



Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario

Estimación del Modelo ZIP Parcialmente Lineal: un Enfoque Robusto Utilizando un Algoritmo tipo EM

Autora: Lic. María José Llop

Directora: Dra. Andrea Bergesio

Co-directora: Dra. Anne-Françoise Yao

Tesis presentada para obtener el grado de Doctora en Estadística
defendida ante el jurado compuesto por:

Dra. Liliana Forzani

Dra. Daniela Rodriguez

Dra. M. Magdalena Lucini

2025

*A Mamá y Papá,
con todo mi amor.*

Agradecimientos

A mi familia: Mamá, Papá, Romi, Ale y Pame, por el amor y el apoyo incondicional. Especialmente a Mamá y Papá por esforzarse incansablemente para que sus hijas e hijo pudiéramos estudiar, por hacer de la educación una prioridad y enseñarme el valor de la dedicación, la responsabilidad y la humildad. A Pame, por ser mi guía y darme la tranquilidad de saber que siempre puedo contar con ella. A los pequeños Emi y Bere por llenarme siempre de amor y alegría.

A Nahuel, por darme su amor y compañía cada día, por ser un pilar esencial en mi vida y brindarme siempre todo lo que necesito; por acompañarme, cuidarme y apoyarme sin condiciones.

A Mariel, Barbi, Ita y Nano, por la amistad y contención diaria. Sin ellos todo habría sido más difícil y, sin dudas, menos divertido.

A Jesi, Yany, Karen, Gise, Stefi, Geni, Agus y Sil, por regalarme su amistad durante todos estos años y por lograr, entre risas y charlas, que el camino transitado sea siempre más feliz.

A Andrea, por la generosidad con la que me compartió su conocimiento, por estar siempre presente cuando la necesité; por la dedicación, la paciencia y el compromiso con los que acompañó y guió todo el proceso de esta tesis.

A Anne, por confiar plenamente en mí desde el comienzo; por brindarme oportunidades de expandir mis horizontes en el mundo de la investigación, transmitiéndome con calidez y generosidad su experiencia.

A los y las docentes de quienes tuve el privilegio de aprender durante toda mi formación académica; a la Facultad de Ciencias Económicas y Estadística (UNR), a los miembros del jurado, a la educación pública y al CONICET.

A todos ellos, mi gratitud más profunda.

Resumen

Los datos de conteo con exceso de ceros son frecuentes en disciplinas como economía, medicina, salud pública, entre otras. El modelo de regresión de Poisson inflado en ceros (ZIP, por sus siglas en inglés *zero-inflated Poisson*) puede aplicarse en estos casos, ya que combina la distribución de Poisson clásica con una distribución degenerada que asigna masa uno al valor cero, permitiendo así capturar adecuadamente la sobreabundancia de ceros observada en la práctica.

Como una extensión más flexible, el modelo de regresión de Poisson parcialmente lineal inflado en ceros (PLZIP, por sus siglas en inglés *partially linear zero-inflated Poisson*) introduce una estructura semiparamétrica que habilita la incorporación de ciertas covariables de manera no paramétrica. Esta formulación amplía la capacidad del modelo para adaptarse a situaciones en que las variables se relacionan de manera compleja. No obstante, los enfoques de estimación disponibles para el modelo PLZIP se basan en la función de verosimilitud, la cual es altamente sensible a valores atípicos y a pequeñas desviaciones de los supuestos del modelo.

Esta tesis propone un procedimiento de estimación robusto para el modelo PLZIP, basado en un algoritmo tipo Esperanza-Maximización (EM) con un procedimiento secuencial de tres pasos que permite estimar las componentes lineales y no paramétricas, capitalizando la naturaleza de mezcla del modelo e incorporando funciones de pérdida robustas para mitigar la influencia de observaciones extremas. Se establecen resultados para la convergencia del algoritmo, la consistencia y la distribución asintótica de los estimadores. Estudios de simulación confirman la robustez y la eficiencia de los estimadores frente a distintos escenarios de contaminación en los datos, así como su mejor desempeño en comparación con métodos clásicos. Finalmente, se ilustra la aplicación de la metodología propuesta mediante ejemplos con datos reales. Esta propuesta de estimación robusta desarrollada específicamente para el modelo PLZIP, constituye un avance significativo en el análisis robusto de datos de conteo inflados en ceros, bajo la presencia de relaciones no estrictamente lineales entre las variables.

Palabras clave: Modelo de Poisson inflado en ceros parcialmente lineal, Algoritmo de tipo EM, Estimación robusta, Propiedades asintóticas.

Resumen

Count data with excess zeros are common in disciplines such as economics, medicine, and public health, among others. The zero-inflated Poisson (ZIP) regression model can be applied in these settings, as it combines the classical Poisson distribution with a degenerate distribution that assigns probability mass one to zero, thereby adequately capturing the overabundance of zeros observed in practice.

As a more flexible extension, the partially linear zero-inflated Poisson (PLZIP) regression model introduces a semiparametric structure that allows certain covariates to enter the model in a nonparametric manner. This formulation expands the model's ability to adapt to situations in which the relationships among variables are complex.

This thesis presents a robust estimation procedure for the PLZIP model, based on an Expectation-Maximization-like algorithm with a three-step sequential procedure that enables the estimation of both linear and nonparametric components. By capitalizing on the mixture structure of the model and incorporating robust loss functions, the proposed approach effectively mitigates the influence of extreme observations. Results on the convergence of the algorithm, as well as on the consistency and asymptotic distribution of the estimators, are established. Simulation studies demonstrate the robustness and efficiency of the proposed estimators under various contamination scenarios, as well as their superior performance compared to classical methods. Finally, the practical utility of the proposed methodology is illustrated through examples of application to real datasets. This work, which develops robust estimation specifically for the PLZIP model, constitutes a significant advance in the robust analysis of zero-inflated count data in the presence of non-strictly linear relationships between variables.

Keywords: Partially linear zero-inflated Poisson regression model, EM-like algorithm, Robust estimation, Asymptotic properties.

Índice general

Índice general	v
Índice de figuras	VII
Índice de cuadros	IX
Índice alfabético	I
Tabla de Notación	2
Introducción	3
Objetivos	8
1 Antecedentes	10
1.1. Modelos Lineales Generalizados	10
1.2. Modelos Parcialmente Lineales Generalizados	20
1.3. Modelo de Poisson Inflado en Ceros	28
1.4. Modelo de Poisson Inflado en Ceros Parcialmente Lineal	40
2 Estimación Robusta del Modelo de Poisson Inflado en Ceros Parcialmente Lineal	44
2.1. Procedimiento de Estimación Basado en Verosimilitud	44
2.2. Procedimiento Robusto de Estimación	46
2.3. Algoritmo Tipo EM	50
2.4. Consistencia	52
2.5. Distribución Asintótica	62
2.6. Test de hipótesis	71

3 Estudios de Simulación y Ejemplos	74
3.1. Estudios de Simulación	74
3.2. Ejemplos de Aplicación a Datos Reales	81
Apéndice A	93
A.1. Resultados clásicos	93
A.2. Algoritmo ES	95
A.3. Procesos empíricos	96
A.4. Propiedades asintóticas	99
A.5. Procedimiento de estimación del modelo ZIP doblemente semiparamétrico .	104
Conclusiones	106
Bibliografía	107

Índice de figuras

1.1. Funciones de pérdida $\phi(u)$ con $c = 1$ (curvas azules) junto a la pérdida cuadrática (curvas punteadas).	15
1.2. Varianza de \sqrt{Y} con $Y \sim \mathcal{P}(\lambda)$, para diferentes valores de λ	19
1.3. Funciones núcleo $K(x)$	23
1.4. Ajustes dados por las distribuciones de Poisson y binomial negativa.	29
3.1. Distancia entre β_0 y los estimadores. Los paneles a), b), d) y e) representan cada esquema de contaminación. c) y f) son vistas ampliadas de los paneles b) y e), respectivamente.	79
3.2. Distancia en norma ℓ_2 entre γ_0 y cada estimador. Los paneles a), b), c) y d) representan cada esquema de contaminación.	80
3.3. RMSE de cada estimador del componente no paramétrico m_0 . Los paneles a), b), d) y e) representan cada esquema de contaminación. c) y f) son vistas ampliadas de los paneles b) y e), respectivamente.	81
3.4. Función verdadera m_0 y sus estimaciones obtenidas a partir de un conjunto de datos simulado específico bajo cada esquema de contaminación.	82
3.5. Gráfico de barras de las abundancias de la bacteria <i>Akkermansia</i> (El 3.8% de los datos presentan valores mayores a 100 y no se muestran en el gráfico) . . .	83
3.6. Panel a): mediana del error cuadrático de predicción de cada estimador, para los $k = 5$ folds. Panel b): ampliación del Panel a).	85
3.7. Estimadores de $m(t)$ basado en datos de APG.	85
3.8. Gráfico de barras del número de días de ausencia en las actividades principales debido a enfermedad en las últimas 4 semanas, según lo auto-reportado por el encuestado.	87
3.9. Mediana del error cuadrático de predicción de cada estimador, para los $k = 5$ folds.	89

3.10. Mediana del error cuadrático de predicción de cada estimador, para los $k = 5$ folds.	89
3.11. Estimadores de $m(t)$ basado en datos de la encuesta de salud pública en Indonesia.	91
3.12. Estimadores de $m_\lambda(t)$ y $m_\pi(t)$ basado en datos de la encuesta de salud pública en Indonesia.	91

Índice de cuadros

1.1. Funciones de pérdida $\phi(u)$	15
1.2. Funciones núcleo $K(x)$	22
3.1. Estimaciones de los coeficientes β y γ para los distintos estimadores aplicados a los datos de <i>Akkermansia</i>	86
3.2. Estimadores de β y γ	90

Tabla de Notación

Símbolo	Significado
\mathbb{R}	Conjunto de los números reales
\mathbb{R}^+	Conjunto de los números reales positivos
\mathbb{R}_0^+	Conjunto de los números reales positivos incluido el 0
\mathbb{N}_0	Conjunto de los números naturales incluido el 0
\mathbf{I}_p	Matriz identidad de dimensión p
\mathbf{A}^\top	Transpuesta de la matriz \mathbf{A}
\mathbf{A}^{-1}	Inversa de la matriz \mathbf{A}
$\text{tr}(\mathbf{A})$	Traza de la matriz \mathbf{A}
$\lambda_p(\mathbf{A})$	p -ésimo autovalor de la matriz \mathbf{A} , ordenados en forma decreciente
$E(\cdot)$	Esperanza matemática
$\text{Var}(\cdot)$	Varianza
$\text{Cov}(\cdot, \cdot)$	Covarianza
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal multivariada p -dimensional con media $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$
$\xrightarrow{\text{c.s.}}$	Convergencia casi segura
$\xrightarrow{\mathcal{D}}$	Convergencia en distribución
\xrightarrow{P}	Convergencia en probabilidad
$o_p(\cdot)$	Orden estocástico
$\mathcal{O}_p(\cdot)$	Orden acotado en probabilidad
$\mathcal{O}(\cdot)$	Orden de magnitud
$\ v(t)\ _{0,\infty}$	$\sup_{t \in \mathcal{T}_0} v(t) $ con \mathcal{T}_0 compacto y $v : \mathbb{R} \rightarrow \mathbb{R}$
$\ v(t)\ _\infty$	$\sup_{t \in \mathcal{T}} v(t) $ con $v : \mathbb{R} \rightarrow \mathbb{R}$
$\ \mathbf{X}\ $	Norma ℓ_2 del vector $\mathbf{X} \in \mathbb{R}^p$
$\ \mathbf{A}\ $	Norma inducida de la matriz $\mathbf{A} \in \mathbb{R}^{m \times p}$
$\mathbb{I}(\cdot)$	Función indicadora
$a \gg b$	a es mucho mayor que b
$a \ll b$	a es mucho menor que b
$a \doteq b$	a se define como b

Introducción

El estudio de datos de conteo es frecuente en diversas áreas científicas, como econometría, biología, epidemiología, industria y salud, entre otras. Los modelos de regresión de Poisson y binomial negativo, que se enmarcan dentro de los modelos lineales generalizados (GLM, por sus siglas en inglés *generalized linear models*), proporcionan un marco metodológico potente para analizar este tipo de datos; detalles adicionales sobre los GLM pueden consultarse en McCullagh and Nelder (1989). Sin embargo, en numerosas aplicaciones, los datos de conteo presentan un exceso de ceros, es decir, una proporción de ceros superior a la esperada bajo los GLM clásicos. Para abordar esta problemática, se han propuesto diversas alternativas. Entre ellas, el modelo de obstáculos (*hurdle*, en inglés) introducido por Cragg (1971) asume que todos los ceros observados provienen de un proceso estructural inherente a los datos. Este enfoque combina una componente de conteo truncada en cero, que modela únicamente los valores positivos, con una componente binaria que describe la probabilidad de superar el “obstáculo” inicial asociado a la ocurrencia de ceros. Otra opción es el modelo inflado en ceros (ZI, por sus siglas en inglés *zero-inflated*), que difiere ligeramente del modelo de obstáculos, ya que es una mezcla de una distribución clásica de conteo y una distribución degenerada, con masa uno en cero. Este modelo asume que los ceros observados pueden ser estructurales, es decir, originados por alguna característica intrínseca del proceso, o bien de muestreo, cuando provienen de la distribución de conteo. Esta capacidad de modelar explícitamente ambas fuentes de ceros constituye una ventaja significativa en aplicaciones donde resulta importante distinguir entre unidades experimentales que nunca experimentan el evento y aquellas que podrían experimentarlo pero no lo hacen en la muestra observada. El modelo ZI puede interpretarse como un modelo de dos niveles: en el primer nivel, una distribución binomial discierne si la observación corresponde a un cero estructural y en el segundo nivel una distribución clásica de conteo modela los ceros no estructurales y los conteos positivos. Cuando se utiliza la distribución de Poisson en un modelo ZI, se obtiene el modelo de regresión de Poisson inflado en ceros (ZIP), discutido por Cohen (1967) y Johnson and Kotz (1969), y posteriormente extendido

al contexto de regresión por Lambert (1992) para analizar defectos de fabricación. Otros modelos inflados en ceros pueden obtenerse al emplear diferentes distribuciones de conteo. Por ejemplo, Greene (1994) utilizó la distribución binomial negativa, originando el modelo binomial negativo inflado en ceros (ZINB), mientras que Hall (2000) propuso el modelo binomial inflado en ceros (ZIB). En la literatura pueden encontrarse numerosas aplicaciones de estos modelos a problemas concretos. Por ejemplo, Dalrymple et al. (2003) examinaron la incidencia del síndrome de muerte súbita del lactante en función del clima mediante modelos ZIP; Rose et al. (2006) compararon el ajuste de diferentes modelos inflados en ceros aplicados a recuentos de eventos adversos de vacunas. Asimismo, Liaqat et al. (2022) estudiaron los factores que afectan el número de ganglios linfáticos axilares en mujeres con cáncer de mama primario; Anggreainy et al. (2022) utilizaron un modelo binomial negativo inflado en ceros para analizar la tasa de mortalidad por COVID-19 en Indonesia en función de indicadores de pobreza. Roy and Hasanuzzaman (2025) estudiaron como la mejora energética, y consecuentemente la reducción del uso de combustibles peligrosos, ayuda a disminuir el riesgo de enfermedades respiratorias en hogares de Bangladesh, utilizando un modelo de regresión ZIP e incluyendo diversos factores socioeconómicos y demográficos; y Tiruye et al. (2025) aplicaron modelos de regresión ZIP para identificar disparidades socioeconómicas y geográficas en los procedimientos posteriores a la radioterapia en hombres con cáncer de próstata.

En casos donde algunas de las variables predictoras se relacionan de manera no lineal con la respuesta, los modelos mencionados anteriormente pueden resultar demasiado restrictivos, debido a su naturaleza lineal. Para capturar de manera más adecuada la verdadera asociación entre las variables, se pueden emplear modelos de regresión semiparamétricos que relajan los supuestos del modelo. Específicamente, el modelo ZIP puede extenderse a una versión parcialmente lineal que incorpora una función no paramétrica, además del término lineal, al modelar la media de la distribución de Poisson. En adelante se utilizarán las siglas PLZIP para hacer referencia a este modelo, debido a su nombre en inglés *partially linear zero-inflates Poisson*. La obtención de estimadores para el modelo PLZIP puede ser desafiante, dado que combina términos paramétricos y no paramétricos, los cuales requieren técnicas de estimación específicas. Lam et al. (2006) introdujeron inicialmente el modelo PLZIP y propusieron estimadores de máxima verosimilitud de tamiz, utilizando funciones lineales a trozos para aproximar la función no paramétrica. Posteriormente otros autores han desarrollado algunas extensiones del modelo PLZIP. Particularmente, He et al. (2010) estudiaron un modelo de regresión ZIP doblemente semiparamétrico y desarrollaron estimadores de máxima verosimilitud de tamiz. Este enfoque aproxima la función no paramétrica mediante un espacio de funciones de dimensión finita que crece con el tamaño de la muestra, en este caso, funciones lineales a trozos. Por otra parte, Wang et al. (2015) propusieron un modelo ZIP con índice único parcialmente lineal y lo estimaron mediante el método de perfiles de verosimilitud. Este enfoque consiste en fijar un valor provisional

para la componente paramétrica y, en función de este, estimar la componente no paramétrica. Una vez reemplazada esta estimación en la verosimilitud original, se la maximiza respecto de los parámetros. Los autores aproximaron la componente no paramétrica con funciones de base B-spline.

A pesar de los mencionados avances metodológicos para el modelo PLZIP, todos los enfoques de estimación existentes se basan en la función de verosimilitud, la cual es conocida por su alta sensibilidad a valores atípicos y a pequeñas desviaciones respecto de los supuestos del modelo. Por este motivo, resulta necesario desarrollar métodos robustos de estimación para el modelo PLZIP que mitiguen estos problemas y proporcionen inferencias más confiables.

Los métodos robustos de estimación han sido estudiados en la literatura para una amplia variedad de modelos. Huber (1964) propuso una generalización del método de máxima verosimilitud, reemplazando la función de *score* en la ecuación de verosimilitud por una función de estimación que reduce el peso asignado a las observaciones extremas, dando lugar a los conocidos M-estimadores. En el marco de los GLM, se han propuesto múltiples enfoques robustos basados en M-estimación. Por ejemplo, Bianco and Yohai (1996) introdujeron M-estimadores para el modelo de regresión logística y estudiaron las condiciones necesarias para garantizar la consistencia de Fisher. Por su parte, Croux and Haesbroeck (2002) complementaron dicho estudio al proponer un algoritmo rápido y estable para calcular estos estimadores. Además, establecieron condiciones para la existencia del estimador en muestras finitas y discutieron la elección de la función de pérdida. Estos estimadores fueron posteriormente generalizados al caso de respuestas faltantes por Bianco et al. (2012) y aplicados a los modelos de regresión de Poisson y Gamma. Por otro lado, Cantoni and Ronchetti (2001) propusieron estimadores robustos basados en la cuasi-verosimilitud, centrándose particularmente en los modelos binomial y de Poisson. Más recientemente, Bergesio and Yohai (2011) presentaron una propuesta robusta basada en la clase de estimadores de proyección para regresión lineal, mientras que Valdora and Yohai (2014) propusieron aplicar una transformación estabilizadora de varianza a la respuesta y luego utilizar un M-estimador.

Si bien es evidente que los M-estimadores clásicos constituyen uno de los enfoques más utilizados y consolidados para lograr robustez en GLM, pueden no ser adecuados en el contexto de modelos de mezcla. Tal como señalan Hall and Shen (2010), en este tipo de modelos muchas observaciones pueden presentar residuos naturalmente grandes respecto de la media global, especialmente cuando los componentes de la mezcla están balanceados y bien separados (por ejemplo, en el caso del modelo ZIP, cuando aproximadamente la mitad de las observaciones son ceros y la media de la distribución de Poisson es elevada). Esto conduce a asignar ponderaciones bajas a observaciones que no necesariamente son extremas con respecto a la componente de la distribución a la que pertenecen, lo que puede

introducir sesgos en la estimación.

Para superar esta limitación, Hall and Shen (2010) propusieron utilizar el algoritmo de Esperanza-Maximización (EM) para identificar primero la pertenencia de cada observación a cada componente y, posteriormente, atenuar la influencia de las observaciones extremas dentro de cada componente. Este algoritmo iterativo permite calcular los estimadores de máxima verosimilitud (MLE, por sus siglas en inglés *maximum likelihood estimators*) en situaciones en las que los datos pueden considerarse incompletos debido a la presencia de variables no observables. Alterna entre el paso de esperanza (E), en el cual se construye la esperanza de la log-verosimilitud condicionada a los datos observables, evaluada con los estimadores actuales; y el paso de maximización (M), donde se obtienen nuevos estimadores que maximizan la log-verosimilitud esperada. Este procedimiento simplifica el problema, al ajustar iterativamente versiones ponderadas de GLM más sencillos, dependientes de distintos subconjuntos de parámetros, que pueden optimizarse de forma separada facilitando notablemente la estimación. Para una cobertura integral del algoritmo EM, incluida su teoría, aplicaciones y extensiones, se recomienda consultar McLachlan and Krishnan (2007).

Específicamente, Hall and Shen (2010) introdujeron el método RES (*robust expectation-solution*) para estimar el modelo de regresión ZIP. Dicho procedimiento consiste en realizar una M-estimación robusta en el paso S de un algoritmo de Esperanza-Solución (ES). El algoritmo ES es similar al EM, pero, en lugar de maximizar la log-verosimilitud esperada, resuelve una ecuación de estimación. Los autores compararon el estimador RES con el propuesto por Lu et al. (2003) para mezclas finitas de modelos de regresión de Poisson, basado en la distancia mínima de Hellinger.

En modelos parcialmente lineales, los valores extremos también pueden afectar la forma y la escala de la estimación de la función no paramétrica. Para afrontar esa dificultad, diversos métodos robustos han sido propuestos en la literatura. Por ejemplo, Boente et al. (2006) analizaron la sensibilidad de los estimadores clásicos en modelos parcialmente lineales generalizados y propusieron estimadores robustos basados en perfiles de verosimilitud generalizados. Como señalan Boente and Rodriguez (2010), la principal desventaja de esta propuesta radica en el alto costo computacional y en la dependencia de las matrices de covarianza de las derivadas de la función de regresión del perfil robusto respecto del término paramétrico, lo que dificulta su estimación. Para superar estas limitaciones, Boente and Rodriguez (2010) introdujeron un procedimiento robusto de tres pasos, basado en la propuesta de Carroll et al. (1997), que permite estimar simultáneamente los parámetros lineales y la función no paramétrica. Este procedimiento resulta menos costoso computacionalmente y evita la dependencia de la función de regresión del perfil respecto del término paramétrico. Los autores demostraron la consistencia y la normalidad asintótica de los estimadores resultantes, y propusieron un procedimiento robusto para el contraste

de hipótesis sobre los parámetros.

Por otro lado, Boente and Rodriguez (2012) estudiaron estimadores robustos en el modelo de índice único parcialmente lineal generalizado, mientras que Boente et al. (2013) propusieron una clase test robustos para la función no paramétrica basada en el procedimiento robusto de tres pasos de Boente and Rodriguez (2010), que permite decidir entre un modelo parcialmente lineal generalizado y un modelo lineal generalizado. Los autores también obtuvieron la distribución asintótica del estadístico bajo la hipótesis nula. Más recientemente, Boente et al. (2018) definieron una clase de estimadores robustos para la función no paramétrica y el parámetro de regresión en un modelo isotónico parcialmente lineal generalizado.

El objetivo principal de esta tesis es desarrollar un procedimiento que permita obtener estimadores robustos del modelo de regresión PLZIP. Para ello, se propone un algoritmo de tipo EM que involucra el procedimiento de tres pasos introducido por Boente and Rodriguez (2010), lo que permite la estimación secuencial de los componentes lineales y no paramétricos. En particular, se incorporan funciones de pérdida robustas en el paso M del enfoque EM, en lugar de las funciones tradicionales de log-verosimilitud, lo que permite asignar peso a las observaciones extremas según el componente al que pertenecen. Al aprovechar la naturaleza de mezcla del modelo, este enfoque aborda tanto la relación posiblemente compleja entre las variables, como la sensibilidad a los valores atípicos. En esta tesis se presenta un método robusto de estimación desarrollado específicamente para el modelo PLZIP, lo que constituye un avance significativo en el análisis robusto de datos de conteo inflados en ceros bajo relaciones no estrictamente lineales entre las variables.

Esta tesis está organizada de la siguiente manera: a continuación se detallan los objetivos generales y específicos de este trabajo. En el Capítulo 1 se contextualiza teóricamente el problema central y se establece la base metodológica sobre la cual se fundamentarán los aportes de la tesis. El Capítulo 2 compone el aporte principal de esta tesis. El mismo se divide en secciones: en la Sección 2.1 se presenta una propuesta de estimación basada en verosimilitud para el modelo de regresión PLZIP que luego en la Sección 2.2 se extiende para obtener estimadores robustos. La Sección 2.3 muestra el algoritmo tipo EM que permite computar los estimadores propuestos, junto con un resultado de convergencia. Las propiedades asintóticas de los estimadores se demuestran en las Secciones 2.4 y 2.5. En particular, en la Sección 2.4 se muestra la consistencia fuerte de los estimadores, mientras que la distribución asintótica de los parámetros lineales se prueban en la Sección 2.5. El Capítulo 3 presenta los estudios de simulación en donde se evalúa el desempeño de los estimadores propuestos bajo diferentes esquemas de contaminación y se los compara con métodos clásicos y propuestas de otros autores. También proporciona ejemplos de aplicación de la metodología utilizando datos reales. Finalmente, se presentan las conclusiones generales de la tesis.

Objetivos

Objetivo General

El objetivo general de esta tesis es desarrollar estimadores robustos para el modelo de regresión de Poisson parcialmente lineal inflado en ceros, fundamentados en métodos robustos previamente propuestos para modelos lineales generalizados, que presenten simultáneamente buenas propiedades de eficiencia y resistencia frente a observaciones atípicas y desviaciones leves de los supuestos del modelo.

Objetivos Específicos

Los Objetivos Específicos de esta tesis son:

- I. Desarrollar un procedimiento robusto de estimación para el modelo PLZIP, que permita estimar tanto los parámetros lineales como el término no paramétrico. Adaptar y combinar estrategias robustas ya establecidas en el contexto de GLM, utilizando funciones de pérdida robustas y técnicas de estimación paramétricas y no paramétricas apropiadas.
- II. Establecer rigurosamente las propiedades teóricas de los estimadores desarrollados en I., incluyendo la consistencia fuerte y la distribución asintótica. Construir, en base a dichas propiedades, procedimientos de prueba robustos para la evaluación de hipótesis sobre los parámetros involucrados en el modelo.
- III. Diseñar e implementar estudios de simulación para comparar el desempeño de los estimadores robustos propuestos con el de estimadores clásicos, analizando especialmente el impacto de diferentes tipos de contaminación.

- IV. Evaluar la capacidad predictiva y la aplicabilidad práctica de los estimadores desarrollados en I. mediante su aplicación a conjuntos de datos reales que presenten exceso de ceros, valores atípicos y relaciones no estrictamente lineales entre las variables, comparando sus resultados con los obtenidos mediante otros enfoques metodológicos existentes en la literatura.

Antecedentes

El presente capítulo introduce y contextualiza de manera progresiva el problema central de esta tesis, estableciendo la base metodológica sobre la cual se fundamentarán los aportes propuestos. Cabe destacar que este capítulo no pretende tener carácter exhaustivo, sino que se acota a los requerimientos teóricos necesarios para que la tesis resulte, en lo posible, autocontenida. Específicamente introduce los conceptos y resultados principales asociados a los modelos lineales y parcialmente lineales generalizados, con especial énfasis en el modelo de regresión de Poisson inflado en ceros y su extensión parcialmente lineal. Se presenta una breve revisión de la literatura sobre enfoques de estimación clásicos y robustos para los modelos mencionados.

1.1. Modelos Lineales Generalizados

Los **modelos de regresión** son ampliamente utilizados para explicar la dependencia de una variable respuesta $Y \in \mathbb{R}$ respecto de una (regresión simple) o varias (regresión múltiple) variables predictoras. Sea $\mathbf{X} \in \mathbb{R}^p$ un vector de p predictoras, se asume que las variables siguen el modelo

$$Y = \nu(\mathbf{X}) + \varepsilon. \quad (1.1)$$

donde $\nu : \mathbb{R}^p \rightarrow \mathbb{R}$ se denomina función de regresión y $\varepsilon \in \mathbb{R}$ es el error aleatorio, independiente de \mathbf{X} , que verifica $E(\varepsilon) = 0$, de manera que $Y|\mathbf{X} \sim F$ con

$$\mu = E(Y|\mathbf{X}) = \nu(\mathbf{X}).$$

Los **modelos de regresión paramétricos** son aquellos que suponen una determinada familia de distribuciones para la variable respuesta. Consecuentemente, un número finito de parámetros que describen dichas distribuciones deben ser estimados. Uno de los modelos paramétricos más simples y clásicos de la literatura es el modelo de **regresión lineal**, en el que la función de regresión toma la forma $\nu(\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}$ donde $\boldsymbol{\beta} \in \mathbb{R}^p$ es un vector de

parámetros a ser estimado y los errores verifican $\varepsilon \sim N(0, \sigma^2)$. Si bien dichos modelos se destacan por su simplicidad y su interpretabilidad, las restricciones que imponen sobre la función de regresión y sobre la distribución de la respuesta pueden no tener en cuenta una variedad de situaciones prácticas. Por ejemplo, cuando se trata de una variable respuesta binaria o de conteo.

Los **modelos lineales generalizados** (GLM, por sus siglas en inglés *generalized linear models*), estudiados por Nelder and Wedderburn (1972) y McCullagh and Nelder (1989), son una herramienta flexible que extiende el modelo lineal clásico para tratar respuestas no necesariamente continuas ni normalmente distribuidas. Los GLM asumen que $Y|\mathbf{X}$ sigue una distribución perteneciente a la familia exponencial, usualmente en la forma canónica.

Las **familias exponenciales** son familias de distribuciones que unifican bajo una estructura en particular muchas de las distribuciones conocidas, tanto discretas como continuas. Su versatilidad las convierte en una herramienta fundamental para modelar distintos tipos de respuesta. Formalmente, se dice que una variable aleatoria Y tiene distribución perteneciente a la familia exponencial si su densidad (o función de probabilidad puntual, en el caso discreto) f se puede escribir como

$$f(y; \theta, \phi) = \exp \left\{ \frac{yb(\theta) - c(\theta)}{a(\phi)} + d(y, \phi) \right\}$$

donde $\theta \in \Theta \subset \mathbb{R}$ es el parámetro natural o canónico y $\phi \in \mathbb{R}$ es el parámetro de dispersión. $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ y $d(\cdot, \cdot)$ son funciones conocidas, con $c(\cdot)$ denominada función cumulante (o *log-partition function*). Cuando ϕ es desconocido se trata de una familia exponencial de dos parámetros.

Se dice que la familia está en la forma **canónica** si $b(\theta) = \theta$, es decir, si el parámetro natural aparece linealmente como coeficiente de y en la exponencial.

En este contexto, los GLM permiten modelar una función de la media condicional $\mu = E(Y|\mathbf{X})$ como combinación lineal de las predictoras, de manera que

$$g(\mu) = \mathbf{X}^\top \boldsymbol{\beta},$$

o bien

$$\mu = g^{-1}(\mathbf{X}^\top \boldsymbol{\beta}),$$

donde $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función monótona y diferenciable denominada **función de enlace** o *link*. Si se adopta la forma canónica de la familia exponencial y se elige como función de enlace la identidad $g(\mu) = \mu$, se dice que se utiliza el **enlace canónico**, en cuyo caso el predictor lineal coincide con el parámetro natural: $\theta = \mathbf{X}^\top \boldsymbol{\beta}$.

A lo largo de esta sección se asumirá que $\boldsymbol{\beta}_0$ es el verdadero valor del parámetro y E_0 es el valor esperado bajo el modelo verdadero, de modo que

$$E_0(Y|\mathbf{X}) = g^{-1}(\mathbf{X}^\top \boldsymbol{\beta}_0). \tag{1.2}$$

A continuación se presentan algunos ejemplos de distribuciones conocidas que pertenecen a la familia exponencial canónica.

Ejemplo 1.1.1. La distribución de Poisson $\mathcal{P}(\lambda)$ se puede expresar en la forma de familia exponencial canónica, ya que

$$\begin{aligned} f(y; \lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp\{y \log(\lambda) - \lambda - \log(y!)\}, \end{aligned}$$

donde $\theta = \log(\lambda)$ es el parámetro canónico o natural, $c(\theta) = e^\theta$, $a(\phi) = 1$ y $d(y, \phi) = -\log(y!)$.

La función de enlace canónica es $g(\lambda) = \log(\lambda)$.

Ejemplo 1.1.2. La distribución binomial $\mathcal{B}(m, \pi)$ se puede expresar como miembro de la familia exponencial, ya que

$$\begin{aligned} f(y; m, \pi) &= \binom{m}{y} \pi^y (1 - \pi)^{m-y} \\ &= \exp \left\{ y \log \left(\frac{\pi}{1 - \pi} \right) + m \log(1 - \pi) + \log \binom{m}{y} \right\}. \end{aligned}$$

donde $\theta = \log \left(\frac{\pi}{1 - \pi} \right)$ es el parámetro canónico o natural, $c(\theta) = m \log(1 + e^\theta)$, $a(\phi) = 1$ y $d(y, \phi) = \log \binom{m}{y}$.

En este caso, la función de enlace canónica viene dada por $g(\pi) = \log \left(\frac{\pi}{1 - \pi} \right)$ y se conoce como **función logística** o **logit**. El modelo lineal generalizado resultante se denomina **modelo de regresión logística**.

Ejemplo 1.1.3. La distribución normal $\mathcal{N}(\mu, \sigma^2)$ puede escribirse en la forma de familia exponencial como

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-(y - \mu)^2}{2\sigma^2} \right\} \\ &= \exp \left\{ \frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2} \left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\} \end{aligned}$$

donde $\theta = \mu$, $\phi = \sigma^2$, $c(\theta) = \frac{\theta^2}{2}$, $a(\phi) = \phi$ y $d(y, \phi) = -\frac{1}{2} \left(\frac{y^2}{\phi} + \log(2\pi\phi) \right)$. En este caso, la función de enlace canónica es la identidad, $g(\mu) = \mu$.

1.1.1. Estimación por Máxima Verosimilitud

Dada una muestra aleatoria $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$, donde para cada i se asume que $Y_i | \mathbf{X}_i$ tiene distribución perteneciente a la familia exponencial en forma canónica y sigue el modelo

(1.2) con g el enlace canónico. La función de **log-verosimilitud empírica**, o simplemente función de **log-verosimilitud**, viene dada por

$$\begin{aligned}\ell(\mathbf{Y}; \boldsymbol{\beta}) &= \sum_{i=1}^n \log f(Y_i, g^{-1}(\mathbf{X}_i^\top \boldsymbol{\beta})) \\ &= \sum_{i=1}^n \frac{Y_i \mathbf{X}_i^\top \boldsymbol{\beta} + c(\mathbf{X}_i^\top \boldsymbol{\beta})}{a(\phi)} + d(Y_i, \phi)\end{aligned}\tag{1.3}$$

y el **estimador de máxima verosimilitud** (MLE, por sus siglas en inglés *Maximum Likelihood Estimator*) de $\boldsymbol{\beta}_0$ resulta

$$\widehat{\boldsymbol{\beta}}_n^{ML} \in \operatorname{argmax}_{\boldsymbol{\beta}} \ell(\mathbf{Y}; \boldsymbol{\beta}).$$

Derivando la expresión (1.3) respecto de $\boldsymbol{\beta}$ se obtienen las siguientes ecuaciones de estimación

$$\sum_{i=1}^n \frac{(Y_i - c'(\mathbf{X}_i^\top \boldsymbol{\beta}))}{a(\phi)} \mathbf{X}_i = 0.$$

donde c' es la derivada de la función cumulante c .

Una formulación alternativa y equivalente del problema de estimación por máxima verosimilitud consiste en minimizar

$$\sum_{i=1}^n D(Y_i, \mu_i),$$

donde $D(Y_i, \mu_i)$, conocida comúnmente como **deviance**, está dada por

$$D(Y_i, \mu_i) = 2 \left[\log f(Y_i, \hat{\mu}_i^{\text{sat}}) - \log f(Y_i, \mu_i) \right],$$

con $\hat{\mu}_i^{\text{sat}}$ el valor que maximiza la verosimilitud en el modelo saturado (es decir, que reproduce exactamente Y_i), y $\mu_i = g^{-1}(\mathbf{X}_i^\top \boldsymbol{\beta})$ el valor bajo el GLM propuesto.

Ejemplo 1.1.4. En el caso de regresión de Poisson con enlace logarítmico, se asume que $Y_i | \mathbf{X}_i \sim \mathcal{P}(\lambda_i)$, donde

$$\log(\lambda_i) = \mathbf{X}_i^\top \boldsymbol{\beta}.$$

Las ecuaciones de estimación para el MLE de $\boldsymbol{\beta}$ son

$$\sum_{i=1}^n (Y_i - e^{\mathbf{X}_i^\top \boldsymbol{\beta}}) \mathbf{X}_i = 0.$$

La deviance, en este caso, se puede calcular explícitamente como

$$D(Y_i, \lambda_i) = 2 \left[Y_i \log \left(\frac{Y_i}{\lambda_i} \right) - (Y_i - \lambda_i) \right],$$

donde $\lambda_i = e^{\mathbf{X}_i^\top \boldsymbol{\beta}}$ es el valor bajo el modelo propuesto. Cuando $Y_i = 0$ se conviene tomar $Y_i \log(Y_i/\lambda_i) = 0$.

1.1.2. Métodos de Estimación Robusta

Si bien el MLE es eficiente bajo el supuesto de que las observaciones siguen el modelo especificado, su desempeño puede deteriorarse significativamente ante la presencia de valores atípicos, tanto en la variable respuesta como en las covariables. Por esta razón, la estimación robusta en el contexto de GLM ha sido extensamente estudiada. En esta sección se describen aquellas propuestas que resultan más relevantes para los objetivos de esta tesis.

1.1.2.1. M-estimadores

El método de estimación por mínimos cuadrados es ampliamente utilizado por su simplicidad y eficiencia bajo supuestos ideales, como la normalidad de los errores. En esencia, consiste en estimar los parámetros desconocidos minimizando la suma de los cuadrados de las diferencias entre los valores observados y los valores ajustados por el modelo. Por ejemplo, en el caso más simple de un modelo de locación, dada una muestra aleatoria $\{Y_i\}_{i=1}^n$, se estima el parámetro de locación μ minimizando $\sum_{i=1}^n (Y_i - \mu)^2$, lo cual conduce a la media muestral $\hat{\mu} = \sum_{i=1}^n Y_i/n$. Sin embargo, este procedimiento no es robusto frente a observaciones atípicas, ya que estas pueden ejercer una influencia desproporcionada sobre el estimador. Para superar esta limitación, Huber (1964) propuso una generalización del método de mínimos cuadrados que consiste en minimizar una función alternativa de los residuos que resulte menos sensible a valores extremos. En particular, propuso obtener estimadores del parámetro de locación minimizando respecto a μ la siguiente expresión

$$\frac{1}{n} \sum_{i=1}^n \phi(Y_i - \mu),$$

donde ϕ se denomina **función de pérdida**. Nótese que si se asume que $Y \sim F$, dicha expresión se obtiene al evaluar el funcional

$$\int \phi(Y - \mu) dF$$

en la distribución empírica \hat{F} de los datos, que asigna peso $1/n$ a cada observación.

Esta idea dió origen a los denominados **M-estimadores**, una extensa clase que incluye, por ejemplo, a la media muestral cuando $\phi(t) = t^2$, a la mediana muestral cuando $\phi(t) = |t|$ y a los estimadores de máxima verosimilitud cuando $\phi(t) = -\log f(t)$, siendo f la densidad asumida para los datos.

La Tabla 1.1 presenta algunos ejemplos de funciones de pérdida robustas. Sus gráficas se pueden observar en la Figura 1.1. Cada función ϕ representa una estrategia diferente para limitar la influencia de observaciones atípicas. Por ejemplo, la función de Huber se comporta de manera similar a la pérdida cuadrática cerca del origen y comienza a

aplanarse a partir del umbral c . La pérdida Bicuadrada de de Tukey es redescendiente, es decir, crece de forma similar a la cuadrática en un entorno del origen pero es constante para observaciones alejadas, eliminando por completo su influencia, mientras que la función de Cauchy atenúa la influencia de las observaciones alejadas del origen de forma continua. El rol de la constante c en las funciones ϕ es controlar la sensibilidad del estimador a los valores atípicos y determinar la compensación (*trade-off*) entre eficiencia y robustez del estimador: valores mayores de c mejoran la eficiencia a costa de la robustez, y valores menores aumentan la robustez pero reducen la eficiencia. Como c está ligada a la robustez del estimador, determina también su **punto de ruptura**, que corresponde a la proporción máxima de datos contaminados que el estimador puede tolerar sin que su estimación se vea afectada significativamente.

Nombre	Función de pérdida
Huber	$\phi(u) = \begin{cases} u^2, & u \leq c, \\ 2c u - c^2, & u > c \end{cases}$
Bicuadrada (Tukey)	$\phi(u) = \begin{cases} \frac{c^2}{6} \left[1 - \left(1 - \left(\frac{u}{c} \right)^2 \right)^3 \right], & \text{si } u \leq c, \\ \frac{c^2}{6}, & \text{si } u > c. \end{cases}$
Cauchy	$\phi(u) = \log\left(1 + \left(\frac{u}{c}\right)^2\right)$

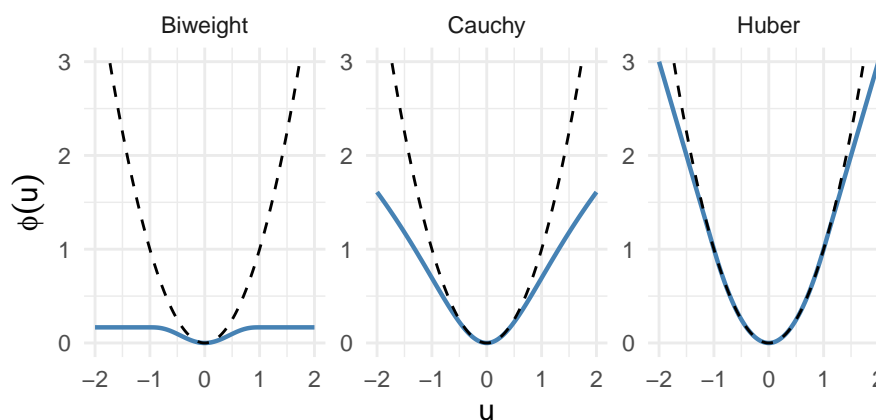
Cuadro 1.1: Funciones de pérdida $\phi(u)$.

Figura 1.1: Funciones de pérdida $\phi(u)$ con $c = 1$ (curvas azules) junto a la pérdida cuadrática (curvas punteadas).

