Métodos de Imputación basados en la Función de Verosimilitud

Subdepartamento de Investigación Estadística

Departamento de Metodologías e Innovación Estadística

Instituto Nacional de Estadísticas

Miguel Alvarado

Noviembre, 2023

Contents

1 Introducción

A partir de esta sección se inicia la presentación de los métodos de imputación basados en modelos. Estos métodos, como veremos en las siguientes dos secciones, se basan en dos conceptos fundamentales: la *función de verosimilitud* de los datos[1] y la *ignorabilidad del mecanismo* que genera la *falta de datos*[2]. La presentación formal de ambos conceptos, así como sus implicancias, se desarrollan en las siguientes secciones; sin embargo, para los propósitos de esta introducción es necesario hacer referencia, al menos en parte, sobre el segundo concepto. En tal sentido, si bien es importante avanzar en la presentación del marco teórico y, en particular, sobre los fundamentos inferenciales de los métodos de imputación basados en modelos, puesto que en estos radica la validez estadística de su implementación; validez que es cuestionada en casi todos los *métodos tradicionales*[3] que fueron presentados en la primera parte de este documento. En efecto, los métodos tradicionales entregan soluciones que, no solo son poco satisfactorias, además, algunos de estos métodos son cuestionables en sí mismos y, por tanto, su aplicación y sus resultados; puesto que tales métodos podrían resultar potencialmente problemáticos debido a que pueden introducir sesgos, independientemente del tipo de mecanismo (Enders 2022, pg.24).

Antes del trabajo de Donal B. Rubin (Rubin 1976), los análisis estadísticos con datos faltantes eran realizados a partir de suponer, implícita o explícitamente, que el *mecanismo* que genera la falta de datos podía ser *ignorado*. Sin embargo, hasta ese entonces, la literatura estadística que estudia esta problemática no había respondido a una pregunta anterior: *¿cuándo es apropiado ignorar el mecanismo que genera los datos faltantes?* (Rubin 1976, pg.581). En este sentido, los métodos tradicionales simplemente asumen que dicho mecanismo puede ser ignorado, al suponer que la falta de datos ocurre de manera *completamente aleatoria*, pero sin discutir sobre la validez de dicho supuesto. De manera más precisa, estos métodos asumen un tipo particular de mecanismo, uno donde la *falta de datos* es *completamente aleatoria* (*Missing Completly At Random*, MCAR)[4]; sin embargo, como se menciona a lo largo de la literatura, tal supuesto resulta sumamente restrictivo (Enders 2022, pg.24) y, a menudo, *poco realista* (Van Buuren 2012, pg.7). Entonces, dado lo poco plausible del supuesto MCAR, los métodos tradicionales y las inferencias que se desprenden de su aplicación, quedan en serio cuestionamiento.

Como se describirá en los siguientes párrafos, los métodos tradicionales dependen de supuestos poco verosímiles y, además, muchos de estos métodos simplemente carecen de algún tipo de fundamento estadístico. Por otro lado, aun cuando la aplicación práctica de los métodos tradicionales es sencilla, algunos de estos métodos pueden dificultar e incluso imposibilitar el cálculo de algunas estimaciones. Por último, en algunos de estos métodos se podrían requerir de la toma de decisiones que quedan al arbitrio de quienes implementan tales métodos. A continuación, de manera sucinta, se discute sobre las limitaciones e inconvenientes que presentan los principales métodos tradicionales.

[1] Dentro de este documento, cuando se utiliza el término "*datos*", nos referimos al *conjunto de datos* o *muestra de datos*, que conforman una *base de datos*. En tal sentido, en este documento se consideran únicamente conjuntos de datos rectangulares; esto es, datos dispuestos como arreglos rectangulares que, en general, pueden representarse a través de *matrices* de datos, donde las filas corresponden a *elementos* (*casos*, *unidades* u *observaciones*, según el contexto), mientras que las columnas corresponden a *variables* que son investigadas en cada uno de los *elementos* que conforman el conjunto de datos. De este modo, las *entradas*, *celdas* o *elementos* de una matriz de datos, que corresponden a números reales, representan los *valores* que fueron informados y/o investigados en cada uno de los elementos, en relación a variables continuas (i.e., ingresos), discretas (i.e., edad); y/o categóricas, que pueden ser ordinales (i.e., nivel de educación) o nominales (i.e., sexo).

[2] Los términos "*falta de datos*" y "*datos faltantes*" son usados de manera indistinta a lo largo de este documento, cuando el *valor* correspondiente en *algunas* variables y para *algunos* elementos del conjunto de datos, es *no observado*.

[3] En algunos textos como (Enders 2022, sec. 1.7), estos métodos se denominan como *métodos antiguos*.

[4] En la Sección (2.2.1) de este documento, se encuentra una presentación formal este concepto.

El método de *análisis de casos completos*, también conocido como *eliminación por lista*, es la forma más simple de lidiar con la falta de datos. En este método se eliminan todos los casos con uno o más datos faltantes en las variables que conforman la muestra de datos. Si el mecanismo es MCAR, el método produce estimaciones insesgadas para las medias, las varianzas y los coeficientes de regresión; no obstante, los errores estándar y niveles de significancia *solo* son correctos para el conjunto reducido de casos completos, pero que a menudo son mayores en relación con todos los datos observados. Una clara desventaja de este método es que potencialmente se puede llegar a eliminar una parte considerable de los casos, especialmente si el número de variables con datos faltantes es grande (Van Buuren 2012, pg.8). En efecto, como se señala en (Little and Rubin 2020, pg.47), las desventajas que se derivan de la posible pérdida de información al descartar casos incompletos tiene dos aspectos: *pérdida de precisión* y *sesgo*, cuando el mecanismo no es del tipo MCAR. El grado de sesgo y pérdida de precisión dependen no solo de la fracción de casos completos y del mecanismo de datos faltantes, sino también de la medida en que las unidades completas e incompletas difieren y de las estimaciones de interés (Little and Rubin 2020, pg.48).

El método de *eliminación por pares*, también denominado *análisis de casos disponibles*, intenta remediar el problema de la pérdida de casos que se produce en el método anterior. En este método, el cálculo de cualquier estimación de interés para alguna variable, es realizado a partir de los casos disponibles en dicha variable. De este modo, las estimaciones de la variable V, se realizan a partir de los casos disponibles en la variable V; las estimaciones de la variable W se realizan a partir de los casos disponibles en la variable W y, análogamente, con el resto de las variables. El método es simple, puesto que usa toda la información disponible y produce estimaciones consistentes para las medias, correlaciones y covarianzas bajo el supuesto MCAR (Van Buuren 2012, pag.10). Sin embargo, cuando estas estimaciones se toman en conjunto, aparecen inconvenientes considerables. En principio, las estimaciones pueden estar sesgadas si el mecanismo no es del tipo MCAR (Van Buuren 2012, pg.10). Por otro lado, existen problemas al momento del cálculo computacional; por ejemplo, la matriz de correlación puede no ser definida positiva, lo cual es un requisito para la mayoría de los procedimientos multivariantes. De igual modo, pueden ocurrir correlaciones que no están en el rango unitario $[-1, +1]$, un problema que proviene de utilizar diferentes subconjuntos de datos para el cálculo de las varianzas y covarianzas. Otro problema es que no queda claro qué tamaño de muestra debe usarse para calcular los errores estándar (Van Buuren 2012, pg.10).

El método de *imputación por el promedio* o *imputación por la media*, es un enfoque de *única* imputación que completa los datos faltantes en alguna variable *continua* con el promedio[5] de los datos observados en la variable. Este método no tiene justificación teórica y distorsiona las estimaciones de parámetros, independiente del *mecanismo* que genera la falta de datos (Enders 2022, pg.25), puesto que este método distorsiona la distribución de los datos de varias maneras (Van Buuren 2012, pg.11). El método es una solución rápida y sencilla para abordar la falta de datos. Sin embargo, este método subestima la varianza, altera las relaciones entre las variables, sesga casi cualquier estimación que no sea la media, pero incluso puede sesgar dicha estimación cuando el mecanismo no es MCAR. Por tanto, el uso de este método debe evitarse en general[6] (Van Buuren 2012, pg.11).

El método de *imputación por regresión*, con el propósito de mejorar la imputación de los datos

[5] En el caso de variables categóricas, la imputación de los datos faltantes es realizada usando la *moda* de los datos observados (Van Buuren 2012, pg.10). El mismo criterio podría ocuparse en el caso de variables numéricas discretas.

[6] Este método de imputación genera sesgos y su uso no suele ser recomendado; no obstante, un refinamiento de este método es imputar a partir del uso de promedios condicionales, dados los valores observados (Little and Rubin 2020, pg.70). Mayor detalle sobre este enfoque se puede encontrar en (Little and Rubin 2020, sec. 4.2.2).

faltantes en la variable de interés, incorpora la información contenida en las otras variables que forman parte del conjunto de datos. El método parte por ajustar un modelo de regresión a partir de los datos observados. Luego, el valor no observado en los datos es reemplazado por los *valores ajustados* (o, *valores estimados*) bajo el modelo ajustado. De este modo, los valores imputados corresponden a los valores más *verosímiles* bajo el modelo ajustado (Van Buuren 2012, pg.12). Sin embargo, al igual que en el método de imputación por la media, el conjunto de valores imputados presenta menor variabilidad que en los valores observados[7]. Si bien es posible que cada uno de los valores individuales imputados sean la mejor "estimación" bajo el modelo, resulta poco probable que los valores reales (pero no observados) de la variable imputada tengan tal distribución. La imputación de los datos faltantes a partir de este método también tiene un efecto sobre la correlación. Dado que la correlación de los datos imputados bajo el modelo ajustado es igual a 1 (Enders 2022, pg.27), la correlación para el conjunto de los datos completos se ve necesariamente incrementada, en consecuencia, las varianzas y correlaciones estimadas quedan sesgadas.

Bajo el supuesto que el mecanismo es del tipo MCAR, la imputación por regresión produce estimaciones insesgadas tanto para las medias (igual que el método de imputación por la media), como para los ponderadores del modelo de regresión ajustado para realizar la imputación de los datos faltantes, esto último si las variables explicativas en el modelo están completas. Por otro lado, como se ha mencionado, la variabilidad de los datos imputados queda subestimada de manera sistemática y el grado de subestimación depende de la varianza explicada y de la proporción de datos faltantes (Van Buuren 2012, pg.12). La idea básica detrás de este método de imputación es intuitivamente atractiva: las variables tienden a estar correlacionadas, por lo que los valores faltantes se *reemplazan* por *estimaciones* que vienen de un modelo que toma prestada información importante de los datos observados. Aunque esta idea tiene sentido, como se ha mencionado, las imputaciones resultantes pueden introducir sesgos, cuya naturaleza y magnitud dependen del mecanismo de datos faltantes y varían según las diferentes estimaciones (Enders 2022, pg.27).

El método de *imputación por regresión estocástica* es un refinamiento del método de imputación por regresión, en el cual se agrega *variabilidad* a las predicciones del modelo ajustado (Van Buuren 2012, pg.13). De este modo, este método también ajusta un modelo de regresión a partir de los datos observados, luego el valor no observado en los datos es reemplazado por los *valores ajustados* bajo el modelo ajustado, pero tomando el paso adicional de *agregar* a cada estimación un término de *ruido aleatorio* (*random noise*) desde una distribución normal. Al agregar estos residuos a los valores ajustado, se reduce la correlación (Van Buuren 2012, pg.13), se restaura la pérdida de variabilidad de los datos y se eliminan los sesgos asociados al método de imputación por regresión (Enders 2022, pg.28).

El método de imputación por regresión estocástica no resuelve todos los problemas y hay muchas sutilezas que deben tenerse presentes[8]. No obstante, el método de imputación por regresión estocástica es el *único* método tradicional que, generalmente, es capaz de producir estimaciones insesgadas de los parámetros de interés cuando la *falta de datos* es *aleatoria* (*Missing At Random*,

[7] La imputación por el promedio se puede considerar como un caso especial del método de imputación por regresión donde las variables explicativas (predictores) son variables indicadoras (*dummies*) para las celdas dentro de las cuales se imputa por el promedio (Little and Rubin 2020, pg.68).

[8] Por ejemplo, al añadir un ruido aleatorio a los valores ajustados bajo el modelo ajustado es posible que para valores localizados en los extremos de la distribución, el valor a imputar quede fuera del rango factible de la variable a imputar. Un ejemplo de esto puede encontrarse en (Van Buuren 2012, pg.13), en cuyo ejemplo, una parte de las imputaciones son valores negativos en circunstancias que la variable a imputar solo puede tomar valores mayores o iguales a cero.

MAR)[9]. Más importante aún, la idea central detrás del método de imputación por regresión estocástica (una imputación es igual a un valor ajustado (o estimado) más un ruido aleatorio) constituye la base de técnicas de imputación más avanzadas y, como se verá más adelante, resurge con los métodos bayesianos e imputación múltiple (Enders 2022, pg.29).

El método de *adelantar la última observación* (*Last Observation Carried Forward*, LOCF) es una técnica de datos faltantes para estudios longitudinales. Utilizar el método LOCF en estudios sociales y del comportamiento es bastante poco frecuente, siendo su uso más común en estudios médicos y ensayos clínicos. Como el nombre del método lo indica, la idea es tomar el último valor observado y *adelantarlo* (*trasladarlo*) en reemplazo de los datos faltantes de la actual muestra de datos. El método LOCF es conveniente en el sentido que genera un conjunto de datos completo; sin embargo, este método asume que no existen cambios desde la última observación realizada y/o durante el período de tiempo en que se genera la nueva medición. La creencia popular indicaría que imputar los datos faltantes con datos *estables* en el tiempo, produciría una estimación conservadora de las diferencias entre los grupos bajo estudio. Sin embargo, la investigación empírica muestra que esto no es necesariamente cierto, ya que el método también puede exagerar las diferencias entre estos grupos. En efecto, la dirección y la magnitud del sesgo que se produce dependen de las características específicas de los datos, pero es probable que el método LOCF produzca estimaciones sesgadas de los parámetros de interés, incluso asumiendo que el mecanismo es del tipo MCAR (Enders 2022, pg.31).

