervención.

Para propósitos de comparar modelos de la variable binaria que se usan en la práctica, consideraremos un modelo logístico y un modelo log binomial.

Dirección General de Bibliotecas UMQ

1.4. Estructura de la tesis

En el capítulo 2 discutiremos los modelos que se utilizarán para la simulación, el método de estimación específico que usaremos (GEE) y algunas definiciones generales de utilidad. En el capítulo 3 se presenta el estudio de simulación, incluyendo los detalles de los diferentes valores de los parámetros con los cuales estamos simulando, y se definirán las medidas de desempeño que se presentarán en los resultados. El capítulo 4 presenta los resultados de los estudios y la discusión de éstos, así como también se explica en breve el impacto de los resultados de esta tesis. Por último en el capítulo 5 se darán las conclusiones a las que se llegó, los logros y aprendizajes de este trabajo.

Dirección General de Bibliotecas UAQ

CAPÍTULO 2

Definiciones y modelos

2.1. Ensayos clínicos aleatorizados

Un ensayo clínico es un experimento controlado donde los sujetos de estudio son pacientes humanos. Su objetivo es evaluar la eficacia, la seguridad y las reacciones adversas de tratamientos o intervenciones médicas contra enfermedades o cualquier problema de salud [1]. En el caso de la prueba de un nuevo medicamento se comparan al menos dos tipos de tratamientos, y siempre debe haber un medicamento de control el cual variará dependiendo de los resultados que se necesiten del estudio:

- Medicamento de control pasivo (negativo). Éste se utiliza cuando se requiere comprobar la efectividad de un tratamiento y cuando los sujetos pueden no ser medicados, pues se utiliza un placebo, es decir, un medicamento inocuo con la misma apariencia que uno activo, o bien puede ser utilizado un medicamento estándar.

- Medicamento de control activo (positivo). Éste se utiliza cuando se necesita mejoría o igualdad con respecto al producto estándar utilizado, además en este caso a los pacientes no se les pude privar del medicamento.

Dentro de la investigación epidemiológica los ensayos clínicos controlados aleatorizados son un paradigma pues son los diseños más cercanos a ser un experimento por el control que se tiene sobre las condiciones de estudio, además que pueden establecer una relación causa-efecto si las siguientes especificaciones se realizan de manera correcta [1]:

1. Asignación de la maniobra de intervención mediante mecanismos de aleatorización en sujetos con características homogéneas que permiten garantizar la comparabilidad de poblaciones [1].

2. El uso de un grupo control (para la comparación no sesgada de dos posibles tratamientos).

3. Para poder comparar información es necesario cegar a los grupos de tratamiento y así minimizar los sesgos de información.

Dirección General de Bibliotecas UAQ

2.1.1. Ensayos clínicos aleatorizados por conglomerados

El tipo de ensayo clínico que nos interesa entender es un diseño de ensayos clínicos aleatorizados por conglomerados (clusters); estos se utilizan para evitar la "contaminación" entre individuos que pertenecen al mismo conglomerado, pues en un ensayo clínico convencional cada persona es una unidad de estudio [2]. En un diseño convencional el otorgamiento de tratamientos se hace aleatoriamente sobre cada individuo.

En dichos ensayos clínicos convencionales suponemos independencia entre los individuos, pues se cree ningún individuo sabe a qué grupo de tratamiento pertence; sin embargo, no siempre es posible cegar a los individuos ante esta situación; por ejemplo, supongamos un caso en el que aleatoriamente se les dan o no a los sujetos una serie de recomendaciones alimenticias para reducir la presión arterial, es claro que los sujetos sabrán si se les dio o no, con lo que es imposible enmascarar el tratamiento.

Para solucionar ese problema se usan los ensayos clínicos aleatorizados por conglomerados, donde se forman grupos de individuos, entonces la unidad de aleatorización [1] son estos grupos, por ejemplo se le podría aleatorizar a un grupo de pacientes de un centro de salud al tratamiento en el que sí se les dan las especificaciones de su dieta, y se le asigna a otro centro de salud los pacientes que pertenecerán al grupo control. Las ventajas en este diseño son que se evita la contaminación entre individuos y que además se espera que los individuos de un mismo grupo se comporten de forma similar.

Dada la naturaleza de este diseño, en la etapa de análisis la conglomeración debe ser tomada en cuenta, ajustando resultados debido a correlación dentro del grupo; también se se toma en cuenta un tamaño de muestra mayor al convencional afectado por el número de grupos y el tamaño de estos.

2.2. Métodos de análisis para variables de conteo

2.2.1. Regresión Poisson

La Regresión Poisson es un Modelo Lineal Generalizado (MLG) [3] donde la variable respuesta es una variable de conteo Y que ajusta a una distribución Poisson, la cual tiene como único parámetro lambda (λ), y ésta coincide con ser la esperanza y la varianza de la distribución, es decir, $E(Y) = Var(Y) = \lambda$.

Sea Y la variable dependiente y $x_i' = (x_{i0}, x_{i1}..., x_{in})'$ un vector de i variables independientes, donde Y sigue una distribución Poisson $Y \sim P(\lambda)$. El parámetro λ puede variar para cada individuo; por ejemplo a través de un vector de variables independientes fijas $x_i' = (x_{i0}, x_{i1}..., x_{in})'$, de modo que

$$\lambda(x_i) = \exp(x_i'\beta) \tag{2.1}$$

donde β es el vector correspondiente de coeficientes de x_i, $\beta = (\beta_0, \beta_1, ..., \beta_n)$.

Por lo tanto, la distribución de Y dadas $x_i = (x_{i0}, x_{i1}..., x_{in})$ es Poisson y viene dada por

$$P(Y_i = y_i|x_i) = \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$$

donde

$$E(Y_i|x_i) = \lambda(x_i) = e^{x_i'\beta} = \exp(\beta_0 + \beta_1 x_{i1}... + \beta_n x_{in})$$

Dirección General de Bibliotecas UAQ

y esta formulación se conoce como Modelo de Regresión Poisson(MRP) [3].

Frecuentemente, cuando la variable respuesta es un conteo, ésta se observa sobre un periodo determinado de tiempo. Para este caso es de interés modelar el número de eventos en el tiempo observado, en otras palabras, la tasa de eventos. Para este propósito se utiliza el término denominado *offset*.

Se define el *offset* A_{ij} como el denominador correspondiente a la tasa de eventos, el cual sirve para escalar la modelación de la media en el MRP [4]. Este término se usa como una variable más en un problema de regresión, con la diferencia de que el coeficiente de esta variable es 1. Este parámetro es utilizado cuando los datos se registran durante un período observado. Por ejemplo, un individuo que se observa durante 15 unidades de tiempo y tiene 5 eventos, tiene un *offset* de 15 y una tasa de $5/15 = 0{,}3$; mientras que un individuo que se observa durante esas mismas 15 unidades de tiempo y tiene 30 eventos, tiene también un *offset* de 15, pero una tasa de $30/15 = 2$. De esta forma, obtenemos la siguiente formulación para el MRP

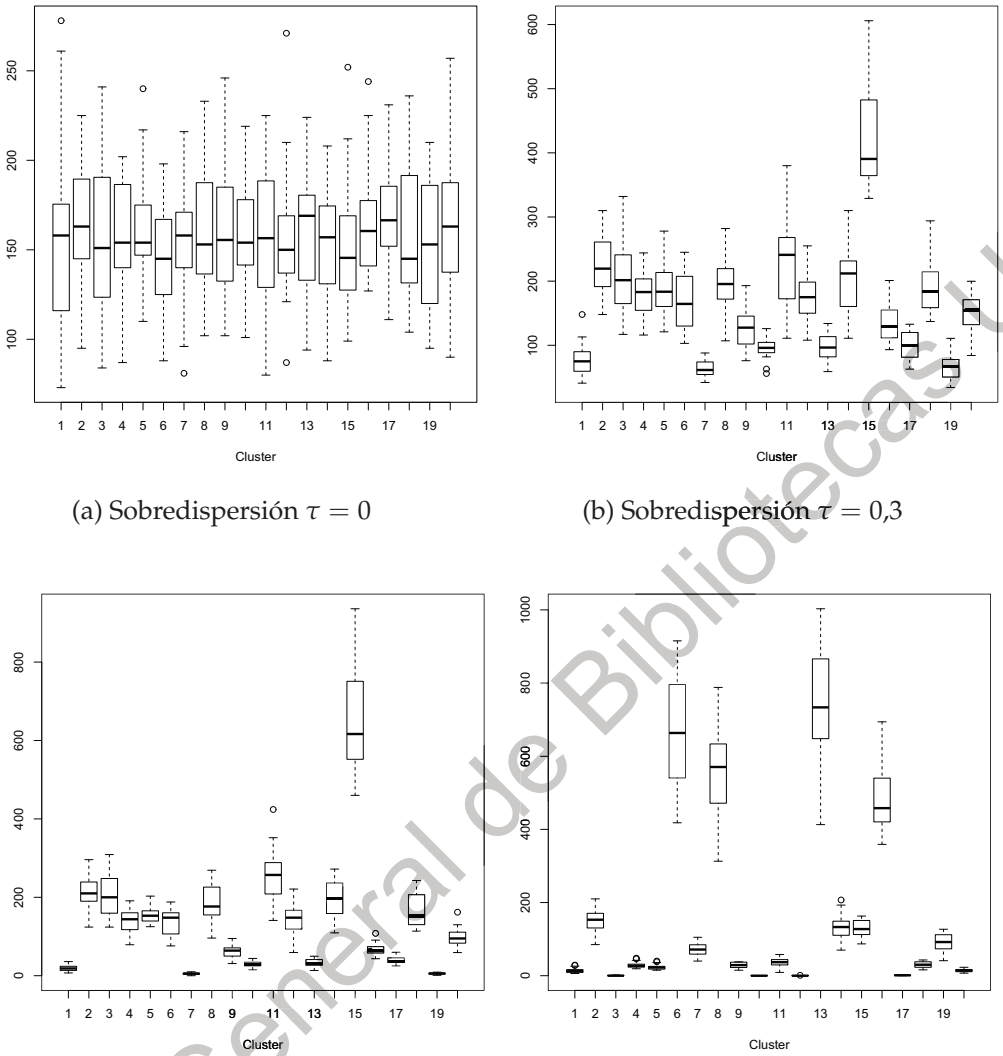
$$E(Y_i|x_i) = A_{ij}e^{x_i'\beta} = A_{ij}\exp(\beta_0 + \beta_1 x_{i1} ... + \beta_n x_{in}). \tag{2.2}$$

El modelo Poisson resulta muy restrictivo en la práctica y es común encontrar fenómenos como:

- Sobredispersión y sub-dispersión: esto ocurre cuando la varianza es mayor a la media (sobredispersión) o bien cuando la varianza es menor a la media (sub-dispersión) [5]. Se han propuesto algunas alternativas para generalizar el supuesto de la media y varianza iguales, por ejemplo el uso de errores estándar robustos (estimación de pseudomáxima verosimilitud), el empleo de un enfoque de cuasi-verosimilitud o los errores estándar bootstrap. En la figura 2.1 podemos observar la diferencia de la distribución de ejemplos de datos mientras la sobredispersión aumenta. Conforme la sobredispersión aumenta hay más variabilidad en los conteos por clúster, que se refleja en algunos clusters teniendo una distribución muy cerrada, mientras que otros tienen una muy dispersa.

- Exceso de ceros: existe una frecuencia de ceros que no es consistente con el modelo Poisson, por ejemplo, al modelar el número de cigarrillos fumados por cada uno de los integrantes de un grupo de personas puede suceder que una cantidad considerable de los individuos no sean fumadores.

Dirección General de Bibliotecas UAQ



(a) Sobredispersión $\tau = 0$

(b) Sobredispersión $\tau = 0{,}3$

(c) Sobredispersión $\tau = 1$

(d) Sobredispersión $\tau = 2{,}5$

Figura 2.1: En la figura se observan los diagramas de caja de los conteos de 20 clusters con 20 individuos cada uno, a diferentes escalas de sobredispersión.

2.3. Modelos con efectos aleatorios

La distribución Poisson es uno de los modelos más utilizados para datos de conteo, sin embargo, en aplicaciones biomédicas es muy raro el caso en que $E(Y) = Var(Y)$ pues normalmente hay sobredispersión, es decir, $E(Y) < Var(Y)$. Una forma de permitir que la varianza sea mayor que la esperanza en un modelo de conteos, es mediante el uso de efectos aleatorios. Los efectos aleatorios permiten modelar variabilidad en variables aleatorias, en particular en nuestro ejemplo del estudio en CEE, denotamos el efecto aleatorio v_j a nivel CEE_j, que sigue una distribución Gamma con varianza $Var(v_j) = \tau$ y esperanza $E(v_j) = 1$ sin pérdida de generalidad; $v_j \sim G(1/\tau)$.

Dirección General de Bibliotecas UAQ

A la varianza de los efectos aleatorios, es decir, a $Var(v_j) = \tau$ se le conoce como el estimador de *heterogeneidad*, esto es porque refleja la variabilidad entre grupos.

El modelo Poisson con efectos aleatorios supone que condicional en v_j, la variable Y_i sigue una distribución Poisson con media

$$\mu_{ij} = v_j \lambda_i = v_j \exp(x_i' \beta), \tag{2.3}$$

donde β corresponde como en (2.1) al coeficiente de regresión del vector de variables x_i.

Los modelos de efectos aleatorios pueden asumir varias distribuciones de v_j, por ejemplo Gamma, en cuyo caso la distribución marginal de los conteos es Binomial Negativa [5], la cual ha demostrado ser muy flexible para modelar conteos. En este caso

$$E(Y_{ij} = 1) = \mu_{ij} \qquad \text{y} \qquad Var(Y_{ij}) = \mu_{ij} + \tau \mu_{ij}^2 \tag{2.4}$$

2.3.1. Ecuaciones Estimadas Generalizadas

En varios modelos la verosimilitud no siempre es fácil de evaluar y maximizar, algunas veces puede ser insoluble e involucra muchos parámetros que no son de interés, por ejemplo la varianza de los efectos aleatorios. Por esta razón, una aproximación razonable cuando se presentan estos problemas, son las Ecuaciones de Estimación Generalizadas (GEE), un análogo multivariado de la cuasi-verosimilitud.