A partir de esta idea, se han desarrollado diversos enfoques robustos, incluyendo extensiones específicas para GLM. Por ejemplo, Pregibon (1982) propuso un método para obtener estimadores robustos de la regresión logística, que se basa en minimizar una función ro-

busta de la deviance. Específicamente,

$$\widehat{\beta}_n = \operatorname{argmin}_{\beta} \sum_{i=1}^n \phi\left(D(Y_i, g^{-1}(\mathbf{X}_i^{\top} \beta))\right). \quad (1.4)$$

Esta modificación atenúa la contribución de observaciones con valores elevados de deviance, reduciendo así su influencia sobre el estimador resultante. No obstante, Bianco and Yohai (1996) mostraron que los estimadores definidos por (1.4) están asintóticamente sesgados, es decir, que cuando el tamaño muestral crece no convergen al verdadero valor del parámetro. Los autores definieron entonces una nueva clase de M-estimadores para el modelo de regresión logística. La propuesta consiste en obtener los estimadores minimizando en β la expresión

$$\sum_{i=1}^n \rho(Y_i, g^{-1}(\mathbf{X}_i^{\top} \beta)), \quad (1.5)$$

donde

$$\rho(y, \mu) = \phi(D(y, \mu)) + G(\mu)$$

y

$$G(\mu) = \int_0^{\mu} \phi'(-\ln(s)) ds.$$

La diferencia entre (1.4) y (1.5) radica en el término $G(g^{-1}(\mathbf{X}_i^{\top} \beta))$ que se incluye para asegurar que se cumpla, para todo β , la siguiente condición

$$E_{\beta}(\Psi(Y, \mathbf{X}^{\top} \beta) | \mathbf{X}) = 0,$$

donde $\Psi(Y, \mathbf{X}^{\top} \beta) = \frac{\partial \rho(Y, \mathbf{X}^{\top} \beta)}{\partial \beta}$. Esta propiedad, denominada **consistencia de Fisher condicional**, es ampliamente reconocida en la literatura de estimación robusta como un criterio esencial que debe satisfacer un funcional de estimación.

Los autores utilizaron la función

$$\phi(u) = \begin{cases} u - \frac{u^2}{2c} & \text{si } u \leq c \\ c/2 & \text{si } u > c \end{cases}$$

donde c , a ser determinada, es la constante que controla la compensación entre robustez y eficiencia del estimador resultante.

Posteriormente Croux and Haesbroeck (2002) complementaron el estudio de Bianco y Yohai al proponer una nueva expresión para ϕ

$$\phi(u) = \begin{cases} ue^{-\sqrt{c}}, & |u| \leq c \\ e^{-\sqrt{c}}(2(1 + \sqrt{c}) + c) - 2e^{-\sqrt{u}}(1 + \sqrt{u}), & |u| > c \end{cases}$$

y derivar criterios para la existencia del estimador en muestras finitas.

Por otra parte, Bergesio and Yohai (2011) presentaron una propuesta de estimación robusta para los GLM basada en la clase de estimadores de proyección (P-estimadores) para modelos lineales. Estos estimadores son altamente robustos pero no son asintóticamente normales. Por lo tanto, los autores propusieron utilizar M-estimadores a un paso usando como estimadores iniciales los P-estimadores. Los estimadores a un paso son asintóticamente normales y conservan las propiedades de robustez de los P-estimadores. Una desventaja de esta propuesta es que su cálculo requiere algoritmos de alta complejidad.

Los estimadores propuestos por Bianco and Yohai (1996) fueron extendidos por Bianco et al. (2012) al contexto de respuestas faltantes y aplicados a los modelos de regresión de Poisson y Gamma. Más precisamente, en el caso de regresión de Poisson, las autoras propusieron obtener estimadores robustos minimizando

$$\sum_{i=1}^n \left\{ \phi \left(\exp(\mathbf{X}_i^\top \boldsymbol{\beta}) - Y_i (\exp(\mathbf{X}_i^\top \boldsymbol{\beta}) + 1 - \ln Y_i) \right) + G(\exp(\mathbf{X}_i^\top \boldsymbol{\beta})) \right\} \omega(\mathbf{X}_i),$$

donde la función ω penaliza las observaciones extremas en las covariables. Denotando $\varphi = \phi'$, las autoras mostraron que para este caso particular la función G tiene la forma específica

$$G(s) = - \int \varphi(s) e^{-s} ds + \sum_{j=1}^{\infty} \int \varphi(s - j(\ln(s) + 1 - \ln j)) \frac{(s^j e^{-s})}{j!} \frac{(j-s)}{s} ds.$$

Por otro lado, Salibian-Barrera et al. (2016) presentaron estadísticos robustos de prueba de razón de verosimilitudes, estimando de manera robusta los parámetros en modelos lineales. Para ello se basaron en los τ -estimadores de regresión, que combinan alta robustez y alta eficiencia. Para computar dichos estimadores utilizaron la familia de funciones introducida por Muler and Yohai (2002), que se define por

$$\phi(u) = \begin{cases} 1.38 \left(\frac{u}{c}\right)^2 & \left|\frac{u}{c}\right| \leq \frac{2}{3} \\ 0.55 - 2.69 \left(\frac{u}{c}\right)^2 + 10.76 \left(\frac{u}{c}\right)^4 - 11.66 \left(\frac{u}{c}\right)^6 + 4.04 \left(\frac{u}{c}\right)^8 & \frac{2}{3} < \left|\frac{u}{c}\right| \leq 1 \\ 1, & \left|\frac{u}{c}\right| > 1. \end{cases}$$

La ventaja que presenta esta familia es que se puede introducir a distintos estimadores robustos, manejando la constante c de manera de obtener estimadores con punto de ruptura 0.5 y con una eficiencia cercana al 100 % en comparación con el estimador de mínimos cuadrados, cuando los errores son normalmente distribuidos. En particular, esta familia ha demostrado proporcionar τ -estimadores con propiedades de robustez superiores a las obtenidas mediante la función bicuadrada de Tukey clásica.

1.1.2.2. MT-estimadores

Valdora and Yohai (2014) propusieron una familia de estimadores robustos para GLM, construidos a partir de M-estimadores redescendentes aplicados a respuestas transforma-

das. Estos, que en adelante se denominarán **MT-estimadores**, buscan mejorar la robustez frente a observaciones atípicas tanto en la variable respuesta como en las covariables.

Una limitación de los M-estimadores tradicionales es que consideran un único parámetro de escala común a todas las observaciones. Sin embargo, en GLM, la dispersión de la variable respuesta puede variar entre observaciones y, además, depender de los parámetros del modelo. Para resolver esta dificultad, los autores propusieron transformar las respuestas mediante una función $t(Y)$ diseñada para estabilizar la varianza, es decir, hacer que $Var(t(Y))$ dependa poco o no dependa en absoluto del parámetro de la distribución y, por lo tanto, de β . Esto permite aplicar procedimientos de estimación robusta que asuman varianza homogénea, facilitando la implementación de M-estimadores sobre los datos transformados.

Formalmente, sea Y una variable aleatoria con distribución F_{λ_0} , con media $\mu(\lambda_0) = \mu_0$ y varianza $v(\lambda_0)$. Al considerar la transformación

$$t(y) = \int_0^y \frac{1}{[v(\mu^{-1}(u))]^{1/2}} du, \quad (1.6)$$

realizando un desarrollo de Taylor de primer orden (Teorema A.1.5 del Apéndice) se tiene que

$$t(y) \approx t(\mu_0) + t'(\mu_0)(y - \mu_0)$$

con $t'(\mu_0) = \frac{1}{[v(\mu^{-1}(\mu_0))]^{1/2}}$. Por lo tanto

$$\begin{aligned} Var(t(Y)) &\approx t'(\mu_0)^2 Var(Y) \\ &= \frac{1}{v(\mu^{-1}(\mu_0))} Var(Y) = 1. \end{aligned}$$

Así, la transformación (1.6) logra que la varianza de la variable transformada sea aproximadamente constante.

En algunas distribuciones en particular se tiene que $V(\lambda) = \mu(\lambda)^k$, en cuyo caso la transformación (1.6) se reduce, salvo una constante, a

$$t(y) = \begin{cases} y^{-(k/2)+1}, & \text{si } k \neq 2 \\ \log(y), & \text{si } k = 2. \end{cases} \quad (1.7)$$

Ejemplo 1.1.5. *En la regresión de Poisson se tiene que $Var(Y_i) = \lambda_i = e^{\mathbf{X}_i^\top \beta}$, de modo que la varianza cambia con \mathbf{X}_i y depende directamente de β . En este caso la transformación dada en (1.7) resulta $t(y) = \sqrt{y}$. En la Figura 1.2 se grafica $Var(t(Y))$ como función de λ . Se puede observar que la varianza es aproximadamente constante, salvo para valores pequeños de λ .*

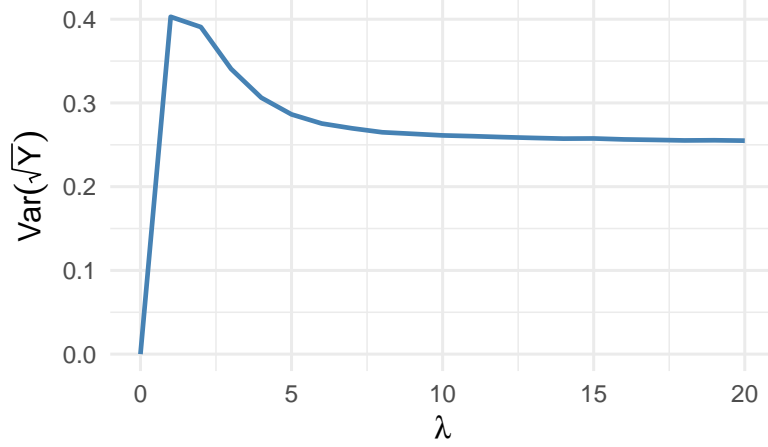


Figura 1.2: Varianza de \sqrt{Y} con $Y \sim \mathcal{P}(\lambda)$, para diferentes valores de λ .

Para introducir los MT-estimadores, los autores consideran una transformación $t : \mathbb{R} \rightarrow \mathbb{R}$ tal que la varianza de $t(Y)$ es casi constante. Sea $\phi : \mathbb{R} \rightarrow \mathbb{R}$ una función continua y acotada con un único mínimo en 0 y $h(\lambda)$ definida como sigue

$$h(\lambda) = \underset{u}{\operatorname{argmin}} E_{\lambda}(\phi(t(Y) - u)).$$

Si $h(\lambda)$ es continua y está unívocamente definida para todo λ , dada una muestra aleatoria $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ independiente e idénticamente distribuidas como $Y|\mathbf{X} \sim F_{\lambda_0}$ que sigue el modelo (1.2), el M-estimador pesado basado en transformaciones (WMT-estimador) de β_0 viene dado por

$$\hat{\beta}_n = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \phi(t(Y_i) - m(g^{-1}(\mathbf{X}_i^{\top} \beta))) \omega(\mathbf{X}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \quad (1.8)$$

con $\hat{\boldsymbol{\mu}}$ y $\hat{\boldsymbol{\Sigma}}$ estimadores consistentes y robustos de posición y matriz de covarianza de \mathbf{X} basados en $\mathbf{X}_1, \dots, \mathbf{X}_n$ y ω es una función de la distancia de Mahalanobis cuyo propósito es penalizar observaciones de alta influencia y se define por

$$\omega(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = w \left(\sqrt{(\mathbf{X} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})} \right),$$

con w una función no negativa y no creciente. Cuando $\omega(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv 1$ el estimador obtenido se denomina MT-estimador.

Notar que los WMT-estimadores resultan naturalmente consistentes Fisher, debido a que la esperanza

$$E_0[\phi(t(Y) - m(g(\mathbf{X}^{\top} \beta))) | \mathbf{X}]$$

se minimiza cuando $\beta = \beta_0$ para toda \mathbf{X} , entonces

$$E_0[\phi(t(Y) - m(g^{-1}(\mathbf{X}^{\top} \beta))) w(\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)]$$

también se minimiza cuando $\beta = \beta_0$.

En relación a la teoría asintótica del estimador definido en (1.8), los autores demostraron bajo ciertos supuestos que $\hat{\beta}_n$ es un estimador fuertemente consistente de β_0 , esto es,

$$\hat{\beta}_n \xrightarrow{c.s.} \beta_0$$

y que

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \Sigma),$$

cuando $n \rightarrow \infty$. Además hallaron una expresión y una estimación consistente para la matriz de covarianza Σ .

1.2. Modelos Parcialmente Lineales Generalizados

Si bien los modelos paramétricos se destacan por su simpleza e interpretabilidad, los supuestos en los que se basan pueden resultar demasiado estrictos para algunas situaciones prácticas, dificultando su aplicación. En contraste, los **modelos de regresión no paramétricos** no realizan suposiciones a priori sobre la función de regresión ν en (1.1) y, por lo tanto, sobre la distribución subyacente de los datos. Por este motivo resultan más flexibles, permitiendo su aplicación a un mayor número de problemas. No obstante, debido a que usan la misma información para realizar estimaciones en dimensión infinita, los métodos no paramétricos producen resultados estadísticamente menos eficientes que los paramétricos. Otra desventaja es que cuando el número de variables predictoras crece, sufren de un problema denominado **maldición de la dimensión**, ya que la cantidad de datos necesarios para que los estimadores tengan un grado de precisión considerable aumenta mucho más rápido que la cantidad de variables involucradas en el modelo. Engle et al. (1986) combinó ambos enfoques para analizar la relación entre el consumo de electricidad y la temperatura, dando origen a los **modelos de regresión semiparamétricos**. Los mismos se caracterizan por reunir la interpretabilidad de los modelos paramétricos con la flexibilidad de los no paramétricos, mitigando el problema de la maldición de la dimensión al incluir covariables de forma paramétrica. Los **modelos parcialmente lineales** son un caso particular de los modelos semiparamétricos, en los cuales la componente paramétrica tiene una estructura lineal en los parámetros.

Los **modelos parcialmente lineales generalizados** (GPLM, por sus siglas en inglés *generalized partially linear models*) flexibilizan los GLM permitiendo incorporar covariables (usualmente continuas) mediante un término no paramétrico, e incluyendo las demás de manera lineal. Esta flexibilidad mejora la capacidad del modelo para capturar de manera más precisa la relación entre la variable respuesta y las predictoras.

Formalmente, sea $T \in \mathbb{R}$ una variable predictora adicional, $Y|(\mathbf{X}, T) \sim F$ sigue un modelo parcialmente lineal generalizado si

$$\mu = E(Y|(\mathbf{X}, T)) = g^{-1}(\mathbf{X}^\top \boldsymbol{\beta} + m(T)) \quad (1.9)$$

donde g es una función de enlace conocida, $\boldsymbol{\beta}$ es un vector de parámetros a ser estimado, $m : \mathbb{R} \rightarrow \mathbb{R}$ una función suave desconocida y, además, se considera que la varianza se puede escribir como función de la media, $Var(Y|(\mathbf{X}, T)) = V(\mu)$.

A lo largo de esta sección se asumirá que $m_0(T)$ y $\boldsymbol{\beta}_0$ son los verdaderos valores de los parámetros y E_0 es el valor esperado bajo el modelo verdadero, de modo que

$$E_0(Y|(\mathbf{X}, T)) = g^{-1}(m_0(T) + \mathbf{X}^\top \boldsymbol{\beta}_0).$$

En algunos casos se tiene que $\mathbf{T} \in \mathbb{R}^s$ y, aunque algunas variables estén modeladas de manera lineal, el modelo sigue expuesto a la maldición de la dimensión cuando s crece. Una opción factible para disminuir la dimensionalidad del problema es que la componente no paramétrica dependa de \mathbf{T} a través de una proyección, de manera que $\boldsymbol{\alpha}^\top \mathbf{T} \in \mathbb{R}$, donde $\boldsymbol{\alpha} \in \mathbb{R}^s$ es un parámetro desconocido a ser estimado. Esto da lugar los **GPLM de índice único** propuestos por Carroll et al. (1997), cuya expresión formal resulta

$$E(Y|(\mathbf{X}, \mathbf{T})) = g^{-1}(\mathbf{X}^\top \boldsymbol{\beta} + m(\boldsymbol{\alpha}^\top \mathbf{T})).$$

La estimación de los GPLM presenta algunos desafíos metodológicos, dado que el término paramétrico y el no paramétrico se encuentran acoplados mediante la función de enlace, que es generalmente no lineal. Esta estructura impone la necesidad de combinar distintos enfoques de estimación: mientras que los parámetros lineales suelen requerir métodos paramétricos de estimación, como máxima verosimilitud, la estimación de la función m requiere técnicas no paramétricas como el suavizado por núcleos o la utilización de funciones spline. En consecuencia, el proceso de estimación resulta computacionalmente costoso, como remarcan Boente and Rodriguez (2010) en su trabajo.

1.2.1. Perfiles de Verosimilitud

El método de **perfiles de verosimilitud**, introducido por Severini and Wong (1992) y posteriormente aplicado a GPLM por Severini and Staniswalis (1994), consiste en reformular la verosimilitud como una función de los parámetros lineales, reemplazando la componente no paramétrica por un estimador que depende de dichos parámetros. Esta estrategia permite construir un perfil de verosimilitud que puede ser optimizado respecto de los parámetros, facilitando su estimación mediante técnicas paramétricas específicas.

Más precisamente, en el contexto de GPLM, se obtiene en primer lugar una estimación de m como función de $\boldsymbol{\beta}$ mediante un método de suavizado apropiado. Esta etapa implica

resolver un problema no paramétrico. Posteriormente, dicha estimación se incorpora en la construcción del perfil de verosimilitud, que depende exclusivamente de β . Finalmente, la estimación de β se obtiene maximizando el perfil construido, lo que constituye un problema puramente paramétrico. En términos formales, para cada valor fijo de la variable T se define el funcional

$$S(\eta, \beta, t) = E \left(-\log f(Y, g^{-1}(\mathbf{X}^\top \beta + \eta)) | T = t \right), \quad (1.10)$$

donde $\eta = m(t) \in \mathbb{R}$. Dada una muestra aleatoria independiente $\{(Y_i, \mathbf{X}_i, T_i)\}_{i=1}^n$ de manera que $Y | (\mathbf{X}, T)$ sigue el modelo (1.9), un estimador de (1.10) basado en la muestra se puede obtener generalizando enfoque de verosimilitud ponderada de Staniswalis (1989), como sigue

$$S_n(\eta, \beta, t) = \sum_{i=1}^n -W_{i,n}(t) \log f(Y_i, g^{-1}(\mathbf{X}_i^\top \beta + \eta)),$$

donde $W_{i,n}$ es la **función peso** de Nadaraya-Watson estudiada por Nadaraya (1964), Watson (1964) y Nadaraya (1965). La misma tiene la forma

$$W_{i,n}(t) = \frac{K\left(\frac{t-T_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{t-T_j}{h_n}\right)}, \quad (1.11)$$

con $h_n > 0$ que se denomina **parámetro de suavizado** y $K : \mathbb{R} \rightarrow \mathbb{R}_0^+$ se conoce como **función núcleo** o *kernel* y verifica $\int K(u)du = 1$ y $\int uK(u)du = 0$. Los núcleos más utilizados se detallan en la Tabla 1.2 y se ilustran en la Figura 1.3.

Nombre	Función núcleo
Uniforme	$K(x) = \frac{1}{2} \mathbb{I}_{[-1,1]}(x)$
Triangular	$K(x) = (x+1) \mathbb{I}_{[-1,0]}(x) + (1-x) \mathbb{I}_{[0,1]}(x)$
Epanechnikov (cuadrático)	$K(x) = \frac{3}{4}(1-x^2) \mathbb{I}_{[-1,1]}(x)$
Gaussiano	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$

Cuadro 1.2: Funciones núcleo $K(x)$.

La extensión de los núcleos al caso multivariado puede ser definida combinando un núcleo en \mathbb{R} con una norma $\|\cdot\|$ en \mathbb{R}^s , es decir

$$\tilde{K}(\mathbf{x}) = K(\|\mathbf{x}\|), \quad \mathbf{x} \in \mathbb{R}^s.$$

La idea central del enfoque basado en una verosimilitud ponderada es que si T_j está próximo a t entonces tanto $m(T_j)$ como Y_j y \mathbf{X}_j aportan información relevante sobre $m(t)$. Así, para cada β y t fijos se define

$$\hat{m}_\beta(t) = \operatorname{argmin}_{\eta \in \mathbb{R}} S_n(\eta, \beta, t). \quad (1.12)$$

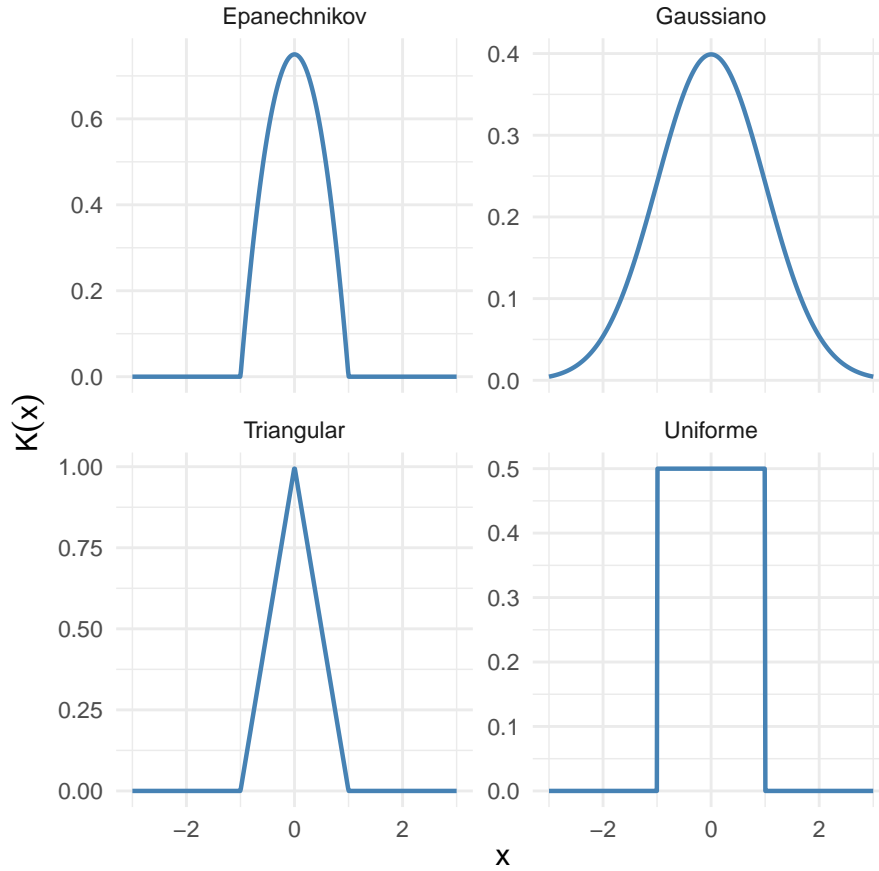


Figura 1.3: Funciones núcleo $K(x)$.

Por otro lado, la función de verosimilitud teórica asociada al modelo puede expresarse mediante el funcional

$$F(\boldsymbol{\beta}, m) = E \left[-\log f \left(Y, g^{-1} \left(\mathbf{X}^\top \boldsymbol{\beta} + m(T) \right) \right) \right].$$

Su versión empírica se define para cada función continua $m : \mathbb{R} \rightarrow \mathbb{R}$ como

$$F_n(\boldsymbol{\beta}, v) = -\frac{1}{n} \sum_{i=1}^n \log f \left(Y_i, g^{-1} \left(\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i) \right) \right). \quad (1.13)$$

Una vez obtenida la estimación de la componente no paramétrica para un valor fijo de $\boldsymbol{\beta}$ como en (1.12), se construye el siguiente perfil de verosimilitud empírico

$$F_n(\boldsymbol{\beta}, \hat{m}_\beta) = -\frac{1}{n} \sum_{i=1}^n \log f \left(Y_i, g^{-1} \left(\mathbf{X}_i^\top \boldsymbol{\beta} + \hat{m}_\beta(T_i) \right) \right).$$

Posteriormente, el estimador del parámetro $\boldsymbol{\beta}_0$ se computa como

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} F_n(\boldsymbol{\beta}, \hat{m}_\beta). \quad (1.14)$$

Cabe señalar que la optimización de (1.13) con respecto a β puede ser computacionalmente costosa, ya que la estimación de la componente no paramétrica \hat{m}_β depende del propio parámetro β , lo que introduce una estructura anidada en el procedimiento de estimación.

Cuando la función de distribución condicional de $Y | (\mathbf{X}, T)$ no está especificada completamente, es posible adaptar el procedimiento anterior utilizando una **función de cuasi-verosimilitud** en lugar de la función de verosimilitud. Esta alternativa permite construir estimadores en contextos donde solo se especifica la media condicional $E(Y | (\mathbf{X}, T)) = \mu$ y la varianza $Var(Y | (\mathbf{X}, T)) = V(\mu)$ como función conocida de la media. En este caso, la función $\log f(y, \mu)$ en (1.13) se reemplaza por la cuasiverosimilitud

$$Q(y, \mu) = \int_{\mu}^y \frac{s - y}{V(s)} ds.$$

Severini and Staniswalis (1994) demostraron, bajo ciertas condiciones de regularidad, que el estimador definido en (1.12) utilizando la cuasiverosimilitud satisface

$$\sup_t |\hat{m}_\beta(t) - m_0(t)| = o_p(n^{-1/4}).$$

Además probaron la normalidad asintótica del estimador que se obtiene al utilizar la cuasiverosimilitud en (2.12).

1.2.2. Métodos de Estimación Robusta

Los estimadores basados en la verosimilitud son conocidos por ser eficientes pero altamente sensibles a observaciones atípicas, tanto en la variable respuesta como en las predictoras. En el contexto de los GPLM, los valores extremos pueden afectar a la estimación de los parámetros lineales, así como a la forma y la escala de la estimación de la componente no paramétrica, propagando dicho efecto a la estimación de los parámetros lineales. Esto motivó el desarrollo de enfoques de estimación robusta con el objetivo de mitigar el impacto de observaciones extremas en los GPLM. En esta sección se presentan dos de ellos.

1.2.2.1. Estimadores Robustos basados en Perfiles de Verosimilitud Generalizados

Boente et al. (2006) propusieron una familia de estimadores robustos para GPLM generalizando la propuesta de Severini and Staniswalis (1994). Más precisamente, sea $\rho(y, \lambda)$ una función de pérdida robusta, se definen los funcionales

$$S(\eta, \beta, t) = E \left[\rho(Y, \mathbf{X}^\top \beta + \eta) \omega_1(\mathbf{X}) | T = t \right]$$

$$F(\beta, v) = E \left[\rho(Y, \mathbf{X}^\top \beta + v(T)) \omega_2(\mathbf{X}) \right],$$

donde ω_1 y ω_2 son funciones (posiblemente iguales) que dan menos peso a las observaciones con \mathbf{X} de alto leverage. Sus respectivas versiones muestrales resultan

$$S_n(\eta, \boldsymbol{\beta}, t) = \sum_{i=1}^n W_{i,n}(t) \rho(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + \eta) \omega_1(\mathbf{X}_i) \quad (1.15)$$

$$F_n(\boldsymbol{\beta}, v) = \frac{1}{n} \sum_{i=1}^n \rho(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + v(t_i)) \omega_2(\mathbf{X}_i), \quad (1.16)$$

donde $W_{i,n}$ son los pesos de tipo núcleo definidos en (1.11).

Luego, el procedimiento de estimación robusta consiste en dos pasos:

(Paso 1). Para cada $T = t$ y $\boldsymbol{\beta}$ fijos, hallar

$$\hat{m}_\beta(t) = \underset{\eta \in \mathbb{R}}{\operatorname{argmin}} S_n(\eta, \boldsymbol{\beta}, t).$$

(Paso 2) El estimador de $\boldsymbol{\beta}$ se define como

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} F_n(\boldsymbol{\beta}, \hat{m}_\beta)$$

y el estimador de $m(t)$ resulta $\hat{m}_{\hat{\boldsymbol{\beta}}_n}(t)$.

Los autores utilizan dos propuestas para la función ρ , la primera de ellas consiste en una modificación de la función de verosimilitud y tiene como objetivo acotar la deviance, mientras que la segunda, propuesta por Cantoni and Ronchetti (2001), es una versión robustificada de la cuasiverosimilitud y tiene como objetivo acotar los residuos de Pearson.

Los autores demostraron bajo ciertas condiciones de regularidad que

$$\sup_{\boldsymbol{\beta}} \|\hat{m}_\beta - m_\beta\|_{0,\infty} \xrightarrow{c.s.} 0 \quad \text{y} \quad \hat{\boldsymbol{\beta}}_n \xrightarrow{c.s.} \boldsymbol{\beta}_0.$$

Además probaron que

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma),$$

y hallaron una expresión para Σ .

1.2.2.2. Estimadores Robustos de Tres Pasos

La principal desventaja de los estimadores propuestos por Boente et al. (2006) es que, dado que se basan en un enfoque de perfiles de verosimilitud generalizados, sus matrices de covarianza asintóticas dependen de las derivadas de la función m_β con respecto a $\boldsymbol{\beta}$, lo que dificulta su estimación. Además, en la práctica, el procedimiento de estimación descrito en la Sección 1.2.2.1 puede implicar un costo computacional considerable, dado que la

componente no paramétrica \widehat{m}_β debe ser calculada para cada valor fijo $T = t$ y para cada candidato de β en una grilla dada, para luego poder construir el perfil robusto (1.16) y seleccionar el valor de la grilla de parámetros que lo minimiza. Esta interdependencia entre ambas etapas genera una estructura iterativa compleja que encarece computacionalmente el procedimiento. La propuesta de Boente and Rodriguez (2010) aborda esta dificultad al disminuir el costo computacional manteniendo buenas propiedades de robustez. Las autoras propusieron una versión robustificada del estimador considerado en Carroll et al. (1997). Específicamente, el procedimiento consiste en obtener el estimador de la componente no paramétrica minimizando la expresión (1.15) simultáneamente respecto a η y β en el Paso 1, es decir, para cada valor fijo de $T = t$, hallar

$$(\tilde{\beta}_n(t), \tilde{m}_n(t)) = \underset{\eta \in \mathbb{R}, \beta \in \mathbb{R}^p}{\operatorname{argmin}} S_n(\eta, \beta, t).$$

Luego, en el Paso 2 se optimiza la expresión (1.16) respecto a β , utilizando la estimación de la componente no paramétrica \tilde{m} obtenida en el Paso 1, es decir

$$\widehat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} F_n(\beta, \tilde{m}_n).$$

Si el interés se centra en estimar β y la componente no paramétrica se considera un parámetro de ruido, aquí podría finalizar el proceso, tomando a $\tilde{m}_\beta(t)$ como estimador de $m_0(t)$. Si es de interés estimar $m_0(t)$ con más precisión, se puede realizar un tercer paso en el que se actualice su estimación utilizando $\widehat{\beta}_n$ como sigue

$$\widehat{m}_n(t) = \underset{\eta}{\operatorname{argmin}} S_n(\eta, \widehat{\beta}_n, t).$$

Las autoras aseguran que este procedimiento resulta computacionalmente más eficiente que el discutido en la Sección 1.2.2.1, ya que en el primer paso no es necesario minimizar S_n para una grilla de valores de β de manera de obtener \widehat{m}_β para cada uno de ellos.

En el Lema 3.1 de Boente and Rodriguez (2010), bajo condiciones de regularidad sobre las funciones de pérdida, los pesos, la función núcleo y la densidad marginal de T , así como supuestos de continuidad y equicontinuidad sobre las funciones involucradas, las autoras probaron la convergencia uniforme casi segura sobre un conjunto compacto tanto del estimador no paramétrico \tilde{m}_n hacia la función verdadera m_0 , como de $\tilde{\beta}_n$ hacia el parámetro verdadero β_0 . Asimismo en el Teorema 3.1, bajo condiciones adicionales sobre la familia de funciones $\{\rho(Y, \mathbf{X}^T \beta + \eta) w_1(\mathbf{X}), \beta \in \mathbb{R}^p, \eta \in \mathbb{R}\}$, mostraron que el funcional $F_n(\beta, \tilde{m}_n)$ converge uniformemente al funcional poblacional $F(\beta, m_0)$. En consecuencia, si este último posee un único mínimo en β_0 , el estimador robusto $\widehat{\beta}_n$ converge casi seguramente a β_0 .

Por otro lado, bajo condiciones de regularidad que aseguran la suavidad y acotación de las funciones involucradas, así como la no singularidad de ciertas matrices fundamentales,

las autoras establecen, en el Teorema 4.1, la normalidad asintótica del estimador del parámetro de regresión $\hat{\beta}_n$. En particular, se requiere que la función de pérdida sea tres veces diferenciable con derivadas continuas y acotadas, que las funciones de pesos y densidad marginal de T posean buenas propiedades de diferenciability y positividad, y que los núcleos utilizados satisfagan condiciones estándar, sean Lipchitz y de variación acotada. Bajo estas hipótesis, junto con la consistencia uniforme de los estimadores \tilde{m}_n y $\tilde{\beta}_n$, las autoras demostraron que el estimador robusto $\hat{\beta}_n$ es asintóticamente normal, con media nula y hallaron una expresión para la matriz de covarianza.

1.2.3. Parámetro de Suavizado

En modelos que incluyen una componente no paramétrica, la elección adecuada del parámetro de suavizado h_n , involucrado en los pesos definidos en (1.11), resulta crucial para lograr una estimación precisa de dicha componente. Este parámetro regula el grado de suavizado aplicado sobre los datos: valores grandes de h_n conducen a una estimación demasiado suave que puede ocultar estructuras relevantes de la función subyacente. Mientras que si h_n es demasiado pequeño se captura el ruido presente en los datos, lo cual resulta en estimaciones muy variables. Por lo tanto, seleccionar un valor apropiado de h_n es fundamental para garantizar una estimación adecuada de la componente no paramétrica.

En el escenario clásico de regresión no paramétrica uno de los procedimientos más utilizados para seleccionar h_n es el método de **validación cruzada** (o *cross validation*), introducida por Rudemo (1982) y Bowman (1984). Este método consiste en elegir el valor de h_n que minimiza la función de validación cruzada que es una medida del error de predicción.

La idea central es evaluar el desempeño predictivo del estimador en subconjuntos de datos no utilizados para su construcción. En la práctica, se emplean principalmente dos variantes de este procedimiento: *k-fold cross validation* y *leave-one-out cross validation*, que se detallan a continuación.

- I. Se divide la muestra aleatoriamente en k subconjuntos (folds) de tamaño aproximadamente igual.
- II. Para cada fold $j = 1, \dots, k$, se calcula el estimador sobre los $k - 1$ subconjuntos restantes. Luego se calculan los errores a través de una función de validación adecuada, evaluada en el fold excluido.
- III. Se promedian los errores obtenidos sobre los k folds.
- IV. Se repite el proceso para distintos valores de h_n en una grilla, y se selecciona como h_n óptimo aquel que minimiza el promedio de errores obtenidos.

El enfoque de leave-one-out consiste en elegir $k = 1$ en el procedimiento descrito anteriormente, es decir, computar los estimadores con todas las observaciones menos una, y luego medir el error de predicción para esa observación.

Una de las funciones de validación más utilizadas es el error cuadrático medio, que consiste en promediar los errores cuadráticos de predicción. Wang and Scott (1994) analizaron el comportamiento de dicha función en regresión no paramétrica y advirtieron que, en presencia de valores atípicos, la misma pierde utilidad como criterio para elegir h_n . Esto se debe a que un solo valor atípico puede generar un error cuadrático desproporcionado que domina la función de validación, llevándola a ser prácticamente constante en todo su dominio y, en consecuencia, inútil para identificar un valor óptimo de suavizado. Los autores propusieron utilizar una función de validación que reemplaza el error cuadrático por el valor absoluto de las desviaciones.

Problemas similares afectan al cómputo del parámetro de suavizado en el caso de modelos parcialmente lineales y, en particular, en GPLM. En esa dirección, Boente et al. (2006) propusieron un procedimiento robusto de validación cruzada que consiste en utilizar como función de validación a la función de pérdida robusta empleada para la estimación. Más precisamente, el procedimiento consiste en los siguientes pasos:

- I. Dado un valor $\alpha \in (0, 1)$ preestablecido, particionar el conjunto de observaciones de la siguiente manera: seleccionar un subconjunto aleatorio de tamaño $100(1 - \alpha)\%$, cuyos índices se denotan como $\mathcal{L}_{1-\alpha}$ y se denomina conjunto de entrenamiento. El resto de los índices se denotan $\mathcal{J}_{1-\alpha}$ y conforman el conjunto de validación.
- II. Para un h_n dado, computar

$$\begin{aligned}\widehat{m}_{\beta}^{(-\alpha)}(t, h) &= \operatorname{argmax}_{\eta \in \mathbb{R}} \sum_{i \in \mathcal{L}_{1-\alpha}} S_n(\eta, \beta, t) \\ \widehat{\beta}_n^{(-\alpha)}(h) &= \operatorname{argmax}_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{L}_{1-\alpha}} F_n(\beta)\end{aligned}$$

- III. Dada una grilla de valores de h_n elegir el valor óptimo como

$$h_n = \operatorname{argmin}_h \sum_{i \in \mathcal{J}_{1-\alpha}} \rho(Y_i, \mathbf{X}_i^\top \widehat{\beta}_n^{(-\alpha)}(h) + \widehat{m}_{\widehat{\beta}_n^{(-\alpha)}}^{(-\alpha)}(T_i, h)) \omega_2(\mathbf{X}_i).$$

1.3. Modelo de Poisson Inflado en Ceros

El modelado estadístico de datos de conteo constituye una herramienta esencial en múltiples áreas de aplicación, como la econometría, la biología, la epidemiología, la industria y la salud. Con frecuencia, estos datos presentan un exceso de ceros, es decir, más ceros

de los que permiten los GLM clásicos. En tales situaciones, los enfoques convencionales muestran limitaciones importantes, ya que no logran describir adecuadamente la estructura subyacente de los datos. Este comportamiento se ilustra en la Figura 1.4, donde se presenta el gráfico de barras de un conjunto de datos con exceso de ceros, junto con las frecuencias esperadas según un modelo de Poisson estándar y una distribución binomial negativa. El gráfico revela una falta de ajuste de los modelos clásicos. En particular, el modelo binomial negativo tiene dificultad para ajustar los valores centrales, mientras que el modelo de Poisson falla al ajustar los valores iniciales, de hecho si $Y \sim \mathcal{P}(\lambda)$ con $\lambda > 0$, la probabilidad de 0 es $P(Y = 0) = e^{-\lambda}$ y, consecuentemente, no es posible que el modelo tenga una alta proporción de ceros, sobre todo si la media es grande.

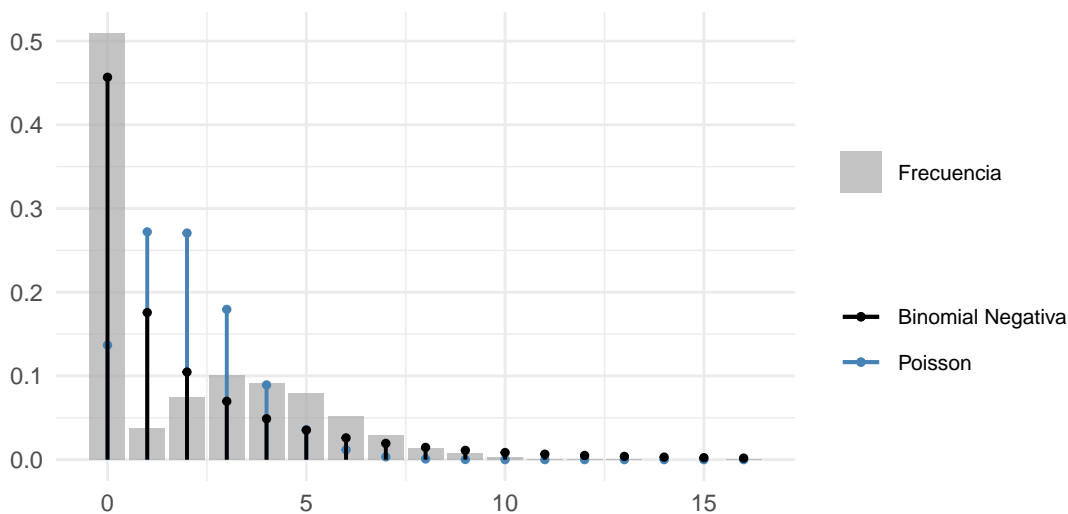


Figura 1.4: Ajustes dados por las distribuciones de Poisson y binomial negativa.

Un ejemplo específico de datos con exceso de ceros se observa frecuentemente en los procesos de producción industrial, ya que el número de defectos en un artículo suele seguir una distribución de Poisson. Sin embargo, a menudo se registran muchos artículos sin defectos. Por ejemplo, Lambert (1992) analizó el número total de cables soldados incorrectamente en placas de circuito impreso y encontró que las 675 observaciones de la variable respuesta tomaron valores entre 0 y 48. De estas, el 81% presentaron cero defectos, el 8% al menos cinco y el 5.2% al menos nueve. Es evidente que la mayoría de las placas no tenían defectos, mientras que aquellas con defectos frecuentemente tenían varios. Los autores encontraron que al ajustar varios modelos log-lineales de Poisson considerando diferentes variables predictoras, el mejor modelo estimó que solo el 71% de las placas no tendrían defectos, subestimando además los recuentos elevados, ya que predijo que solo el 3.7% de las placas tendrían al menos nueve defectos.

Una manera de lidiar con los datos que presentan excesos de ceros es considerarlos como una mezcla entre una masa puntual en cero y una distribución de conteo clásica. Este tipo de modelos se denominan **modelos inflados en ceros**. Particularmente cuando la distribución de conteo es de Poisson se obtiene el **modelo de Poisson inflado en ceros** (ZIP, por sus siglas en inglés *Zero-Inflated Poisson*) propuesto por Mullahy (1986). Más precisamente se dice que una variable de conteo $Y \in \mathbb{N}_0$ sigue el modelo $ZIP(\pi, \lambda)$ si

$$Y \sim \begin{cases} 0 & \text{con probabilidad } \pi \\ P(\lambda) & \text{con probabilidad } 1 - \pi, \end{cases}$$

donde $Y \sim 0$ indica que la variable sigue un proceso de ceros estructurales debido a alguna configuración subyacente de los datos. Mientras que $P(\lambda)$ es una distribución de Poisson clásica con parámetro λ . Equivalentemente se puede escribir

$$\begin{aligned} P(Y = 0) &= \pi + (1 - \pi)e^{-\lambda} \\ P(Y = y) &= (1 - \pi) \frac{e^{-\lambda} \lambda^y}{y!} \quad \text{para } y > 0. \end{aligned} \tag{1.17}$$

Como puede observarse, el modelo $ZIP(\pi, \lambda)$ se reduce a la distribución clásica de Poisson $P(\lambda)$ cuando $\pi = 0$ y a una distribución degenerada de masa puntual en 0 cuando $\pi = 1$, mientras que para $0 < \pi < 1$ se puede considerar una mezcla de ambas distribuciones.

Lambert (1992) extendió el modelo ZIP al caso de regresión, en el que las variables predictoras $\mathbf{X} \in \mathbb{R}^p$ y $\mathbf{Z} \in \mathbb{R}^q$ están relacionadas con la variable respuesta a través de los enlaces logístico y logarítmico, esto es

$$\begin{aligned} \text{logit}(\pi) &= \mathbf{Z}^\top \boldsymbol{\gamma} \\ \log(\lambda) &= \mathbf{X}^\top \boldsymbol{\beta} \end{aligned} \tag{1.18}$$

donde $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^p \times \mathbb{R}^q$ es el parámetro desconocido de la regresión.

