Evaluación de la presencia de confusión en algunos miembros de la familia exponencial

Lina María Acosta Avena

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2014

Evaluación de la presencia de confusión en algunos miembros de la familia exponencial

Lina Maria Acosta Avena

Tesis de grado presentado como requisito parcial para optar al título de:
Magíster en Ciencias - Estadística

Director:
Juan Carlos Salazar Uribe, Ph.D. en Estadística

Líneas de Investigación:
Bioestadística

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadistica
Medellin, Colombia
2014

A todos los miembros de las familias: Zuluaga Avena, Acosta Correa, Ortiz Patrón.

Agradecimientos

En primera instancia, agradezco enormemente al profesor Juan Carlos Salazar Uribe, por ser mi asesor y consejero, por su acompañamiento, ideas y orientaciones, las cuales fueron vitales para el desarrollo de este trabajo.

Quiero agradecer de manera especial a Roger Jesús Tovar Falon, por ser la persona que me motivó hacer esta maestría y por darme el coraje para terminarla, por todo el apoyo académico - emocional, a pesar de la distancia.

También agradezco a todos los profesores de la Escuela de Estadística quienes contribuyeron con sus valiosas enseñanzas durante mi proceso de formación durante estos dos años, de manera especial a Sergio Yañez, Juan Carlos Correa, Nelfi González, Victor López, Juan Carlos Salazar y Elkin Castaño, quienes fueron mis tutores en los diferentes cursos vistos en la Maestría.

También quiero agradecer a dos personas muy valiosas en la escuela, Diana Arboleda y Olga Bustos, quienes tuvieron mucha paciencia durante mi proceso de inscripción y matrícula, además porque siempre estuvieron atentas y dispuestas a colaborar ante cualquier situación de dificultad que se presentó en el transcurso de la maestría.

Finalmente, a quienes fueron mis compañeros de estudio, en especial a mis grandes amigos Javier Lozano y Elizabeth Estrada, por ese apoyo incondicional, principalmente en los momentos difíciles. Muchas Gracias!

Resumen

En modelamiento, el fenómeno de la confusión puede ser problemático y por lo tanto se debe prestar atención a su detección. En este trabajo es de particular interés estudiar una de las técnicas recomendadas en la literatura para detectar confusión cuando se trabaja con modelos de regresión. El objetivo es determinar si el criterio propuesto es válido para algunos modelos de la familia exponencial, particularmente el modelo lineal clásico, el logístico y el Poisson. Un criterio alternativo es propuesto para el caso del modelo logístico.

Palabras clave: confusión, familia exponencial, estadística, regresión.

Abstract

In modeling, the phenomenon of confounding can be problematic and therefore one should pay attention to its detection. In this paper it is of particular interest to study one of the recommended techniques in the literature to detect confounding when working with regression models. The goal is to assess whether the criteria proposed is valid for some models of the exponential family, particularly the classical linear, the logistic and Poisson models. An alternative criteria is proposed for the logistic model case.

Keywords: confounding, exponential family, statistics, regression.

Contenido

Agradecimientos vii

Resumen ix

1. Introducción 1

2. Una introducción a las variables de confusión 5
2.1. Variables de confusión . . . 5
2.2. Tipos de confusores . . . 6
2.3. Identificación de variables de confusión . . . 8
2.4. Algunos métodos para evaluar confusión . . . 11
2.4.1. Métodos analíticos tradicionales . . . 11
2.4.2. Métodos comunes para controlar variables de confusión . . . 13

3. Confusión en modelos de regresión 17
3.1. Modelo lineal generalizado (MLG) . . . 17
3.1.1. Componentes del MLG . . . 17
3.1.2. Deviance . . . 18
3.2. Familia exponencial . . . 18
3.2.1. Modelo de regresión lineal clásico . . . 19
3.2.2. Regresión logística . . . 19
3.2.3. Regresión Poisson . . . 22
3.3. Modelo de riesgo proporcional . . . 22
3.4. Confusión en modelos de regresión . . . 23
3.4.1. Criterio del cambio porcentual . . . 23
3.4.2. Criterio propuesto . . . 25

4. Estudio de simulación 27
4.1. Diseño del estudio de simulación . . . 27
4.2. Resultados . . . 30
4.2.1. Modelo logístico y Modelo Poisson . . . 31

4.2.2. Modelo lineal clásico 50

5. Conclusiones y recomendaciones **53**

A. Programa en R para el estudio de simulación **55**
A.1. Modelo logístico 55
A.2. Modelo Poisson 56
A.3. Modelo lineal clásico 57

Lista de Tablas

4-1. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$ . . . 33
4-2. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$ . . . 33
4-3. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$ . . . 34
4-4. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$ . . . 34
4-5. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$ . . . 36
4-6. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$ . . . 37
4-7. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$ . . . 37
4-8. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$ . . . 38
4-9. Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$ . . . 40
4-10.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$ . . . 40
4-11.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$ . . . 41
4-12.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$ . . . 41
4-13.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$ . . . 44
4-14.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$ . . . 44
4-15.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$ . . . 45
4-16.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$ . . . 45
4-17.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$ . . . 48
4-18.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$ . . . 49
4-19.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$ . . . 49
4-20.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$ . . . 50
4-21.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$ . . . 51
4-22.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$ . . . 52
4-23.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$ . . . 52
4-24.Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$ . . . 52

Lista de Figuras

2-1. Clasificación de los confusores 7
2-2. Situaciones en las que C es un factor de confusión 9
2-3. Situaciones en las que C no es un factor de confusión 9
2-4. Diagrama de ruta para la asociación consumo de ginseng - cáncer gástrico . 10
2-5. Diagráma de ruta para la asociación fibrinógeno - CHD 11

4-1. Escenarios para el modelo logístico 28
4-2. Escenarios para el modelo Poisson (a) Escenario 1, (b) Escenario 2, (c) Escenario 3, (d) Escenario 4, (e) Escenario 5 29
4-3. Comportamiento de los criterios (a) $\Delta\hat{\beta}\,\%$ y (b) CP, en el modelo logístico cuando $\rho = 0$. 32
4-4. Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en (a) modelo Poisson cuando $\delta = 0.05$ y $p = -0.8, -0.6, -0.4, -0.2, 0$, (b) modelo logístico cuando $\rho = 0$. 35
4-5. Comportamiento del criterio $\Delta\hat{\beta}\,\%$ cuando $\delta = 0.20$ y $p = 0.8, 0.6, 0.4, 0.2, 0$ en (a) Modelo logístico, (b) Modelo Poisson. 39
4-6. Comportamiento de los criterios (a) $\Delta\hat{\beta}\,\%$ y (b) CP, en el modelo logistico cuando $\rho = 0$. 42
4-7. Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo Poisson cuando (a) $\rho = -0.8$, (b) $\rho = -0.6$, (c) $\rho = -0.4$, (d) $\rho = -0.2$, (e) $\rho = 0$ 43
4-8. Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo logistico cuando (a)$\rho = 0.8$,(b)$\rho = 0.6$,(c)$\rho = 0.4$,(d)$\rho = 0.2$,(e)$\rho = 0$. 46
4-9. Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo Poisson cuando (a)$\rho = 0.8$,(b)$\rho = 0.6$,(c)$\rho = 0.4$,(d)$\rho = 0.2$,(e)$\rho = 0$. 47
4-10.Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo lineal clásico cuando $\rho = 0$. 51

1. Introducción

En la práctica estadística usualmente es de interés para el investigador mirar la relación entre las variables que se estudian. Por ejemplo, en estudios epidemiológicos, los investigadores muchas veces están interesados en explorar la relación entre una enfermedad de interés y el grado de exposición a un cierto factor de riesgo. En ocasiones un tercer factor puede tener una influencia importante en la relación aparente entre esas dos variables (Woodward 1999); eventualmente, estos tipos de factores no son medidos y controlados como parte del estudio. Si este es un factor de riesgo considerado potencial para la enfermedad y que además se mezcla con la exposición al otro factor de riesgo para distorsionar la relación observada entre la enfermedad y la exposición, entonces el segundo se considera que es un factor de confusión (Schelesselman 1982).

La consideración de los factores de confusión es fundamental para el diseño y análisis de estudios de efectos causales (Greenland & Pearl 1999). La importancia de detectar las variables de confusión descansa en el hecho de que están asociadas tanto con la causa probable y el resultado, lo cuál puede conducir a la detección de una asociación espúrea, así que la identificación de los factores de confusión son determinantes para la inferencia causal ya que distorsionan la verdadera relación entre una exposición con la enfermedad. De hecho, estas variables son la mayor amenaza para la validación (interna) de las inferencias hechas sobre causa y efecto (Pourhoseingholi et al. 2012).

La interpretación estadística del término "confusión" ha sido un tema que ha generado controversia en la literatura (Wickramaratne & Holford 1987). Infortunadamente, la palabra inglesa "confounding" ha sido usada para referirse al menos a tres conceptos distintos. El uso más antiguo se refiere a un tipo de sesgo en la estimación de efectos causales, el cuál a veces es descrito como una mezcla entre los factores extraños y el factor de interés, este uso predomina en investigaciones no experimentales, especialmente en epidemiología y sociología (Greenland & Pearl 1999). Más recientemente, "confounding" es usado como un sinónimo de "noncollapsability" específicamente en aquellas situaciones en las que el parámetro de interés es un efecto causal (Greenland & Pearl 1999). Finalmente, el término se emplea en la literatura del diseño experimental como sinónimo de "aliasing", partícular-

mente en el análisis de varianza, para referirse a la inseparabilidad de los efectos principales y las interacciones de un diseño en particular (Greenland & Pearl 1999).

Métodos analíticos tradicionales como estimaciones y pruebas de hipótesis, son frecuentemente utilizados para evaluar confusión (Woodward 1999); entre los más comunes se encuentran el riesgo relativo, las razones de odds y las estandarizaciones, los cuales se encargan de variables de confusión con muchos niveles.

Pourhoseingholi et al. (2012), Newman (2001) y Kamangar (2012) indican que existen varios caminos para modificar un estudio excluyendo o controlando variables de confusión, empleando mecanismos como aleatorización, restricción y matching. Con la aleatorización se trata de romper el vínculo entre la exposición y la variable de confusión asignando aleatoriamente los sujetos del estudio en las categorías de la exposición, con esto se reduce la posibilidad de confusión porque los grupos generados son bastante comparables con respecto a las variables de confusión conocidas y desconocidas. La restricción elimina la variación en la confusión. Matching es comúnmente usado en estudios de casos y controles; este método obliga a que estos grupos sean muy similares con respecto a los factores de riesgo importantes, y por lo tanto hace que las comparaciones de casos y controles sean menos sujetas a factores de confusión. Estos métodos son aplicables en el diseño de estudio y antes del proceso de recolección de los datos. Una discusión sobre el uso de estos métodos se puede ver Mickey & Greenland (1989) y Kamangar (2012).

Otros procedimientos empleados para controlar los factores de confusión en los análisis son, la estratificación y los modelos multivariados. Con la estratificación se trata de buscar grupos dentro de los cuales el factor de confusión no varíe y así evaluar la asociación entre la exposición y la enfermedad dentro de cada estrato del confusor, y luego se emplea el estimador de Mantel - Haenszel para ajustar los resultados para cada estrato (Pourhoseingholi et al. 2012), con este método se determina que hay confusión cuando se encuentran diferencias entre los resultados crudos y los ajustados de acuerdo al estrato. Sin embargo, este método se vuelve poco manejable cuando se tiene un gran número de estratos o de covariables, o inclusive un número grande de posibles confusores. En estas situaciones es de gran utilidad implementar entonces modelos multivariados.

Los efectos que tienen estas variables en modelos de regresión han sido estudiado por algunos investigadores en diferentes situaciones. A continuación se presentan algunos trabajos relacionados con el tema.

Chen & Winkler (1999), llevan a cabo un estudio de simulación en el que se observa un modelo de regressión Poisson bajo la presencia de un factor de confusión. Los resultados obtenidos en dicho estudio, indican que la presencia del factor de confusión afecta la estimación de los coeficientes y sus niveles de confianza, el estudio además muestra cómo este efecto cambia con los rangos de las covariables y los tamaños de las muestras.

Un aporte importante en el estudio de variables de confusión en modelos lineales generalizados, está dado en Austin & Brunner (2004). Estos autores muestran en este estudio que la probabilidad de cometer un error tipo I en un modelo de regresión logística aumenta cuando una variable de confusión de tipo continuo es categorizada. Específicamente, mediante un proceso de simulación se encuentra que la probabilidad de cometer un error tipo I crece a medida que la correlación entre la exposición y la variable de confusión crece, y también a medida que se aumenta el tamaño de la muestra y que disminuye a medida que aumenta el número de categorías de la variable de confusión.

Otra investigación sobre este fenómeno en la regresión logística se encuentra en Becher (1992); aquí el autor estudia el comportamiento de los residuales de confusión cuando una variable de confusión es categorizada. El efecto residual de confusión es definido como

$$\phi = \exp\{\tilde{\beta}_1 - \beta_1\}$$

donde $\exp\{\tilde{\beta}_1\}$ es el odds para el factor de riesgo después de la estratificación del confusor y $\exp\{\beta_1\}$ es el odds para el factor de riesgo sin estratificar al confusor. Becher (1992) recomienda que en caso de que la estratificación del confusor sea necesaria se deben usar cuatro o cinco niveles, esto también es sugerido por Cochran (1968) cuando se trabaja con un modelo lineal; ellos además sugieren que dicotomizar un confusor continuo no es recomendado. En ambos estudios las covariables se toman de una distribución normal bivariada.

En la literatura se encuentran otras investigaciones sobre el efecto que tienen las variables de confusión en modelos de regresión, tales como Frank (2000), Wilson & Gordon (1986), Chao & Young (1997), entre otros.

Es claro que en estos trabajos el interés es determinar los efectos que tiene la presencia de confusión en algunos modelos conocidos, sin embargo, cuando los investigadores tiene el interés de modelar el comportamiento de una variable de interés a través de un conjunto de variables explicativas por medio de un modelo estadístico que describa tal relación, antes de ejecutar el modelo cuastionan estas relaciones al punto de preguntarse ¿la relación es real, o es falsa? De manera que para el caso de los modelos de regresión es de gran importancia

estudiar los criterios para detectar estas falsas asociaciones.

Entre los criterios para detectar confusión en modelos de regresión que se encuentran en la literatura está el propuesto por Hosmer & Lemeshow (1999), quienes recomiendan usar el porcentaje de cambio en la estimación como una posible medida de confusión, generando un estimador puntual de cambio que en general lo definen como el cociente entre la diferencia de los estimadores ajustados del modelo que no incluye la potencial confusión y uno que si la incluye. En la práctica, encuentran que un cambio mayor que $15\,\% - 20\,\%$ indica que la confusión está presente, y por lo tanto que el ajuste es necesario. En general, toman un nivel de referencia del $20\,\%$ como un importante cambio en el coeficiente, así que ellos recomiendan que si el valor del criterio es inferior a éste la confusión no está presente.

La deducción de este criterio está basada en un modelo de riesgos proporcionales, sin embargo ¿este criterio es eficiente para cualquier modelo de regresión?. La idea central de este trabajo es dar respuesta a esta pregunta, para esto se estudia vía simulación, el comportamiento del criterio del cambio porcentual en algunos miembros de la familia exponencial, particularmente el modelo lineal clásico, el logístico y el Poisson, los cuales son más comúnmente usados en epidemiología (Kamangar 2012), para el estudio de datos de enfermedades (Mcnamee 2005). El objetivo principal de este trabajo es determinar si el criterio es válido para estos modelos. Para esto se asumieron dos variables regresoras de una distribución normal multivariada y en las simulaciones se establecieron diferentes escenarios en los parámetros y tamaños muestrales. Para determinar la eficiencia del criterio en modelos considerados, se usaron diferentes niveles de referencia.

Este trabajo está organizado en 5 Capítulos y un Apéndice. En el Capítulo 2 se presenta una breve introducción a las variables de confusión y se muestran algunos ejemplos para lograr un mejor entendimiento del trabajo; así como una revisión de los métodos que según la literatura son los más comúnmente empleados por los investigadores cuando se evalúa la presencia del fenómeno de la confusión. El criterio propuesto por Hosmer & Lemeshow (1999) para evaluar la presencia de confusión en modelos de regresión, es abordado en el Capítulo 3, y se presenta un nuevo criterio para evaluar este fenómeno en el modelo logístico. También en este capítulo se hace una revisión de los modelos: lineal clásico, logístico y Poisson, puesto que el desempeño del criterio del cambio porcentual es evaluado en ellos. El estudio de simulación que se implementó para evaluar tal desempeño es presentado en el Capítulo 4. Las conclusiones y perspectivas futuras de investigación son presentadas en el Capítulo 5.

2. Una introducción a las variables de confusión

Gran parte de los estudios de investigación implican en algunos casos la recolección de grandes cantidades de datos y como tal, una amplia gama de variables. El proceso de selección de las variables es tal vez una de las tareas más difíciles en el análisis estadístico, debido a que en ocasiones éstas pueden estar relacionadas de una forma u otra y con ello llegar a distorsionar la medida de asociación entre otras variables dando un resultado espúreo. Las variables que ocasionan este tipo de efectos son conocidas en investigaciones clínicas y epidemiológicas como variables de confusión. A continuación se hace una introducción a estas variables.

2.1. Variables de confusión

Una variable es considerada confundente, de confusión o factor extraño, cuando de manera total o parcial, explica el efecto observado o asociación de una exposición (factor de riesgo) sobre un resultado (enfermedad). El efecto aquí se refiere a una aparente relación o una aparente falta de relación. La aparente relación se refiere a que la variable de confusión es la causante de la relación; en el caso de la aparente falta de relación, la variable de confusión está "enmascarando" la relación entre la exposición o el factor de riesgo (Woodward 1999).

Pueden ser variable de confusión alguna característica, rasgo u otro factor que no fue considerado inicialmente en el diseño del estudio, lo cual puede conllevar a conclusiones erróneas o errores de información en los reportes. Algunas de las variables de confusión más frecuentes e importantes son: edad, sexo, raza, educación, entre otras. La edad es la variable de confusión más común en investigaciones epidemiológicas (Woodward 1999).

Con el objetivo de ejemplificar este concepto, considere un caso hipotético en donde el interés es evaluar si consumir café tiene o no un efecto en el cáncer de pulmón; para ello se establecen dos grupos: Los que consumen café y los que no lo consume. En este caso se está ignorando la relación entre el consumo de café y el tabaquismo, pues las personas que

fuman tienden a tomar café, lo cual sí puede estar asociado de una manera más razonable a la presencia de cáncer de pulmón. Así que si en este caso se ignora el consumo de tabaco existiría una falsa asociación entre tomar café y cáncer de pulmón y ésta se debe a la presencia de un factor de confusión que en este caso es fumar cigarrillo

2.2. Tipos de confusores

Los confusores pueden ser clasificados en dos categorías: una cualitativa y otra cuantitativa. Las primeras son aquellas que después de que se haga el ajuste, la asociación entre exposición y resultado desaparece por completo o se invierte la dirección de esta relación, esto significa que la calidad o naturaleza de la asociación cambia (Kamangar 2012), considere por ejemplo, las siguientes situaciones:

1. La obesidad, el estilo de vida sedentaria, la contaminación del aire y fumar hacen que la vida sea más corta. Así que, ¿por qué es que la gente tenía un lapso de vida más corto hace 500 años, cuando eran más delgados, eran más activos físicamente, se respiraban un aire más limpio, y se fumaba menos? la respuesta se encuentra en el efecto confusor de los avances de la vida moderna, como por ejemplo una mejor higiene y el desarrollo de vacunas y antibióticos.