El enfoque GEE está basado en el concepto de *ecuaciones estimantes* y proporciona una aproximación muy general y uniforme para analizar variables respuesta correlacionadas que pueden ser discretas o continuas[6]. La idea esencial detrás de las GEE es generalizar y extender las ecuaciones de verosimilitud usuales para un modelo lineal generalizado de respuesta univariada, esto por medio de la incorporación de la matriz de covarianza del vector de respuestas Y_i. En términos simples, las GEE consisten en encontrar un estimador insesgado del parámetro de interés, el cual generalmente se encuentra simplificando el problema, y en (2.2) ajustar la varianza con base en un estimador robusto del *sándwich* [7] para que refleje mejor la estructura de covarianza de los datos.

En la ausencia de una función conveniente de la verosimilitud con la cual trabajar, es natural estimar β, el parámetro principal de interés en presencia de un parámetro secundario (α, que en nuestro caso corresponde a τ, el parámetro de heterogeneidad o sobredispersión), resolviendo un análogo multivariado de la función cuasi-score:

$$S_\beta(\beta, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta} \right)' Var(Y_i)^{-1} (Y_i - \mu_i) = 0 \tag{2.5}$$

donde i representa al individuo y m corresponde al número total de individuos.

En el caso multivariado, está la complicación adicional de que S_β depende de α tanto como de β ya que $Var(Y_i) = Var(Y_i; \beta, \alpha)$. Esto se puede superar reemplazando α en la ecuación anterior por una estimación consistente en $m^{1/2}$, $\hat{\alpha}(\hat{\beta})$. Algunos autores han demostrado que la solución de la ecuación resultante es asintóticamente tan eficiente como si α fuera conocida.

Los parámetros α pueden ser estimados por medio de la solución simultánea de $S_\beta = 0$ y

$$S_\alpha(\beta, \alpha) = \sum_{i=1}^{m} \left(\frac{\partial \eta_i}{\partial \alpha} \right)' H_i^{-1} (W_i - \eta_i) = 0, \tag{2.6}$$

Dirección General de Bibliotecas UAQ

donde $W_i = (R_{i1}R_{i2}, R_{i1}R_{i3}, ..., R_{in_{i-1}}R_{in_i}, R_{i1}^2, R_{i2}^2, ..., R_{in}^2)'$, el conjunto de todos los productos por pares de los residuales y los residuales al cuadrado con $R_{ij} = \{Y_{ij} - \mu_{ij}\}/\{v(\mu_{ij})\}^{1/2}$ y $\eta_i = E(W_i; \beta, \alpha)$. La elección de la matriz de peso, H_i, depende del tipo de respuestas. Para respuestas binarias, los últimos n_i componentes de W_i pueden ser ignorados ya que la varianza de una respuesta binaria es determinada por su media. En este caso podemos usar

$$H_i = \begin{pmatrix} Var(R_{i1}R_{i2}) & & & 0 \\ & Var(R_{i1}R_{i3}) & & \\ & & \ddots & \\ 0 & & & Var(R_{in_i-1}R_{in_i}) \end{pmatrix}$$

lo cual asegura que S_α depende de β y α solamente. Para datos de conteo, se sugiere el uso de la matriz de identidad $n_i^* \times n_i^*$ para H_i cuando se resuelve $S_\alpha = 0$, donde

$$n_i^* = \binom{n_i}{2} + n_i,$$

es el número de elementos de W_i.

La experiencia en aplicaciones ha sido que la elección de H_i tiene un impacto pequeño en la inferencia para β cuando m es grande. Además, para usar el peso adecuado, llamado $Var(W_i)$, se necesitaría hacer otras suposiciones acerca del tercer y cuarto momentos conjuntos de Y_i.

La solución, $(\hat{\beta}, \hat{\alpha})$, de $S_\beta = 0$ y $S_\alpha = 0$ es asintóticamente Gaussiana, con la varianza consistentemente estimada como

$$\left(\sum_{i=1}^{m} C_i' B_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{m} C_i' B_i^{-1} V_{0i} B_i^{-1} C_i\right) \left(\sum_{i=1}^{m} D_i' B_i^{-1} C_i\right)^{-1}, \tag{2.7}$$

evaluada en $(\hat{\beta}, \hat{\alpha})$, donde

$$C_i = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ 0 & \frac{\partial \eta_i}{\partial \alpha} \end{pmatrix}, \quad B_i = \begin{pmatrix} Var(Y_i) & 0 \\ 0 & H_i \end{pmatrix}, \quad D_i = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & \frac{\partial \mu_i}{\partial \alpha} \\ \frac{\partial \eta_i}{\partial \beta} & \frac{\partial \eta_i}{\partial \alpha} \end{pmatrix},$$

además

$$V_{0i} = \begin{pmatrix} y_i - \mu_i \\ \omega_i - \eta_i \end{pmatrix} \begin{pmatrix} y_i - \mu_i \\ \omega_i - \eta_i \end{pmatrix}'.$$

La ecuación (2.7) proporciona la fórmula del *estimador robusto del sándwich*, estimador que se usa frecuentemente para ajustar la varianza de datos correlacionados (conglomerados). De este modo la estimación del modelo se realiza de forma simplificada, pero al final (después del ajuste de varianza), la varianza refleja la naturaleza correlacionada de los datos.

2.3.2. Modelo de Regresión Logística

El modelo de regresión Logística es un tipo de modelo lineal generalizado (MLG)[8], éste usa la función logística para relacionar una variable respuesta dicotómica [9] a un conjunto de variables explicativas x_i. La forma de la función logística es:

Dirección General de Bibliotecas UAQ

$$P(Y=1) = \frac{\exp(\alpha + x_i'\beta)}{1+\exp(\alpha + x_i'\beta)}$$

de la cual se deriva el modelo de regresión logística.

Para nuestro ejemplo de motivación en particular, sea Y_{ij}^* la variable respuesta, donde $Y_{ij}^* = 1$ si la variable tiene el evento y $Y_{ij}^* = 0$ si no; el modelo de regresión logística con efectos aleatorios [6] está representado por

$$\log\left(\frac{P(Y_{ij}^*=1)}{1-P(Y_{ij}^*=1)}\right) = x_i'\beta + a_i \tag{2.8}$$

donde a_i es el efecto aleatorio y se asume que $a_i \sim N(0, \sigma_a^2)$, donde σ_a^2 es la varianza del efecto aleatorio.

2.3.3. Modelo de Regresión Log-Binomial

El modelo de regresión Log-Binomial es muy similar al modelo de regresión Poisson, pero siendo Y_{ij}^* la variable respuesta binaria, tendremos $Y_{ij}^* = 1$ si la variable tiene el evento y $Y_{ij}^* = 0$ si no; entonces

$$\log(P(Y_{ij}^*=1)) = x_i'\beta. \tag{2.9}$$

Sin embargo, cuando se aplica a este tipo de datos (binomiales), se debe usar un procedimiento llamado varianza de error robusto, o también conocido como estimación del sándwich, esto conduce a un modelo que llamaremos de regresión Poisson modificada [10].

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

CAPÍTULO 3

Estudio de simulación

3.1. Modelo de simulación

Usamos el modelo Poisson con efectos aleatorios para simular datos (2.4) entonces, sea v_j el efecto aleatorio que explica la heterogeneidad en el clúster j y sigue una distribución $G_v(\cdot)$. Sin pérdida de generalidad, se puede tomar $E(v_j) = 1$ y sea $Var(v_j) = \tau$. Condicional en el efecto aleatorio v_j, se define n_{ij} como el número de conteos en el clúster j para el individuo i y sigue una distribución Poisson con media

$$\mu_{ij} = A_{ij} \exp(\beta_0 + x_i' \beta_1) \tag{3.1}$$

donde $x_i = 1$ si el individuo está en el grupo tratamiento, y 0 si no; además A_{ij} es el número de días de exposición al tratamiento y β corresponde a los parámetros del vector x_i. Se llevarán a cabo $I = 1000$ réplicas (conjuntos de datos simulados) para cada uno de los escenarios.

3.1.1. Modelos a usar

Para comparar la potencia de un modelo en datos binarios con la potencia del modelo Poisson con efectos aleatorios (el usado para generar los datos). En datos por conglomerados usamos los siguientes modelos:

1. **Modelo Logístico con ajuste GEE:** (2.8) donde la variable binaria ha sido definida por la media o la mediana, sin tomar en cuenta los conglomerados.

2. **Modelo Log Binomial con ajuste GEE:** (2.9) donde la variable binaria ha sido definida por la media o la mediana, sin tomar en cuenta los conglomerados.

Para simplificar notación, a partir de esta sección omitimos referencia al ajuste GEE, pero todos los modelos usados (Poisson, Logístico y Log binomial) llevan ajuste GEE debido a correlación de individuos dentro del mismo clúster.

3.1.2. Parámetros de simulación

Se examinan distintos valores para los diferentes parámetros que se definen para el modelo (3.1)

Dirección General de Bibliotecas UAQ

1. Nivel de la media de conteos en el grupo control β_0. Que $\beta_0 = 2$ implica que la media de conteos del grupo control sea grande, es decir, habrá muchos más eventos; en otro escenario $\beta_0 = -2$ implica menos eventos, y este valor refleja con más cercanía el nivel de eventos del ejemplo de motivación. Por lo tanto se elige $\beta_0 = \{-2, 0, 2\}$. Como ejemplo la figura 3.1 muestra tres conjuntos de datos que corresponden a niveles de conteo bajos ($\beta_0 = -2$), moderados ($\beta_0 = 0$) y altos ($\beta_0 = 2$).

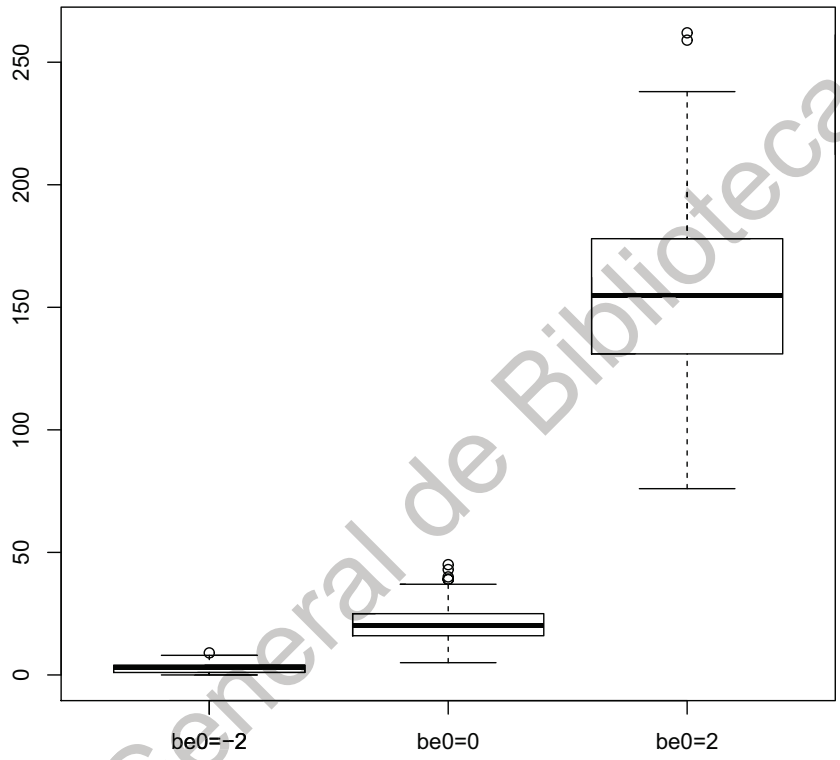


Figura 3.1: Ejemplos de conjuntos de datos con niveles de media de conteo bajos ($\beta_0 = -2$), moderados ($\beta_0 = 0$) y altos ($\beta_0 = 2$). La mediana y rango de los datos con conteos bajos son $med(Q1 - Q3)$: $3(1 - 4)$; conteos medios $20(16 - 25)$ y conteos altos $155(131 - 178)$.

2. Efecto del tratamiento/intervención β_1. Cuando el efecto del tratamiento es grande, la potencia bajo todos los modelos es cercana, y como la potencia está acotada por 1, entonces es difícil distinguir la diferencia; por esta razón se decide $\beta_1 : [0, 3, 5]$

3. El número de días n_{ij} se construye con una variable Poisson con media $\lambda = 21$, ya que de acuerdo al ejemplo de motivación la media de los días que los individuos pasan en los CEE es de 21 aproximadamente; el tamaño de n_{ij} se determina de acuerdo a la cantidad de clusters j e individuos por clúster i.

Dirección General de Bibliotecas UAQ

4. Número de clusters m : $20, 30$. Se decide de esta forma ya que, en el ejemplo de motivación, hay aproximadamente 30 CEE en el programa, además también se requiere visualizar un escenario donde hay menos clusters (20 en este caso).

5. Número de personas por clúster M. En el ejemplo de motivación está documentado que se aceptan aproximadamente entre 15 y 50 personas por CEE; es necesario en este sentido, ver al menos dos escenarios, uno con una cantidad pequeña de individuos y otra con una cantidad más grande. Para este caso se tomará M : $20, 40$.

6. Varianza de los efectos aleatorios τ. En estudios reales es común encontrar valores de sobredispersión alrededor de 1, sin embargo es bueno considerar al menos tres diferentes valores, uno sin sobredispersión, otro con un valor de sobredispersión cercano al 1, y otro donde la varianza sea mayor. Se considerarán $\tau = \{0, 1{,}5, 2{,}5\}$.