El método de imputación *Hot-Deck* imputa los valores faltantes utilizando los valores observados en casos "*similares*" en el conjunto de datos, estos últimos usualmente denominados como "*donantes*"[10]. Este método es común en la práctica de las encuestas y puede implicar esquemas muy elaborados para seleccionar los casos donantes[11]. La ventaja del método Hot-Deck es que, a diferencia del método de imputación por la media, la distribución de los valores de la variable a imputar no queda distorsionada por las imputaciones; sin embargo, el incremento en la varianza que produce el método Hot-Deck puede ser no despreciable. Aun cuando se pueden lograr reducciones en la varianza adicional que se produce con el método Hot-Deck, por ejemplo mediante una selección más eficiente del esquema de muestreo, poniendo restricciones en el número de veces que un caso actúa como donante, usando los valores observados en la variable para formar estratos de muestreo para donantes o mediante el uso de un *Hot-Deck Secuencial*; los *métodos de imputación múltiple* se deben preferir por sobre este método, puesto que los métodos de imputación múltiple no solo que pueden reducir el incremento de la varianza del muestreo a niveles insignificantes, sino que también proporcionan errores estándar válidos que tienen en cuenta la incertidumbre del proceso de imputación. Las estimaciones que se derivan del uso del método Hot-Deck son insesgadas solo bajo el supuesto que el mecanismo es del tipo MCAR; supuesto que, generalmente, es poco realista (Little and Rubin 2020, pg.78).

El método de imputación *Cold-Deck* imputa los valores faltantes de una variable por un valor constante que proviene de una fuente externa; por ejemplo, a partir de los datos de una encuesta anterior. La aplicación práctica de este método suele tratar los datos resultantes como una muestra completa, ignorando las consecuencias de la imputación. Una teoría satisfactoria para el análisis de datos obtenidos mediante el método de imputación Cold-Deck es inexistente (Little and Rubin 2020,

[9] En la Sección (2.2.2) de este documento, se encuentra una presentación formal este concepto.

[10] En este método, la imputación de los valores faltantes de un caso es realizada con los valores observados en algún otro caso *similar* al que se busca imputar. Sin embargo, cuando existen dos o más casos *similares*, pero con valores observados diferentes en las variables a imputar, la decisión sobre cuál caso tomar como *donante*, queda al arbitrio de quien realiza la imputación.

[11] En (Little and Rubin 2020, sec. 4.3.2) se puede encontrar mayor detalle sobre variantes del método Hot-Deck.

pg.69; Van Buuren 2012, pg.7).

A manera de síntesis, se puede señalar que una limitación importante de los *métodos tradicionales* de imputación descritos en esta introducción es que los estimadores de la varianza de muestreo que son aplicados a los datos *completados* mediante estos métodos de imputación, al no tener en cuenta la incertidumbre asociada al proceso de imputación, a excepción del método de imputación por regresión estocástica, finalmente subestiman sistemáticamente la verdadera varianza de muestreo de las estimaciones. Por tanto, los errores estándar calculados a partir de los datos *completados* también se subestiman sistemáticamente, lo que implica que los *p-values* de las pruebas sean demasiado significativos y los intervalos de confianza sean demasiado estrechos (Little and Rubin 2020, pg.81). Lo anterior ocurre incluso si el modelo utilizado para generar las imputaciones es el correcto, algo que, salvo para el caso antes mencionado, depende de asumir que el mecanismo que genera la falta de datos es del tipo MCAR, supuesto que, como ya se ha mencionado, generalmente es poco realista (Little and Rubin 2020, pg.78).

Dado que los métodos tradicionales presentan limitaciones importantes que resultan insalvables y dado que estos métodos dependen de supuestos inverosímiles, lo que a continuación sigue en este documento es la presentación del marco teórico y los robustos fundamentos inferenciales en los que se basan los métodos de imputación basados en modelos; los cuales no presentan las limitaciones de los métodos tradicionales, ni dependen de supuestos inverosímiles.

2 Mecanismo de Datos Faltantes

Como se ha mencionado, hasta antes del trabajo de Rubin (Rubin 1976), los análisis estadísticos con datos faltantes eran realizados a partir de suponer, implícita o explícitamente, que el *mecanismo* que genera los datos faltantes podía ser *ignorado*, pero sin dar respuesta a la importante pregunta sobre: *¿cuándo es apropiado ignorar el mecanismo que genera los datos faltantes?* (Rubin 1976, pg.581). Rubin, sin embargo, logró establecer las *condiciones necesarias* (*weakest conditions*) sobre el mecanismo que genera los datos faltantes, tal que *siempre* es apropiado *ignorar* dicho mecanismo, al momento de realizar inferencia sobre la distribución de los datos (Rubin 1976, pg.582). Esto, dentro la literatura de datos faltantes, se denomina *ignorabilidad del mecanismo*.

En esta sección, se introducen los conceptos de *mecanismo de datos faltantes*, *tipos de mecanismo* y, a partir de estos, se señalan las condiciones necesarias que dan paso al importante concepto de *ignorabilidad del mecanismo*.

2.1 Patrón y Mecanismo de Datos Faltantes

Dentro de la literatura de datos faltantes, los conceptos de *patrón* y *mecanismo de datos faltantes* suelen prestarse a confusión. El *patrón de datos faltantes* se refiere a la configuración o disposición de los datos *observados* y los *no observados* (*missing*), dentro de un conjunto de datos. En tanto, el *mecanismo de datos faltantes* describe las posibles relaciones entre los datos y la *propensión* que estos tienen de ser o no observados. De un modo más simple; mientras que el patrón de datos faltantes describe *dónde* están los "*missing*" (las celdas vacías) en los datos, el mecanismo de datos faltantes describe *cómo* se generan los "*missing*" en los datos (Enders 2022, pg.2). Respecto al *patrón de datos faltantes*, dentro de la literatura existe concenso en cuanto a reconocer seis tipos de patrones, los que se distinguen según la configuración que emerge de la localización de los datos faltantes dentro del conjunto de datos[12]. En relación al *mecanismo de datos faltantes*, dentro de la literatura se reconocen tres tipos de mecanismo, que dependen según si la falta de datos está relacionada con los valores subyacentes de las variables del conjunto de datos[13].

Distinguir de manera clara entre lo que conceptualmente representra, por un lado, el *patrón* y, por otro, el *mecanismo* de datos faltantes, así como reconocer que *tipos de patrones* de datos faltantes están presenten dentro del conjunto de datos; son consideraciones importantes que se deben tener presente para alcanzar una adecuada comprensión del contenido de esta y la siguiente sección; no obstante, comprender las implicancias que tienen los diferentes tipos de mecanismo, resulta simplemente crucial. Diferenciar entre entre lo que representan el patrón y el mecanismo de datos faltantes, puede no ser del todo simple, más cuando ambos suelen representarse a través de un mismo objeto matemático y es solo el contexto de su uso lo que permite diferenciarlos; por tanto, siempre que sea necesario se debe regresar al inicio de esta sección. En tanto, reconocer el tipo de patrón de datos faltantes, permite concentrarse en algunos pocos métodos que están diseñados para un particular tipo de patrón; por tanto, se suguiere revisar (Little and Rubin 2020, sec. 1.2) si se considera prudente ahondar en un caso particular. En relación a los tipos de mecanismo y las distintas implicancias que cada uno de estos tiene, como veremos más adelante; la validez de los métodos de imputación basados en modelos depende, en gran medida, de la naturaleza de las

[12] Para un mayor detalle sobre los diferentes tipos de patrones de datos faltantes, se puede consultar (Little and Rubin 2020, sec. 1.2) y (Enders 2022, sec. 1.2).

[13] En las Secciones (2.2.1), (2.2.2) y (2.2.3) de este documento, se describen estos tres tipos de mecanismo de datos faltantes. Para un mayor detalle sobre estos tipos de mecanismo de datos faltantes, se puede consultar (Little and Rubin 2020, sec. 1.3), (Enders 2022, sec. 1.3), (Van Buuren 2012, sec. 1.3) y (Schafer 1997, sec. 2.2).

dependencias al interior del *mecanismo* (Little and Rubin 2020, pg.13); por tanto, el énfasis de esta sección esta puesto en este último punto.

Supongamos un conjunto de datos que está conformado por *valores* que corresponden a la información registrada de n elementos, respecto a p variables de interés. Conviene recordar que, en términos prácticos, los conjunto de datos que son de nuestro interés, corresponden a bases de datos rectangulares; es decir, conjuntos de datos dispuestos por filas y columnas; por tanto, lo natural es representar un conjunto de datos mediante un arreglo rectangular por filas y columnas; es decir, mediante una matriz de datos. Con el propósito de formalizar los conceptos mencionados al inicio de esta sección y, en particular, las implicancias que tienen los diferentes tipos de mecanismo, consideremos la siguiente notación[14]:

Sea $\mathbf{Y}$ una matriz de dimensión $n \times p$, cuyos elementos se pueden denotar por y_{ij}, con $i = 1, \ldots, n$, y $j = 1, \ldots, p$. Luego, el elemento en la posición (i, j) de la matriz $\mathbf{Y}$; esto es, el elemento y_{ij}, corresponde al *valor* registrado en el elemento i y la variable j. De este modo, el conjunto de datos queda representado por la matriz $\mathbf{Y}$[15]. Si suponemos que se cuenta con un conjunto de datos *completo* o, dicho de modo más simple, un conjunto de datos sin datos faltantes; en la matriz $\mathbf{Y}$, el valor correspondiente para y_{ij} es *observado* en *todos* los elementos del conjunto de datos, $\mathbf{Y}$. Si por el contrario, *no* se cuenta con un conjunto de datos completo; es decir, el conjunto de datos tiene datos faltantes; en la matriz $\mathbf{Y}$, el valor correspondiente para y_{ij} es *observado* en *algunos* elementos del conjunto de datos, $\mathbf{Y}$[16].

Se ha señalado que el *patrón* de datos faltantes se refiere a la disposición (localización) de los datos *observados* y *no observados* dentro del conjunto de datos. En tal sentido, para describir la localización de los datos *observados* dentro del conjunto de datos, $\mathbf{Y}$; se define la *matriz indicadora de respuesta*, que suele denotarse por $\mathbf{R}$. Sea $\mathbf{R}$ una matriz *indicadora* de dimensión $n \times p$, cuyos elementos se pueden denotar por r_{ij}, con $i = 1, \ldots, n$, y $j = 1, \ldots, p$; donde:

$$r_{ij} = \begin{cases} 0 & , \quad y_{ij} \text{ es no observado} \\ 1 & , \quad y_{ij} \text{ es observado} \end{cases} \tag{1}$$

[14] En este documento, salvo que de manera explícita se señale lo contrario, se sigue la convención de denotar las variables aleatorias con letras mayúsculas y los valores observados, que son la realización de variables aleatorias, con las correspondientes letras minúsculas; por ejemplo, $Z_1, \ldots, Z_n$ denotan una secuencia de variables aleatorias, mientras que la realización de las variables aleatorias de la secuencia, se denotan por $z_1, \ldots, z_n$, respectivamente. En tanto, se utilizan letras griegas para denotar parámetros y cuando a estos se superpone el símbolo $\widehat{\ }$, estos denotan un estimador y una estimación del parámetro; por ejemplo, si algún parámetro es denotado por γ, entonces $\widehat{\gamma}$ denota un estimador y una estimación de γ. Por otro lado, los vectores y matrices, ya sean aleatorios o no, se denotan con letras minúsculas y mayúsculas en negrita, respectivamente; por ejemplo, los vectores:

$$\begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}$$

representan, un vector de observaciones y de variables aleatorias, respectivamente; los que, por convención, se denotan por $\mathbf{z}$; de igual modo, $\boldsymbol{\gamma}$ denota un vector de parámetros; mientras que $\mathbf{X}$, por convención, denota una matriz de observaciones o de variables aleatorias. Finalmente, el superíndice $^\intercal$ se utiliza para denotar la transposición de vectores y matrices; por ejemplo, un vector columna (como los antes descritos) se escriben como vectores fila al denotar $\mathbf{z} = [z_1, \ldots, z_n]^\intercal$ y $\mathbf{z} = [Z_1, \ldots, Z_n]^\intercal$, respectivamente. La notación adoptada en esta sección, se encuentra en (Van Buuren 2012, pg.30), (Schafer 1997, secs. 2.1–2.2) y (He, Zhang, and Hsu 2021, pg.7) y, con algunas diferencias, en (Enders 2022, pg.4–5) y (Little and Rubin 2020, pg.8–9.).

[15] En la Sección (3) de este documento, se aborda con mayor detalle esta misma descripción.

[16] En tal sentido, dentro de la literatura se suele decir que se tiene un conjunto de datos *incompleto* (*incomplete data*).

$\forall\,(i,j)$. De este modo, la matriz indicadora de respuesta $\mathbf{R}$, describe la localización de los datos *observados* (y la de los datos *no observados*) dentro del conjunto de datos, $\mathbf{Y}$.

Por otro lado, también se ha señalado que el *mecanismo* de datos faltantes describe las posibles relaciones entre los datos y la *propensión* que estos tienen de ser o no observados. En tal sentido, siguiendo el trabajo de Rubin; si consideramos $\mathbf{R}$ como una variable aleatoria, entonces se le puede asignar una distribución de probabilidades (Little and Rubin 2020, pg.13). Sea $P(\cdot)$, el proceso aleatorio que gobierna la probabilidad de *observar el valor de los datos* y, en tal caso, $P(\cdot)$ se denomina *mecanismo de respuesta* (Van Buuren 2012, pg.6). No obstante, dada la dualidad de la variable aleatoria, $P(\cdot)$ también permite representar el proceso aleatorio que gobierna la probabilidad de *no observar el valor de los datos*; por tanto, $P(\cdot)$ puede también denominarse *mecanismo de falta de respuesta* o, lo que es lo mismo, *mecanismo de datos faltantes*[17]. De este modo, de un modo general, el *mecanismo* puede ser formulado como un modelo estadístico para $\mathbf{R}$, dado el conjunto de datos, $\mathbf{Y}$[18]. Por tanto, sin pérdida de generalidad, el *mecanismo de falta de respuesta* puede ser caracterizado mediante:

$$P(\mathbf{R} \mid \mathbf{Y}, \boldsymbol{\psi}) \tag{2}$$

donde, $P(\cdot)$ denota la distribución condicional de $\mathbf{R}$ dado $\mathbf{Y}$, y $\boldsymbol{\psi}$ denota un vector de parámetros desconocidos del modelo formulado para $\mathbf{R}$ (Little and Rubin 2020, pg.13).