2. Algunas investigaciones muestran que mujeres que han tenido múltiples embarazos, tienen un alto riesgo de tener hijos con síndrome de Down en el último de estos. Sin embago, se sabe que esta asociación no se encuentra relacionada al número de partos, sino a la edad de la madre; por ejemplo el décimo niño de una madre que tiene 26 años de edad en el embarazo puede tener un menor riesgo de síndrome de Down que el primer niño nacido de una madre que tiene 39 años. En este caso la edad está asociada con la exposición (partos) y el resultado (síndrome de Down) pero no viene entre ellos.

En el primer ejemplo, el confusor reversa la asociación real. Si no se hace el ajuste por los avances en la vida moderna, la conclusión al comparar la vida de ahora con los de hace 500 años podría conducir a una conclusión que fue exactamente opuesta a la verdadera. Para el segundo ejemplo, se tiene que el confusor es la edad y cuando se hace el ajuste por ésta, la asociación entre la exposición y el resultado desaparece.

Cuando se hace el ajuste con confusores cuantitativos, éste cambia la magnitud de la asociación (positiva o negativa) pero no su naturaleza. Los confusores positivos son aquellos que aumentan más allá de su tamaño real, es decir hacen parecer la asociación más grande

de lo que es; por ejemplo un estudio de cohorte en Irán mostró una asociación entre el consumo de opio[1] y la muerte (Kamangar 2012). En el estudio se encontró que el consumo de opio aumentaba el riesgo de morir en un 126 % (riesgo relativo de 2.26). Un potencial confusor en este caso sería el uso de tabaco, puesto que éste está asociado con la exposición (consumo de opio) dado que a mayor uso de tabaco, aumenta la probabilidad del consumo de opio y al mismo tiempo es más probable morir. Después del ajuste por fumador, edad y sexo, la asociación entre la exposición y el resultado es menos fuerte, ya que el riesgo relativo es de 1.86, pero no desaparece.
En el caso de los confusores negativos, hacen parecer la asociación más pequeña de lo que es.

En la **Figura 2-1**, se presenta un resumen de la clasificación de los confusores.

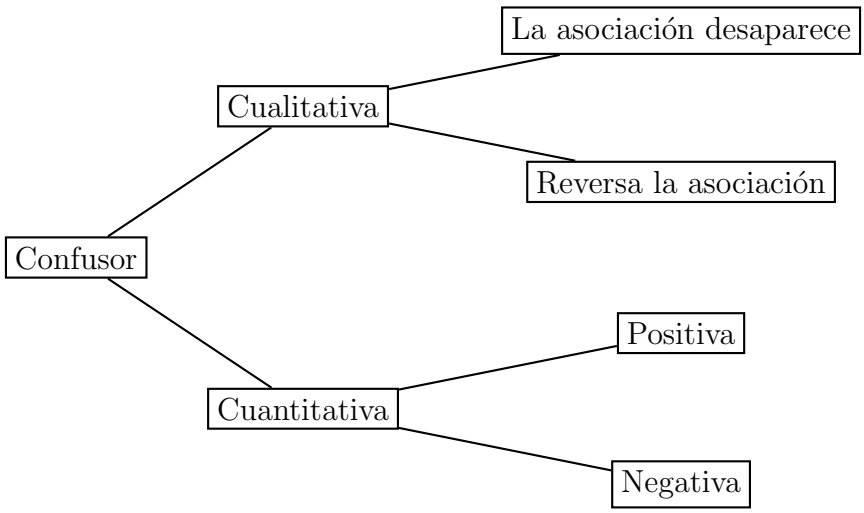

Figura 2-1.: Clasificación de los confusores

[1]Mezcla compleja de sustancias que contiene morfina

2.3. Identificación de variables de confusión

Lo que hace que una variable sea considerada confusora depende de las relaciones observadas a partir de los datos y también de un conocimiento a priori del supuesto proceso biológico, es decir, se requiere que el "experto" (quien tiene mayor conocimiento del tema) y el analista, realicen una lista de todas las variables potenciales que influyen en el estudio e identifiquen las posibles relaciones causales entre éstas.

Woodward (1999) y Kamangar (2012) indican que es necesario considerar las siguientes condiciones antes de que una variable sea considerada como factor de confusión.

- Un factor de confusión debe ser un factor de riesgo causal para el resultado, pero no puede ser consecuencia de éste. Por ejemplo, existe una relación entre colesterol alto y enfermedad coronaria, así que el colesterol alto podría ser una factor de confusión, pero en la relación tomar aspirina - infarto de miocardio es claro que la primera no podría ser un factor de confusión ya que es una consecuencia de la enfermedad.

- Debe estar asociado causalmente con el potencial factor de riesgo, más no una causa intermedia entre la relación exposición - resultado. En otras palabras, el factor de confusión no debe ser un "camino causal" entre el factor de riesgo y el resultado. Suponga por ejemplo que el factor de riesgo es dieta baja en vitamina C y que la persona no consume frutas; como el no consumir frutas implica una dieta baja en vitamina C, el no consumir frutas podría ser un factor de confusión. Pero si el factor de riesgo es fumar y el "confusor" es fibrinógeno[2], el factor de riesgo implica al "confusor" y entonces, en realidad, éste no sería un confusor, ya que es una consecuencia del factor de riesgo.

Para aclarar mejor las condiciones planteadas anteriormente por Woodward (1999) y Kamangar (2012), se utilizarán los denominados diagramas de ruta implementados por Schelesselman (1982) como una ilustración de las mismas. Para ello se adoptará la siguiente notación:
Asociación causal (F causa a E):

$$\mathrm{F} \longrightarrow \mathrm{E}$$

Asociación (puede ser causal o no causal):

$$F \longleftrightarrow E$$

[2]Proteína soluble del plasma sanguíneo producida por el higado que ayuda a detener el sangrado al favorecer la formación de coagulos en la sangre

donde E y F hacen referencia a enfermedad y factor de riesgo, respectivamente; la variable de confusión se notará con la letra C. Vale la pena resaltar que esta ilustración es para identificar si un factor o variable es o no un factor de confusión de la relación factor de riesgo - enfermedad.

La **Figura 2-2** muestra tres posibles situaciones en las que C es un factor de confusión. En la primera situación se puede ver que tanto el factor de riesgo como la enfermedad son a causa del confusor, de manera que satisface las dos condiciones planteadas en Woodward (1999), análogamente en las dos siguientes situaciones.

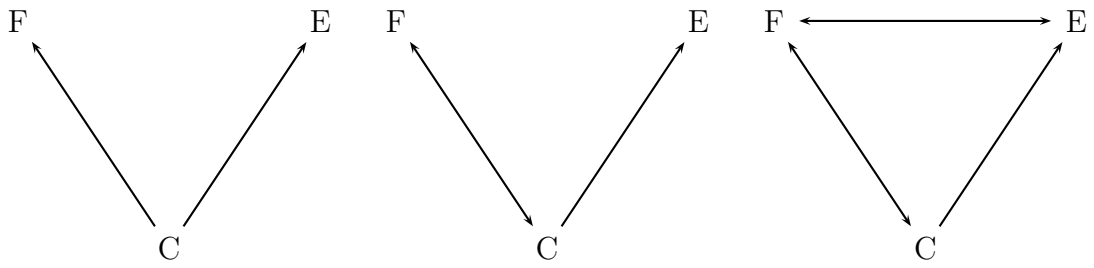

Figura 2-2.: Situaciones en las que C es un factor de confusión

En la **Figura 2-3** se ilustra tres situaciones en las que C no es un factor de confusión, particularmente, note en la tercera situación que C es una consecuencia de F, por lo tanto C no se considera como confusor para la relación $F - E$.

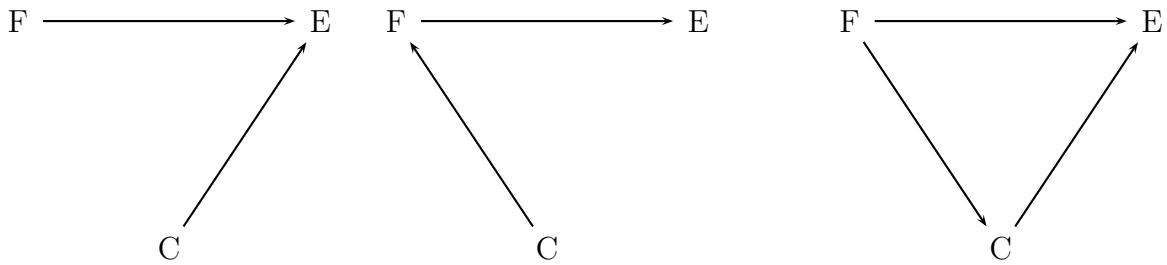

Figura 2-3.: Situaciones en las que C no es un factor de confusión

Ejemplos ilustrativos

A continuación se muestran algunos ejemplos reales como ilustración de los diagramas de rutas.

1. El ginseng es una hierba, principalmente cultivada en China y Corea, la cual es usada para propósitos medicinales. Algunas personas creen que puede fortalecer el cuerpo y prevenir enfermedades. Un estudio de cohorte que investigó la asociación de ginseng (F) con el cáncer gástrico (E) en China encontró que contrariamente a las espectativas, el ginseng incrementa el riesgo de cáncer gástrico en un 40 % (riesgo relativo de 1.40) (Kamangar 2012). Sin embargo, después de ajustar por edad (C), la asociación desaparece por completo y el ginseng no aumenta ni disminuye el riesgo, personas de avanzada edad fueron más propensos a usar ginseng y tenían más probabilidad de desarrollar cáncer gástico.

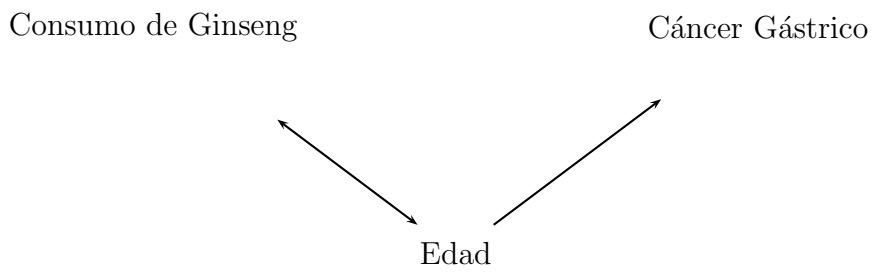

Figura 2-4.: Diagrama de ruta para la asociación consumo de ginseng - cáncer gástrico

2. El tabaquismo (F) y el fibrinógeno (C) son factores de riesgo para las enfermedades coronarias (Coronary Heart Disease, CHD, siglas en Inglés) (E) (Woodward 1999), pero fumar promueve el aumento de fibrinógeno, es decir, niveles altos de fibrinógeno son a causa del consumo de cigarrillo, así que los niveles altos de fibrinógeno son una consecuencia del consumo de cigarrillo y por lo tanto no es un factor de confusión.

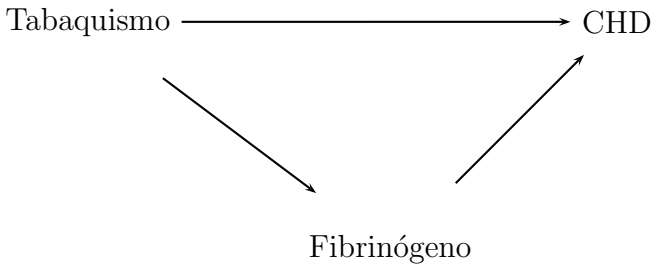

Figura 2-5.: Diagráma de ruta para la asociación fibrinógeno - CHD

2.4. Algunos métodos para evaluar confusión

Existen diferentes mecanismos para evaluar el efecto de alguna variable que se sospecha como un potencial confusor. Algunos de éstos, se basan en métodos analíticos tradicionales como estimadores y pruebas de hipótesis; otros métodos son implementados en el diseño del estudio antes de la recolección de los datos para controlar variables de confusión. En este capítulo se abordan los métodos más comunes para evaluar la presencia de estas variables.

2.4.1. Métodos analíticos tradicionales

Una vez los datos han sido recolectados se pueden utilizar métodos analíticos para evaluar el efecto de cualquier variable que es un potencial confusor. Al igual que en otros problemas, se opta por utilizar técnicas analíticas basadas en la estimación o pruebas de hipótesis. El primero es preferible, porque es el más sencillo (Woodward 1999).

Usando estimación

Los factores de confusión pueden ser evaluados mediante la estimación del efecto del factor de riesgo con y sin los factores de confusión (Woodward 1999), por ejemplo, suponga que

R_1 es el riesgo relativo [3] sin la presencia de la variable de confusión y que R_2 es el riesgo relativo en presencia de la variable de confusión; entonces la estimación del efecto de la variable de confusión se calcula como R_2/R_1. Un problema con este enfoque de razones de riesgo relativos es que la respuesta dependerá de la medida de probabilidad comparativa de enfermedad (expuestos versus no expuestos) usada. En lugar de usar el riesgo relativo se pueden usar razones de odds pero los resultados obtenidos pueden variar a menos que se esté estudiando una enfermedad "rara" (de baja incidencia) en cuyo caso la evaluación será similar.

Ahora bien, cuando los riesgos relativos para los distintos niveles del factor de confusión son iguales y diferentes al riesgo relativo cuando se ignora la variable de confusión, se dice que la confusión es perfecta (Woodward 1999). En la práctica es muy poco frecuente esta situación pero cuando los riesgos por estrato se parecen mucho al global la confusión no es importante.

Usando pruebas de hipótesis

Como se ha visto en algunas situaciones, las pruebas de hipótesis por sí mismas son de uso limitado. Aquí la situación es peor, porque no hay ninguna prueba directa para confusión. Lo que se puede hacer es evaluar si el factor de riesgo (F) y la enfermedad (E) están relacionados después de realizar el ajuste por la variable de confusión (C), esto a menudo se interpreta como un efecto ajustado de F en E. Es decir, se puede evaluar por una relación no ajustada entre F y E. Luego, se pueden comparar estos resultados con una prueba similar antes de hacer el ajuste. Si por ejemplo F está relacionado significativamente con E antes del ajuste pero no después del ajuste, hay evidencia de que el confusor candidato tiene un efecto significativo. Similarmente, cuando aún hay un efecto significativo de F sobre E después del ajuste de C se puede inferir que F tiene un efecto sobre E por encima de cualquier efecto de C. Este efecto significativo se denomina en epidemiología **efecto independiente**, esta independencia no es la misma independencia estadística, lo que significa es que el riesgo de enfermedad dada la exposición al factor de riesgo, es realmente el mismo si una persona esta expuesta al factor de confusión o no (Woodward 1999).

[3] Comparación de riesgos. el riesgo relativo se usa para decidir si la exposición al factor de riesgo tiene un efecto sustancial sobre la aparición de la enfermedad.

2.4.2. Métodos comunes para controlar variables de confusión

Pourhoseingholi et al. (2012), Newman (2001) y Kamangar (2012) indican que los métodos más comúnmente usados para controlar (por ajuste) confusión son: aleatorización, estratificación, restricción, apareamiento, y regresión. Aleateorización, restricción y apareamiento, son implementados en el diseño de estudio como estrategia para minimizar confusión. La estratificación y regresión son usados durante el análisis (Kamangar 2012), éstos dos últimos son los dos mayores enfoques para controlar confusión en estudios observacionales (prospectivos o retrospectivos) (Wunsch 2007).

Aleatorización

Este método permite que las variables se distribuyan homogéneamente, con ello desaparece la asociación entre la posible confusora y la exposición estudiada, lo cuál es una de las condiciones para que una variable sea de confusión. Wunsch (2007) afirma que la aleatorización es la mejor manera para evitar sesgos de confusión en estudios prospectivos.

El método de aleatorización consiste en asignar aleatoriamente los sujetos a los grupos de tratamiento y control, esto asegura en alta medida que los grupos difieran, solo porque uno de ellos recibe el tratamiento y el otro no; así la relación observada entre el tratamiento y el resultado no es afectada por el confusor, debido al hecho de que los dos grupos son similares en todos los aspectos excepto a que se recibe y no el tratamiento, particularmente en esta situación no hay sesgo de selección. La ausencia de éste último no significa sin embargo que el grupo etiquetado es homogéneo debido al efecto del tratamiento; el grupo puede contener sub grupos los cuales pueden variar ampliamente en respuesta al tratamiento (Wunsch 2007).

Una desventaja de este método es que cuando los experimentos son aleatorizados no se muestra cómo el tratamiento se usa en el mundo real, es decir no se puede responder a interrogantes como ¿quién desea el tratamiento? ¿quién lo optará?. Otra desventaja de la utilización de éste método, recae en que por razones morales no se pueden asignar aleatoriamente a los sujetos, sin embargo la aleatorización es bastante común en estudios clínicos (Wunsch 2007).

Restricción

Una de las condiciones necesarias para que se produzca confusión es que el factor de confusión sea distribuido de manera desigual entre los grupos que se comparan; así que podría pensarse, que una estrategia para evitar confusión es incluir en el estudio sólo al grupo de

sujetos que tienen los mismos niveles de los factores de confusión. Por ejemplo, considere un caso hipotético en el que se estudia la asociación entre la actividad física y las enfermedades del corazón, y además suponga que la edad y el género son los dos factores que fueron potenciales confusores en dicho estudio. Si es así, la confusión de estos factores podrían haberse evitado, por ejemplo asegurándose de que todos los sujetos fueran hombres con edades entre 30-50. Esto asegurará entonces que las distribuciones de edad son similares en los grupos que se comparan, por lo tanto la confusión se minimizará.

De acuerdo con lo anterior, con este procedimiento se trata de eliminar la variación en el factor de confusión restringiendo la población bajo estudio a una sola categoría de éste, además se consigue romper la distribución heterogénea del factor de confusión entre los grupos comparados (Kamangar 2012). Este método se conoce en la literatura como restricción.

Ahora bien, al reducir el número de individuos en el estudio se tendrían problemas con el tamaño de la muestra, y por lo tanto una pérdida en la potencia del mismo. Otra desventaja de la restricción es que limita la participación en el estudio a sujetos que son similares con respecto a la variable de confusión, lo cual limita la capacidad de generalizar los resultados del estudio y además no sería posible evaluar la relación de interés en los diferentes niveles del factor de confusión.