3.2. Medidas de desempeño.

Por cada conjunto de datos simulados se evalúa el desempeño de los estimadores basado en el criterio que evalúa tanto estimadores puntuales como de intervalo. Los estimadores se evalúan por simulación con base en $I = 1000$ datos simulados.

1. Sesgo: el sesgo del efecto del tratamiento es calculado como $Bias(\hat{\beta}_1) = \hat{\beta}_1 - \beta_1$ donde $\hat{\beta}_1$ es el promedio del parámetro estimado para el efecto del tratamiento.

2. Varianza asintótica: $\hat{Var}(\hat{\beta}_1)$ es la varianza de los $I = 1000$ parámetros estimados del efecto del tratamiento.

3. Cobertura de los intervalos de confianza (IC): el desempeño de la estimación por intervalo del efecto del tratamiento estimado es evaluado por la cobertura del IC, construído mediante el cálculo del porcentaje de simulaciones cubriendo el verdadero β_1 dentro del rango del intervalo de confianza empírico del 95 %. El desempeño del modelo de análisis seleccionado será considerado bueno cuando la cobertura del 95 %−IC es cercana al 95 %.

4. Diferencia en Error Estándar (SE) ($\triangle SE$): esta medida es calculada como

$$\sqrt{\frac{\sum_{i=1}^{I}(SE(\hat{\beta}_{1,i}))^2}{I}} - \hat{SD}(\hat{\beta}_1)$$

   donde $SE(\hat{\beta}_{1,i})$ es el error estándar estimado de la $i-$ésima estimación del efecto del tratamiento, y $\hat{SD}(\hat{\beta}_1)$ es la desviación estándar de las I estimaciones del efecto del tratamiento.

5. Potencia: es la probabilidad de rechazar la hipótesis nula de que no hay efecto del tratamiento H_o: $\beta_1 = 0$. La potencia se calcula como el número de veces que se rechaza la hipótesis nula dividido por el número total de simulaciones. Cuando $\beta = 0$ este valor es el valor de significancia α.

Las primeras cuatro medidas sólo las calculamos para el modelo Poisson debido a que no tienen relevancia en los modelos binarios, dado que estos últimos manejan diferentes estimadores. Sin embargo, como la pregunta de la potencia para detectar el efecto del tratamiento sí tiene relevancia y es comparable en todos los modelos, nos concentramos en ésta. Dado que los datos son

Dirección General de Bibliotecas UAQ

simulados con base en un modelo Poisson con efectos aleatorios y el modelo Poisson de ajuste es GEE, es importante evaluar si los estimadores se desempeñan bien bajo Poisson GEE dado que estos dan la base para comparar con los modelos binarios GEE.

Dirección General de Bibliotecas UAQ

CAPÍTULO 4

Resultados y discusión

4.1. Resultados

En esta sección se presentan los resultados de cuatro distintos escenarios, todos fueron elegidos de acuerdo al ejemplo de motivación donde tenemos 30 clusters y una media de conteos baja, a partir de eso se incluyeron resultados donde se pueda observar qué ocurre cuando la cantidad de individuos por clúster es pequeña ($m = 20$) y cuando es más grande ($m = 40$); también donde se pueda observar qué pasa cuando se disminuye el número de clusters y cuando la media de conteos es alta ($\beta_0 = 2$).

Los modelos Poisson con ajuste GEE se desempeñan bien, sobretodo cuando la heterogeneidad es baja ($\tau = 0$). En este caso el sesgo es pequeño en todos los escenarios, la cobertura es ligeramente baja siendo alrededor de 92 % en lugar de 95 % (como se esperaría), y la diferencia entre el error estándar estimado y el empírico es pequeña conforme la sobredispersión aumenta. La varianza de los estimadores aumenta, como era de esperarse, dado que hay más variabilidad. Sin embargo, la cobertura resulta aún más subestimada con coberturas de alrededor de 90 % para $\tau = 1{,}5$ y 88 % para $\tau = 2{,}5$. La diferencia entre el error estándar estimado y el empírico es más grande indicando que el error estándar es probablemente subestimado, lo cual a su vez causa algunos problemas de cobertura. Es común encontrar ligeros problemas de cobertura cuando se usan GEE, sin embargo, en general y en este caso en particular no son tan serios y el modelo Poisson GEE permite establecer la base para comparar los otros modelos.

La figura 4.1 presenta la potencia para detectar el efecto del tratamiento β_1 para tres escenarios de sobredispersión $\tau = 0, 1{,}5, 2{,}5$, todos los casos con 30 clusters ($m = 30$), 20 individuos por clúster ($M = 20$) y una media baja de conteos en el grupo control ($\beta_0 = -2$). Cuando no hay sobredispersión ($\tau = 0$), la potencia de los modelos binarios es muy similar al modelo Poisson (M1). Es decir, que no se pierde mucha información al convertir la variable Poisson, además ambos modelos binarios basados en la media y la mediana son similares. La potencia disminuye conforme la sobredispersión aumenta en todos los casos de uso de modelos, y la potencia del Modelo Poisson (M1) es más grande que la de los otros modelos Log Binomial (M2) y Logístico (M3). Sin embargo, la diferencia es mas pronunciada en el caso de sobredispersión mas alto (tau=2.5). Por ejemplo, cuando la sobredispersión $\tau = 1{,}5$ y el efecto del tratamiento $\beta_1 = 1{,}5$, la potencia de M1 es 0.865 y la de M2 y M3 son 0.769 y 0.800, respectivamente. Mientras que cuando $\tau = 2{,}5$, cuando el efecto del tratmiento $\beta_1 = 1{,}5$ la potencia de M1 es 0.733, y la potencia de M2 y M3 son 0.519 y 0.560,

Dirección General de Bibliotecas UAQ

respectivamente. La potencia de M2 y M3 es similar pero la del modelo logístico (M3) es ligeramente mayor que la del modelo Log Binomial (M2). Los modelos binarios definidos con base en la media se desempeñan un poco mejor que con base en la mediana (ver tablas 4.1, 4.2 y 4.3).

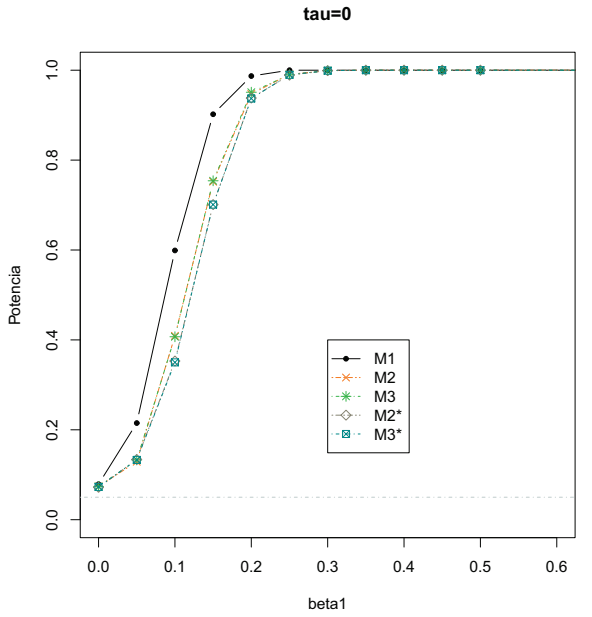
En la figura 4.2 tenemos una media de conteos más alta $\beta_0 = 2$; en el caso $\tau = 0$ podemos notar que la potencia aumenta con más rapidez hacia 1, además hay poca diferencia entre los tres modelos. En general, en este caso se observa también que la diferencia entre el M1 contra los otros dos modelos es más grande y que esta diferencia se acentúa mientras τ crece; podemos ver que, nuevamente la potencia del modelo Logístico (M3) se desempeña mejor que el Log binomial (M2). La disimilitud es mucho más pronunciada entre los modelos M2 y M3 respecto a M1 que cuando los conteos son bajos, además conforme la dispersión aumenta, esta disimilitud se acentúa. Veamos por ejemplo, cuando la sobredispersión $\tau = 1{,}5$ y el efecto del tratamiento $\beta_1 = 1{,}5$, la potencia de M1 es 0.91 y la de M2 y M3 son 0.58 y 0.68, respectivamente. Mientras que cuando $\tau = 2{,}5$, cuando el efecto del tratmiento $\beta_1 = 1{,}5$ la potencia de M1 es 0.791, y la potencia de M2 y M3 son 0.359 y 0.450, respectivamente. Similar al caso de conteos bajos, los estimadores de los modelos binarios se desempeñan un poco mejor con base en el corte de la media que con base en la mediana (ver tablas 4.4, 4.5 y 4.6).

En la figura 4.3 tenemos una media de conteos baja $\beta_0 = -2$, con número de clusters $M = 30$ y número de individuos por clúster más alto $M = 40$; en el caso $\tau = 0$ es notorio que la potencia aumenta con más rapidez hacia 1, además la diferencia es mínima entre los tres modelos. La disimilitud entre M1 y los dos modelos binarios es más grandre mientras τ crece; por ejemplo, cuando la sobredispersión $\tau = 1{,}5$ y el efecto del tratamiento $\beta_1 = 1{,}5$, la potencia de M1 es 0.886 y la de M2 y M3 son 0.773 y 0.808, respectivamente. Mientras que cuando $\tau = 2{,}5$, y el efecto del tratmiento es $\beta_1 = 1{,}5$ la potencia de M1 es 0.757, y la potencia de M2 y M3 son 0.542 y 0.571, respectivamente. Similar a los casos anteriores, la potencia del modelo log binomial (M3) se desempeña mejor que el Logístico (M2). Así como en los casos anteriores, los estimadores de los modelos binarios se desempeñan un poco mejor con base en el corte de la media que con base en la mediana (ver tablas 4.7, 4.8 y 4.9).

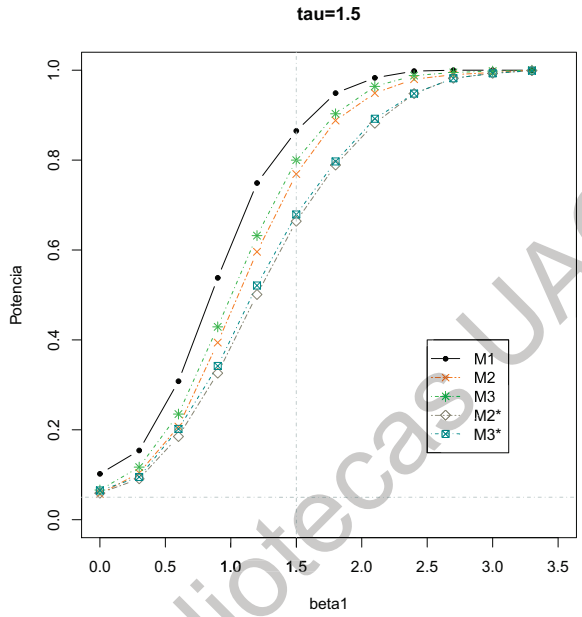
En la figura 4.4 tenemos una media de conteos baja $\beta_0 = -2$, con número de clusters bajo $m = 20$ y número de individuos por clúster $M = 20$. Podemos observar que cuando no hay sobredispersión ($\tau = 0$) la potencia de los tres modelos es muy similar, ésta incrementa rápidamente y llega a 1 cuando $\beta_0 = 0{,}3$. Cuando la sobredispersión aumenta a $\tau = 1{,}5$, vemos que sí hay diferencia entre M1 y los dos modelos binarios, se puede observar que cuando $\beta_1 = 1{,}5$ la potencia de M1 es 0.751 mientras que la de M2 y M3 es 0.583 y 0.685 respectivamente. Con lo anterior podemos constatar que además que la potencia del modelo log binomial (M3) se desempeña mejor que da del modelo Logístico (M2). En este caso no se presentó la gráfica para un nivel de sobredispersión mayor $\tau = 2{,}5$ ya que existía un exceso de ceros en los datos y muchas de las simulaciones no permitían estimación de los parámetros de manera estable (ver tablas 4.4 y 4.11).

Los resultados correspondientes al resto de los escenarios muestran tendencias similares a las discutidas en los párrafos anteriores. La potencia de M2 y M3 es más baja que M1, con M3 ligeramente mayor que M2 y los estimadores se desempeñan mejor cuando el corte es con base en la media que cuando es con base en la mediana (ver figuras A.1 - A.8 en el apéndice A).

