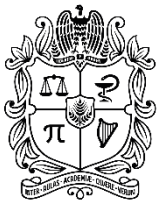

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Estimación simultánea de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables

Jessica Nathaly Pulzara Mora

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2022

Estimación simultánea de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables

Jessica Nathaly Pulzara Mora

Tesis presentada como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director:
Juan Carlos Correa Morales Ph.D., en estadística

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2022

Dedicatoria

Al $\boldsymbol{A}$ *y la* $\boldsymbol{\Omega}$*, el mismo de ayer, hoy y siempre.*
Totus tuus ego sum, et omnia mea tua sunt.
Accipio te in mea omnia. Praebe mihi cor tuum, Maria!

Agradecimientos

A mis padres y a mi hermano por todo el apoyo que me brindaron, por el tiempo que se tomaron para escucharme, darme una palabra de aliento, y compartirme sus experiencias y enseñanzas.

A todas las personas que me apoyaron a través de lo académico, moral, emocional y/o espiritual.

Al profesor Juan Carlos Correa por su paciencia, por compartirme de su conocimiento, experiencia y en especial por la oportunidad de trabajar con él.

A todos los profesores que hicieron parte de mi formación académica, profesional y personal.

Y por último pero no menos importante a Dios, por haber soñado esto para mí y haber permitido todas las condiciones necesarias para poder llegar a este punto.

Resumen

Estimación simultánea de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables

La sensibilidad y especificidad son utilizadas para evaluar el rendimiento de pruebas de diagnóstico en áreas de la medicina tales como epidemiología, psicología, genética, y también existen aplicaciones en finanzas y agronomía. La sensibilidad señala la proporción de casos positivos que son bien detectados por la prueba, en otras palabras, la sensibilidad mide la efectividad de la prueba cuando se usa en individuos positivos, mientras que la especificidad señala la proporción de casos negativos que son bien detectados por la prueba, es decir, mide la efectividad de la prueba cuando se usa en individuos negativos. Para la estimación de ambas cantidades varios autores han propuesto diferentes métodos tales como la prueba "estándar de oro", aproximación bayesiana, máxima verosimilitud, o por medio de modelos logísticos. Sin embargo, éstas pruebas solo dan estimaciones de tipo marginal.
En el presente trabajo se desarrolla un procedimiento para estimar de forma simultánea la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables.

Palabras clave: Pruebas diagnósticas, sensibilidad, especificidad, estadística, metodología GSK.

Abstract

Simultaneous estimation of sensitivity and specificity using the GSK methodology in the presence of covariates

Sensitivity and specificity are used to evaluate the performance of diagnostic tests in areas of medicine such as epidemiology, psychology, genetics, and there are also applications in finance and agronomy. Sensitivity indicates the proportion of positive cases that are well detected by the test, in other words, it measures how the test is effective when used on positive individuals, while specificity indicates the proportion of negative cases that are well detected by the test; that is, it measures how the test is effective when used on negative individuals. For the estimation of both quantities several authors have proposed different methods such as the "gold standard" test, Bayesian approximation, maximum likelihood, or by means of logistic models. However, these tests only give marginal estimates.
In this paper, a procedure is developed to simultaneously estimate sensitivity and specificity using the GSK methodology in the presence of covariates.

Keywords: Diagnostic tests, sensitivity, specificity, statistics, GSK methodology.

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

La sensibilidad y la especificidad son las medidas estadísticas de rendimiento de una prueba de clasificación binaria introducidas por el bioestadístico estadounidense Jacob Yerushalmy (1947). La sensibilidad mide la proporción de positivos reales que se clasifican como tales (por ejemplo, el porcentaje de personas enfermas que se identifican como que tienen la afección); y la especificidad mide la proporción de negativos que se identifican correctamente (por ejemplo, el porcentaje de personas sanas identificadas como que no tienen la afección) (Sharma et al., 2009). En otras palabras, la sensibilidad se refiere a la probabilidad de que se muestre verdadero lo que es verdadero y la especificidad a la probabilidad de que aparezca falso lo que es falso.

En la investigación epidemiológica, los estudios de validación a menudo tienen como objetivo la determinación de la sensibilidad y especificidad de una "prueba" para detectar la presencia de un factor de riesgo (Tosteson et al., 1994).

La sensibilidad y la especificidad de una prueba de detección pueden estimarse a partir de las frecuencias y los totales marginales de una tabla de contingencia de dos por dos definida por la presencia o ausencia de la enfermedad de interés, como lo revela el "estándar de oro" y los resultados de las pruebas de detección (Coughlin et al., 1992). Sin embargo, debido a que la sensibilidad y la especificidad son estimados a partir del mismo estudio, estos parámetros deben ser estimados simultáneamente, permitiendo la realización de inferencias de funciones de estos parámetros que consideren la correlación entre estas dos medidas.