Matching

El matching (o pareo) es un enfoque popular para controlar confusión en estudios de casos y controles. Este método obliga a que estos grupos sean similares con respecto a los factores de riesgos más importantes y por lo tanto hace que las comparaciones de casos y controles estén menos sujetas a factores de confusión.
El primer paso en matching es identificar un caso. Los investigadores luego seleccionan de la población de origen uno o más controles potenciales que tienen los mismos valores que el caso tiene para cada factor de matching.

El matching es común en estudios clínicos, en particular cuando la enfermedad de interés es muy poco frecuente, donde hay un número pequeño de casos posibles y un gran número de posibles controles. El matching puede aumentar la eficiencia estadística de las comparaciones de casos y controles y así lograr un determinado nivel de potencia estadística con un tamaño de muestra pequeño.
Entre las ventajas del matching se destacan:

1. Hay un control directo de los factores de confusión. Esto significa que el ajuste de la relación entre el factor de riesgo y la enfermedad es logrado automática e intuitivamente, lo cuál conduce a una clara interpretación.

2. Asegura que el ajuste es posible. En circunstancias excepcionales puede no haber coincidencia entre los casos y un conjunto de controles muestreados al azar. Por ejemplo, todos los casos pueden ser ancianos, pero los controles no incluir ninguna persona anciana. Así el ajuste por edad, usando el método de Mantel - Haenszel o algún otro procedimiento no sería posible (Woodward 1999).

3. Bajo ciertas condiciones el matching prueba la eficiencia de la investigación.

Estratificación

Con la estratificación se busca fijar los niveles de los factores de confusión y producir grupos dentro de los cuales el confusor no varíe, y entonces evaluar la asociación entre el resultado y la exposición dentro de cada estrato del confusor. Así que dentro de cada estrato, el factor de confusión no puede confundir, ya que no varían a través de la relación exposición - resultado (Pourhoseingholi et al. 2012).

Luego de que se realice la estratificación, se emplea el estimador de Mantel - Haenszel para probar el resultado del ajuste de acuerdo con los estratos. Si existe una diferencia entre el estimador crudo (sin la estratificación) y el ajustado (producto de los estratos) la confusión es bastante probable; en el caso de que el segundo estimador no difiera mucho del primero, la confusión no es probable.

Existen varios inconvenientes en la aplicación de este método, uno de ellos es que pueden resultar tablas con celdas cuyos valores sean pequeños o incluso cero, esto puede ocurrir cuando un gran número de factores de confusión tiene que ser controlado de forma simultánea. Otro inconveniente de la estratificación cuando hay que controlar muchos factores de confusión, sobre todo cuando son variables de tipo continuo, es el planteamiento de los estratos, es decir, si se presenta esta situación ¿cómo se deben crear los estratos?. Una solución cuando se presenta este tipo de situaciones se basa en el puntaje de propensión (Rosenbaum & Rubin (1983), Joffe & Rosenbaum (1999)), que es una función definida en términos de factores de riesgo conocidos (aparte de la exposición de interés) que da la probabilidad de que un individuo pertenezca a la población expuesta. Si los estratos son creados mediante la agrupación de individuos con el mismo puntaje de propensión, los temas expuestos y no expuestos en cada estrato se equilibrará en los factores de riesgo conocidos (Newman 2001). Estas puntuaciones pueden estimarse a través de modelos de

regresión. Pese a todos los inconvenientes de la estratificación, ésta se utiliza casi siempre en las etapas exploratorias de un análisis de datos epidemiológicos.

Modelos multivariados

Los modelos multivariados son de gran utilidad cuando se tienen una gran cantidad de estratos o un gran número de covariables, inclusive un gran número de confusores, simultáneamente (Pourhoseingholi et al. 2012), esto hace que este método sea más eficiente que el método de estratificación, ya que con el mismo tamaño de muestra se obtienen estimaciones más precisas y con un número mayor de variables. El análisis de regresión es el método más común para el ajuste por factores de confusión en estudios observacionales (Kamangar 2012).

Debido a que el análisis de regresión permite estimar la asociación entre una variable independiente y el resultado manteniendo las demás variables constantes, proporciona una forma de ajustar por potenciales confusores que han sido incluídos en el modelo. Así que el efecto del factor de riesgo (o exposición) se puede estimar mientras se retienen los niveles del confusor.

Una vez que la variable es identificada como confusora, ésta se incluye en el modelo y estima nuevamente la asociación entre la exposición y el resultado. Luego se realiza una prueba de significancia estadística para el coeficiente asociado a la exposición para determinar si la asociación entre éste y el resultado sigue siendo estadísticamente significativa.

3. Confusión en modelos de regresión

En modelos de regresión, la confusión es un fenómeno asociado a la multicolinealidad, ya que ésta ocurre cuando las covariables están altamente correlacionadas y por tanto los efectos individuales sobre la variable respuesta son casi imposibles de distinguir; por lo tanto la colinealidad o multicolinealidad puede verse como un caso extremo de confusión. En la primera parte de este capítulo se presentan los modelos de regresión que se utilizarán en este trabajo, y luego se presentan los métodos o técnicas que se exponen en la literatura para determinar si la confusión está presente o no en un modelo de regresión.

3.1. Modelo lineal generalizado (MLG)

Los modelos lineales generalizados son una extensión de los modelos lineales clásicos basados en la distribución normal, la extensión hace referencia a dos aspectos principalmente, la parte lineal de los modelos clásicos y una variedad de distribuciones seleccionadas de una familia especial, la cual desde los tiempos de Fisher se conoce como familia exponencial (Lindsey 2007). Estos modelos son de gran utilidad en situaciones en donde la variable respuesta no necesariamente sigue una distribución normal o cuando las variables explicativas y los valores esperados de la respuesta no tienen una relación lineal, pero que se pueden linealizar por medio de una función que se conoce como función de enlace.

3.1.1. Componentes del MLG

Las tres componentes que especifican un MLG son: una componente aleatoria que identifica a la variable respuesta y su distribución de probabilidad; una componente sistemática que especifica a las variables explicativas usadas en una función lineal, y una función de enlace (link, conexión o vínculo) que relaciona las medias del modelo con las componentes sistemáticas (Agresti 2002).

- El conjunto de variables aleatorias independientes ($\mathbf{y}$) pertenecientes a la familia exponencial.
- Una matriz de diseño ($\mathbf{X}$) y un vector de parámetros ($\boldsymbol{\theta}$).

- Una función de enlace, vínculo, conexión o link (g) que relaciona las medias del modelo lineal.

Dentro de los modelos que pertenecen a la familia exponencial se encuentran el modelo lineal clásico, el modelo loglineal, modelo binomial, la regresión Poisson, entre otros.

La construcción de un MLG puede hacerse escogiendo adecuadamente la función de enlace y la distribución de probabilidad (Dobson 2001). Por ejemplo, en el modelo lineal clásico la distribución de probabilidad es la normal y la función de enlace es la identidad. En el modelo de regresión logística, la distribución de probabilidad es la binomial y la función de enlace es la transformación logit:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$$

Para el modelo de regresión Poisson la distribución es la Poisson y la función de enlace es $g(\mu) = \ln(\mu)$, donde ln se refiere al logaritmo natural.

3.1.2. Deviance

El deviance es un estadístico para evaluar el ajuste de un modelo lineal generalizado. El deviance (D) es definido de la siguiente forma:

$$D = 2[l(\beta, y) - l(\hat{\beta}, y)] \tag{3-1}$$

donde $l(\beta, y)$ es el logaritmo natural de la función de verosimilitud para el modelo saturado (completo) evaluado en β, y $l(\hat{\beta}, y)$ denota el máximo valor de la función de verosimilitud para el modelo de interés. Entonces, el deviance es escencialmente una medida de discrepancia entre el modelo postulado y el modelo completo o saturado. Valores grandes de D sugieren que el modelo postulado no es adecuado para los datos.

La distribución de D es $D \sim \chi_{n-p}$, donde n corresponde al número de parámetros en el modelo saturado y p es el número de parámetros en el modelo de interés (Dobson 2001).

3.2. Familia exponencial

Sea Y una variable aleatoria, cuya distribución de probabilidad sólo depende del parámetro θ. Se dice que la distribución de Y pertenece a la familia exponencial, si la función de densidad de probabilidad (f.d.p) se puede escribir de la siguiente forma, (Dobson 2001):

$$f(y;\theta) = s(y)t(\theta)\exp\{a(y)b(\theta)\} \tag{3-2}$$

donde a, b, s y t son funciones conocidas. Así, (3-2) puede reescribirse como

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\} \tag{3-3}$$

donde $s(y) = \exp\{d(y)\}$ y $t(\theta) = \exp\{c(\theta)\}$.
Si $a(y) = y$ se dice que la distribución está en la forma canónica o estándar, y $b(\theta)$ es llamado parámetro natural de la distribución.

Una versión más general de (3-2) es

$$f(y;\theta) = s(y)t(\theta)\exp\left\{\sum_{i=1}^{n} b_i(\theta)a_i(y)\right\} \tag{3-4}$$

Si además del parámetro de interés (θ) hay otros parámetros adicionales a estos, son considerados como parámetros de perturbación y forman parte de las funciones de a, b, c y d y se tratan como si fueran conocidos (o que pueden ser previamente estimados).

3.2.1. Modelo de regresión lineal clásico

Son modelos de la forma: (modelo individual)

$$E(Y_i) = \mu_i = \mathbf{x_i^t}\boldsymbol{\beta} \tag{3-5}$$

donde $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ $, i = 1, 2, .., n$. La función de enlace es la identidad, esto es $g(\mu_i) = \mu_i$. Usualmente este modelo se escribe como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3-6}$$

donde, $\mathbf{X}$ se conoce como matriz de diseño, $\boldsymbol{\beta}$ es el vector de parámetros y los $\mathbf{e}$ son variables aleatorias independientes, idénticamente distribuidas con $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I)$ para $i = 1, 2, ..., n$. Los modelos de regresión lineal múltiple, el análisis de varianza (ANOVA) y el análisis de covarianza (ANCOVA) son de esta forma y algunas veces son llamados **modelo lineal general** (Dobson 2001).

3.2.2. Regresión logística

Los métodos de regresión se han convertido en una componente integral del análisis de datos que trate de describir la relación entre una variable respuesta y una o más variables

explicativas (Hosmer & Lemeshow 1999).

En estos métodos la cantidad clave es el valor de la media de la variable respuesta, dado el valor de la variable independiente. Esta cantidad es llamada media condicional y es expresada como $E(Y|x)$, donde Y denota a la variable respuesta y x denota el valor de la variable independiente. En regresión lineal se asume que $E(Y|x)$ puede ser expresada como una ecuación lineal en x (o de alguna transformación de x o Y) de la forma

$$E(Y|x) = \beta_0 + \beta_1 x \tag{3-7}$$

Recuerde que (3-7) se conoce como modelo de regresión lineal y que la variable respuesta es asumida como continua. Ahora bien, es frecuente encontrarse con el caso donde la naturaleza de la variable respuesta sea de tipo categórica o discreta, por ejemplo, considere el caso donde el análisis busca determinar los éxitos y los fracasos, los cuales han sido codificados como 1 y 0, respectivamente. Si en este caso se implementara la ecuación (3-7), es posible que $E(Y|x)$ tome valores entre $-\infty$ y ∞, lo cuál no tendría ningún sentido, puesto que Y solo toma valores de cero y uno. Sin embargo, para este caso se podría estar interesado en modelar la probabilidad de un resultado exitoso dado un valor de x, esto es $P(Y = 1|x) = \pi(x)$, de esta forma se pensaría que $Y|x \sim Ber(\pi(x))$ y de esta forma el modelo de regresión lineal simple para $\pi(x)$, sería

$$E(Y|x) = \pi(x) = \beta_0 + \beta_1 x \tag{3-8}$$

Los modelos de este tipo son llamados modelos de probabilidad lineal (Friendly 2000), sin embargo estos modelos tienen el grave defecto de producir ($\pi(x) < 0$) para x muy pequeños, y $\pi(x) > 1$ cuando x es muy grande (asumiendo $\beta > 0$); y lo que se quiere es que el modelo proporcione directamente la probabilidad de un éxito ($Y = 1$) dado un valor de x.

Una solución a este problema es transformar a la variable respuesta de algún modo para garantizar que la respuesta prevista este entre cero y uno. Por lo tanto, se debe hallar una función F, tal que:

$$\pi(x) = F(\beta_0 + \beta_1 x) \tag{3-9}$$

garantice que $0 \leq \pi(x) \leq 1$.

Recuerde que la función de distribución pertenece a la clase de funciones no decrecientes, acotadas entre cero y uno, así que el problema se resuelve tomando como F cualquier función de distribución.

Particularmente se toma a F como la función de distribución logística, la cuál está dada por la expresión $1/(1+\exp\{-(\beta_0+\beta_1 x)\})$, y por lo tanto

$$\pi(x) = \frac{1}{1+\exp\{-(\beta_0+\beta_1 x)\}} \tag{3-10}$$

La expresión dada en (3-10) se conoce como modelo logístico, en su forma más simple (Diaz & Morales 2009). En el caso más general, que involucra k variables explicativas $(x_1, x_2, \ldots, x_k)$, es el siguiente

$$\pi = \frac{1}{1+\exp\{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)\}} \tag{3-11}$$

Donde $\beta_0, \beta_1, \ldots, \beta_k$ son los parámetros del modelo. Las covariables $x^{'}(s)$ pueden ser de tipo discreto o continuo.

Por otra parte, la probabilidad de que el resultado sea un fracaso en el modelo logístico simple, está dada por:

$$\begin{aligned} 1-\pi(x) &= 1-1/(1+\exp\{-(\beta_0+\beta_1 x)\}) \\ &= (1+\exp\{-(\beta_0+\beta_1 x)\})/(1+\exp\{-(\beta_0+\beta_1 x)\}) - 1/(1+\exp\{-(\beta_0+\beta_1 x)\}) \\ &= \frac{\exp\{-(\beta_0+\beta_1 x)\}}{1+\exp\{-(\beta_0+\beta_1 x)\}} \end{aligned} \tag{3-12}$$

Una interpretación adecuada de los coeficientes $\beta^{'}(s)$ de un modelo de regresión logística, va de la mano con los conceptos de riesgo relativo, odds [1] y razones de odds.
Para este caso, en la expresión (3-13), expresa el odds como un modelo multiplicativo

$$\begin{aligned} \frac{\pi(x)}{1-\pi(x)} &= \frac{1/(1+\exp\{-(\beta_0+\beta_1 x)\})}{\exp\{-(\beta_0+\beta_1 x)\}/(1+\exp\{-(\beta_0+\beta_1 x)\})} \\ &= \frac{1}{\exp\{-(\beta_0+\beta_1 x)\}} \\ &= \exp\{\beta_0+\beta_1 x\} \end{aligned} \tag{3-13}$$

Una transformación del odds fundamental en la regresión logística es la transformación logit $(g(x))$

[1] El odds es un número que expresa cuánto más probable es que se produzca un evento frente a que no se produzca.

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

La importancia de ésta es que comparte algunas propiedades deseables de un modelo de regresión lineal, como es la linealidad en sus parámetros, puede ser continua y puede variar en un rango de $-\infty$ a $+\infty$, dependiento del rango de x (Hosmer & Lemeshow 1999). Así, la función de enlace para este modelo es $g(x)$ que corresponde al logaritmo natural del odds.

3.2.3. Regresión Poisson

La regresión Poisson se usa cuando la respuesta varía de acuerdo a varios niveles o cantidades de exposición y éstos deben tenerse en cuenta cuando se modela la tasa promedio de eventos (Dobson, 2001).

Considere $Y_1, \ldots, Y_n$ variables aleatorias independientes, donde Y_i denota el número de eventos observados en n_i exposiciones para el i-ésimo patrón de covariables, el valor esperado de Y_i puede escribirse como

$$E(Y_i) = n_i\theta_i$$

La dependencia de θ_i con las variables explicativas usualmente se modela por

$$\theta_i = \exp\{\mathbf{x_i^t}\boldsymbol{\beta}\} \tag{3-14}$$

y por lo tanto el MLG es

$$E(Y_i) = \mu_i = n_i \exp\{\mathbf{x_i^t}\boldsymbol{\beta}\} \tag{3-15}$$

donde $Y_i \sim \text{Poisson}(\mu_i)$. La función enlace en este modelo, es el logaritmo natural:

$$\ln(\mu_i) = \ln(n_i) + \mathbf{x_i^t}\boldsymbol{\beta}$$

3.3. Modelo de riesgo proporcional

Este modelo también conocido en la literatura como modelo de Cox (Hosmer & Lemeshow 1999) permite analizar datos de sobrevida, es decir, datos donde la respuesta es el tiempo hasta la ocurrencia de un evento de interés. La expresión general del modelo está dada por la siguiente ecuación:

$$h(t, x, \beta) = h_0(t) r(x, \beta) \tag{3-16}$$

Note que este modelo es el producto de dos funciones, una relacionada con los tiempos y otra que relaciona las covariables. Han sido consideradas algunas parametrizaciones para $r(x, \boldsymbol{\beta})$, sin embargo la más implementada es $r(x, \boldsymbol{\beta}) = \exp\{x\beta\}$ (Hosmer & Lemeshow 1999), la cuál, al igual que el modelo (**??**) fue propuesta por Cox (1972) y permite gran facilidad a la hora de interpretar los coeficientes β además, garantiza que $h(t, x, \beta)$ siempre será positiva. Por lo tanto el modelo de Cox es de la forma

$$h(t, x, \beta) = h_0(t) e^{x\beta} \tag{3-17}$$

Cuando $x = 0$, $h(t, x, \beta) = h_0(t)$ y entonces la función $h_0(t)$ es llamada función de base o función baseline.
Por otro lado, la linealización del modelo (3-17) se logra a través de la transformación logaritmo natural, ésta será denotada por $g(t, x, \beta)$ y que en adelante llamaremos función log-hazard, está dada por la siguiente ecuación

$$g(t, x, \beta) = \ln[h_0(t)] + x\beta \tag{3-18}$$

3.4. Confusión en modelos de regresión

Cuando se observa una asociación, se cuestiona si se trata de una asociación real o de una falsa asociación; y si la asociación observada es real, ¿se considera también causal? Estas y otras preguntas son las que se puede plantear un investigador que intenta explicar un resultado a través de una exposición.
En la literatura se encuentran propuestos varios mecanismos para detectar falsas asociaciones en los modelos de regresión. En este apartado se presenta un resumen de algunos de estos mecanismos.

3.4.1. Criterio del cambio porcentual

Considere el estudio del UMARU IMPACT (University of Massachusetts Aids Research Unit) mostrado en Hosmer & Lemeshow (1999), el cual tiene como propósito comparar los programas de tratamiento de diferentes duraciones planeadas designadas para reducir el abuso de las drogas y evitar comportamientos de alto riesgo de HIV (Human Immunodeficiency Virus) . En dicho estudio, se recolectan dos variables: edad (a) y uso histórico de la droga (d). El uso histórico de la droga fue recolectado en una variable dicótoma

$d(1 = \text{Si}, 0 = \text{No})$. Se asume el logaritmo del hazard como una función lineal en la covariables, cuyo principal objetivo es estimar la razón de asociación del hazard con el uso de la droga, d.