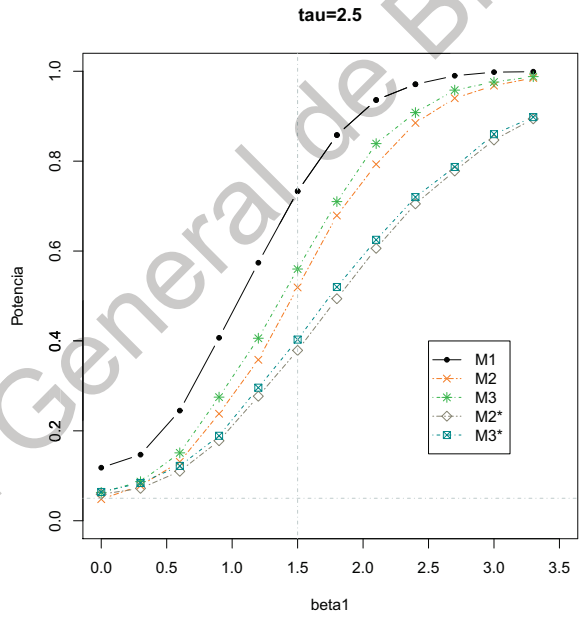
Dirección General de Bibliotecas UAQ



(a)



(b)



(c)

Figura 4.1: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = -2$, número de clusters $m = 30$ y número de individuos por clúster $M = 20$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

Dirección General de Bibliotecas UAQ

Cuadro 4.1: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 0$, con $m = 30$ clusters y $M = 20$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{Var}(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

							Media		Mediana	
β_1	Sesgo	$\hat{Var}(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	-5.21E-05	0.00218	0.921	0.0498	-0.00307	0.079	0.074	0.074	0.072	0.073
0.05	-2.30E-04	0.00213	0.921	0.0491	-0.00293	0.215	0.129	0.133	0.134	0.133
0.1	-2.64E-04	0.00208	0.925	0.0483	-0.00267	0.599	0.409	0.407	0.353	0.350
0.15	-5.00E-05	0.00203	0.924	0.0476	-0.00260	0.902	0.752	0.754	0.701	0.701
0.2	-1.56E-04	0.00199	0.923	0.0471	-0.00252	0.987	0.947	0.951	0.938	0.937
0.25	-1.10E-04	0.00194	0.923	0.0469	-0.00283	1.000	0.991	0.990	0.989	0.990
0.3	-6.94E-05	0.00190	0.922	0.0463	-0.00278	1.000	0.999	0.999	0.999	0.999
0.35	-1.57E-04	0.00186	0.922	0.0456	-0.00253	1.000	1.000	1.000	1.000	1.000
0.4	1.01E-04	0.00182	0.922	0.0454	-0.00268	1.000	1.000	1.000	1.000	1.000
0.45	1.20E-04	0.00179	0.918	0.0449	-0.00267	1.000	1.000	1.000	1.000	1.000
0.5	-5.60E-05	0.00175	0.922	0.0446	-0.00272	1.000	1.000	1.000	1.000	1.000
1	-1.76E-04	0.00149	0.917	0.0416	-0.00297	1.000	1.000	1.000	1.000	1.000
1.5	3.74E-04	0.00134	0.923	0.0390	-0.00243	1.000	1.000	1.000	1.000	1.000
2	8.92E-04	0.00124	0.922	0.0371	-0.00188	1.000	1.000	1.000	1.000	1.000
2.5	5.62E-04	0.00118	0.919	0.0365	-0.00210	1.000	1.000	1.000	1.000	1.000
3	7.26E-04	0.00115	0.919	0.0359	-0.00204	1.000	1.000	1.000	1.000	1.000
3.5	6.74E-04	0.00112	0.914	0.0355	-0.00204	1.000	1.000	1.000	1.000	1.000

Dirección General de Bibliotecas UAQ

Cuadro 4.2: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 1{,}5$, con $m = 30$ clusters y $M = 20$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{V}ar(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

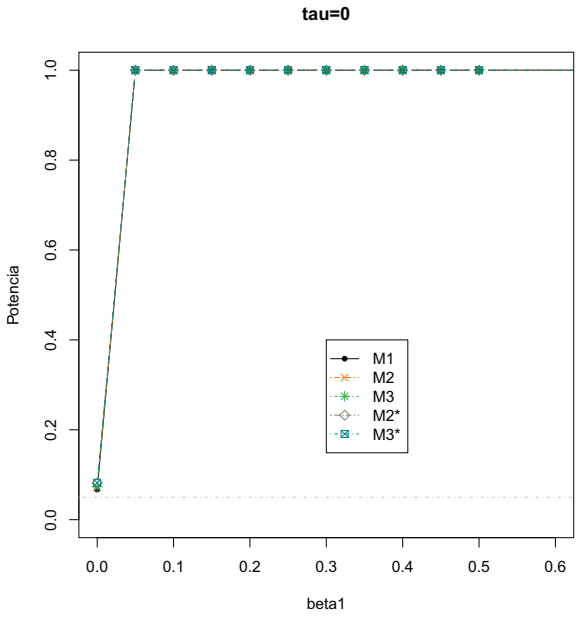
						Media			Mediana	
β_1	Sesgo	$\hat{V}ar(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0265	0.2073	0.898	0.5084	-0.05318	0.102	0.057	0.066	0.059	0.065
0.25	0.0330	0.2088	0.903	0.5107	-0.05369	0.147	0.093	0.100	0.079	0.090
0.5	0.0263	0.2117	0.893	0.5208	-0.06070	0.246	0.166	0.183	0.154	0.165
0.75	0.0297	0.2124	0.901	0.5219	-0.06112	0.410	0.298	0.319	0.241	0.257
1	0.0325	0.2174	0.901	0.5274	-0.06109	0.609	0.451	0.487	0.377	0.400
1.25	0.0334	0.2209	0.904	0.5295	-0.05948	0.764	0.630	0.665	0.535	0.549
1.5	0.0334	0.2266	0.900	0.5450	-0.06900	0.865	0.769	0.800	0.664	0.679
1.75	0.0397	0.2325	0.899	0.5507	-0.06854	0.941	0.864	0.891	0.766	0.778
2	0.0429	0.2385	0.901	0.5514	-0.06303	0.982	0.929	0.952	0.853	0.865
2.25	0.0501	0.2450	0.907	0.5545	-0.05952	0.994	0.971	0.981	0.925	0.925
2.5	0.0517	0.2518	0.910	0.5639	-0.06212	0.997	0.984	0.993	0.959	0.962
2.75	0.0527	0.2592	0.909	0.5650	-0.05584	1.000	0.990	0.995	0.982	0.983
3	0.0573	0.2661	0.916	0.5747	-0.05889	1.000	0.995	0.998	0.993	0.993
3.25	0.0612	0.2733	0.912	0.5810	-0.05821	1.000	0.998	1.000	0.997	0.998
3.5	0.0650	0.2799	0.911	0.5824	-0.05332	1.000	1.000	1.000	0.999	0.999

Dirección General de Bibliotecas UAQ

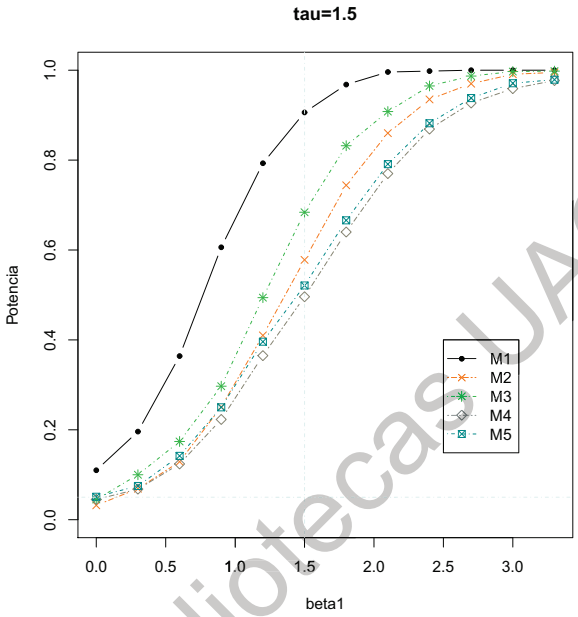
Cuadro 4.3: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 2{,}5$, con $m = 30$ clusters y $M = 20$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{V}ar(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE ; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

						Media			Mediana	
β_1	Sesgo	$\hat{V}ar(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0196	0.3240	0.882	0.6752	-0.1060	0.118	0.048	0.063	0.059	0.064
0.3	0.0166	0.3224	0.878	0.6791	-0.1113	0.147	0.078	0.087	0.072	0.084
0.6	0.0263	0.3230	0.880	0.6895	-0.1212	0.245	0.131	0.151	0.110	0.122
0.9	0.0360	0.3269	0.886	0.6863	-0.1146	0.407	0.238	0.275	0.178	0.189
1.2	0.0271	0.3328	0.881	0.6997	-0.1228	0.574	0.358	0.406	0.277	0.296
1.5	0.0373	0.3382	0.882	0.7101	-0.1285	0.733	0.519	0.560	0.379	0.403
1.8	0.0422	0.3451	0.884	0.7165	-0.1290	0.858	0.679	0.710	0.494	0.520
2.1	0.0415	0.3516	0.886	0.7257	-0.1327	0.936	0.793	0.839	0.606	0.625
2.4	0.0464	0.3618	0.887	0.7317	-0.1302	0.971	0.885	0.908	0.705	0.720
2.7	0.0515	0.3746	0.893	0.7374	-0.1253	0.990	0.940	0.958	0.778	0.787
3	0.0567	0.3837	0.891	0.7420	-0.1225	0.998	0.968	0.976	0.847	0.860
3.3	0.0640	0.3973	0.895	0.7413	-0.1110	0.999	0.985	0.988	0.894	0.898

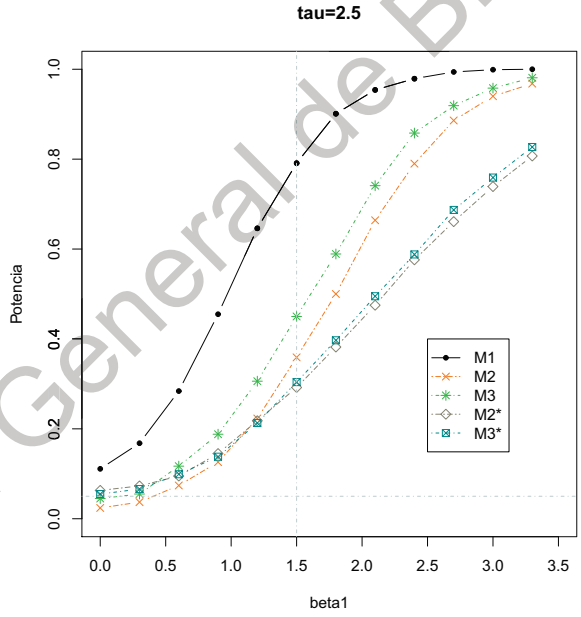
Dirección General de Bibliotecas UAQ



(a)



(b)



(c)

Figura 4.2: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = 2$, número de clusters $m = 30$ y número de individuos por clúster $M = 20$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

Dirección General de Bibliotecas UAQ

Cuadro 4.4: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 0$, con $m = 30$ clusters y $M = 20$ individuos por clúster con una media de conteos alta $\beta_0 = 2$. Se presenta el sesgo, la varianza asintótica $\hat{Var}(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

						Media			Mediana	
β_1	Sesgo	$\hat{Var}(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.00021	4.02E-05	0.933	0.00661	-0.00027	0.067	0.072	0.073	0.082	0.082
0.05	0.00024	3.91E-05	0.929	0.00654	-0.00029	1.000	1.000	1.000	1.000	1.000
0.1	0.00026	3.81E-05	0.935	0.00642	-0.00025	1.000	1.000	1.000	1.000	1.000
0.15	0.00021	3.74E-05	0.932	0.00639	-0.00028	1.000	1.000	1.000	1.000	1.000
0.2	0.00021	3.65E-05	0.929	0.00630	-0.00026	1.000	1.000	1.000	1.000	1.000
0.25	0.00026	3.56E-05	0.933	0.00620	-0.00023	1.000	1.000	1.000	1.000	1.000
0.3	0.00024	3.48E-05	0.932	0.00615	-0.00024	1.000	1.000	1.000	1.000	1.000
0.35	0.00024	3.41E-05	0.932	0.00610	-0.00026	1.000	1.000	1.000	1.000	1.000
0.4	0.00021	3.34E-05	0.934	0.00599	-0.00021	1.000	1.000	1.000	1.000	1.000
0.45	0.00024	3.26E-05	0.932	0.00595	-0.00024	1.000	1.000	1.000	1.000	1.000
0.5	0.00023	3.20E-05	0.935	0.00588	-0.00022	1.000	1.000	1.000	1.000	1.000
1	0.00023	2.72E-05	0.931	0.00546	-0.00025	1.000	1.000	1.000	1.000	1.000
1.5	0.00023	2.43E-05	0.933	0.00520	-0.00027	1.000	1.000	1.000	1.000	1.000
2	0.00024	2.25E-05	0.931	0.00498	-0.00024	1.000	1.000	1.000	1.000	1.000
2.5	0.00022	2.14E-05	0.932	0.00485	-0.00022	1.000	1.000	1.000	1.000	1.000
3	0.00022	2.08E-05	0.930	0.00476	-0.00020	1.000	1.000	1.000	1.000	1.000
3.5	0.00021	2.04E-05	0.932	0.00470	-0.00019	1.000	1.000	1.000	1.000	1.000

Dirección General de Bibliotecas UAQ

Cuadro 4.5: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 1{,}5$, con $m = 30$ clusters y $M = 20$ individuos por clúster con una media de conteos alta $\beta_0 = 2$. Se presenta el sesgo, la varianza asintótica $\hat{Var}(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

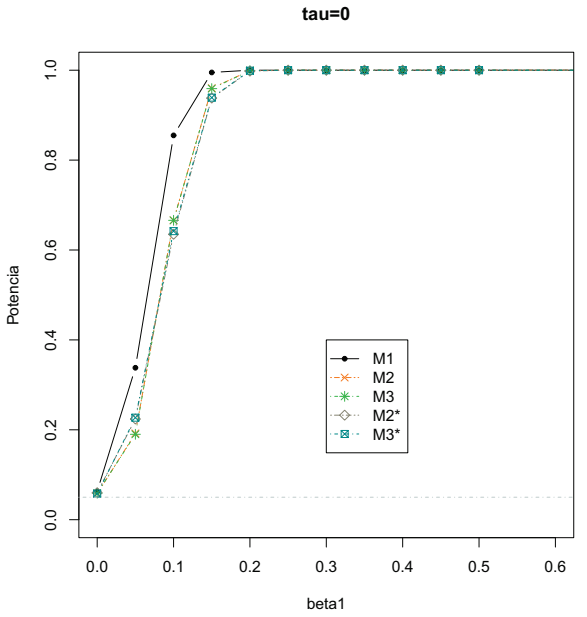
						Media			Mediana	
β_1	Sesgo	$\hat{Var}(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.02303	0.17661	0.890	0.4960	-0.0758	0.11	0.03	0.05	0.05	0.05
0.3	0.02043	0.17606	0.890	0.4960	-0.0764	0.20	0.07	0.10	0.07	0.08
0.6	0.02184	0.17607	0.896	0.4949	-0.0753	0.36	0.13	0.17	0.12	0.14
0.9	0.02142	0.17567	0.889	0.4974	-0.0783	0.61	0.25	0.30	0.22	0.25
1.2	0.02195	0.17582	0.888	0.4962	-0.0769	0.79	0.41	0.49	0.37	0.40
1.5	0.02249	0.17571	0.885	0.4960	-0.0768	0.91	0.58	0.68	0.50	0.52
1.8	0.02167	0.17553	0.888	0.4962	-0.0773	0.97	0.74	0.83	0.64	0.67
2.1	0.02235	0.17544	0.890	0.4959	-0.0770	1.00	0.86	0.91	0.77	0.79
2.4	0.02314	0.17534	0.889	0.4953	-0.0766	1.00	0.94	0.97	0.87	0.88
2.7	0.02334	0.17527	0.888	0.4954	-0.0767	1.00	0.97	0.99	0.93	0.94
3	0.02337	0.17532	0.891	0.4953	-0.0766	1.00	0.99	1.00	0.96	0.97
3.3	0.02327	0.17531	0.890	0.4952	-0.0765	1.00	1.00	1.00	0.98	0.98