Dado el conjunto de covariables, la esperanza de la respuesta está dada por

$$E_{\boldsymbol{\theta}}(Y | (\mathbf{X}, \mathbf{Z})) = (1 - \pi)\lambda = \frac{\exp(\mathbf{X}^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{Z}^\top \boldsymbol{\gamma})}$$

donde el subíndice indica que la esperanza se toma bajo el modelo correspondiente con parámetro $\boldsymbol{\theta}$.

A lo largo de esta sección se asumirá que $Y | (\mathbf{X}, \mathbf{Z})$ sigue el modelo ZIP dado por (1.17) y (1.18) con $\boldsymbol{\gamma}_0$ y $\boldsymbol{\beta}_0$ los verdaderos valores de los parámetros. Además se denotará E_0 al valor esperado bajo el modelo verdadero.

En el caso en que los vectores \mathbf{X} y \mathbf{Z} sean los mismos, Lambert (1992) menciona que es natural reducir el número de parámetros considerando π como una función de λ , y propone

la siguiente parametrización

$$\begin{aligned}\log(\lambda) &= \mathbf{X}^\top \boldsymbol{\beta} \\ \text{logit}(\pi) &= -\tau \mathbf{X}^\top \boldsymbol{\beta},\end{aligned}$$

donde $\tau \in \mathbb{R}$ es un parámetro desconocido.

En la literatura se pueden encontrar modelos inflados en ceros que utilizan distribuciones diferentes a la Poisson. Por ejemplo, Hall (2000) que propone la binomial inflada con ceros (ZIB) o Greene (1994) que introduce la binomial negativa inflada en ceros (ZIBN).

1.3.1. El modelo ZIP como familia exponencial

Teorema 1.3.1. *La familia ZIP(λ, π) con $(\lambda, \pi) \in \mathbb{R}_{>0} \times (0, 1]$ es una familia exponencial con parámetros naturales $\zeta_1(\lambda, \pi) = \log\left(\frac{e^\lambda \pi + (1-\pi)}{1-\pi}\right)$ y $\zeta_2(\lambda, \pi) = \log \lambda$; y estadísticos suficientes $T_1(y) = R$ y $T_2(y) = y(1-R)$, con $R = \mathbb{I}(y=0)$.*

Demostración. Sea $f(y; \lambda, \pi)$ la función de probabilidad puntual de $Y \sim \text{ZIP}(\lambda, \pi)$, entonces

$$\begin{aligned}f(y; \lambda, \pi) &= \left(\pi + (1-\pi)e^{-\lambda}\right)^{\mathbb{I}(y=0)} \left((1-\pi)\frac{e^{-\lambda}\lambda^y}{y!}\right)^{\mathbb{I}(y>0)} \\ &= \left(\pi + (1-\pi)e^{-\lambda}\right)^R (1-\pi)\frac{e^{-\lambda}\lambda^y}{y!} \left((1-\pi)\frac{e^{-\lambda}\lambda^y}{y!}\right)^{-R} \\ &= (1-\pi)e^{-\lambda} \left(\pi + (1-\pi)e^{-\lambda}\right)^R \lambda^y \left((1-\pi)e^{-\lambda}\lambda^y\right)^{-R} \frac{1}{y!^{1-R}} \\ &= (1-\pi)e^{-\lambda} \exp\left\{\log\left(\left(\pi + (1-\pi)e^{-\lambda}\right)^R \lambda^y \left((1-\pi)e^{-\lambda}\lambda^y\right)^{-R}\right)\right\} \frac{1}{y!^{1-R}} \\ &= (1-\pi)e^{-\lambda} \exp\left\{\log\left(\left(\frac{\pi}{1-\pi} + e^{-\lambda}\right)^R \lambda^{y(1-R)} e^{\lambda R}\right)\right\} \frac{1}{y!^{1-R}} \\ &= (1-\pi)e^{-\lambda} \exp\left\{R \log\left(\frac{\pi}{1-\pi} + e^{-\lambda}\right) + y(1-R) \log \lambda + \lambda R\right\} \frac{1}{y!^{1-R}} \\ &= (1-\pi)e^{-\lambda} \exp\left\{R \left(\log\left(\frac{\pi}{1-\pi} + e^{-\lambda}\right) + \log(e^\lambda)\right) + y(1-R) \log \lambda\right\} \frac{1}{y!^{1-R}} \\ &= (1-\pi)e^{-\lambda} \exp\left\{R \log\left(\frac{\pi e^\lambda + (1-\pi)}{1-\pi}\right) + y(1-R) \log \lambda\right\} \frac{1}{y!^{1-R}}\end{aligned}$$

Por lo tanto los estadísticos suficientes vienen dados por $T_1(y) = R$ y $T_2(y) = y(1-R)$, mientras que los parámetros naturales quedan definidos por $\zeta_1(\lambda, \pi) = \log\left(\frac{e^\lambda \pi + (1-\pi)}{1-\pi}\right)$ y $\zeta_2(\lambda, \pi) = \log \lambda$.

□

Notar que si bien el modelo ZIP pertenece a una familia exponencial, la misma no está dada en forma canónica, y por lo tanto no se puede aplicar el procedimiento clásico de estimación basado en GLM.

1.3.2. Prueba de inflación de ceros

Para determinar si la desviación de los datos respecto al modelo teórico de Poisson es estadísticamente significativa se requiere llevar a cabo un test de hipótesis. Las pruebas de razón de verosimilitud (LR) se emplean generalmente para determinar si un modelo más grande es significativamente mejor que un modelo competidor restringido, que está anidado dentro de él. Cuando el modelo restringido no se encuentra en el borde del espacio de parámetros del modelo más grande, es bien sabido que la distribución del estadístico de prueba LR (bajo el modelo restringido) sigue una distribución χ^2 , con grados de libertad iguales al número de parámetros por los que difieren los dos modelos. Sin embargo, en el caso de modelos inflados en ceros el espacio paramétrico de π es $[0, 1]$, por ende resultan un escenario donde el modelo restringido con $\pi = 0$ se encuentra en el borde de ese espacio. Como alternativa Vuong (1989) propuso una prueba para modelos estrictamente no anidados. Sin embargo, como mencionan Wilson (2015) esa prueba resulta metodológicamente inadecuada cuando se aplica para detectar inflación de ceros.

Van den Broek (1995) desarrollaron un test de score para la inflación de ceros en el contexto del modelo ZIP para el caso en que las probabilidades de inflación de ceros no dependen de covariables. Jansakul and Hinde (2002) extendieron esa prueba al caso de regresión. El test de score tiene una ventaja sobre el de razón de verosimilitud y el test de Wald en que solo requiere las estimaciones de los parámetros bajo la hipótesis nula. Sin embargo, estas pruebas se basan en propiedades asintóticas y, a menudo, no diferencian claramente entre inflación y deflación de ceros.

Wilson and Einbeck (2018) introdujeron una nueva prueba que permite detectar modificación, inflación o deflación de ceros en modelos para datos de conteo, particularmente en el contexto de la regresión de Poisson. El término modificación con ceros es utilizado para referirse a la presencia de demasiados o muy pocos ceros en los datos, en relación con el modelo de datos de conteo especificado. Las hipótesis consideradas por los autores son:

H_0 : Los datos siguen una distribución de Poisson.

$H_1^{(a)}$: Los datos siguen una distribución de Poisson con modificación con ceros.

$H_1^{(b)}$: Los datos siguen una distribución de Poisson con inflación con ceros.

$H_1^{(c)}$: Los datos siguen una distribución de Poisson con deflación con ceros.

La prueba asume que los ceros son generados por un proceso de Bernoulli con probabilidad variable y utilizan el número total de ceros observados, N_0 , como estadístico de prueba,

el cual sigue una distribución Poisson-binomial. Esta distribución generaliza la binomial cuando las probabilidades de éxito varían entre las observaciones (por ejemplo, al incluir covariables).

A diferencia de los métodos clásicos, esta prueba distingue de forma clara entre inflación y deflación con ceros. Además, es válida incluso para tamaños de muestra moderados o pequeños debido a que no se basa en aproximaciones asintóticas. Bajo la hipótesis nula se utiliza únicamente el modelo ajustado sin requerir un modelo alternativo explícito, similar a una prueba de score. Los estudios de simulación realizados en el artículo demuestran que la prueba alcanza una potencia competitiva y tiene niveles de significancia correctos en comparación con los enfoques basados en LR.

1.3.3. Estimación clásica: Máxima Verosimilitud

Dada una muestra $\{(Y_i, \mathbf{X}_i^\top, \mathbf{Z}_i^\top)\}_{i=1}^n \in \mathbb{N}_0 \times \mathbb{R}^p \times \mathbb{R}^q$ que sigue el modelo de regresión ZIP dado por (1.17) y (1.18), la función de log-verosimilitud está definida por

$$\begin{aligned}
 \ell(\mathbf{Y}; \boldsymbol{\theta}) &= \sum_{i=1}^n \log f(Y_i, \boldsymbol{\theta}) & (1.19) \\
 &= \sum_{Y_i=0} \log(\pi_i + (1 - \pi_i)e^{-\lambda_i}) + \sum_{Y_i>0} \log\left((1 - \pi_i) \frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!}\right) \\
 &= \sum_{Y_i=0} \log\left(\frac{\pi_i}{1 - \pi_i} + e^{-\lambda_i}\right) + \sum_{i=1}^n \log(1 - \pi_i) \\
 &\quad + \sum_{Y_i>0} \left\{ \log(e^{-\lambda_i} \lambda_i^{Y_i}) - \log(Y_i!) \right\} \\
 &= \sum_{Y_i=0} \log(e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}} + \exp(-e^{\mathbf{X}_i^\top \boldsymbol{\beta}})) - \sum_{i=1}^n \log(1 + e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}) \\
 &\quad + \sum_{Y_i>0} \left\{ (-e^{\mathbf{X}_i^\top \boldsymbol{\beta}} + Y_i \mathbf{X}_i^\top \boldsymbol{\beta}) - \log(Y_i!) \right\} \\
 &= \sum_{Y_i=0} (\mathbf{Z}_i^\top \boldsymbol{\gamma} - e^{\mathbf{X}_i^\top \boldsymbol{\beta}}) - \sum_{i=1}^n \log(1 + e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}) \\
 &\quad + \sum_{Y_i>0} \left\{ (-e^{\mathbf{X}_i^\top \boldsymbol{\beta}} + Y_i \mathbf{X}_i^\top \boldsymbol{\beta}) - \log(Y_i!) \right\}.
 \end{aligned}$$

donde se considera $\mathbf{Y} = (Y_1, \dots, Y_n)$.

Luego, el MLE de $\boldsymbol{\theta}_0$ viene dado por

$$\hat{\boldsymbol{\theta}}_n^{ML} \in \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\mathbf{Y}; \boldsymbol{\theta}).$$

Si bien la log-verosimilitud del modelo ZIP puede maximizarse directamente mediante métodos numéricos generales, la estructura del modelo permite reformular el problema

de manera más conveniente mediante el algoritmo EM, que se presentará en la sección a continuación. Este algoritmo explota la naturaleza de mezcla del modelo y conduce a pasos de actualización más simples. Detalles sobre el algoritmo se pueden encontrar en McLachlan and Krishnan (2007).

1.3.4. Algoritmo EM

El **algoritmo EM** (*Expectation-Maximization*) propuesto por Dempster et al. (1977) es una buena alternativa para hallar el MLE aprovechando la naturaleza del modelo ZIP, ya que permite ajustar iterativamente versiones ponderadas de GLM más simples.

Seguendo Lambert (1992), si se supone que la fuente de ceros (ceros estructurales o distribución de Poisson) es conocida, es decir, que la variable auxiliar (también conocida como latente o faltante)

$$W = \begin{cases} 1 & \text{si } Y \sim 0, \\ 0 & \text{si } Y \sim \mathcal{P}(\lambda) \end{cases}$$

es observable. Dada la muestra aleatoria completa $\{(Y_i, W_i, \mathbf{X}_i^\top, \mathbf{Z}_i^\top)\}_{i=1}^n$ independiente y sea $\mathbf{W} = (W_1, \dots, W_n)$, la función de log-verosimilitud empírica completa resulta

$$\begin{aligned} \ell_c(\mathbf{Y}, \mathbf{W}; \boldsymbol{\theta}) &= \sum_{i=1}^n \log f(Y_i, W_i, \boldsymbol{\theta}) & (1.20) \\ &= \sum_{i=1}^n \log f(Y_i, \boldsymbol{\theta} | W_i) + \sum_{i=1}^n \log f(W_i, \boldsymbol{\theta}) \\ &= \sum_{i:W_i=0} \log \left(\frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!} \right) + \sum_{i:W_i=1} \log(1) + \sum_{i=1}^n \log \left(\pi_i^{W_i} (1 - \pi_i)^{1-W_i} \right) \\ &= \sum_{i=1}^n (1 - W_i) \log \left(e^{-\lambda_i} \lambda_i^{Y_i} \right) - \sum_{i=1}^n (1 - W_i) \log(Y_i!) \\ &\quad + \sum_{i=1}^n (W_i \log \pi_i + (1 - W_i) \log(1 - \pi_i)) \\ &= \sum_{i=1}^n (1 - W_i) \log \left(e^{-\lambda_i} \lambda_i^{Y_i} \right) - \sum_{i=1}^n (1 - W_i) \log(Y_i!) \\ &\quad + \sum_{i=1}^n \left(W_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) \right) \\ &= \ell_1(\mathbf{Y}, \mathbf{W}; \boldsymbol{\beta}) + \ell_2(\mathbf{W}; \boldsymbol{\gamma}), \end{aligned}$$

donde

$$\begin{aligned}\ell_1(\mathbf{Y}, \mathbf{W}; \boldsymbol{\beta}) &= \sum_{i=1}^n (1 - W_i) \left(Y_i \mathbf{X}_i^\top \boldsymbol{\beta} - e^{\mathbf{X}_i^\top \boldsymbol{\beta}} - \log(Y_i!) \right), \\ \ell_2(\mathbf{W}; \boldsymbol{\gamma}) &= \sum_{i=1}^n \left(W_i \mathbf{Z}_i^\top \boldsymbol{\gamma} - \log(1 + e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}) \right).\end{aligned}$$

A diferencia de (1.19), ℓ_1 depende solo de $\boldsymbol{\beta}$ y ℓ_2 solo de $\boldsymbol{\gamma}$. Por lo tanto, los dos términos pueden optimizarse por separado, facilitando la estimación. Las ecuaciones de estimación resultan entonces

$$\sum_{i=1}^n \left(W_i - \left(1 + e^{-\mathbf{Z}_i^\top \boldsymbol{\gamma}} \right)^{-1} \right) \mathbf{Z}_i^\top = 0, \quad (1.21)$$

$$\sum_{i=1}^n (1 - W_i) \left(Y_i - e^{\mathbf{X}_i^\top \boldsymbol{\beta}} \right) \mathbf{X}_i^\top = 0. \quad (1.22)$$

El algoritmo EM aborda el problema de optimizar la función de log-verosimilitud de datos incompletos (1.19) indirectamente al optimizar de forma iterativa la función de log-verosimilitud de los datos completos (1.20). Sin embargo, dicha función no es observable (debido a que la variable W no lo es). La propuesta consiste entonces en maximizar la esperanza condicional de la log-verosimilitud completa dados los datos incompletos observables $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i)\}_{i=1}^n$, bajo la estimación actual de los parámetros $\hat{\boldsymbol{\theta}}_n$. Con la finalidad de simplificar la notación, a lo largo de la tesis denotaremos esa esperanza $E_{\hat{\boldsymbol{\theta}}_n}(\cdot | \mathbf{Y})$. Específicamente, el algoritmo EM consiste en los siguientes pasos:

(Inicialización) Obtener estimadores iniciales de los parámetros de interés $\hat{\boldsymbol{\theta}}_n^{(0)}$.

(Paso *E*) Calcular la log-verosimilitud completa esperada como

$$Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_n^{(r)}) \doteq E_{\hat{\boldsymbol{\theta}}_n^{(r)}}(\ell_c(\mathbf{Y}, \mathbf{W}; \boldsymbol{\theta}) | \mathbf{Y}). \quad (1.23)$$

(Paso *M*) Maximizar la log-verosimilitud completa esperada con respecto a los parámetros de interés, esto es

$$\hat{\boldsymbol{\theta}}_n^{(r+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

Este procedimiento iterativo se repite hasta que se satisfaga algún criterio de convergencia predefinido, como una diferencia pequeña en la log-verosimilitud empírica para estimadores consecutivos, es decir $\ell(\mathbf{Y}; \hat{\boldsymbol{\theta}}_n^{(r+1)}) - \ell(\mathbf{Y}; \hat{\boldsymbol{\theta}}_n^{(r)}) < \delta$ para $\delta > 0$ suficientemente chico.

Notar que ℓ_c es lineal en W_i y al condicionar en los datos observables el paso *E* para el modelo ZIP se reduce a actualizar los elementos de \mathbf{W} con su esperanza condicional dados los datos observables y la estimación actual de los parámetros en la iteración r , como sigue

$$W_i^{(r)} = E_{\hat{\boldsymbol{\theta}}_n^{(r)}}(W_i | Y_i) \quad i = 1, \dots, n$$

con la cual resulta

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_n^{(r)}) = \ell_c(\mathbf{Y}, \mathbf{W}^{(r)}; \boldsymbol{\theta}).$$

La esperanza condicional de la variable W se puede calcular fácilmente usando la regla de Bayes, en efecto

$$P(Y = 0|W) = \begin{cases} 1 & \text{si } W = 1 \\ e^{-\lambda} & \text{si } W = 0 \end{cases}$$

y si $k > 0$

$$P(Y = k|W) = \begin{cases} 0 & \text{si } W = 1 \\ e^{-\lambda} \lambda^k / k! & \text{si } W = 0 \end{cases}$$

por lo tanto

$$\begin{aligned} E_{\boldsymbol{\theta}}(W|Y) &= P(W = 1|Y) = \frac{P(Y|W = 1)P(W = 1)}{P(Y)} & (1.24) \\ &= \frac{P(Y|W = 1)P(W = 1)}{P(Y|W = 1)P(W = 1) + P(Y|W = 0)P(W = 0)} \\ &= \begin{cases} \frac{\pi}{\pi + e^{-\lambda}(1 - \pi)} & \text{si } Y = 0, \\ 0 & \text{si } Y > 0 \end{cases} \\ &= \begin{cases} \frac{1}{1 + \exp\{-\mathbf{Z}^\top \boldsymbol{\gamma} - \exp \mathbf{X}^\top \boldsymbol{\beta}\}} & \text{si } Y = 0 \\ 0 & \text{si } Y > 0. \end{cases} \end{aligned}$$

Así, en el contexto del modelo ZIP, los pasos del algoritmo EM se reducen simplemente a:

(Inicialización) Obtener estimadores iniciales de los parámetros de interés $\widehat{\boldsymbol{\beta}}_n^{(0)}$ y $\widehat{\boldsymbol{\gamma}}_n^{(0)}$.

(Paso *E*) Actualizar según (1.24) los elementos de $\mathbf{W}^{(r)}$ por su esperanza condicional bajo la estimación actual, es decir

$$W_i^{(r)} = \begin{cases} \frac{1}{1 + \exp\{-\mathbf{Z}_i^\top \widehat{\boldsymbol{\gamma}}_n^{(r)} - \exp \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{(r)}\}} & \text{si } Y_i = 0 \\ 0 & \text{si } Y_i > 0 \end{cases} \quad i = 1, \dots, n.$$

(Paso *M*) Maximizar la función de log-verosimilitud empírica completa esperada con respecto a los parámetros de interés, lo cual es equivalente a computar

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n^{(r+1)} &= \operatorname{argmax}_{\boldsymbol{\beta}} \ell_1(\mathbf{Y}, \mathbf{W}^{(r)}; \boldsymbol{\beta}) \\ \widehat{\boldsymbol{\gamma}}_n^{(r+1)} &= \operatorname{argmax}_{\boldsymbol{\gamma}} \ell_2(\mathbf{W}^{(r)}; \boldsymbol{\gamma}). \end{aligned}$$

1.3.4.1. ¿Por qué funciona el Algoritmo EM?

El algoritmo EM fue introducido por Dempster et al. (1977), quienes establecieron, entre otros resultados, las conocidas propiedades de monotonía del algoritmo. Demostraron particularmente que la log-verosimilitud con datos incompletos $\ell(\mathbf{Y}; \boldsymbol{\theta})$ no disminuye después de una iteración del algoritmo. Dicha propiedad se presenta en el siguiente resultado:

Proposición 1.3.1. *Sea $\{\widehat{\boldsymbol{\theta}}_n^{(r)}\}$ una sucesión de estimadores obtenidos mediante el algoritmo EM, esto es, tales que*

$$\widehat{\boldsymbol{\theta}}_n^{(r+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} Q(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}_n^{(r)}).$$

Entonces,

$$\ell(\mathbf{Y}; \widehat{\boldsymbol{\theta}}_n^{(r+1)}) \geq \ell(\mathbf{Y}; \widehat{\boldsymbol{\theta}}_n^{(r)}) \quad \forall r = 0, 1, 2, \dots \quad (1.25)$$

Demostración. Para probarlo notar que mediante la regla de Bayes se obtiene que

$$\log f(\mathbf{Y}, \boldsymbol{\theta}) = \log f(\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta}) - \log f(\mathbf{W} | \mathbf{Y}, \boldsymbol{\theta}).$$

Luego,

$$\ell(\mathbf{Y}; \boldsymbol{\theta}) = \ell_c(\mathbf{Y}, \mathbf{W}; \boldsymbol{\theta}) - \sum_{i=1}^n \log f(W_i | Y_i, \boldsymbol{\theta}).$$

Tomando a ambos lados de la ecuación la esperanza condicional de $\mathbf{W} | \mathbf{Y}$ bajo la estimación actual $\widehat{\boldsymbol{\theta}}_n^{(r)}$, como el lado izquierdo no depende de \mathbf{W} , se tiene

$$\ell(\mathbf{Y}; \boldsymbol{\theta}) = Q(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}_n^{(r)}) - H(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}_n^{(r)}), \quad (1.26)$$

donde

$$H(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}_n^{(r)}) \doteq E_{\widehat{\boldsymbol{\theta}}_n^{(r)}} \left(\sum_{i=1}^n \log f(W_i | Y_i, \boldsymbol{\theta}) | \mathbf{Y} \right)$$

y $Q(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}_n^{(r)})$ está definida en (1.23). De (1.26) se tiene que

$$\begin{aligned} \ell(\mathbf{Y}; \widehat{\boldsymbol{\theta}}_n^{(r+1)}) - \ell(\mathbf{Y}; \widehat{\boldsymbol{\theta}}_n^{(r)}) &= \left(Q(\widehat{\boldsymbol{\theta}}_n^{(r+1)} | \widehat{\boldsymbol{\theta}}_n^{(r)}) - Q(\widehat{\boldsymbol{\theta}}_n^{(r)} | \widehat{\boldsymbol{\theta}}_n^{(r)}) \right) \\ &\quad - \left(H(\widehat{\boldsymbol{\theta}}_n^{(r+1)} | \widehat{\boldsymbol{\theta}}_n^{(r)}) - H(\widehat{\boldsymbol{\theta}}_n^{(r)} | \widehat{\boldsymbol{\theta}}_n^{(r)}) \right). \end{aligned} \quad (1.27)$$

La primer diferencia es positiva porque $\widehat{\boldsymbol{\theta}}_n^{(r+1)}$ se elige de manera que

$$Q(\widehat{\boldsymbol{\theta}}_n^{(r+1)} | \widehat{\boldsymbol{\theta}}_n^{(r)}) \geq Q(\boldsymbol{\theta} | \widehat{\boldsymbol{\theta}}_n^{(r)}) \quad \forall \boldsymbol{\theta} \in \Theta. \quad (1.28)$$

Ahora, para cualquier θ

$$\begin{aligned}
 H(\theta|\hat{\theta}_n^{(r)}) - H(\hat{\theta}_n^{(r)}|\hat{\theta}_n^{(r)}) &= E_{\hat{\theta}_n^{(r)}} \left(\sum_{i=1}^n \log \frac{f(W_i|Y_i;\theta)}{f(W_i|Y_i;\hat{\theta}_n^{(r)})} \mid \mathbf{Y} \right) \\
 &\leq \log E_{\hat{\theta}_n^{(r)}} \left(\frac{\prod_{i=1}^n f(W_i|Y_i;\theta)}{\prod_{i=1}^n f(W_i|Y_i;\hat{\theta}_n^{(r)})} \mid \mathbf{Y} \right) \\
 &= \log(1) = 0
 \end{aligned} \tag{1.29}$$

donde la desigualdad en (1.29) es una consecuencia de la desigualdad de Jensen (Teorema A.1.4) y de la concavidad de la función logarítmica. Finalmente, por (1.27), (1.28) y (1.29) se verifica (1.25). \square

Una consecuencia de este resultado es la propiedad de auto-consistencia, que establece que si $\hat{\theta}_n^{ML}$ maximiza globalmente a $\ell(\mathbf{Y};\theta)$, entonces

$$\hat{\theta}_n^{ML} \in \operatorname{argmax}_{\theta} Q(\theta|\hat{\theta}_n^{ML}).$$

En efecto, por (1.27) y (1.29) vale que

$$\begin{aligned}
 0 &\geq \ell(\mathbf{Y};\theta) - \ell(\mathbf{Y};\hat{\theta}_n^{ML}) \\
 &\geq Q(\theta|\hat{\theta}_n^{ML}) - Q(\hat{\theta}_n^{ML}|\hat{\theta}_n^{ML}) \quad \forall \theta \in \Theta.
 \end{aligned}$$

Notar que la Proposición 1.3.1 implica que si $\{\ell(\mathbf{Y};\hat{\theta}_n^{(r)})\}$ es una sucesión acotada, entonces converge monótonamente a algún valor ℓ^* , pero eso no garantiza que ℓ^* sea un máximo global, uno local o incluso un punto estacionario de $\ell(\mathbf{Y};\theta)$. En este sentido Wu (1983) estableció algunos de los resultados de convergencia más generales para el algoritmo EM. Por ejemplo, el Teorema 2 del citado artículo establece que si $Q(\theta|\theta')$ es continua en θ y θ' , entonces el algoritmo converge a un punto estacionario ℓ^* de $\ell(\mathbf{Y};\theta)$. Además, el Corolario 1 señala que, si la función de log-verosimilitud tiene un único punto estacionario y la primera derivada de $Q(\theta|\theta')$ con respecto a θ es continua en θ y θ' , la sucesión de estimadores que resulta de las iteraciones del algoritmo EM converge al único máximo de $\ell(\mathbf{Y};\theta)$. Lambert (1992) demostró que estas hipótesis se verifican en el caso del modelo de regresión ZIP y, en ese caso, la consistencia del estimador obtenido mediante el algoritmo EM puede garantizarse por la del MLE.

Cuando la función de verosimilitud es multimodal lo mejor que se puede garantizar es la convergencia a algún óptimo local de la verosimilitud. En general, si tiene varios máximos y valores estacionarios, la convergencia de la sucesión de estimadores depende de la elección del punto de partida. Entre los resultados más recientes, el trabajo de Balakrishnan et al. (2017) caracteriza un entorno alrededor del verdadero parámetro poblacional tal que si

el algoritmo EM se inicializa dentro de ese entorno, entonces se garantiza que convergerá con precisión estadística a un óptimo global de la verosimilitud poblacional. Esto implica la consistencia del estimador obtenido. En el libro de McLachlan and Krishnan (2007) se puede encontrar una explicación completa sobre estos aspectos teóricos del algoritmo.

1.3.5. Estimador Robusto para el modelo ZIP

El enfoque clásico de estimación robusta consiste en reducir la influencia de las observaciones extremas utilizando funciones de score acotadas en lugar de la función de log-verosimilitud. No obstante, en modelos de mezcla como el ZIP, este enfoque puede ser insuficiente ya que numerosas observaciones legítimas pueden estar alejadas de la media global debido a la naturaleza del modelo. Para superar estas limitaciones Hall and Shen (2010) propusieron el método **Robust Expectation Solution** (RES) en el cual utilizan el algoritmo **Expectation-Solution** (ES) que es similar al EM pero el paso M se modifica para resolver ecuaciones de estimación en lugar de maximizar un funcional. Cuando las ecuaciones de estimación del algoritmo ES vienen dadas por las derivadas de la log-verosimilitud, los algoritmos ES y EM son equivalentes. Este procedimiento permite robustificar cada componente de la mezcla de distribuciones por separado, clasificando las observaciones como extremas o no en función de la componente a la que pertenecen. Específicamente los autores propusieron reemplazar las ecuaciones de estimación (1.21) y (1.22) por las versiones robustas

$$\sum_{i=1}^n \left\{ W_i - \frac{e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}}{1 + e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}} \right\} \omega(\mathbf{Z}_i) \mathbf{Z}_i = 0,$$

$$\sum_{i=1}^n (1 - W_i) \{ \psi_c(Y_i) - o_i(\boldsymbol{\beta}, c) \} \omega(\mathbf{X}_i) \mathbf{X}_i = 0,$$

donde

$$\psi_c(y) = \begin{cases} j_1 & \text{si } y < j_1 \\ y & \text{si } y \in [j_1, j_2] \\ j_2 & \text{si } y > j_2, \end{cases}$$

con j_1 y j_2 los cuantiles c y $1 - c$ de la distribución de Poisson, respectivamente, y

$$\begin{aligned} o_i(\boldsymbol{\beta}, c) &= E(\psi_c(Y_i)) \\ &= j_1 P(Y_i < j_1) + \lambda_i P(j_1 \leq Y_i \leq j_2) + j_2 P(Y_i > j_2), \end{aligned}$$

donde las probabilidades se calculan en base a la componente de Poisson, asumiendo que $Y_i \sim \mathcal{P}(\lambda_i)$ con $\lambda_i = e^{\mathbf{X}_i^\top \boldsymbol{\beta}}$. La función ω da peso pequeño a las observaciones de alta influencia. Los autores consideran $\omega(\mathbf{Z}_i) = \sqrt{1 - \bar{h}_i}$ con h_i el leverage de \mathbf{Z}_i , es decir, el i -ésimo elemento de la diagonal de la matriz $\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}^\top$ donde $\tilde{\mathbf{Z}} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)^\top$

es una matriz de tamaño $n \times q$. Para $\omega(\mathbf{X}_i)$ se realiza una elección análoga. Los autores utilizan $c = 0.01$ para la constante que controla la compensación entre la robustez y la eficiencia.

Los autores destacan que la robustez de la estimación de γ está garantizada por la función de pesos ω_2 y una precisa estimación de W_i (que viene dada por la estimación robusta de β).

El algoritmo resulta entonces:

(Inicialización) Obtener estimadores iniciales de los parámetros de interés $\hat{\beta}_n^{(0)}$ y $\hat{\gamma}_n^{(0)}$.

(Paso *E*) En la iteración r estimar los elementos de $\mathbf{W}^{(r)}$ como

$$W_i^{(r)} = \begin{cases} \frac{1}{1 + \exp\{-\mathbf{Z}_i^\top \hat{\gamma}_n^{(r)} - \exp \mathbf{X}_i^\top \hat{\beta}_n^{(r)}\}} & \text{si } Y_i = 0 \\ 0 & \text{si } Y_i > 0 \end{cases} \quad i = 1, \dots, n.$$

(Paso *S*) Hallar $\hat{\gamma}_n^{(r+1)}$ y $\hat{\beta}_n^{(r+1)}$ que resuelven

$$\sum_{i=1}^n \left\{ W_i^{(r)} - \frac{e^{\mathbf{Z}_i^\top \hat{\gamma}_n^{(r+1)}}}{1 + e^{\mathbf{Z}_i^\top \hat{\gamma}_n^{(r+1)}}} \right\} \omega(\mathbf{Z}_i) \mathbf{Z}_i = 0,$$

$$\sum_{i=1}^n (1 - W_i^{(r)}) \left\{ \psi_c(Y_i) - o_i(\hat{\beta}_n^{(r+1)}, c) \right\} \omega(\mathbf{X}_i) \mathbf{X}_i = 0.$$

Los autores demostraron que los estimadores obtenidos mediante el método RES resultan consistentes y asintóticamente normales.

1.4. Modelo de Poisson Inflado en Ceros Parcialmente Lineal

El **modelo de Poisson inflado en ceros parcialmente lineal** (PLZIP) es una extensión del modelo ZIP especificado en 1.17 y 1.18 que permite incluir predictoras de manera no paramétrica. Dicha componente puede estar incluida en la media de Poisson, en la probabilidad de ceros estructurales, o en ambas. Más precisamente, una variable aleatoria $Y \in \mathbb{N}_0$ sigue el modelo PLZIP si

$$P(Y = 0) = \pi + (1 - \pi)e^{-\lambda} \tag{1.30}$$

$$P(Y = y) = (1 - \pi) \frac{e^{-\lambda} \lambda^y}{y!} \quad \text{para } y > 0$$

y dadas las variables predictoras $(\mathbf{X}, \mathbf{Z}, T) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}$ se asume

$$\begin{aligned}\log(\lambda) &= \boldsymbol{\beta}^T \mathbf{X} + m(T) \\ \text{logit}(\pi) &= \mathbf{Z}^T \boldsymbol{\gamma},\end{aligned}\tag{1.31}$$

donde $\boldsymbol{\beta} \in \mathbb{R}^p$ y $\boldsymbol{\gamma} \in \mathbb{R}^q$ son los parámetros desconocidos de la regresión y $m : \mathbb{R} \rightarrow \mathbb{R}$ una función continua desconocida. El espacio paramétrico para el modelo resulta entonces

$$\Theta = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, m) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathcal{C}\},\tag{1.32}$$

donde \mathcal{C} es el conjunto de las funciones continuas.

Otra alternativa parcialmente lineal del modelo ZIP considera la componente no paramétrica en el enlace logístico, de manera que

$$\begin{aligned}\log(\lambda) &= \mathbf{X}^T \boldsymbol{\beta} \\ \text{logit}(\pi) &= \mathbf{Z}^T \boldsymbol{\gamma} + m(T),\end{aligned}$$

mientras que el caso más general, denominado por algunos autores como **modelo ZIP doblemente semiparamétrico**, viene dado por

$$\begin{aligned}\log(\lambda) &= \mathbf{X}^T \boldsymbol{\beta} + m_\lambda(T) \\ \text{logit}(\pi) &= \boldsymbol{\gamma}^T \mathbf{X} + m_\pi(T).\end{aligned}$$

Una extensión del modelo PLZIP que permite incluir un vector de variables predictoras de forma no paramétrica (en lugar de una sola variable real) sin incurrir en la maldición de la dimensión es el **modelo ZIP de índice único** estudiado por Wang et al. (2015). Los autores estiman el modelo (1.30) con una configuración parcialmente lineal similar a (1.31), pero donde m es una función univariada aplicada a la proyección del vector $\mathbf{T} \in \mathbb{R}^s$ en alguna dirección, es decir

$$\begin{aligned}\log(\lambda) &= \boldsymbol{\beta}^T \mathbf{X} + m(\mathbf{T}^T \boldsymbol{\alpha}) \\ \text{logit}(\pi) &= \mathbf{Z}^T \boldsymbol{\gamma}\end{aligned}\tag{1.33}$$

donde ahora $\boldsymbol{\alpha}$ es un vector en \mathbb{R}^r , $\mathbf{T}^T \boldsymbol{\alpha} \in \mathbb{R}$ y $m : \mathbb{R} \rightarrow \mathbb{R}$.

1.4.1. Estimación Basada en Verosimilitud

Dada una muestra aleatoria $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i, T_i)\}_{i=1}^n$ que sigue el modelo (1.30) con la estructura parcialmente lineal dada en (1.31), la función de log-verosimilitud viene dada

por

$$\begin{aligned}
 \ell(\mathbf{Y}; \boldsymbol{\theta}) &= \sum_{i=1}^n \log f(Y_i, \boldsymbol{\theta}) \\
 &= \sum_{Y_i=0} \log \left(\pi_i + (1 - \pi_i)e^{-\lambda_i} \right) + \sum_{Y_i>0} \log \left((1 - \pi_i) \frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!} \right) \\
 &= \sum_{Y_i=0} \log \left(\frac{\pi_i}{1 - \pi_i} + e^{-\lambda_i} \right) + \sum_{i=1}^n \log(1 - \pi_i) + \sum_{Y_i>0} \log \left(e^{-\lambda_i} \lambda_i^{Y_i} \right) - \sum_{Y_i>0} \log(Y_i!) \\
 &= \sum_{Y_i=0} \log \left(e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}} + \exp(-e^{\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i)}) \right) - \sum_{i=1}^n \log(1 - e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}) \\
 &+ \sum_{Y_i>0} \left(-e^{\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i)} + Y_i \left(\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i) \right) \right) - \sum_{Y_i>0} \log(Y_i!).
 \end{aligned}$$

La función de verosimilitud para el modelo PLZIP de índice único es análoga a la del caso escalar, reemplazando $m(T_i)$ por $m(\mathbf{T}_i^\top \boldsymbol{\alpha})$. En ambos escenarios, la maximización de la verosimilitud exige combinar técnicas paramétricas y no paramétricas, dado que deben estimarse simultáneamente los parámetros lineales y la función m . A continuación se presenta un método de estimación basado en la función de log-verosimilitud para el modelo PLZIP de índice único, propuesto por Wang et al. (2015).

1.4.1.1. Perfiles de Verosimilitud basados en Splines

En Wang et al. (2015) los autores propusieron estimadores del modelo PLZIP de índice único definido en (1.30) y (1.33). Los estimadores están basados en perfiles de verosimilitud, estimando en primer lugar el término no paramétrico utilizando una base B-spline y luego construyendo el perfil de verosimilitud para estimar los demás parámetros. Más precisamente, aproximaron $m(t) \approx \mathbf{B}(t)^\top \mathbf{a}$, donde $\mathbf{B}(t) = (B_1(t), \dots, B_{k_n+v+1}(t))^\top$ un conjunto de funciones base B-spline de orden $v+1$ con k_n nodos internos casi uniformes y $\mathbf{a} = (a_1, \dots, a_{k_n+v+1})^\top$ es el vector de coeficientes spline, más detalles se pueden encontrar en Schumaker (1981). Sean $\boldsymbol{\beta}, \boldsymbol{\gamma}$ fijos y $\hat{\mathbf{a}}_n$ el maximizador de la función

$$\begin{aligned}
 \ell(\mathbf{a}) &= \sum_{Y_i=0} \log \left(\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} + \frac{\exp(-\exp(\mathbf{B}(\mathbf{T}_i^\top \boldsymbol{\alpha})^\top \mathbf{a} + \mathbf{Z}_i^\top \boldsymbol{\gamma}))}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} \right) \\
 &+ \sum_{Y_i>0} \left\{ \log \left(\frac{1}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} \right) - \exp \left(\mathbf{B}(\mathbf{T}_i^\top \boldsymbol{\alpha})^\top \mathbf{a} + \mathbf{Z}_i^\top \boldsymbol{\gamma} \right) \right. \\
 &\left. + Y_i \left(\mathbf{B}(\mathbf{T}_i^\top \boldsymbol{\alpha})^\top \mathbf{a} + \mathbf{Z}_i^\top \boldsymbol{\gamma} \right) \right\},
 \end{aligned} \tag{1.34}$$

el estimador de $m(t)$ resulta $\hat{m}_n(t) = \mathbf{B}(t)^\top \hat{\mathbf{a}}_n$ y el perfil de verosimilitud para estimar $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ está dado por

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = & \sum_{Y_i=0} \log \left(\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} + \frac{\exp(-\exp(\hat{m}_n(\mathbf{T}_i^\top \boldsymbol{\alpha}) + \mathbf{Z}_i^\top \boldsymbol{\gamma}))}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} \right) \quad (1.35) \\ & + \sum_{Y_i>0} \left\{ \log \left(\frac{1}{1 + \exp(\mathbf{X}_i^\top \boldsymbol{\beta})} \right) - \exp \left(\hat{m}_n(\mathbf{T}_i^\top \boldsymbol{\alpha}) + \mathbf{Z}_i^\top \boldsymbol{\gamma} \right) \right. \\ & \left. + Y_i \left(\hat{m}_n(\mathbf{T}_i^\top \boldsymbol{\alpha}) + \mathbf{Z}_i^\top \boldsymbol{\gamma} \right) \right\}. \end{aligned}$$

El procedimiento de estimación actualiza iterativamente $\hat{\boldsymbol{\theta}}_n(\cdot)$ mediante el cálculo de $\hat{\mathbf{a}}_n$ usando el algoritmo de punto fijo de Cui et al. (2011), y los estimadores de los parámetros $\hat{\boldsymbol{\beta}}_n$ y $\hat{\boldsymbol{\gamma}}_n$ mediante el algoritmo de Newton-Raphson. El procedimiento iterativo es el siguiente:

(Inicialización) Obtener estimadores iniciales de $\hat{\boldsymbol{\beta}}_n^{(0)}$, $\hat{\boldsymbol{\gamma}}_n^{(0)}$ y $\hat{\boldsymbol{\alpha}}$ como el MLE de un modelo ZIP lineal, es decir, considerando m como la identidad. Debido a la restricción sobre $\boldsymbol{\alpha}$, el valor inicial de $\boldsymbol{\alpha}$ se normaliza como $\hat{\boldsymbol{\alpha}}^{(0)} / \|\hat{\boldsymbol{\alpha}}^{(0)}\|$.

(Paso 1) Sustituir $\hat{\boldsymbol{\beta}}_n^{(r)}$, $\hat{\boldsymbol{\gamma}}_n^{(r)}$ y $\hat{\boldsymbol{\alpha}}^{(r)}$ en la log-verosimilitud no paramétrica (1.34) y obtener su maximizador $\hat{\mathbf{a}}_n^{(r)}$.

(Paso 2) Sustituir $\hat{\mathbf{a}}_n^{(r)}$ en el perfil de log-verosimilitud (1.35) y obtener los estimadores actualizados $\hat{\boldsymbol{\beta}}_n^{(r+1)}$, $\hat{\boldsymbol{\gamma}}_n^{(r+1)}$ y $\hat{\boldsymbol{\alpha}}^{(r+1)}$ bajo la restricción $\|\boldsymbol{\alpha}\| = 1$.

Repetir los pasos 1 y 2 hasta que se cumpla un criterio de convergencia predeterminado.

Los autores demostraron bajo ciertas condiciones de regularidad que los estimadores son débilmente consistentes y asintóticamente normales, además una expresión para la matriz de covarianza asintótica.

El procedimiento de estimación descrito anteriormente es implementado en el Capítulo 3, donde se estudia el desempeño de los estimadores propuestos en esta tesis.

Estimación Robusta del Modelo de Poisson Inflado en Ceros Parcialmente Lineal

Este capítulo constituye el principal aporte de esta tesis. En el mismo se introduce un nuevo procedimiento robusto de estimación para el modelo PLZIP, se propone un algoritmo tipo EM que permite obtener los estimadores y un resultado de convergencia de dicho algoritmo. Asimismo, se presentan resultados asintóticos: se demuestra la consistencia fuerte tanto de los parámetros lineales como de la función no paramétrica, y se establece la distribución asintótica de los estimadores de los parámetros lineales. Esta última se emplea para construir un estadístico tipo Wald que permite realizar inferencias sobre los parámetros lineales de regresión.