Inicialmente se asume un modelo, que contiene sólo el uso de la droga, donde la función log-hazard es

$$g(t, d, \theta_1) = \ln[h_0(t)] + d\theta_1$$

Ahora bien, al hacer la diferencia en la función log-hazard, teniendo en cuenta la codificación del uso de la droga se tiene

$$\begin{aligned} g(t, d = 1, \theta_1) - g(t, d = 0, \theta_1) &= \{\ln[h_0(t)] + 1\theta_1\} - \{\ln[h_0(t)] + 0\theta_1\} \\ &= \theta_1 \end{aligned} \tag{3-19}$$

Considere un segundo modelo, el cuál además del uso de la droga, contiene la edad; entonces la función log-hazard es

$$g(t, d, a, \boldsymbol{\beta}) = \ln[h_0(t)] + d\beta_1 + a\beta_2$$

Luego, la diferencia en la función log-hazard es

$$\begin{aligned} g(t, d = 1, a, \boldsymbol{\beta}) - g(t, d = 0, a, \boldsymbol{\beta}) &= \{\ln[h_0(t)] + 1\beta_1 + a\beta_2\} - \{\ln[h_0(t)] + 0\beta_1 + a\beta_2\} \\ &= \beta_1 + a\beta_2 - a\beta_2 \\ &= \beta_1 \end{aligned}$$

Con este resultado se asume entonces, que la diferencia de la función log-hazard es constante para todas las edades.

De esta forma se tienen dos estimadores para la función log-hazard; uno llamado estimador crudo (θ_1) y otro, el ajustado notado por β_1. En el caso de que los estimadores β_1 y θ_1 sean similares, el ajuste por la edad sería innecesario. Para el caso opuesto, si los estimadores son diferentes, el ajuste fue necesario y entonces se considera que la variable edad es un confusor de la razón hazard para d.

Suponga que el modelo correcto, es el que contiene la edad y denota el promedio de la edad de los sujetos con y sin los grupos de uso histórico de la droga con $\bar{a}_1$ y $\bar{a}_0$, respectivamente.

Una aproximación para el promedio del logaritmo de la función de hazard para los dos grupos de uso de la droga es

$$g(t, d = 0, \beta) = \ln[h_0(t)] + \bar{a}_0\beta_2$$

$$g(t, d = 1, \beta) = \ln[h_0(t)] + \beta_1 + \bar{a}_1\beta_2$$

Ahora

$$\begin{aligned} g(t, d = 1, \beta) - g(t, d = 0, \beta) &= \ln[h_0(t)] + \beta_1 + \bar{a}_1\beta_2 - \{\ln[h_0(t)] + \bar{a}_0\beta_2\} \\ &= \beta_1 + \bar{a}_1\beta_2 - \bar{a}_0\beta_2 \\ &= \beta_1 + (\bar{a}_1 - \bar{a}_0)\beta_2 \end{aligned} \tag{3-20}$$

De (3-19) y (3-20) se tiene que

$$\hat{\theta}_1 \approx \hat{\beta}_1 + (\bar{a}_1 - \bar{a}_0)\hat{\beta}_2 \tag{3-21}$$

Si en (3-21) $\bar{a}_1 - \bar{a}_0 = 0$ o si los coeficientes de la edad son iguales a cero, los estimadores serían aproximadamente iguales. Ahora, los dos estimadores podrían diferir si el tamaño de al menos uno de los dos es grande o moderado.

Hosmer & Lemeshow (1999) recomiendan que el porcentaje de cambio en la estimación ajustada se calcula como una medida de la cantidad de ajuste. El estimador del porcentaje de cambio, en general es definido como:

$$\Delta\hat{\beta}\,\% = 100\frac{\hat{\theta} - \hat{\beta}}{\hat{\beta}} \tag{3-22}$$

donde $\hat{\theta}$ denota el estimador crudo del modelo que no contiene el potencial confusor y $\hat{\beta}$ denota el estimador ajustado del modelo que incluye el potencial confusor.

A pesar de que no existen reglas en el porcentaje de cambio como medida de confusión, Hosmer & Lemeshow (1999) sugieren que ésta está presente en el modelo si $\Delta\hat{\beta}\,\%$ es más grande que $15 - 20$ por ciento, por recomendaciones de Mickey & Greenland (1989) se usa un nivel convencional del $20\,\%$.

3.4.2. Criterio propuesto

Woodward (1999) presenta dos posibles alternativas para determinar la presencia de confusión en modelos de regresión. La primera consiste en comparar los deviances, para el

modelo con el factor de riesgo contra el modelo que contiene el confusor y el factor de riesgo, esto indica si el factor de riesgo sigue siendo importante después del ajuste por el confusor; y la segunda se basa en la comparación de las razones de odds de estos modelos. Pese a que la última no es una prueba formal de significancia, se concluye que la confusión no está presente en el estudio cuando las razones de odds, sin ajuste y con ajuste, son muy similares.

Partiendo de la segunda propuesta de Woodward (1999) y dada la importancia interpretativa del odds en un modelo de regresión logística, se considera más interesante mirar cuanto se ve afectado el odds cuando una variable de confusión se encuentra presente en el modelo. Así que se considera el criterio del cambio porcentual propuesto por Hosmer & Lemeshow (1999) como medida de confusión, pero en lugar de utilizar los coeficientes del modelo, se usa el odds en presencia y ausencia de confusión. Por lo tanto el criterio propuesto está dado por la siguiente expresión

$$CP = \left| \frac{\exp\{\theta\} - \exp\{\beta\}}{\exp\{\beta\}} \right| \tag{3-23}$$

donde $\exp\{\theta\}$ y $\exp\{\beta\}$, denotan los odds del modelo sin ajuste por confusión y con ajuste, respectivamente.

4. Estudio de simulación

El interés es evaluar vía simulación estadística, el comportamiento del criterio de cambio porcentual dado en (3-22), como medida de confusión en modelos de regresión, para algunos miembros de la familia exponencial.
En todos los casos se simulan situaciones similares a un estudio de corte transversal y por efectos de ilustración se asumieron dos variables explicativas; la primera que hace referencia al factor de riesgo y la segunda al factor de confusión. Vale la pena resaltar además, que en los escenarios simulados, el factor de confusión estaba asociado con el factor de riesgo y el resultado.

4.1. Diseño del estudio de simulación

Los tamaños de muestras asumidos en el estudio de simulación fueron $n = 50, 100, 200, 500$ y 1000. Las variables explicativas, fueron generadas de una distribución normal bivariada con vector de medias cero y correlación $\rho = -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8$. Las muestras generadas para la variable respuesta, se tomaron de una distribución normal, Bernoulli y Poisson, para los modelos lineal clásico, logístico y Poisson, respectivamente.

El mecanismo usado para generar la dependencia entre la variable respuesta y las variables explicativas está basado en la función de enlace de los modelos considerados. Dado que en cada una de estas funciones se requería del conocimiento de los parámetros del modelo, se asignaron diferentes valores a éstos como se muestra a continuación:

1. Para el modelo logístico dado en la expresión (3-11), los valores conocidos de los parámetros $\beta_0, \beta_1, \beta_2$ se tomaron de Austin & Brunner (2004), quienes consideran los siguientes escenarios de acuerdo a un estudio realizado sobre la relación entre el confusor, el factor de riesgo y la probabilidad de observar el resultado de interés.
    - Escenario 1: $\beta_0 = 0, \beta_1 = 0, \beta_2 = 3$
    - Escenario 2: $\beta_0 = -0, \beta_1 = 0, \beta_2 = 1$
    - Escenario 3: $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0.5$

- Escenario 4: $\beta_0 = -2.9, \beta_1 = 0, \beta_2 = 1$
- Escenario 5: $\beta_0 = -2.9, \beta_1 = 0, \beta_2 = 0.5$

Como se observa en la Figura **4-1**, con los escenarios 1 y 2 la probabilidad del resultado permite variar en un rango de muy bajo a muy alto, particularmente con el primer escenario se permiten más probabilidades extremas. El tercer escenario, permitía probabilidades moderadas para la mayoría de los sujetos. Finalmente, en los dos últimos escenarios la probabilidad del resultado es baja para la mayoría de los sujetos.

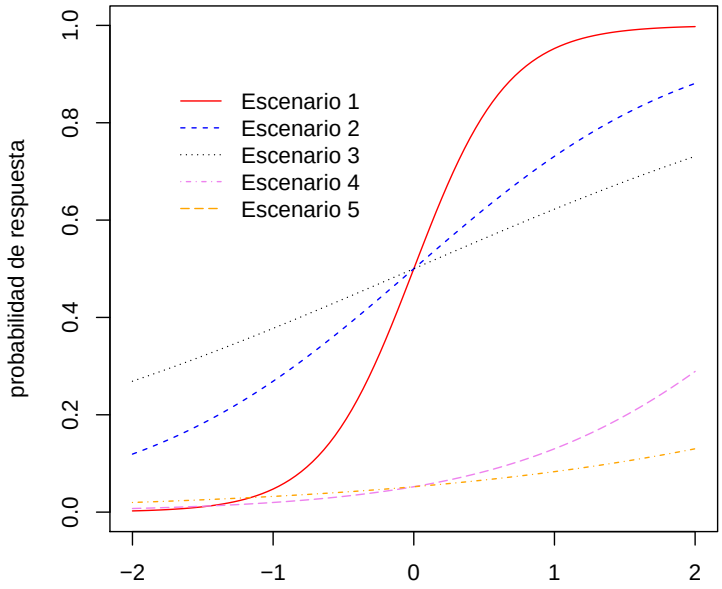

Figura 4-1.: Escenarios para el modelo logístico

2. El conjunto de escenarios en los $\beta^{'}(s)$ para el modelo Poisson dado en (3-14), se hace en forma análoga al modelo logístico, puesto que con éstos se permitía que la media λ de la distribución Poisson tuviera menos variabilidad a medida que se avanzara en los escenarios. Para ejemplificar, note que en la Figura **4-2**(a) λ varía entre 0 y 50, y en **4-2**(c) varía entre 0 y 20.

   - Escenario 1: $\beta_0 = 2.9, \beta_1 = -0.3, \beta_2 = -0.2$

- Escenario 2: $\beta_0 = 3, \beta_1 = 0.02, \beta_2 = -0.2$
- Escenario 3: $\beta_0 = 1.9, \beta_1 = 0.1, \beta_2 = 0.4$
- Escenario 4: $\beta_0 = 1.2, \beta_1 = 0.09, \beta_2 = 0.05$
- Escenario 5: $\beta_0 = -0.7, \beta_1 = 0.2, \beta_2 = 0.5$

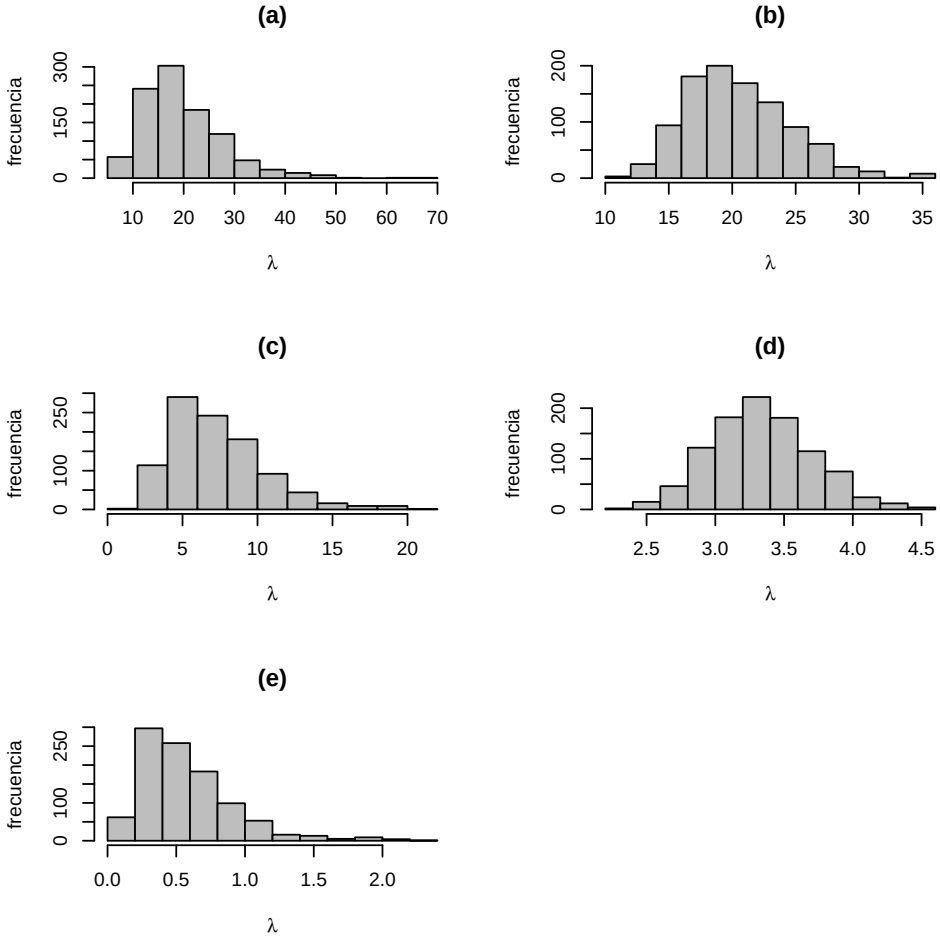

Figura 4-2.: Escenarios para el modelo Poisson (a) Escenario 1, (b) Escenario 2, (c) Escenario 3, (d) Escenario 4, (e) Escenario 5

3. A diferencia de los dos modelos anteriores, en el modelo lineal clásico no se puede pensar en la variable respuesta como "algo" grande, mediano o pequeño, debido a la naturaleza de ésta; así que se trabajó con un sólo escenario y los valores de los $\beta'(s)$ fueron tomados arbitrariamente.

$$\beta_0 = -0.6, \beta_1 = 1.3, \beta_2 = 1.4$$

Luego de haber generado las variables explicativas ($x^{'}(s)$) y la variable respuesta (y), se ajustan dos modelos. Inicialmente, el primer ajuste se realizó con la primera variable explicativa (x_1), la cuál correspondía al factor de riesgo, y de este ajuste se extrae el parámetro estimado de esta variable; a éste se le llamará estimador crudo (θ); posteriormente, se ajusta otro modelo con las dos variables explicativas: el factor de riesgo (x_1) y el confusor (x_2), del qué se extrae nuevamente el parámetro estimado de x_1, a éste se le llama estimador ajustado (β). Finalmente, se calcula $\Delta\hat{\beta}\,\%$. Este procedimiento se repite $N = 5000$ veces para cada uno de los modelos estudiados, bajo cada uno de los diferentes escenarios de n y ρ.

En las simulaciones, el desempeño de (3-22) se evaluó utilizando diferentes niveles de referencia $\delta = 0.05, 0.10, 0.15, 0.20$; este desempeño se determina calculando la proporción de veces en que el criterio es inferior a estos niveles.
Debido a que se están asumiendo diferentes grados de correlación entre las variables explicativas, por lo tanto niveles bajos y altos de confusión en el modelo, se determina que el criterio es válido en los modelos mencionados anteriormente si la proporción de veces en la que $\Delta\hat{\beta}\,\% < \delta$:

- es cercana a cero y $\rho \to \pm 1$, es decir cuando existe un alto grado de confusión en el modelo.
- es cercana a uno y $\rho \to 0$, es decir cuando la confusión no esté presente en el modelo.

4.2. Resultados

En este apartado se presentan los resultados del proceso de simulación antes descrito con el fin de evaluar el criterio del cambio en la estimación de los parámetros en los modelos lineal clásico, logístico y Poisson.

Los programas usados para el estudio de simulación se desarrollaron usando el paquete estadístico **R** (R Development Core Team) versión 3.1.1. En el apéndice se encuentra documentado el programa de simulación que se implementó.

La presentación de los resultados se desarrolla en dos partes, una para los modelos logístico y Poisson y otra para el modelo lineal clásico. Los resultados en los dos primeros modelos estudiados se encuentran a su vez divididos en los cinco escenarios considerados, así mismo,

éstos estan divididos por cada nivel de referencia δ.
En cada tabla, la primera columna está etiquetada con modelo, indicando los modelos considerados en el estudio; la segunda columna, etiquetada por ρ representa las correlaciones impartidas en el estudio. Las cinco columnas siguientes conforman los tamaños muestrales (n) asumidos en las simulaciones.

En todas las tablas, los resultados representan la proporción de veces en las que el criterio del cambio porcentual fue inferior a los niveles de referencia asumidos $(0.05, 0.10, 0.15, 0.20)$ para cada modelo, y los resultados que se encuentran entre paréntesis, corresponden a el criterio propuesto para el modelo logístico.

4.2.1. Modelo logístico y Modelo Poisson

Escenario 1

En las **Tablas 4-1** a **4-4** se presentan los resultados obtenidos para este escenario, el cuál permitía en el modelo logístico variar la probabilidad de la respuesta de un rango muy bajo a muy alto; y para el modelo Poisson también se permitía un amplio rango de valores para la variable respuesta (ver **Figura 4-2**(a)).

En estas tablas se puede observar que para el modelo Poisson, la proporción de veces en las que $\Delta\hat{\beta}\,\% < \delta$ fue exactamente 1, esto sin importar el grado de correlación entre las variables, el tamaño de la muestra y el nivel de referencia δ, indicando con esto que la presencia del fenómeno de la confusión no estaba presente en el modelo cuando realmente lo estaba, por lo tanto, para este modelo y bajo este escenario el criterio no funciona. Sin embargo, cuando $\rho = 0$ osea cuando no hay confusión, el criterio acertó en todas los modelos que se ajustaron.

En el modelo logístico, los criterios $\Delta\hat{\beta}\,\%$ y CP se comportan de manera similar y tienen buen desempeño, principalmente cuando existe algún grado de correlación entre las covariables y los tamaños en las muestras aumentan, pues en estos casos las proporciones son cercanas a cero y por tanto logran detectar la presencia de confusión en el modelo. Sin embargo cuando $\rho = \pm 0.2$ y $\delta \geq 0.10$ el criterio $\Delta\hat{\beta}\,\%$ es más efectivo que el propuesto, puesto que las proporciones son más bajas y por lo tanto percibe con mayor exactitud la presencia del fenómeno en el modelo.

La **Figura 4-3** ilustra el comportamiento de los criterios $\Delta\hat{\beta}\,\%$ y CP cuando en el modelo logístico se simularon covariables con $\rho = 0$. Se puede observar en la **Figura 4-3**(a) que la proporción de veces en las que $\Delta\hat{\beta}\,\% < \delta$ es menor a 0.15, señalando con esto que el fenómeno de la confusión se encuentra presente en aproximadamente el 80 % de los modelos, cuando realmente no lo está. En la **Figura 4-3**(b) las proporciones en las que $CP < \delta$ son acercanas a 1 a medida que n y δ aumentan, esto indica que este criterio es más efectivo que el anterior.