Dirección General de Bibliotecas UAQ

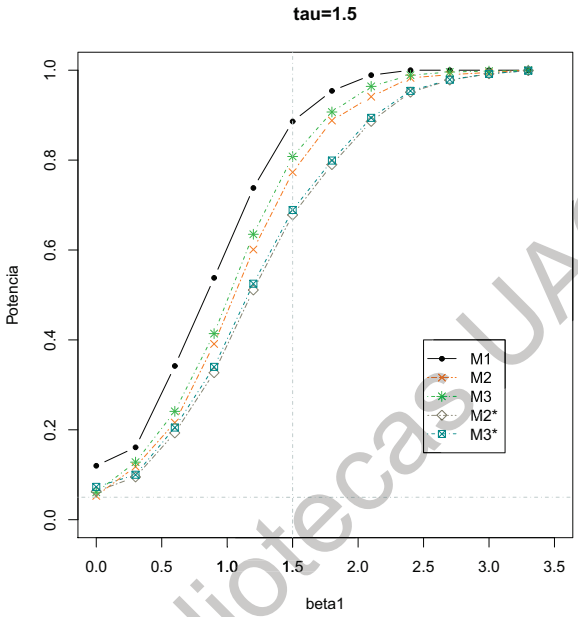
Cuadro 4.6: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 2{,}5$, con $m = 30$ clusters y $M = 20$ individuos por clúster con una media de conteos alta $\beta_0 = 2$. Se presenta el sesgo, la varianza asintótica $\hat{V}ar(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE ; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

						Media			Mediana	
β_1	Sesgo	$\hat{V}ar(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0289	0.2733	0.889	0.6298	-0.1070	0.111	0.024	0.045	0.063	0.055
0.3	0.0273	0.2726	0.887	0.6343	-0.1122	0.168	0.037	0.055	0.073	0.066
0.6	0.0288	0.2721	0.882	0.6322	-0.1105	0.284	0.074	0.117	0.095	0.099
0.9	0.0285	0.2718	0.880	0.6324	-0.1111	0.455	0.126	0.188	0.145	0.138
1.2	0.0289	0.2719	0.882	0.6324	-0.1109	0.646	0.223	0.306	0.218	0.213
1.5	0.0307	0.2714	0.885	0.6317	-0.1107	0.791	0.359	0.450	0.292	0.304
1.8	0.0299	0.2716	0.884	0.6310	-0.1099	0.901	0.500	0.589	0.382	0.397
2.1	0.0301	0.2716	0.880	0.6315	-0.1104	0.954	0.664	0.741	0.475	0.495
2.4	0.0291	0.2718	0.883	0.6315	-0.1101	0.979	0.790	0.858	0.576	0.588
2.7	0.0292	0.2716	0.883	0.6313	-0.1102	0.994	0.886	0.919	0.661	0.687
3	0.0302	0.2716	0.882	0.6311	-0.1099	0.999	0.940	0.958	0.739	0.759
3.3	0.0298	0.2716	0.885	0.6308	-0.1097	1.000	0.968	0.981	0.807	0.827

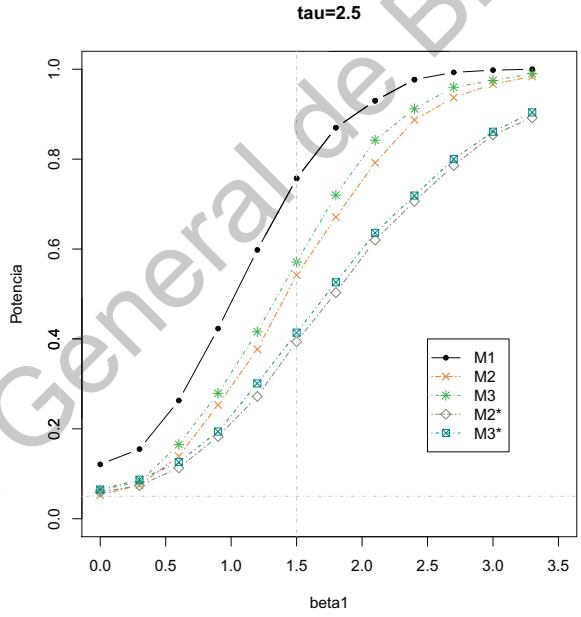
Dirección General de Bibliotecas UAQ



(a)



(b)



(c)

Figura 4.3: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = -2$, número de clusters $m = 30$ y número de individuos por clúster $M = 40$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

Dirección General de Bibliotecas UAQ

Cuadro 4.7: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 0$, con $m = 30$ clusters y $M = 40$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{Var}(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

							Media		Mediana	
β_1	Sesgo	$\hat{Var}(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	-0.00086	0.00110	0.937	0.0335	-0.0002	0.063	0.058	0.059	0.060	0.058
0.05	-0.00070	0.00108	0.945	0.0332	-0.0004	0.338	0.192	0.190	0.224	0.227
0.1	-0.00073	0.00105	0.933	0.0330	-0.0006	0.855	0.666	0.666	0.635	0.642
0.15	-0.00077	0.00102	0.938	0.0325	-0.0005	0.995	0.960	0.959	0.938	0.939
0.2	-0.00082	0.00100	0.933	0.0320	-0.0004	1.000	0.999	0.999	0.998	0.999
0.25	-0.00066	0.00098	0.939	0.0316	-0.0003	1.000	1.000	1.000	1.000	1.000
0.3	-0.00063	0.00096	0.940	0.0314	-0.0004	1.000	1.000	1.000	1.000	1.000
0.35	-0.00051	0.00094	0.935	0.0311	-0.0005	1.000	1.000	1.000	1.000	1.000
0.4	-0.00073	0.00092	0.938	0.0308	-0.0005	1.000	1.000	1.000	1.000	1.000
0.45	-0.00053	0.00090	0.938	0.0304	-0.0004	1.000	1.000	1.000	1.000	1.000
0.5	-0.00041	0.00088	0.938	0.0303	-0.0006	1.000	1.000	1.000	1.000	1.000
1	0.00016	0.00075	0.933	0.0281	-0.0007	1.000	1.000	1.000	1.000	1.000
1.5	0.00128	0.00067	0.928	0.0267	-0.0008	1.000	1.000	1.000	1.000	1.000
2	0.00088	0.00062	0.928	0.0257	-0.0008	1.000	1.000	1.000	1.000	1.000
2.5	0.00083	0.00059	0.920	0.0251	-0.0008	1.000	1.000	1.000	1.000	1.000
3	0.00073	0.00057	0.925	0.0247	-0.0008	1.000	1.000	1.000	1.000	1.000
3.5	0.00065	0.00056	0.917	0.0244	-0.0007	1.000	1.000	1.000	1.000	1.000

Dirección General de Bibliotecas UAQ

Cuadro 4.8: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 1{,}5$, con $m = 30$ clusters y $M = 40$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{Var}(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

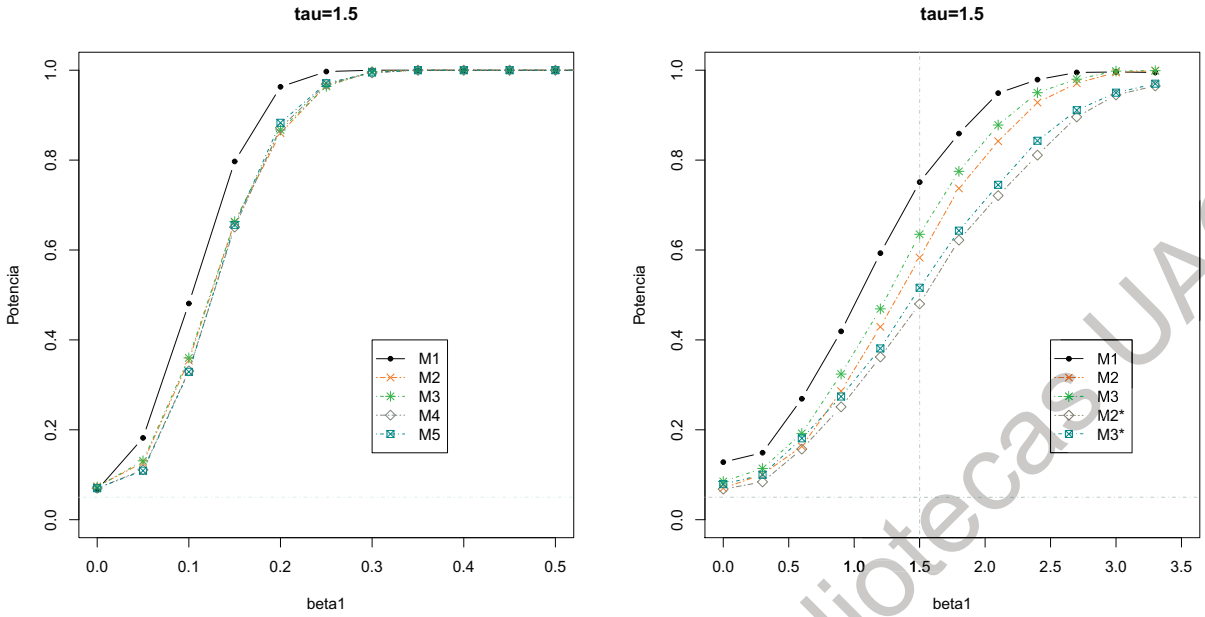
							Media		Mediana	
β_1	Sesgo	$\hat{Var}(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0050	0.2032	0.880	0.5422	-0.0914	0.120	0.053	0.061	0.063	0.073
0.05	0.0045	0.2043	0.889	0.5338	-0.0818	0.109	0.057	0.067	0.066	0.076
0.1	0.0013	0.2031	0.880	0.5419	-0.0912	0.121	0.069	0.072	0.068	0.075
0.15	0.0023	0.2033	0.882	0.5374	-0.0865	0.119	0.076	0.086	0.064	0.075
0.2	0.0005	0.2042	0.887	0.5405	-0.0886	0.139	0.093	0.099	0.071	0.080
0.25	0.0119	0.2035	0.886	0.5414	-0.0903	0.143	0.107	0.110	0.081	0.089
0.3	0.0105	0.2030	0.891	0.5348	-0.0843	0.161	0.118	0.128	0.095	0.100
0.35	0.0021	0.2035	0.885	0.5375	-0.0864	0.185	0.126	0.140	0.110	0.118
0.4	0.0058	0.2039	0.888	0.5403	-0.0887	0.212	0.145	0.157	0.124	0.129
0.45	0.0076	0.2020	0.885	0.5435	-0.0941	0.239	0.163	0.175	0.138	0.155
0.5	0.0088	0.2019	0.881	0.5329	-0.0835	0.257	0.177	0.193	0.152	0.164
1	0.0138	0.2048	0.875	0.5435	-0.0910	0.614	0.457	0.497	0.391	0.415
1.5	0.0159	0.2087	0.888	0.5460	-0.0891	0.886	0.773	0.808	0.678	0.689
2	0.0161	0.2129	0.888	0.5537	-0.0923	0.979	0.930	0.944	0.859	0.869
2.5	0.0258	0.2170	0.887	0.5541	-0.0883	1.000	0.985	0.993	0.961	0.966
3	0.0257	0.2219	0.891	0.5624	-0.0913	1.000	0.995	0.998	0.992	0.992
3.5	0.0265	0.2258	0.892	0.5672	-0.0920	1.000	1.000	1.000	0.999	0.999

Dirección General de Bibliotecas UAQ

Cuadro 4.9: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 2{,}5$, con $m = 30$ clusters y $M = 40$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{V}ar(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

						Media			Mediana	
β_1	Sesgo	$\hat{V}ar(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0101	0.3145	0.879	0.6890	-0.1282	0.121	0.053	0.064	0.059	0.065
0.3	0.0143	0.3116	0.882	0.6865	-0.1282	0.155	0.075	0.082	0.074	0.087
0.6	0.0258	0.3113	0.881	0.6777	-0.1197	0.263	0.139	0.165	0.113	0.126
0.9	0.0275	0.3123	0.884	0.6814	-0.1226	0.423	0.253	0.279	0.183	0.194
1.2	0.0326	0.3104	0.882	0.6909	-0.1338	0.598	0.377	0.416	0.272	0.301
1.5	0.0350	0.3140	0.877	0.6931	-0.1328	0.757	0.542	0.571	0.394	0.414
1.8	0.0353	0.3156	0.877	0.6886	-0.1268	0.870	0.671	0.720	0.503	0.526
2.1	0.0367	0.3177	0.877	0.6924	-0.1287	0.930	0.792	0.842	0.620	0.636
2.4	0.0399	0.3188	0.875	0.6917	-0.1272	0.977	0.887	0.912	0.706	0.719
2.7	0.0405	0.3208	0.879	0.6957	-0.1293	0.993	0.937	0.960	0.786	0.800
3	0.0413	0.3233	0.881	0.6986	-0.1299	0.998	0.967	0.975	0.854	0.861
3.3	0.0419	0.3255	0.877	0.7017	-0.1312	1.000	0.984	0.990	0.892	0.904