Finalmente, si en el conjunto de datos $\mathbf{Y}$, denotamos los datos *observados* y los *no observados* por $\mathbf{Y}_{obs}$ e $\mathbf{Y}_{mis}$, respectivamente; los *datos completos* se pueden escribir como: $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. De este modo, se tiene un modelo que establece la relación entre $\mathbf{R}$, que es completamente observado, e $\mathbf{Y}$, donde una parte es *observada*, $\mathbf{Y}_{obs}$; y otra es *no observada*, $\mathbf{Y}_{mis}$. De este modo, la distribución de probabilidades descrita en (2), se puede escribir como:

$$P(\mathbf{R} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}) \tag{3}$$

2.2 Tipos de Mecanismo

Tomando el trabajo de Rubin (Rubin 1976), Little y Rubin (Little and Rubin 2020, sec. 1.3) introdujeron un sistema de clasificación para el mecanismo de datos faltantes que es virtualmente universal en la literatura. Este trabajo describe tres tipos de *mecanismos* o *procesos aleatorios* que describen diferentes maneras en que la probabilidad de los datos faltantes se relaciona con los datos: *Falta de Datos Completamente Aleatoria* (*Missing Completly At Random*, MCAR), *Falta de Datos Aleatoria* (*Missing At Random*, MAR) y *Falta de Datos No Aleatoria* (*Missing Not At Random*, MNAR). Desde una perspectiva práctica, estos diferentes tipos de mecanismo son de vital importancia, puesto que funcionan como supuestos estadísticos en el análisis de datos faltantes (Enders 2022, pg.3–4); lo que hace importante un análisis formal de cada uno de estos.

[17] Como se menciona en (Little and Rubin 2020, pg.9), alternativamente, se puede definir la *matriz indicadora de falta de respuesta*, denotada por $\mathbf{M}$, la cual describe la localización de los datos *no observados* dentro del conjunto de datos $\mathbf{Y}$. Sea $\mathbf{M}$ una matriz *indicadora* de dimensión $n \times p$, cuyos elementos se pueden denotar por m_{ij}, con $i = 1, \ldots, n$, y $j = 1, \ldots, p$; donde:

$$m_{ij} = \begin{cases} 0 & , \quad y_{ij} \text{ es observado} \\ 1 & , \quad y_{ij} \text{ es no observado} \end{cases}$$

$\forall\,(i,j)$. En este caso, al proceso aleatorio se suele denominar *mecanismo de falta de respuesta* o, lo que es lo mismo, *mecanismo de datos faltantes*. El uso de $\mathbf{M}$ se encuentra, entre otros textos, en (Little and Rubin 2020, pg.9) y (Enders 2022, pg.5); en tanto, el uso de $\mathbf{R}$, y que sigue este documento, se encuentra en (Van Buuren 2012, pg.30), (Schafer 1997, sec. 2.2) y (He, Zhang, and Hsu 2021, pg.7), entre otros.

[18] Otro concepto que suele mencionarse dentro la literatura es el de *modelo de respuesta* o *modelo de falta de datos* y se refiere al modelo particular del *mecanismo* (Van Buuren 2012, pg.6).

2.2.1 Falta de Datos Completamente Aleatoria (MCAR)

Si la probabilidad de no observar el valor de un dato; es decir, ser un dato faltante ($\mathbf{R} = 0$), es la misma para todas las observaciones, se dice que la *falta de datos es completamente aleatoria*, esto es, el mecanismo es del tipo MCAR (Van Buuren 2012, pg.7). Entonces, el mecanismo del tipo MCAR establece que la probabilidad de ser un dato faltante *no* está relacionada con los *datos* (i.e., ni con los datos observados, como tampoco con los no observados) (Enders 2022, pg.6). Considerando la definición formal que involucra la distribución condicional de $\mathbf{R}$ dado $\mathbf{Y}$; la distribución para un mecanismo MCAR (He, Zhang, and Hsu 2021, pg.13), viene dada por[19]:

$$P\left(\mathbf{R} = 0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}\right) \quad = \quad P\left(\mathbf{R} = 0 \mid \boldsymbol{\psi}\right) \tag{4}$$

Esto es, la probabilidad de los datos faltantes *no* está relacionada con los datos $\mathbf{Y}$ y solo depende de los parámetros $\boldsymbol{\psi}$. En palabras simples, el lado derecho de la ecuación (4) dice que todos los elementos tienen la misma probabilidad de ser un dato faltante, dados los parámetros $\boldsymbol{\psi}$ (Enders 2022, pg.6). Una consecuencia muy importante de un proceso de este tipo es que se pueden ignorar muchas de las complejidades que surgen debido a la falta de datos, fuera de la pérdida obvia de información. No obstante, como ya se ha mencionado, aun cuando esta situación resulta sumamente conveniente, el mecanismo MCAR es una situación poco realista (Little and Rubin 2020, pg.78; Van Buuren 2012, pg.7).

2.2.2 Falta de Datos Aleatoria (MAR)

Si la probabilidad de no observar el valor de un dato; es decir, ser un dato faltante, es la misma solo dentro de grupos definidos por los datos *observados*, se dice que la *falta de datos es aleatoria*, esto es, el mecanismo es del tipo MAR (Van Buuren 2012, pg.7). Entonces, el mecanismo del tipo MAR establece que la probabilidad de ser un dato faltante está relacionada con los *datos observados*, pero *no* con los *datos no observados* (Enders 2022, pg.8). Considerando la definición formal que involucra la distribución condicional de $\mathbf{R}$ dado $\mathbf{Y}$; la distribución para un mecanismo MAR (He, Zhang, and Hsu 2021, pg.13), viene dada por[20]:

$$P\left(\mathbf{R} = 0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}\right) \quad = \quad P\left(\mathbf{R} = 0 \mid \mathbf{Y}_{obs}, \boldsymbol{\psi}\right) \tag{5}$$

Esto es, la probabilidad de los datos faltantes está relacionada *solo* con la parte observada de los datos $\mathbf{Y}_{obs}$ y los parámetros $\boldsymbol{\psi}$. En palabras simples, el lado derecho de la ecuación (5) dice que los valores que se hubieran observado en $\mathbf{Y}_{mis}$ no contiene información adicional sobre los datos faltantes, distinta a la aportada por los datos observados $\mathbf{Y}_{obs}$ (Enders 2022, pg.8). Este mecanismo es más general que el primero y resulta un supuesto más realista que suponer un mecanismo MCAR. Como veremos, los métodos modernos de imputación, generalmente, suponen que la falta de datos es generada por un mecanismo del tipo MAR.

[19] Si se utiliza la *matriz indicadora de falta de respuesta* $\mathbf{M}$, equivalentemente, la distribución para un mecanismo MCAR (Enders 2022, pg.6), viene dada por:

$$P\left(\mathbf{M} = 1 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}\right) \quad = \quad P\left(\mathbf{M} = 1 \mid \boldsymbol{\psi}\right)$$

[20] Si se utiliza la *matriz indicadora de falta de respuesta* $\mathbf{M}$, equivalentemente, la distribución para un mecanismo MAR (Enders 2022, pg.8), viene dada por:

$$P\left(\mathbf{M} = 1 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}\right) \quad = \quad P\left(\mathbf{M} = 1 \mid \mathbf{Y}_{obs}, \boldsymbol{\psi}\right)$$

2.2.3 Falta de Datos No Aleatoria (MNAR)

Si la probabilidad de no observar el valor de un dato; es decir, ser un dato faltante, *no* es la misma para todas las observaciones, se dice que la *falta de datos es no aleatoria*, esto es, el mecanismo es del tipo MNAR. Enronces, el mecanismo del tipo MNAR establece que la probabilidad de ser un dato faltante está relacionada con los *datos observados* y, también, con los *datos no observados* (Enders 2022, pg.11). Considerando la definición formal que involucra la distribución condicional de $\mathbf{R}$ dado $\mathbf{Y}$; la distribución para un mecanismo MNAR (Van Buuren 2012, pg.31), viene dada por[21]:

$$P\left(\mathbf{R} = 0 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}\right) \tag{6}$$

A diferencia de los mecanismos MCAR y MAR; en el caso del mecanismo MNAR, la distribución condicional de $\mathbf{R}$ dado $\mathbf{Y}$ no se simplifica.

2.3 Ignorabilidad del Mecanismo

Hasta ahora poco se ha dicho sobre el conjunto de parámetros del modelo formulado para $\mathbf{R}$, esto es, $\boldsymbol{\psi}$. La razón es bastante simple, tales parámetros no tienen algún valor en sí mismos y, generalmente, son además desconocidos. En tal sentido, el análisis de los datos faltantes se simplificaría si, dichos parámetros, simplemente se pudieran *ignorar*. En este sentido, la importancia práctica de haber realizado una distinción clara entre los diferentes tipos de mecanismo y, más importante aún, sobre lo que cada uno implica, como veremos más adelante dentro de este documento, clarifica las condiciones bajo las cuales es posible estimar los parámetros de nuestro modelo estadístico, sin la necesidad de conocer el conjunto de parámetros $\boldsymbol{\psi}$.

La última parte del anterior párrafo, en buena medida, resume el propósito de la siguiente sección y, en cierto modo, destaca la relevancia práctica que se espera sea desarrollada dentro de este documento. Por mientras, es suficiente comentar que en el trabajo desarrollado por Rubin (Rubin 1976), se presentan dos modelos: el modelo que es el foco del análisis y un modelo que describe el mecanismo de datos faltantes. Sin pérdida de generalidad, supongamos que estos modelos tienen conjuntos de parámetros denotador por $\boldsymbol{\theta}$ y $\boldsymbol{\psi}$, respectivamente. Los parámetros en $\boldsymbol{\psi}$ son esencialmente una *molestia*, porque no están relacionados con los objetivos que motivaron la investigación de las unidades que conforman el conjunto de datos, $\mathbf{Y}$. Entonces, cabe preguntarse: *¿en qué situaciones podemos simplemente estimar $\boldsymbol{\theta}$ a partir de los datos observados, sin preocuparnos de estimar el modelo para los datos faltantes o los parámetros en $\boldsymbol{\psi}$?*. Esta es la esencia del concepto de *ignorabilidad del mecanismo*; es decir, regresamos a la importante pregunta planteada por Rubin: *¿cuándo es apropiado ignorar el mecanismo que genera los datos faltantes?* (Rubin 1976, pg.581).

El trabajo de Rubin logró establecer las *condiciones necesarias* (*weakest conditions*) sobre el *mecanismo* que genera los datos faltantes, tal que *siempre* es apropiado *ignorar* el mecanismo al momento de realizar inferencia sobre la distribución de los datos (Rubin 1976, pg.582). De este modo, se dice que el *mecanismo es ignorable*, si:

i. El mecanismo que genera los datos faltantes es del tipo MAR, y

ii. Los parámetros $\boldsymbol{\psi}$ no contienen información sobre los parámetros de interés $\boldsymbol{\theta}$; es decir, $\boldsymbol{\psi}$ y $\boldsymbol{\theta}$ son parámetros *distintos*.

[21] Si se utiliza la *matriz indicadora de falta de respuesta*, $\mathbf{M}$, equivalentemente, la distribución para un mecanismo MNAR (Enders 2022, pg.11), viene dada por:

$$P\left(\mathbf{M} = 1 \mid \mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \boldsymbol{\psi}\right)$$

Como se verá en las secciones siguientes, el concepto de *ignorabilidad del mecanismo* tiene implicancias muy importantes en cuanto a la aplicación de los métodos de imputación basados en modelos. En este sentido, las condiciones que dan paso al concepto de *ignorabilidad del mecanismo*, son igual de importantes.

3 Métodos Basados en la Función de Verosimilitud

Los métodos estadísticos modernos para el análisis e imputación de datos faltantes, bajo ciertos supuestos, se basan en la *función de verosimilitud*. En este sentido, el propósito fundamental de esta sección es presentar la teoría básica de la inferencia basada en la función de verosimilitud. Específicamente, se revisan el método de *Estimación por Máxima Verosimilitud* y su aplicación en la estimación e inferencia de los *Modelos Lineales Generalizados*. Presentado el marco teórico, se describe como estos pueden ser implementados, bajo el supuesto de *ignorabilidad del mecanismo*, cuando el conjunto de datos tiene datos faltantes. Finalmente, se presenta el *Algoritmo de Esperanza-Maximización*.

Antes de comenzar con la presentación de la teoría básica de la inferencia basada en la función de verosimilitud, conviene presentar algunos conceptos que, además de ser mencionados de manera recurrente, resultan fundamentales para el adecuado entendimiento de los métodos de imputación modernos.

3.1 Definiciones Generales

Sea $\mathbf{Y}$ una matriz de dimensión $n \times p$, donde cada fila puede ser modelada como la realización desde alguna distribución de probabilidad *multivariada*. De este modo, el conjunto de datos (representado por $\mathbf{Y}$) es la realización de una secuencia de n vectores aleatorios, que pueden ser denotados por $\mathbf{y}_1^\intercal, \ldots, \mathbf{y}_n^\intercal$. Por tanto, la matriz $\mathbf{Y}$ puede denotarse como $\mathbf{Y} = [\mathbf{y}_1^\intercal, \ldots, \mathbf{y}_n^\intercal]^\intercal$; donde $\mathbf{y}_i^\intercal$, con $i = 1, \ldots, n$; es un vector aleatorio de dimensión $1 \times p$; que puede denotarse por $\mathbf{y}_i^\intercal = [Y_{i1}, \ldots, Y_{ip}]$; donde Y_{ij}, con $j = 1, \ldots, p$; es una variable aleatoria. Luego, la secuencia de vectores aleatorios, según como se describa su soporte, pueden representar variables aleatorias discretas o continuas.

Sin pérdida de generalidad, con el fin de simplificar la presentación teórica, supongamos que el conjunto de datos corresponde a la información registrada en n elementos, para una variable de interés. En tal caso, el conjunto de datos puede representarse por medio de un vector columna de n filas[22]. Sea $\mathbf{y}$ un vector de dimensión $n \times 1$, donde cada fila puede ser modelada como la realización desde alguna distribución de probabilidad *univariada*[23]. De este modo, el conjunto de datos (representados por $\mathbf{y}$) es la realización de una secuencia de n variables aleatorias, denotadas por $Y_1, \ldots, Y_n$; por cuanto, el vector $\mathbf{y}$ puede denotarse como $\mathbf{y} = [Y_1, \ldots, Y_n]^\intercal$. En tanto, en términos prácticos, el conjunto de datos corresponde a la *realización* de la secuencia de variables aleatorias $Y_1, \ldots, Y_n$; las que pueden denotarse, respectivamente, por $y_1, \ldots, y_n$. De este modo, la secuencia de realizaciones, por convención, es denotado por $\mathbf{y} = [y_1, \ldots, y_n]^\intercal$.

Notemos que $\mathbf{y}$ denota, al mismo tiempo, un vector de variables aleatorias y un vector de realizaciones; esto es, $\mathbf{y} = [Y_1, \ldots, Y_n]^\intercal$ e $\mathbf{y} = [y_1, \ldots, y_n]^\intercal$, respectivamente. Si bien esto no es un inconveniente, puesto que el contexto en que es utilizado el vector $\mathbf{y}$, permite diferenciar entre uno y otro[24]. No obstante, dado que la notación que corresponde al caso univariado podría generar algún tipo de confunsión y, dado que las definiciones, conceptos y métodos que se presentan en esta sección, pueden extenderse al caso multivariado; se adopta la siguiente *excepción*: El conjunto de datos, que se asume puede ser modelado como la realización de una secuencia de variables aleatorias, denotadas por $Y_1, \ldots, Y_n$; *excepcionalmente* será denotado por $\mathbf{Y} = [Y_1, \ldots, Y_n]^\intercal$, donde $\mathbf{Y}$ es una *matriz* de

[22] Recordemos que, por convensión, los vectores (aleatorios o no) se denotan mediante letras minúsculas en negrita.