2.1. Procedimiento de Estimación Basado en Verosimilitud

En el contexto del modelo PLZIP introducido previamente en (1.30) y (1.31), sea W la variable que indica si la respuesta Y proviene de la distribución degenerada, es decir, si corresponde a un cero estructural ($W = 1$), o de la distribución Poisson ($W = 0$). Entonces, dada la muestra aleatoria completa $\{(Y_i, W_i, \mathbf{X}_i^\top, \mathbf{Z}_i^\top)\}_{i=1}^n$, la log-verosimilitud para los datos completos (\mathbf{Y}, \mathbf{W}) resulta

$$\begin{aligned} \ell_n^c(\mathbf{Y}, \mathbf{W}, \boldsymbol{\theta}) &= \sum_{i=1}^n (1 - W_i) \left(Y_i \left(\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i) \right) - e^{\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i)} - \log(Y_i!) \right) \\ &\quad + \sum_{i=1}^n \left(W_i \mathbf{Z}_i^\top \boldsymbol{\gamma} - \log(1 + e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}}) \right). \end{aligned} \quad (2.1)$$

Notar que la primera sumatoria depende únicamente de β y m , mientras que la segunda depende solo de γ . Esta separabilidad permite optimizar ambos términos de manera independiente, simplificando el procedimiento de estimación.

El hecho de que el primer término en (2.1) involucre una componente lineal en β y la función m , requiere combinar técnicas de estimación paramétricas y no paramétricas. Para abordar este punto, se introduce un procedimiento de estimación basado en la propuesta de Boente and Rodriguez (2010). Específicamente, se procede mediante los siguientes pasos:

(Paso 1) Para cada valor fijo $T = t$ sea $\eta = m(t) \in \mathbb{R}$, calcular

$$(\tilde{\beta}_n(t), \tilde{m}_n(t)) = \underset{\beta, \eta}{\operatorname{argmin}} \sum_{i=1}^n W_{i,n}(t)(1 - W_i) \left(e^{\mathbf{X}_i^\top \beta + \eta} - Y_i \left(\mathbf{X}_i^\top \beta + \eta \right) + \log(Y_i!) \right),$$

donde $W_{i,n}(t)$ son los pesos de tipo núcleo de Nadaraya-Watson definidos en (1.11).

(Paso 2) Una vez obtenida una estimación inicial de la componente no paramétrica, la misma se utiliza para estimar el parámetro lineal β , de manera que no dependa de t , de la siguiente manera

$$\hat{\beta}_n = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (1 - W_i) \left(e^{\mathbf{X}_i^\top \beta + \tilde{m}_n(T_i)} - Y_i \left(\mathbf{X}_i^\top \beta + \tilde{m}_n(T_i) \right) + \log(Y_i!) \right).$$

Mientras que el estimador de γ se obtiene como

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n \left(\log(1 + e^{\mathbf{Z}_i^\top \gamma}) - W_i \mathbf{Z}_i^\top \gamma \right).$$

(Paso 3) La estimación $\hat{\beta}_n$ obtenida en el Paso 2 se utiliza para mejorar la estimación de m , como sigue

$$\hat{m}_n(t) = \underset{\eta}{\operatorname{argmin}} \sum_{i=1}^n W_{i,n}(t)(1 - W_i) \left(e^{\mathbf{X}_i^\top \hat{\beta}_n + \eta} - Y_i \left(\mathbf{X}_i^\top \hat{\beta}_n + \eta \right) + \log(Y_i!) \right),$$

donde, nuevamente, esta estimación se realiza para cada valor fijo $T = t$.

El procedimiento descrito anteriormente se basa en la suposición de que la variable auxiliar W_i es observable, sin embargo, no es posible determinar si un cero en los datos es estructural o proviene del proceso de Poisson. En este contexto se utiliza la idea principal del algoritmo EM, que procede calculando la esperanza condicional de (2.1) dado el valor de la respuesta y luego maximizándola. Debido a que la log-verosimilitud de los datos completos es lineal en W_i , este paso es equivalente a reemplazar directamente W_i por su esperanza condicional $E(W_i|Y_i)$ en (2.1). Utilizando la regla de Bayes y un razonamiento análogo al realizado en (1.24), es sencillo demostrar que

$$E_{\theta}(W_i|Y_i) = \begin{cases} \frac{1}{1 + \exp\{-\mathbf{Z}_i^\top \gamma - e^{\mathbf{X}_i^\top \beta + m(T_i)}\}} & \text{si } Y_i = 0, \\ 0 & \text{si } Y_i > 0. \end{cases}$$

2.2. Procedimiento Robusto de Estimación

Para obtener estimadores robustos, se propone reemplazar las funciones de estimación utilizadas en el procedimiento de tres pasos basado en verosimilitud, por alternativas robustas. En esa dirección, sea $\rho : \mathbb{N}_0 \times \mathbb{R} \rightarrow \mathbb{R}$ una función de pérdida (sobre la cual se imponen condiciones en la Sección 2.4), $\omega_1 : \mathbb{R}^p \rightarrow \mathbb{R}$ y $\omega_2 : \mathbb{R}^q \rightarrow \mathbb{R}$ funciones de penalización que controlan, respectivamente, las predictoras de alta influencia \mathbf{X} y \mathbf{Z} . Se definen los siguientes funcionales de estimación

$$Q_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta}) = E_0 \left[(1 - E_{\boldsymbol{\theta}}(W|Y))\rho(y, \mathbf{X}^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X})|T = t \right], \quad (2.2)$$

$$Q_2(m, \boldsymbol{\beta}|\boldsymbol{\theta}) = E_0 \left[(1 - E_{\boldsymbol{\theta}}(W|Y))\rho(y, \mathbf{X}^\top \boldsymbol{\beta} + m(T))\omega_1(\mathbf{X}) \right], \quad (2.3)$$

$$Q_3(\boldsymbol{\gamma}|\boldsymbol{\theta}) = E_0 \left[\left(\log \left(1 + e^{\mathbf{Z}^\top \boldsymbol{\gamma}} \right) - E_{\boldsymbol{\theta}}(W|Y)\mathbf{Z}^\top \boldsymbol{\gamma} \right) \omega_2(\mathbf{Z}) \right], \quad (2.4)$$

donde E_0 denota la esperanza bajo el modelo verdadero, es decir, cuando el valor verdadero del parámetro es $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, m_0) \in \Theta$, donde Θ está definido en (1.32).

Se asume que las funciones ρ , ω_1 y ω_2 son tales que

$$\begin{aligned} (\tilde{\boldsymbol{\beta}}_0(t), m_0(t)) &= \underset{\eta, \boldsymbol{\beta}}{\operatorname{argmin}} Q_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta}), \\ \boldsymbol{\beta}_0 &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} Q_2(m, \boldsymbol{\beta}|\boldsymbol{\theta}), \\ \boldsymbol{\gamma}_0 &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} Q_3(\boldsymbol{\gamma}|\boldsymbol{\theta}), \end{aligned}$$

y la función de pérdida robusta ρ se selecciona de manera que se garantice la unicidad de $\boldsymbol{\beta}_0$ y de $m_0(\cdot)$.

La esperanza condicional a $T = t$ en (2.2) se puede estimar a través de un promedio pesado, cuyos pesos dependan de t . Así, la expresión empírica de (2.2) se construye mediante una versión robusta del enfoque de verosimilitud ponderada propuesto por Staniswalis (1989) y Severini and Staniswalis (1994), como sigue

$$Q_{n,1}(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta}) = \sum_{i=1}^n W_{i,n}(t)(1 - E_{\boldsymbol{\theta}}(W_i|Y_i))\rho(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X}_i). \quad (2.5)$$

Mientras que las versiones empíricas de (2.3) y (2.4) se obtienen reemplazando la verdadera distribución en el operador de esperanza por la distribución empírica de los datos observados, lo cual conduce a las siguientes expresiones

$$\begin{aligned} Q_{n,2}(m, \boldsymbol{\beta}|\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i))\rho(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i))\omega_1(\mathbf{X}_i), \\ Q_{n,3}(\boldsymbol{\gamma}|\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \left(\log \left(1 + e^{\mathbf{Z}_i^\top \boldsymbol{\gamma}} \right) - E_{\boldsymbol{\theta}}(W_i|Y_i)\mathbf{Z}_i^\top \boldsymbol{\gamma} \right) \omega_2(\mathbf{Z}_i). \end{aligned} \quad (2.6)$$

Cabe destacar que no se precisa una función de pérdida robusta en (2.6), ya que reducir la influencia de observaciones con covariables de alta influencia mediante la función de pesos ω_2 y asegurar una estimación precisa de W_i a través de su esperanza condicional (obtenida mediante una estimación robusta de β y m) es suficiente para preservar la robustez del estimador de γ .

Definición 2.2.1. Dada la muestra aleatoria $\{(Y_i, \mathbf{X}_i^\top, \mathbf{Z}_i^\top, T_i)\}_{i=1}^n \in \mathbb{N}_0 \times \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}$, que sigue el modelo PLZIP con parámetro $\theta_0 = (\beta_0, \gamma_0, m_0) \in \Theta$, el **estimador robusto de tres pasos** $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n, \hat{m}_n)$ para θ_0 queda determinado por:

(Paso 1)

$$(\tilde{\beta}_n(t), \tilde{m}_n(t)) = \underset{\beta, \eta}{\operatorname{argmin}} Q_{n,1}(\eta, \beta, t | \theta).$$

(Paso 2)

$$\hat{\beta}_n = \underset{\beta}{\operatorname{argmin}} Q_{n,2}(\tilde{m}_n, \beta | \theta), \quad (2.7)$$

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} Q_{n,3}(\gamma | \theta).$$

(Paso 3)

$$\hat{m}_n(t) = \underset{\eta}{\operatorname{argmin}} Q_{n,1}(\eta, \hat{\beta}_n, t | \theta).$$

2.2.1. Ecuaciones de Estimación

Los estimadores introducidos en la Definición 2.2.1 también se pueden obtener a través de las ecuaciones de estimación empíricas que se derivan de $Q_{n,1}$, $Q_{n,2}$ y $Q_{n,3}$. Específicamente, sea

$$S_n(\theta) = (S_{n,1}(\eta, \beta, t | \theta)^\top, S_{n,2}(m, \beta | \theta)^\top, S_{n,3}(\gamma | \theta)^\top)^\top$$

con

$$\begin{aligned} S_{n,1}(\eta, \beta, t | \theta) &= \sum_{i=1}^n W_{i,n}(t) (1 - E_\theta(W_i | Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \beta + \eta) \omega_1(\mathbf{X}_i) \mathbf{a}_i, \\ S_{n,2}(m, \beta | \theta) &= \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i | Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \beta + m(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i, \\ S_{n,3}(\gamma | \theta) &= \frac{1}{n} \sum_{i=1}^n \left(\left(1 + e^{-\mathbf{Z}_i^\top \gamma} \right)^{-1} - E_\theta(W_i | Y_i) \right) \omega_2(\mathbf{Z}_i) \mathbf{Z}_i, \end{aligned}$$

donde $\Psi(y, u) = \frac{\partial}{\partial u} \rho(y, u)$ y $\mathbf{a}_i = (1, \mathbf{X}_i^\top)^\top$. Entonces, el **sistema de ecuaciones de estimación empíricas** que da lugar a los estimadores es

$$S_n(\boldsymbol{\theta}) = \mathbf{0}. \quad (2.8)$$

Las ecuaciones de estimación bajo el modelo verdadero están dadas por

$$S_0(\boldsymbol{\theta}) = (S_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta})^\top, S_2(m, \boldsymbol{\beta}|\boldsymbol{\theta})^\top, S_3(\boldsymbol{\gamma}|\boldsymbol{\theta})^\top),$$

donde

$$\begin{aligned} S_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta}) &= E_0[(1 - E_\theta(W|Y))\Psi(y, \mathbf{X}^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X})\mathbf{a}|T = t], \\ S_2(m, \boldsymbol{\beta}|\boldsymbol{\theta}) &= E_0[(1 - E_\theta(W|Y))\Psi(y, \mathbf{X}^\top \boldsymbol{\beta} + m(T))\omega_1(\mathbf{X})\mathbf{X}^\top], \\ S_3(\boldsymbol{\gamma}|\boldsymbol{\theta}) &= E_0 \left[\left((1 + e^{-\mathbf{Z}^\top \boldsymbol{\gamma}})^{-1} - E_\theta(W|Y) \right) \omega_2(\mathbf{Z})\mathbf{Z}^\top \right], \end{aligned}$$

con $\mathbf{a} = (1, \mathbf{X}^\top)^\top$. Entonces, el **sistema de ecuaciones de estimación bajo el modelo verdadero** resulta

$$S_0(\boldsymbol{\theta}) = \mathbf{0}. \quad (2.9)$$

Nótese que si la función ρ es derivable con derivada Ψ y se elige de manera que se satisfaga la consistencia de Fisher, es fácil mostrar que $\boldsymbol{\theta}_0$ verifica (2.9). En efecto,

$$\begin{aligned} S_1(m_0(t), \boldsymbol{\beta}_0, t|\boldsymbol{\theta}_0) &= E_0[(1 - E_0(W|Y))\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))\omega_1(\mathbf{X})\mathbf{a}|T = t] \\ &= E_0[(1 - W)\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))\omega_1(\mathbf{X})\mathbf{a}|T = t] \\ &= (1 - \pi)E_0[(1 - W)\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))\omega_1(\mathbf{X})\mathbf{a}|(T = t, W = 0)] \\ &= (1 - \pi)E_0[E_0[\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))|(\mathbf{X}, T)]\omega_1(\mathbf{X})\mathbf{a}|T = t] = 0 \end{aligned}$$

y

$$\begin{aligned} S_2(m_0, \boldsymbol{\beta}_0|\boldsymbol{\theta}_0) &= E_0[(1 - E_0(W|Y))\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))\omega_1(\mathbf{X})\mathbf{X}^\top] \\ &= E_0((1 - W)\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))\omega_1(\mathbf{X})\mathbf{X}^\top) \\ &= (1 - \pi)E_0((1 - W)\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))\omega_1(\mathbf{X})\mathbf{X}^\top|W = 0) \\ &= (1 - \pi)E_0[E_0[\Psi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T))|(\mathbf{X}, T)]\omega_1(\mathbf{X})\mathbf{X}^\top] = 0. \end{aligned}$$

Además, independientemente de la elección de la función de pérdida, se verifica

$$\begin{aligned} S_3(\boldsymbol{\gamma}_0|\boldsymbol{\theta}_0) &= E_0 \left[\left((1 + e^{-\mathbf{Z}^\top \boldsymbol{\gamma}_0})^{-1} - E_0(W|Y) \right) \omega_2(\mathbf{Z})\mathbf{Z}^\top \right] \\ &= E_0 \left[\left((1 + e^{-\mathbf{Z}^\top \boldsymbol{\gamma}_0})^{-1} - W \right) \omega_2(\mathbf{Z})\mathbf{Z}^\top \right] \\ &= E_0 \left[\left((1 + e^{-\mathbf{Z}^\top \boldsymbol{\gamma}_0})^{-1} - E_0(W|\mathbf{Z}) \right) \omega_2(\mathbf{Z})\mathbf{Z}^\top \right] = 0. \end{aligned}$$

2.2.2. Función de Pérdida

Para asegurar la consistencia de Fisher de los funcionales de estimación propuestos, la función de pérdida ρ considerada en esta tesis se define como

$$\rho(y, \mu) = \phi(D(y, \mu)) + G(\mu) \quad (2.10)$$

donde $\phi(\cdot)$ es una función acotada y no decreciente para $x > 0$, con derivada continua $\phi(\cdot)$, y $G(\mu)$ es el término de corrección y $D(y, \mu)$ es la deviance, que en el caso de la distribución de Poisson viene dada por

$$\begin{aligned} D(y, \mu) &= \log f(y, y) - \log f(y, \mu) \\ &= \mu - y(\log \mu + 1 - \log y) \end{aligned}$$

cuya derivada respecto de μ es

$$\frac{\partial D(y, \mu)}{\partial \mu} = \frac{\mu - y}{\mu}.$$

Por otro lado, la derivada de la pérdida resulta

$$\Psi(y, \mu) = \phi'(D(y, \mu)) \frac{\partial D(y, \mu)}{\partial \mu} + G'(\mu).$$

Luego, tomando esperanza,

$$G'(\mu) = -E \left(\phi'(D(y, \mu)) \frac{\partial D(y, \mu)}{\partial \mu} \right).$$

Recordando que la función de probabilidad puntual del modelo de Poisson es $f(y, \mu) = \frac{e^{-\mu} \mu^y}{y!}$ se tiene

$$\begin{aligned} G'(\mu) &= -E \left(\phi(\mu - y(\log \mu + 1 - \log y)) \frac{\mu - y}{\mu} \right) \\ &= \sum_{y=0}^{\infty} \phi(\mu - y(\log \mu + 1 - \log y)) \frac{(y - \mu) e^{-\mu} \mu^y}{\mu y!}. \end{aligned}$$

Finalmente, integrando con respecto a μ ,

$$\begin{aligned} G(\mu) &= - \int \varphi(\mu) e^{-\mu} d\mu \\ &\quad + \sum_{y=1}^{\infty} \int \varphi(\mu - y(\ln(\mu) + 1 - \ln y)) \frac{(\mu^y e^{-\mu})}{y!} \frac{(y - \mu)}{\mu} d\mu. \end{aligned} \quad (2.11)$$

Haciendo el cambio de variable $v = \mu - y(\log \mu + 1 - \log y)$ es posible simplificar la expresión de la siguiente manera

$$G(\mu) = - \int \varphi(\mu) e^{-\mu} d\mu - \sum_{y=1}^{\infty} \frac{e^{-y} y^y}{y!} \int \phi(v) dv.$$

donde la integral en v depende del cambio de variable realizado.

2.3. Algoritmo Tipo EM

A continuación se introduce un algoritmo tipo EM combinado con el procedimiento de tres pasos que permite calcular el estimador propuesto en la Definición 2.2.1.

(Inicialización) Obtener un estimador inicial $\hat{\boldsymbol{\theta}}_n^{(0)} = (\hat{\boldsymbol{\beta}}_n^{(0)}, \hat{\boldsymbol{\gamma}}_n^{(0)}, \hat{m}_n^{(0)})$.

(Paso E) En la r -ésima iteración estimar W_i mediante

$$W_i^{(r)} = E_{\hat{\boldsymbol{\theta}}_n^{(r)}}(W_i | Y_i),$$

donde la esperanza se toma bajo el estimador actual del parámetro $\hat{\boldsymbol{\theta}}_n^{(r)}$.

(Paso M) Realizar el procedimiento de tres pasos:

(Paso 1) Para cada valor fijo $T = t$, computar

$$(\tilde{\boldsymbol{\beta}}_n(t), \tilde{m}_n(t)) = \underset{\boldsymbol{\beta}, \eta}{\operatorname{argmin}} Q_{n,1}(\eta, \boldsymbol{\beta}, t | \hat{\boldsymbol{\theta}}_n^{(r)}),$$

donde se denota $\eta = m(t)$.

(Paso 2) Calcular el estimador de $\boldsymbol{\beta}$ usando \tilde{m}_n obtenido en el Paso 1,

$$\hat{\boldsymbol{\beta}}_n^{(r+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} Q_{n,2}(\tilde{m}_n, \boldsymbol{\beta} | \hat{\boldsymbol{\theta}}_n^{(r)}) \quad (2.12)$$

y obtener el estimador de $\boldsymbol{\gamma}$ como

$$\hat{\boldsymbol{\gamma}}_n^{(r+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} Q_{n,3}(\boldsymbol{\gamma} | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

(Paso 3) Usando $\hat{\boldsymbol{\beta}}_n^{(r+1)}$ del Paso 2, actualizar el estimador de m como

$$\hat{m}_n^{(r+1)}(t) = \underset{\eta}{\operatorname{argmin}} Q_{n,1}(\eta, \hat{\boldsymbol{\beta}}_n^{(r+1)}, t | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

El estimador inicial se puede calcular con un estimador robusto para el modelo ZIP, por ejemplo el propuesto por Hall and Shen (2010). Mientras que el estimador inicial de m se puede derivar del residuo entre el logaritmo de la respuesta y el ajuste lineal.

Cabe observar que este enfoque difiere de aquel basado en perfiles de verosimilitud generalizada de Boente et al. (2006). Si se adaptara dicho enfoque al modelo PLZIP, para cada valor fijo de $\boldsymbol{\beta} \in \mathbb{R}^p$ se debería calcular

$$\tilde{m}_{\boldsymbol{\beta},n}(t) = \underset{\eta}{\operatorname{argmin}} Q_{n,1}(\eta, \boldsymbol{\beta}, t | \hat{\boldsymbol{\theta}}_n).$$

En consecuencia, sería necesario computar $\tilde{m}_{\beta,n}$ en cada punto de datos T_i sobre una grilla de posibles valores de β , con el fin de calcular la función objetivo $Q_{n,2}(\beta, \tilde{m}_{\beta,n}|\hat{\theta}_n)$, lo que implica un costo computacional considerable. En contraste, la estimación simultánea de β y η en el Paso 1 del algoritmo propuesto evita optimizaciones anidadas, ya que β y η se estiman conjuntamente para cada punto T_i , sin necesidad de hacerlo sobre una grilla de valores de β , reduciendo así el costo computacional.

2.3.1. Convergencia del algoritmo

En Hall and Shen (2010) los autores propusieron ecuaciones de estimación robustas para obtener los estimadores del modelo ZIP mediante el algoritmo ES. Rosen et al. (2000) mostró que, bajo ciertas condiciones de regularidad, si el algoritmo ES converge, lo hace hacia una solución de una función de estimación insesgada. A continuación, se demuestra un resultado similar para las ecuaciones de estimación definidas en (2.8).

Teorema 2.3.1. *Si existe un punto $\theta^* \in \Theta$ tal que $\lim_{r \rightarrow \infty} \hat{\theta}^{(r)} = \theta^*$, donde $\hat{\theta}^{(r)}$, $r = 0, 1, 2, \dots$ es una sucesión de estimadores generada por el algoritmo ES que deriva del algoritmo tipo EM propuesto, y ρ es tal que su derivada Ψ es continua y satisface $E_{\theta}(\Psi(Y, \mathbf{X}^T \beta + m(T)) | (\mathbf{X}, T)) = 0$, entonces $S_n(\theta^*) = \mathbf{0}$. Además, S_n es una ecuación de estimación insesgada, es decir, $E_{\theta}(S_n(\theta)) = 0$ para cada $\theta \in \Theta$.*

Demostración. Usando la Proposición A.2.1, es suficiente mostrar que se cumplen las siguientes condiciones:

- I. $E_{\theta}(S_{n,1}(\eta, \beta, t|\theta)) = \mathbf{0}$, $E_{\theta}(S_{n,2}(m, \beta|\theta)) = \mathbf{0}$ y $E_{\theta}(S_{n,3}(\gamma|\theta)) = \mathbf{0}$.
- II. $S_n(\theta)$ es una función continua en Θ .

La condición I se satisface para $S_{n,1}$, en efecto

$$\begin{aligned} E_{\theta}(S_{n,1}(\eta, \beta, t|\theta)) &= \sum_{i=1}^n E_{\theta}[W_{i,n}(t)(1 - W_i)\Psi(Y_i, \mathbf{X}_i^T \beta + \eta)\omega_1(\mathbf{X}_i)\mathbf{a}_i] \\ &= \sum_{i=1}^n E_{\theta}[E_{\theta}(W_{i,n}(t)(1 - W_i)\Psi(Y_i, \mathbf{X}_i^T \beta + \eta)\omega_1(\mathbf{X}_i)\mathbf{a}_i | W_i)] \\ &= \sum_{i=1}^n (1 - \pi_i) E_{\theta}(W_{i,n}(t)(1 - W_i)\Psi(Y_i, \mathbf{X}_i^T \beta + \eta)\omega_1(\mathbf{X}_i)\mathbf{a}_i | W_i = 0) \\ &= \sum_{i=1}^n (1 - \pi_i) E_{\theta}[W_{i,n}(t)E_{\theta}(\Psi(Y_i, \mathbf{X}_i^T \beta + \eta) | \mathbf{X}_i)\omega_1(\mathbf{X}_i)\mathbf{a}_i]. \end{aligned}$$

Como $E_{\theta}(\Psi(Y_i, \mathbf{X}_i^T \beta + \eta) | \mathbf{X}_i) = E_{\theta}[E_{\theta}(\Psi(Y_i, \mathbf{X}_i^T \beta + \eta) | (\mathbf{X}_i, T_i)) | \mathbf{X}_i] = 0$, se tiene que $E_{\theta}(S_{n,1}(\eta, \beta, t|\theta)) = 0$.

Para $S_{n,2}(m, \boldsymbol{\beta}|\boldsymbol{\theta})$ se puede reproducir un razonamiento similar,

$$\begin{aligned} E_{\boldsymbol{\theta}}(S_{n,2}(m, \boldsymbol{\beta}|\boldsymbol{\theta})) &= \frac{1}{n} \sum_{i=1}^n E_{\boldsymbol{\theta}}[(1 - E_{\boldsymbol{\theta}}(W_i|Y_i))\Psi(Y_i, \mathbf{X}_i^{\top} \boldsymbol{\beta} + m(T_i))\omega_1(\mathbf{X}_i) \mathbf{X}_i^{\top}] \\ &= \frac{1}{n} \sum_{i=1}^n (1 - \pi_i) E_{\boldsymbol{\theta}}((1 - W_i)\Psi(Y_i, \mathbf{X}_i^{\top} \boldsymbol{\beta} + m(T_i))\omega_1(\mathbf{X}_i) \mathbf{X}_i^{\top} | W_i = 0) \\ &= \frac{1}{n} \sum_{i=1}^n (1 - \pi_i) E_{\boldsymbol{\theta}}[E_{\boldsymbol{\theta}}(\Psi(Y_i, \mathbf{X}_i^{\top} \boldsymbol{\beta} + m(T_i)) | (\mathbf{X}_i, T_i))\omega_1(\mathbf{X}_i) \mathbf{X}_i^{\top}] = 0. \end{aligned}$$

Para $S_{n,3}(\boldsymbol{\gamma}|\boldsymbol{\theta})$ resulta trivial al condicionar en \mathbf{Z}_i .

Por otro lado, la condición II se satisface dado que ρ fue elegida como una función continuamente diferenciable.

□

Si bien el Teorema 2.3.1 no garantiza la convergencia del algoritmo, establece que si la sucesión generada por el algoritmo converge y se satisface la consistencia de Fisher, entonces su límite necesariamente verifica el sistema de ecuaciones de estimación $S_n(\boldsymbol{\theta}) = 0$, que resultan ser insesgadas.

Por otro lado, la unicidad de la solución del sistema de ecuaciones empírico no está garantizada, de modo que distintas inicializaciones del algoritmo pueden conducir (si el algoritmo converge) a límites diferentes. Por eso, en la práctica, la elección de una buena inicialización resulta esencial.

En las siguientes secciones de este capítulo se fundamentan las propiedades teóricas de los estimadores propuestos. En particular, se muestran resultados de consistencia fuerte, normalidad asintótica y se presenta un test de hipótesis.

2.4. Consistencia

En esta sección se muestran los resultados de consistencia fuerte del estimador robusto de tres pasos definido en 2.2.1. Los mismos se obtiene bajo las siguientes condiciones de regularidad, que se generalizan de las enunciadas por Boente and Rodriguez (2010).

- C1. $\rho(y, u)$ es continuamente diferenciable y acotada; $\omega_1(\cdot)$ y $\omega_2(\cdot)$ son funciones acotadas.
- C2. La función núcleo $K : \mathbb{R} \rightarrow \mathbb{R}$ es par, no negativa, acotada y continua con derivada acotada, y satisface $\int K(u)du = 1$, $\int u^2 K(u)du < \infty$ y $|u|K(u) \rightarrow 0$ si $|u| \rightarrow 0$.

- C3. La sucesión h_n satisface $h_n \rightarrow 0$ y $nh_n/\log(n) \rightarrow \infty$.
- C4. La densidad marginal de T , f_T , es una función acotada en \mathcal{T} . Sea $\mathcal{T}_0 \subset \mathcal{T}$ un conjunto compacto y sea $\mathcal{T}_\delta \subset \mathcal{T}$ la clausura de una δ -vecindad de \mathcal{T}_0 , existe una constante positiva $A_1(\mathcal{T}_\delta)$ tal que $A_1(\mathcal{T}_\delta) < f_T(t) \forall t \in \mathcal{T}_\delta$.
- C5. $Q_1(\eta, \beta, t|\theta)$ satisface la siguiente condición de equicontinuidad: dado $\varepsilon > 0 \exists \delta > 0$ tal que $|t_1 - t_2| < \delta$ para $t_i \in \mathcal{T}_0$ y $\|\beta_1 - \beta_2\| < \delta$ para $\beta_i \in \mathcal{K}$, $i = 1, 2$. Entonces

$$\sup_{\eta \in \mathbb{R}} |Q_1(\eta_1, \beta_1, t_1|\theta) - Q_1(\eta_2, \beta_2, t_2|\theta)| < \varepsilon.$$

Aquí $\eta_i = m(t_i)$ para $i = 1, 2$.

- C6. $Q_1(\eta, \beta, t|\theta)$ y $m_0(t)$ son funciones continuas.

A continuación se enuncian dos definiciones presentadas en Pollard (1984) que resultan relevantes para los resultados de esta sección. La primera de ellas brinda nociones sobre la complejidad de una familia de funciones, cuantificando las funciones que se necesitan para aproximar uniformemente a toda la clase con un error dado. La segunda definición analiza cómo una colección de conjuntos puede fraccionar o particionar un conjunto finito de N puntos. En particular, mide cuántas formas distintas tiene la clase de separar esos puntos cuando se la restringe a ese conjunto finito. Si el número de fraccionamientos posibles crece sólo de manera polinómica en N entonces la clase posee una complejidad controlada, en el sentido de que no puede generar una cantidad desmedida de particiones distintas.

Definición 2.4.1. Dado $\varepsilon > 0$ y una medida de probabilidad Q , el **número de cubrimiento** de una familia de funciones \mathcal{F} es el menor valor de $m = \mathcal{N}(\varepsilon, \mathcal{F}, L^1(Q))$ para el cual existen g_1, \dots, g_m (no necesariamente en \mathcal{F}) tales que

$$\min_j \int |f - g_j| dQ \leq \varepsilon \quad \forall f \in \mathcal{F}.$$

Si tal m no existe, se define $\mathcal{N}(\varepsilon, \mathcal{F}, L^1(Q)) = \infty$.

Definición 2.4.2. Sea \mathcal{D} una clase de subconjuntos de algún espacio S . Se dice que \mathcal{D} tiene **discriminación polinómica** (de grado v) si existe un polinomio $p(\cdot)$ de grado v tal que, para todo conjunto de N puntos en S , la clase selecciona a lo sumo $p(N)$ subconjuntos distintos. Formalmente, si S_0 consiste en N puntos, entonces hay como mucho $p(N)$ conjuntos distintos de la forma $S_0 \cap D$ con $D \in \mathcal{D}$.

$p(\cdot)$ se denomina **polinomio discriminante** para \mathcal{D} , y las clases que tienen discriminación polinómica se denominan **clases polinómicas**.

El Lema de Aproximación A.3.5 del Apéndice A relaciona estas dos definiciones mostrando, bajo ciertas condiciones, que si los gráficos de la familia de funciones \mathcal{F} forman una clase polinómica, entonces, para cada $0 < \varepsilon < 1$ el número de cubrimiento de la familia está acotado por $A\varepsilon^{-W}$, donde las constantes A y W dependen únicamente del polinomio discriminante de la clase de gráficos. Más detalles sobre estos tópicos pueden encontrarse en Pollard (1984) y Van der Vaart and Wellner (1996).

En lo que sigue se presenta una serie de resultados que establecen la consistencia fuerte, tanto de la componente no paramétrica como de los parámetros lineales del modelo. Específicamente, el Lema 2.4.1 demuestra que los funcionales empíricos que intervienen en la estimación convergen uniformemente hacia sus contrapartes poblacionales. Esta convergencia garantiza que los estimadores preliminares $\tilde{\beta}_n$ y \tilde{m}_n se aproximan uniformemente a las cantidades verdaderas β_0 y m_0 . El Teorema 2.4.1 muestra que, establecidas dichas convergencias uniformes, el funcional asociado al parámetro lineal β también converge uniformemente a su contraparte poblacional y ello implica la convergencia casi segura del estimador $\hat{\beta}_n$ hacia β_0 . Finalmente, en el Corolario 2.4.1, la consistencia del estimador $\hat{\beta}_n$ combinada con la convergencia uniforme del estimador preliminar \tilde{m}_n , asegura que la estimación final de la componente no paramétrica \hat{m}_n también converge uniformemente a la función verdadera m_0 . En conjunto, estos resultados establecen la consistencia fuerte de los componentes $\tilde{\beta}_n$ y \tilde{m}_n , uniformemente en t , y de $\hat{\beta}_n$ y \hat{m}_n , involucrados en el procedimiento de estimación.

Lema 2.4.1. *Sea $\mathcal{K} \subset \mathbb{R}^p$ un conjunto compacto, se supone que C1-C6 se cumplen y que K es una función de variación acotada. Se considera la familia de funciones*

$$\mathcal{F} = \{f_{\eta,\beta}(Y, \mathbf{X}|\boldsymbol{\theta}) = (1 - E_{\boldsymbol{\theta}}(W|Y))\rho(y, \mathbf{X}^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X}), \beta \in \mathcal{K}, \eta \in \mathbb{R}\}.$$

Si existen constantes A y W tales que \mathcal{F} tiene un número de cubrimiento que satisface

$$\sup_Q \mathcal{N}(\varepsilon, \mathcal{F}, L^1(Q)) \leq A\varepsilon^{-W}$$

para $0 < \varepsilon < 1$, entonces,

$$(a) \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|Q_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}) - Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})\|_{0,\infty} \xrightarrow{c.s.} 0.$$

$$(b) \|\tilde{m}_n - m_0\|_{0,\infty} \xrightarrow{c.s.} 0 \text{ y } \|\tilde{\beta}_n - \beta_0\|_{0,\infty} \xrightarrow{c.s.} 0.$$

Demostración. Sea

$$R_{n,1}(\eta, \beta, t|\boldsymbol{\theta}) = (nh_n)^{-1} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i))\rho(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X}_i)K((T_i - t)/h_n)$$

$$R_{n,0}(t) = (nh_n)^{-1} \sum_{j=1}^n K\left(\frac{T_j - t}{h_n}\right)$$

y $Q_{n,1}(\eta, \beta, t|\boldsymbol{\theta}) = R_{n,1}(\eta, \beta, t|\boldsymbol{\theta})/R_{n,0}(t)$. Entonces,

$$\begin{aligned} & \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|Q_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}) - Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})\|_{0,\infty} = \\ & \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}) - Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})R_{n,0}(\cdot)\|_{0,\infty} \left(\inf_{t \in \mathcal{T}_0} R_{n,0}(t) \right)^{-1} \leq \\ & \left\{ \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}) - E_0(R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}))\|_{0,\infty} \right. \\ & + \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|E_0(R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta})) - Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})E_0(R_{n,0}(\cdot))\|_{0,\infty} \\ & \left. + \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})\|_{0,\infty} \|R_{n,0}(\cdot) - E_0(R_{n,0}(\cdot))\|_{0,\infty} \right\} \left(\inf_{t \in \mathcal{T}_0} R_{n,0}(t) \right)^{-1}. \end{aligned}$$

La condición C4 implica que $E_0(R_0(t)) = \int K(u) f_T(t - uh_n) du > A_1(\mathcal{T}_\delta)$. Por lo tanto, basta probar las siguientes convergencias:

- (A) $\sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}) - E_0(R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta}))\|_{0,\infty} \xrightarrow{c.s.} 0$,
- (B) $\sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|E_0(R_{n,1}(\eta, \beta, \cdot|\boldsymbol{\theta})) - Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})E_0(R_{n,0}(\cdot))\|_{0,\infty} \xrightarrow{c.s.} 0$,
- (C) $\sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|Q_1(\eta, \beta, \cdot|\boldsymbol{\theta})\|_{0,\infty} \|R_{n,0}(\cdot) - E_0(R_{n,0}(\cdot))\|_{0,\infty} \xrightarrow{c.s.} 0$.

Para demostrar (A), se aplica el Teorema A.3.2 a la familia de funciones

$$\mathcal{F}_n = \left\{ f_{\eta, \beta, t}^n(Y, \mathbf{X}, T|\boldsymbol{\theta}) = \frac{(1 - E_{\boldsymbol{\theta}}(W|Y))\rho(y, \mathbf{X}^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X})K((T - t)/h_n)}{\|\rho\|_\infty \|\omega_1\|_\infty \|K\|_\infty}, \right. \\ \left. \beta \in \mathcal{K}, \eta \in \mathbb{R}, t \in \mathcal{T}_0 \right\}$$

donde $\|\rho\|_\infty = \sup_{y,u} |\rho(y, u)|$, $\|\omega_1\|_\infty = \sup_{\mathbf{X}} |\omega_1(\mathbf{X})|$ y $\|K\|_\infty = \sup_v |K(v)|$. Nótese que \mathcal{F}_n depende de n a través de h_n . Para garantizar las condiciones del teorema se considera la familia

$$\mathcal{K}_n = \left\{ k_t^n(T) = K\left(\frac{T - t}{h_n}\right), t \in \mathcal{T}_0 \right\}$$

donde K es una función de variación acotada en \mathbb{R} que satisface $K \leq \|K\|_\infty$. Entonces, por el Lema A.3.3, los gráficos de funciones en \mathcal{K}_n tienen discriminación polinómica. Por lo tanto, por el Lema de Aproximación A.3.5, existen valores reales A_1 y W_1 (que dependen solo del polinomio discriminante) tales que para todo $0 < \varepsilon < 1$ y una medida de probabilidad Q ,

$$\mathcal{N}(\varepsilon \|K\|_\infty, \mathcal{K}_n, L^1(Q)) \leq A_1 \varepsilon^{-W_1}.$$

Así, existen constantes A_2 y W_2 tales que

$$\sup_Q \mathcal{N}(\varepsilon \|K\|_\infty, \mathcal{F}_n, L^1(Q)) < A_2 \varepsilon^{-W_2}, \quad \forall 0 < \varepsilon < 1.$$

Además, toda $f_{\eta,\beta,t}^n \in \mathcal{F}_n$ está acotada por 1 ya que $|1 - E_{\theta}(W_i|Y_i)| < 1$. Más aún,

$$\begin{aligned}
 E_0(f_{\eta,\beta,t}^n(Y, \mathbf{X}, T|\boldsymbol{\theta}))^2 &\leq E_0(f_{\eta,\beta,t}^n(Y, \mathbf{X}, T|\boldsymbol{\theta})) \\
 &= E_0\left(\frac{(1 - E_{\theta}(W|Y))\rho(y, \mathbf{X}^\top \boldsymbol{\beta} + \eta)\omega_1(\mathbf{X})K((T-t)/h_n)}{\|\rho\|_\infty \|\omega_1\|_\infty \|K\|_\infty}\right) \\
 &\leq E_0\left(\frac{K((T-t)/h_n)}{\|K\|_\infty}\right) \\
 &= \frac{1}{\|K\|_\infty} \int K\left(\frac{T-t}{h_n}\right) f_T(T) dT \\
 (l = (T-t)/h_n) &= \frac{h_n}{\|K\|_\infty} \int K(l) f_T(lh_n + t) dl \\
 &\leq \frac{h_n \|f_T\|_\infty}{\|K\|_\infty} \int K(l) dl \\
 &= h_n \|f_T\|_\infty \|K\|_\infty^{-1}
 \end{aligned}$$

Tomando $\alpha_n = 1$ y $\delta_n^2 = h_n \|f_T\|_\infty \|K\|_\infty^{-1}$ se tiene que $n\delta_n^2 \alpha_n^2 \gg \log(n)$ y

$$\sup_{\mathcal{F}_n} \left\| \frac{1}{n} \sum_{i=1}^n f_{\eta,\beta,t}^n(Y_i, \mathbf{X}_i, T_i|\boldsymbol{\theta}) - E_0(f_{\eta,\beta,t}^n(Y, \mathbf{X}, T|\boldsymbol{\theta})) \right\| \ll \delta_n^2 \alpha_n \quad \text{casi seguramente.}$$

Esto implica que

$$\frac{1}{\|f_T\|_\infty \|\rho\|_\infty \|\omega_1\|_\infty} \sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|R_{n,1}(\eta, \boldsymbol{\beta}, \cdot|\boldsymbol{\theta}) - E_0(R_{n,1}(\eta, \boldsymbol{\beta}, \cdot|\boldsymbol{\theta}))\|_{0,\infty} \xrightarrow{c.s.} 0,$$

y el primer factor es finito, por lo que (A) queda demostrado.

Para demostrar (B), nótese que

$$\begin{aligned}
 &|E_0(R_{n,1}(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta})) - Q_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta})E_0(R_{n,0}(t))| \\
 &= h_n^{-1} \left| \int Q_1(\eta, \boldsymbol{\beta}, T|\boldsymbol{\theta}) K\left(\frac{T-t}{h_n}\right) f_T(T) dT \right. \\
 &\quad \left. - \int Q_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta}) K\left(\frac{T-t}{h_n}\right) f_T(T) dT \right| \\
 &= \left| \int (Q_1(\eta, \boldsymbol{\beta}, lh_n + t|\boldsymbol{\theta}) - Q_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta})) K(l) f_T(lh_n + t) dl \right| \\
 &\leq \|f_T\|_\infty \int |Q_1(\eta, \boldsymbol{\beta}, lh_n + t|\boldsymbol{\theta}) - Q_1(\eta, \boldsymbol{\beta}, t|\boldsymbol{\theta})| K(l) dl.
 \end{aligned}$$

El hecho de que $h_n \rightarrow 0$ y la condición C5 implican que

$$\sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|E_0(R_{n,1}(\eta, \boldsymbol{\beta}, \cdot|\boldsymbol{\theta})) - Q_1(\eta, \boldsymbol{\beta}, \cdot|\boldsymbol{\theta})E_0(R_{n,0}(\cdot))\|_{0,\infty} \rightarrow 0.$$

Para demostrar (C) es suficiente ver que

$$\|R_{n,0}(\cdot) - E_0(R_{n,0}(\cdot))\|_{0,\infty} \xrightarrow{c.s.} 0.$$

Dado que $|1 - E_{\theta}(W|Y)| < 1$,

$$\sup_{\beta \in \mathcal{K}, \eta \in \mathbb{R}} \|Q_1(\eta, \beta, \cdot | \boldsymbol{\theta})\|_{0, \infty} \leq \|\rho\|_{\infty} \|\omega_1\|_{\infty}$$

con ρ y ω_1 acotadas.

Las hipótesis del Teorema A.3.2 se verifican para la familia \mathcal{K}_n . En efecto, sin pérdida de generalidad, si la función núcleo está acotada por 1,

$$\begin{aligned} \int K((T-t)/h_n)^2 f_T(T) dT &\leq \int K((T-t)/h_n) f_T(T) dT \\ &= h_n \int K(l) f_T(lh_n + t) dl \\ &\leq h_n \|f_T\|_{\infty} \end{aligned}$$

y tomando $\alpha_n = 1$, por hipótesis $n\delta_n^2\alpha_n^2 \gg \log(n)$. Así,

$$\sup_{\mathcal{K}_n} \left| \frac{1}{n} \sum_{i=1}^n K((T_i - t)/h_n) - E_0(K((T-t)/h_n)) \right| \ll \|f_T\|_{\infty} h_n \quad c.s$$

y luego

$$\sup_{\mathcal{K}_n} |R_{n,0}(t) - E_0(R_{n,0}(t))| \xrightarrow{c.s.} 0$$

lo que concluye la prueba de (a).