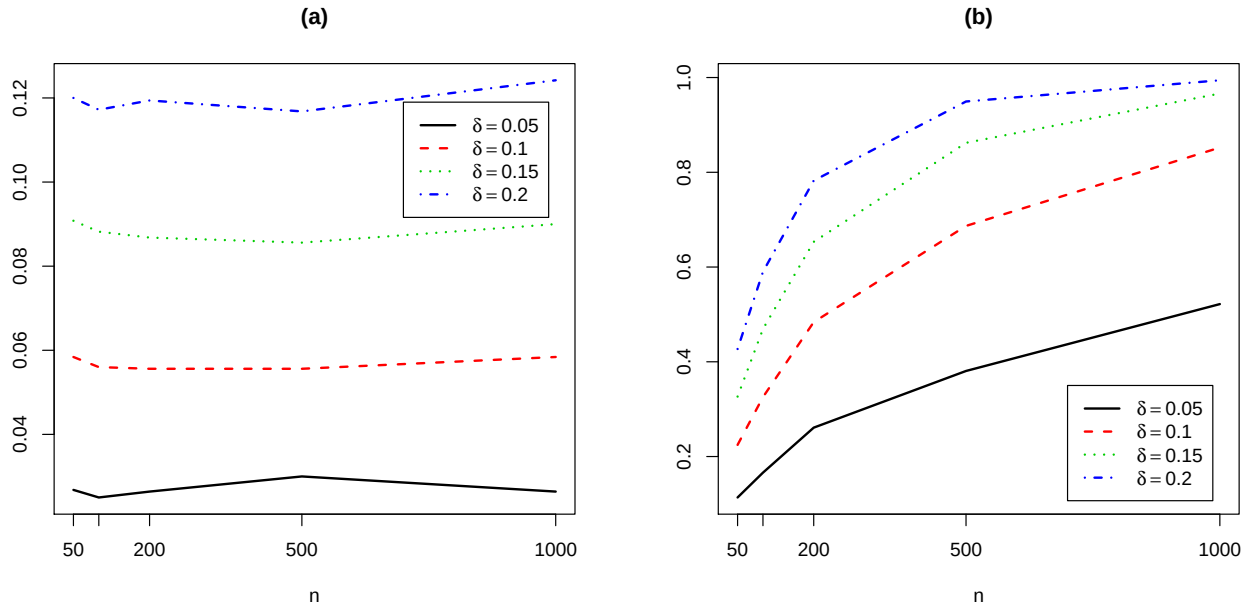

Figura 4-3.: Comportamiento de los criterios (a) $\Delta\hat{\beta}\,\%$ y (b) CP, en el modelo logístico cuando $\rho = 0$.

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0006(0.0002)	0(0)	0(0)	0(0)	0(0)
	−0.6	0.0044(0.002)	0.0006(0.0004)	0(0)	0(0)	0(0)
	−0.4	0.0186(0.0258)	0.0048(0.0076)	0.0006(0.0008)	0(0)	0(0)
	−0.2	0.0256(0.075)	0.024(0.0772)	0.0142(0.0494)	0.0016(0.0092)	0(0.0004)
	0.0	0.0268(0.1136)	0.025(0.166)	0.0264(0.261)	0.03(0.3806)	0.0264(0.5218)
	0.2	0.0274(0.0802)	0.0226(0.0758)	0.0144(0.0482)	0.0014(0.008)	0(0.0002)
	0.4	0.015(0.0208)	0.0054(0.0084)	0.0002(0.0004)	0(0)	0(0)
	0.6	0.0046(0.0042)	0.0002(0.0002)	0(0)	0(0)	0(0)
	0.8	0.0012(0.0008)	0(0)	0(0)	0(0)	0(0)
Poisson	−0.8	1	1	1	1	1
	−0.6	1	1	1	1	1
	−0.4	1	1	1	1	1
	−0.2	1	1	1	1	1
	0.0	1	1	1	1	1
	0.2	1	1	1	1	1
	0.4	1	1	1	1	1
	0.6	1	1	1	1	1
	0.8	1	1	1	1	1

Tabla 4-1.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.003(0.0006)	0(0)	0(0)	0(0)	0(0)
	−0.6	0.0088(0.0064)	0.001(0.0006)	0(0)	0(0)	0(0)
	−0.4	0.0322(0.048)	0.0096(0.016)	0.001(0.002)	0(0)	0(0)
	−0.2	0.054(0.1584)	0.0456(0.1658)	0.0272(0.1186)	0.0036(0.042)	0.0002(0.0066)
	0.0	0.0584(0.2246)	0.056(0.3254)	0.0556(0.4838)	0.0556(0.6868)	0.0584(0.8522)
	0.2	0.0514(0.152)	0.0452(0.1438)	0.03(0.1062)	0.0028(0.0304)	0(0.0052)
	0.4	0.0306(0.0404)	0.0108(0.0164)	0.0012(0.002)	0(0)	0(0)
	0.6	0.0092(0.0072)	0.0006(0.0002)	0(0)	0(0)	0(0)
	0.8	0.0028(0.001)	0(0)	0(0)	0(0)	0(0)
Poisson	−0.8	1	1	1	1	1
	−0.6	1	1	1	1	1
	−0.4	1	1	1	1	1
	−0.2	1	1	1	1	1
	0.0	1	1	1	1	1
	0.2	1	1	1	1	1
	0.4	1	1	1	1	1
	0.6	1	1	1	1	1
	0.8	1	1	1	1	1

Tabla 4-2.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0042(0.0006)	0(0)	0(0)	0(0)	0(0)
	−0.6	0.0136(0.0116)	0.0014(0.0008)	0(0)	0(0)	0(0)
	−0.4	0.0488(0.0778)	0.0138(0.0304)	0.0012(0.0052)	0(0)	0(0)
	−0.2	0.087(0.2384)	0.066(0.2616)	0.0412(0.2154)	0.007(0.1166)	0.0004(0.0452)
	0.0	0.0908(0.3266)	0.0882(0.4692)	0.0868(0.6532)	0.0856(0.8622)	0.09(0.9662)
	0.2	0.0758(0.2214)	0.0696(0.2286)	0.0442(0.1774)	0.0056(0.0792)	0.0002(0.0222)
	0.4	0.0452(0.0642)	0.018(0.0272)	0.0016(0.004)	0(0)	0(0)
	0.6	0.0158(0.012)	0.0008(0.0008)	0(0)	0(0)	0(0)
	0.8	0.0044(0.0014)	0(0)	0(0)	0(0)	0(0)
Poisson	−0.8	1	1	1	1	1
	−0.6	1	1	1	1	1
	−0.4	1	1	1	1	1
	−0.2	1	1	1	1	1
	0.0	1	1	1	1	1
	0.2	1	1	1	1	1
	0.4	1	1	1	1	1
	0.6	1	1	1	1	1
	0.8	1	1	1	1	1

Tabla 4-3.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0066(0.0018)	0(0)	0(0)	0(0)	0(0)
	−0.6	0.0184(0.0178)	0.0018(0.002)	0(0)	0(0)	0(0)
	−0.4	0.0656(0.1078)	0.0206(0.0526)	0.0028(0.0106)	0(0)	0(0)
	−0.2	0.113(0.327)	0.0866(0.3636)	0.0566(0.346)	0.0104(0.2848)	0.0004(0.2006)
	0.0	0.12(0.427)	0.1172(0.59)	0.1194(0.7824)	0.1168(0.9494)	0.1242(0.9942)
	0.2	0.1038(0.2828)	0.0906(0.3084)	0.056(0.2624)	0.0082(0.1666)	0.0004(0.0846)
	0.4	0.0602(0.0902)	0.0246(0.039)	0.003(0.0068)	0(0)	0(0)
	0.6	0.0214(0.0152)	0.0018(0.001)	0(0)	0(0)	0(0)
	0.8	0.007(0.0022)	0(0)	0(0)	0(0)	0(0)
Poisson	−0.8	1	1	1	1	1
	−0.6	1	1	1	1	1
	−0.4	1	1	1	1	1
	−0.2	1	1	1	1	1
	0.0	1	1	1	1	1
	0.2	1	1	1	1	1
	0.4	1	1	1	1	1
	0.6	1	1	1	1	1
	0.8	1	1	1	1	1

Tabla 4-4.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$

Escenario 2

En las Tablas de la **4-5** a la **4-8** se presentan los resultados de la proporción de veces en que $\Delta\hat{\beta}\,\% < \delta$ para los datos simulados de un modelo de regresión Poisson con parámetro λ entre 10 y 30, y un modelo logístico donde $0.15 < \pi < 0.8$, como se ve en las Figuras **4-2**(b) y **4-1**, respectivamente.

En estas tablas se puede ver que el desempeño de $\Delta\hat{\beta}\,\%$ en comparación con el escenario anterior mejora notablemente en el modelo Poisson, puesto que acierta en aproximadamente el 90 % de los casos simulados con confusión cuando $n \geq 500$; sin embargo llama particularmente la atención (1) que para $\rho = \pm 0.8$ las proporciones se encuentran entre 0.15 y 0.35 y (2) el desempeño disminuye cuando $\rho = 0$. Para ver esto con mayor claridad se muestra como caso particular en la **Figura 4-4**(a) cuando $\delta = 0.05$ y $p = -0.8, -0.6, -0.4, -0.2, 0$.

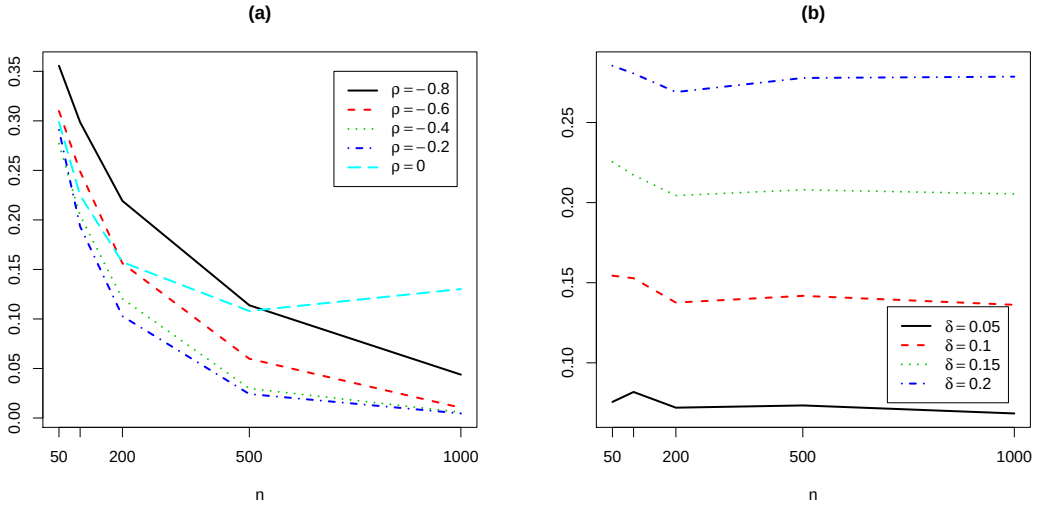

Figura 4-4.: Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en (a) modelo Poisson cuando $\delta = 0.05$ y $p = -0.8, -0.6, -0.4, -0.2, 0$, (b) modelo logístico cuando $\rho = 0$.

En la **Figura 4-4**(b), se presenta la proporción de veces en las que $\Delta\hat{\beta}\,\% < \delta$ cuando se trabaja con un modelo logístico sin confusión ($\rho = 0$). En esta gráfica se nota que a medida de que aumentan δ las proporciones aumentan, decrece cuando $50 \leq n \leq 100$ y se mantine constante cuando $n \geq 200$; sin embargo las proporciones se encuentra por debajo de 0.30. En forma general, el comportamiento del criterio bajo este modelo y bajo este escenario es similar al visto el escenario anterior; esto se deba tal vez al hecho de que la probabilidad

de la respuesta no difiera mucho a la del caso anterior. Particularmente en la **Tabla 4-6** cuando $\delta \geq 0.10$ y $n \geq 200$, se encuentra que $\Delta\hat{\beta}\,\%$ determina correctamente la confusión en todas los modelos simulados bajo este fenómeno.

En cuanto al criterio propuesto, se observa que el comportamiento es similar al $\Delta\hat{\beta}\,\%$ y nuevamente las diferencias entre estos se encuentra en $\rho = 0$ y $\rho = \pm 0.2$. Cuando $\rho = 0$ las proporciones con CP están entre $(0.30, 1)$, mientras que con $\Delta\hat{\beta}\,\%$ se tienen proporciomes entre $(0.02, 0.29)$, lo que hace a CP más efectivo en este caso. Sin embargo, cuando $\rho = \pm 0.2$ $\Delta\hat{\beta}\,\%$ es más eficiente,puesto que las proporciones son inferiores a 0.2 mientras que con CP llegan a superar 0.9 (**Tabla 4-8**).

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0122(0.0162)	0.0024(0.0026)	0(0.0002)	0(0)	0(0)
	−0.6	0.0036(0.0076)	0.0002(0.0002)	0(0)	0(0)	0(0)
	−0.4	0.0078(0.018)	0.0008(0.0022)	0(0)	0(0)	0(0)
	−0.2	0.0464(0.155)	0.0188(0.0964)	0.0046(0.0404)	0(0.0022)	0(0)
	0.0	0.0756(0.3008)	0.0818(0.4242)	0.072(0.5438)	0.0734(0.7686)	0.0684(0.9134)
	0.2	0.045(0.1466)	0.0204(0.1034)	0.003(0.0328)	0(0.0014)	0(0.0002)
	0.4	0.008(0.019)	0.0004(0.0008)	0(0)	0(0)	0(0)
	0.6	0.0062(0.0076)	0.0006(0.0006)	0(0)	0(0)	0(0)
	0.8	0.0154(0.0164)	0.003(0.0034)	0(0)	0(0)	0(0)
Poisson	−0.8	0.3556	0.2986	0.2192	0.1138	0.0438
	−0.6	0.31	0.249	0.1562	0.0598	0.0104
	−0.4	0.277	0.2044	0.1204	0.0298	0.0058
	−0.2	0.2908	0.1934	0.1028	0.0242	0.0046
	0	0.2988	0.2248	0.1574	0.108	0.1302
	0.2	0.2866	0.2108	0.1088	0.0256	0.0028
	0.4	0.2876	0.2122	0.1232	0.034	0.0036
	0.6	0.3046	0.2354	0.164	0.0486	0.0138
	0.8	0.358	0.306	0.2262	0.1142	0.0418

Tabla 4-5.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0272(0.0308)	0.0056(0.008)	0.0002(0.0004)	0(0)	0(0)
	−0.6	0.0086(0.02)	0.0004(0.0014)	0(0)	0(0)	0(0)
	−0.4	0.015(0.0548)	0.0012(0.0098)	0(0.0002)	0(0)	0(0)
	−0.2	0.0886(0.318)	0.0362(0.2718)	0.0098(0.189)	0(0.0752)	0(0.0226)
	0.0	0.1544(0.5322)	0.1528(0.7007)	0.1376(0.8536)	0.1418(0.98)	0.1362(0.9996)
	0.2	0.0894(0.2978)	0.0434(0.2362)	0.008(0.1442)	0(0.0388)	0(0.0072)
	0.4	0.0178(0.05)	0.0006(0.0066)	0(0.0002)	0(0)	0(0)
	0.6	0.011(0.018)	0.0006(0.0012)	0(0)	0(0)	0(0)
	0.8	0.0268(0.032)	0.0068(0.0088)	0.0002(0.0002)	0(0)	0(0)
Poisson	−0.8	0.356	0.2986	0.2192	0.1138	0.0438
	−0.6	0.31	0.249	0.1562	0.0598	0.0104
	−0.4	0.2778	0.2044	0.1204	0.0298	0.0058
	−0.2	0.3044	0.2012	0.104	0.0242	0.0046
	0.0	0.336	0.2688	0.2194	0.1904	0.249
	0.2	0.2982	0.2158	0.1098	0.0256	0.0028
	0.4	0.288	0.2122	0.1232	0.034	0.0036
	0.6	0.3046	0.2354	0.164	0.0486	0.0138
	0.8	0.358	0.306	0.2262	0.1142	0.0418

Tabla 4-6.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0386(0.0526)	0.0088(0.0162)	0.0004(0.0004)	0(0)	0(0)
	−0.6	0.0152(0.0358)	0.0008(0.0044)	0(0.0002)	0(0)	0(0)
	−0.4	0.0254(0.135)	0.0028(0.0456)	0(0.0066)	0(0)	0(0)
	−0.2	0.1294(0.4846)	0.06(0.4958)	0.0162(0.4796)	0(0.4642)	0(0.452)
	0.0	0.2254(0.6996)	0.2172(0.8668)	0.2044(0.9644)	0.208(0.9986)	0.2054(1)
	0.2	0.131(0.4372)	0.0636(0.4076)	0.0134(0.3578)	0.0002(0.2632)	0(0.192)
	0.4	0.0282(0.1058)	0.0014(0.0244)	0(0.0024)	0(0)	0(0)
	0.6	0.0172(0.0304)	0.0014(0.0028)	0(0)	0(0)	0(0)
	0.8	0.0412(0.0486)	0.0112(0.0144)	0.0004(0.0014)	0(0)	0(0)
Poisson	−0.8	0.3562	0.2986	0.2192	0.1138	0.0438
	−0.6	0.31	0.249	0.1562	0.0598	0.0104
	−0.4	0.2779	0.2044	0.1204	0.0298	0.0058
	−0.2	0.3176	0.209	0.1056	0.0242	0.0046
	0.0	0.3706	0.3118	0.2716	0.275	0.3556
	0.2	0.3108	0.2206	0.112	0.0256	0.0028
	0.4	0.2884	0.2122	0.1232	0.034	0.0036
	0.6	0.3046	0.2354	0.164	0.0486	0.0138
	0.8	0.358	0.306	0.2262	0.1142	0.0418

Tabla 4-7.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0522(0.0794)	0.0144(0.027)	0.0004(0.0026)	0(0)	0(0)
	−0.6	0.022(0.0692)	0.0016(0.0194)	0(0.0014)	0(0)	0(0)
	−0.4	0.04(0.2458)	0.005(0.0514)	0(0.0606)	0(0.0078)	0(0)
	−0.2	0.1702(0.6418)	0.0792(0.703)	0.0252(0.7788)	0.0006(0.8944)	0(0.9606)
	0.0	0.2854(0.8092)	0.2806(0.9462)	0.269(0.9924)	0.2778(1)	0.2786(1)
	0.2	0.1714(0.5594)	0.085(0.5748)	0.0214(0.5946)	0.0002(0.6424)	0(0.6988)
	0.4	0.0452(0.1798)	0.0036(0.0626)	0(0.017)	0(0.0006)	0(0)
	0.6	0.0242(0.046)	0.0018(0.0072)	0(0.0008)	0(0)	0(0)
	0.8	0.0516(0.0642)	0.0146(0.0206)	0.0008(0.0014)	0(0)	0(0)
Poisson	−0.8	0.3566	0.2986	0.2192	0.1138	0.0438
	−0.6	0.31	0.249	0.1562	0.0598	0.0104
	−0.4	0.2798	0.2044	0.1204	0.0298	0.0058
	−0.2	0.3334	0.2174	0.1086	0.0242	0.0046
	0.0	0.4048	0.3542	0.3264	0.3516	0.4508
	0.2	0.3258	0.2288	0.1156	0.0256	0.0028
	0.4	0.2892	0.2122	0.1232	0.034	0.0036
	0.6	0.3046	0.2354	0.164	0.0486	0.0138
	0.8	0.358	0.306	0.2262	0.1142	0.0418

Tabla 4-8.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$

Escenario 3

A continuación se presentan los resultados obtenidos bajo este escenario, en el cuál se simularon respuestas de una distribución binomial con probabilidades moderadas, y respuestas de una distribución Poisson con media $\lambda \in (0, 20)$.