[23] No obstante, las definiciones, conceptos y métodos que se presentan en esta sección, pueden extenderse al caso multivariado de manera simple y natural.

[24] Esto asumiendo que el contexto en que es utilizado $\mathbf{y}$ es el apropiado y, además, que la interpretación del uso dado en todo contexto, es también la apropiada.

dimensión $n \times 1$ (i.e. un vector columna de dimensión $n \times 1$). De este modo, $\mathbf{Y}$ denota una matriz (columna) de variables aleatorias, mientras que $\mathbf{y}$, siguiendo la convención, denota el vector de realizaciones de la secuencia de variables aleatorias, $\mathbf{Y}$; esto es, $\mathbf{y} = [y_1, \ldots, y_n]^\intercal$. La excepción antes descrita, se mantiene para el resto de esta sección y las siguientes secciones.

Definición 3.1 (Función de probabilidad conjunta) *Sea $Y_1, \ldots, Y_n$, una secuencia de variables aleatorias discretas; entonces, la función de probabilidad conjunta de $Y_1, \ldots, Y_n$ es la función $f(y_1, \ldots, y_n) : \mathbb{R}^n \longrightarrow [0, 1]$, tal que:*

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n) &= f(y_1, \ldots, y_n) \\ &= f(\mathbf{y}) \end{aligned} \tag{7}$$

$\forall \mathbf{y} : \ \mathbf{y} = [y_1, \ldots, y_n]^\intercal \in \mathbb{R}^n$.

De modo más general, si A es un evento cualesquiera tal que $A \subset \mathbb{R}^n$, entonces la *función de probabilidad conjunta* $f(\mathbf{y})$, viene dada por:

$$P(\mathbf{Y} \in A) = \sum_{\mathbf{y} \in A} f(\mathbf{y}) \tag{8}$$

Definición 3.2 (Función de densidad conjunta) *Sea $Y_1, \ldots, Y_n$, una secuencia de variables aleatorias continuas; entonces, la función de densidad conjunta de $Y_1, \ldots, Y_n$ es la función $f(y_1, \ldots, y_n) : \mathbb{R}^n \longrightarrow [0, \infty)$, tal que:*

$$\begin{aligned} P(Y_1 \leq y_1, \ldots, Y_n \leq y_n) &= \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f(u_1, \ldots, u_n)\, du_1 \cdots du_n \\ &= f(y_1, \ldots, y_n) \\ &= f(\mathbf{y}) \end{aligned} \tag{9}$$

$\forall \mathbf{y} : \ \mathbf{y} = [y_1, \ldots, y_n]^\intercal \in \mathbb{R}^n$.

De modo más general, si A es un evento cualesquiera tal que $A \subset \mathbb{R}^n$, entonces la *función de densidad conjunta* $f(\mathbf{y})$, viene dada por:

$$\begin{aligned} P(\mathbf{Y} \in A) &= \int \cdots \int_A f(y_1, \ldots, y_n)\, dy_1 \cdots dy_n \\ &= \int \cdots \int_A f(\mathbf{y})\, d\mathbf{y} \end{aligned} \tag{10}$$

Entonces, dada una secuencia de variables aleatorias discretas o continuas $Y_1, \ldots, Y_n$; que puede denotarse como un vector aleatorio por $\mathbf{Y} = [Y_1, \ldots, Y_n]^\intercal$; la *función de probabilidad conjunta* (3.1) y la *función de densidad conjunta* (3.2), respectivamente, se representan indistintamente por $f(y_1, \ldots, y_n)$ o, de manera compacta, $f(\mathbf{y})$.

Obsérvese que en la definición de función de probabilidad (densidad) conjunta no se está suponiendo particularmente algo sobre la secuencia de variables aleatorias $Y_1, \ldots, Y_n$; sin embargo, si se asume que las variables aleatorias provienen de alguna distribución de probabilidad (que puede no ser la misma) y estas distribuyen *independientemente*, cada una con función de probabilidad (densidad) $f_{Y_i}(y_i)$, la función de probabilidad (densidad) conjunta de $Y_1, \ldots, Y_n$; viene dada por:

$$\begin{aligned} f(\mathbf{y}) &= f(y_1, \ldots, y_n) \\ &= f_{Y_1}(y_1) \cdot \ \cdots \ \cdot f_{Y_n}(y_n) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i) \end{aligned} \tag{11}$$

Si, además, se asume que la secuencia de variables aleatorias $Y_1, \ldots, Y_n$, distribuyen *idénticamente*, todas con función de probabilidad (densidad) $f(y_i)$, la función de probabilidad (densidad) conjunta de $Y_1, \ldots, Y_n$; viene dada por:

$$\begin{aligned} f(\mathbf{y}) &= f(y_1, \ldots, y_n) \\ &= f(y_1) \cdot \cdots \cdot f(y_n) \\ &= \prod_{i=1}^{n} f(y_i) \end{aligned} \tag{12}$$

Respecto a la distribución de probabilidad, serán de nuestro interés particular las distribuciones de probabilidad parametrizadas, puesto que los modelos estadísticos, por lo general, se describen a partir de este tipo de distribuciones de probabilidad. Estas distribuciones de probabilidad, como cualquier otra, quedan completamente especificadas, ya sea por su función de distribución o por su función de probabilidad (densidad); no obstante, a estas las caracteriza un número finito de parámetros. De este modo, si se asume que la secuencia de variables aleatorias $Y_1, \ldots, Y_n$; proviene de alguna distribución de probabilidad parametrizada, cuya función de probabilidad (densidad) es denotada por $f_{Y_i}(y_i)$ o $f(y_i)$; es caracterizada por un conjunto finito de parámetros denotados por $\boldsymbol{\theta} \in \Theta$, donde $\Theta \subset \mathbb{R}^k$ denota el espacio paramétrico de $\boldsymbol{\theta}$ y $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]^\intercal$, denota un vector de dimensión $k \times 1$. Entonces, para enfatizar la dependencia de la función de probabilidad (densidad) sobre el conjunto de parámetros $\boldsymbol{\theta} \in \Theta$, denotaremos esto por, $f_{Y_i}(y_i \mid \boldsymbol{\theta})$ o $f(y_i \mid \boldsymbol{\theta})$; y, en consecuencia, la función de probabilidad (densidad) conjunta de $Y_1, \ldots, Y_n$, de manera general, se denota por:

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = f(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \tag{13}$$

Sin pérdida de generalidad, en lo sucesivo, tanto $f(\mathbf{y} \mid \boldsymbol{\theta})$, como $f(y_1, \ldots, y_n \mid \boldsymbol{\theta})$, representarán a alguna función de *densidad* conjunta de $Y_1, \ldots, Y_n$.

Definición 3.3 (Función de verosimilitud) *Sea $Y_1, \ldots, Y_n$, una secuencia de variables aleatorias continuas, cuya función de densidad conjunta depende de un conjunto de parámetros desconocidos $\boldsymbol{\theta} \in \Theta$; es decir, $f(y_1, \ldots, y_n \mid \boldsymbol{\theta})$. Entonces, la función de verosimilitud evaluada en los datos observados $y_1, \ldots, y_n$; denota por $L(\theta \mid y_1, \ldots, y_n)$; se define como la función de densidad conjunta de $Y_1, \ldots, Y_n$; es decir:*

$$\begin{aligned} L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) &= f(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \\ L(\boldsymbol{\theta} \mid \mathbf{y}) &= f(\mathbf{y} \mid \boldsymbol{\theta}) \end{aligned} \tag{14}$$

Se debe notar que la función de verosimilitud $L(\boldsymbol{\theta} \mid \mathbf{y})$ es función del conjunto de parámetros $\boldsymbol{\theta} \in \Theta$, dados los datos observados $\mathbf{y}$; mientras que, la función de densidad conjunta $f(\mathbf{y} \mid \boldsymbol{\theta})$ es función de $\mathbf{y}$, dados valores fijos de $\boldsymbol{\theta}$. En este sentido, la función de verosimilitud mide cuan bien el modelo estadístico logra explicar los datos observados $\mathbf{y}$.

Si se asume que la secuencia de variables aleatorias $Y_1, \ldots, Y_n$; distribuyen *independientemente*, cada una con función de densidad $f_{Y_i}(y_i \mid \theta)$; la función de verosimilitud viene dada por:

$$\begin{aligned} L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) &= f(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \\ &= f_{Y_1}(y_1 \mid \boldsymbol{\theta}) \cdot \cdots \cdot f_{Y_n}(y_n \mid \boldsymbol{\theta}) \\ L(\boldsymbol{\theta} \mid \mathbf{y}) &= \prod_{i=1}^{n} f_{Y_i}(y_i \mid \boldsymbol{\theta}) \end{aligned} \tag{15}$$

En tanto, si se asume que la secuencia de variables aleatorias $Y_1, \ldots, Y_n$; distribuyen *independiente* e *idénticamente*, todas con función de densidad $f(y_i \mid \boldsymbol{\theta})$; la función de verosimilitud viene dada por:

$$\begin{aligned} L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) &= f(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \\ &= f(y_1 \mid \boldsymbol{\theta}) \cdot \cdots \cdot f(y_n \mid \boldsymbol{\theta}) \\ L(\boldsymbol{\theta} \mid \mathbf{y}) &= \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}) \end{aligned} \tag{16}$$

Definición 3.4 (Función de log-verosimilitud) *La función de log-verosimilitud, denota por $\ell(\boldsymbol{\theta} \mid y_1, \ldots, y_n)$; se define como el logaritmo natural de la función de verosimilitud; es decir:*

$$\begin{aligned} \ell(\boldsymbol{\theta} \mid y_1, \ldots, y_n) &= \log L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) \\ \ell(\boldsymbol{\theta} \mid \mathbf{y}) &= \log L(\boldsymbol{\theta} \mid \mathbf{y}) \end{aligned} \tag{17}$$

De este modo, si se asume que la secuencia de variables aleatorias $\mathbf{Y} = [Y_1, \ldots, Y_n]^\intercal$ distribuyen *independiente* e *idénticamente* (*i.i.d.*), todas con función de densidad $f(y_i \mid \boldsymbol{\theta})$; la función de log-verosimilitud viene dada por:

$$\begin{aligned} \ell(\boldsymbol{\theta} \mid y_1, \ldots, y_n) &= \log L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) \\ &= \log \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\theta}) \\ \ell(\boldsymbol{\theta} \mid \mathbf{y}) &= \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{\theta}) \end{aligned} \tag{18}$$

3.2 Estimación por Máxima Verosimilitud

El método de *Estimación por Máxima Verosimilitud* (MLE), introducido por R. A. Fischer, es un método de estimación de parámetros de una distribución o, de manera más general, de un modelo estadístico. El método de MLE, *dados los datos observados* $\mathbf{y} = [y_1, \ldots, y_n]^\intercal$, *estima* valores para los parámetros del modelo $\boldsymbol{\theta} \in \Theta$; tales que estos *maximizan* la función de verosimilitud $L(\boldsymbol{\theta} \mid \mathbf{y})$. De este modo, a través de maximizar la función de verosimilitud $L(\boldsymbol{\theta} \mid \mathbf{y})$, se busca el valor de $\boldsymbol{\theta} \in \Theta$ que hace más verosímil la muestra de datos observada $\mathbf{y}$.

Definición 3.5 (Estimador de máxima verosimilitud) *Sea $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(Y_1, \ldots, Y_n)$ una estadística; es decir, una función de los datos observados $\mathbf{y}$, que permite estimar $\boldsymbol{\theta} \in \Theta$. Si el estimador $\widehat{\boldsymbol{\theta}}$, al ser evaluado en los datos observados, maximiza la función de verosimilitud $L(\boldsymbol{\theta} \mid \mathbf{y})$, entonces se denomina Estimador de Máxima Verosimilitud (MLE) de $\boldsymbol{\theta}$. En tanto, la estimación $\widehat{\boldsymbol{\theta}}_{MLE} = \widehat{\boldsymbol{\theta}}(y_1, \ldots, y_n)$, se denomina estimación máximo verosimil de $\boldsymbol{\theta}$. De este modo, se tiene que:*

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{MLE} &= \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) \\ &= \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \mid \mathbf{y}) \end{aligned} \tag{19}$$

Normalmente es más fácil trabajar con la función de log-verosimilitud $\ell(\boldsymbol{\theta} \mid \mathbf{y})$, la cual tiene el mismo máximo que la función de verosimilitud $L(\boldsymbol{\theta} \mid \mathbf{y})$, ya que la función logaritmo es una función continua y estrictamente monótona. Además, cualquier constante aditiva que involucre posiblemente los datos observados $\mathbf{y}$ pero no a $\boldsymbol{\theta}$, puede omitirse de la función de log-verosimilitud sin cambiar

la ubicación de su máximo o las diferencias entre los valores de la función de log-verosimilitud en diferentes valores de $\boldsymbol{\theta}$. De este modo, equivalentemente, se tiene que:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{MLE} &= \arg\max_{\boldsymbol{\theta}\in\Theta} \ \ell\left(\boldsymbol{\theta} \mid y_1, \ldots, y_n\right) \\ &= \arg\max_{\boldsymbol{\theta}\in\Theta} \ \ell\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \end{aligned} \tag{20}$$

La función de verosimilitud $L\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$ y la función de log-verosimilitud $\ell\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$, de manera más simple, se suelen denotar, respectivamente, por $L\left(\boldsymbol{\theta}\right)$ y $\ell\left(\boldsymbol{\theta}\right)$. De este modo, siguiendo la definición (3.5), el MLE del parámetro $\boldsymbol{\theta}$, denotado por $\widehat{\boldsymbol{\theta}}_{MLE}$, corresponde al valor de $\boldsymbol{\theta} \in \Theta$ que, dados los datos observados, hace máximo el valor de la función de verosimilitud $L\left(\boldsymbol{\theta}\right)$. Entonces, $\widehat{\boldsymbol{\theta}}_{MLE}$ resulta de resolver un problema de optimización donde la función objetivo a maximizar es $L\left(\boldsymbol{\theta}\right)$ o, equivalentemente, $\ell\left(\boldsymbol{\theta}\right)$; mientras que $\boldsymbol{\theta}$ es la variable de elección. Es decir, se busca resolver el siguiente problema:

$$\max \ \ell\left(\boldsymbol{\theta}\right) \quad ; \quad \texttt{s.a.} \ \boldsymbol{\theta} \in \Theta \tag{21}$$

Un enfoque sistemático para maximizar la función de log-verosimilitud $\ell\left(\boldsymbol{\theta}\right)$ es mediante el uso del cálculo diferencial; puesto que el valor buscado del parámetro $\boldsymbol{\theta} \in \Theta$ corresponderá a la raíz de la derivada de $\ell\left(\boldsymbol{\theta}\right)$ con respecto de $\boldsymbol{\theta}$. No obstante, se debe verificar que la raiz hallada corresponda a un máximo para la función $\ell\left(\boldsymbol{\theta}\right)$.