Usando (a) y la condición C6, la aplicación del Lema A.1.4 concluye la prueba de (b). En la notación del Lema A1, se toma $f = Q_1$ con minimizador único $(m_0(t), \beta_0) \in \mathcal{K} \times \mathcal{C}$ y $f_n = Q_{n,1}$ con un minimizador $(\tilde{\beta}_n(t), \tilde{m}_n(t))$. Nótese que la continuidad de m implica que existe un conjunto compacto $\mathcal{A} \in \mathbb{R}$ tal que $m(t) \in \mathcal{A}$ para todo $t \in \mathcal{T}_0$. □

Teorema 2.4.1. *Sea Ψ acotada, si las condiciones del Lema 2.4.1 se satisfacen y la familia de funciones*

$$\mathcal{H} = \{h_{\beta}(Y, \mathbf{X}, T | \boldsymbol{\theta}) = (1 - E_{\theta}(W|Y))\rho(Y, \mathbf{X}^{\top} \beta + m_0(T)), \beta \in \mathbb{R}^p\}$$

tiene número de cubrimiento tal que $\log \mathcal{N}(\varepsilon, \mathcal{H}, L^1(P_n)) = o_p(n)$, donde P_n es la distribución empírica de (Y, \mathbf{X}, T) . Entonces, se satisfacen las siguientes condiciones,

(a) $\sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(\tilde{m}_n, \beta | \boldsymbol{\theta}) - Q_2(m_0, \beta | \boldsymbol{\theta})| \xrightarrow{c.s.} 0$.

(b) *Si $Q_2(m_0, \beta | \boldsymbol{\theta})$ tiene un único mínimo β_0 , \mathbf{X} tiene segundo momento finito y se cumple que $\liminf_{\|\beta\| \rightarrow \infty} Q_2(m_0, \beta | \boldsymbol{\theta}) > Q_2(m_0, \beta_0 | \boldsymbol{\theta})$, entonces $\hat{\beta}_n \xrightarrow{c.s.} \beta_0$.*

Demostración. Para probar (a), nótese que

$$\begin{aligned} \sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(\tilde{m}, \beta|\theta) - Q_2(m_0, \beta|\theta)| &\leq \sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(\tilde{m}, \beta|\theta) - Q_{n,2}(m_0, \beta|\theta)| \\ &\quad + \sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(m_0, \beta|\theta) - Q_2(m_0, \beta|\theta)|. \end{aligned}$$

Por lo tanto, es suficiente mostrar las convergencias

$$(A) \sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(\tilde{m}, \beta|\theta) - Q_{n,2}(m_0, \beta|\theta)| \xrightarrow{c.s.} 0,$$

$$(B) \sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(m_0, \beta|\theta) - Q_2(m_0, \beta|\theta)| \xrightarrow{c.s.} 0.$$

Sea $\mathcal{T}_0 \in \mathcal{T}$ un conjunto compacto tal que $P(T_i \notin \mathcal{T}_0) < \varepsilon$. El desarrollo de Taylor (Teorema A.1.5) de primer orden de ρ alrededor de $\mathbf{X}_i^\top \beta + m_0(T_i)$ para cada $i = 1, \dots, n$ implica que

$$\begin{aligned} &\sup_{\beta \in \mathbb{R}^p} |Q_{n,2}(\tilde{m}, \beta|\theta) - Q_{n,2}(m_0, \beta|\theta)| \\ &\leq \sup_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n |\Psi(Y_i, u_i)| |\tilde{m}(T_i) - m_0(T_i)| |\omega_1(\mathbf{X}_i)| \mathbb{I}(T_i \in \mathcal{T}_0) \\ &\quad + \sup_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \left(|\rho(Y_i, \mathbf{X}_i^\top \beta + \tilde{m}(T_i))| + |\rho(Y_i, \mathbf{X}_i^\top \beta + m_0(T_i))| \right) |\omega_1(\mathbf{X}_i)| \mathbb{I}(T_i \notin \mathcal{T}_0) \\ &\leq \|\Psi\|_\infty \|\tilde{m} - m_0\|_{0,\infty} \|\omega_1\|_\infty + 2\|\rho\|_\infty \|\omega_1\|_\infty \frac{1}{n} \sum_{i=1}^n \mathbb{I}(T_i \notin \mathcal{T}_0), \end{aligned}$$

donde $u_i \in (\mathbf{X}_i^\top \beta + \tilde{m}(T_i), \mathbf{X}_i^\top \beta + m_0(T_i)) \forall i = 1, \dots, n$.

Dado que Ψ y ω_1 son acotadas, y $\|\tilde{m} - m_0\|_{0,\infty} \xrightarrow{c.s.} 0$ en \mathcal{T}_0 , el primer término tiende a 0 casi seguramente. Respecto al segundo término, dado que ρ es acotada y $E(\mathbb{I}(T_i \notin \mathcal{T}_0)) = P(T_i \notin \mathcal{T}_0) < \varepsilon$, la Ley Fuerte de los Grandes Números (Teorema A.1.1) implica que el segundo término tiende a 0 casi seguramente. Por lo tanto (A) se cumple.

Por otra parte, la hipótesis sobre el número de cubrimiento de \mathcal{H} , el hecho de que las funciones en \mathcal{H} tienen la cota uniforme $\|\rho\|_\infty \|\omega_1\|_\infty$ y el Teorema A.3.1 garantizan que (B) se cumple, lo que concluye la demostración de (a).

Para probar (b), sea $\hat{\beta}_{n_k}$ una subsucesión convergente de $\hat{\beta}_n$ con límite $\bar{\beta}$. Se supone sin pérdida de generalidad que $\hat{\beta}_n \xrightarrow{c.s.} \bar{\beta}$.

Si $\|\bar{\beta}\| < \infty$, de (a) se tiene que

$$Q_{n,2}(\hat{m}_n, \beta_0|\theta) - Q_2(m_0, \beta_0|\theta) \xrightarrow{c.s.} 0.$$

Además,

$$\begin{aligned} |Q_{n,2}(\tilde{m}, \hat{\beta}_n|\theta) - Q_2(m_0, \bar{\beta}|\theta)| &\leq |Q_{n,2}(\tilde{m}, \hat{\beta}_n|\theta) - Q_2(m_0, \hat{\beta}_n|\theta)| \\ &\quad + |Q_2(m_0, \hat{\beta}_n|\theta) - Q_2(m_0, \bar{\beta}|\theta)|. \end{aligned}$$

El primer término tiende a 0 casi seguramente por (a). En cuanto al segundo término, el desarrollo de Taylor de primer orden de ρ alrededor de $\mathbf{X}^\top \bar{\beta} + m_0(T)$ implica que

$$\begin{aligned} |Q_2(m_0, \hat{\beta}_n | \boldsymbol{\theta}) - Q_2(m_0, \bar{\beta} | \boldsymbol{\theta})| &\leq E_0 \left(|1 - E_{\boldsymbol{\theta}}(w|y)| |\rho(y, \mathbf{X}^\top \hat{\beta}_n + m_0(T)) \right. \\ &\quad \left. - \rho(Y, \mathbf{X}^\top \bar{\beta} + m_0(T)) | |\omega_1(\mathbf{X})| \right) \\ &\leq E_0 \left(|\Psi(Y, u)| |\mathbf{X}^\top (\hat{\beta}_n - \bar{\beta})| |\omega_1(\mathbf{X})| \right) \\ &\leq \|\Psi\|_\infty E_0(\|\mathbf{X}^\top\|) \|\hat{\beta}_n - \bar{\beta}\| \|\omega_1\|_\infty, \end{aligned}$$

donde $u \in (\mathbf{X}^\top \hat{\beta}_n + \eta, \mathbf{X}^\top \bar{\beta} + \eta)$, Ψ y ω_1 son funciones acotadas, \mathbf{X} tiene segundo momento finito, y $\hat{\beta}_n \xrightarrow{c.s.} \bar{\beta}$. Por lo tanto,

$$Q_2(m_0, \hat{\beta}_n | \boldsymbol{\theta}) - Q_2(m_0, \bar{\beta} | \boldsymbol{\theta}) \xrightarrow{c.s.} 0,$$

lo cual implica

$$Q_{n,2}(\tilde{m}, \hat{\beta}_n | \boldsymbol{\theta}) - Q_2(m_0, \bar{\beta} | \boldsymbol{\theta}) \xrightarrow{c.s.} 0.$$

Nótese que $Q_{n,2}(\tilde{m}, \hat{\beta}_n | \boldsymbol{\theta}) \leq Q_{n,2}(\tilde{m}, \beta_0 | \boldsymbol{\theta})$, ya que

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} Q_{n,2}(\tilde{m}, \beta | \boldsymbol{\theta}).$$

Entonces, tomando límite casi seguro, se tiene que

$$Q_2(m_0, \bar{\beta} | \boldsymbol{\theta}) \leq Q_2(m_0, \beta_0 | \boldsymbol{\theta}),$$

pero (β_0, m_0) es el único mínimo de Q_2 , por lo que debe cumplirse $\bar{\beta} = \beta_0$.

Si $\|\bar{\beta}\| = \infty$, por (a) y la continuidad de Q_2 ,

$$Q_{n,2}(\tilde{m}, \beta_0 | \boldsymbol{\theta}) - Q_2(m_0, \beta_0 | \boldsymbol{\theta}) \xrightarrow{c.s.} 0 \quad \text{y} \quad Q_{n,2}(\tilde{m}, \hat{\beta}_n | \boldsymbol{\theta}) - Q_2(m_0, \bar{\beta} | \boldsymbol{\theta}) \xrightarrow{c.s.} 0.$$

Además, $Q_{n,2}(\tilde{m}, \hat{\beta}_n | \boldsymbol{\theta}) \leq Q_{n,2}(\tilde{m}, \beta_0 | \boldsymbol{\theta})$. Por lo tanto

$$\begin{aligned} 0 &\geq Q_{n,2}(\tilde{m}, \hat{\beta}_n | \boldsymbol{\theta}) - Q_{n,2}(\tilde{m}, \beta_0 | \boldsymbol{\theta}) \geq Q_{n,2}(\tilde{m}, \hat{\beta}_n | \boldsymbol{\theta}) - Q_2(m_0, \bar{\beta} | \boldsymbol{\theta}) \\ &\quad - (Q_{n,2}(\tilde{m}, \beta_0 | \boldsymbol{\theta}) - Q_2(m_0, \beta_0 | \boldsymbol{\theta})) \\ &\quad + Q_2(m_0, \bar{\beta} | \boldsymbol{\theta}) - Q_2(m_0, \beta_0 | \boldsymbol{\theta}). \end{aligned}$$

Tomando límite casi seguro,

$$\lim_{\|\beta\| \rightarrow \infty} Q_2(m_0, \beta | \boldsymbol{\theta}) \geq Q_2(m_0, \beta_0 | \boldsymbol{\theta}),$$

lo cual, por hipótesis, es absurdo. Por lo tanto, debe cumplirse $\|\bar{\beta}\| < \infty$.

□

Observación 2.4.1. *Aplicando resultados de la Sección A.3 se puede ver que la condición sobre el número de cubrimiento de la familia \mathcal{F} en el Lema 2.4.1 se cumple bajo condiciones generales cuando ρ está definida como en (2.10). En efecto, si ϕ y $G(e^u)$ tienen variación acotada, entonces, por el Lema A.3.3, el número de cubrimiento de la clase $\{G(e^{\mathbf{X}^\top \beta + \eta}), \beta \in \mathbb{R}^q, \eta \in \mathbb{R}\}$ y de la clase $\{\phi(e^{\mathbf{X}^\top \beta + \eta} - Y(\mathbf{X}^\top \beta + \eta + 1 - \ln Y)), \beta \in \mathbb{R}^p, \eta \in \mathbb{R}\}$ crecen a una tasa polinómica. Luego se puede aplicar el Lema A.3.1 para concluir que la suma de las clases es polinómica. Asimismo, los lemas A.3.6 y A.3.2 permiten concluir que $\{(1 - E_\theta(W|Y)), \beta \in \mathbb{R}^p, \eta \in \mathbb{R}\}$ es una clase polinómica. Por lo tanto, del Lema A.3.4 se tiene el resultado.*

Un razonamiento similar puede aplicarse a la condición sobre el número de cubrimiento de la familia de funciones \mathcal{H} en el Teorema 2.4.1.

Corolario 2.4.1. *Bajo las hipótesis del Teorema 2.4.1 se tiene que $\|\widehat{m}_n - m_0\|_{0,\infty} \xrightarrow{c.s.} 0$.*

Demostración. Nótese que

$$\begin{aligned} \sup_{\eta \in \mathbb{R}} \|Q_{n,1}(\eta, \widehat{\beta}_n, \cdot | \boldsymbol{\theta}) - Q_1(\eta, \beta_0, \cdot | \boldsymbol{\theta})\|_{0,\infty} &\leq \sup_{\eta \in \mathbb{R}} \|Q_{n,1}(\eta, \widehat{\beta}_n, \cdot | \boldsymbol{\theta}) - Q_{n,1}(\eta, \beta_0, \cdot | \boldsymbol{\theta})\|_{0,\infty} \\ &\quad + \sup_{\eta \in \mathbb{R}} \|Q_{n,1}(\eta, \beta_0, \cdot | \boldsymbol{\theta}) - Q_1(\eta, \beta_0, \cdot | \boldsymbol{\theta})\|_{0,\infty}. \end{aligned}$$

El segundo término converge casi seguramente a 0 por el Lema 2.4.1, ya que, al cumplirse para el supremo en β , se cumple en particular para β_0 . Respecto del primer término, la expansión de Taylor de primer orden de ρ alrededor de $\mathbf{X}_i \beta_0 + \eta$ para cada $i = 1, \dots, n$ deja

$$\begin{aligned} &\|Q_{n,1}(\eta, \widehat{\beta}_n, \cdot | \boldsymbol{\theta}) - Q_{n,1}(\eta, \beta_0, \cdot | \boldsymbol{\theta})\|_{0,\infty} \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i | Y_i)) \Psi(Y_i, u_i) \mathbf{X}_i^\top (\widehat{\beta}_n - \beta_0) \omega_1(\mathbf{X}_i) \right\|_{0,\infty} \\ &\leq \|\Psi\|_\infty \|\widehat{\beta}_n - \beta_0\| \|\omega_1\|_\infty \frac{1}{n} \sum_{i=1}^n \|\mathbf{X}_i\| \end{aligned}$$

donde $u_i \in (\mathbf{X}_i \widehat{\beta}_n + \eta, \mathbf{X}_i \beta_0 + \eta)$ para cada $i = 1, \dots, n$.

Por hipótesis, el segundo momento de \mathbf{X} es finito. Entonces, debido a la ley fuerte de los grandes números, $\sum_{i=1}^n \|\mathbf{X}_i\|$ es finito para $n > N$. Además, dado que Ψ y ω_1 son acotadas y $\widehat{\beta}_n \xrightarrow{c.s.} \beta_0$ por el Teorema 2.4.1, se verifica que

$$\sup_{\eta \in \mathbb{R}} \|Q_{n,1}(\eta, \widehat{\beta}_n, \cdot | \boldsymbol{\theta}) - Q_1(\eta, \beta_0, \cdot | \boldsymbol{\theta})\|_{0,\infty} \xrightarrow{c.s.} 0.$$

Finalmente, una aplicación del Lema A.1.4 concluye la demostración. □

El siguiente teorema establece la consistencia de $\hat{\gamma}_n$ y se basa en un resultado análogo al de Huber (1967) para la consistencia del estimador de máxima verosimilitud bajo condiciones más débiles de lo habitual.

Teorema 2.4.2. *Se supone que existe N_0 tal que $\hat{\gamma}_n \in \mathcal{K}_1$ para todo $n > N_0$, siendo \mathcal{K}_1 un conjunto compacto. Se supone además que se cumplen las siguientes condiciones:*

I. *Sea \mathcal{B} un entorno abierto de γ' ,*

$$E_0 \left[\inf_{\gamma \in \mathcal{B}} \left(\log \left(1 + e^{\mathbf{Z}^T \gamma} \right) - E_{\theta}(W|Y) \mathbf{Z}^T \gamma \right) \omega_2(\mathbf{Z}) \right] \rightarrow Q_3(\gamma'|\theta)$$

cuando $\mathcal{B} \rightarrow \gamma'$.

II. *Existe γ_0 tal que $Q_3(\gamma_0|\theta) < Q_3(\gamma|\theta)$ para todo γ .*

Entonces, $\hat{\gamma}_n \xrightarrow{c.s.} \gamma_0$ cuando $n \rightarrow \infty$.

Demostración. Sea \mathcal{B}_0 un entorno abierto de γ_0 , de modo que $\mathcal{K}_1 \setminus \mathcal{B}_0$ es un conjunto compacto. Se probará que

$$\inf_{\gamma \in \mathcal{B}_0} Q_{n,3}(\gamma|\theta) \leq \inf_{\gamma \in \mathcal{K}_1 \setminus \mathcal{B}_0} Q_{n,3}(\gamma|\theta) \quad c.s \quad (2.13)$$

Nótese que $\mathcal{K}_1 \setminus \mathcal{B}_0$ tiene un cubrimiento finito de entornos abiertos \mathcal{B}_s centrados en γ_s , con s entero, $1 \leq s \leq S$. Entonces, el ínfimo de $Q_{n,3}(\gamma|\theta)$ se alcanza en $\mathcal{K}_1 \setminus \mathcal{B}_0$ y, en particular, existe s' tal que el ínfimo se alcanza en $\mathcal{B}_{s'}$, de modo que

$$\begin{aligned} \inf_{\gamma \in \mathcal{K}_1 \setminus \mathcal{B}_0} Q_{n,3}(\gamma|\theta) &= \inf_{\gamma \in \mathcal{B}_{s'}} Q_{n,3}(\gamma|\theta) \\ &\geq \frac{1}{n} \sum_{i=1}^n \inf_{\gamma \in \mathcal{B}_{s'}} \left(\log \left(1 + e^{\mathbf{Z}_i^T \gamma} \right) - E_{\theta}(W_i|Y_i) \mathbf{Z}_i^T \gamma \right) \omega_2(\mathbf{Z}_i) \\ &\xrightarrow{c.s.} E_0 \left[\inf_{\gamma \in \mathcal{B}_{s'}} \left(\log \left(1 + e^{\mathbf{Z}^T \gamma} \right) - E_{\theta}(W|Y) \mathbf{Z}^T \gamma \right) \omega_2(\mathbf{Z}) \right] \end{aligned}$$

donde la convergencia casi segura se cumple por el Teorema A.3.1.

Luego, por I, si $\mathcal{B}_{s'} \rightarrow \{\gamma_{s'}\}$,

$$E_0 \left[\inf_{\gamma \in \mathcal{B}_{s'}} \left(\log \left(1 + e^{\mathbf{Z}^T \gamma} \right) - E_0(W|Y) \mathbf{Z}^T \gamma \right) \omega_2(\mathbf{Z}) \right] \rightarrow Q_3(\gamma_{s'}|\theta) > Q_3(\gamma_0|\theta).$$

Por lo tanto,

$$Q_3(\gamma_0|\theta) < \inf_{\gamma \in \mathcal{K}_1 \setminus \mathcal{B}_0} Q_{n,3}(\gamma|\theta) \quad c.s \quad (2.14)$$

Por otro lado, dado que $\gamma_0 \in \mathcal{B}_0$,

$$\inf_{\gamma \in \mathcal{B}_0} Q_{n,3}(\gamma|\boldsymbol{\theta}) \leq Q_{n,3}(\gamma_0|\boldsymbol{\theta}) \xrightarrow{c.s.} Q_3(\gamma_0|\boldsymbol{\theta})$$

nuevamente por el Teorema A.3.1.

Así,

$$\inf_{\gamma \in \mathcal{B}_0} Q_{n,3}(\gamma|\boldsymbol{\theta}) \leq Q_3(\gamma_0|\boldsymbol{\theta}) \quad c.s \quad (2.15)$$

Las desigualdades (2.14) y (2.15) garantizan (2.13). Por lo tanto,

$$\hat{\gamma}_n = \arg \min_{\gamma} Q_{n,3}(\gamma|\boldsymbol{\theta})$$

debe estar en \mathcal{B}_0 entorno de γ_0 , lo que concluye la demostración. □

2.5. Distribución Asintótica

Esta sección se dedica a establecer la distribución asintótica de los estimadores definidos en 2.2.1. A lo largo del análisis se asumirá que \mathcal{T} es un conjunto compacto y se denotará

$$\chi(y, u) = \frac{\partial}{\partial u} \Psi(y, u), \quad y \quad \chi_1(y, a) = \frac{\partial^2}{\partial u^2} \Psi(y, u),$$

y se define $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ como

$$\begin{aligned} \boldsymbol{\Sigma} = E_0 \left((1 - E_0(W|Y))^2 \Psi^2(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T)) \right. \\ \left. [\omega_1(\mathbf{X})\mathbf{X} + D(\mathbf{X}, T)\boldsymbol{\gamma}(T)f_T(T)] [\omega_1(\mathbf{X})\mathbf{X} + D(\mathbf{X}, T)\boldsymbol{\gamma}(T)f_T(T)]^\top \right), \end{aligned} \quad (2.16)$$

donde $D(\mathbf{X}, t)$ es el primer elemento de $\mathbf{M}(t)^{-1} \omega_1(\mathbf{X})(1, \mathbf{X}^\top)^\top$, con

$$\mathbf{M}(t) = E_0 \left(\chi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T)) \omega_1(\mathbf{X})(1, \mathbf{X}^\top)^\top (1, \mathbf{X}^\top) | T = t \right).$$

Además, sea

$$G_i(t) = G(Y_i, \mathbf{X}_i, t) = (1 - E_0(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_0 + m_0(t)) D(\mathbf{X}_i, t),$$

esto es, el primer elemento de $\mathbf{M}(t)^{-1} (1 - E_0(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_0 + m_0(t)) \omega_1(\mathbf{X}_i)(1, \mathbf{X}_i^\top)^\top$ y sea

$$\boldsymbol{\kappa}(t) = E_0 \left((1 - E_0(W|Y)) \chi(Y, \mathbf{X}^\top \boldsymbol{\beta}_0 + m_0(T)) \omega_1(\mathbf{X}) \mathbf{X} f_T^{-1}(T) | T = t \right).$$

La distribución asintótica de $\hat{\boldsymbol{\beta}}_n$ se obtiene bajo las siguientes hipótesis, que se generalizan de las condiciones requeridas por Boente and Rodriguez (2010) para hallar la distribución asintótica de los estimadores,

- N1. m_0 y \tilde{m}_n son funciones continuamente diferenciables, tales que $\|\tilde{m}'_n - m'_0\|_\infty \xrightarrow{P} 0$.
- N2. Las funciones Ψ , χ , χ_1 , ω_1 , ω_2 y $\psi_1(\mathbf{X}) = \mathbf{X}\omega_1(\mathbf{X})$ son acotadas y continuas. Además, $\omega_1(\cdot)$ y $\omega_2(\cdot)$ son funciones no negativas.
- N3. I. La matriz $\mathbf{A} = E_0\left((1 - E_0(W|Y))\chi(Y, \mathbf{X}^\top\beta_0 + m_0(T))\mathbf{X}\mathbf{X}^\top\omega_1(\mathbf{X})\right)$ es no singular.
 II. La matriz $\mathbf{M}(t)$ definida anteriormente es no singular y $\inf_{t \in \mathcal{T}} \lambda_{p+1}(\mathbf{M}(t)) > 0$.
- N4. Σ es una matriz definida positiva.
- N5. $E_0\left(\Psi(Y, \mathbf{X}^\top\beta_0 + m_0(T)) | (\mathbf{X}, T)\right) = 0$.
- N6. $E_0(\omega_1(\mathbf{X})\|\mathbf{X}\|^2) < \infty$.
- N7. La densidad marginal de T , f_T , es acotada y continuamente diferenciable con derivada acotada y continua. Además, $\inf_{t \in \mathcal{T}} f_T(t) > 0$.
- N8. I. Las funciones $m_1(t_1, t_2) = E_0(G_1(T_2)|T_1 = t_1, T_2 = t_2)$,
 $m'_1(t_1, t_2) = \partial m_1(t_1, t_2)/\partial t_2$ y $m''_1(t_1, t_2) = \partial^2 m_1(t_1, t_2)/\partial t_2^2$ son continuas y acotadas.
 II. Las funciones $m_{is}(t_1, t_2, t_3, t_4) = E_0[(G_i(T_j) - G_i(T_i))(G_s(T_l) - G_s(T_s))|T_i = t_1, T_j = t_2, T_s = t_3, T_l = t_4]$, $m'_{is}(t_1, t_2, t_3, t_4) = \partial m_{is}(t_1, t_2, t_3, t_4)/\partial t_l$ y $m''_{is,lr}(t_1, t_2, t_3, t_4) = \partial^2 m_{is}(t_1, t_2, t_3, t_4)/\partial t_l \partial t_s$ son continuas y acotadas.
 III. $\kappa(t)$ es continuamente diferenciable con derivada continua.
- N9. El núcleo $K : \mathbb{R} \rightarrow \mathbb{R}$ es una función par, no negativa, acotada con variación acotada. Más aún, satisface la condición de Lipchitz de orden uno y $\int K(u)du = 1$, $\int uK(u)du = 0$ y $\int u^2K(u)du < \infty$.

A continuación se enuncian dos lemas que se utilizarán en la demostración de la distribución asintótica de $\hat{\beta}_n$.

Lema 2.5.1. *Asumiendo que las condiciones N1-N3, N7 y N9 se cumplen. Si además $\omega_1(\mathbf{x})\|\mathbf{x}\|^3$ está acotado, $nh_n^4 \rightarrow 0$ y $\lim_{n \rightarrow \infty} nh_n^2/\log^2(1/h_n) = +\infty$, entonces:*

$$(a) \ n^{1/4}\|\tilde{\beta}_n - \beta_0\|_\infty \xrightarrow{P} 0, \quad y \quad n^{1/4}\|\tilde{m}_n - m_0\|_\infty \xrightarrow{P} 0.$$

$$(b) \ \sup_{t \in \mathcal{T}} \left| \tilde{m}_n(t) - m_0(t) - \frac{1}{nh_n f_T(t)} \sum_{j=1}^n K\left(\frac{t-T_j}{h_n}\right) G_j(t) \right| = o_p(n^{-1/2}).$$

Para demostrar este lema se puede replicar la demostración del Lema 4.1 de Boente and Rodriguez (2010), utilizando un razonamiento similar al realizado para el análisis del término $\mathbf{B}_n^{(6)}$ en la demostración del Teorema 2.5.1 a continuación.

Lema 2.5.2. *Asumiendo que las condiciones N2 y N6 se cumplen y que las variables T_i tienen soporte compacto \mathcal{T} . Sea $\hat{\beta}_n$ un estimador débilmente consistente de β_0 y sea \tilde{m}_n tal que $\|\tilde{m}_n - m_0\|_\infty \xrightarrow{P} 0$, entonces*

$$\hat{\mathbf{A}}(\tilde{m}_n, \hat{\beta}_n) \xrightarrow{P} \mathbf{A},$$

donde

$$\hat{\mathbf{A}}(m, \beta) = \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi(Y_i, \mathbf{X}_i^\top \beta + m(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top$$

y \mathbf{A} está definida en N3.

Demostración. Realizando una expansión de Taylor de primer orden, es fácil ver que $\hat{\mathbf{A}}(\tilde{m}_n, \hat{\beta}_n)$ se puede escribir como

$$\begin{aligned} \hat{\mathbf{A}}(\tilde{m}_n, \hat{\beta}_n) &= \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi(Y_i, \mathbf{X}_i^\top \hat{\beta}_n + m_0(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top \\ &\quad + \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi_1(Y_i, \mathbf{X}_i^\top \hat{\beta}_n + \xi_{ni}) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top (\tilde{m}_n(T_i) - m_0(T_i)) \\ &= \mathbf{A}_n^{(1)} + \mathbf{A}_n^{(2)}, \end{aligned}$$

donde $\xi_{ni} = \delta_i \tilde{m}_n(T_i) + (1 - \delta_i) m_0(T_i)$ con $0 \leq \delta_i \leq 1$, $i = 1, \dots, n$.

Notar que $|1 - E_\theta(W_i|Y_i)| < 1$ y la función χ_1 es acotada por N2. Además, por N6 $E_0(\omega_1(\mathbf{X}_i) \|\mathbf{X}_i \mathbf{X}_i^\top\|) \leq E_0(\omega_1(\mathbf{X}_i) \|\mathbf{X}_i\|^2) < \infty$, lo que permite aplicar la ley de los grandes números (Teorema A.1.1) para concluir que $\frac{1}{n} \sum_{i=1}^n \omega_1(\mathbf{X}_i) \|\mathbf{X}_i \mathbf{X}_i^\top\| = \mathcal{O}_p(1)$. Luego por el hecho de que $\|\tilde{m}_n - m_0\|_\infty \xrightarrow{P} 0$, el Lema A.1.1 asegura $\mathbf{A}_n^{(2)} \xrightarrow{P} 0$.

Por otro lado, N2 implica que $E_0\left((1 - E_\theta(W|Y)) \chi(Y, \mathbf{X}^\top \beta + m_0(T)) \omega_1(\mathbf{X}) \mathbf{X} \mathbf{X}^\top\right)$ es continua como función de β . Dado que $\hat{\beta}_n \xrightarrow{P} \beta_0$ se tiene que

$$E_0 \left[(1 - E_\theta(W|Y)) \chi \left(Y, \mathbf{X}^\top \hat{\beta}_n + m_0(T) \right) \omega_1(\mathbf{X}) \mathbf{X} \mathbf{X}^\top \right] - \mathbf{A} \rightarrow 0.$$

Por lo tanto, para ver que $\mathbf{A}_n^{(1)} \xrightarrow{P} \mathbf{A}$ es suficiente demostrar que

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi \left(Y_i, \mathbf{X}_i^\top \hat{\beta}_n + m_0(T_i) \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top \\ &\quad - E_0 \left[(1 - E_\theta(W|Y)) \chi \left(Y, \mathbf{X}^\top \hat{\beta}_n + m_0(T) \right) \omega_1(\mathbf{X}) \mathbf{X} \mathbf{X}^\top \right] \xrightarrow{P} 0. \end{aligned} \tag{2.17}$$

El Teorema A.3.1 aplicado a la familia de funciones

$$\mathcal{G} = \{g_\beta(Y, \mathbf{X}, T|\theta) = (1 - E_\theta(W|Y)) \chi(Y, \mathbf{X}^\top \beta + m_0(T)) \omega_1(\mathbf{X}) \mathbf{X} \mathbf{X}^\top, \beta \in \mathbb{R}^p\},$$

implica que vale (2.17). □

Teorema 2.5.1. *Asumiendo que T_i tiene soporte compacto \mathcal{T} y que se cumplen N1-N9. Sea $\boldsymbol{\theta} = (\hat{\boldsymbol{\beta}}_n, \hat{\gamma}_n, \tilde{m}_n)$ un estimador fuertemente consistente de $\boldsymbol{\theta}_0$. Si se verifica $\sup_{\mathbf{X}, Z, T} |E_0(W|Y) - E_{\boldsymbol{\theta}}(W|Y)| = o_p(n^{-1/2})$, $nh_n^4 \rightarrow 0$ y la conclusión del Lema 2.5.1 se cumple, entonces*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-1}),$$

donde \mathbf{A} y $\boldsymbol{\Sigma}$ están definidas en N3 y (2.16), respectivamente.

Demostración. $\hat{\boldsymbol{\beta}}_n$ y $\boldsymbol{\theta}$ son tales que se verifica

$$\frac{1}{n} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_n + \tilde{m}_n(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i = 0.$$

Haciendo una expansión de Taylor de primer orden alrededor de $\boldsymbol{\beta}_0$, se obtiene

$$\begin{aligned} \mathbf{0} &= \frac{1}{n} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_0 + \tilde{m}_n(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i \\ &\quad + \frac{1}{n} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i)) \chi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_n^* + \tilde{m}_n(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \end{aligned} \quad (2.18)$$

donde $\boldsymbol{\beta}_n^* = \boldsymbol{\beta}_0 + \hat{\boldsymbol{\beta}}_n(1 - \delta)$ con $\delta \in (0, 1)$.

De (2.18) se tiene entonces que $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = -\mathbf{A}_n^{-1} \mathbf{B}_n$ con

$$\mathbf{A}_n = \frac{1}{n} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i)) \chi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_n^* + \tilde{m}_n(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top$$

y

$$\mathbf{B}_n = \frac{\sqrt{n}}{n} \sum_{i=1}^n (1 - E_{\boldsymbol{\theta}}(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_0 + \tilde{m}_n(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i.$$

Por lo tanto, es suficiente demostrar que $\mathbf{A}_n \xrightarrow{P} \mathbf{A}$ (donde \mathbf{A} es la matriz constante definida en N3) y que $\mathbf{B}_n \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Una vez garantizadas esas convergencias, por el teorema de Slutsky (Teorema A.1.2) se tiene que $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-1})$.

La convergencia $\mathbf{A}_n \xrightarrow{P} \mathbf{A}$ se da por la conclusión de Lema 2.5.2, ya que $\mathbf{A}_n = \hat{\mathbf{A}}(\tilde{m}, \boldsymbol{\beta}_n^*)$, $\hat{\boldsymbol{\beta}}_n$ es un estimador consistente de $\boldsymbol{\beta}_0$ y $\boldsymbol{\beta}_n^*$ es un punto intermedio entre $\boldsymbol{\beta}_0$ y $\hat{\boldsymbol{\beta}}_n$.

Para probar la convergencia en distribución $\mathbf{B}_n \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, se considera

$$\begin{aligned} \mathbf{C}_n &= \frac{\sqrt{n}}{n} \sum_{i=1}^n \left\{ (1 - E_0(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_0 + m_0(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i + \boldsymbol{\kappa}(T_i) G_i(T_i) f_T(T_i) \right\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - E_0(W_i|Y_i)) \Psi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta}_0 + m_0(T_i)) [\omega_1(\mathbf{X}_i) \mathbf{X}_i + \boldsymbol{\kappa}(T_i) D(X_i, T_i) f_T(T_i)]. \end{aligned}$$

Por la condición N5 se tiene que \mathbf{C}_n tiene esperanza cero y matriz de covarianza $\boldsymbol{\Sigma}$. Por el Teorema Central del Límite A.1.6 se puede concluir que \mathbf{C}_n tiene distribución asintótica

$\mathcal{N}_p(\mathbf{0}, \Sigma)$. Por lo tanto bastaría con ver que \mathbf{B}_n y \mathbf{C}_n son asintóticamente equivalentes, esto es $\mathbf{B}_n - \mathbf{C}_n \xrightarrow{P} 0$.

Es fácil ver que $\mathbf{B}_n - \mathbf{C}_n = \mathbf{B}_n^{(1)} + \mathbf{B}_n^{(2)} - \mathbf{B}_n^{(3)} + \mathbf{B}_n^{(4)}$, donde

$$\begin{aligned}\mathbf{B}_n^{(1)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (E_0(W_i|Y_i) - E_\theta(W_i|Y_i)) \Psi \left(Y_i, \mathbf{X}_i^\top \beta_0 + m_0(T_i) \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i, \\ \mathbf{B}_n^{(2)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi \left(Y_i, \mathbf{X}_i^\top \beta_0 + m_0(T_i) \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i [\tilde{m}_n(T_i) - m_0(T_i)], \\ \mathbf{B}_n^{(3)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \kappa(T_i) G_i(T_i) f_T(T_i), \\ \mathbf{B}_n^{(4)} &= \frac{1}{2n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi_1 \left(Y_i, \mathbf{X}_i^\top \beta_n^* + \xi_{ni} \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i \left(n^{1/4} [\tilde{m}_n(T_i) - m_0(T_i)] \right)^2.\end{aligned}$$

Aquí, $\xi_{ni} = \delta_i \tilde{m}_n(T_i) + (1 - \delta_i) m_0(T_i)$ con $\delta_i \in (0, 1)$ es un punto intermedio entre $\tilde{m}_n(T_i)$ y $m_0(T_i)$ para cada $i = 1, \dots, n$.

Bajo la hipótesis $\sup_{\mathbf{X}, \mathbf{Z}, T} |E_0(W|Y) - E_\theta(W|Y)| = o_p(n^{-1/2})$, por la Ley de los Grandes Números es posible concluir que $\mathbf{B}_n^{(1)} \xrightarrow{P} 0$.

Por N2 y la conclusión del Lema 2.5.1 (a) se tiene que $\mathbf{B}_n^{(4)} \xrightarrow{P} 0$. Así, para concluir la prueba bastará con demostrar que

$$\mathbf{B}_n^{(2)} - \mathbf{B}_n^{(3)} \xrightarrow{P} 0.$$

El Lema 2.5.1 (b) implica que

$$\begin{aligned}\mathbf{B}_n^{(2)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi \left(Y_i, \mathbf{X}_i^\top \beta_0 + m_0(T_i) \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i \cdot \\ &\quad \left\{ \frac{1}{nh_n f_T(T_i)} \sum_{j=1}^n G_j(T_i) K \left(\frac{T_i - T_j}{h_n} \right) + o_p(n^{-1/2}) \right\} \\ &= \frac{1}{h_n n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n (1 - E_\theta(W_i|Y_i)) \chi \left(Y_i, \mathbf{X}_i^\top \beta_0 + m_0(T_i) \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i \cdot \\ &\quad f_T^{-1}(T_i) G_j(T_i) K \left(\frac{T_i - T_j}{h_n} \right) + o_p(1) \\ &= \mathbf{B}_n^{(5)} + o_p(1).\end{aligned}$$

Por lo tanto, es suficiente probar que $\mathbf{B}_n^{(5)} - \mathbf{B}_n^{(3)} \xrightarrow{P} 0$.

Sea

$$R_\theta(Y_i, \mathbf{X}_i, T_i) = (1 - E_\theta(W_i|Y_i)) \chi \left(Y_i, \mathbf{X}_i^\top \beta_0 + m_0(T_i) \right) \omega_1(\mathbf{X}_i) \mathbf{X}_i f_T^{-1}(T_i),$$

que se denota $R_0(Y_i, \mathbf{X}_i, T_i)$ cuando $\theta = \theta_0$. Es fácil ver que

$$\mathbf{B}_n^{(5)} - \mathbf{B}_n^{(3)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{1}{nh_n} \sum_{j=1}^n R_\theta(Y_j, \mathbf{X}_j, T_j) G_i(T_j) K \left(\frac{T_i - T_j}{h_n} \right) - \kappa(T_i) G_i(T_i) f_T(T_i) \right\}$$

se puede escribir como

$$\mathbf{B}_n^{(5)} - \mathbf{B}_n^{(3)} = \mathbf{B}_n^{(6)} + \mathbf{B}_n^{(7)} + \mathbf{B}_n^{(8)} + \mathbf{B}_n^{(9)} + \mathbf{B}_n^{(10)}$$

donde

$$\begin{aligned} \mathbf{B}_n^{(6)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) [R_{\theta}(Y_j, \mathbf{X}_j, T_j) - R_0(Y_j, \mathbf{X}_j, T_j)] G_i(T_j) \\ \mathbf{B}_n^{(7)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) [R_0(Y_j, \mathbf{X}_j, T_j) - \kappa(T_j)] G_i(T_j) \\ \mathbf{B}_n^{(8)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) [\kappa(T_j) - \kappa(T_i)] G_i(T_j) \\ \mathbf{B}_n^{(9)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) [G_i(T_j) - G_i(T_i)] \kappa(T_i) \\ \mathbf{B}_n^{(10)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) - f_T(T_i) \right] \kappa(T_i) G_i(T_i). \end{aligned}$$

Resta ver que cada término converge a 0 en probabilidad.

Para comprobar la convergencia de $\mathbf{B}_n^{(10)}$ notar que

$$\begin{aligned} \mathbf{B}_n^{(10)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) - f_T(T_i) \right] \kappa(T_i) G_i(T_i) \\ &\leq \sup_{t \in \mathcal{T}} \left| \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{t - T_j}{h_n} \right) - f_T(t) \right| \frac{1}{\sqrt{n}} \sum_{i=1}^n \kappa(T_i) G_i(T_i). \end{aligned}$$

Notar que $\frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{t - T_j}{h_n} \right)$ es el estimador basado en núcleos de la densidad f_T y por el Lema A.4.1 se puede concluir que

$$\sup_{t \in \mathcal{T}} \left| \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{t - T_j}{h_n} \right) - f_T(t) \right| \xrightarrow{P} 0.$$

Además, la condición N5 implica que

$$\begin{aligned} E_0(\kappa(T_i) G_i(T_i)) &= E_0[\kappa(T_i) E_0(G_i(T_i) | (\mathbf{X}_i, T_i))] \\ &= E_0[\kappa(T_i) D(\mathbf{X}_i, T_i) (1 - \pi_i) E_0(\Psi(Y_i, \mathbf{X}_i^\top \beta_0 + m_0(T_i)) | (\mathbf{X}_i, T_i))] = 0, \end{aligned}$$

y de las condiciones N2, N3II, N6 y N7 se sigue que $\kappa(T_i) G_i(T_i)$ tiene covarianza finita. Luego, se puede aplicar el Teorema Central del Límite para concluir que $\frac{1}{\sqrt{n}} \sum_{i=1}^n \kappa(T_i) G_i(T_i)$ converge en distribución a una normal con media 0 y matriz de covarianza finita. Consecuentemente por Lema A.1.2

$$\mathbf{B}_n^{(10)} = \mathcal{O}_p(1) o_p(1) = o_p(1).$$

Para analizar la convergencia del término $\mathbf{B}_n^{(9)}$, por Lema A.1.3, es suficiente probar que $\|E_0(\mathbf{B}_n^{(9)})\|^2$ y $\text{tr}(\text{Cov}(\mathbf{B}_n^{(9)}, \mathbf{B}_n^{(9)}))$ convergen a 0 cuando $n \rightarrow \infty$. En ese sentido, notar que por linealidad de la esperanza y por definición de m_1 en N8 se tiene

$$E_0(G_1(T_2) - G_1(T_1)|(T_1, T_2)) = m_1(T_1, T_2) - E_0(G_1(T_1)|(T_1, T_2))$$

pero T_2 es independiente de (Y_1, \mathbf{X}_1, T_1) , entonces $E_0(G_1(T_1)|(T_1, T_2)) = E_0(G_1(T_1)|T_1) = 0$ por N5. Así

$$m_1(T_1, T_2) = E_0(G_1(T_2) - G_1(T_1)|(T_1, T_2))$$

y consecuentemente $m_1(T_1, T_1) = 0$. Por lo tanto

$$\begin{aligned} E_0(\mathbf{B}_n^{(9)}) &= \frac{n(n-1)}{n\sqrt{n}} E_0\left(\frac{1}{h_n} K\left(\frac{T_1 - T_2}{h_n}\right) E_0[G_1(T_2) - G_1(T_1)|(T_1, T_2)] \boldsymbol{\kappa}(T_1)\right) \\ &= \frac{n-1}{\sqrt{n}} E_0\left(\frac{1}{h_n} K\left(\frac{T_1 - T_2}{h_n}\right) m(T_1, T_2) \boldsymbol{\kappa}(T_1)\right) \\ &= \frac{n-1}{\sqrt{n}} \int \boldsymbol{\kappa}(u) f_T(u) \left(\int m_1(u, u - h_n v) K(v) f_T(u - h_n v) (-dv)\right) du, \end{aligned}$$

donde en la última igualdad se realizó el cambio de variables $u = T_1$ y $v = \frac{T_1 - T_2}{h_n}$.