Dirección General de Bibliotecas UAQ



(a) (b)

Figura 4.4: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0, 1, 5\}$ respectivamente, ambas con media de conteos $\beta_0 = -2$, número de clusters $m = 20$ y número de individuos por clúster $M = 20$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

Dirección General de Bibliotecas UAQ

Cuadro 4.10: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 0$, con $m = 20$ clusters y $M = 20$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{Var}(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

						Media			Mediana	
β_1	Sesgo	$\hat{Var}(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0025	0.0032	0.934	0.05702	-0.00086	0.066	0.074	0.073	0.070	0.070
0.05	0.0027	0.0031	0.940	0.05602	-0.00052	0.182	0.127	0.131	0.111	0.109
0.1	0.0031	0.0030	0.940	0.05517	-0.00040	0.481	0.354	0.360	0.331	0.329
0.15	0.0031	0.0029	0.940	0.05447	-0.00028	0.797	0.662	0.663	0.651	0.656
0.2	0.0028	0.0029	0.939	0.05410	-0.00043	0.963	0.860	0.867	0.872	0.883
0.25	0.0024	0.0028	0.937	0.05335	-0.00030	0.997	0.966	0.964	0.969	0.971
0.3	0.0026	0.0027	0.933	0.05306	-0.00064	1.000	0.997	0.997	0.994	0.995
0.35	0.0026	0.0027	0.934	0.05267	-0.00077	1.000	1.000	1.000	1.000	1.000
0.4	0.0029	0.0026	0.935	0.05183	-0.00045	1.000	1.000	1.000	1.000	1.000
0.45	0.0028	0.0026	0.941	0.05123	-0.00035	1.000	1.000	1.000	1.000	1.000
0.5	0.0028	0.0025	0.939	0.05087	-0.00048	1.000	1.000	1.000	1.000	1.000
0.6	0.0027	0.0025	0.935	0.05019	-0.00067	1.000	1.000	1.000	1.000	1.000
1.1	0.0018	0.0021	0.924	0.04783	-0.00192	1.000	1.000	1.000	1.000	1.000
1.6	0.0009	0.0019	0.921	0.04543	-0.00176	1.000	1.000	1.000	1.000	1.000
2.1	0.0016	0.0018	0.913	0.04352	-0.00127	1.000	1.000	1.000	1.000	1.000
2.6	0.0016	0.0017	0.921	0.04268	-0.00134	1.000	1.000	1.000	1.000	1.000
3.1	0.0017	0.0017	0.918	0.04201	-0.00122	1.000	1.000	1.000	1.000	1.000

Dirección General de Bibliotecas UAQ

Cuadro 4.11: Resultado de la simulación de 1000 conjuntos de datos, bajo un modelo de regresión Poisson con efectos aleatorios $\tau = 1{,}5$, con $m = 20$ clusters y $M = 20$ individuos por clúster con una media de conteos baja $\beta_0 = -2$. Se presenta el sesgo, la varianza asintótica $\hat{V}ar(\hat{\beta}_1)$, la cobertura, el error estándar(SE), y diferencia en el SE; p_1 corresponde al p-valor del estimador del efecto del tratamiento bajo un modelo Poisson, p_2 y p_3 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la media de la variable de conteo. Por último p_4 y p_5 son la potencia del estimador del efecto del tratamiento bajo un modelo Log Binomial y un modelo Logístico respectivamente, con una variable respuesta binaria construida a través de un punto de corte definido por la mediana de la variable de conteo.

						Media			Mediana	
β_1	Sesgo	$\hat{V}ar(\hat{\beta}_1)$	Cobertura	SE	$\triangle SE$	p_1	p_2	p_3	p_4	p_5
0	0.0106	0.2876	0.872	0.6477	-0.1114	0.128	0.070	0.085	0.068	0.079
0.3	0.0049	0.2897	0.864	0.6554	-0.1172	0.149	0.101	0.114	0.084	0.100
0.6	0.0086	0.2941	0.876	0.6670	-0.1247	0.269	0.163	0.192	0.157	0.181
0.9	0.0163	0.3013	0.877	0.6705	-0.1216	0.419	0.286	0.324	0.251	0.274
1.2	0.0247	0.3118	0.878	0.6786	-0.1202	0.593	0.429	0.469	0.362	0.381
1.5	0.0251	0.3202	0.873	0.6901	-0.1242	0.751	0.583	0.635	0.480	0.516
1.8	0.0307	0.3397	0.878	0.7090	-0.1261	0.859	0.737	0.775	0.622	0.643
2.1	0.0384	0.3552	0.893	0.7165	-0.1205	0.949	0.842	0.878	0.721	0.745
2.4	0.0474	0.3805	0.893	0.7341	-0.1172	0.979	0.928	0.950	0.811	0.843
2.7	0.0486	0.4008	0.894	0.7412	-0.1081	0.995	0.971	0.980	0.896	0.911
3	0.0588	0.4322	0.900	0.7533	-0.0959	0.996	0.995	0.998	0.945	0.950
3.3	0.0718	0.9170	0.903	0.7809	0.1767	0.995	0.999	0.999	0.965	0.970

Dirección General de Bibliotecas UAQ

4.2. Discusión

En esta tesis investigamos la potencia de los modelos Log binomial y Logśtico respescto al Poisson con ajuste GEE por correlación debido a datos por conglomerados; para esto se construyó una variable binaria a partir de 1) la media y 2) la mediana de la tasa de los conteos de la variable Poisson. El estudio de simulación para comparar la potencia de los modelos reflejó que los modelos Log binomial (M2) y Logístico (M3) se desempeñan bien cuando no hay sobredispersión, i.e. cuando no hay varianza explicada por los clusters ($\tau = 0$), pero conforme la heterogeneidad de los clusters aumenta, la potencia de los modelos M2 y M3 baja sustancialmente; también se observó que el modelo M3 se desempeña ligeramente mejor que el M2 y que la variable binaria definida con base en la media se desempeña mejor que cuando se define con base en la mediana. Todas estas tendencias son consistentes para todos los casos, incluyendo cuando la media de conteos es alta o baja. Sin embargo, cuando la media de conteos es baja la diferencia es menos pronunciada.

Cuando la media de conteos es baja, intuitivamente hay más ceros, entonces los conteos son más pequeños, de esta forma, la aproximación de una variable poisson por una variable binaria es más cercana, así el modelo Log bonomial y Logístico no hacen tanta diferencia pues estos modelos usan variables respuestas definidas por ceros y unos.

Entre más heterogeneidad de los conglomerados más diferencias habrá entre los modelos, es decir, supongamos el caso de homogeneidad, en éste la media es igual para todos los clusters, entonces el desempeño de los tres modelos se reflejará muy similar.

Como en el modelo Poisson con ajuste GEE se tuvieron coberturas ligeramente bajas respecto a los niveles deseados (95 %), este modelo puede ser mejorado usando ajustes de GEE para muestras pequeñas. Este problema es parte de los desarrollos que se pueden explorar en un futuro. Otro problema que se puede abordar es ilustrar desde el punto de vista de diseño, es el cómo podría planearse un estudio usando información preliminar con base en una variable binaria (que es más fácil de comprender y elicitar) para un estudio que analizará los datos usando modelos Poisson.

Los resultados de este estudio pueden ser útiles para los analistas que diseñan y analizan datos de estudios por conglomerados; por ejemplo, para calcular la potencia para detectar el efecto de un tratamiento. El cálculo de la potencia es de extrema importancia. Hoy día la mayoría de las agencias que proveen fondos de investigación requieren una planeación detallada que incluya, entre otras cosas, especificaciones del diseño muestral como el tamaño de muestra y la potencia para detectar el efecto del tratamiento. El uso óptimo de recursos depende sustancialmente de las suposiciones realizadas para planear el estudio. Si el tamaño de muestra es muy pequeño, podría no detectarse ninguna diferencia. Por ejemplo, si se planea un estudio con cierto tamaño de muestra con una potencia del 80 % ($\alpha = 0{,}05$), pero resulta que hay una pérdida de potencia y en realidad este valor es del 60 %, es posible que no se observe una diferencia significativa. Entonces la conclusión del estudio sería que no hay evidencia de que el tratamiento funciona. Esto significaría que otros estudios tienen que ser realizados con este mismo tratamiento para obtener conclusiones definitivas al respecto. Por otro lado, si el tamaño de muestra es más grande de lo necesario, podría haber un desperdicio de recursos.

Se espera que el trabajo realizado sea de apoyo a diversos analistas: para asegurarse de que se realicen los cálculos correctos de potencia y tamaño de muestra, para proporcionar evidencia a los analistas de que el análisis necesita tomar en cuenta la agrupación y que, por lo general, no se recomienda su simplificación mediante una variable binaria.

Dirección General de Bibliotecas UAQ

Bibliografía

[1] E. Lazcano-Ponce, E. Salazar-Martínez, P. Gutiérrez-Castrellón, A. Angeles-Llerenas, A. Hernández-Garduño, and J. L. Viramontes, "Ensayos clínicos aleatorizados: variantes, métodos de aleatorización, análisis, consideraciones éticas y regulación," *Salud Pública de México*, vol. 46, pp. 559–584, Dec. 2004.

[2] G. Perman, "Ensayos clínicos por conglomerados (clusters)," p. 4, 2017.

[3] M. Alcaide Delgado, "Modelo de regresión binominal negativa," 2015.

[4] "Offset Definition | Statistics Dictionary | MBA Skool-Study.Learn.Share.."

[5] A. Salinas-Rodríguez, B. Manrique-Espinoza, and S. G. Sosa-Rubí, "Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud," Oct. 2009.

[6] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*. No. 25, Great Britain: Oxford Statistical Science Series, second edition ed., 2002.

[7] M. L. Nores, "CONSTRUCCIÓN DE MODELOS GEE PARA VARIABLES CON DISTRIBUCIÓN SIMÉTRICA," p. 21.

[8] T. Williamson, M. Eliasziw, and G. H. Fick, "Log-binomial models: exploring failed convergence," *Emerging Themes in Epidemiology*, vol. 10, p. 14, Dec. 2013.

[9] L.-A. McNutt, C. Wu, X. Xue, and J. P. Hafner, "Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes," *American Journal of Epidemiology*, vol. 157, pp. 940–943, May 2003.

[10] G. Zou, "A Modified Poisson Regression Approach to Prospective Studies with Binary Data," *American Journal of Epidemiology*, vol. 159, pp. 702–706, Apr. 2004.

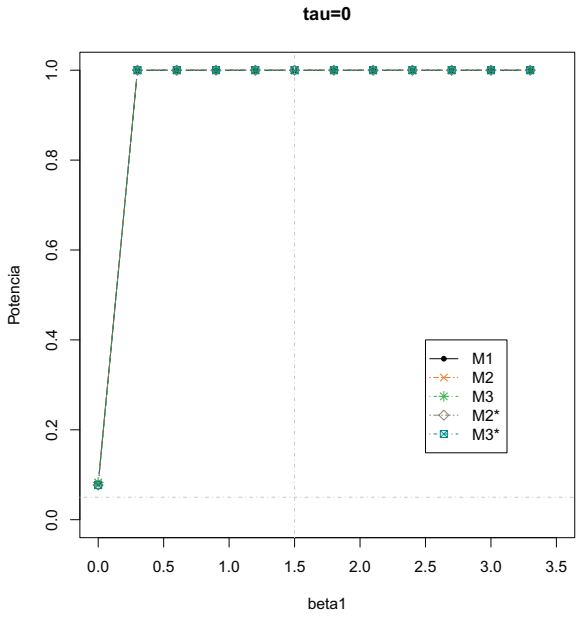
Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ

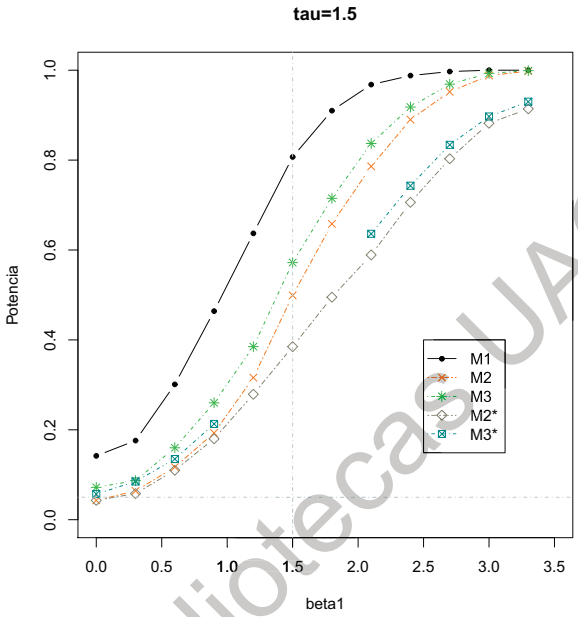
APÉNDICE A

Figuras

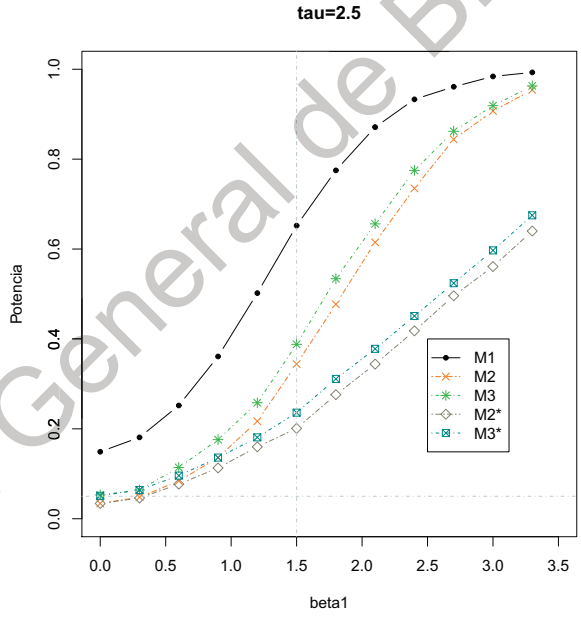
Dirección General de Bibliotecas UAQ



(a)