3.2.1 Estimación e Inferencia: $k = 1$

Sea $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)^{\intercal}$ el conjunto de parámetros desconocidos del modelo estadístico. Si $k = 1$, entonces $(\boldsymbol{\theta} = \theta)$; la condición necesaria que debe satisfacer el máximo de $\ell\left(\theta\right)$, es:

$$\frac{d\ell\left(\theta\right)}{d\theta} \ = \ \ell'\left(\theta\right) \ = \ 0 \tag{22}$$

Luego, la raíz de la ecuación (22); es decir, el valor de $\theta \in \Theta$ que anula $\ell'\left(\theta\right)$[25], es candidato a máximo de $\ell\left(\theta\right)$. En general, la raíz de una función no es necesariamente el máximo global; podría ser simplemente un máximo local o incluso un mínimo local. En tal caso, para verificar que $\ell\left(\theta\right)$ alcanza un máximo en algún punto que verifica la ecuación (22), digamos $\theta = \widehat{\theta}$; debe verificarse que la segunda derivada de $\ell\left(\theta\right)$ respecto de θ es negativa en $\theta = \widehat{\theta}$. Es decir, se debe veriricar que:

$$\left.\frac{d^2\ell\left(\theta\right)}{d\theta^2}\right|_{\theta=\widehat{\theta}} \ = \ \left.\frac{d\ell'\left(\theta\right)}{d\theta}\right|_{\theta=\widehat{\theta}} \ = \ \ell''\left(\theta = \widehat{\theta}\right) \ < \ 0 \tag{23}$$

Entonces, si el punto $\theta = \widehat{\theta}$ verifica las ecuaciones (22) y (23), la función de log-verosimilitud $\ell\left(\theta\right)$ alcanza un máximo en tal punto. Luego, $\theta = \widehat{\theta}$ es el MLE de θ, el cual se suele denotar por $\widehat{\theta}_{MLE}$.

El MLE tiene una distribución muestral puesto que depende de la realización de las variables aleatorias $Y_1, \ldots, Y_n$. Este estimador puede ser sesgado o insesgado para θ, pero en condiciones bastante generales es asintóticamente insesgado cuando $n \longrightarrow \infty$. En tanto, la varianza muestral del MLE depende de la curvatura promedio de la función de log-verosimilitud $\ell\left(\theta\right)$: cuando $\ell\left(\theta\right)$ es muy empinada, la ubicación del máximo se conoce con mayor precisión. De este modo, la segunda

[25] Dentro de la inferencia clásica, $\ell'\left(\theta\right)$ suele denotarse por $U\left(\theta\right)$ y esta expresión se denomina *función score*. En tanto, la ecuación (22), cuando es escrita como $U\left(\theta\right) = 0$, se denomina *ecuación score*.

derivada de la función de log-verosimilitud; esto es $\ell''(\theta)$, permite medir qué tan bien determinado esta el MLE. Denotemos por $\mathcal{J}(\theta)$ a $-\ell''(\theta)$; es decir:

$$\mathcal{J}(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2} = -\frac{d\ell'(\theta)}{d\theta} = -\ell''(\theta) \tag{24}$$

Luego, considerando (24), $\mathcal{J}(\theta)$ debe ser positivo cerca del MLE, $\widehat{\theta}_{MLE}$. Si $\mathcal{J}(\theta)$ es grande, la pendiente de $\ell(\theta)$, descrita por $\ell'(\theta)$, está cambiando rápidamente cerca del estimador, lo que significa que la cima de la función de log-verosimilitud es muy pronunciada alrededor de $\widehat{\theta}_{MLE}$ y, por tanto, la estimación del parámetro, $\widehat{\theta}_{MLE}$, está bien definida. En esta situación, un pequeño cambio en la estimación de θ, cambiará sustancialmente el valor de la función de log-verosimilitud; por tanto, $\widehat{\theta}_{MLE}$ es una estimación muy precisa de θ. Por otro lado, si $\mathcal{J}(\theta)$ es cercano a cero, la pendiente de $\ell(\theta)$ está cambiando lentamente cerca del estimador, lo que significa que la cima de la función de log-verosimilitud es relativamente plana alrededor de $\widehat{\theta}_{MLE}$ y, por tanto, la estimación del parámetro, $\widehat{\theta}_{MLE}$, no está tan bien determinada. En esta situación, $\widehat{\theta}_{MLE}$ es un estimador menos preciso de θ. Por tanto, $\mathcal{J}(\theta)$ es una medida de la precisión de $\widehat{\theta}_{MLE}$; es decir, $\mathcal{J}(\theta)$ mide cuánta información está disponible para estimar θ.

La expresión $\mathcal{J}(\theta) = -\ell''(\theta)$, se denomina *información observada*. En tanto, $\mathcal{I}(\theta) = \mathbb{E}\{\mathcal{J}(\theta)\}$, se define como la *información esperada*, también denominada *información de Fisher*. Mientras que $\mathcal{J}(\theta)$ es una función de los datos observados, $\mathcal{I}(\theta)$ es una propiedad del modelo. Esta mide la información promedio que se observará para el parámetro del modelo y el valor especificado para este.

Se puede demostrar que $\mathcal{I}(\theta) = \mathbb{E}\{\ell'(\theta)^2\} = Var(\ell'(\theta))$. Lo anterior indica exactamente cómo la información esperada mide la tasa de cambio en la derivada de la función de log-verosimilitud alrededor del valor verdadero del parámetro. Una expansión lineal de la serie de Taylor de la función de log-verosimilitud alrededor de $\theta = \widehat{\theta}_{MLE}$, muestra además que:

$$Var\left(\widehat{\theta}_{MLE}\right) \approx \frac{1}{\mathcal{I}(\theta)} \tag{25}$$

Por tanto, la información esperada es una medida de la precisión del MLE; específicamente, la varianza del MLE es inversamente proporcional a la *información de Fisher* $\mathcal{I}(\theta)$ para el parámetro. Luego, la varianza estimada para $\widehat{\theta}_{MLE}$, viene dada por:

$$\widehat{Var}\left(\widehat{\theta}_{MLE}\right) = \frac{1}{\mathcal{I}\left(\widehat{\theta}_{MLE}\right)} \tag{26}$$

Finalmente, la desviación estándar (error estándar) estimada de $\widehat{\theta}_{MLE}$, viene dada por:

$$\widehat{se}\left(\widehat{\theta}_{MLE}\right) = \frac{1}{\sqrt{\mathcal{I}\left(\widehat{\theta}_{MLE}\right)}} \tag{27}$$

3.2.1.1 Distribución Weibull

Sea $Y_1, \ldots, Y_n$; una secuencia de variables aleatorias (v.a.) *i.i.d.* de una distribución Weilbull con función de densidad:

$$f(y \mid a, \sigma) = \left(\frac{a}{\sigma}\right)\left(\frac{y}{\sigma}\right)^{a-1} \exp\left(-\left(\frac{y}{\sigma}\right)^{a}\right) \tag{28}$$

para $y > 0$; donde $a > 0$ es el parámetro de forma y $\sigma > 0$ es el parámetro de escala.

Se puede mostrar que la función de log-verosimilitud $\ell(a, \sigma)$ de (28), viene dada por:

$$\begin{aligned} \ell(a,\sigma) &= \sum_{i=1}^{n} \log f(y_i \mid a, \sigma) \\ &= \sum_{i=1}^{n} \log \left\{ \left(\frac{a}{\sigma}\right) \left(\frac{y_i}{\sigma}\right)^{a-1} \exp\left(-\left(\frac{y_i}{\sigma}\right)^{a}\right) \right\} \\ &= n \log a - na \log \sigma + (a-1) \sum_{i=1}^{n} \log y_i - \sum_{i=1}^{n} \left(\frac{y_i}{\sigma}\right)^{a} \end{aligned} \tag{29}$$

Si se asume que se conoce el valor del parámetro de forma a; de las ecuaciones (22 y 29), se tiene que el MLE de σ viene de resolver, para σ, la ecuación:

$$\frac{d\ell(\sigma)}{d\sigma} = 0 \tag{30}$$

No es difícil mostrar que la solución de la ecuación (30), corresponde a:

$$\widehat{\sigma}_{MLE} = \left(\frac{1}{n} \sum_{i=1}^{n} {y_i}^{a}\right)^{1/a} \tag{31}$$

Puede remitirse al Anexo 1.1. Distribución Weibull, donde se presenta una aplicación para este caso.

3.2.2 Estimación e Inferencia: $k > 1$

Sea $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)^{\intercal}$ el conjunto de parámetros desconocidos del modelo estadístico. Si $k > 1$, la condición necesaria que debe satisfacer el máximo de $\ell(\boldsymbol{\theta})$, es:

$$\nabla \ell(\boldsymbol{\theta}) = \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k}\right)^{\intercal} = \mathbf{0} \tag{32}$$

donde, $\nabla \ell(\boldsymbol{\theta})$[26] es el vector gradiente de $\ell(\boldsymbol{\theta})$ y $\mathbf{0}$ es el vector de ceros, ambos vectores de dimensión $k \times 1$. Luego, la raíz del sistema de ecuaciones (32); es decir, el valor del vector $\boldsymbol{\theta} \in \Theta$ que anula $\nabla \ell(\boldsymbol{\theta})$, es candidato a máximo de $\ell(\boldsymbol{\theta})$ y, al igual que cuando $k = 1$, se debe verificar que $\ell(\boldsymbol{\theta})$ alcanza un máximo en el punto que verifica el sistema de ecuaciones (32), digamos $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$; para esto debe verificarse que la matriz Hessiana de $\ell(\boldsymbol{\theta})$, denotada por $\mathbf{H}(\ell(\boldsymbol{\theta}))$, es (semi) definida negativa en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. La matriz Hessiana, también denotada por $\mathbf{H}_{\ell}$, es una matriz cuadrada de dimensión $k \times k$, definida como:

$$\mathbf{H}(\ell(\boldsymbol{\theta})) = \left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right] \tag{33}$$

$\forall i, j : \; i = 1, \ldots, k \; ; \; j = 1, \ldots, k.$

A partir de las expresiones descritas cuando $k = 1$, se pueden deducir las expresiones correspondientes para la *matriz de información observada* $\mathcal{J}(\boldsymbol{\theta})$, la *matriz de información esperada* o *matriz de información de Fisher* $\mathcal{I}(\boldsymbol{\theta})$ y la *matriz de varianzas y covarianzas* $\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)$. En esta última, la varianza estimada para cada parámetro estimado se encuentra en la diagonal de la inversa de la matriz de información de Fisher.

[26] El vector $\nabla \ell(\boldsymbol{\theta})$ suele denotarse por $U(\boldsymbol{\theta}) = (U(\boldsymbol{\theta}_1), \ldots, U(\boldsymbol{\theta}_k))^{\intercal}$ y, en tal caso, $U(\boldsymbol{\theta})$ se denominan las *funciones score* y el sistema de ecuaciones (29), cuando es escrito como $U(\boldsymbol{\theta}) = \mathbf{0}$, se denomina como el *sistema de ecuaciones score*.

3.2.2.1 Distribución Normal

Sea $Y_1, \ldots, Y_n$; una secuencia de v.a. *i.i.d.* de una distribución normal con función de densidad:

$$f\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{34}$$

para $-\infty < y < \infty$; donde μ y σ^2 son la media y la varianza, respectivamente, de la distribución.

Se puede mostrar que la función de log-verosimilitud $\ell\left(\mu, \sigma^2\right)$ de (34), viene dada por:

$$\begin{aligned} \ell\left(\mu, \sigma^2\right) &= \sum_{i=1}^{n} \log f\left(y_i \mid \mu, \sigma^2\right) \\ &= \sum_{i=1}^{n} \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)\right\} \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \end{aligned} \tag{35}$$

Si se asume que no se conocen los valores de ambos parámetros, μ y σ^2; como se señala en (32), el MLE para μ y σ^2 vienen de resolver, para μ y σ^2, el sistema de ecuaciones:

$$\nabla\ell\left(\mu, \sigma^2\right) = \left(\frac{\partial\ell\left(\mu, \sigma^2\right)}{\partial\mu}, \frac{\partial\ell\left(\mu, \sigma^2\right)}{\partial\sigma^2}\right)^{\intercal} = \mathbf{0} \tag{36}$$

No es difícil mostrar que la solución del sistema de ecuaciones (36), corresponde a:

$$\widehat{\mu}_{MLE} = \bar{y} \tag{37}$$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{38}$$

donde $\bar{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$.

Finalmente, para cerrar este apartado referido al método de Estimación por Máxima Verosimilitud, la siguiente proposición describe una de las propiedades más importantes del MLE: la propiedad de *invarianza*.

Proposición 3.1 *Sea $g(\boldsymbol{\theta})$ una función continua y estrictamente monótona que depende de $\boldsymbol{\theta}$. Si $\widehat{\boldsymbol{\theta}}_{MLE}$ es MLE de $\boldsymbol{\theta}$, entonces $g\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)$ es MLE de $g(\boldsymbol{\theta})$.*

3.3 Modelos Lineales Generalizados

Nuestra discusión sobre la función de verosimilitud hasta ahora solo ha considerado la distribución de probabilidades del conjunto de datos $\mathbf{Y}$. Sin embargo, la mayoría de las situaciones en las que se requiere algún tipo de modelización estadística son más complejas de lo que se puede describir mediante una distribución de probabilidades; situación ampliamente descrita en la sección precedente. Por el contrario, si estamos interesados en incorporar algún mecanismo que nos permita analizar Y como función de un otro conjunto de variables explicativas (covariables y factores) de la primera,

esto supone el uso de modelos estadísticos, específicamente, modelos de regresión que consideran un componente aleatorio y un componente sistemático.