Al expandir f_T y m_1 usando polinomios de Taylor de primer y segundo orden, respectivamente, se obtiene

$$\begin{aligned} m_1(u, u - h_n v) &= m_1(u, u) - m_1'(u, u) h_n v + m_1''(u, \xi_1) v^2 h_n^2 \\ f_T(u - h_n v) &= f_T(u) - f_T'(\xi_2) h_n v. \end{aligned}$$

Reemplazando las expresiones anteriores en la integral respecto de v , teniendo en cuenta que $m_1(u, u) = 0$ se tiene

$$\begin{aligned} &\int m_1(u, u - h_n v) K(v) f_T(u - h_n v) (-dv) = \\ &\quad - \int [m_1'(u, u) h_n v - m_1''(u, \xi_1) h_n^2 v^2] K(v) [f_T(u) - f_T'(\xi_2) h_n v] dv \\ &= h_n^2 \left[m_1'(u, u) \int v^2 K(v) f_T'(\xi_2) dv + f_T(u) \int v^2 K(v) m_1''(u, \xi_1) dv \right. \\ &\quad \left. - h \int v^3 K(v) m_1''(u, \xi_1) f_T'(\xi_2) dv \right], \end{aligned}$$

ya que N7, N8 y N9 se cumplen y \mathcal{T} es un conjunto compacto. Esto implica que

$$\|E_0(\mathbf{B}_n^{(9)})\|^2 = \mathcal{O}(nh^4),$$

que converge a 0 por hipótesis.

Para el cálculo de la covarianza de cada componente de $\mathbf{B}_n^{(9)}$, notar que

$$\mathbf{B}_n^{(9)} = \frac{1}{n^{3/2}} \sum_{i \neq j} \frac{1}{h_n} K\left(\frac{T_i - T_j}{h_n}\right) [G_i(T_j) - G_i(T_i)] \boldsymbol{\kappa}(T_i) = \frac{1}{n^{3/2}} \sum_{i \neq j} \mathbf{V}_{ij}.$$

Entonces,

$$\text{Cov}(\mathbf{B}_n^{(9)}, \mathbf{B}_n^{(9)}) = \frac{1}{n^3} \sum_{i \neq j} \sum_{l \neq s} \text{Cov}(\mathbf{V}_{ij}, \mathbf{V}_{ls}),$$

y por lo tanto

$$\begin{aligned} \text{Cov}(\mathbf{B}_n^{(9)}, \mathbf{B}_n^{(9)}) &= \frac{n(n-1)}{n^3} \text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12}) \\ &+ \frac{n(n-1)(n-2)}{n^3} [\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13}) + \text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{32})]. \end{aligned}$$

En lo que sigue, se analiza la traza de cada uno de los términos involucrados.

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) &= \text{tr}(E_0(\mathbf{V}_{12}\mathbf{V}_{12}^\top) - E_0(\mathbf{V}_{12})E_0(\mathbf{V}_{12}^\top)) \tag{2.19} \\ &\leq \text{tr}(E_0(\mathbf{V}_{12}\mathbf{V}_{12}^\top)) \\ &= E_0 \left[\frac{1}{h_n^2} K^2 \left(\frac{T_1 - T_2}{h_n} \right) (G_1(T_2) - G_1(T_1))^2 \text{tr}(\boldsymbol{\kappa}(T_1)\boldsymbol{\kappa}(T_1)^\top) \right] \\ &= E_0 \left[\frac{1}{h_n^2} K^2 \left(\frac{T_1 - T_2}{h_n} \right) E_0 \left[(G_1(T_2) - G_1(T_1))^2 | (T_1, T_2) \right] \cdot \right. \\ &\quad \left. \text{tr}(\boldsymbol{\kappa}(T_1)\boldsymbol{\kappa}(T_1)^\top) \right] \\ &= E_0 \left[\frac{1}{h_n^2} K^2 \left(\frac{T_1 - T_2}{h_n} \right) m_{11}(T_1, T_2, T_1, T_2) \text{tr}(\boldsymbol{\kappa}(T_1)\boldsymbol{\kappa}(T_1)^\top) \right] \\ &= -\frac{1}{h_n} \int K^2(v) m_{11}(u, u - h_n v, u, u - h_n v) \text{tr}(\boldsymbol{\kappa}(u)\boldsymbol{\kappa}(u)^\top) \cdot \\ &\quad f_T(u) f_T(u - h_n v) dudv \end{aligned}$$

donde en la última expresión se realizó el cambio de variables $u = T_1$ y $v = \frac{T_1 - T_2}{h_n}$.

Mediante un desarrollo de Taylor de segundo orden se obtiene

$$\begin{aligned} m_{11}(u, u - h_n v, u, u - h_n v) &= m_{11}(u, u - h_n v, u, u) \tag{2.20} \\ &- m'_{11}(u, u - h_n v, u, u) h_n v \\ &+ m''_{11,44}(u, u - h_n v, u, \xi) \frac{h_n^2 v^2}{2} \end{aligned}$$

donde $\xi \in (u - h_n v, u)$ es un punto intermedio. Luego reemplazando en (2.19) y teniendo en cuenta que por definición $m_{11}(u, u - h_n v, u, u) = 0$,

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) &= \tag{2.21} \\ &\int K^2(v) m'_{11}(u, u - h_n v, u, u) v \text{tr}(\boldsymbol{\kappa}(u)\boldsymbol{\kappa}(u)^\top) f_T(u) f_T(u - h_n v) dudv \\ &- \int K^2(v) m''_{11,44}(u, u - h_n v, u, \xi) \frac{h_n^2 v^2}{2} \text{tr}(\boldsymbol{\kappa}(u)\boldsymbol{\kappa}(u)^\top) f_T(u) f_T(u - h_n v) dudv. \end{aligned}$$

Al realizar un desarrollo de Taylor de $m'_{11}(u, u - h_n v, u, u) f_T(u - h_n v)$ en $h_n v$ alrededor de 0, se tiene

$$m'_{11}(u, u - h_n v, u, u) f_T(u - h_n v) = m'_{11}(u, u, u, u) f_T(u) + O(h_n v).$$

Sustituyendo en la primer integral de (2.21),

$$\begin{aligned} \int K^2(v)m'_{11}(u, u - h_nv, u, u)v f_T(u - h_nv)dv &= m'_{11}(u, u, u, u) f_T(u) \int K^2(v)v dv \\ &+ O(h_n) \int K^2(v)v^2 dv. \end{aligned}$$

El primer término se anula porque $\int K^2(v)v dv = 0$ al ser una función impar. El segundo término es de orden $O(h_n)$, ya que $\int v^2 K^2(v)dv < \infty$. La segunda integral en (2.21) también es $O(h_n)$.

Por lo tanto,

$$\frac{n(n-1)}{n^3} \text{tr}(\text{Cov}(V_{12}, V_{12})) = O(h_n n^{-1}).$$

Por otro lado,

$$\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13}) = E_0(\mathbf{V}_{12}\mathbf{V}_{13}^\top) - E_0(\mathbf{V}_{12})E_0(\mathbf{V}_{13})^\top. \quad (2.22)$$

Es fácil ver que $E(\mathbf{V}_{12}) = O(h_n)$ y $E(\mathbf{V}_{13}) = O(h_n)$ usando los mismos argumentos que para el cálculo de la esperanza de $\mathbf{B}_n^{(9)}$.

Para calcular $\text{tr}(E_0(\mathbf{V}_{12}\mathbf{V}_{13}^\top))$ se realiza el cambio de variable $z = \frac{T_1 - T_3}{h_n}$, de manera que

$$\begin{aligned} \text{tr}(E_0(\mathbf{V}_{12}\mathbf{V}_{13}^\top)) &= \int K(v)K(z) \text{tr}(\boldsymbol{\kappa}(u)\boldsymbol{\kappa}(u)^\top) m_{11}(u, u - h_nv, u, u - h_nz) \\ &f_T(u) f_T(u - h_nv) f_T(u - h_nz) dudz dv \end{aligned}$$

y con un desarrollo de Taylor análogo al realizado en (2.20) se tiene que

$$\text{tr}(E_0(\mathbf{V}_{12}\mathbf{V}_{13}^\top)) = O(h_n^2).$$

De forma análoga se puede probar que

$$\text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{32})) = O(h_n^2),$$

lo cual implica que

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{B}_n^{(9)}, \mathbf{B}_n^{(9)})) &= \frac{n-1}{n^2} \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) \\ &+ \frac{n-1}{n} [\text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13})) + \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{32}))] \\ &= \frac{n-1}{n^2} O(h_n) + \frac{n-1}{n} O(h_n^2) \\ &= O(h_n n^{-1} + h_n^2). \end{aligned}$$

La convergencia de $\mathbf{B}_n^{(7)}$ y $\mathbf{B}_n^{(8)}$ a cero se puede probar mediante argumentos similares, calculando las medias y la traza de las matrices de covarianza. Las respectivas demostraciones se encuentran en la Sección A.4 del Apéndice.

Finalmente, notar que $\mathbf{B}_n^{(6)}$ se puede escribir como

$$\begin{aligned}\mathbf{B}_n^{(6)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K\left(\frac{T_i - T_j}{h_n}\right) [R_\theta(Y_j, \mathbf{X}_j, T_j) - R_0(Y_j, \mathbf{X}_j, T_j)] G_i(T_j) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K\left(\frac{T_i - T_j}{h_n}\right) [E_0(W_j|Y_j) - E_\theta(W_j|Y_j)] \cdot \\ &\quad \chi\left(Y_j, \mathbf{X}_j^\top \beta_0 + m_0(T_j)\right) \omega_1(\mathbf{X}_j) \mathbf{X}_j f_T^{-1}(T_j) G_i(T_j)\end{aligned}$$

Por hipótesis, la diferencia de las esperanzas condicionales converge uniformemente a cero en probabilidad. Además, como $\frac{1}{h_n n} \sum_{j=1}^n K\left(\frac{t-T_j}{h_n}\right)$ es el estimador basado en núcleos de la densidad f_T , por el Lema A.4.1 se cumple la convergencia uniforme. N5 permite mostrar que $G_i(t)$ tiene media cero, además es acotado, pues por N3 $\mathbf{M}(t)$ es invertible y $\inf_{t \in \mathcal{T}} \lambda_{p+1}(\mathbf{M}(t)) > 0$, entonces $\|\mathbf{M}(t)^{-1}\| \leq \sqrt{p+1}/\lambda_{p+1}(\mathbf{M}(t))$ con $\lambda_{p+1}(\mathbf{M}(t)) < \infty$ por la continuidad de los autovalores y la compacidad de \mathcal{T} . Luego se tiene que $\mathbf{B}_n^{(6)} \xrightarrow{P} 0$. \square

2.6. Test de hipótesis

En numerosos contextos aplicados resulta de interés determinar si las covariables \mathbf{X} influyen de manera significativa sobre la respuesta Y . Este objetivo exige evaluar estadísticamente el parámetro de regresión β , lo que puede abordarse mediante el siguiente contraste de hipótesis

$$\begin{aligned}H_0 &: \beta = \beta_0 \\ H_1 &: \beta \neq \beta_0,\end{aligned}$$

el cual permite determinar si las predictoras aportan información relevante sobre la respuesta o si, por el contrario, su efecto es estadísticamente indistinguible de cero.

Para este propósito resulta natural utilizar un test de Wald, cuya construcción se apoya en la distribución asintótica del estimador de β . Específicamente, el estadístico de Wald se define como

$$Q(\hat{\beta}_n, \hat{\Sigma}_{\hat{\beta}_n}, H_0) = (\hat{\beta}_n - \beta_0)^\top \hat{\Sigma}_{\hat{\beta}_n}^{-1} (\hat{\beta}_n - \beta_0), \quad (2.23)$$

donde $\hat{\Sigma}_{\hat{\beta}_n}$ es una estimación de $\Sigma_{\beta_0} \doteq \mathbf{A}^{-1} \Sigma \mathbf{A}^{-1}$, que se define a continuación.

Sea

$$\hat{\mathbf{A}}(m, \beta) = \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i)) \chi(Y_i, \mathbf{X}_i^\top \beta + m(T_i)) \omega_1(\mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^\top$$

definida en el Lema 2.5.2 y la matriz $\hat{\Sigma}(m, \beta)$ definida como sigue

$$\hat{\Sigma}(m, \beta) = \frac{1}{n} \sum_{i=1}^n (1 - E_\theta(W_i|Y_i))^2 \Psi^2\left(Y_i, \mathbf{X}_i^\top \beta + m(T_i)\right) \hat{\mathbf{u}}_i(m, \beta) \hat{\mathbf{u}}_i(m, \beta)^\top,$$

donde

$$\begin{aligned}\widehat{\mathbf{u}}_i(m, \boldsymbol{\beta}) &= \omega_1(\mathbf{X}_i)\mathbf{X}_i + \widehat{D}_i(T_i, m, \boldsymbol{\beta})\widehat{\boldsymbol{\kappa}}(T_i, m, \boldsymbol{\beta})\widehat{f}_T(T_i), \\ \widehat{\boldsymbol{\kappa}}(t, m, \boldsymbol{\beta}) &= \sum_{i=1}^n W_{i,n}(t)\chi\left(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i)\right)\omega_1(\mathbf{X}_i)\mathbf{X}_i\widehat{f}_T^{-1}(T_i),\end{aligned}$$

con $\widehat{f}_T(t)$ un estimador núcleo de la densidad $f_T(t)$ y $\widehat{D}_i(t, \boldsymbol{\beta}, m)$ el primer elemento de $\widehat{\mathbf{M}}(t, m, \boldsymbol{\beta})^{-1}\omega_1(\mathbf{X}_i)(1, \mathbf{X}_i^\top)^\top$ con

$$\widehat{\mathbf{M}}(t, m, \boldsymbol{\beta}) = \sum_{i=1}^n W_{i,n}(t)\chi(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i))\omega_1(\mathbf{X}_i)(1, \mathbf{X}_i^\top)^\top(1, \mathbf{X}_i^\top).$$

Se consideran $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n)$ y $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n)$ como estimadores de $\boldsymbol{\Sigma}$ y \mathbf{A} , respectivamente, de manera que

$$\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n} = \widehat{\mathbf{A}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{A}}^{-1}.$$

Teorema 2.6.1. *Sea $\{(Y_i, \mathbf{X}_i^\top, T_i)^\top\}_{i=1}^n$ una muestra aleatoria que satisface el modelo PLZIP. Se supone que T_i tiene soporte compacto \mathcal{T} y que las condiciones N1-N9 se cumplen. Sea $\widehat{\boldsymbol{\beta}}_n$ una solución de (2.7) que proporciona un estimador consistente de $\boldsymbol{\beta}_0$. Si $nh_n^4 \rightarrow 0$ y se cumple el resultado del Lema 2.5.2, entonces*

- (a) Bajo $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, se tiene que $nQ(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}, H_0) \xrightarrow{\mathcal{D}} \chi_p^2$.
- (b) Bajo $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $nQ(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}, H_0) \rightarrow \infty$, para cualquier $\boldsymbol{\beta}$ fija.

Demostración. Para demostrar (a), notar que argumentos análogos a los considerados en el Lema 2.5.2 permiten concluir que bajo H_0 , $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}$, la demostración se encuentra en la Sección A.4 del Apéndice. Luego, por el Teorema 2.5.1 y el Teorema de Slutsky A.1.2

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n} \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} W \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_0} \end{pmatrix}$$

con $W \sim \mathcal{N}_p(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-1})$, donde \mathbf{A} y $\boldsymbol{\Sigma}$ están definidas en N3I y (2.16), respectivamente. Considerando $f(w, S) = w^\top S^{-1}w$, se tiene que

$$nQ(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}, H_0) = f\left(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}\right) \xrightarrow{\mathcal{D}} f(W, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}) = W^\top \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}W,$$

pero $W^\top \mathbf{A}\boldsymbol{\Sigma}^{-1}\mathbf{A}W = Z^\top Z$, con $Z \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ y, por el Teorema A.1.7, $nQ(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}, H_0)$ converge en distribución a una variable aleatoria con distribución χ_p^2 .

Para demostrar (b), notar que si el verdadero valor del parámetro es $\boldsymbol{\beta}_1$ se tiene

$$nQ(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}, H_0) = n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_1)^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}^{-1}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_1) + n(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}^{-1}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1).$$

Cuando $n \rightarrow \infty$, por un razonamiento análogo al realizado en (a), el primer término converge en distribución a una variable aleatoria con distribución χ_p^2 . Por otro lado,

$$(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}^{-1} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1) \xrightarrow{P} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)^\top \boldsymbol{\Sigma} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1),$$

y por lo tanto el segundo término, al estar multiplicado por n , tiende a infinito. \square

Dado que el estadístico de Wald definido en (2.23) es no negativo y, de acuerdo con el Teorema 2.6.1, bajo H_0 converge en distribución a una χ^2 con p grados de libertad, la prueba rechaza H_0 a favor de H_1 con un nivel de significancia aproximado α cuando el estadístico observado verifica

$$nQ(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}, H_0) > \chi_{p,\alpha}^2,$$

donde $\chi_{p,\alpha}^2$ denota el percentil $(1 - \alpha)$ de una distribución χ^2 con p grados de libertad.

A partir de la distribución asintótica del estimador, también es posible abordar test de hipótesis más generales. En efecto, si se desea testear hipótesis simultáneas de la forma

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

donde \mathbf{R} es una matriz de rango completo de dimensión $q \times p$ y $\mathbf{r} \in \mathbb{R}^q$, la teoría asintótica garantiza que

$$\sqrt{n}(\mathbf{R}\widehat{\boldsymbol{\beta}}_n - \mathbf{r}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{R}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}\mathbf{R}^\top).$$

En consecuencia, bajo H_0 el estadístico de Wald generalizado

$$n(\mathbf{R}\widehat{\boldsymbol{\beta}}_n - \mathbf{r})^\top [\mathbf{R}\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}\mathbf{R}^\top]^{-1} (\mathbf{R}\widehat{\boldsymbol{\beta}}_n - \mathbf{r})$$

converge en distribución a una χ_q^2 . Esta propiedad permite evaluar la significancia estadística de combinaciones arbitrarias de los coeficientes del modelo.

Estudios de Simulación y Ejemplos

3.1. Estudios de Simulación

En esta sección se realiza un estudio de simulación para muestras finitas con el fin de evaluar la robustez y el desempeño de los estimadores propuestos en esta tesis, introducidos en la Definición 2.2.1. Los datos se simulan bajo cuatro escenarios diferentes. Para cada muestra simulada, se estima el modelo PLZIP utilizando el procedimiento desarrollado en esta tesis y presentado en la Sección 2.3, que consiste en la implementación del algoritmo tipo EM, con diferentes funciones de pérdida robustas: la que da lugar a los MT estimadores, introducida por Valdora and Yohai (2014) y aquella introducida por Croux and Haesbroeck (2002), ambas se definen en la siguiente sección. Los estimadores obtenidos con las mencionadas funciones de pérdida se denotan MT y CH, respectivamente. También se calculan estimadores del modelo PLZIP mediante el procedimiento presentado en la Sección 2.1, que consiste en el algoritmo tipo EM utilizando la función de log-verosimilitud estándar. Los mismos se denotan ML.

Con el propósito de comparar estos métodos con alternativas existentes, se incluyen también estimadores propuestos por otros autores. En particular, se consideran los estimadores del modelo PLZIP de índice único introducidos por Wang et al. (2015), detallados en la Sección 1.4.1.1. Los mismos se denotan BS-ML, donde BS indica el uso de funciones de base B-spline. Para su implementación computacional se utiliza la función `bs` del paquete `splines`.

Adicionalmente, se consideran estimadores del modelo de regresión lineal ZIP, donde la covariable T se introduce como una predictora lineal. En este contexto, se incluye el estimador clásico de máxima verosimilitud, introducido en la Sección 1.3.3, así como el estimador RES de Hall and Shen (2010), descrito en la Sección 1.3.5. Se denota a dichos estimadores ZIP-ML y RES, respectivamente. Para computar el estimador ZIP-ML se utiliza la función `zeroinfl` del paquete `pscl`, mientras que el estimador RES ha sido

implementado, usando la función `multiroot` del paquete `rootSolve` para resolver las ecuaciones de estimación. Debido a que los datos se simulan bajo un modelo PLZIP, el desempeño de estos estimadores puede verse afectado, ya que estiman un modelo mal especificado.

En resumen, el presente estudio de simulación incluye los estimadores MT, CH y ML desarrollados en esta tesis, así como otros estimadores propuestos en la literatura utilizados con fines comparativos. Entre ellos, los estimadores BS-ML del modelo PLZIP de índice único y los estimadores ZIP-ML y RES correspondientes al modelo de regresión lineal ZIP.

3.1.1. Funciones de Pérdida Robusta

Las funciones de pérdida empleadas en este estudio tienen la forma general dada en (2.10), donde el término G , que asegura la consistencia de Fisher, se define en (2.11). Por lo tanto, la única elección restante es la especificación de la función ϕ .

Se realizan dos elecciones diferentes de ϕ , lo que resulta en dos estimadores robustos para cada parámetro lineal y para la función no paramétrica del modelo PLZIP.

Una de las funciones de pérdida robusta ϕ consideradas es la que se utiliza en los MT-estimadores propuestos por Valdora and Yohai (2014) y resulta

$$\rho(y, u) = \phi_{MT}(\sqrt{y} - f(e^u)),$$

con

$$\phi_{MT}(s) = \begin{cases} 1 - \left(1 - \left(\frac{s}{c}\right)^2\right)^4, & |s| \leq c, \\ 1, & |s| > c, \end{cases}$$

y

$$f(\lambda) = \arg \min_u E_\lambda(\phi_{MT}(\sqrt{y} - u)).$$

Como indican los autores, la constante se elige $c = 2.9$ para obtener una eficiencia entre 75 % y 90 %. Como se mencionó previamente, los estimadores obtenidos usando esta función de pérdida robusta se denotan MT.

La segunda función de pérdida robusta ϕ es la propuesta por Croux and Haesbroeck (2002), que está dada por

$$\phi_{CH}(s) = \begin{cases} se^{-\sqrt{c}}, & |s| \leq c \\ e^{-\sqrt{c}}(2(1 + \sqrt{c}) + c) - 2e^{-\sqrt{s}}(1 + \sqrt{s}), & |s| > c, \end{cases}$$

donde la constante c se elige siguiendo la recomendación de los autores, quienes sugieren seleccionar $c = 0.5$ para balancear el compromiso entre robustez y eficiencia. Los estimadores obtenidos usando esta función de pérdida robusta se denotan CH, en referencia a los autores Croux y Haesbroeck.

Los códigos para calcular los estimadores han sido implementados en el software estadístico R Core Team (2013). Las optimizaciones requeridas para los diferentes estimadores se llevan a cabo mediante las funciones `optim` con `method="Nelder-Mead"` (para la optimización multivariada) y `optimize` (para la optimización unidimensional) del paquete `stats` de R.

Siguiendo a Valdora and Yohai (2014), las funciones de peso ω_1 y ω_2 se definen como una función de la distancia de Mahalanobis. Específicamente,

$$\omega_1(\mathbf{X}) = g\left((\mathbf{X} - \hat{\mu}_{\mathbf{X}})^\top \hat{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{X} - \hat{\mu}_{\mathbf{X}})^{1/2}\right),$$

donde $\hat{\mu}_{\mathbf{X}}$ y $\hat{\Sigma}_{\mathbf{X}}$ corresponden a los estimadores robustos de localización y matriz de covarianza de \mathbf{X} calculados con la función `SestCov`, `method="bisquare"` del paquete `rrcov` de R. Una definición análoga se utiliza para $\omega_2(\mathbf{Z})$, de manera que

$$\omega_2(\mathbf{Z}) = g\left((\mathbf{Z} - \hat{\mu}_{\mathbf{Z}})^\top \hat{\Sigma}_{\mathbf{Z}}^{-1}(\mathbf{Z} - \hat{\mu}_{\mathbf{Z}})^{1/2}\right),$$

donde $\hat{\mu}_{\mathbf{Z}}$ y $\hat{\Sigma}_{\mathbf{Z}}$ corresponden a los estimadores robustos de localización y matriz de covarianza de \mathbf{Z} . La función g se define como

$$g(t) = \begin{cases} 1, & t \leq \chi_{0.035,p}^2 \\ \frac{\chi_{0.025,p}^2 - t}{\chi_{0.025,p}^2 - \chi_{0.035,p}^2}, & \chi_{0.035,p}^2 < t \leq \chi_{0.025,p}^2 \\ 0, & t > \chi_{0.025,p}^2 \end{cases}$$

donde $\chi_{\alpha,p}^2$ denota el percentil $(1 - \alpha)$ de la distribución chi-cuadrado con p grados de libertad, y p representa el número de covariables involucradas.

3.1.2. Escenarios de Simulación

Para evaluar la robustez de los estimadores propuestos, se consideran cuatro escenarios diferentes, inspirados en el estudio de simulación de Hall and Shen (2010). El primero no presenta contaminación, mientras que los restantes tienen un 10% de los datos contaminados, ya sea en la variable respuesta, en las predictoras, o en ambas.

Los parámetros lineales se consideran $\beta = (2, 2)^\top$ y $\gamma = (-1, 1)^\top$. Las covariables se generan de la siguiente manera: $X_1 = 1$ para las primeras $n/2$ observaciones y $X_1 = 0$ en los demás casos, $X_2 \sim U[0, 1]$, $T \sim U[-2, 2]$, $Z_1 \sim U[0, 1]$, y $Z_2 \sim N(0, 1)$. La componente

no paramétrica se define como $m(t) = \sin(\frac{\pi}{2}t)$. Para esta configuración de parámetros, se generaron $N = 100$ conjuntos de datos con tamaño $n = 500$.

Los esquemas de contaminación considerados en este estudio son los siguientes:

Escenario C_0 : No contaminado.

Escenario C_1 : El 10% de la variable respuesta se contamina sumando un valor grande fijo $Y_0 = 70$.

Escenario C_2 : El 10% de la covariable X_2 se contamina, generándose ahora con distribución $U[1, 2]$. En las mismas observaciones, la respuesta se establece como $Y = 0$. Esto genera ceros falsos, ya que para estas observaciones contaminadas, $\mathbf{X}_i^\top \boldsymbol{\beta}$ es grande y, por lo tanto, la media esperada de Poisson $\lambda_i = e^{\mathbf{X}_i^\top \boldsymbol{\beta} + m(T_i)}$ sugeriría valores altos. Sin embargo, se observan ceros. Para este esquema de contaminación, es necesario incorporar también el peso $\omega_2(\mathbf{X}_i)$ en (2.6) para manejar adecuadamente la presencia de ceros falsos.

Escenario C_3 : Este escenario combina los esquemas C_1 y C_2 . Específicamente, el 5% de la covariable X_2 se contamina con distribución $U[1, 2]$, y en las mismas observaciones la respuesta se establece en $Y = 0$. Además, el 5% de la variable respuesta se contamina sumando un valor constante $Y_0 = 70$.

3.1.3. Medidas de Desempeño

Para cuantificar el error de \hat{m}_n , se calcula el conocido error cuadrático medio (RMSE, por sus siglas en inglés)

$$RMSE(\hat{m}) = \left(\frac{1}{n} \sum_{i=1}^n (\hat{m}_n(t_i) - m_0(t_i))^2 \right)^{1/2}.$$

En cuanto a los parámetros lineales, se calculan las distancias en norma ℓ_2 entre los estimadores obtenidos en cada muestra simulada y el valor verdadero del parámetro, es decir, $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|$ y $\|\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|$.

3.1.4. Parámetro de Suavizado

Los estimadores propuestos en la presente tesis requieren la especificación del parámetro de suavizado h_n , involucrado en los pesos del kernel de Nadaraya-Watson utilizados en (2.5). Para computarlo se utiliza un enfoque robusto de validación cruzada K -fold con $K = 5$, empleando como función de validación la misma pérdida robusta que se utiliza para computar los estimadores. El procedimiento viene dado por los siguientes pasos:

- I. Se divide la muestra aleatoriamente en K subconjuntos de tamaño aproximadamente igual. El conjunto de validación se conforma por un fold fijo, cuyos índices se denotan \mathcal{J}_α con $\alpha = 1/K$. Mientras que el conjunto de entrenamiento se conforma por los $K - 1$ folds restantes y sus índices se denotan $\mathcal{L}_{1-\alpha}$.
- II. Para un h fijo dado, se utiliza el conjunto de entrenamiento $\mathcal{L}_{1-\alpha}$ para computar los estimadores con el procedimiento descrito en la Sección 2.3. Dichos estimadores se denotan $\boldsymbol{\theta}^{(-\alpha)} = (\widehat{\boldsymbol{\beta}}_n^{(-\alpha)}(h), \widehat{\boldsymbol{\gamma}}_n^{(-\alpha)}(h), \widehat{m}_n^{(-\alpha)}(\cdot, h))$, donde el supraíndice $(-\alpha)$ indica que se computaron con el $(1 - \alpha)100\%$ de los datos.
- III. Se utiliza el conjunto de validación \mathcal{J}_α para calcular el valor de la función de validación como sigue

$$CV_k(h) = \sum_{i \in \mathcal{J}_\alpha} (1 - E_{\boldsymbol{\theta}^{(-\alpha)}}(W_i | Y_i)) \rho(Y_i, \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_n^{(-\alpha)}(h) + \widehat{m}_n^{(-\alpha)}(T_i, h)) \omega_1(\mathbf{X}_i).$$

El subíndice k indica el fold considerado como conjunto de validación.

- IV. Los ítems II. y III. se repiten utilizando cada uno de los K folds como conjunto de validación, obteniendo K valores de la función de validación para cada h , que luego se promedian, como sigue

$$CV(h) = \frac{1}{K} \sum_{k=1}^K CV_k(h).$$

- v. Se repiten los ítems II., III. y IV. para una grilla de valores de h y luego se elige el valor óptimo como

$$h_n = \underset{h}{\operatorname{argmin}} CV(h).$$

Notar que en el paso III. se precisa un método de aproximación (por ejemplo, utilizando splines) para evaluar $\widehat{m}_n^{(-\alpha)}(t, h)$ en T_i para $i \in \mathcal{J}_\alpha$, siempre que no se pueda hallar su forma explícita.

En el estudio de simulación se realizó el procedimiento de validación cruzada descrito anteriormente para los primeros diez conjuntos de datos simulados. Dada la similitud de los valores de h_n óptimos obtenidos y con el fin de reducir el costo computacional global, se adoptó el promedio de los diez valores óptimos como valor de h_n fijo para los conjuntos de datos restantes. Los valores del parámetro de suavizado utilizados para cada estimador fueron $h_{n,ML} = 0.126$, $h_{n,MT} = 0.135$ y $h_{n,CH} = 0.159$ para los estimadores ML, MT y CH, respectivamente.

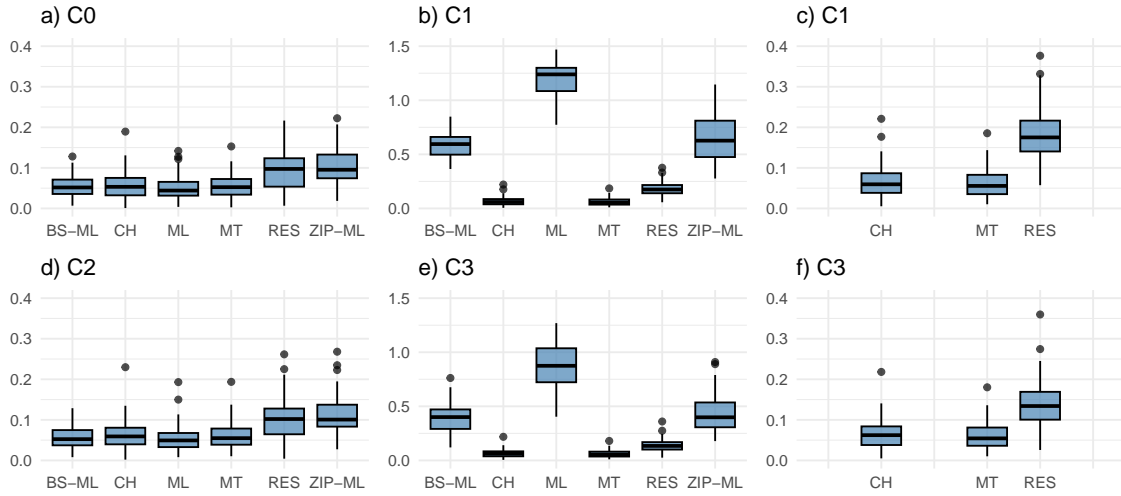


Figura 3.1: Distancia entre β_0 y los estimadores. Los paneles a), b), d) y e) representan cada esquema de contaminación. c) y f) son vistas ampliadas de los paneles b) y e), respectivamente.

3.1.5. Resultados

En esta subsección se presentan los resultados del estudio de simulación.

La Figura 3.1 muestra la distancia entre cada estimador de β y el valor verdadero del parámetro β_0 , para cada método y esquema de contaminación. En el panel a) se observa que bajo el esquema C_0 el estimador ML propuesto en esta tesis presenta el mejor desempeño, seguido por el estimador BS-ML y los estimadores robustos CH y MT. El comportamiento de los estimadores robustos no difiere significativamente del de los estimadores ML en este escenario, demostrando su eficiencia. En contraste, los estimadores RES y ZIP-ML presentan un desempeño inferior, ya que estiman un modelo ZIP cuando los datos fueron simulados bajo un modelo PLZIP. El panel d) muestra que el esquema C_2 no parece afectar la estimación de β , ya que el comportamiento de los estimadores es muy similar al del esquema C_0 . Esto es razonable, ya que la estimación de W_i por su esperanza condicional hace que los falsos ceros participen principalmente la estimación de γ . Por otro lado, los paneles b) y e) muestran que en los esquemas C_1 y C_3 , donde existe contaminación en la variable respuesta, los estimadores basados en máxima verosimilitud presentan un desempeño deficiente, como es de esperar. En este caso, los estimadores BS-ML y ZIP-ML son más conservadores que el ML, ya que este último ofrece mayor flexibilidad proporcionada por la componente no paramétrica estimada mediante la función de log-verosimilitud ponderada, permitiendo que el estimador se adapte más a los datos. Sin embargo, cuando los datos contienen valores extremos, esta flexibilidad puede volverse contraproducente. Los paneles c) y f) son ampliaciones de b) y e), respectivamente, que permiten examinar con mayor claridad el comportamiento de los estimadores robustos. Se observa que en ambos

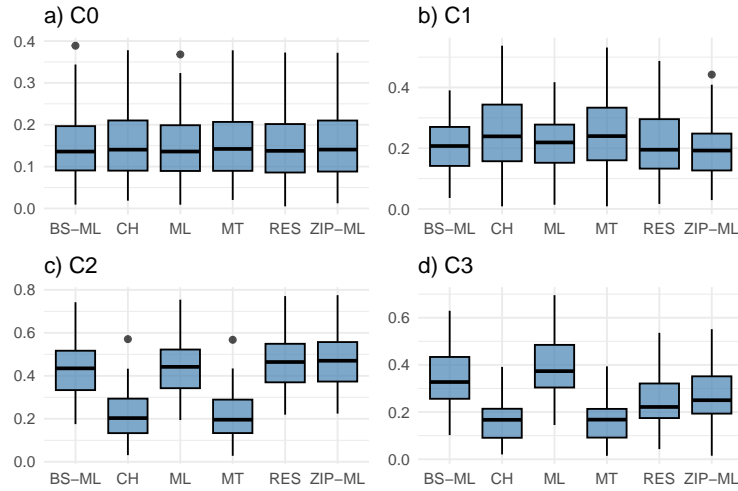


Figura 3.2: Distancia en norma ℓ_2 entre γ_0 y cada estimador. Los paneles a), b), c) y d) representan cada esquema de contaminación.

esquemas C_1 y C_3 , el estimador RES presenta el peor resultado, mientras que el estimador MT muestra el mejor desempeño, superando ligeramente a CH. En general, los estimadores robustos MT y CH propuestos en esta tesis muestran buen desempeño en datos no contaminados y fuerte resistencia a valores atípicos en datos contaminados.

En la Figura 3.2 se grafican las distancias en norma ℓ_2 entre los estimadores de γ y el valor verdadero del parámetro γ_0 . En el panel a) se observa que todos los estimadores presentan un desempeño similar. En este caso, los estimadores RES y ZIP-ML no se ven afectados por la especificación del modelo ya que el mismo no incluye una componente no paramétrica en π_i . El panel b) muestra que la contaminación bajo el esquema C_1 induce una ligera ineficiencia en los estimadores robustos MT, CH y RES. En los escenarios con contaminación en la covariable, C_2 y C_3 , los estimadores robustos MT y CH propuestos en esta tesis superan claramente a los demás métodos. Este resultado es razonable, ya que los pesos utilizados para controlar valores extremos de las covariables previenen que los ceros falsos afecten a los estimadores MT y CH. El estimador RES tiene un desempeño deficiente bajo la presencia de falsos ceros, posiblemente debido a la falta de pesos que controlen los valores extremos en \mathbf{X} .

La Figura 3.3 muestra los valores el RMSE de los estimadores de la función m , para aquellos métodos que permiten su estimación. En el esquema C_0 , el estimador BS-ML presenta un desempeño ligeramente mejor que los demás. Sin embargo, se puede observar que los estimadores robustos muestran un desempeño similar a aquellos basados en verosimilitud. En los esquemas restantes, el comportamiento de los estimadores de m es muy similar al de los estimadores de β_0 descrito en la Figura 3.1, ya que el parámetro β y la función m se estiman minimizando la misma función objetivo y, por lo tanto, los datos contaminados

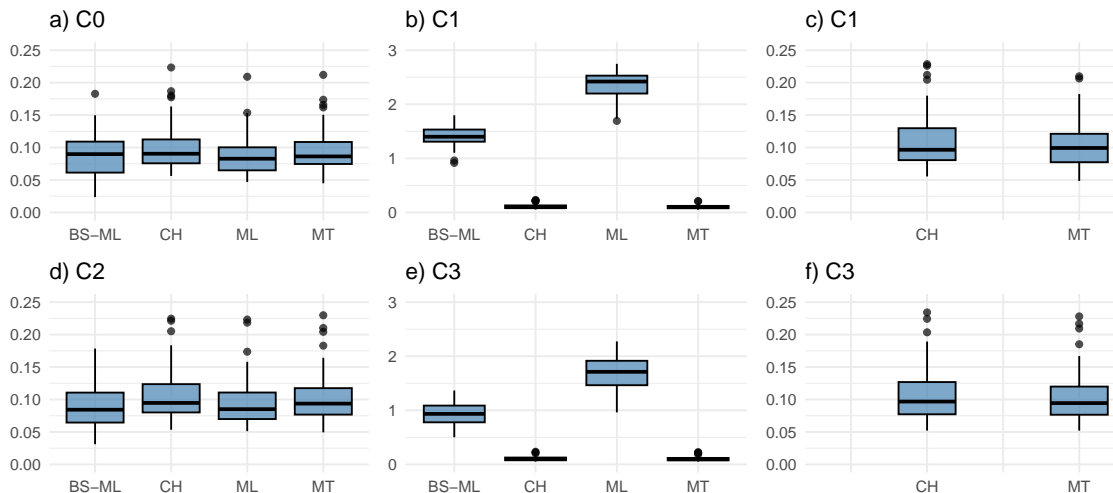


Figura 3.3: RMSE de cada estimador del componente no paramétrico m_0 . Los paneles a), b), d) y e) representan cada esquema de contaminación. c) y f) son vistas ampliadas de los paneles b) y e), respectivamente.

afectan ambos estimadores de manera similar. Específicamente, se observa que bajo el escenario C_2 , el desempeño de los estimadores es similar al de C_0 , mientras que en los esquemas C_1 y C_3 , los estimadores no robustos se ven muy afectados por valores extremos, mientras que los estimadores robustos propuestos en esta tesis, MT y CH, presentan un mejor desempeño. En los paneles c) y f), que muestran ampliaciones de los paneles b) y e) respectivamente, se puede observar que el comportamiento de ambos estimadores es comparable.

La Figura 3.4 ilustra un ejemplo de la función m estimada para una muestra particular. Se observa en los paneles a) y d) que bajo los esquemas C_0 y C_2 , todos los estimadores muestran un comportamiento similar. Particularmente cuando no hay contaminación todos logran estimar la función con precisión, mientras que bajo contaminación en la variable predictora (escenario C_2) los estimadores se ven levemente afectados. En contraste, se observa en los paneles b) y e) (cuyas ampliaciones se muestran en c) y f)) que bajo los escenarios C_1 y C_3 los estimadores ML y BS-ML basados en verosimilitud presentan un desempeño deficiente, mientras que los robustos mantienen un buen desempeño.

3.2. Ejemplos de Aplicación a Datos Reales

En esta sección se presentan ejemplos utilizando datos reales para evaluar el poder predictivo de los estimadores propuestos y sus competidores. Se llevan a cabo análisis detallados de dos conjuntos de datos: el primero se centra en el análisis de abundancia de la bacteria intestinal *Akkermansia* utilizando datos del *American Gut Project* (APG), mientras que

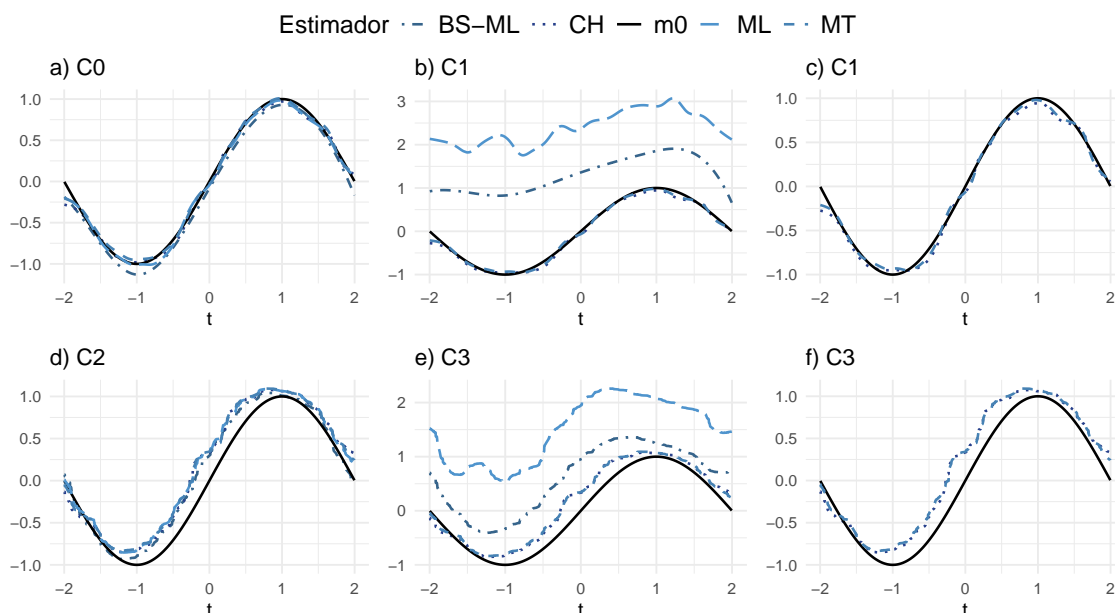


Figura 3.4: Función verdadera m_0 y sus estimaciones obtenidas a partir de un conjunto de datos simulado específico bajo cada esquema de contaminación.

el segundo se basa en una encuesta de salud pública realizada en Indonesia, destinada a analizar los días de ausencia a actividades principales por enfermedad en personas en edad laboral.