En la Tablas de la **4-9** a **4-12** se observa que en el modelo Poisson, la proporción de veces en las que $\Delta\hat{\beta}\,\% < \delta$ disminuye a medida de aumenta n y que el grado de correlación entre las variables es más fuerte; esto tal vez se deba a que el modelo se simula con una varianza menor a la que se consideró en los dos escenarios anteriores.

La **Figura 4-5** muestra el comportamiento del criterio del cambio porcentual propuesto por Hosmer & Lemeshow (1999) en ambos modelos cuando $\delta = 0.20$ y correlaciones positivas entre las covariables. Note que en estos casos y para $n \geq 100$ el criterio es bastante eficiente en los dos modelos, sin embargo, en el modelo logístico se encuentra que cuando $\rho = 0$ el criterio no tiene buen desempeño, ya que las proporciones son inferiores a 0.5.

En relación al criterio propuesto se observa que es sensible cuando existe algún grado de

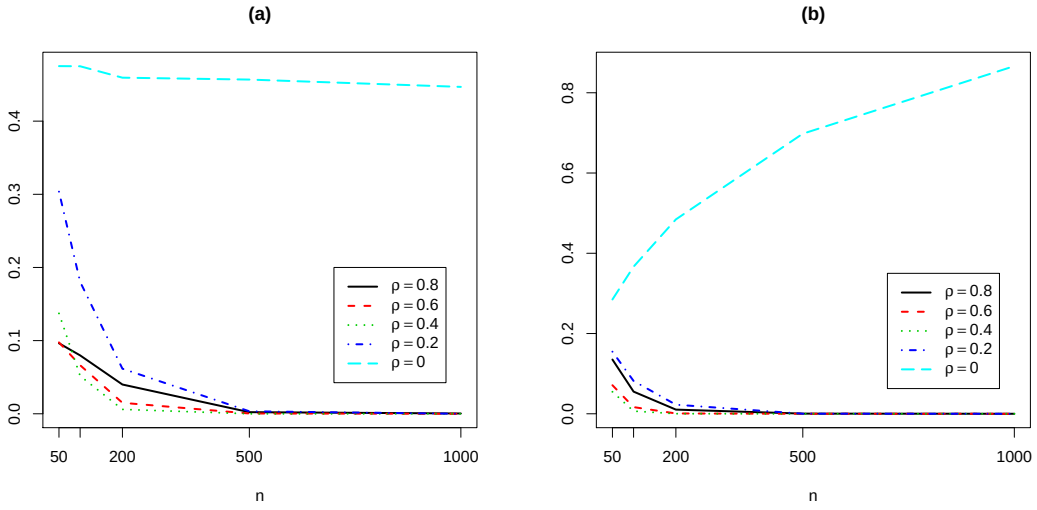

Figura 4-5.: Comportamiento del criterio $\Delta\hat{\beta}\,\%$ cuando $\delta = 0.20$ y $p = 0.8, 0.6, 0.4, 0.2, 0$ en (a) Modelo logístico, (b) Modelo Poisson.

correlación entre las covariables y cuando δ aumenta, puesto que la proporción de veces en las que fue menor a δ es bastante alta, indicando que el fenómeno se encuentra en el modelo cuando realmente no está. A manera de ilustración, oberve que cuando $\rho = \pm 0.2$, sin importar el tamaño de la muestra el criterio no detecta la confusión en más de la mitad de los modelos que fueron realmente simulados con ese fenómeno.

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0254(0.0612)	0.021(0.0504)	0.0074(0.0248)	0.0002(0.0014)	0(0)
	−0.6	0.0238(0.066)	0.0126(0.0464)	0.0018(0.012)	0(0)	0(0)
	−0.4	0.033(0.1248)	0.0108(0.0598)	0.0006(0.0102)	0(0)	0(0)
	−0.2	0.0936(0.3216)	0.0438(0.261)	0.007(0.1606)	0(0.0466)	0(0.0072)
	0.0	0.1852(0.551)	0.1606(0.6896)	0.1422(0.8354)	0.1328(0.9728)	0.1326(0.9978)
	0.2	0.1036(0.328)	0.0428(0.249)	0.0046(0.1426)	0(0.0304)	0(0.002)
	0.4	0.0342(0.1126)	0.0088(0.0522)	0.0008(0.0068)	0(0.0002)	0(0)
	0.6	0.0266(0.069)	0.0102(0.0362)	0.0024(0.0108)	0(0.0002)	0(0)
	0.8	0.026(0.058)	0.0178(0.049)	0.0082(0.026)	0.0004(0.0012)	0(0.0004)
Poisson	−0.8	0.1384	0.0594	0.0128	0	0
	−0.6	0.0718	0.0182	0.0018	0	0
	−0.4	0.0478	0.01	0.0006	0	0
	−0.2	0.0592	0.0192	0.0036	0	0
	0.0	0.0994	0.1012	0.1314	0.2078	0.2894
	0.2	0.0624	0.024	0.0048	0	0
	0.4	0.046	0.0066	0.0002	0	0
	0.6	0.0712	0.0166	0.0008	0	0
	0.8	0.1346	0.0552	0.0104	0.0004	0

Tabla 4-9.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.054(0.1258)	0.0428(0.1052)	0.0176(0.0566)	0.0006(0.007)	0(0)
	−0.6	0.046(0.1462)	0.0292(0.114)	0.0068(0.039)	0(0.0028)	0(0)
	−0.4	0.0648(0.2762)	0.0246(0.1918)	0.0022(0.0898)	0(0.019)	0(0.002)
	−0.2	0.1752(0.5678)	0.082(0.6024	0.0164(0.6142)	0(0.664)	0(0.7726)
	0.0	0.3048(0.7902)	0.291(0.9144)	0.2716(0.9828)	0.2576(0.9998)	0.2546(1)
	0.2	0.181(0.5454)	0.089(0.5458)	0.0146(0.5414)	0(0.5226)	0(0.5224)
	0.4	0.068(0.2352)	0.022(0.1562)	0.0016(0.0666)	0(0.0066)	0(0.0002)
	0.6	0.0502(0.139)	0.023(0.084)	0.0048(0.0346)	0(0.0022)	0(0.0002)
	0.8	0.0492(0.1108)	0.0386(0.0994)	0.0196(0.0546)	0.0006(0.006)	0.0002(0.0006)
Poisson	−0.8	0.1384	0.0594	0.0128	0	0
	−0.6	0.0718	0.0182	0.0018	0	0
	−0.4	0.0488	0.0102	0.0006	0	0
	−0.2	0.0832	0.0356	0.0084	0	0
	0.0	0.168	0.191	0.2478	0.402	0.5494
	0.2	0.093	0.043	0.0096	0	0
	0.4	0.0486	0.0066	0.0002	0	0
	0.6	0.0712	0.0166	0.0008	0	0
	0.8	0.1348	0.0552	0.0104	0.0004	0

Tabla 4-10.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0796(0.1794)	0.0632(0.1698)	0.027(0.1002)	0.0012(0.0264)	0(0.0036)
	−0.6	0.0688(0.2462)	0.0436(0.2074)	0.0112(0.1248)	0(0.0294)	0(0.0038)
	−0.4	0.1038(0.4388)	0.042(0.4132)	0.0044(0.3476)	0(0.2546)	0(0.1742)
	−0.2	0.2436(0.7468)	0.1248(0.833)	0.032(0.9092)	0.0024(0.982)	0(0.9984)
	0.0	0.4028(0.8938)	0.3964(0.9756)	0.3768(0.998)	0.3632(1)	0.3584(1)
	0.2	0.2452(0.6952)	0.13(0.7642)	0.0358(0.8274)	0.0008(0.927)	0(0.9804)
	0.4	0.1034(0.3678)	0.0348(0.2996)	0.0034(0.2254)	0(0.1064)	0(0.039)
	0.6	0.0752(0.2156)	0.0346(0.1544)	0.0106(0.0836)	0(0.0138)	0(0.0006)
	0.8	0.0746(0.1676)	0.0612(0.1462)	0.0286(0.0856)	0.0014(0.0172)	0.0002(0.0028)
Poisson	−0.8	0.1384	0.0594	0.0128	0	0
	−0.6	0.0718	0.0182	0.0018	0	0
	−0.4	0.0502	0.0102	0.0006	0	0
	−0.2	0.1094	0.0526	0.015	0	0
	0.0	0.225	0.2814	0.3704	0.5718	0.7366
	0.2	0.123	0.0602	0.0144	0.0002	0
	0.4	0.051	0.0066	0.0002	0	0
	0.6	0.0712	0.0166	0.0008	0	0
	0.8	0.135	0.0552	0.0104	0.0004	0

Tabla 4-11.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.1028(0.2462)	0.0792(0.2424)	0.036(0.1788)	0.002(0.074)	0(0.0238)
	−0.6	0.0932(0.3552)	0.0578(0.3408)	0.0132(0.2724)	0(0.1694)	0(0.081)
	−0.4	0.5926(0.5926)	0.0562(0.6502)	0.0078(0.679)	0(0.7736)	0(0.854)
	−0.2	0.2996(0.8646)	0.174(0.9448)	0.0574(0.9894)	0(1)	0(1)
	0.0	0.4752(0.994)	0.475(0.9932)	0.4594(1)	0.4568(1)	0.4468(1)
	0.2	0.3036(0.7948)	0.1804(0.8776)	0.0616(0.9528)	0.0034(0.9952)	0(1)
	0.4	0.1372(0.4936)	0.0518(0.4786)	0.006(0.456)	0(0.431)	0(0.3988)
	0.6	0.0976(0.2934)	0.0664(0.238)	0.015(0.159)	0.0004(0.063)	0(0.0126)
	0.8	0.0964(0.2234)	0.0798(0.1974)	0.04(0.1228)	0.0022(0.0418)	0.0004(0.0084)
Poisson	−0.8	0.1384	0.0594	0.0128	0	0
	−0.6	0.0718	0.0182	0.0018	0	0
	−0.4	0.053	0.0102	0.0006	0	0
	−0.2	0.14	0.0706	0.0234	0.0008	0
	0.0	0.2848	0.3666	0.4844	0.698	0.8664
	0.2	0.1548	0.0818	0.0226	0.0002	0
	0.4	0.0548	0.0066	0.0002	0	0
	0.6	0.0712	0.0166	0.0008	0	0
	0.8	0.1352	0.0552	0.0104	0.0004	0

Tabla 4-12.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$

Escenario 4

A continuación se presentan las proporciones en las que $\Delta\hat{\beta}\,\% < \delta$, cuando se simularon los modelos logístico y Poisson con respuestas bajas.

La Tabla **4-13** muestra los resultados de la simulación cuando el nivel de referencia $\delta = 0.05$. Aquí se puede ver que, en el modelo logístico las proporciones se encuentran en la mayoría de los casos por debajo de 0.05, con esto se asume que $\Delta\hat{\beta}\,\%$ tiene un buen desempeño en este modelo, bajo este escenario. En el modelo Poisson por su parte, las proporciones son superiores a 0.10 para $n \leq 200$.

En las Tablas **4-14**, **4-15** y **4-16** se nota que el criterio es efectivo en el modelo Poisson cuando $\rho = 0$, pero cuándo la confusión está presente en el modelo el criterio tiende a ser menos efectivo. Este mismo comportamiento se presenta en el modelo logístico cuando se emplea CP, ya que no es efectivo en $\rho = \pm 0.2$.

Sin importar el nivel de referencia, se observa que cuando no existe correlación entre las covariables la proporción de veces en las que $\Delta\hat{\beta}\,\% < \delta$ es inferior a 0.5, asi que en este caso ($\rho = 0$) el criterio no tiene un buen desempeño en el modelo logístico; esto se puede ver claramente en la **Figura 4-6**(a), mientras que en la **Figura 4-6**(b) se observa que con el nuevo criterio las proporciones son cercanas a uno, principalmente cuando $n \geq 100$ y $\delta \geq 0.10$.

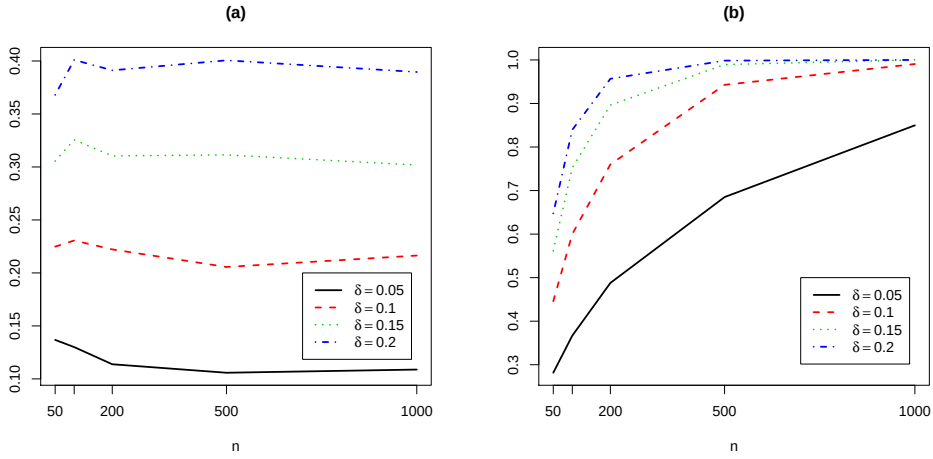

Figura 4-6.: Comportamiento de los criterios (a) $\Delta\hat{\beta}\,\%$ y (b) CP, en el modelo logistico cuando $\rho = 0$.

En la **Figura 4-7** se muestra el comportamiento del criterio de cambio porcentual en el modelo Poisson cuando $\rho = -0.8, -0.6, -0.4, -0.2, 0$, aqui se ve claramente que este criterio es sensible en este modelo a medida de que el grado de correlación entre las covariables tiende a ser bajo. Cuando $\rho = 0$ y $\rho = -0.8$ (**Figura 4-7**(a),(e)) el criterio tiene buen desempeño, principalmente cuando $\delta = 0.05$ y $\delta = 0.20$, respectivamente. Por tanto, en caso de trabajar con un modelo Poisson donde $\lambda \in (2.5, 4.5)$, se recomienda utilizar $\delta = 0.05$, pero sólo cuando $n \geq 500$, en caso contrario no se recomienda utilizar $\Delta\hat{\beta}\,\%$.

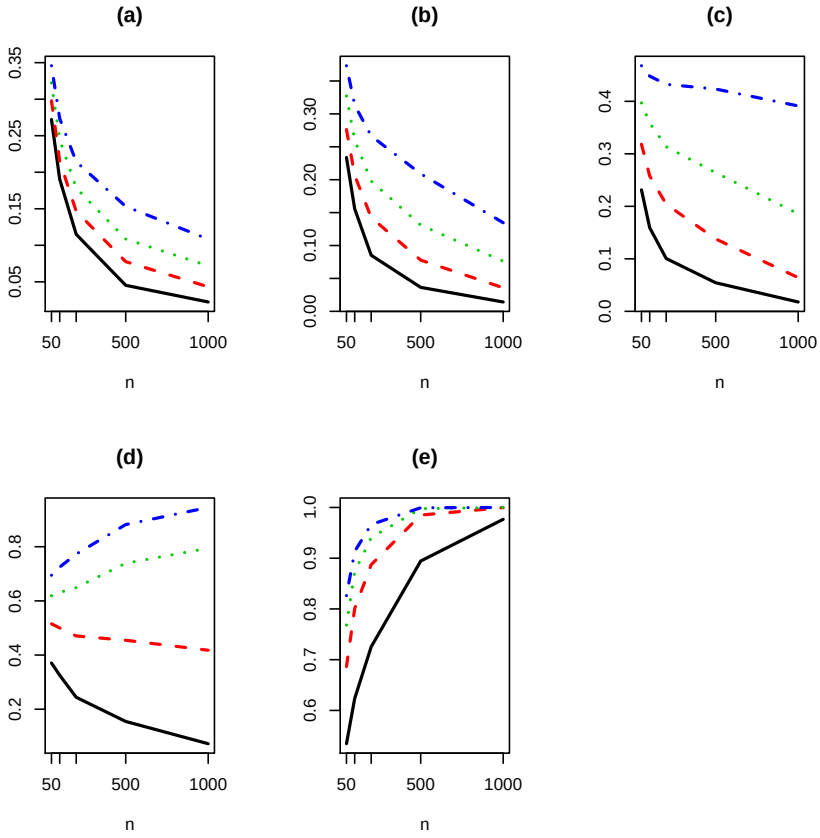

Figura 4-7.: Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo Poisson cuando (a) $\rho = -0.8$, (b) $\rho = -0.6$, (c) $\rho = -0.4$, (d) $\rho = -0.2$, (e) $\rho = 0$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.025(0.042)	0.0216(0.0222)	0.0108(0.0124)	0.0004(0.0006)	0(0)
	−0.6	0.027(0.0558)	0.0108(0.0216)	0.0028(0.005)	0.0002(0.0002)	0(0)
	−0.4	0.0472(0.086)	0.012(0.0236)	0.0012(0.0038)	0(0)	0(0)
	−0.2	0.1008(0.1998)	0.0538(0.1382)	0.0148(0.0624)	0.0008(0.0042)	0(0.0002)
	0.0	0.1368(0.2816)	0.13(0.3668)	0.1138(0.4878)	0.1058(0.685)	0.1088(0.8498)
	0.2	0.1034(0.2028)	0.0598(0.1372)	0.0168(0.0572)	0(0.0038)	0(0)
	0.4	0.0436(0.078)	0.0156(0.0294)	0.001(0.0028)	0(0)	0(0)
	0.6	0.028(0.0546)	0.0134(0.0192)	0.0026(0.005)	0(0)	0(0)
	0.8	0.0264(0.0432)	0.0188(0.0248)	0.0106(0.0128)	0.0004(0.0006)	0(0)
Poisson	−0.8	0.2722	0.1904	0.115	0.0452	0.0224
	−0.6	0.2338	0.1556	0.0852	0.0364	0.0142
	−0.4	0.2314	0.1592	0.1006	0.0544	0.0178
	−0.2	0.3706	0.3254	0.244	0.1548	0.0726
	0.0	0.5344	0.6236	0.7254	0.8944	0.9766
	0.2	0.4058	0.3426	0.2824	0.1732	0.088
	0.4	0.2754	0.2052	0.149	0.079	0.0282
	0.6	0.251	0.1806	0.1274	0.0756	0.0266
	0.8	0.2914	0.2302	0.162	0.0888	0.0496