De manera muy general, un modelo de regresión supone que $\mathbf{y} = [Y_1, \ldots, Y_n]^\intercal$, es una secuencia de variables respuesta que se asumen *independientes* entre sí y que la variable respuesta de la observación i, la media condicional $\mathbb{E}(Y_i|\mathbf{x}_i^\intercal) = \mu_i$, depende de p variables explicativas descritas por $\mathbf{x}_i^\intercal = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ y un conjunto de p parámetros de regresión, denotados por $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\intercal$, a través de alguna función general, digamos f. Se asume, además, que el conjunto de variables explicativas $\mathbf{x}_i^\intercal$ y los parámetros $\boldsymbol{\beta}$, combinan linealmente los efectos de las primeras. De este modo, el componente sistemático a menudo se describe como[27]:

$$\begin{aligned} \mathbb{E}(Y_i|\mathbf{x}_i^\intercal) \;=\; \mu_i \;&=\; f(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \\ &=\; f\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) \\ &=\; f(\mathbf{x}_i^\intercal \boldsymbol{\beta}) \end{aligned} \tag{39}$$

para $i = 1, \ldots, n$; donde, el componente $\mathbf{x}_i^\intercal \boldsymbol{\beta}$ suele denominarse el *predictor lineal* de la regresión.

La ecuación (39), para el conjunto de datos, se puede escribir en notación matricial, como:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) \;=\; \boldsymbol{\mu} \;=\; f(\mathbf{X}\boldsymbol{\beta}) \tag{40}$$

donde, $\mathbf{X} = (\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_n^\intercal)^\intercal$ es la *matriz de diseño*, matriz de dimensión $n \times p$, $\boldsymbol{\mu}$ es el vector de medias condicionales del vector $\mathbf{Y}$, ambos vectores de dimensión $n \times 1$, y $\boldsymbol{\beta}$ es el vector de parámetros desconocidos de dimensión $p \times 1$; siendo p es el número de parámetros desconocidos del modelo, incluyendo el intercepto que, en esta presentación, viene dado por β_1. Luego, $\mathbf{Y}$ contiene la parte aleatoria, mientras que $\mathbf{X}\boldsymbol{\beta}$ es el *predictor lineal*.

Los modelos de regresión como el descrito en (39 y 40), son modelos de regresión lineales en los parámetros $\boldsymbol{\beta}$. Dentro de este tipo de modelos de regresión, destacan los *Modelos Lineales* (LMs), que son el método estadístico más utilizado en el análisis de regresión, esto debido a que son simples de construir e interpretar. Sin embargo, los supuestos estadísticos en que se basan los LMs, limitan (de manera considerable) el análisis estadístico al momento de utilizar modelos estadísticos para analizar la variable respuesta $\mathbf{Y}$, como función de un otro conjunto de variables explicativas $\mathbf{X}$. Con el propósito de evitar que las limitaciones propias de los LMs acoten los alcances de nuestro estudio, se introduce una clase de modelos más amplia: los *Modelos Lineales Generalizados* (GLMs).

Los GLMs fueron popularizados en el trabajo de (Nelder and Wedderburn 1972), donde muestran que:

i. Los modelos de regresión lineales más comunes de la estadística clásica, incluidos los LMs, son en realidad miembros de una misma familia de modelos y que pueden tratarse de la misma manera.

ii. El método de Estimación por Máxima Verosimilitud puede ser utilizado para todos estos modelos y los MLEs de los parámetros desconocidos de los GLMs se pueden obtener utilizando el mismo algoritmo: el algoritmo de *Mínimos Cuadrados Ponderados Iterados* (IWLS).

Antes de realizar una presentación formal de los GLMs, conviene realizar una breve descripción de los LMs y, además, introducir una importante familia de distribuciones para los GLMs.

[27] En el conjunto de variables explicativas $\mathbf{x}_i^\intercal$, en esta presentación, se asume que el primer término es constante e igual a uno; es decir, $x_{i1} = 1\,, \forall i:\; i = 1, \ldots, n$.

3.3.1 Modelos Lineales

El supuesto básico de los LMs es asumir una *relación lineal* entre la media condicional de la variable respuesta $\mathbf{Y}$ y el predictor lineal $\mathbf{X}\boldsymbol{\beta}$ y, además, los valores de la variable respuesta son continuos y siguen una distribución normal condicional en $\mathbf{X}$, con varianza constante.

Sea $\mathbf{Y} = [Y_1, \ldots, Y_n]^\intercal$ una secuencia de variables respuesta que distribuyen *independientes* condicionales en $\mathbf{X}$ desde un distribución normal con media μ_i, para $i = 1, \ldots, n$; y varianza constante σ^2; es decir, $Y_i|\mathbf{x}_i^\intercal \sim N(\mu_i, \sigma^2)$; $\forall i$. Además, suponga que $\mathbf{x}_i^\intercal = [1, x_{i2}, \ldots, x_{ip}]$ es un vector con p variables explicativas para la observación i y $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\intercal$, es el conjunto de p parámetros desconocidos; donde el parámetro desconocido β_j está asociado a la correspondiente variable explicativa $x_{\cdot j}$, para $j = 1, \ldots, p$. Entonces, el LM clásico puede ser escrito como:

$$\begin{aligned} \mathbb{E}(Y_i|\mathbf{x}_i^\intercal) \;=\; \mu_i \;&=\; \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \\ &=\; \sum_{j=1}^{p} \beta_j x_{ij} \\ &=\; \mathbf{x}_i^\intercal \boldsymbol{\beta} \end{aligned} \tag{41}$$

para, $i = 1, \ldots, n$.

La ecuación (41), para el conjunto de datos, se puede escribir en notación matricial, como:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) \;=\; \boldsymbol{\mu} \;=\; \mathbf{X}\boldsymbol{\beta} \tag{42}$$

Para estimar los parámetros desconocidos $\boldsymbol{\beta}$ de un LM, en general, se utiliza el *Método de Mínimos Cuadrados*[28]. Este método proporciona el estimador más eficiente entre todos los estimadores insesgados, porque tiene varias propiedades estadísticas deseables y, en tal sentido, se lo denomina el "*Mejor Estimador Lineal Insesgado*" (BLUE). No obstante, como se ha mencionado, los LMs presentan una serie de limitaciones para el análisis. En particular, a continuación, se discuten sobre dos limitaciones que resultan relevantes para los fines de nuestro estudio.

Una importante limitación del modelo lineal es que este tipo de modelos solo son adecuados cuando la variable respuesta $\mathbf{Y}$ (o alguna transformación de $\mathbf{Y}$) distribuye normal con varianza constante; situación que limita el análisis al caso en que la variable respuesta $\mathbf{Y}$, corresponda a datos continuos que provienen de una distribución normal con varianza constante. En tal sentido, los LMs no serían adecuados en el caso que la variable respuesta $\mathbf{Y}$ corresponda, por ejemplo, a datos binarios que provienen de una distribución Bernoulli o datos discretos que provienen de una distribución Binomial. Los LMs tienen incluso dificultades con datos continuos, pero cuyo soporte esta restringido a un subconjunto de los reales, como es el caso de datos que provienen de una distribución Beta. Otra importante limitación de los LMs, es asumir una relación lineal entre la media condicional de la variable respuesta $\mathbf{Y}$ y el predictor lineal $\mathbf{X}\boldsymbol{\beta}$; es decir, si la relación antes descrita no puede describirse de forma lineal, los LMs tampoco serían adecuados en tal situación.

Los GLMs se llaman así porque generalizan los modelos lineales clásicos, incluidos los LMs, basados en la distribución normal. Esta generalización, básicamente, tiene dos aspectos:

i. Para la variable respuesta $\mathbf{Y}$, es posible considerar distribuciones distintas a la distribución normal; pudiendo considerar distribuciones para variables aleatorias binarias, categóricas, discretas, así como una amplia gama de distribuciones continuas, incluida la distribución normal.

[28] Una descripción de este método puede encontrarse en (Dobson and Barnett 2018, sec. 1.6.3) y (Dunn and Gordon 2018, sec. 2.3.1).

En particular, estos modelos pueden involucrar una variedad de distribuciones seleccionadas desde una familia de distribuciones especial: la *familia de dispersión exponencial*[29].

ii. La asociación entre la media de variable respuesta y el predictor lineal, no tiene que ser tomar la simple forma lineal descrita en (41 y 42). Estos modelos involucran transformaciones de la media $\boldsymbol{\mu}$, a través de lo que se denomina la *función de enlace*; la cual, como veremos, permite vincular los componentes sistemático y aleatorio del modelo de regresión, incluso de forma no lineal.

Realizada esta breve descripción a cerca de los LMs, mencionando los supuestos de estos modelos, pero además realizando un breve comentario sobre dos importantes limitaciones que tienen estos; como se mencionó al inicio de este apartado, conviene introducir una importante familia de distribuciones para los GLMs, puesto que los miembros de tal familia de distribuciones poseen propiedades estadísticas deseables. Esta familia de distribuciones corresponde a la generalización de la *familia exponencial*: la *familia de dispersión exponencial*.

3.3.2 La Familia Exponencial y de Dispersión Exponencial

Suponga que se tiene una secuencia de variables respuesta aleatorias *independientes* Z_i, $\forall i : i = 1, \ldots, n$; cada una con función de probabilidad si Z es discreta, o función de densidad si Z es continua; $f(z_i \mid \xi_i)$ que puede ser escrita en la forma:

$$\begin{aligned} f(z_i|\xi_i) &= r(z_i)\, s(\xi_i) \exp\{t(z_i)\, u(\xi_i)\} \\ &= \exp\{t(z_i)\, u(\xi_i) + v(z_i) + w(\xi_i)\} \end{aligned} \tag{43}$$

donde, $r(), s(), t()$ y $u()$; son funciones conocidas, con $r(z_i) = \exp\{v(z_i)\}$ y $s(\xi_i) = \exp\{w(\xi_i)\}$. Además, ξ_i es un *parámetro de localización* que indica la posición donde se encuentra la distribución dentro del rango de posibles valores para la variable respuesta. Cualquier distribución que pueda escribirse en la forma que se describe en (43), se dice es miembro de la *familia exponencial*.

La *forma canónica* de la variable aleatoria, la del parámetro y la de la distribución, se obtiene haciendo $y = t(z)$ y $\theta = u(\xi)$. Si estas son transformaciones uno a uno, estas se simplifican, pero no cambian fundamentalmente. Luego, el modelo ahora se convierte en:

$$f(y_i|\theta_i) = \exp\{y_i\theta_i - b(\theta_i) + c(y_i)\} \tag{44}$$

donde, $b(\theta_i)$ es la constante de normalización de la distribución. Ahora, Y_i, $\forall i : i = 1, \ldots, n$; es un conjunto de variables aleatorias *independientes* con medias, digamos μ_i, de modo que podríamos, clásicamente, escribir $y_i = \mu_i + \varepsilon_i$.

Como se verá más adelante, $b(\theta)$ es una función muy importante, puesto que de sus derivadas se obtienen las funciones para la media y la varianza.

La familia exponencial puede generalizarse al incluir un *parámetro de escala* (constante), digamos ϕ, en la distribución, tal que:

$$f(y_i \mid \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\} \tag{45}$$

[29]Esta restricción, en realidad, surge por razones puramente técnicas. La razón es que el algoritmo numérico comúnmente utilizado para la estimación de los GLMs, el algoritmo de *Mínimos Cuadrados Ponderados Iterados* (IWLS), solo funciona dentro de esta familia de distribuciones. Sin embargo, con los computadores modernos, esta limitación puede eliminarse fácilmente (Lindsey 1997, pg.9).

donde, θ_i sigue siendo la forma canónica del parámetro de localización, alguna función de la media, μ_i. Entonces, se dice que $f(y_i \mid \theta_i, \phi)$ es miembro de la *familia de dispersión exponencial*. Por último, es claro que cualquier miembro de la familia exponencial es también miembro de la familia de dispersión exponencial, con $a_i(\phi) = 1$.

3.3.3 Estructura de los Modelos Lineales Generalizados

(Nelder and Wedderburn 1972, pg.372), caracterizan a través de tres componentes, lo que ellos denominan, "*un modelo que produce el modelo lineal generalizado*". En este sentido, los GLMs pueden describirse a partir de los siguientes tres componentes:

1. *Distribución de la variable respuesta* o *estructura del error*: $\mathbf{Y}$

Se asume que la secuencia $Y_1, \ldots, Y_n$; son v.a. *independientes* entre sí, que vienen de alguna distribución de la familia de dispersión exponencial. De este modo, las v.a. Y_i, $\forall i : i = 1, \ldots, n$, si bien comparten la misma distribución de la familia de dispersión exponencial, con un parámetro de escala constante, cada observación tiene su propia media, μ_i. Es decir, se asume que el conjunto de datos $\mathbf{Y}$ distribuyen, desde una misma distribución de la familia de dispersión exponencial, de manera *independiente*, pero no idénticamente.

2. *Predictor Lineal*: $\mathbf{X}\boldsymbol{\beta}$

Se asume que se cuenta con un conjunto de variables explicativas conocidas $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_n^\intercal]^\intercal$, también denominada *matriz de diseño* y un conjunto de p parámetros desconocidos $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^\intercal$, tal que:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{46}$$

donde $\mathbf{X}\boldsymbol{\beta}$ es el predictor lineal. Este describe cómo cambia la localización de la distribución de la respuesta con las variables explicativas $\mathbf{X}$.

3. *Función de Enlace*: $g_i(\mu_i)$

Si se toma la forma canónica del parámetro de localización θ_i de la distribución definida en el primer componente, tal que: $\theta_i = \eta_i$, la estructura de nuestro *GLM* está completo. Sin embargo, una generalización adicional a transformaciones no canónicas de la media requiere un componente adicional si se quiere mantener la idea de una estructura lineal. La relación entre la media de la i-ésima observación y su predictor lineal vendrá dada por una *función de enlace*, $g_i(\cdot)$:

$$\begin{aligned} \eta_i &= g_i(\mu_i) \\ &= \mathbf{x}_i^\intercal \boldsymbol{\beta} \end{aligned} \tag{47}$$

La *función de enlace* $g_i(\cdot)$ debe ser monótona y diferenciable. Normalmente se utiliza la misma función de enlace para todas las observaciones. Entonces, la *función de enlace canónica* es aquella función que transforma la media μ en la forma canónica del parámetro de localización θ_i, de la distribución de la familia de dispersión exponencial, definida en el primer componente.

3.3.4 Estimación e Inferencia

Supongamos que el conjunto de datos corresponden a una muestra de v.a. *independientes* que viene dada por $(y_i, \mathbf{x}_i^\intercal)$, para $i = 1, \ldots, n$; donde para cada elemento de la muestra $y_i|\mathbf{x}_i^\intercal$ se tiene una

función de densidad que pertenece a la familia de dispersión exponencial, tal como se definió en (45):

$$f\left(y_i \mid \theta_i, \phi\right) \quad = \quad \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{a_i\left(\phi\right)} + c\left(y_i, \phi\right)\right\}$$

para $i = 1, \ldots, n$. En la presentación, se tiene fija una distribución que pertenece a alguna familia de dispersión exponencial subyacente y un parámetro de dispersión común ϕ, pero se permite que cada elemento de la muestra $y_i|\mathbf{x}_i^\intercal$, tenga su propio parámetro natural: θ_i.