3.2.1. Ejemplo 1: Análisis de la abundancia de *Akkermansia*

En este ejemplo se analiza la abundancia de la bacteria intestinal *Akkermansia* (Y) en función del sexo ($X_1 = 1$ para femenino), la edad (X_2) y el índice de masa corporal (IMC) (T), utilizando datos del proyecto American Gut Project. La abundancia de la bacteria intestinal *Akkermansia* se determina a partir de datos de secuenciación del gen 16S rRNA. Cada muestra fecal es secuenciada y las secuencias obtenidas se asignan a unidades taxonómicas operativas (OTUs), permitiendo cuantificar cuántas lecturas corresponden a *Akkermansia*. En el conjunto de datos las abundancias varían entre 0 y 688, con media de 16.1 y varianza 2424.3. Dado que muchas muestras presentan cero conteos de esta bacteria, la distribución de los datos presenta un exceso de ceros, como puede apreciarse en la Figura 3.5. Específicamente, la proporción de ceros es aproximadamente 0.45. Además existen conteos particularmente extremos. En este contexto, un modelo inflado en ceros permite modelar simultáneamente la presencia o ausencia estructural de la bacteria y el número de lecturas en las muestras donde *Akkermansia* está presente, capturando adecuadamente la naturaleza de los datos.

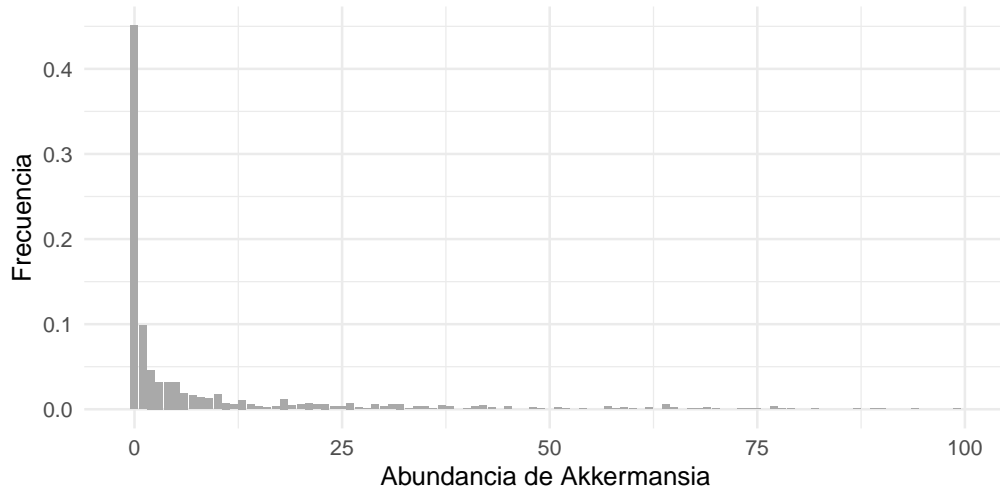


Figura 3.5: Gráfico de barras de las abundancias de la bacteria *Akkermansia* (El 3.8 % de los datos presentan valores mayores a 100 y no se muestran en el gráfico)

La selección de las variables predictoras a utilizar se fundamenta en los resultados reportados por el estudios de Zhou et al. (2020), donde se estudió el efecto de la obesidad (medida a través del índice de masa corporal), en la abundancia de *Akkermansia* en el intestino. Dicho efecto puede verse influido por factores como sexo, edad y hábitos alimentarios, que también han sido considerados por los autores. En particular, encontraron que la asociación entre la abundancia de *Akkermansia* y el riesgo de obesidad es no lineal y que disminuye con la edad. Asimismo, se identificó que el sexo y la edad son factores relevantes que afectan a la abundancia de la bacteria. Por otra parte, los autores observaron que la asociación permaneció prácticamente inalterada tras ajustar por consumo de tabaco, consumo de alcohol, dieta y país.

En esta tesis se evalúa la relación entre las covariables y la abundancia de *Akkermansia* a través del modelo PLZIP, donde el IMC se modela mediante una función no paramétrica

$$\begin{aligned}\text{logit}(\pi) &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 T, \\ \log(\lambda) &= \beta_1 X_1 + \beta_2 X_2 + m(T)\end{aligned}\tag{3.1}$$

utilizando los estimadores robustos MT y CH propuestos en esta tesis, así como la propuesta basada en verosimilitud, ML. Se utiliza también el estimador introducido por Wang et al. (2015), denotado BS-ML.

Además, se considera el modelo (3.2) cuando m es una función lineal, lo que resulta en el siguiente modelo ZIP

$$\begin{aligned}\text{logit}(\pi) &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 T, \\ \log(\lambda) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 T,\end{aligned}\tag{3.2}$$

que se estima a través de máxima verosimilitud (ZIP-ML) y mediante el estimador RES de Hall (2000). Asimismo, se realiza su estimación a través de un algoritmo tipo EM, utilizando como funciones de pérdida robustas para la estimación de β las propuestas por Valdora and Yohai (2014) y Croux and Haesbroeck (2002). Dichos estimadores se denotan ZIP-MT y ZIP-CH, respectivamente. Concretamente, la diferencia entre MT y ZIP-MT radica en que los estimadores MT corresponden a un modelo PLZIP mientras que ZIP-MT corresponden a un modelo ZIP lineal. Análogamente los estimadores CH estiman un modelo PLZIP mientras que ZIP-CH estiman un modelo ZIP lineal.

Para proporcionar una medida del error de predicción, se realizó una validación cruzada con $k = 5$ folds y se calculó la mediana del error cuadrático de predicción (*Root Median Square Error*) dentro de cada fold. Es decir,

$$RMedSE = \left(\text{Mediana} \{ (\hat{Y}_i - Y_i)^2 \}_{i=1}^{n_k} \right)^{1/2}$$

donde n_k representa la cantidad de observaciones en el k -ésimo fold. Se recurre a la mediana del error cuadrático ya que necesariamente los estimadores robustos predicen mal aquellas observaciones que son valores extremos. La media, a diferencia de la mediana, se vería demasiado afectada por los errores asociados a dichas observaciones.

En la Figura 3.6 se presentan los errores de predicción para todos los estimadores. En el Panel a) se observa que los modelos basados en verosimilitud presentan errores considerablemente mayores, indicando su sensibilidad a posibles valores atípicos en los datos. El estimador robusto RES muestra una ligera mejora, mientras que los estimadores robustos MT, ZIP-MT, CH y ZIP-CH alcanzan los menores errores, destacándose como los métodos más adecuados para este conjunto de datos. El Panel b) presenta un acercamiento a los errores de predicción de los estimadores robustos. Se aprecia que los estimadores MT y CH que estiman m de manera no paramétrica, superan respectivamente a sus versiones lineales correspondientes: ZIP-MT y ZIP-CH. En particular, el estimador MT alcanza los menores errores. Esto resalta la ventaja de combinar estimadores robustos con el modelado flexible, a través de la función m , de la covariable IMC.

La Figura 3.7 muestra las estimaciones de la función no paramétrica del IMC para los estimadores MT, CH, BS-ML y ML. Se observa que la estimación de ML sigue un patrón aproximadamente lineal, mientras que BS-ML presenta fluctuaciones considerables y una tendencia decreciente con el IMC. En contraste, los estimadores robustos MT y CH estiman una función aproximadamente constante, seguida de una disminución abrupta a partir de $IMC = 30$, el cual representa el límite de obesidad. Esto sugiere que la relación entre IMC y la abundancia de *Akkermansia* no puede capturarse adecuadamente mediante un modelo estrictamente lineal.

En el Cuadro 3.1 se observan los coeficientes de regresión estimados. En cuanto al efecto de la variable sexo, las estimaciones de β_1 son positivas en la mayoría de los modelos, lo

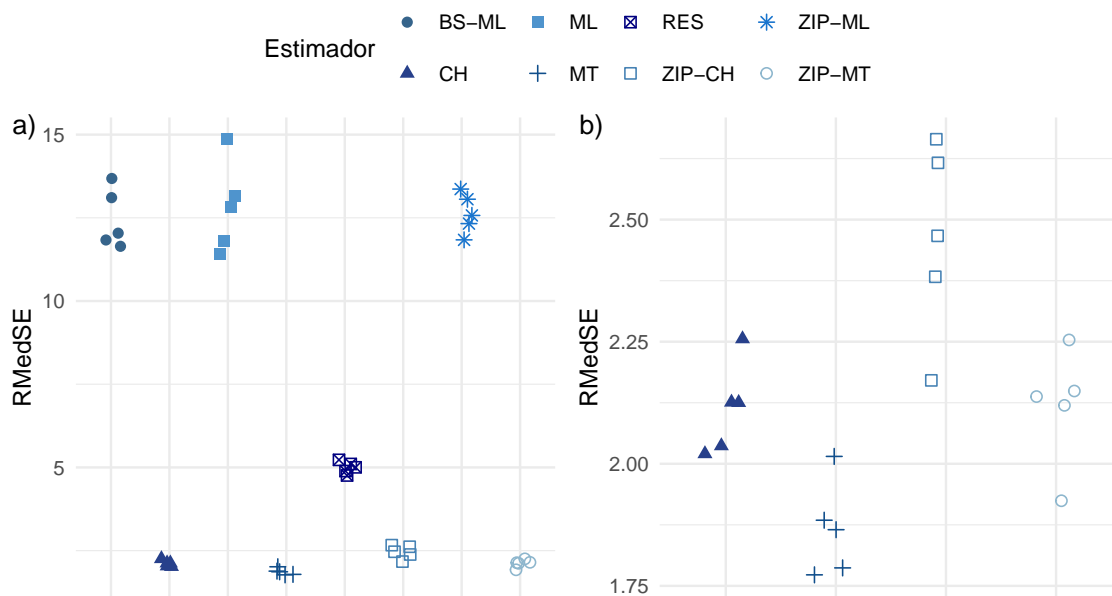


Figura 3.6: Panel a): mediana del error cuadrático de predicción de cada estimador, para los $k = 5$ folds. Panel b): ampliación del Panel a).

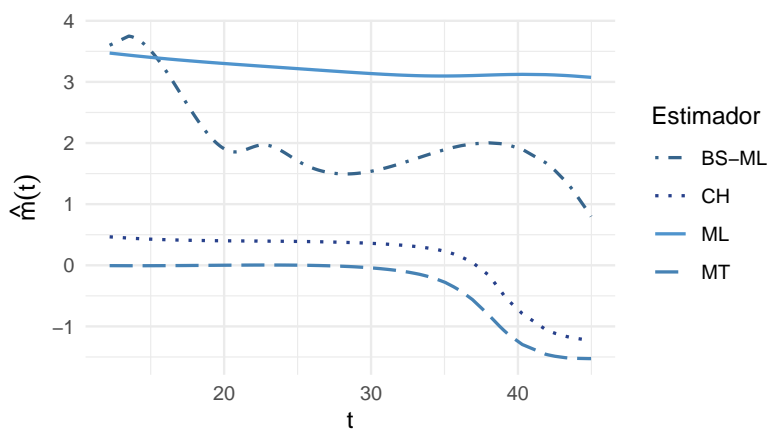


Figura 3.7: Estimadores de $m(t)$ basado en datos de APG.

que indicaría, en caso de que la variable sea significativa, que el sexo femenino ($X_1 = 1$) se asocia con una mayor abundancia de *Akkermansia*. Sin embargo, el estimador MT arroja un valor de $\hat{\beta}_1 = 0.15$ con un error estándar de 5.2, lo que evidencia que para el modelo PLZIP la variable sexo no estaría asociada a la abundancia de *Akkermansia*. En relación con la edad, los estimadores de β_2 en el modelo PLZIP presentan valores positivos, lo que sugeriría, en caso de tratarse de una variable significativa, que el aumento en la edad tiende a estar asociado con un aumento en la abundancia de *Akkermansia*. Específicamente, el estimador MT arroja un valor de $\hat{\beta}_2 = 0.027$ con un error estándar de 0.03, lo que indica una asociación de la edad con la abundancia de *Akkermansia*. Con respecto al IMC, los modelos que lo incluyen de manera lineal presentan estimaciones de β_3 negativas, indicando una disminución en la abundancia de *Akkermansia* cuando el IMC aumenta. Por su parte, las estimaciones de γ_1 son negativas de manera consistente, lo que refuerza la idea de que tener sexo femenino reduce la probabilidad de ceros estructurales en la abundancia de esta bacteria. Además, las estimaciones robustas de los coeficientes γ_2 y γ_3 son negativas, lo que indica que tanto la condición de joven como el bajo IMC incrementan la probabilidad de ceros estructurales, en línea con la evidencia de que el efecto de *Akkermansia* disminuye con el envejecimiento y con valores altos de IMC. Estos resultados mantienen coherencia con aquellos obtenidos en el estudio de Zhou et al. (2020).

	ZIP-ML	ML	BS-ML	RES	MT	CH	ZIP-MT	ZIP-CH
β_0	3.971	-	-	2.678	-	-	2.103	3.322
β_1	0.532	0.445	0.628	0.158	0.153	0.125	0.037	-0.101
β_2	-0.005	-0.005	0.027	-0.011	0.027	0.022	0.003	-0.009
β_3	-0.033	-	-	-0.004	-	-	-0.041	-0.061
γ_0	0.47	0.47	0.47	0.472	2.139	2.301	2.225	2.254
γ_1	-0.541	-0.541	-0.541	-0.541	-0.584	-0.578	-0.616	-0.625
γ_2	-0.011	-0.011	-0.011	-0.011	-0.03	-0.033	-0.027	-0.028
γ_3	0.001	0.001	0.001	0.001	-0.045	-0.044	-0.051	-0.049

Cuadro 3.1: Estimaciones de los coeficientes β y γ para los distintos estimadores aplicados a los datos de *Akkermansia*.

En conjunto, estos resultados muestran que los estimadores robustos del modelo PLZIP, y en particular el estimador MT propuesto en esta tesis, proporcionan un mejor ajuste y predicción frente a los métodos basados en verosimilitud y frente a sus versiones completamente lineales. La inclusión de la componente no paramétrica para el IMC es fundamental para reflejar la relación no lineal existente entre esta variable y la abundancia de *Akkermansia*, resaltando la importancia de emplear métodos robustos y flexibles en estudios de microbioma con posibles valores atípicos que no son fácilmente detectables.

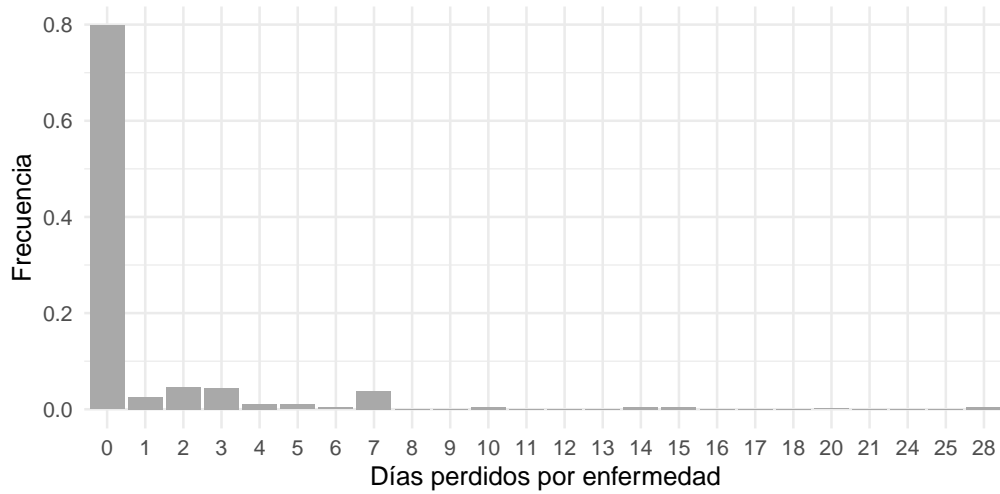


Figura 3.8: Gráfico de barras del número de días de ausencia en las actividades principales debido a enfermedad en las últimas 4 semanas, según lo auto-reportado por el encuestado.

3.2.2. Ejemplo 2: Días de Ausencia por Enfermedad

Este ejemplo se basa en datos de una encuesta de salud pública realizada en Indonesia en 1997. La variable respuesta de interés es el número de días que los individuos en edad laboral (18 a 60 años) dejaron de realizar sus actividades principales debido a enfermedad en las cuatro semanas previas, autorreportado por el encuestado. Las variables predictoras son: sexo ($X_1 = 1$ para femenino y 0 en caso contrario), ingreso anual per cápita del hogar (X_2 , en miles de rupias), índice de higiene del hogar (X_3 que varía de 0 a 5, de mejor a peor), y edad (T , de 18 a 60 años). Los entrevistadores calificaron 5 condiciones de higiene del hogar, entre ellas, si alrededor del hogar se encontraron desechos humanos y animales, montones de basura o agua estancada, si hay un establo debajo o junto a la casa, y si los jardines se encuentran bien mantenidos y limpios. El índice de higiene del hogar indica el número de condiciones que los entrevistadores calificaron negativamente. Para evitar posibles correlaciones entre los hogares encuestados, se seleccionó aleatoriamente una observación por hogar.

El número de días que los individuos dejaron de realizar sus actividades principales debido a enfermedad en las cuatro semanas previas varía entre 0 y 28, con media 1.05 y varianza 10.16. La varianza es mucho mayor que la media, lo que indica una posible sobredispersión. Como se observa en la Figura 3.8, los datos contienen un exceso de ceros. Específicamente, la proporción de ceros es aproximadamente 0.8. Además, existen algunos conteos particularmente extremos como 20 y 28.

Estos datos también son analizados por Lam et al. (2006), quienes asumen que la población se divide en un grupo de bajo riesgo, que genera ceros estructurales, y un grupo de

alto riesgo, cuyos conteos pueden modelarse mediante un modelo de regresión Poisson. Así, utilizan un modelo de regresión ZIP semiparamétrico. Los autores mencionan que la variable edad generalmente no es significativa cuando se asume una relación lineal en un contexto paramétrico, principalmente porque los efectos de la edad pueden ser muy diferentes entre distintos grupos etarios. Por lo tanto, postulan que la variable edad tiene una relación no lineal con el logaritmo natural de la media de días de ausencia en actividades principales debido a enfermedad entre los sujetos del grupo de alto riesgo.

En este ejemplo, se plantea el modelo PLZIP para estudiar la relación entre el número de días de ausencia en actividades principales y las predictoras. En base a la suposición de Lam et al. (2006), se incluye la variable edad en la componente no paramétrica, de modo que el modelo resulta

$$\begin{aligned}\text{logit}(\pi) &= \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 T, \\ \log(\lambda) &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + m(T).\end{aligned}$$

La mediana de los errores cuadráticos de predicción se presentan en la Figura 3.9. Las predicciones obtenidas con modelos estimados mediante máxima verosimilitud presentan mayores errores que las predicciones realizadas mediante métodos robustos, lo que sugiere la presencia de valores atípicos en los datos.

Entre los estimadores basados en verosimilitud (ZIP-ML, BS-ML y ML), ZIP-ML, que no incluye una componente no paramétrica, muestra el peor desempeño. El estimador BS-ML, que estima la componente no paramétrica mediante B-splines, parece ser una opción ligeramente mejor que ML. Esto podría deberse a que ML se adapta demasiado a los datos, debido a la flexibilidad que ofrecen los estimadores basados en núcleos, lo cual puede ser problemático en presencia de valores extremos en los datos.

Por otro lado, entre los estimadores robustos (RES, MT, ZIP-MT, CH y ZIP-CH), las predicciones obtenidas con el modelo PLZIP utilizando los estimadores MT y CH propuestos en esta tesis presentan errores más pequeños que aquellas realizadas con el modelo ZIP utilizando estimadores RES. Si bien el estimador MT presenta errores levemente menores, las diferencias no parecen ser significativas en comparación con CH y con las versiones completamente lineales ZIP-MT y ZIP-CH. Por ese motivo y para una mayor exploración de la relación entre las variables, se lleva a cabo también un análisis doblemente semiparamétrico. Específicamente, se plantea un modelo PLZIP que incorpora una componente no paramétrica tanto en la media de Poisson como también en la probabilidad de ceros estructurales. El modelo viene dado por

$$\begin{aligned}\log(\lambda) &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + m_\lambda(T) \\ \text{logit}(\pi) &= \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + m_\pi(T).\end{aligned}\tag{3.3}$$

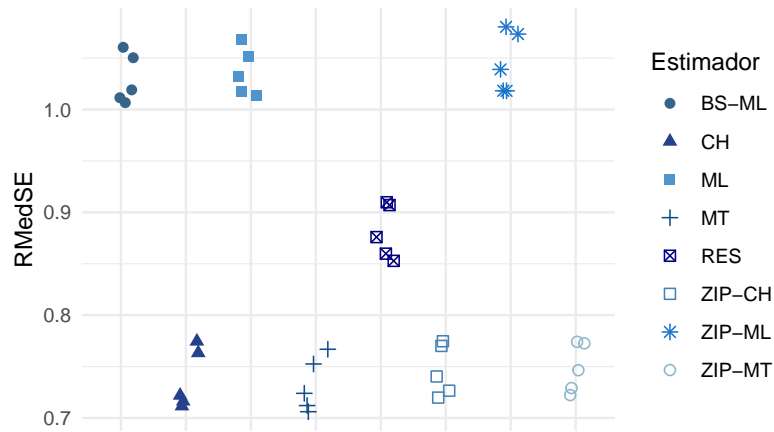


Figura 3.9: Mediana del error cuadrático de predicción de cada estimador, para los $k = 5$ folds.

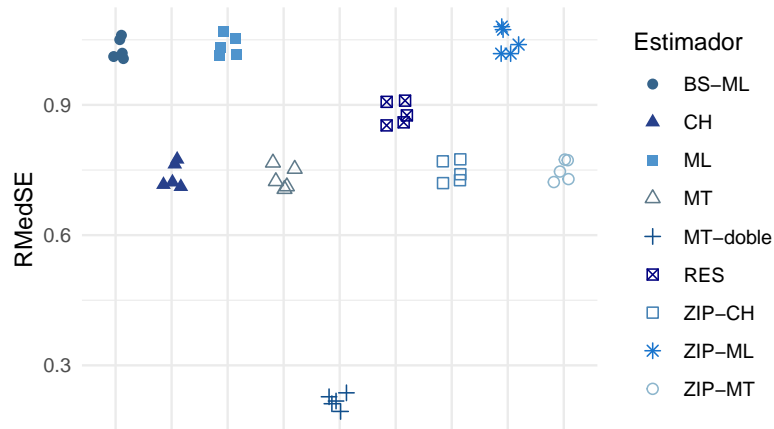


Figura 3.10: Mediana del error cuadrático de predicción de cada estimador, para los $k = 5$ folds.

El procedimiento de estimación para este modelo se detalla en la Sección A.5 del Apéndice. La estimación se realiza utilizando estimadores MT, que mostraron buenas propiedades en los estudios de simulación y en las aplicaciones a datos reales, y se denotan MT-doble. En la Figura 3.10 se adicionaron los errores del modelo doblemente semiparamétrico, que resultan considerablemente menores que los demás. Esto sugiere que el modelo 3.3 resulta una especificación más adecuada para el modelo.

El Cuadro 3.2 presenta las estimaciones de los parámetros de regresión β y γ obtenidas para los diferentes modelos y procedimientos de estimación. Se pueden hacer varias observaciones. En primer lugar, los signos y magnitudes de las estimaciones son en general

consistentes entre métodos. En cuanto a las estimaciones de β , el ingreso (X_1) muestra un efecto negativo, lo que indica que si se trata de una variable significativa, un mayor nivel socioeconómico se asocia con menos días de ausencia por enfermedad. En particular, el estimador MT arroja un valor de $\hat{\beta}_1 = -0.27$ con un error estándar de 0.07. El sexo (X_2) en general muestra un efecto positivo, sugiriendo, en caso de que la variable sea significativa, que las personas de sexo femenino tienden a reportar más ausencias por enfermedad. Específicamente, el estimador MT presenta un valor de $\hat{\beta}_2 = 0.167$ con un error estándar de 0.04. El índice de higiene (X_3) también presenta estimaciones positivas del coeficiente, lo cual sugiere que peores condiciones de higiene aumentan las ausencias. Para esta variable el estimador MT arroja $\hat{\beta}_3 = 0.003$ con un error estándar de 0.03. Los resultados indican que, según el modelo PLZIP propuesto utilizando el estimador MT, tanto el ingreso, como el sexo y la condición de higiene están asociadas al número de días de ausencia por enfermedad. Por otro lado, los estimadores de γ muestran patrones similares en todos los métodos: un mayor ingreso (X_1) incrementa la probabilidad de pertenecer al grupo de bajo riesgo, mientras que tener sexo femenino ($X_2 = 1$), las malas condiciones de higiene del hogar (X_3) y la edad (T) la disminuyen. Cabe destacar que estos resultados mantienen coherencia con los obtenidos por Lam et al. (2006).

	ZIP-ML	ML	BS-ML	RES	MT	CH	ZIP-MT	ZIP-CH	doble-MT
β_0	1.633	-	-	1.393	-	-	1.076	1.068	-
β_1	-0.017	-0.008	-0.017	-0.018	-0.27	-0.302	-0.187	-0.22	0.653
β_2	-0.009	-0.002	0.05	0.048	0.167	0.176	0.152	0.158	0.304
β_3	0.013	0.015	0.016	0.019	0.003	0.006	0.008	0.009	-0.087
β_4	0.098	-	-	0.084	-	-	0.046	0.064	-
γ_0	1.606	1.606	1.604	1.592	1.546	1.549	1.545	1.542	-
γ_1	0.04	0.041	0.04	0.04	0.075	0.075	0.052	0.046	0.724
γ_2	-0.396	-0.396	-0.394	-0.393	-0.371	-0.372	-0.379	-0.378	-0.317
γ_3	-0.048	-0.048	-0.048	-0.047	-0.048	-0.047	-0.059	-0.059	-0.117
γ_4	-0.109	-0.11	-0.11	-0.106	-0.098	-0.096	-0.095	-0.092	-

Cuadro 3.2: Estimadores de β y γ .

La Figura 3.11 muestra las estimaciones de $m(t)$ para cada método parcialmente lineal con componente no paramétrica en la media de Poisson. Se observa que los estimadores basados en verosimilitud fluctúan, pero en general muestran una tendencia creciente en el efecto de la edad sobre la respuesta. En cambio, los estimadores robustos muestran que el efecto de la edad disminuye hasta los 30 años, luego aumenta y no presenta demasiadas variaciones entre los 50 y 60 años. Esto corrobora que la edad tiene un efecto no lineal sobre la respuesta y sugiere incluir $m(t)$ como componente no paramétrico en la especificación del modelo, lo que permite que el efecto de la edad varíe de manera flexible entre grupos etarios, lo cual explica el mejor desempeño de los estimadores robustos parcialmente lineales. Por otro lado, en la Figura 3.12 se graficaron las estimaciones MT de m_λ y m_π del modelo ZIP

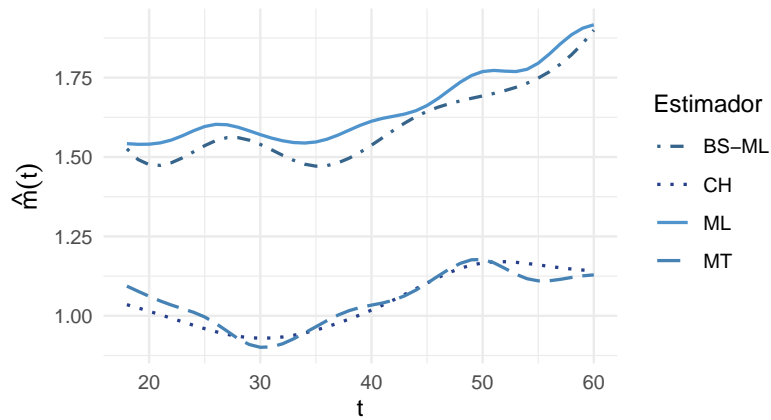


Figura 3.11: Estimadores de $m(t)$ basado en datos de la encuesta de salud pública en Indonesia.

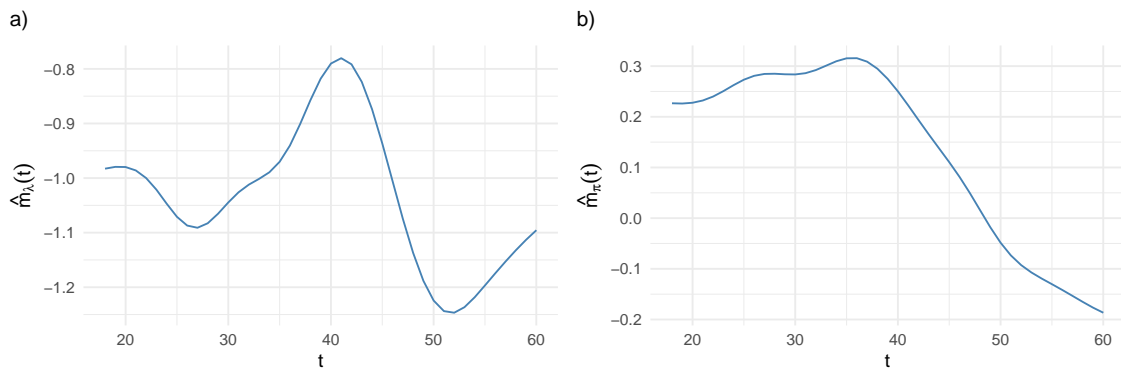


Figura 3.12: Estimadores de $m_\lambda(t)$ y $m_\pi(t)$ basado en datos de la encuesta de salud pública en Indonesia.

doblemente semiparamétrico.

En el caso doblemente semiparamétrico la interpretación de los estimadores de los coeficientes y las funciones no paramétricas estimadas se vuelve un poco más compleja. En principio, en los parámetros estimados se observan efectos opuestos del ingreso y del índice de higiene. En cuanto a la estimación de las componentes no paramétricas, \hat{m}_λ indica que la edad tiene un efecto mayor en los 40 años; mientras que \hat{m}_π indica que el efecto de la edad en la probabilidad de pertenecer al grupo de bajo riesgo es aproximadamente constante hasta los 40 años y luego disminuye. En términos generales, se puede concluir que los grupos etarios de jóvenes de entre 18 y 30-40 años tienen un efecto diferente al del grupo de personas mayores de entre 30-40 y 60 años.

Este análisis preliminar de datos reales muestra que los métodos de estimación robusta

son una herramienta útil que permite estimar la variable de respuesta de forma eficiente cuando la presencia de valores atípicos no es fácilmente detectable. Esto sugiere que las técnicas propuestas de estimación robusta podrían ser herramientas valiosas en conjuntos de datos con valores atípicos, mejorando la fiabilidad del análisis.

A.1. Resultados clásicos

Teorema A.1.1 (Ley fuerte de los Grandes Números). *Sea $\{\mathbf{X}_i\}_{i \geq 1}$ una sucesión de vectores aleatorios independientes e idénticamente distribuidos en \mathbb{R}^d tales que $E\|\mathbf{X}_1\| < \infty$. Sea $\boldsymbol{\mu} = E(\mathbf{X}_1)$ y $\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$. Entonces*

$$\bar{\mathbf{X}}_n \xrightarrow{c.s.} \boldsymbol{\mu}.$$

Teorema A.1.2 (Teorema de Slutsky 1). *Sean $\{\mathbf{X}_n\}_{n \geq 1}$ e $\{\mathbf{Y}_n\}_{n \geq 1}$ sucesiones de vectores aleatorios. Entonces:*

(a) *Si $\mathbf{X}_n \in \mathbb{R}^d$, $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$, y $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ es tal que $P(\mathbf{X} \in C(f)) = 1$, entonces $f(\mathbf{X}_n) \xrightarrow{\mathcal{D}} f(\mathbf{X})$.*

(b) *Si $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ y $\mathbf{X}_n - \mathbf{Y}_n \xrightarrow{P} 0$, entonces $\mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{X}$.*

(c) *Si $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ y $\mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{c}$, entonces $\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix}$.*

Teorema A.1.3 (Teorema de Slutsky 2). *Sean $\{\mathbf{X}_n\}_{n \geq 1}$ e $\{\mathbf{Y}_n\}_{n \geq 1}$ sucesiones de vectores aleatorios. Entonces:*

(a) *Si $\mathbf{X}_n \in \mathbb{R}^d$, $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, y $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ es tal que $P(\mathbf{X} \in C(f)) = 1$, entonces $f(\mathbf{X}_n) \xrightarrow{P} f(\mathbf{X})$.*

(b) *Si $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ y $\mathbf{X}_n - \mathbf{Y}_n \xrightarrow{P} 0$, entonces $\mathbf{Y}_n \xrightarrow{P} \mathbf{X}$.*

(c) *Si $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ y $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$, entonces $\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$.*

También son válidos los resultados para convergencia casi segura.

Teorema A.1.4 (Desigualdad de Jensen). *Sea (Ω, \mathcal{A}, P) un espacio de probabilidad, $f : \Omega \rightarrow \mathbb{R}$ una variable aleatoria integrable y sea $\phi : \mathbb{R} \rightarrow \mathbb{R}$ una función convexa, entonces se cumple*

$$\phi\left(\int_{\Omega} f dP\right) \leq \int_{\Omega} \phi(f) dP.$$

Teorema A.1.5 (Teorema de Taylor con resto en forma de Lagrange). *Sea $I \subset \mathbb{R}$ un intervalo abierto y $f : I \rightarrow \mathbb{R}$ una función de clase $\mathcal{C}^{k+1}(I)$. Sean $a, x \in I$. Entonces existe $\xi = a + \theta(x - a)$ para algún $\theta \in (0, 1)$ tal que*

$$f(x) = \sum_{j=0}^k \frac{f^{(j)}(a)}{j!} (x - a)^j + R_{k+1}(x),$$

donde el resto en la forma de Lagrange está dado por

$$R_{k+1}(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!} (x - a)^{k+1}.$$

Teorema A.1.6 (Teorema Central del Límite para vectores aleatorios iid). *Sea $\{\mathbf{X}_i\}_{i \geq 1}$ una sucesión de vectores aleatorios independientes e idénticamente distribuidos en \mathbb{R}^d . Supóngase que $E(\|\mathbf{X}_1\|^2) < \infty$ y que la sucesión de vectores tienen media común $\boldsymbol{\mu} = E(\mathbf{X}_1) \in \mathbb{R}^d$ y matriz de covarianza común $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}_1) \in \mathbb{R}^{d \times d}$. Entonces, la media muestral $\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$ cumple que*

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{cuando } n \rightarrow \infty.$$

Lema A.1.1. *Sean $\{\mathbf{X}_n\}_{n \geq 1}$ e $\{\mathbf{Y}_n\}_{n \geq 1}$ sucesiones de vectores aleatorios en \mathbb{R}^d . Si $\mathbf{X}_n = O_p(1)$ y $\mathbf{Y}_n = o_p(1)$, entonces $\mathbf{X}_n^\top \mathbf{Y}_n = o_p(1)$.*

Lema A.1.2. *Sea $\{\mathbf{X}_n\}$ una sucesión de vectores aleatorios en \mathbb{R}^d . Si $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$, para algún vector aleatorio \mathbf{X} en \mathbb{R}^d . Entonces $\mathbf{X}_n = O_p(1)$, es decir, la sucesión $\{\mathbf{X}_n\}$ es acotada en probabilidad.*

Lema A.1.3 (Desigualdad de Markov). *Sea \mathbf{X} un vector aleatorio en \mathbb{R}^d y sea $\|\cdot\|$ una norma en \mathbb{R}^d . Entonces, para todo $\varepsilon > 0$ se cumple que*

$$P(\|\mathbf{X}\| \geq \varepsilon) \leq \frac{E(\|\mathbf{X}\|^p)}{\varepsilon^p},$$

para cualquier $p > 0$ tal que $E(\|\mathbf{X}\|^p) < \infty$.

Observación A.1.1. *Si $\|\cdot\|$ es la norma euclídea y $E(\|\mathbf{X}\|^2) < \infty$, entonces para $p = 2$, se tiene que*

$$P(\|\mathbf{X}\| \geq \varepsilon) \leq \frac{\|E(\mathbf{X})\|^2 + \text{tr}(\text{Cov}(\mathbf{X}))}{\varepsilon^2},$$

ya que

$$\begin{aligned} E(\|\mathbf{X}\|^2) &= E(\mathbf{X}^\top \mathbf{X}) \\ &= E[(\mathbf{X} - E(\mathbf{X}))^\top (\mathbf{X} - E(\mathbf{X})) + 2E(\mathbf{X})^\top \mathbf{X} - E(\mathbf{X})^\top E(\mathbf{X})] \\ &= E[(\mathbf{X} - E(\mathbf{X}))^\top (\mathbf{X} - E(\mathbf{X}))] + \|E(\mathbf{X})\|^2, \end{aligned}$$

luego tomando traza y usando su propiedad cíclica se tiene

$$E(\|\mathbf{X}\|^2) = \text{tr}(\text{Cov}(\mathbf{X})) + \|E(\mathbf{X})\|^2.$$

Teorema A.1.7. Sea \mathbf{Z} , una variable aleatoria con distribución normal multivariada estándar de dimensión k , $\sim N_k(\mathbf{0}, I_k)$. Entonces la forma cuadrática

$$Q = \mathbf{Z}^\top \mathbf{Z} = \sum_{i=1}^k Z_i^2$$

sigue una distribución chi-cuadrado con k grados de libertad, es decir $Q \sim \chi_k^2$.

Más generalmente, si $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma}$ definida positiva, se cumple que

$$Q = (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2.$$

Lema A.1.4. Sean C y D conjuntos compactos en \mathbb{R}^d y \mathbb{R}^p respectivamente, y sea $f(\mathbf{X}, \boldsymbol{\theta})$ una función continua en $\boldsymbol{\theta} \in C$ y $\mathbf{X} \in D$. Supongamos que $\hat{\boldsymbol{\theta}}(\mathbf{X}) \in C$ es continua en $\mathbf{X} \in D$ y es el único maximizador de $f(\mathbf{X}, \boldsymbol{\theta})$. Sea $\hat{\boldsymbol{\theta}}_n(\mathbf{X}) \in C$ un maximizador de $f_n(\mathbf{X}, \boldsymbol{\theta})$.

Si

$$\sup_{\boldsymbol{\theta} \in C, \mathbf{X} \in D} |f_n(\mathbf{X}, \boldsymbol{\theta}) - f(\mathbf{X}, \boldsymbol{\theta})| \rightarrow 0,$$

entonces

$$\sup_{\mathbf{X} \in D} |\hat{\boldsymbol{\theta}}_n(\mathbf{X}) - \hat{\boldsymbol{\theta}}(\mathbf{X})| \rightarrow 0, \quad \text{cuando } n \rightarrow \infty.$$

Este resultado corresponde al Lema A1 de Carroll et al. (1997) pág. 486.

A.2. Algoritmo ES

Sean $\{(Y_i, W_i)\}_{i=1}^n$ los datos completos, donde Y_i es observable y W_i no lo es. Sea $s(Y, W; \boldsymbol{\theta})$ la función utilizada para definir la ecuación de estimación.

A partir de un valor inicial $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\theta}$, para $r = 0, 1, 2, \dots$, se computa la iteración $r + 1$ como

(Paso E)

$$S_n(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(r)}) = \sum_{i=1}^n E_{\boldsymbol{\theta}^{(r)}}[s(Y_i, W_i; \boldsymbol{\theta}')|Y_i].$$

(Paso S) $\boldsymbol{\theta}^{(r+1)}$ se obtiene resolviendo la ecuación

$$S_n(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{\theta}^{(r)}) = 0.$$

Proposición A.2.1. *Si existe un punto $\widehat{\boldsymbol{\theta}} \in \Theta$ tal que*

$$\lim_{r \rightarrow \infty} \boldsymbol{\theta}^{(r)} = \widehat{\boldsymbol{\theta}},$$

donde $\{\boldsymbol{\theta}^{(r)}\}$ es la sucesión generada por el algoritmo ES. Si se verifican las condiciones

I. $s(\cdot, \cdot; \boldsymbol{\theta})$ es una función de estimación insesgada, es decir, para todo $\boldsymbol{\theta} \in \Theta$ se cumple

$$E_{\boldsymbol{\theta}}[s(Y, W; \boldsymbol{\theta})] = 0$$

II. La función

$$S_n(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \sum_{i=1}^n E_{\boldsymbol{\theta}}[s(Y_i, W_i; \boldsymbol{\theta}')]$$

es continua en $(\boldsymbol{\theta}', \boldsymbol{\theta}) \in \Theta \times \Theta$.

Entonces se cumple que

(a) $\widehat{\boldsymbol{\theta}}$ satisface $S_n(\widehat{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}}) = 0$.

(b) $S_n(\boldsymbol{\theta}|\boldsymbol{\theta})$ define una ecuación de estimación insesgada, es decir,

$$E_{\boldsymbol{\theta}}[S_n(\boldsymbol{\theta}|\boldsymbol{\theta})] = 0, \quad \text{para todo } \boldsymbol{\theta} \in \Theta.$$

Para más detalles sobre el algoritmo ES consultar Rosen et al. (2000). Específicamente la demostración de esta proposición se encuentra en el apéndice, pág. 14.

A.3. Procesos empíricos

Para una muestra independiente proveniente de una función de distribución F , la ley fuerte de los grandes números establece que la proporción de puntos en un intervalo $(-\infty, t]$ converge casi seguramente a $F(t)$. El clásico teorema de Glivenko-Cantelli refuerza este resultado al añadir que la convergencia es uniforme en todo t . En esta Sección se introducen los resultados necesarios para la convergencia uniforme sobre clases de conjuntos más generales que intervalos en la recta real, por ejemplo, clases de funciones.

En lo que sigue se presentan resultados útiles para determinar cuando una familia de funciones es polinómica.

Definición A.3.1. Sea \mathcal{F} una familia de funciones, se denomina **envolvente** para la familia \mathcal{F} a una función medible F que verifica $|f| \leq F$ para toda $f \in \mathcal{F}$.

Sea Q una medida de probabilidad y $p \geq 1$, se denota $Q(F^p) = \int F^p dQ$.

Lema A.3.1. Si \mathcal{F} y \mathcal{G} son clases de funciones con envolventes F y G , y Q es una medida con $Q(F^p) < \infty$ y $Q(G^p) < \infty$ para algún $p \geq 1$, entonces la clase

$$\mathcal{F} + \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\},$$

con envolvente $F + G$ satisface

$$\mathcal{N}(2\varepsilon_1 + 2\varepsilon_2, \mathcal{F} + \mathcal{G}, L^1(Q)) \leq \mathcal{N}(\varepsilon_1, \mathcal{F}, L^1(Q))\mathcal{N}(\varepsilon_2, \mathcal{G}, L^1(Q)).$$

Este resultado corresponde al lema 16 (pág. 796) de Nolan and Pollard (1987). Allí se puede encontrar su respectiva demostración.