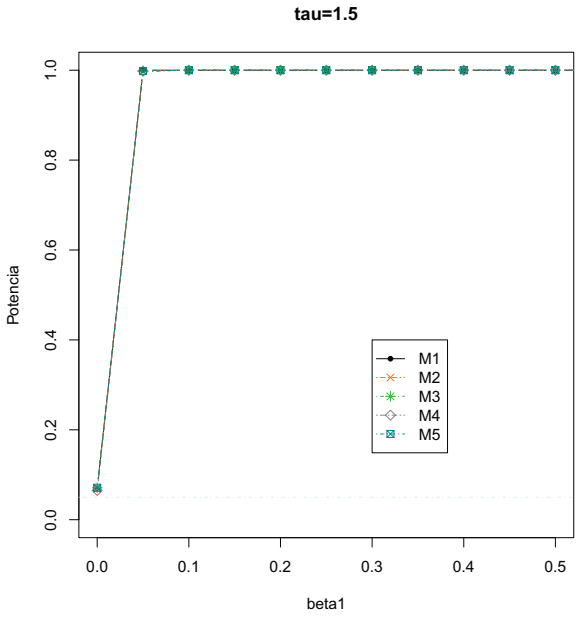
(b)



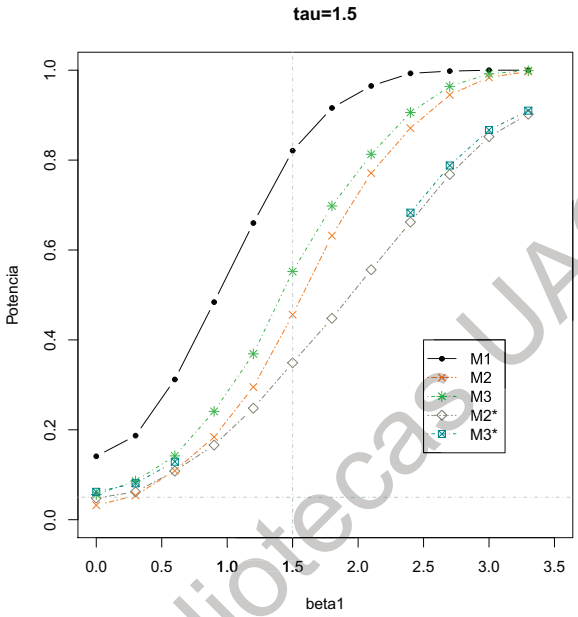
(c)

Figura A.1: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = 0$, número de clusters $m = 20$ y número de individuos por clúster $M = 20$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

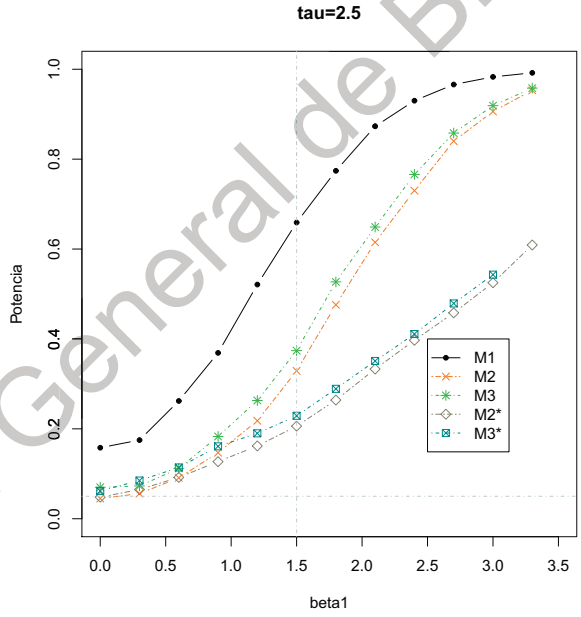
Dirección General de Bibliotecas UAQ



(a)



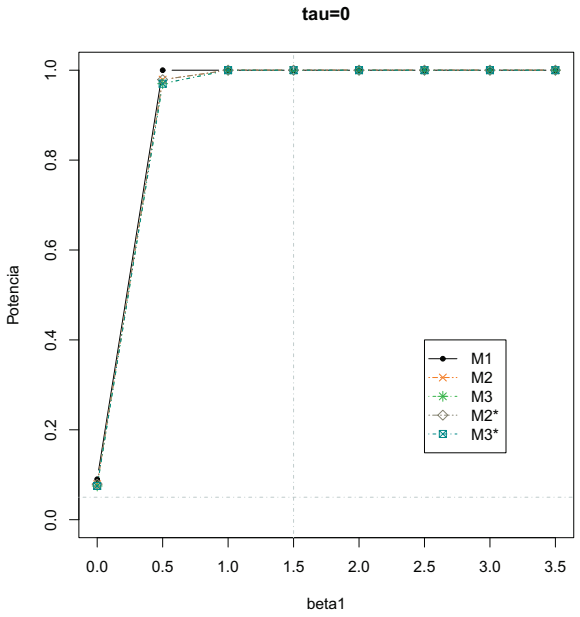
(b)



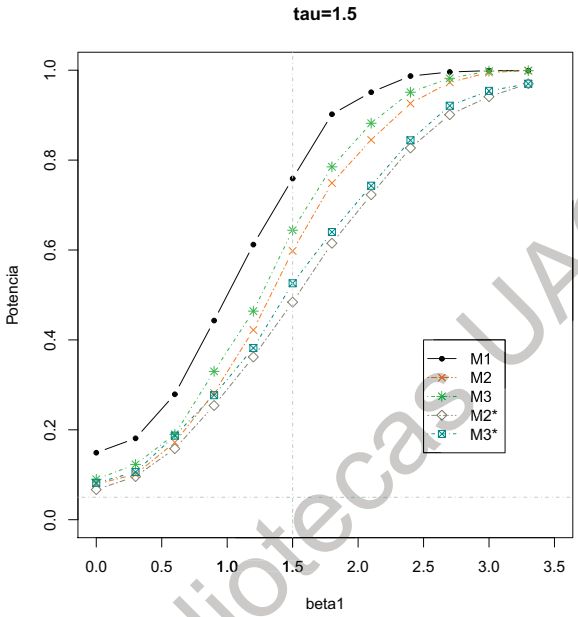
(c)

Figura A.2: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = 2$, número de clusters $m = 20$ y número de individuos por clúster $M = 20$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

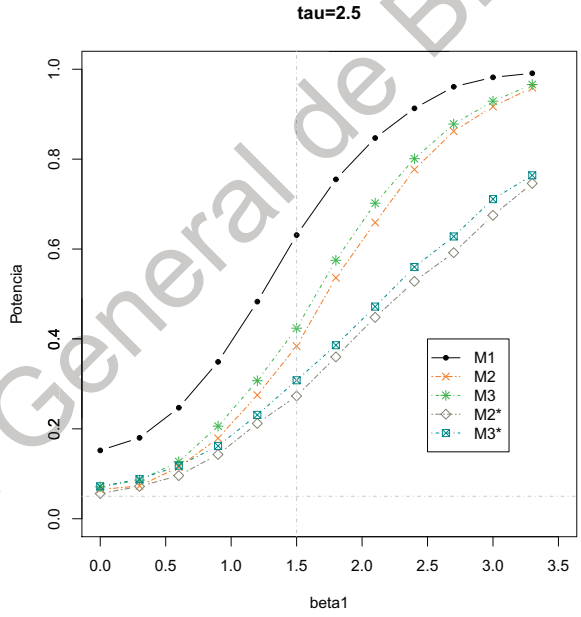
Dirección General de Bibliotecas UAQ



(a)



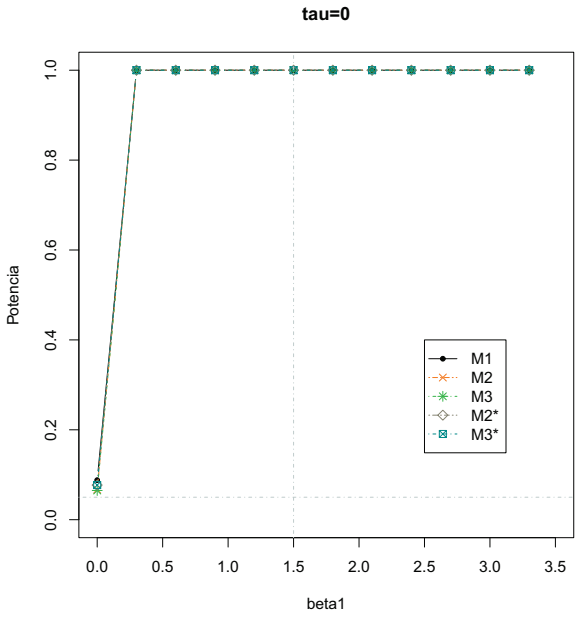
(b)



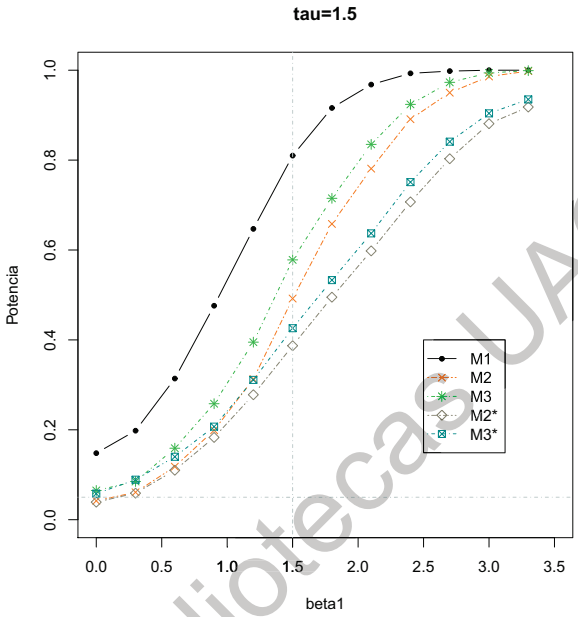
(c)

Figura A.3: Las gráficas (a), (b) y (c) presentan la potencia del efecto del tratamiento para sobre-dispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = -2$, número de clusters $m = 20$ y número de individuos por clúster $M = 40$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

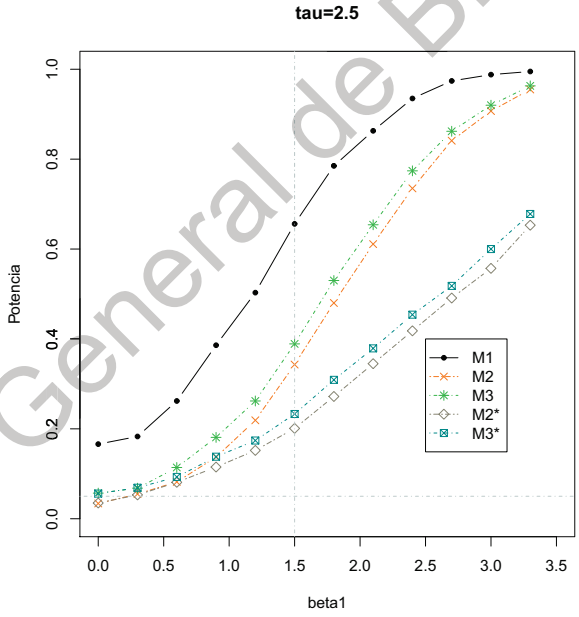
Dirección General de Bibliotecas UAQ



(a)



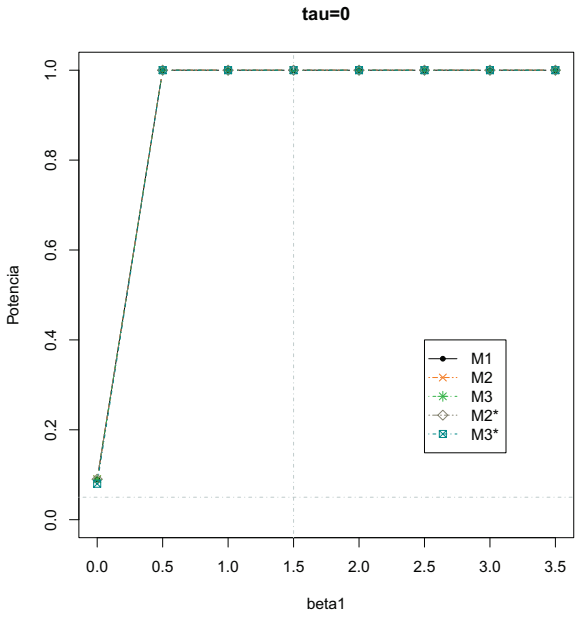
(b)



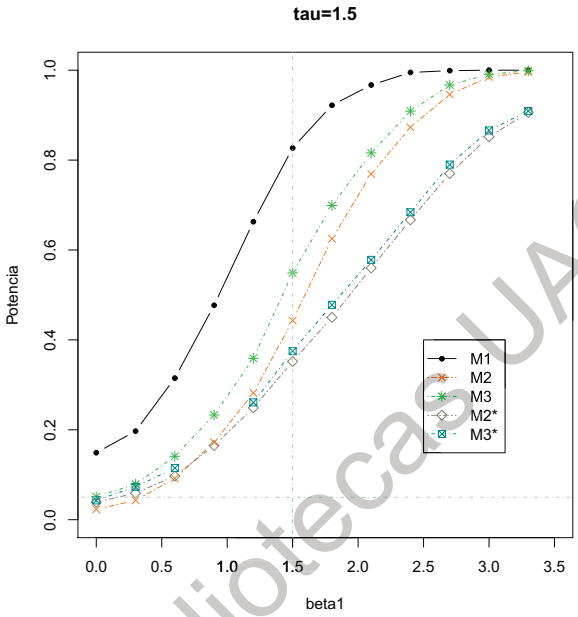
(c)