Nuestro objetivo es estimar las medias $\mu_i = \mathbb{E}\left(Y_i|\mathbf{x}_i^\intercal\right)$, para $i = 1, \ldots, n$. Recordemos, de la ecuación (47), que tenemos una función de enlace $\eta_i = g_i(\mu_i)$, que conecta la media μ_i con el parámetro $\eta_i = \mathbf{x}_i^\intercal\boldsymbol{\beta}$. Por lo tanto, primero se pueden estimar los coeficientes $\boldsymbol{\beta}$, digamos $\widehat{\boldsymbol{\beta}}$, y luego se pueden usar estas estimaciones de forma que:

$$g\left(\widehat{\mu_i}\right) \quad = \quad \mathbf{x}_i^\intercal\widehat{\boldsymbol{\beta}} \tag{48}$$

o, equivalentemente:

$$\widehat{\mu_i} \quad = \quad g^{-1}\left(\mathbf{x}_i^\intercal\widehat{\boldsymbol{\beta}}\right) \tag{49}$$

para $i = 1, \ldots, n$.

Para calcular $\widehat{\boldsymbol{\beta}}$ podemos utilizar el método de *Estimación por Máxima Verosimilitud*. La verosimilitud del conjunto de datos $(y_i, \mathbf{x}_i^\intercal)$, para $i = 1, \ldots, n$; condicional en $\mathbf{x}_i^\intercal$, es:

$$\begin{aligned} L\left(\boldsymbol{\theta}\right) \quad &= \quad \prod_{i=1}^{n} f\left(y_i \mid \theta_i, \phi\right) \\ &= \quad \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{a_i\left(\phi\right)} + c\left(y_i, \phi\right)\right\} \end{aligned} \tag{50}$$

función que se escribe como una función de $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^\intercal$, para denotar la dependencia del parámetro natural. Luego, la función de log-verosimilitud, viene dada por:

$$\begin{aligned} \ell\left(\boldsymbol{\theta}\right) \quad &= \quad \log L\left(\boldsymbol{\theta}\right) \\ &= \quad \sum_{i=1}^{n} \log f\left(y_i \mid \theta_i, \phi\right) \\ &= \quad \sum_{i=1}^{n} \frac{y_i\theta_i - b\left(\theta_i\right)}{a_i\left(\phi\right)} + \sum_{i=1}^{n} c\left(y_i, \phi\right) \end{aligned} \tag{51}$$

De este modo, si se busca maximizar la función de log-verosimilitud (51), sobre todas las posibles elecciones de coeficientes $\boldsymbol{\beta} \in \mathbb{R}^p$; que es verdaderamente una función de $\boldsymbol{\beta}$, porque cada parámetro natural θ_i puede escribirse en términos de la media μ_i de la distribución que pertenece a alguna familia de dispersión exponencial, y $\mu_i = \mathbf{x}_i^\intercal\boldsymbol{\beta}$, para $i = 1, \ldots, n$. Por tanto podemos escribir:

$$\ell\left(\boldsymbol{\beta}\right) \quad = \quad \sum_{i=1}^{n} y_i\theta_i - b\left(\theta_i\right) \tag{52}$$

donde se han descartado términos que no dependen de θ_i, para $i = 1, \ldots, n$.

Para ser más concretos, supongamos que consideramos una función de enlace canónica g, recordando que esta es la función de enlace que establece $\theta_i = \eta_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}$, para $i = 1, \ldots, n$. Entonces, la función de log-verosimilitud a maximizar sobre $\boldsymbol{\beta}$, es:

$$\ell(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} y_i \mathbf{x}_i^\intercal \boldsymbol{\beta} - b(\mathbf{x}_i^\intercal \boldsymbol{\beta}) \tag{53}$$

Excepto por el caso de mínimos cuadrados gaussianos; en general, no existe una forma cerrada para la solución de la maximización de $\ell(\boldsymbol{\beta})$. Por tanto, se debe recurrir a algoritmos de optimización para calcular su maximizador; es decir, se deben aplicar métodos numéricos.

Aun cuando, en el caso que la función de densidad involucrada pertenezca a la familia de dispersión exponencial, es posible maximizar $\ell(\boldsymbol{\beta})$, a través de realizar repetidamente regresiones de mínimos cuadrados ponderados; es decir, aplicar el método de *Mínimos Cuadrados Ponderados Iterados* (IWLS). Si bien el método es computacionalmente conveniente (y eficiente), el algoritmo no puede aplicarse a situaciones más generales donde, en particular, la función de densidad involucrada no pertenece a la familia de dispersión exponencial. No obstante, el *Método de Newton*, entre otros, son una alternativa al problema general.

3.4 Métodos Numéricos

Muchos modelos estadísticos realistas inducen funciones de verosimilitud y de log-verosimilitud que, distinto a los ejemplos anteriores, no pueden optimizarse analíticamente. Esto ocurre cuando, por ejemplo, $\ell'(\theta) = 0$ es una ecuación no lineal y, por tanto, la solución no puede ser determinada analíticamente. Similar situación acontece cuando, $\nabla \ell(\boldsymbol{\theta}) = \mathbf{0}$ es un sistema de ecuaciones no lineales. En ambas situaciones, un método extremadamente eficiente para encontrar raíces es el método de Newton[30].

3.4.1 El Método de Newton

Supongamos que se busca la raíz de g, función de valor real no lineal en x; es decir, se busca el valor de x tal que $g(x) = 0$. Supongamos que la raíz buscada es $x = x^*$; entonces, $g(x^*) = 0$. Supongamos, además, que g es continuamente diferenciable y que $g'(x) \neq 0$. Si $g(x)$ es diferenciable en el punto $x = x^{(t)}$, mediante la expansión lineal de primer orden de la serie de Taylor, la función $g(x)$ tiene una aproximación lineal, alrededor del punto $x = x^{(t)}$, la cual viene dada por:

$$g(x) \quad \approx \quad g\left(x^{(t)}\right) + g'\left(x^{(t)}\right)\left(x - x^{(t)}\right) \tag{54}$$

Si bien no es posible encontrar una solución analítica para la ecuación $g(x) = 0$, esto debido a la no linealidad de g en x; $g(x^*)$ puede ser aproximada por la aproximación lineal de $g(x)$; es decir:

$$g\left(x^{(t)}\right) + g'\left(x^{(t)}\right)\left(x^* - x^{(t)}\right) \quad = \quad 0 \tag{55}$$

Dado que $g(x^*)$ está siendo aproximada por su recta tangente alrededor del punto $x = x^{(t)}$, parece razonable aproximar la raíz de $g(x)$; esto es x^*, por la raíz de la recta tangente. Por lo tanto, resolviendo para x^*, se tiene:

$$x^* \quad = \quad x^{(t)} - \left\{g'\left(x^{(t)}\right)\right\}^{-1} g\left(x^{(t)}\right) \tag{56}$$

[30] Este método es también conocido en aplicaciones univariadas como el método de iteración de Newton-Raphson.

De este modo, iterando esta estrategia, se obtiene la ecuación de actualización del método de Newton:

$$x^{(t+1)} = x^{(t)} - \left\{g'\left(x^{(t)}\right)\right\}^{-1} g\left(x^{(t)}\right) \tag{57}$$

Cuando el problema de optimización corresponde a un problema de MLE donde se busca una solución a la ecuación no lineal $\ell'(\theta) = 0$, la ecuación de actualización para el método de Newton es:

$$\theta^{(t+1)} = \theta^{(t)} - \left\{\ell''\left(\theta^{(t)}\right)\right\}^{-1} \ell'\left(\theta^{(t)}\right) \tag{58}$$

En tanto, cuando el problema de optimización corresponde a un problema de MLE donde se busca una solución al sistema de ecuaciones no lineales $\nabla\ell(\boldsymbol{\theta}) = \mathbf{0}$, a partir de la aproximación multivariada de la serie de Taylor para $\nabla\ell(\boldsymbol{\theta})$, el algoritmo de actualización para el método de Newton es:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left\{\mathbf{H}\left(\ell\left(\boldsymbol{\theta}^{(t)}\right)\right)\right\}^{-1} \nabla\ell\left(\boldsymbol{\theta}^{(t)}\right) \tag{59}$$

donde, $\left\{\mathbf{H}\left(\ell\left(\boldsymbol{\theta}^{(t)}\right)\right)\right\}^{-1}$ es la inversa de la matriz Hessiana $\mathbf{H}$.

De manera similar, es posible aplicar el método de Newton, cuando el problema de optimización corresponde a un problema de MLE donde se busca optimizar $\ell(\boldsymbol{\beta})$.

3.4.1.1 Distribución Gamma

Sea $Y_1, \ldots, Y_n$; una secuencia de v.a. *i.i.d.* de una distribución Gamma con función de densidad que viene dada por:

$$f(y \mid \alpha, \sigma) = \frac{1}{\sigma^{\alpha}\Gamma(\alpha)} y^{\alpha-1} \exp\left(-\frac{y}{\sigma}\right) \tag{60}$$

para $y > 0$; donde $\alpha > 0$ es el parámetro de forma y $\sigma > 0$ es el parámetro de escala.

Se puede mostrar que la función de log-verosimilitud $\ell(\alpha, \sigma)$, viene dada por:

$$\begin{aligned} \ell(\alpha, \sigma) &= \sum_{i=1}^{n} \log f(y_i \mid \alpha, \sigma) \\ &= \sum_{i=1}^{n} \log\left\{\frac{1}{\sigma^{\alpha}\Gamma(\alpha)} {y_i}^{\alpha-1} \exp\left(-\frac{y_i}{\sigma}\right)\right\} \\ &= -n\log\Gamma(\alpha) - n\alpha\log\sigma + (\alpha-1)\sum_{i=1}^{n}\log y_i - \frac{1}{\sigma}\sum_{i=1}^{n} y_i \end{aligned} \tag{61}$$

Si se asume que se conoce el valor del parámetro de escala σ, el estimador de máxima verosimilitud para α, se obtiene de resolver, para α, la siguiente ecuación:

$$\frac{d\ell(\alpha)}{d\alpha} = 0 \tag{62}$$

No es difícil mostrar que la ecuación (49) corresponde a:

$$-n\left(\psi(\alpha) + \log\sigma\right) + \sum_{i=1}^{n}\log y_i = 0 \tag{63}$$

donde, $\psi(\alpha) = \dfrac{d \log \Gamma(\alpha)}{d\alpha}$ es la función digamma. La ecuación (50) es no lineal en α; por tanto, no es posible obtener una solución analítica. No obstante, el método de Newton-Rapshon descrito en (45), puede aplicarse a este caso y, de este modo, obtener una solución numérica para $\widehat{\alpha}_{MLE}$.

Puede remitirse al Anexo 1.2. Distribución Gamma, donde se presenta una aplicación para este caso.

3.4.1.2 Distribución Weibull

La función de log-verosimilitud de una muestra aleatoria que proviene de una distribución Weibull $\ell(a, \sigma)$, quedó descrita en (25) tal como sigue:

$$\ell(a, \sigma) \quad = \quad n \log a - na \log \sigma + (a-1) \sum_{i=1}^{n} \log y_i - \sum_{i=1}^{n} \left(\frac{y_i}{\sigma} \right)^a$$

Si se asume que no se conocen los valores de ambos parámetros, a y σ; el estimador de máxima verosimilitud para a y σ, se obtienen de resolver, para a y σ, el siguiente sistema de ecuaciones:

$$\nabla \ell(a, \sigma) \quad = \quad \left(\frac{\partial \ell(a, \sigma)}{\partial a}, \frac{\partial \ell(a, \sigma)}{\partial \sigma} \right)^{\mathsf{T}} = \mathbf{0} \tag{64}$$

No es difícil mostrar que el sistema de ecuaciones (51) corresponde a:

$$\frac{n}{a} - n \log \sigma + \sum_{i=1}^{n} \log y_i - \frac{1}{\sigma^a} \sum_{i=1}^{n} {y_i}^a \log \frac{y_i}{\sigma} \quad = \quad 0 \tag{65}$$

$$-\frac{na}{\sigma} + \frac{a}{\sigma^{a+1}} \sum_{i=1}^{n} {y_i}^a \quad = \quad 0 \tag{66}$$

El sistema de ecuaciones (65 y 64) es no lineal en a y σ; por tanto, no es posible obtener una solución analítica. No obstante, el método de Newton descrito en (46), puede aplicarse a este caso y, de este modo, obtener una solución numérica para $\widehat{a}_{MLE}$ y $\widehat{\sigma}_{MLE}$.

Puede remitirse al Anexo 1.3. Distribución Weibull, donde se presenta una aplicación para este caso.

3.5 Estimación con Datos Incompletos

Supongamos que $\mathbf{Y}$ denota el conjunto de datos completos, donde $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ y, por su parte, $\boldsymbol{\theta}$ denota el conjunto de parámetros de interés; de foma que $P(\mathbf{Y} \mid \boldsymbol{\theta})$, denota la función de densidad conjunta de los datos completos. Por otro lado, $\mathbf{R}$ denota la matriz de respuesta y $\boldsymbol{\psi}$ denota es el conjunto de parámetros del modelo estadístico para $\mathbf{R}$; es decir, $P(\mathbf{R} \mid \mathbf{Y}, \boldsymbol{\psi})$, denota el mecanismo de falta de respuesta.

Dado que la información observada en los datos incluye los datos observados y la matriz de respuesta; esto es $\mathbf{Y}_{obs}$ y $\mathbf{R}$; respectivamente, la función de verosimilitud de los datos observados se puede expresar como:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) \quad \propto \quad P(\mathbf{Y}_{obs}, \mathbf{R} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{67}$$

Luego, por definición de función de densidad marginal y probabilidad conjunta, se puede escribir:

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}) &\propto P(\mathbf{Y}_{obs}, \mathbf{R} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \int P(\mathbf{Y}, \mathbf{R} \mid \boldsymbol{\theta}, \boldsymbol{\psi})\, d\mathbf{Y}_{mis} \\ &= \int P(\mathbf{Y} \mid \boldsymbol{\theta})\, f(\mathbf{R} \mid \mathbf{R}, \boldsymbol{\psi})\, d\mathbf{Y}_{miss} \\ &= \int P(\mathbf{Y}_{obs}, \mathbf{Y}_{miss} \mid \boldsymbol{\theta})\, f(\mathbf{R} \mid \mathbf{Y}_{obs}, \mathbf{Y}_{miss}, \boldsymbol{\psi})\, d\mathbf{Y}_{miss} \end{aligned} \tag{68}$$

Si es posible asumir que el *mecanismo es ignorable*, se tiene que.