Tabla 4-13.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0504(0.069)	0.0366(0.0436)	0.0202(0.0254)	0.001(0.0012)	0(0)
	−0.6	0.0548(0.095)	0.0288(0.0456)	0.0058(0.011)	0.0002(0.0002)	0(0)
	−0.4	0.0806(0.1514)	0.03(0.0726)	0.0032(0.0176)	0(0)	0(0)
	−0.2	0.1678(0.3298)	0.1026(0.3038)	0.0324(0.2162)	0.0014(0.0758)	0(0.0174)
	0.0	0.2248(0.4458)	0.2306(0.6002)	0.2222(0.76)	0.2056(0.9428)	0.2164(0.9906)
	0.2	0.171(0.328)	0.1018(0.272)	0.037(0.1766)	0.0008(0.0558)	0(0.0084)
	0.4	0.081(0.1454)	0.0338(0.061)	0.0026(0.0084)	0(0)	0(0)
	0.6	0.0578(0.0912)	0.029(0.0448)	0.006(0.0094)	0(0)	0(0)
	0.8	0.0508(0.072)	0.038(0.0458)	0.0218(0.0276)	0.0012(0.0012)	0(0)
Poisson	−0.8	0.2978	0.2152	0.1466	0.0776	0.0432
	−0.6	0.2762	0.2066	0.1426	0.0778	0.0358
	−0.4	0.3182	0.2578	0.204	0.1378	0.0642
	−0.2	0.5154	0.5	0.4706	0.4546	0.4178
	0.0	0.6864	0.8008	0.8872	0.985	0.9996
	0.2	0.561	0.5254	0.5	0.4578	0.436
	0.4	0.3878	0.3292	0.2696	0.1714	0.0996
	0.6	0.3184	0.2758	0.2204	0.1472	0.074
	0.8	0.336	0.2866	0.233	0.1586	0.0992

Tabla 4-14.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0772(0.0942)	0.0516(0.0672)	0.0284(0.0392)	0.0012(0.0036)	0(0.0004)
	−0.6	0.0822(0.1338)	0.0422(0.074)	0.009(0.023)	0.0002(0.0004)	0(0)
	−0.4	0.1116(0.2252)	0.0476(0.1374)	0.0068(0.0496)	0(0.0028)	0(0)
	−0.2	0.2228(0.4518)	0.156(0.4796)	0.0584(0.4538)	0.0024(0.3668)	0(0.2996)
	0.0	0.3056(0.5618)	0.3256(0.7524)	0.3104(0.8962)	0.3114(0.9894)	0.3018(1)
	0.2	0.2292(0.4306)	0.1492(0.4156)	0.0556(0.3522)	0.0026(0.2324)	0(0.1306)
	0.4	0.1168(0.2042)	0.096(0.1108)	0.006(0.0286)	0(0.0012)	0(0)
	0.6	0.087(0.1242)	0.0444(0.0678)	0.0092(0.0168)	0.0002(0.0002)	0(0)
	0.8	0.074(0.0988)	0.0566(0.0666)	0.0332(0.0392)	0.0014(0.0024)	0(0)
Poisson	−0.8	0.3224	0.245	0.1784	0.1086	0.0718
	−0.6	0.3272	0.2602	0.1978	0.131	0.0762
	−0.4	0.397	0.3586	0.3134	0.2642	0.1858
	−0.2	0.6192	0.6312	0.649	0.739	0.7944
	0.0	0.7682	0.8734	0.9406	0.9976	1
	0.2	0.648	0.6438	0.6566	0.6934	0.7732
	0.4	0.4816	0.4402	0.3904	0.3024	0.2468
	0.6	0.3896	0.3582	0.2956	0.2274	0.1446
	0.8	0.3806	0.3486	0.305	0.2202	0.1526

Tabla 4-15.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.1(0.1214)	0.0686(0.0936)	0.0366(0.0586)	0.002(0.008)	0.0002(0.0006)
	−0.6	0.1036(0.179)	0.0542(0.111)	0.011(0.0444)	0.0002(0.0028)	0(0)
	−0.4	0.1522(0.3046)	0.0662(0.2316)	0.0108(0.1316)	0(0.0318)	0(0.0036)
	−0.2	0.2732(0.5496)	0.2008(0.6402)	0.0872(0.6822)	0.0058(0.7338)	0.0004(0.8154)
	0.0	0.368(0.6476)	0.401(0.8396)	0.3912(0.9568)	0.4006(0.9984)	0.3896(1)
	0.2	0.2803(0.5068)	0.1926(0.5324)	0.079(0.5228)	0.0076(0.514)	0(0.4802)
	0.4	0.1446(0.2646)	0.072(0.174)	0.0122(0.0608)	0(0.0068)	0(0.0002)
	0.6	0.1116(0.1586)	0.0578(0.0968)	0.0134(0.0302)	0.0002(0.001)	0(0)
	0.8	0.0962(0.1252)	0.0772(0.094)	0.0412(0.0516)	0.0024(0.0054)	0(0)
Poisson	−0.8	0.3458	0.2734	0.2144	0.1532	0.1084
	−0.6	0.3732	0.3134	0.2664	0.209	0.1346
	−0.4	0.4678	0.448	0.4324	0.4234	0.3912
	−0.2	0.6948	0.7248	0.7728	0.8818	0.945
	0.0	0.8262	0.9134	0.9664	0.9994	1
	0.2	0.712	0.7242	0.7532	0.842	0.9182
	0.4	0.559	0.5184	0.4898	0.4434	0.2468
	0.6	0.4548	0.4252	0.3698	0.3024	0.2388
	0.8	0.43	0.411	0.3698	0.2824	0.2114

Tabla 4-16.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$

Escenario 5

En las Tablas de la **4-17** a la **4-20** se presentan los resultados de la proporción de veces en que $\Delta\hat{\beta}\,\% < \delta$, cuando en el modelo logístico la probabilidad del resultado era baja y la respuesta en el modelo Poisson tomaba valores entre 0 y 2; y además los resultados en que $CP < \delta$ cuando se trabaja con un modelo logístico.

En las Tablas **4-17** y **4-18** para el modelo logístico, se observa que a medida que el grado de correlación entre las covariables es más fuerte, la proporción de veces en que $\Delta\hat{\beta}\,\% < \delta$ disminuye, y que cuando se aumenta el tamaño de la muestra tienden a ser nula. Este comportamiento también se observa en el modelo Poisson cuando $n \geq 500$.

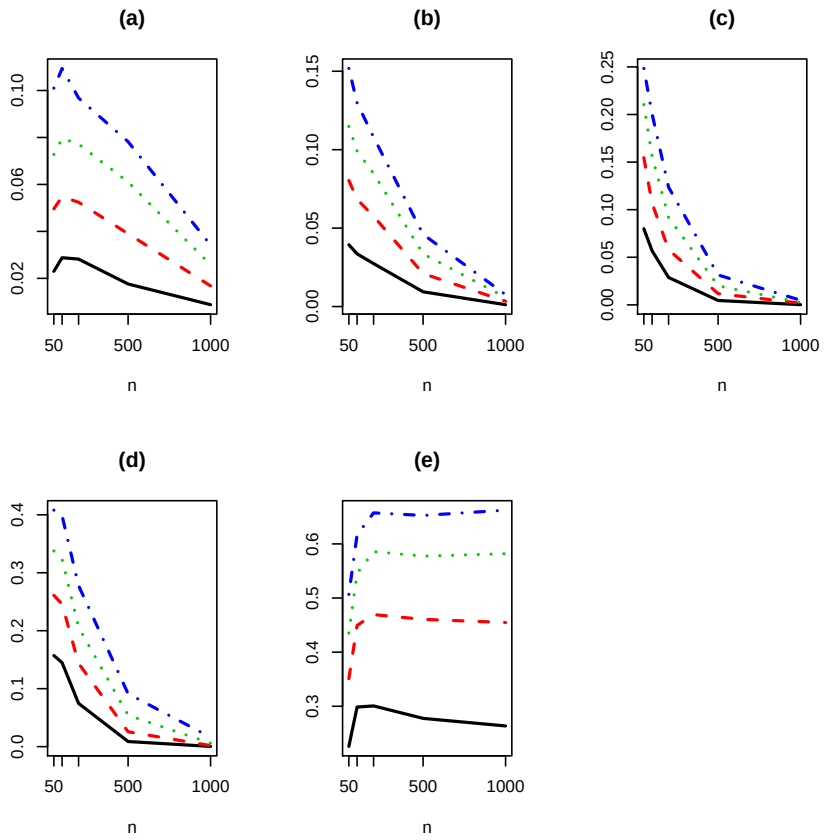

Figura 4-8.: Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo logistico cuando (a)$\rho = 0.8$,(b)$\rho = 0.6$,(c)$\rho = 0.4$,(d)$\rho = 0.2$,(e)$\rho = 0$.

En cuanto a la ausencia de correlación entre las covariables ($\rho = 0$), se encuentran propor-

ciones más altas en el modelo Poisson, pero al igual que en el modelo logístico, éstas son inferiores a 0.5 cuando $\delta = 0.05$ o $\delta = 0.10$ (**Figura 4-8**(e) y **Figura 4-9**(e), respectivamente), excepto en el modelo Poisson cuando $n \geq 500$ y $\delta = 0.10$; es decir en estos casos $\Delta\hat{\beta}\,\%$ indica que la confusión está presente en dicho modelo, en más de la mitad de los modelos simulados sin este fenómeno.

En la **Figura 4-8**, se nota que el criterio $\Delta\hat{\beta}\,\%$ mantiene un buen desempeño en el modelo logístico y que además mejora cuando $\rho = 0$, pues a medida de que se aumenta el tamaño de la muestra y el nivel de referencia δ, las proporciones son cada vez más grandes. Cuando $p = 0.8, 0.6, 0.4$ en el modelo Poisson, el criterio tiene el mismo comportamiento (**Figura 4-9**(a),(b),(c)), mostrando que para $n \geq 500$ este criterio tiene un desempeño en este modelo.

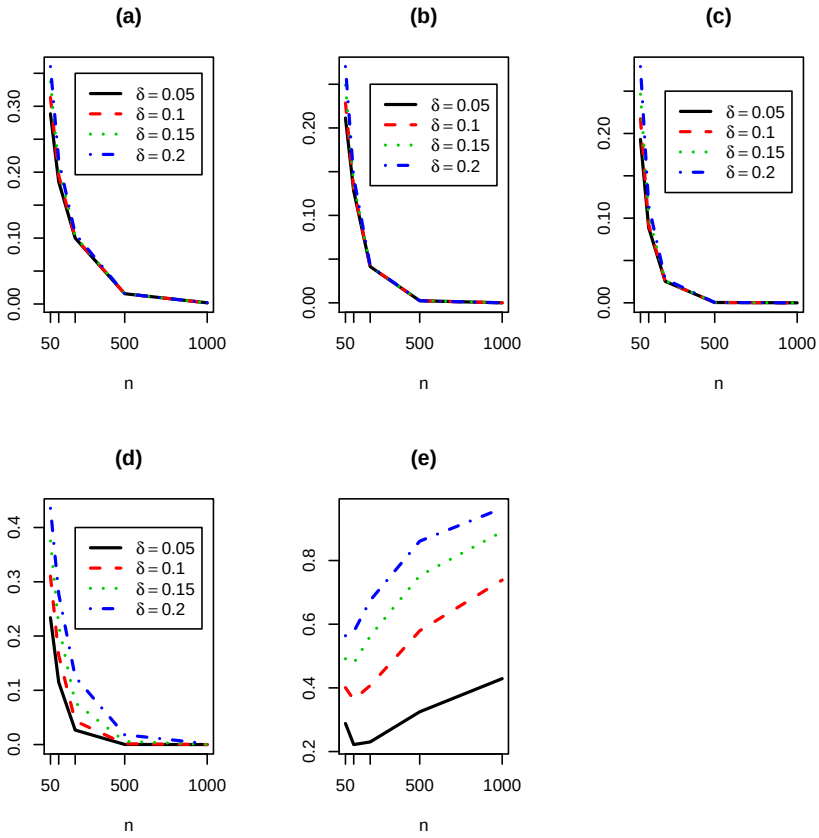

Figura 4-9.: Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo Poisson cuando (a)$\rho = 0.8$,(b)$\rho = 0.6$,(c)$\rho = 0.4$,(d)$\rho = 0.2$,(e)$\rho = 0$.

Los criterios en estudio tienen comportamientos muy similares cuando en el modelo logístico las covariables se simulan con algún grado de correlación, puesto que en la mayoría de los casos la diferencia entre las proporciones tiende a ser nula, salvo en $\rho = -0.2$ donde las proporciones de CP llegan a ser superiores a 0.5, mientras que con el criterio $\Delta\hat{\beta}\,\%$ éstas tienden a cero. Para $\rho = 0$ las proporciones aumentan cuando n y δ aumentan pero con el criterio propuesto se obtienen valores más cercanos a 1.

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0256(0.09)	0.0262(0.0512)	0.0216(0.0562)	0.0148(0.0434)	0.0086(0.0254)
	−0.6	0.0408(0.1134)	0.0378(0.0728)	0.0284(0.076)	0.0126(0.0404)	0.0014(0.0062)
	−0.4	0.0724(0.171)	0.0584(0.1332)	0.0348(0.1072)	0.0092(0.0398)	0(0.0058)
	−0.2	0.1584(0.3194)	0.1398(0.3294)	0.0068(0.273)	0.0122(0.145)	0.002(0.0572)
	0.0	0.2258(0.437)	0.2986(0.6008)	0.3006(0.757)	0.2776(0.9268)	0.2636(0.986)
	0.2	0.1572(0.3166)	0.1448(0.3224)	0.0746(0.2654)	0.0088(0.1282)	0.0004(0.0468)
	0.4	0.08(0.1762)	0.0568(0.1274)	0.0288(0.1016)	0.0046(0.0316)	0.0002(0.0066)
	0.6	0.0394(0.113)	0.0336(0.0724)	0.0274(0.075)	0.0094(0.034)	0.0012(0.0074)
	0.8	0.023(0.0862)	0.0288(0.0562)	0.0282(0.059)	0.0176(0.0526)	0.0088(0.0226)
Poisson	−0.8	0.27	0.1976	0.1086	0.029	0.0042
	−0.6	0.2172	0.122	0.0522	0.0052	0.0002
	−0.4	0.1868	0.0962	0.0266	0.0018	0
	−0.2	0.221	0.1096	0.029	0.0008	0
	0.0	0.2882	0.222	0.2302	0.3248	0.429
	0.2	0.2338	0.1148	0.0266	0.0002	0
	0.4	0.193	0.0886	0.0254	0.0004	0
	0.6	0.2114	0.128	0.0414	0.0024	0
	0.8	0.2886	0.1854	0.1	0.0156	0.0018

Tabla 4-17.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0468(0.1212)	0.0484(0.1008)	0.0446(0.1096)	0.035(0.0914)	0.019(0.0536)
	−0.6	0.0812(0.1746)	0.0718(0.1546)	0.0554(0.156)	0.0236(0.0938)	0.0026(0.0254)
	−0.4	0.1408(0.2814)	0.1078(0.266)	0.0672(0.2336)	0.0186(0.141)	0.0012(0.0636)
	−0.2	0.2552(0.4738)	0.2416(0.5504)	0.1476(0.5864)	0.032(0.5928)	0.0018(0.6082)
	0.0	0.351(0.6054)	0.4494(0.7942)	0.4694(0.933)	0.4606(0.9956)	0.4548(1)
	0.2	0.261(0.4698)	0.2458(0.5274)	0.1436(0.531)	0.0256(0.4924)	0.0018(0.484)
	0.4	0.1546(0.29)	0.1068(0.2454)	0.058(0.2132)	0.0118(0.1164)	0.0018(0.0448)
	0.6	0.0804(0.1726)	0.0684(0.1528)	0.0576(0.1512)	0.021(0.0818)	0.0034(0.0248)
	0.8	0.0496(0.121)	0.0548(0.1052)	0.0524(0.11)	0.039(0.1008)	0.0168(0.0456)
Poisson	−0.8	0.2706	0.1976	0.1086	0.029	0.0042
	−0.6	0.2196	0.1222	0.0522	0.0052	0.0002
	−0.4	0.1974	0.097	0.0266	0.0018	0
	−0.2	0.2748	0.1464	0.0406	0.0012	0
	0.0	0.4008	0.363	0.4074	0.5792	0.7388
	0.2	0.3104	0.1646	0.0434	0.0014	0
	0.4	0.2174	0.0924	0.0256	0.0004	0
	0.6	0.2286	0.1306	0.0416	0.0024	0
	0.8	0.3134	0.1952	0.1006	0.0156	0.0018

Tabla 4-18.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0688(0.157)	0.0748(0.1518)	0.0688(0.172)	0.0502(0.1532)	0.0284(0.0982)
	−0.6	0.1186(0.2404)	0.106(0.2354)	0.0818(0.244)	0.0372(0.1798)	0.005(0.0886)
	−0.4	0.102(0.384)	0.159(0.4002)	0.1006(0.4016)	0.0264(0.3456)	0.0026(0.2824)
	−0.2	0.333(0.5962)	0.3214(0.707)	0.212(0.8028)	0.058(0.9008)	0.0056(0.9656)
	0.0	0.4354(0.7004)	0.546(0.8792)	0.5858(0.9782)	0.5774(1)	0.582(1)
	0.2	0.3376(0.5774)	0.3234(0.6704)	0.2058(0.731)	0.0542(0.8144)	0.006(0.8908)
	0.4	0.2104(0.3844)	0.1562(0.3594)	0.0918(0.3426)	0.02(0.2494)	0.0024(0.1626)
	0.6	0.1148(0.2282)	0.0992(0.223)	0.0848(0.22)	0.0336(0.1442)	0.0058(0.0598)
	0.8	0.0728(0.1582)	0.0796(0.1548)	0.0772(0.1644)	0.061(0.1482)	0.0258(0.0786)
Poisson	−0.8	0.2722	0.1978	0.1086	0.029	0.0042
	−0.6	0.2216	0.1226	0.0522	0.0052	0.0002
	−0.4	0.2096	0.0986	0.0266	0.0018	0
	−0.2	0.3258	0.1918	0.0644	0.0052	0
	0.0	0.4914	0.4772	0.5624	0.7524	0.8904
	0.2	0.3752	0.2186	0.0772	0.0058	0.0002
	0.4	0.2464	0.1024	0.026	0.0004	0
	0.6	0.249	0.1358	0.0418	0.0024	0
	0.8	0.337	0.2066	0.103	0.0156	0.0018

Tabla 4-19.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$

Modelo	ρ	n				
		50	100	200	500	1000
logístico	−0.8	0.0906(0.194)	0.1002(0.203)	0.0942(0.2474)	0.0676(0.229)	0.0358(0.1672)
	−0.6	0.1548(0.303)	0.1408(0.3244)	0.1074(0.3474)	0.0438(0.3086)	0.0072(0.234)
	−0.4	0.2558(0.4794)	0.203(0.541)	0.126(0.576)	0.038(0.6266)	0.0044(0.6714)
	−0.2	0.3958(0.6814)	0.3954(0.8186)	0.2718(0.9162)	0.0918(0.9868)	0.0142(0.999)
	0.0	0.5068(0.77)	0.6168(0.9286)	0.6574(0.9904)	0.6526(1)	0.6624(1)
	0.2	0.408(0.6592)	0.3982(0.7722)	0.2778(0.8554)	0.0912(0.944)	0.0162(0.986)
	0.4	0.2484(0.4662)	0.2006(0.4592)	0.1236(0.475)	0.0314(0.4302)	0.0048(0.398)
	0.6	0.1518(0.2842)	0.1298(0.2926)	0.108(0.2856)	0.046(0.2124)	0.0076(0.1302)
	0.8	0.101(0.1954)	0.1094(0.2038)	0.0968(0.2216)	0.0782(0.195)	0.0342(0.113)
Poisson	−0.8	0.273	0.1978	0.1086	0.029	0.0042
	−0.6	0.2254	0.1236	0.0522	0.0052	0.0002
	−0.4	0.2258	0.1048	0.0266	0.0018	0
	−0.2	0.3786	0.2464	0.1026	0.0172	0.0006
	0.0	0.5638	0.576	0.6746	0.8608	0.964
	0.2	0.4354	0.2778	0.1248	0.018	0.0016
	0.4	0.279	0.1148	0.028	0.0004	0
	0.6	0.2702	0.1436	0.0422	0.0024	0
	0.8	0.3602	0.218	0.1066	0.0156	0.0018

Tabla 4-20.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$

4.2.2. Modelo lineal clásico

En las tablas de la **Tablas 4-21** a la **4-24**, se muestra el comportamiento del criterio $\Delta\hat{\beta}\,\%$ cuando se trabaja con un modelo lineal clásico.