Figura A.4: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0, 1{,}5, 2{,}5\}$ respectivamente, todas con media de conteos $\beta_0 = 0$, número de clusters $m = 20$ y número de individuos por clúster $M = 40$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

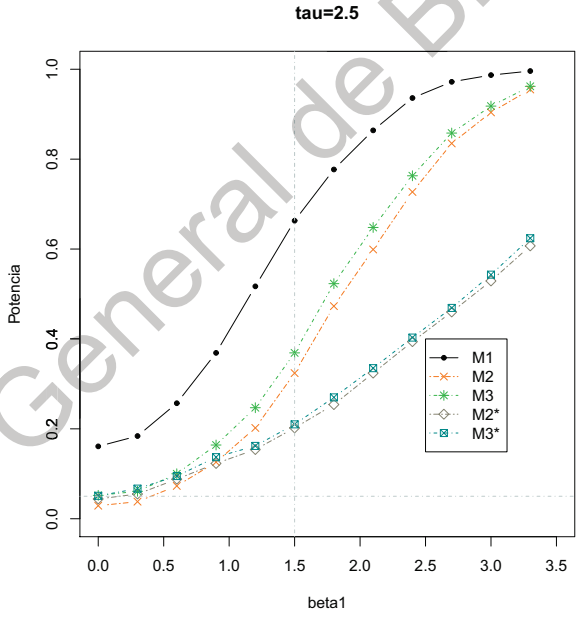
Dirección General de Bibliotecas UAQ



(a)



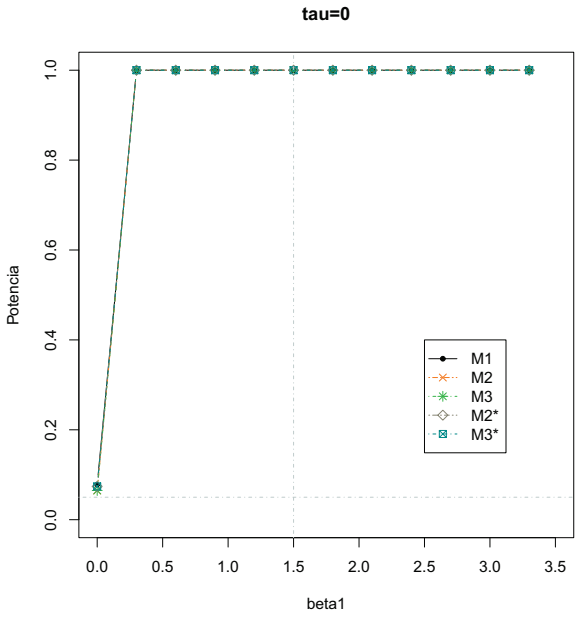
(b)



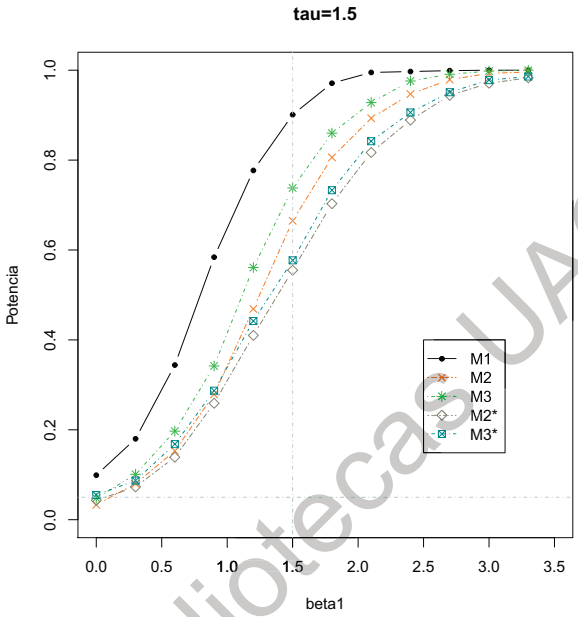
(c)

Figura A.5: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0, 1{,}5, 2{,}5\}$ respectivamente, todas con media de conteos $\beta_0 = 2$, número de clusters $m = 20$ y número de individuos por clúster $M = 40$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

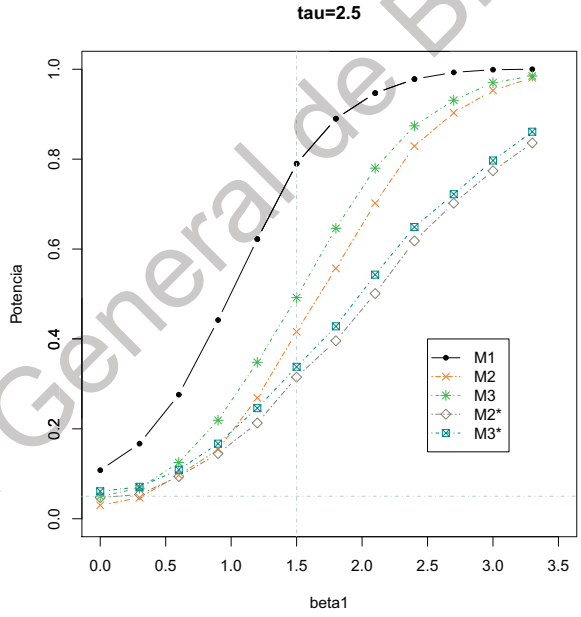
Dirección General de Bibliotecas UAQ



(a)



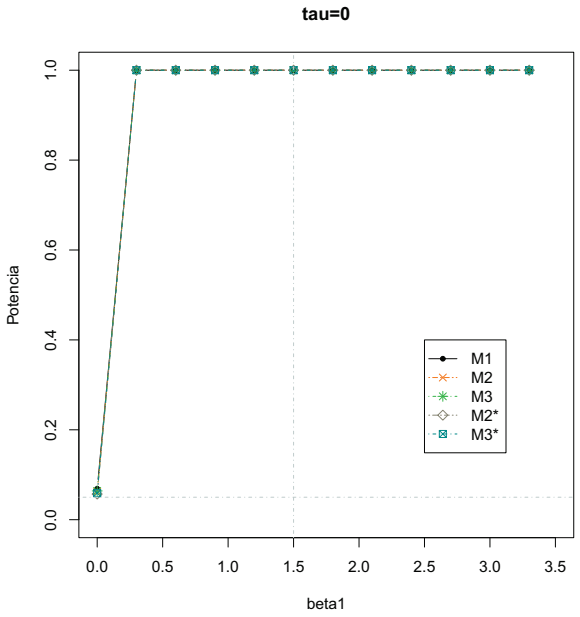
(b)



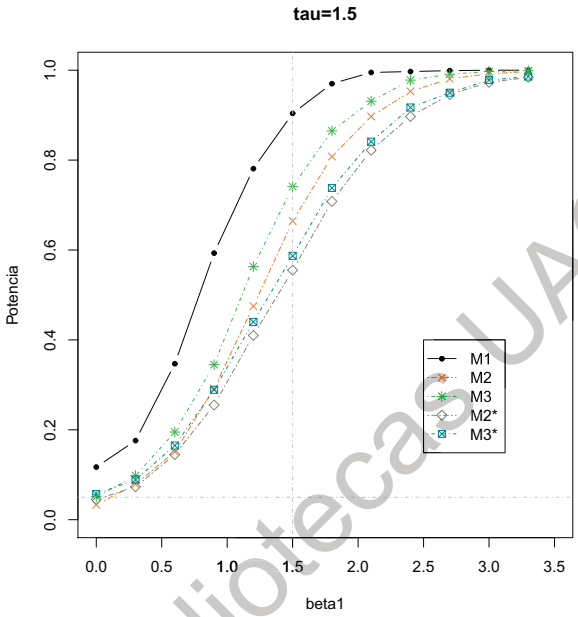
(c)

Figura A.6: Las gráficas presentan la potencia del efecto del tratamiento para valores de sobredispersión $\tau = \{0, 1{,}5, 2{,}5\}$ respectivamente, todas con media de conteos $\beta_0 = 0$, número de clusters $m = 30$ y número de individuos por clúster $M = 20$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

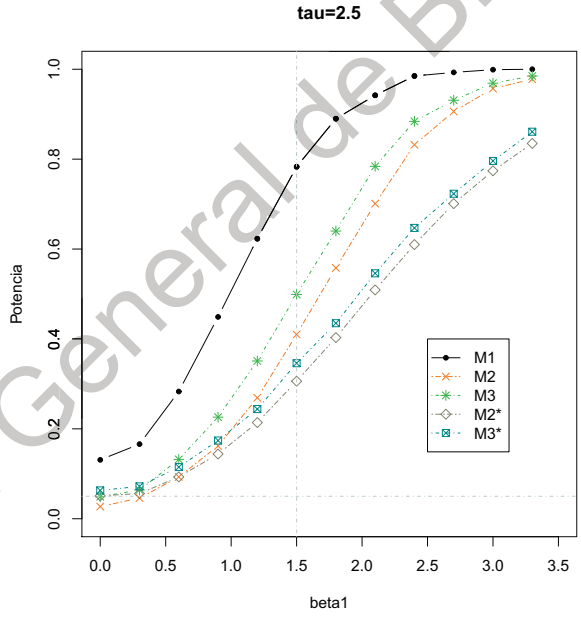
Dirección General de Bibliotecas UAQ



(a)



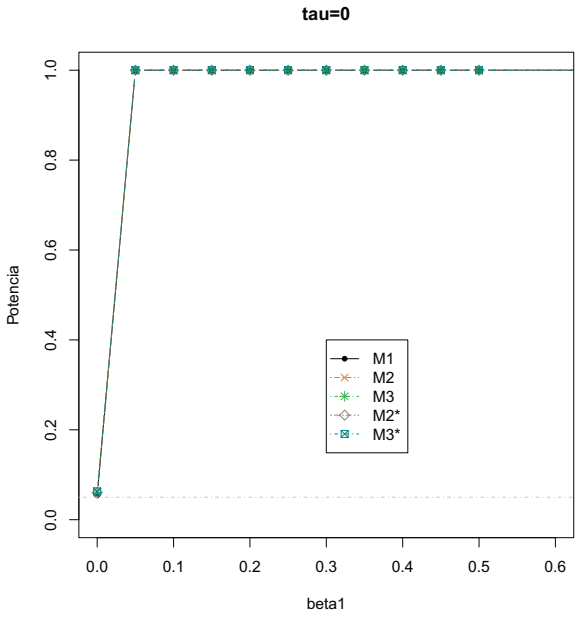
(b)



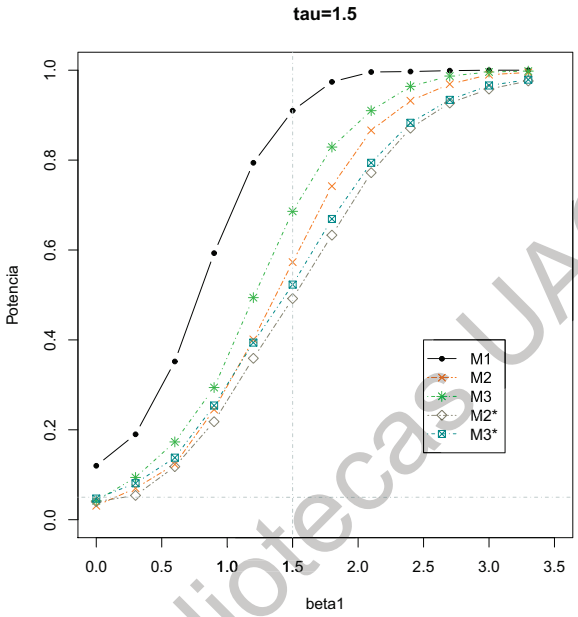
(c)

Figura A.7: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0,1,5,2,5\}$ respectivamente, todas con media de conteos $\beta_0 = 0$, número de clusters $m = 30$ y número de individuos por clúster $M = 40$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

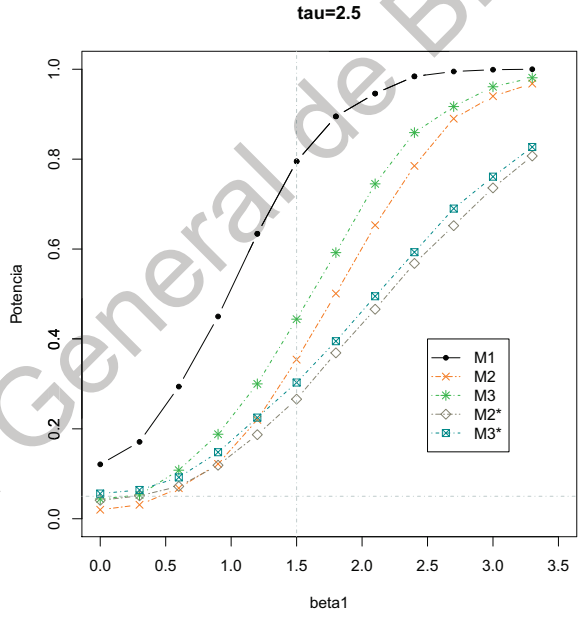
Dirección General de Bibliotecas UAQ



(a)



(b)



(c)

Figura A.8: Las gráficas presentan la potencia del efecto del tratamiento para sobredispersión $\tau = \{0, 1{,}5, 2{,}5\}$ respectivamente, todas con media de conteos $\beta_0 = 2$, número de clusters $m = 30$ y número de individuos por clúster $M = 40$, con base en un Modelo Poisson (M1), en un Log Binomial (M2 y M4) y en un Logístico (M3 y M5). La variable binaria para M2 y M3 está definida con base en la media. La variable binaria para M4 y M5 está definida con base en la mediana.

Dirección General de Bibliotecas UAQ

Dirección General de Bibliotecas UAQ