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\psi}) &= \int P(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta})\, f(\mathbf{R} \mid \mathbf{Y}_{obs}, \boldsymbol{\psi})\, d\mathbf{Y}_{mis} \\ &= P(\mathbf{R} \mid \mathbf{Y}_{obs}, \boldsymbol{\psi}) \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis} | \boldsymbol{\theta})\, d\mathbf{Y}_{mis} \\ &= P(\mathbf{R} | \mathbf{Y}_{obs}, \boldsymbol{\psi})\, f(\mathbf{Y}_{obs} \mid \boldsymbol{\theta}) \end{aligned} \tag{69}$$

La ecuación (69) muestra que, cuando el *mecanismo es ignorable*, para realizar inferencias sobre $\boldsymbol{\theta}$, solo es necesario trabajar con $P(\mathbf{Y}_{obs} \mid \boldsymbol{\theta})$ en lugar de $P(\mathbf{Y}_{obs}, \mathbf{R} \mid \boldsymbol{\theta}, \boldsymbol{\psi})$. Es decir, es suficiente trabajar con la función de verosimilitud de los datos observados, ignorando el mecanismo de datos faltantes y, con esto, la falta de datos en si misma. De este modo, se tiene que:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) \propto P(\mathbf{Y}_{obs} \mid \boldsymbol{\theta}) \tag{70}$$

3.5.1 Distribución Exponencial

Sea $Y_1, \ldots, Y_n$; una secuencia de v.a. *i.i.d.* de una distribución exponencial con función de densidad que viene dada por:

$$f(y \mid \lambda) = \lambda \exp(-\lambda y) \tag{71}$$

para $y > 0$; donde $\lambda > 0$ es el parámetro de la distribución.

La función de log-verosimilitud $\ell(\lambda)$ de los datos completos, viene dada por:

$$\begin{aligned} \ell(\lambda) &= \sum_{i=1}^{n} \log f(y_i \mid \lambda) \\ &= \sum_{i=1}^{n} \log \{\lambda \exp(-\lambda y_i)\} \\ &= n \log \lambda - \lambda \sum_{i=1}^{n} y_i \end{aligned} \tag{72}$$

Si una parte de los datos es observada (Y_{obs}) y otra es no observada (Y_{mis}); y, además, el *mecanismo es ignorable*, la función de log-verosimilitud $\ell(\lambda)$ de los datos observados, viene dada por:

$$\begin{aligned} \ell(\lambda) &= \sum_{obs} \log f(y_i \mid \lambda) \\ &= \sum_{obs} \log \{\lambda \exp(-\lambda y_i)\} \\ &= n_{obs} \log \lambda - \lambda \sum_{obs} y_i \end{aligned} \tag{73}$$

El estimador de máxima verosimilitud para λ, se obtiene de resolver, para λ, la siguiente ecuación:

$$\frac{d\ell(\lambda)}{d\lambda} = 0 \tag{74}$$

No es difícil mostrar que la solución de la ecuación (74), corresponde a:

$$\widehat{\lambda} = \frac{1}{\bar{y}_{obs}} \tag{75}$$

donde, $\bar{y}_{obs} = \dfrac{1}{n_{obs}} \sum_{obs} y_i$.

Puede remitirse al Anexo 2.1. Distribución Exponencial, donde se presenta una aplicación para este caso.

3.5.2 Distribución Binomial

Sea $Y_1, \ldots, Y_n$; una secuencia de v.a. *i.i.d.* de una distribución binomial con función de probabilidad que viene dada por:

$$f(y \mid k, \pi) = \binom{k}{y} \pi^y (1-\pi)^{k-y} \tag{76}$$

para $y \in \mathbb{N}$; donde $0 \leq \pi \leq 1$ es la probabilidad de éxito y $k \in \mathbb{N} - \{0\}$ es el número de ensayos Bernoulli.

La función de log-verosimilitud $\ell(k, \pi)$ de los datos completos, viene dada por:

$$\begin{aligned} \ell(k,\pi) &= \sum_{i=1}^{n} \log f(y_i \mid k, \pi) \\ &= \sum_{i=1}^{n} \log \left\{ \binom{k}{y_i} \pi^{y_i} (1-\pi)^{k-y_i} \right\} \\ &= \log \pi \sum_{i=1}^{n} y_i + nk \log(1-\pi) - \log(1-\pi) \sum_{i=1}^{n} y_i \end{aligned} \tag{77}$$

Si una parte de los datos es observada (Y_{obs}) y otra es no observada (Y_{mis}); y, además, el *mecanismo es ignorable*, la función de log-verosimilitud $\ell(k, \pi)$ de los datos observados, viene dada por:

$$\begin{aligned} \ell(\pi) &= \sum_{obs} \log f(y_i \mid k, \pi) \\ &= \sum_{obs} \log \left\{ \binom{k}{y_i} \pi^{y_i} (1-\pi)^{k-y_i} \right\} \\ &= \log \pi \sum_{obs} y_i + nk \log(1-\pi) - \log(1-\pi) \sum_{obs} y_i \end{aligned} \tag{78}$$

Si se asume que se conoce el valor del parámetro k; de las ecuaciones (32 y 78), se tiene que el estimador para π, se obtiene de resolver, para π, la siguiente ecuación:

$$\frac{d\ell(\pi)}{d\pi} = 0 \tag{79}$$

No es difícil mostrar que la solución de la ecuación (79), corresponde a:

$$\widehat{\pi} \quad = \quad \frac{\bar{y}_{obs}}{k} \tag{80}$$

donde, $\bar{y}_{obs} = \dfrac{1}{n_{obs}} \sum_{obs} y_i$.

Puede remitirse al Anexo 2.2. Distribución Binomial, donde se presenta una aplicación para este caso.

3.6 Algoritmo de Esperanza-Maximización

El *Algoritmo de Esperanza-Maximización*[31] (EM), presenta una técnica iterativa general para realizar Estimación por Máxima Verosimilitud (MLE) de parámetros, en problemas que pueden presentarse como uno de datos *incompletos*. El algoritmo EM puede utilizarse incluso con datos completos pero su mayor utilidad ocurre cuando la función a optimizar es difícil de abordar, en general, debido a la incompletitud de los datos. En términos muy simples, el algoritmo busca iterativamente maximizar la función de verosimilitud $L(\boldsymbol{\theta})$ con respecto a $\boldsymbol{\theta}$.

Denotemos $\boldsymbol{\theta}^{(t)}$ el estimador maximizado en la iteración t, para $t = 0, 1, \ldots$. Se define la función $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$ como la *esperanza* de la función de log-verosimilitud conjunta de los datos completos $\mathbf{Y}$; condicional en los datos observados $\mathbf{Y}_{obs} = \mathbf{y}_{obs}$. Es decir:

$$\begin{aligned} Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) &= \mathbb{E}\left(\log L(\boldsymbol{\theta}) \mid \mathbf{y}_{obs}, \boldsymbol{\theta}^{(t)}\right) \\ &= \mathbb{E}\left(\log f_Y(\mathbf{y} \mid \boldsymbol{\theta}) \mid \mathbf{y}_{obs}, \boldsymbol{\theta}^{(t)}\right) \\ &= \int \log f_Y(\mathbf{y} \mid \boldsymbol{\theta})\, f_{Y_{mis} \mid Y_{obs}}\left(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\theta}^{(t)}\right) d\mathbf{Y}_{mis} \end{aligned} \tag{81}$$

El algoritmo EM es inicializado para $\boldsymbol{\theta}^{(0)}$, luego se alterna entre dos pasos: E de *Esperanza* y M de *Maximización*. El algoritmo EM se resume como sigue:

1. **Paso E:** Se calcula $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$.
2. **Paso M:** Se maximiza $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$ con respecto a $\boldsymbol{\theta}$. Se establece $\boldsymbol{\theta}^{(t+1)}$ igual al estimador que maximiza Q.
3. Se regresa al paso E, a menos que algún criterio de parada haya sido alcanzado[32].

3.6.1 Distribución Exponencial

Considerando la función de log-verosimilitud de los datos completos dada en la ecuación (72):

$$\ell(\lambda) \quad = \quad n \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

[31] Para un detalle completo del *Algoritmo de Esperanza-Maximización* (EM), se puede revisar (Dempster, Laird, and Rubin 1977).

[32] En el presente caso, los criterios de parada se construyen en base a: $\left(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\right)^{\intercal} \left(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\right)$ o $|Q\left(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}\right) - Q\left(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}\right)|$.

La función $Q\left(\lambda \mid \lambda^{(t)}\right)$ viene dada por:

$$\begin{aligned} Q\left(\lambda \mid \lambda^{(t)}\right) &= \mathbb{E}\left(\ell\left(\lambda\right) \mid \mathbf{y}_{obs}, \lambda^{(t)}\right) \\ &= \mathbb{E}\left(n\log\lambda - \lambda\sum_{i=1}^{n} y_i \mid \mathbf{y}_{obs}, \lambda^{(t)}\right) \\ &= n\log\lambda - \lambda\sum_{obs} y_i - n_{mis}\frac{\lambda}{\lambda^{(t)}} \end{aligned} \tag{82}$$

Ahora se debe maximizar la función $Q\left(\lambda \mid \lambda^{(t)}\right)$, descrita en la ecuación (82). Es decir, se debe resolver:

$$\frac{dQ\left(\lambda \mid \lambda^{(t)}\right)}{d\lambda} = 0 \tag{83}$$

No es difícil mostrar que la solución de la ecuación (83), corresponde a:

$$\lambda^{*} = \frac{n\lambda^{(t)}}{\lambda^{(t)}\sum_{obs} y_i + n_{mis}} \tag{84}$$

Luego, se tiene que:

$$\lambda^{(t+1)} = \frac{n\lambda^{(t)}}{\lambda^{(t)}\sum_{obs} y_i + n_{mis}} \tag{85}$$

Puede remitirse al Anexo 2.3. Distribución Exponencial, donde se presenta una aplicación para este caso.

3.6.2 Distribución Binomial

Considerando la función de log-verosimilitud de los datos completos dada en la ecuación (78):

$$\ell\left(\pi\right) = \log\pi\sum_{i=1}^{n} y_i + nk\log\left(1-\pi\right) - \log\left(1-\pi\right)\sum_{i=1}^{n} y_i$$

La función $Q\left(\pi \mid \pi^{(t)}\right)$ viene dada por:

$$\begin{aligned} Q\left(\pi \mid \pi^{(t)}\right) &= \mathbb{E}\left(\ell\left(\pi\right) \mid \mathbf{y}_{obs}, \pi^{(t)}\right) \\ &= \mathbb{E}\left(\log\pi\sum_{i=1}^{n} y_i + nk\log\left(1-\pi\right) - \log\left(1-\pi\right)\sum_{i=1}^{n} y_i \mid \mathbf{y}_{obs}, \pi^{(t)}\right) \\ &= \log\left(\frac{\pi}{1-\pi}\right)\sum_{obs} y_i + \log\left(\frac{\pi}{1-\pi}\right) n_{mis}k\pi^{(t)} + nk\log\left(1-\pi\right) \end{aligned} \tag{86}$$

Ahora se debe maximizar la función $Q\left(\pi \mid \pi^{(t)}\right)$, descrita en la ecuación (86). Es decir, se debe resolver:

$$\frac{dQ\left(\pi \mid \pi^{(t)}\right)}{d\pi} = 0 \tag{87}$$

No es difícil mostrar que la solución de la ecuación (87), corresponde a:

$$\pi^* = \frac{\sum_{obs} y_i + n_{mis} k \pi^{(t)}}{nk} \tag{88}$$

Luego, se tiene que:

$$\pi^{(t+1)} = \frac{\sum_{obs} y_i + n_{mis} k \pi^{(t)}}{nk} \tag{89}$$

Puede remitirse al Anexo 2.4. Distribución Binomial, donde se presenta una aplicación para este caso.

4 Aplicación de Métodos basados en la Función de Verosimilitud

4.1 Datos Categóricos

Modelo GLM datos categóricos.

Puede remitirse al Anexo 3.1. GLM - Datos Categóricos, donde se presenta una aplicación para este caso.

4.2 Datos Discretos

Modelo GLM datos discretos.

Puede remitirse al Anexo 3.2. GLM - Datos Discretos, donde se presenta una aplicación para este caso.

4.3 Datos Continuos

Modelo GLM datos continuos.

Puede remitirse al Anexo 3.3. GLM - Datos Continuos, donde se presenta una aplicación para este caso.

4.4 Mezclas de Distribuciones

Modelo mezclas de distribuciones (posiblemente).

5 Estimación Bayesiana con Datos Incompletos

5.1 Estimación Bayesiana: Conceptos básicos

5.2 Métodos Bayesianos con Datos Incompletos: Marco teórico general

Anexos

Anexo 1. Datos Completos

En los siguientes anexos ....

Anexo 1.1. Distribución Weibull

En el Anexo.

Anexo 1.2. Distribución Gamma

En este anexo.

Anexo 1.3. Distribución Weibull

En este anexo.

Anexo 2. Datos Incompletos

En los siguientes anexos ....

Anexo 2.1. Distribución Exponencial

En este anexo.

Anexo 2.2. Distribución Binomial

En este anexo.

Anexo 2.3. Distribución Exponencial

En este anexo.

Anexo 2.4. Distribución Binomial

En este anexo.

Anexo 3. Datos Incompletos

En los siguientes anexos ....

Anexo 3.1. GLM - Datos Categóricos

En este anexo.

Anexo 3.2. GLM - Datos Discretos

En este anexo.

Anexo 3.3. GLM - Datos Continuos

En este anexo.

Bibliografía

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22.

Dobson, Annette J., and Adrian G. Barnett. 2018. *An Introduction to Generalized Linear Models*. Texts in Statistical Science Series. CRC Press.

Dunn, Peter K., and K. Smyth Gordon. 2018. *Generalized Linear Models with Examples in R*. Springer Texts in Statistics. Springer.

Enders, Craig K. 2022. *Applied Missing Data Analysis*. Guilford Publications.

He, Yulei, Guangyu Zhang, and Chiu-Hsieh Hsu. 2021. *Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies*. CRC Press.

Lindsey, James. K. 1997. *Applying Generalized Linear Models*. Springer Texts in Statistics. Springer.

Little, Roderick J. A., and Donald B. Rubin. 2020. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. Chapman; Hall, London.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke. 2015. *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society* A135 (3): 370–84.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. CRC press.

Tan, Ming T., Guo L. Tian, and Kai W. Ng. 2010. *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*. Chapman & Hall/CRC Biostatistics Series. CRC Press.

Van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press.