Lema A.3.2. Sea \mathcal{F} una clase polinómica de funciones definidas en un conjunto \mathcal{X} , y sea $\phi : \mathbb{R} \rightarrow \mathbb{R}$ una función monótona fija, entonces $\phi \circ \mathcal{F}$ es una clase polinómica.

Este lema se sigue de la Definición 7.11, el Teorema 7.12 y Lema 7.19 (viii) de Van der Vaart and Wellner (1996).

Lema A.3.3. Sea $\lambda(\cdot)$ una función de valor real de variación acotada en \mathbb{R} . La clase de todas las funciones en \mathbb{R}^d de la forma

$$f(x) = \lambda(\boldsymbol{\alpha}^\top x + \beta),$$

con $\boldsymbol{\alpha} \in \mathbb{R}^d$ y $\beta \in \mathbb{R}$, es polinómica para una envolvente constante.

La demostración de este lema se encuentra en Nolan and Pollard (1987) (Lema 22 (ii), pág. 797).

Lema A.3.4. Sea $r \geq 1$, \mathcal{F} y \mathcal{G} clases de funciones medibles con envolventes F y G , respectivamente. Entonces, para todo $0 < \varepsilon < 1$, se cumple que

$$\sup_Q \mathcal{N}(2\varepsilon \|FG\|_{Q,2}, \mathcal{F} \cdot \mathcal{G}, L_2(Q)) \leq \sup_Q \mathcal{N}(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \sup_Q \mathcal{N}(\varepsilon \|G\|_{Q,2}, \mathcal{G}, L_2(Q))$$

donde $\|f\|_{Q,2} = (\int |f|^2 dQ)^{1/2}$ y

$$\mathcal{F} \cdot \mathcal{G} := \{fg : f \in \mathcal{F}, g \in \mathcal{G}\},$$

y los supremos se toman sobre todos los subconjuntos apropiados de medidas de probabilidad finitamente discretas Q .

La demostración de este lema se puede encontrar en las notas A Gentle Introduction to Empirical Process Theory and Applications (Lema 7.21 (ii)).

A continuación, siguiendo Pollard (1984), se presentan resultados adicionales que permiten determinar si una clase es polinómica. Asimismo, se incluyen resultados que garantizan la convergencia uniforme de distribuciones empíricas sobre clases de funciones. De acuerdo al autor, se denomina **permisible** a una clase que satisface las condiciones de regularidad necesarias para evitar dificultades de medibilidad en el análisis de clases polinómicas de conjuntos.

Teorema A.3.1. *Sea \mathcal{F} una clase permisible de funciones con envolvente F . Supongamos que $PF < \infty$. Si P_n se obtiene mediante muestreo independiente de la medida de probabilidad P y si*

$$\log N_1(\varepsilon, P_n, \mathcal{F}) = o_p(n), \quad \text{para cada } \varepsilon > 0 \text{ fijo,}$$

entonces

$$\sup_{f \in \mathcal{F}} |P_n f - P f| \rightarrow 0 \quad \text{casi seguramente.}$$

Este resultado corresponde al Teorema 24 (pág. 25) de Pollard (1984).

Lema A.3.5 (Lema de Aproximación). *Sea \mathcal{F} una clase de funciones en un conjunto S con envolvente F , y sea Q una medida de probabilidad en S con $0 < QF < \infty$. Si los gráficos de funciones en \mathcal{F} forman una clase polinómica de conjuntos, entonces*

$$N_1(\varepsilon QF, Q, \mathcal{F}) \leq A\varepsilon^{-W}, \quad 0 < \varepsilon < 1,$$

donde las constantes A y W dependen únicamente del polinomio discriminante de la clase de gráficos.

Este resultado corresponde al Teorema 25 (pág. 27) de Pollard (1984).

Lema A.3.6. *Sea \mathcal{F} un espacio vectorial de dimensión finita de funciones reales en S . La clase de gráficos de funciones en \mathcal{F} tiene discriminación polinómica.*

Este resultado corresponde al Teorema 28 (pág. 30) de Pollard (1984).

Teorema A.3.2. *Para cada n , sea \mathcal{F}_n una clase permisible de funciones cuyos números de recubrimiento satisfacen*

$$\sup_Q N_1(\varepsilon, Q, \mathcal{F}_n) \leq A\varepsilon^{-W} \quad \text{para } 0 < \varepsilon < 1$$

con constantes A y W que no dependen de n . Sea $\{\alpha_n\}$ una sucesión no creciente de números positivos para la cual $n\delta_n^2\alpha_n^2 \gg \log n$. Si $|f| \leq 1$ y $\|f\|_{P,2} \leq \delta_n$ para cada $f \in \mathcal{F}_n$, entonces

$$\sup_{f \in \mathcal{F}_n} |P_n f - P f| \ll \delta_n^2 \alpha_n \quad \text{casi seguramente.}$$

Este resultado corresponde al Teorema 37 (pág. 34) de Pollard (1984).

A.4. Propiedades asintóticas

Lema A.4.1. Sea $(T_1, Z_1), \dots, (T_n, Z_n)$ una muestra i.i.d. de vectores aleatorios, donde Z_i son variables aleatorias escalares tales que $|Z_i| \leq A$. Sea K una función positiva y acotada, que satisface una condición de Lipschitz de orden uno y tal que $\int K(u)du = 1$. Sea \mathcal{T} el soporte de T un conjunto compacto y sea f_T la densidad de T una función acotada, entonces:

(a) Para todo $\epsilon > 0$ y n suficientemente grande tal que $\theta_n \leq 1$, se cumple

$$\begin{aligned} P \left(\theta_n^{-1} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n K \left(\frac{t - T_i}{h_n} \right) Z_i - E \left[K \left(\frac{t - T_i}{h_n} \right) Z_i \right] \right| > 2\epsilon \right) \\ \leq 2C_1 \delta_n^{-1} \exp \left\{ -(4C_2)^{-1} A_1^2 \epsilon^2 n h_n \theta_n^2 \right\}, \end{aligned}$$

donde $\delta_n < \epsilon \theta_n h_n^2 / (2AC_K)$, $C_2 = 2 \left(1 + \frac{1}{3} \epsilon A_1 \right) A^2 \|K\|_\infty \|f_T\|_\infty$, $C_1 = \text{diam}(\mathcal{T}) + 1$ y C_K denota la constante de Lipschitz del núcleo K .

(b) Si, además, $nh^4 \rightarrow 0$ y $nh_n^2 / \log^2(1/h) \rightarrow \infty$, entonces

$$n^{1/4} \sup_{t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^n K \left(\frac{t - T_i}{h_n} \right) Z_i - E \left[K \left(\frac{t - T_i}{h_n} \right) Z_i \right] \right| \xrightarrow{p} 0.$$

Este resultado corresponde al Lema A.1. de Boente and Rodriguez (2010). En el artículo se puede encontrar una detallada demostración.

Convergencia $B_n^{(8)} \xrightarrow{P} 0$:

$$B_n^{(8)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h_n n} \sum_{j=1}^n K \left(\frac{T_i - T_j}{h_n} \right) [\kappa(T_j) - \kappa(T_i)] G_i(T_j).$$

Mediante el cambio de variables $u = T_1$, $v = \frac{T_1 - T_2}{h_n}$ se tiene que

$$\begin{aligned} E_0(\mathbf{B}_n^{(8)}) &= \frac{n(n-1)}{n\sqrt{n}} E_0 \left(\frac{1}{h_n} K \left(\frac{T_1 - T_2}{h_n} \right) [\boldsymbol{\kappa}(T_1) - \boldsymbol{\kappa}(T_2)] G_1(T_2) \right) \\ &= \frac{n-1}{\sqrt{n}} E_0 \left(\frac{1}{h_n} K \left(\frac{T_1 - T_2}{h_n} \right) [\boldsymbol{\kappa}(T_1) - \boldsymbol{\kappa}(T_2)] m_1(T_1, T_2) \right) \\ &= -\frac{n-1}{\sqrt{n}} \int K(v) [\boldsymbol{\kappa}(u) - \boldsymbol{\kappa}(u - h_n v)] m_1(u, u - h_n v) f_T(u) f_T(u - h_n v) du dv. \end{aligned} \quad (\text{A.1})$$

Mediante un desarrollos de Taylor de primer orden y usando que $m_1(u, u) = 0$ tenemos

$$\begin{aligned} \boldsymbol{\kappa}(u - h_n v) &= \boldsymbol{\kappa}(u) - \boldsymbol{\kappa}'(\xi_1) h_n v \\ m_1(u, u - h_n v) &= -m_1'(u, \xi_2) h_n v \end{aligned}$$

con $\xi_1, \xi_2 \in (u - h_n v, u)$ puntos intermedios. Luego reemplazando en (A.1)

$$E_0(\mathbf{B}_n^{(8)}) = \frac{n-1}{\sqrt{n}} \int K(v) \boldsymbol{\kappa}'(\xi_1) m_1'(u, \xi_2) h_n^2 v^2 f_T(u) f_T(u - h_n v) dv du.$$

lo cual usando N7, N8 y N9 implica que $\|E_0(\mathbf{B}_n^{(8)})\|^2 = \mathcal{O}(nh_n^4) \rightarrow 0$. Usando argumentos análogos se puede probar que la traza de su matriz de covarianza converge a cero. En efecto,

$$\mathbf{B}_n^{(8)} = \frac{1}{n^{3/2}} \sum_{i \neq j} \frac{1}{h_n} K \left(\frac{T_i - T_j}{h_n} \right) [\boldsymbol{\kappa}(T_j) - \boldsymbol{\kappa}(T_i)] G_i(T_j) = \frac{1}{n^{3/2}} \sum_{i \neq j} \mathbf{V}_{ij}$$

entonces

$$\begin{aligned} \text{Cov}(\mathbf{B}_n^{(8)}, \mathbf{B}_n^{(8)}) &= \frac{1}{n^3} \sum_{i \neq j} \sum_{l \neq s} \text{Cov}(\mathbf{V}_{ij}, \mathbf{V}_{ls}) \\ &= \frac{n(n-1)}{n^3} \text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12}) \\ &\quad + \frac{n(n-1)(n-2)}{n^3} [\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13}) + \text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{32})]. \end{aligned}$$

Para analizar el término $\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})$ se define $m_2(T_1, T_2) = E_0(G_1(T_2)^2 | (T_1, T_2))$. Luego, haciendo un desarrollo de Taylor se tiene

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) &\leq \text{tr}(E(\mathbf{V}_{12} \mathbf{V}_{12}^\top)) \\ &= \frac{1}{h_n^2} E_0 \left(K \left(\frac{T_1 - T_2}{h_n} \right)^2 \text{tr} \left([\boldsymbol{\kappa}(T_1) - \boldsymbol{\kappa}(T_2)] [\boldsymbol{\kappa}(T_1) - \boldsymbol{\kappa}(T_2)]^\top \right) \cdot \right. \\ &\quad \left. E_0(G_1(T_2)^2 | T_1, T_2) \right) \\ &= -\frac{1}{h_n} \int K(v)^2 \text{tr} \left([\boldsymbol{\kappa}(u) - \boldsymbol{\kappa}(u - h_n v)] [\boldsymbol{\kappa}(u) - \boldsymbol{\kappa}(u - h_n v)]^\top \right) \cdot \\ &\quad m_2(u, u - h_n v) f_T(u) f_T(u - h_n v) du dv \\ &= -\frac{1}{h_n} \int K(v)^2 \text{tr} \left(\boldsymbol{\kappa}(\xi) \boldsymbol{\kappa}(\xi)^\top \right) h_n^2 v^2 m_2(u, u - h_n v) \cdot \\ &\quad f_T(u) f_T(u - h_n v) du dv, \end{aligned}$$

donde $\xi \in (u - h_n v, u)$ es un punto intermedio. Como $G_i(t)$ es acotada, se tiene que $m_2(T_1, T_2)$ es acotada. Luego por N7, N9 y por el hecho de que \mathcal{T} es compacto, $\text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) = \mathcal{O}(h_n)$.

Para calcular $\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13})$ (2.22) es de utilidad. En primer lugar, es fácil ver que $E(\mathbf{V}_{12}) = \mathcal{O}(h_n)$ y $E(\mathbf{V}_{13}) = \mathcal{O}(h_n)$ usando los mismos argumentos que para el cálculo de la esperanza de $\mathbf{B}_n^{(8)}$. Además, definiendo $m_3(T_1, T_2, T_3) = E_0(G_1(T_2)G_1(T_3)|T_1, T_2, T_3)$ y haciendo un cambio de variables

$$E(\mathbf{V}_{12}, \mathbf{V}_{13}^\top) = \int K(v)K(z)\boldsymbol{\kappa}'(\xi_1)\boldsymbol{\kappa}'(\xi_2)^\top h_n^2 v z m_3(u, u - h_n v, u - h_n z) \cdot f_T(u)f_T(u - h_n v)f_T(u - h_n z) dudvdz$$

donde $\xi_1 \in (u - h_n v, u)$ y $\xi_2 \in (u - h_n z, u)$, lo cual implica que $\text{tr}(E(\mathbf{V}_{12}, \mathbf{V}_{13}^\top)) = \mathcal{O}(h_n^2)$. Consecuentemente,

$$\text{tr}(\text{Cov}(\mathbf{B}_n^{(8)}, \mathbf{B}_n^{(8)})) = \frac{n(n-1)}{n^3} \mathcal{O}(h_n) + \frac{n(n-1)(n-2)}{n^3} \mathcal{O}(h_n^2) = \mathcal{O}(h_n n^{-1} + h_n^2).$$

Convergencia $\mathbf{B}_n^{(7)} \xrightarrow{P} 0$:

Separando los casos $j \neq i$ y $j = i$

$$\begin{aligned} \mathbf{B}_n^{(7)} &= \frac{1}{\sqrt{n}} \frac{1}{h_n n} \left(\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K\left(\frac{T_i - T_j}{h_n}\right) [R_0(Y_j, \mathbf{X}_j, T_j) - \boldsymbol{\kappa}(T_j)] G_i(T_j) \right. \\ &\quad \left. + \sum_{i=1}^n K(0) [R_0(Y_i, \mathbf{X}_i, T_i) - \boldsymbol{\kappa}(T_i)] G_i(T_i) \right). \end{aligned}$$

Se consideran las esperanzas de ambos términos:

$$\begin{aligned} E_0 \left(\frac{1}{\sqrt{n}} \frac{1}{h_n n} \sum_{i=1}^n K(0) (R_0(Y_i, \mathbf{X}_i, T_i) - \boldsymbol{\kappa}(T_i)) G_i(T_i) \right) \\ &= \frac{1}{\sqrt{n}} \frac{1}{h_n n} n K(0) E_0 [(R_0(Y_1, \mathbf{X}_1, T_1) - \boldsymbol{\kappa}(T_1)) G_1(T_1)] \\ &= \frac{K(0)}{h_n \sqrt{n}} E_0 [(R_0(Y_1, \mathbf{X}_1, T_1) - \boldsymbol{\kappa}(T_1)) G_1(T_1)] = \mathcal{O}(h_n^{-1} n^{-1/2}) \end{aligned}$$

Para $i \neq j$, como $\boldsymbol{\kappa}(T) = E_0(R_0(Y, \mathbf{X}, T)|T)$, se tiene

$$\begin{aligned} \frac{n(n-1)}{n\sqrt{n}} E_0 \left(\frac{1}{h_n} K\left(\frac{T_1 - T_2}{h_n}\right) [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2)] G_1(T_2) \right) \\ &= \frac{n-1}{\sqrt{n}} E_0 \left(\frac{1}{h_n} K\left(\frac{T_1 - T_2}{h_n}\right) E_0 [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2) | (Y_1, \mathbf{X}_1, T_1, T_2)] G_1(T_2) \right) \\ &= \frac{n-1}{\sqrt{n}} E_0 \left(\frac{1}{h_n} K\left(\frac{T_1 - T_2}{h_n}\right) [E_0(R_0(Y_2, \mathbf{X}_2, T_2)|T_2) - \boldsymbol{\kappa}(T_2)] G_1(T_2) \right) = 0. \end{aligned} \tag{A.2}$$

Para analizar la covarianza, se define

$$\mathbf{V}_{ij} = \frac{1}{h_n} K \left(\frac{T_i - T_j}{h_n} \right) [R_0(Y_j, \mathbf{X}_j, T_j) - \boldsymbol{\kappa}(T_j)] G_i(T_j)$$

luego

$$\begin{aligned} \text{Cov}(\mathbf{B}_n^{(7)}, \mathbf{B}_n^{(7)}) &= \frac{1}{n^3} \left(n \text{Cov}(\mathbf{V}_{11}, \mathbf{V}_{11}) + n(n-1) \text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12}) \right. \\ &\quad \left. + n(n-1)(n-2) [\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13}) + \text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{32})] \right). \end{aligned}$$

Analicemos la traza de cada uno de los términos por separado,

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{V}_{11}, \mathbf{V}_{11})) &\leq E_0(\|\mathbf{V}_{11}\|^2) \\ &= \frac{1}{h_n^2} K(0)^2 E_0 \left(G_1^2(T_1) \|R_0(Y_1, \mathbf{X}_1, T_1) - \boldsymbol{\kappa}(T_1)\|^2 \right) = \mathcal{O}(h_n^{-2}) \end{aligned}$$

ya que G_1 , R_0 y $\boldsymbol{\kappa}$ son funciones acotadas por N2, N3 y N7.

Por el mismo razonamiento realizado en (A.2) tenemos que $E_0(\mathbf{V}_{ij}) = 0$ para $i \neq j$, luego

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) &= \text{tr}(E_0(\mathbf{V}_{12} \mathbf{V}_{12}^\top)) \\ &= \text{tr} \left(E_0 \left[\frac{1}{h_n^2} K^2 \left(\frac{T_1 - T_2}{h_n} \right) [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2)] \cdot \right. \right. \\ &\quad \left. \left. [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2)]^\top G_1(T_2)^2 \right] \right) \\ &= \int \frac{1}{h_n^2} K(v)^2 \text{tr}([R_0(Y_2, \mathbf{X}_2, u - h_n v) - \boldsymbol{\kappa}(u - h_n v)] \cdot \\ &\quad [R_0(Y_2, \mathbf{X}_2, u - h_n v) - \boldsymbol{\kappa}(u - h_n v)]^\top \cdot \\ &\quad G_1(u - h_n v)^2 f_T(u) f_T(u - h_n v) (-h_n) dudv). \end{aligned}$$

Por N2, N3 y N7 se tiene que $\text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{12})) = \mathcal{O}(h_n^{-1})$.

Por otro lado,

$$\begin{aligned} \text{tr}(\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{13})) &= \text{tr}(E_0(\mathbf{V}_{12} \mathbf{V}_{13}^\top)) \\ &= \text{tr} \left(E_0 \left[\frac{1}{h_n^2} K \left(\frac{T_1 - T_2}{h_n} \right) K \left(\frac{T_1 - T_3}{h_n} \right) \cdot \right. \right. \\ &\quad \left. \left. [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2)] [R_0(Y_3, \mathbf{X}_3, T_3) - \boldsymbol{\kappa}(T_3)]^\top G_1(T_2) G_1(T_3) \right] \right) \\ &= \text{tr} \left(E_0 \left[\frac{1}{h_n^2} K \left(\frac{T_1 - T_2}{h_n} \right) K \left(\frac{T_1 - T_3}{h_n} \right) [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2)] \cdot \right. \right. \\ &\quad \left. \left. E_0 [R_0(Y_3, \mathbf{X}_3, T_3) - \boldsymbol{\kappa}(T_3) | (Y_1, \mathbf{X}_1, T_1, Y_2, \mathbf{X}_2, T_2, T_3)]^\top G_1(T_2) G_1(T_3) \right] \right) \\ &= \text{tr} \left(E_0 \left[\frac{1}{h_n^2} K \left(\frac{T_1 - T_2}{h_n} \right) K \left(\frac{T_1 - T_3}{h_n} \right) [R_0(Y_2, \mathbf{X}_2, T_2) - \boldsymbol{\kappa}(T_2)] \cdot \right. \right. \\ &\quad \left. \left. [E_0(R_0(Y_3, \mathbf{X}_3, T_3) | T_3) - \boldsymbol{\kappa}(T_3)]^\top G_1(T_2) G_1(T_3) \right] \right) = 0 \end{aligned}$$

Análogamente se puede mostrar que $\text{Cov}(\mathbf{V}_{12}, \mathbf{V}_{32}) = 0$.

Finalmente se tiene que

$$\text{tr}(\text{Cov}(\mathbf{B}_n^{(7)}, \mathbf{B}_n^{(7)})) = O((nh)^{-1}),$$

concluyendo así la demostración.

Convergencia $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n} \xrightarrow{P} \boldsymbol{\Sigma}_{\beta_0}$:

La expresión de $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n}$ es

$$\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n} = \widehat{\mathbf{A}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n)^{-1} \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n) \widehat{\mathbf{A}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n)^{-1}.$$

Por Lema 2.5.2 $\widehat{\mathbf{A}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n) \xrightarrow{P} \mathbf{A}$, luego para concluir que $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}_n} \xrightarrow{P} \boldsymbol{\Sigma}_{\beta_0} = \mathbf{A}^{-1} \boldsymbol{\Sigma} \mathbf{A}^{-1}$, es suficiente ver que $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_n, \widehat{m}_n) \xrightarrow{P} \boldsymbol{\Sigma}$.

Para ello, en primer lugar notar que

$$\widehat{\mathbf{M}}(t, m, \boldsymbol{\beta}) \xrightarrow{P} \mathbf{M}(t). \quad (\text{A.3})$$

En efecto, un desarrollo de Taylor de primer orden permite escribir $\widehat{\mathbf{M}}(t, \widehat{m}_n, \widehat{\boldsymbol{\beta}}_n) - \mathbf{M}(t) = \mathbf{M}^{(1)} + \mathbf{M}^{(2)} + \mathbf{M}^{(3)}$ con

$$\begin{aligned} \mathbf{M}^{(1)} &= \sum_{i=1}^n W_{i,n}(t) \chi(Y_i, \mathbf{X}_i^\top \widehat{\boldsymbol{\beta}}_n + m_0(T_i)) \omega_1(\mathbf{X}_i) (1, \mathbf{X}_i^\top)^\top (1, \mathbf{X}_i^\top) \\ &\quad - E_0 \left(\chi(Y, \mathbf{X}^\top \widehat{\boldsymbol{\beta}}_n + m_0(T)) \omega_1(\mathbf{X}) (1, \mathbf{X}^\top)^\top (1, \mathbf{X}^\top) | T = t \right), \\ \mathbf{M}^{(2)} &= \sum_{i=1}^n W_{i,n}(t) \chi_1(Y_i, \mathbf{X}_i^\top \boldsymbol{\beta} + \xi_i) \omega_1(\mathbf{X}_i) (1, \mathbf{X}_i^\top)^\top (1, \mathbf{X}_i^\top) (\widehat{m}_n(T_i) - m_0(T_i)), \\ \mathbf{M}^{(3)} &= E_0 \left(\chi(Y, \mathbf{X}^\top \widehat{\boldsymbol{\beta}}_n + m_0(T)) \omega_1(\mathbf{X}) (1, \mathbf{X}^\top)^\top (1, \mathbf{X}^\top) | T = t \right) - \mathbf{M}(t). \end{aligned}$$

Por el Teorema A.3.2 se puede concluir que $\mathbf{M}^{(1)} \xrightarrow{P} 0$. Asimismo $\mathbf{M}^{(2)} \xrightarrow{P} 0$ por N2, N6 y el hecho de que $\|\widehat{m}_n - m_0\|_\infty \xrightarrow{P} 0$. Finalmente, por la consistencia de $\widehat{\boldsymbol{\beta}}_n$ y la continuidad de $E_0 \left(\chi(Y, \mathbf{X}^\top \boldsymbol{\beta} + m_0(T)) \omega_1(\mathbf{X}) (1, \mathbf{X}^\top)^\top (1, \mathbf{X}^\top) | T = t \right)$ como función de $\boldsymbol{\beta}$ se tiene $\mathbf{M}^{(3)} \xrightarrow{P} 0$.

Un razonamiento análogo al anterior y la consistencia del estimador de núcleos de la densidad \widehat{f}_T permiten mostrar que

$$\widehat{\boldsymbol{\kappa}}(t, m, \boldsymbol{\beta}) \xrightarrow{P} \boldsymbol{\kappa}(t). \quad (\text{A.4})$$

Finalmente, es fácil ver que $\widehat{\Sigma}(\widehat{m}_n, \widehat{\beta}_n) - \Sigma = \mathbf{S}^{(1)} + \mathbf{S}^{(2)} + \mathbf{S}^{(3)}$ con

$$\begin{aligned}\mathbf{S}^{(1)} &= \widehat{\Sigma}(\widehat{m}_n, \widehat{\beta}_n) - \frac{1}{n} \sum_{i=1}^n (1 - E_{\theta}(W|Y))^2 \Psi^2 \left(Y_i, \mathbf{X}_i^{\top} \widehat{\beta}_n + \widehat{m}_n(T_i) \right) \mathbf{u}_i \mathbf{u}_i^{\top} \\ \mathbf{S}^{(2)} &= \frac{1}{n} \sum_{i=1}^n (1 - E_{\theta}(W|Y))^2 \Psi^2 \left(Y_i, \mathbf{X}_i^{\top} \widehat{\beta}_n + \widehat{m}_n(T_i) \right) \mathbf{u}_i \mathbf{u}_i^{\top} \\ &\quad - E_0 \left((1 - E_{\theta}(W|Y))^2 \Psi^2(Y, \mathbf{X}^{\top} \widehat{\beta}_n + \widehat{m}_n(T)) \mathbf{u} \mathbf{u}^{\top} \right) \\ \mathbf{S}^{(3)} &= E_0 \left((1 - E_{\theta}(W|Y))^2 \Psi^2(Y, \mathbf{X}^{\top} \widehat{\beta}_n + \widehat{m}_n(T)) \mathbf{u} \mathbf{u}^{\top} \right) - \Sigma\end{aligned}$$

donde $\mathbf{u}_i = \omega_1(\mathbf{X}_i) \mathbf{X}_i + D(\mathbf{X}_i, T_i) \boldsymbol{\gamma}(T_i) f_T(T_i)$.

Las convergencias (A.3) y (A.4), el hecho de que $|1 - E_{\theta}(W|Y)| < 1$ y la condición N2 aseguran que $\mathbf{S}^{(1)} \xrightarrow{P} 0$. El Teorema A.3.1 implica que $\mathbf{S}^{(2)} \xrightarrow{P} 0$ y la continuidad de $E_0 \left((1 - E_{\theta}(W|Y))^2 \Psi^2(Y, \mathbf{X}^{\top} \boldsymbol{\beta} + m(T)) \mathbf{u} \mathbf{u}^{\top} \right)$ como función de $(\boldsymbol{\beta}, m)$ junto con la consistencia de los estimadores aseguran que $\mathbf{S}^{(3)} \rightarrow 0$. Así $\widehat{\Sigma}(\widehat{m}_n, \widehat{\beta}_n) - \Sigma \xrightarrow{P} 0$.

A.5. Procedimiento de estimación del modelo ZIP doblemente semiparamétrico

El procedimiento de estimación desarrollado en el Capítulo 2 de esta tesis puede extenderse naturalmente al marco doblemente semiparamétrico dado por

$$\begin{aligned}\log(\lambda) &= \mathbf{X}^{\top} \boldsymbol{\beta} + m_{\lambda}(T) \\ \text{logit}(\pi) &= \mathbf{Z}^{\top} \boldsymbol{\gamma} + m_{\pi}(T).\end{aligned}$$

En este contexto, la función de enlace logística incorpora una componente no paramétrica, lo que flexibiliza los supuestos lineales y permite una mayor flexibilidad de modelado.

La implementación de esta extensión solo requiere adaptar el paso 1 del procedimiento de estimación de la misma manera que para la estimación de $\boldsymbol{\beta}$ y m . Específicamente, se define un nuevo funcional empírico

$$Q_{n,4}(\eta, \boldsymbol{\gamma}, t|\boldsymbol{\theta}) = \sum_{i=1}^n W_{i,n}(t) \left(\log \left(1 + e^{\mathbf{Z}_i^{\top} \boldsymbol{\gamma} + \eta} \right) - E_{\theta}(W_i|Y_i)(\mathbf{Z}_i^{\top} \boldsymbol{\gamma} + \eta) \right) \omega_2(\mathbf{Z}_i)$$

que permite estimar simultáneamente $\boldsymbol{\gamma}$ y m_{π} en el Paso 1. Además se redefine $Q_{n,3}$ para incluir la función m_{π} como

$$Q_{n,3}(m_{\pi}, \boldsymbol{\gamma}|\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \left(\log \left(1 + e^{\mathbf{Z}_i^{\top} \boldsymbol{\gamma} + m_{\pi}(T_i)} \right) - E_{\theta}(W_i|Y_i)(\mathbf{Z}_i^{\top} \boldsymbol{\gamma} + m_{\pi}(T_i)) \right) \omega_2(\mathbf{Z}_i).$$

Así, el algoritmo de tipo EM que permite la estimación del modelo ZIP doblemente semiparamétrico consiste en los siguientes pasos:

(Inicialización) Obtener un estimador inicial $\hat{\boldsymbol{\theta}}_n^{(0)} = (\hat{\boldsymbol{\beta}}_n^{(0)}, \hat{\boldsymbol{\gamma}}_n^{(0)}, \hat{m}_{n,\lambda}^{(0)}, \hat{m}_{n,\pi}^{(0)})$.

(Paso E) En la r -ésima iteración estimar W_i mediante

$$W_i^{(r)} = E_{\hat{\boldsymbol{\theta}}_n^{(r)}}(W_i|Y_i) = \begin{cases} \frac{1}{1 + \exp\{-(\mathbf{Z}_i^\top \hat{\boldsymbol{\gamma}}_n^{(r)} + \hat{m}_{n,\pi}^{(r)}(T_i)) - e^{\mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_n^{(r)} + \hat{m}_{n,\lambda}^{(r)}(T_i)}\}} & \text{si } Y_i = 0, \\ 0 & \text{si } Y_i > 0. \end{cases}$$

donde la esperanza se toma bajo el estimador actual del parámetro $\hat{\boldsymbol{\theta}}_n^{(r)}$.

(Paso M) Realizar el procedimiento de tres pasos:

(Paso 1)

Para cada valor fijo $T = t$, computar

$$(\tilde{\boldsymbol{\beta}}_n(t), \tilde{m}_{n,\lambda}(t)) = \underset{\boldsymbol{\beta}, \eta}{\operatorname{argmin}} Q_{n,1}(\eta, \boldsymbol{\beta}, t | \hat{\boldsymbol{\theta}}_n^{(r)}),$$

donde se denota $\eta = m_\lambda(t)$.

Computar además

$$(\tilde{\boldsymbol{\gamma}}_n(t), \tilde{m}_{n,\pi}(t)) = \underset{\boldsymbol{\gamma}, \eta}{\operatorname{argmin}} Q_{n,4}(\eta, \boldsymbol{\gamma}, t | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

Aquí se denota $\eta = m_\pi(t)$.

(Paso 2)

Calcular el estimador de $\boldsymbol{\beta}$ usando $\tilde{m}_{n,\lambda}$ obtenido en el Paso 1,

$$\hat{\boldsymbol{\beta}}_n^{(r+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} Q_{n,2}(\tilde{m}_{n,\lambda}, \boldsymbol{\beta} | \hat{\boldsymbol{\theta}}_n^{(r)})$$

y obtener el estimador de $\boldsymbol{\gamma}$ usando $\tilde{m}_{n,\pi}$ obtenido en el Paso 1 como

$$\hat{\boldsymbol{\gamma}}_n^{(r+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} Q_{n,3}(\tilde{m}_{n,\pi}, \boldsymbol{\gamma} | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

(Paso 3)

Usando $\hat{\boldsymbol{\beta}}_n^{(r+1)}$ del Paso 2, actualizar el estimador de m_λ como

$$\hat{m}_{n,\lambda}(t) = \underset{\eta}{\operatorname{argmin}} Q_{n,1}(\eta, \hat{\boldsymbol{\beta}}_n^{(r+1)}, t | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

Actualizar también el estimador de m_π usando $\hat{\boldsymbol{\gamma}}_n^{(r+1)}$ del Paso 2, como sigue

$$\hat{m}_{n,\pi}(t) = \underset{\eta}{\operatorname{argmin}} Q_{n,4}(\eta, \hat{\boldsymbol{\gamma}}_n^{(r+1)}, t | \hat{\boldsymbol{\theta}}_n^{(r)}).$$

Esta modificación resulta en una herramienta más versátil, aplicable a una mayor variedad de problemas en escenarios de análisis de datos complejos.

Conclusiones

El objetivo central de esta tesis fue desarrollar un procedimiento de estimación robusto para el modelo de regresión de Poisson parcialmente lineal inflado en ceros (PLZIP), con el fin de superar las limitaciones de los enfoques tradicionales en presencia de valores atípicos y leves apartamientos de los supuestos del modelo.

Una de las contribuciones de esta tesis fue desarrollar un sólido marco teórico para la metodología propuesta. Se demostró la consistencia casi segura de los estimadores del modelo, tanto de los parámetros lineales como de la componente no paramétrica. En particular, para esta última se demostró la consistencia uniforme. Se mostró además que los parámetros lineales asociados a la media de Poisson tienen normalidad asintótica y se encontró una expresión para la matriz de covarianza, lo que permitió luego desarrollar un resultado de test de hipótesis robusto para dichos parámetros.

Además del desarrollo teórico, esta tesis también aportó un enfoque práctico. A través de exhaustivos estudios de simulación se mostró la robustez y flexibilidad de los estimadores propuestos frente a diferentes escenarios de contaminación, en comparación con otros enfoques de estimación para modelos inflados en ceros presentes en la literatura. Se evidenció que los estimadores robustos propuestos superan ampliamente a las metodologías convencionales en presencia de valores atípicos, ofreciendo estimaciones más precisas y confiables. Además, bajo escenarios sin contaminación presentan un comportamiento similar a los estimadores clásicos basados en verosimilitud.

Los métodos desarrollados se destacaron también por su flexibilidad, robustez y aplicabilidad en escenarios reales y complejos. Esto representa un avance significativo para el análisis de datos de conteo inflados en ceros, especialmente en áreas como la salud, la economía y la biología, donde la presencia de ceros excesivos es común.

En resumen, los estimadores propuestos no solo ofrecen una mayor robustez y precisión, sino que también aportan un marco teórico sólido que puede ser la base para futuras investigaciones y aplicaciones.

Bibliografía

- M. Anggreainy, A. Ilyasu, H. Musyaffa, and F. Kansil. Analysis of factors influencing the covid-19 mortality rate in indonesia using zero inflated negative binomial model. *International Journal of Advanced Computer Science and Applications*, 13(4), 2022. doi: 10.14569/IJACSA.2022.0130485.
- S. Balakrishnan, M. Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1), 2017. doi: 10.1214/16-AOS1435.
- A. Bergesio and V. Yohai. Projection estimators for generalized linear models. *Journal of the American Statistical Association*, 106(494):661–671, 2011. doi: 10.1198/jasa.2011.tm09774.
- A. Bianco and V. Yohai. *Robust Estimation in the Logistic Regression Model*. Springer, New York, 1996. doi: 10.1007/978-1-4612-2380-1_2.
- A. Bianco, G. Boente, and D. Rodriguez. Resistant estimators in poisson and gamma models with missing responses and an application to outlier detection. *Journal of Multivariate Analysis*, 114:209–226, 2012. doi: 10.1016/j.jmva.2012.08.008.
- G. Boente and D. Rodriguez. Robust inference in generalized partially linear models. *Computational Statistics and Data Analysis*, 54(12):2942–2966, 2010. doi: 10.1016/j.csda.2010.05.025.
- G. Boente and D. Rodriguez. Robust estimates in generalized partially linear single-index models. *Test*, 21(2):386–411, 2012. doi: 10.1007/s11749-011-0249-z.
- G. Boente, X. He, and J. Zhou. Robust estimates in generalized partially linear models. *Annals of Statistics*, 34(6):2856–2878, 2006. doi: 10.1214/009053606000000858.

- G. Boente, R. Cao, W. González Manteiga, and D. Rodríguez. Testing in generalized partially linear models: A robust approach. *Statistics & Probability Letters*, 83(1):203–212, 2013. doi: 10.1016/j.spl.2012.08.031.
- G. Boente, D. Rodríguez, and P. Vena. Robust estimators in a generalized partly linear regression model under monotony constraints. *TEST*, 29(1), 2018. doi: 10.1007/s11749-019-00629-7.
- A. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 72(2):353–360, 1984. doi: 10.1093/biomet/71.2.353.
- E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030, 2001. doi: 10.1198/016214501753209004.
- R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997. doi: 10.1080/01621459.1997.10474001.
- A. Cohen. Estimation in mixtures of two normal distributions. *Technometrics*, 9(1):15–28, 1967. doi: 10.1080/00401706.1967.10490438.
- J. Cragg. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844, 1971. doi: 10.2307/1909582.
- C. Croux and G. Haesbroeck. Implementing the bianco and yohai estimator for logistic regression. *Computational Statistics & Data Analysis*, 44(1):273–295, 2002. doi: 10.1016/S0167-9473(03)00042-2.
- X. Cui, W. Härdle, and L. Zhu. The efm approach for single-index models. *Annals of Statistics*, 39(3):1658–1688, 2011. doi: 10.1214/10-AOS871.
- M. L. Dalrymple, I. L. Hudson, and R. P. K. Ford. Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Comput. Stat. Data Anal.*, 41(3–4):491–504, 2003. doi: 10.1016/S0167-9473(02)00187-1.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.
- R. Engle, C. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):310–320, 1986. doi: 10.2307/2289218.

- W. Greene. *Some Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*. Department of Economics, New York University, New York, 1994.
- D. Hall and J. Shen. Robust estimation for zero-inflated poisson regression. *Scandinavian Journal of Statistics*, 37(2):237–252, 2010. doi: 10.1111/j.1467-9469.2009.00657.x.
- D. B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, 2000. doi: 10.2307/2677034.
- X. He, H. Xue, and N. Z. Shi. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated poisson models. *Journal of multivariate analysis*, 101(9):2026–2038, 2010. doi: 10.1016/j.jmva.2010.05.003.
- P. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.
- P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–223, 1967.
- N. Jansakul and J. P. Hinde. Score tests for zero-inflated poisson models. *Computational Statistics & Data Analysis*, 40(1):75–96, 2002. doi: 10.1016/S0167-9473(01)00104-9.
- N. Johnson and S. Kotz. *Distributions in Statistics—Discrete Distributions*. Wiley, New York, 1969. doi: 10.1002/0471715816.
- K. F. Lam, H. Xue, and Y. B. Cheung. Semiparametric analysis of zero-inflated count data. *Biometrics*, 62(4):996–1003, 2006. doi: 10.1111/j.1541-0420.2006.00575.x.
- D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992. doi: 10.2307/1269547.
- M. Liaqat, S. Kamal, F. Fischer, and N. Zia. Zero-inflated and hurdle models with an application to the number of involved axillary lymph nodes in primary breast cancer. *Journal of King Saud University - Science*, 34(4), 2022. doi: 10.1016/j.jksus.2022.101932.
- Z. Lu, Y. V. Hui, and A. H. Lee. Minimum hellinger distance estimation for finite mixtures of poisson regression models and its applications. *Biometrics*, 59(4):1016–1026, 2003. doi: 10.1111/j.0006-341x.2003.00117.x.
- P. McCullagh and J. Nelder. *Generalized Linear Models. 2nd Edition*. Chapman and Hall, London, 1989. doi: 10.1007/978-1-4899-3242-6.
- G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2007.

- N. Muler and V. Yohai. Robust estimates for arch processes. *Journal of Time Series Analysis*, 23(3):341–375, 2002. doi: 10.1111/1467-9892.00268.
- J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986. doi: 10.1016/0304-4076(86)90002-3.
- E. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964. doi: 10.1137/1109020.
- E. Nadaraya. On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1):186–190, 1965. doi: 10.1137/1110024.
- J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. doi: 10.2307/2344614.
- D. Nolan and D. Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15 (2):780–799, 1987. doi: 10.1214/aos/1176350374.
- D. Pollard. *Convergence of stochastic processes*. Springer, New York ; Berlin ; Tokyo, 1984.
- D. Pregibon. Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, 38(2):485–498, 1982. doi: 10.2307/2530463.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL <http://www.R-project.org/>.
- C. E. Rose, S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, 16(4):463–481, 2006. doi: 10.1080/10543400600719384.
- O. Rosen, W. Jiang, and M. A. Tanner. Mixtures of marginal models. *Biometrika*, 87(2): 391–404, 2000. doi: 10.1093/biomet/87.2.391.
- A. Roy and S. Hasanuzzaman. Role of energy stacking in minimizing respiratory health risks in bangladeshi households: Evidence from zero-inflated poisson regression model. *World Medical & Health Policy*, 2025. doi: 10.1002/wmh3.70033.
- M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78, 1982.
- M. Salibian-Barrera, S. Van Aelst, and V. Yohai. Robust tests for linear regression models based on τ -estimates. *Computational Statistics & Data Analysis*, 93:436–455, 2016. doi: 10.1016/j.csda.2014.09.012.
- L. Schumaker. *Spline Functions: Basic Theory*. John Wiley and Son, New York, 1981.

- T. Severini and J. Staniswalis. Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89(426):501–511, 1994. doi: 10.1080/01621459.1994.10476774.
- T. A. Severini and W. H. Wong. Profile likelihood and conditionally parametric models. *The Annals of Statistics*, 20(4):1768–1802, 1992. doi: 10.1214/aos/1176348889.
- J. Staniswalis. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283, 1989. doi: 10.1080/01621459.1989.10478766.
- T. Tiruye, B. Higgs, M. O’Callaghan, L. FitzGerald, D. Roder, and K. Beckmann. Socioeconomic and geographic disparities in anorectal and urinary procedures following radiotherapy for prostate cancer. *Cancer Medicine*, 14(15):e71135, 2025. doi: 10.1002/cam4.71135.
- V. Valdora and V. Yohai. Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146:31–48, 2014. doi: 10.1016/j.jspi.2013.09.016.
- J. Van den Broek. A score test for zero inflation in a poisson distribution. *Biometrics*, 51(2):738–743, 1995.
- A. Van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996. doi: 10.1007/978-1-4757-2545-2.
- Q. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.
- F. Wang and D. Scott. The l_1 method for robust nonparametric regression. *Journal of the American Statistical Association*, 89:65–76, 1994. doi: 10.1080/01621459.1994.10476446.
- X. Wang, J. Zhang, L. Yu, and G. Yin. Generalized partially linear single-index model for zero-inflated count data. *Statistics in medicine*, 34(4):876–886, 2015. doi: 10.1002/sim.6382.
- G. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.
- P. Wilson. The misuse of the vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127(C):51–53, 2015. doi: 10.1016/j.econlet.2014.12.029.
- P. Wilson and J. Einbeck. A new and intuitive test for zero modification. *Statistical Modelling*, 19(4):341–361, 2018. doi: 10.1177/1471082X18762277.
- C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983. doi: 10.1214/aos/1176346060.

- Q. Zhou, Y. Zhang, X. Wang, R. Yang, X. Zhu, Y. Zhang, C. Chen, H. Yuan, Z. Yang, and L. Sun. Gut bacteria *akkermansia* is associated with reduced risk of obesity: evidence from the american gut project. *Nutrition & metabolism*, 17(1):90, 2020. doi: 10.1186/s12986-020-00516-1.