Cuando ρ toma valores ± 0.8 y ± 0.6, la proporción de veces en la que $\Delta\hat{\beta}\,\% < \delta$ es nula, esto significa que para los $N = 5000$ modelos simulados bajo un alto grado de confusión, este criterio detectó que en todas el fenómeno realmente estaba presente.

En la **Figura 4-10**, se observa que a medida que los tamaños muestrales (n) y el nivel de referencia (δ) aumentan y que no existe confusión en el modelo ($\rho = 0$), el desempeño del criterio mejora. Note por ejemplo en la **Tabla 4-23** donde $\delta = 0.20$, el criterio detecta efectivamente la ausencia de confusión en más del 80 % de los modelos simulados, esto sin importar el tamaño en las muestras.

Para $\rho \pm 0.4$ y $n = 50$, se observa que $\Delta\hat{\beta}\,\%$ tiende a equivocarse aproximadamente entre el 1 % y el 6 % de las simulaciones, aunque podría pensarse que este "pequeño error" sucede porque el tamaño de muestra no es lo suficientemente grande, debido a que cuando

$n \geq 100$, el criterio funciona perfectamente. Análogamente sucede cuando $\rho \pm 0.2$, donde el desempeño del criterio es cada vez más bajo cuando se aumenta el nivel de referencia.

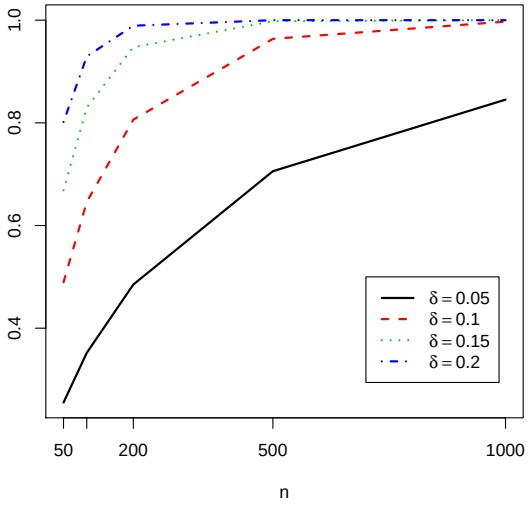

Figura 4-10.: Comportamiento del criterio $\Delta\hat{\beta}\,\%$ en el modelo lineal clásico cuando $\rho = 0$.

ρ	n				
	50	100	200	500	1000
-0.8	0	0	0	0	0
-0.6	0	0	0	0	0
-0.4	0.0042	0	0	0	0
-0.2	0.0968	0.0484	0.0146	0.0002	0
0	0.2552	0.352	0.485	0.7058	0.8452
0.2	0.1028	0.0546	0.0134	0	0
0.4	0.0046	0	0	0	0
0.6	0	0	0	0	0
0.8	0	0	0	0	0

Tabla 4-21.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.05$

ρ	n				
	50	100	200	500	1000
−0.8	0	0	0	0	0
−0.6	0	0	0	0	0
−0.4	0.0108	0.0004	0	0	0
−0.2	0.2076	0.1306	0.0594	0.0068	0.0004
0	0.4894	0.646	0.8062	0.9636	0.9972
0.2	0.2158	0.1398	0.0558	0.0062	0.0002
0.4	0.0128	0.0004	0	0	0
0.6	0	0	0	0	0
0.8	0	0	0	0	0

Tabla 4-22.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.10$

ρ	n				
	50	100	200	500	1000
−0.8	0	0	0	0	0
−0.6	0	0	0	0	0
−0.4	0.0244	0.0024	0	0	0
−0.2	0.3368	0.2622	0.1892	0.0814	0.0256
0	0.6686	0.83	0.9474	0.998	1
0.2	0.3386	0.2758	0.1884	0.0844	0.0236
0.4	0.029	0.0036	0	0	0
0.6	0.0002	0	0	0	0
0.8	0	0	0	0	0

Tabla 4-23.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.15$

ρ	n				
	50	100	200	500	1000
−0.8	0	0	0	0	0
−0.6	0.0006	0	0	0	0
−0.4	0.0556	0.0104	0.0002	0	0
−0.2	0.4518	0.4434	0.4088	0.3694	0.3184
0	0.8016	0.9294	0.9892	1	1
0.2	0.4596	0.4564	0.416	0.3756	0.3218
0.4	0.0596	0.0122	0.0014	0	0
0.6	0.0004	0	0	0	0
0.8	0	0	0	0	0

Tabla 4-24.: Proporción de veces en las que el cambio porcentual $\Delta\hat{\beta}\,\% < \delta = 0.20$

5. Conclusiones y recomendaciones

En este trabajo, fue estudiado el criterio de cambio porcentual propuesto por Hosmer & Lemeshow (1999), como una posible medida de confusión en modelos de regresión. La propuesta hecha por ellos y dada por la ecuación (3-22), fue deducida a partir de un modelo de riesgos proporcionales.

El eje central de esta investigación se basó en las siguientes preguntas: ¿tiene ese criterio un buen desempeño en cualquier modelo de regresión? ¿cuál era el nivel de referencia adecuado para decidir si el fenómeno de la confusión estaba presente o no en cada modelo? Para dar respuesta a estos interrogantes, se estudió -vía simulación- el comportamiento de $\Delta\hat{\beta}\,\%$ en tres miembros de la familia exponencial (3-4), que son el modelo lineal clásico, el modelo logístico y el modelo Poisson; y se implementaron cuatro niveles de referencia δ con el objetivo de evaluar el desempeño del criterio.

De acuerdo con los resultados del estudio de simulación, se encontró que independientemente de la probabilidad del resultado, el nivel de referencia, el tamaño de la muestra y el grado de confusión, el criterio $\Delta\hat{\beta}\,\%$ mostró tener buen desempeño en el modelo logístico, sin embargo éste tiende a fallar cuando no existe correlación entre las covariables ($\rho = 0$), principalmente en los escenarios 1 y 2, donde la probabilidad del resultado se permitía variar en un rango de muy bajo a muy alto. Al igual que el criterio del cambio porcentual, el criterio propuesto tiene un desempeño excelente a lo largo de los escenarios simulados, pero es más confiable para detectar la ausencia de confusión; cuando se emplee este criterio como medida de confusión, se recomienda utilizar un nivel de referencia $\delta = 0.10$.

El principal hallazgo es que cuando se simula la mayor variabilidad en la variable respuesta del modelo Poisson, es decir bajo el primer escenario, el criterio falló completamente en todos los modelos simulados con confusión, incluso con un tamaño de muestra grande. En cambio, el criterio fue efectivo en todos los modelos simulados cuyas covariables tenian $\rho = 0$. Estas dos interpretaciones se pueden comprobar con los resultados expuestos en las Tablas **4-1** a **4-4**.

Por otra parte, cuando se considera el escenario 2 en el modelo Poisson, se recomienda usar el criterio con cualquiera de los δ que se utilizaron en este estudio, siempre que $n \geq 500$. En situaciones similares a los escenarios 3 y 5 y $n \geq 100$, un nivel de referencia $\delta = 0.10$ es más adecuado cuando se utilice a $\Delta\hat{\beta}\,\%$ como medida de confusión, ya que con éste la proporción de veces en las que $\Delta\hat{\beta}\,\% < \delta$ es cercana a cero cuando la confusión realmente está presente.

En forma general, el criterio estudiado mostró a lo largo de los escenarios de simulación, tener mejor desempeño en el modelo lineal clásico, donde un nivel de referencia del 5 % es suficiente para que dicho criterio fuera capaz de detectar en más del 90 % de las situaciones simuladas con el fenómeno de la confusión; pero para garantizar la veracidad del criterio en situaciones donde no haya confusión, se recomienda utilizar $\delta = 0.15$.

Para trabajos futuros se sugiere que los ejercicios de simulación en el estudio del desempeño del criterio del cambio porcentual sean aplicados a más escenarios, donde se consideren otras distribuciones de la familia exponencial o de algunas otras de colas más pesadas,o usando modelos gerárquicos o datos longitudinales con respuesta continua y categórica. También se podrían generar escenarios con estructuras más complejas para el vector de parámetros que incluyan múltiples confusores (discretos y continuos)y la generación de variables explicativas podría ser a partir de distribuciones multivariadas diferentes a la normal bivariada.

En el caso de la regresión lineal tradicional, la evaluación de la confusión podría ser realizada usando estimadores sesgados de la familia de los estimadores contraídos, tal como el estimador Ridge o el estimador de componentes principales, que bien elegidos, pueden arrojar estimadores de bajo sesgo y mayor precisión.

Resultaría muy interesante conocer la distribución de $\Delta\hat{\beta}\,\%$, una vía para determinarla es traves de métodos de remuestreo, tales como el Bootstrap y el jakknife, y apartir de ahí evaluar el desempeño del criterio propuesto por Hosmer & Lemeshow (1999) mirando la proporcion de veces en que se rechaza la hipótesis nula $H_0 : \ \Delta\hat{\beta}\,\% = \delta$ cuándo ésta realmente es cierta o hallando intervalos de confianza bootstrap para este criterio.

A. Programa en R para el estudio de simulación

A.1. Modelo logístico

```
library(MASS)
f<-function(bv,n,rho){
                      mu<-c(0,0)
                      sigma<-matrix(c(1,rho,rho,1),2,2)
                      x<-mvrnorm(n,mu,sigma)
                      m.d<-cbind(rep(1,n),x)
                      p<-exp(m.d%*%bv)/(1+exp(m.d%*%bv))
                      y<-rbinom(n,1,p)
                      #Ajuste sin confusor
                      mod.sin<-glm(y~x[,1],family=binomial)
                      theta<-summary(mod.sin)$coefficients[2]

                      #Ajuste con confusor
                      mod.con<-glm(y~x[,1]+x[,2],family=binomial())
                      bta<-summary(mod.con)$coefficients[2]

                      #Criterio de Hosmer and Lemeshows
                      C.HL<-abs((theta-bta)/bta)
                      #Nuevo criterio
                      New.C<- abs((exp(theta)-exp(bta))/exp(bta))
                      return(c(C.HL,New.C))
                     }
rta<-replicate(5000,f(c(-2.9,0,0.5),1000,0.8))

#Criterio Hosmer and Lemeshow
length(table(which(rta[1,]<0.05)))/5000
```

```
length(table(which(rta[1,]<0.10)))/5000
length(table(which(rta[1,]<0.15)))/5000
length(table(which(rta[1,]<0.20)))/5000

#Criterio propuesto
length(table(which(rta[2,]<0.05)))/5000
length(table(which(rta[2,]<0.10)))/5000
length(table(which(rta[2,]<0.15)))/5000
length(table(which(rta[2,]<0.20)))/5000
```

A.2. Modelo Poisson

```
library(MASS)
f<-function(bv,n,rho){
                      mu<-c(0,0)
                      sigma<-matrix(c(1,rho,rho,1),2,2)
                      x<-mvrnorm(n,mu,sigma)
                      xd<-cbind(rep(1,n),x)
                      lambda<-exp(xd%*%bv)
                      y<-rpois(n,lambda)

                      #Ajuste sin confusor
                      mod.sin<-glm(y~x[,1],family=poisson())
                      theta<-summary(mod.sin)$coefficients[2]

                      #Ajuste con confusor
                      mod.con<-glm(y~x[,1]+x[,2],family=poisson())
                      bta<-summary(mod.con)$coefficients[2]

                      #Criterio de Hosmer and Lemeshow
                      criterio<-abs(theta-bta)/bta
                      return(criterio)
                      }
rta<-replicate(5000,f(c(-0.7,0.2,0.5),1000,-0.8))
length(table(which(rta<0.05)))/5000
length(table(which(rta<0.10)))/5000
length(table(which(rta<0.15)))/5000
```

```
length(table(which(rta<0.20)))/5000
```

A.3. Modelo lineal clásico

```
library(MASS)
f<-function(bv,n,rho){
                      mu<-c(0,0)
                      sigma<-matrix(c(1,rho,rho,1),2,2)
                      x<-mvrnorm(n,mu,sigma)
                      m.d<-cbind(rep(1,n),x)
                      btas<-as.matrix(bv)
                      y<-m.d%*%btas

                      #Ajuste sin confusor
                      mod.sin<-glm(y~m.d[,2],family=gaussian)
                      theta<-summary(mod.sin)$coefficients[2]

                      #Ajuste con confusor
                      mod.con<-glm(y~m.d[,2]+m.d[,3],family=gaussian)
                      bta<-summary(mod.con)$coefficients[2]

                      #Criterio de Hosmer and Lemeshow
                      criterio<-abs((theta-bta)/bta)
                      return(criterio)
                     }
rta<-replicate(5000,f(c(-0.6,1.3,1.4),1000,0.8))
length(table(which(rta<0.05)))/5000
length(table(which(rta<0.10)))/5000
length(table(which(rta<0.15)))/5000
length(table(which(rta<0.20)))/5000
```

Bibliografía

Agresti, A. (2002), *Categorical Data Analysis*, 2 edn, John Wiley & Sons, Inc., New York.

Austin, P. & Brunner, L. (2004), 'Inflation of the Type I Error Rate When a Continuous Confounding Variable is Categorized in Logistic Regression Analysis', *Statistics in Medicine* **23**, 1159–1178.

Becher, H. (1992), 'The concept of Residual Confounding in Regression Models and some applications', *Statistics in Medicine* **11**, 1747–1758.

Chao, W., P. M. & Young, T. (1997), 'Efect of omitted Confounders on the Analysis of Correlated Binary Data.', *Biometrics* **53**, 678–689.

Chen, C., C. D. & Winkler, S. (1999), 'A Simulation Study of Confounding in Generalized Linear Models for Air Pollution Epidemiology', *Environmental Health Perspectives* **107**(3), 217–222.

Cochran, W. (1968), 'The effectiveness of adjustment by subclassification in removing bias in observational studies', *Biometrics* **24**, 295–313.

Cox, D. (1972), 'Regression models and life tables (with discussion)', *Journal of Royal Statistical Society* (34), 187–220.

Diaz, L. G. & Morales, M. A. (2009), *Análisis de datos Categóricos*, 1 edn, Editorial Universidad Nacional de Colombia, Bogotá.

Dobson, A. (2001), *An introduction to Generalized Linear Models*, 2 edn, Chapman & Hall/CRC, Londres.

Frank, K. (2000), 'Impact of a Confounding Variable on the Inference of a Regression Coefficient', *Sociological Methods and Research* **29**(2), 147–194.

Friendly, M. (2000), *Visualizing Categorical Data*, 1 edn, Cary, NC: SAS Institute Inc.,, North Carolina, USA.

Greenland, S., R. J. & Pearl, J. (1999), 'Confounding and collapsability in causal inference', *Statistics in Science* **14**(3), 29–46.

Hosmer, D. W. & Lemeshow, S. (1999), *Applied Survival Analysis*, John Wiley & Sons, Inc., New York.

Joffe, M. M. & Rosenbaum, P. R. (1999), 'Invited commentary: Propensity scores', *American Journal of Epidemiology* **150**, 327–333.

Kamangar, F. (2012), 'Confounding Variables in Epidemiologic Studies: Basics and Beyond', *Arch Iran Med* **15**(6), 508–516.

Lindsey, J. (2007), *Applying Generalized Linear Models*, Limburg Universitair Centrum: Diepembeek.

Mcnamee, R. (2005), 'Regression modelling and other methods to control confounding', *Occup Environ Med* **62**, 500–506.

Mickey, J. & Greenland, S. (1989), 'A study of the impact of confounder selecction criteria of effect estimation.', *American Journal of epidemiology* **129**, 125–137.

Newman, S. C. (2001), *Biostatistical Methods in Epidemiology*, John Wiley & Sons, Inc., New York.

Pourhoseingholi, M. A., Baghestani, A. R. & Vahedi, M. (2012), 'How to Control Confounding Effects by Statistical Analysis', *Gastroenterol Hepatol Bed Bench* **5**(2), 79–83.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**, 41–55.

Schelesselman, J. (1982), *Case - Control Studies. Designs, Conduct and Analysis*, University Press:New York.

Wickramaratne, P. J. & Holford, T. R. (1987), 'Confounding in Epidemiologic Studies: The Adequacy of the Control Group as a Measure of Confounding', *Biometrics* **43**(4), 751–765.

Wilson, S. & Gordon, I. (1986), 'Calcutaing Sample Sizes in the Presence of Confounding Variables', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **35**(2), 207–213.

Woodward, M. (1999), *Epidemiology. Study Design and Data Analisys*, 2 edn, Boca Ratón, Chapman & Hall/CRC, .

Wunsch, G. (2007), 'Confounding and control', *Demographic research* **16**(4), 97–120.