Después de realizar una extensa búsqueda bibliográfica en el estado del arte acerca de la estimación de la sensibilidad y la especificidad de manera conjunta determinamos que este tema no ha sido muy explorado en la literatura ya que se encontró únicamente la investigación de Puggioni et al. (2008). De aquí nuestro interés por desarrollar una alternativa para estimar ambas medidas de manera simultanea y adicionalmente incluir covariables. La metodología GSK, es una metodología de carácter general para el análisis de tablas de conteo que permite respuestas correlacionadas y no requiere varianza constante (Correa, 2016). Por tanto, en esta tesis se propone el desarrollo de un método utilizando la metodología GSK, que permite resolver el problema de la estimación simultánea de la sensibilidad y especificidad y admite la inclusión de covariables de manera directa, ilustrado con datos pseudo reales y una simulación para analizar el efecto que el tamaño de la muestra tiene en la estimación de los

parámetros del modelo.

En el desarrollo de este trabajo inicialmente se presenta una breve revisión de algunos aspectos teóricos asociados al procedimiento diagnóstico estándar de oro, tabla de confusión, distancia de Kullback-Leibler, y la metodología GSK. En el capítulo 3 se desarrolla la propuesta para la estimación simultánea de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables. En el capítulo 4 se ilustran las propuestas planteadas en el capitulo 3 con datos pseudo-reales. El capítulo 5 contiene un estudio de simulación diseñado para determinar el tamaño de muestra mínimo tal que se puedan garantizar los resultados de la metodología, dado que está se basa en distribuciones normales asintóticas, así como el efecto que tiene el tamaño de la muestra en la estimación de los parámetros del modelo, en la sensibilidad y la especificidad utilizando la metodología GSK cuando la función respuesta se plantea específicamente con una de las funciones propuestas en el capítulo 3. Finalmente, en el capítulo 6 se presentan nuestras principales conclusiones y recomendaciones.

2. Marco Teórico

La investigación sobre pruebas diagnósticas tiene como objetivos estimar la capacidad discriminatoria de una prueba diagnóstica entre enfermos y no enfermos (sensibilidad-especificidad), determinar el rendimiento de la misma (valores predictivos) o evaluar la utilidad y satisfacción de un procedimiento diagnóstico. La utilidad de una prueba diagnóstica depende de su capacidad de producir los mismos resultados cada vez que se aplica en similares condiciones (fiabilidad) y de que sus mediciones reflejen exactamente el fenómeno que se intenta medir (validez o exactitud), pero también de su rendimiento clínico y de su coste (Ochoa et al., 2007).

Toda investigación sobre pruebas diagnósticas parte del conocimiento de que una enfermedad existe o no en un grupo de individuos de una muestra. Se necesita, pues, algo que defina este punto. Concretamente, se precisa una prueba diagnóstica suficientemente acreditada en ese momento que puntualice la existencia real de la enfermedad. Es lo que se llama prueba patrón de referencia, estándar de oro o gold standard, que a veces es una sola prueba, una serie de pruebas, el resultado del seguimiento de los casos, entre otros. No siempre este patrón es completamente fiable, como se desearía. A él se enfrenta la prueba diagnóstica en evaluación que suele llevar el nombre de test. Para definir el resultado de un test es preciso aplicar un criterio diagnóstico, cuya menor o mayor claridad y facilidad de interpretación son también cruciales para el resultado de la evaluación. Muchas veces, tanto el patrón como el test presentarán un resultado dicotómico, positivo o negativo, con lo cual su enfrentamiento se podrá resumir en una tabla de contingencia 2×2 (2 filas y 2 columnas), que toma el nombre de tabla de confusión y se muestra en la 3.1. La prueba patrón suele colocarse en columnas y el test en filas, para mayor claridad. La combinación de positivos y negativos del patrón y del test configura los verdaderos y los falsos resultados del propio test. La combinación por columnas de verdaderos y falsos da lugar a los índices diagnósticos fundamentales que definen a un test: la sensibilidad y la especificidad (Escrig-Sos et al., 2006).

2.1. Estándar de oro

El diccionario inglés Oxford establece que es una "medida a la que otros se ajustan o por la cual se juzga la precisión de los demás ... lo que sirve como base para la comparación". El estándar de oro ("Gold Standard") no es la prueba perfecta, sino simplemente la mejor prueba disponible (Versi, 1992).

El estándar de oro es un término para el procedimiento de diagnóstico más definitivo (p. ej. examen microscópico de una muestra de tejido), o la mejor prueba de laboratorio disponible (p. ej. anticuerpos séricos contra el VIH). A veces puede referirse a una evaluación clínica integral (p. ej. evaluación clínica de la artritis) (Alexander et al., 2014).
Se hace referencia al término "ausencia de un estándar de oro", cuando se desconoce el verdadero estado de la enfermedad (no infectado o infectado) (Engel et al., 2006).

2.2. Tabla de confusión

La matriz de confusión es una tabla con dos dimensiones, "Real" y "Prueba", y conjuntos de clases en ambas dimensiones. Los resultados del estándar de oro que se utilizan para definir la dimensión "Real" de la tabla suele colocarse en columnas y la "Prueba" en filas.

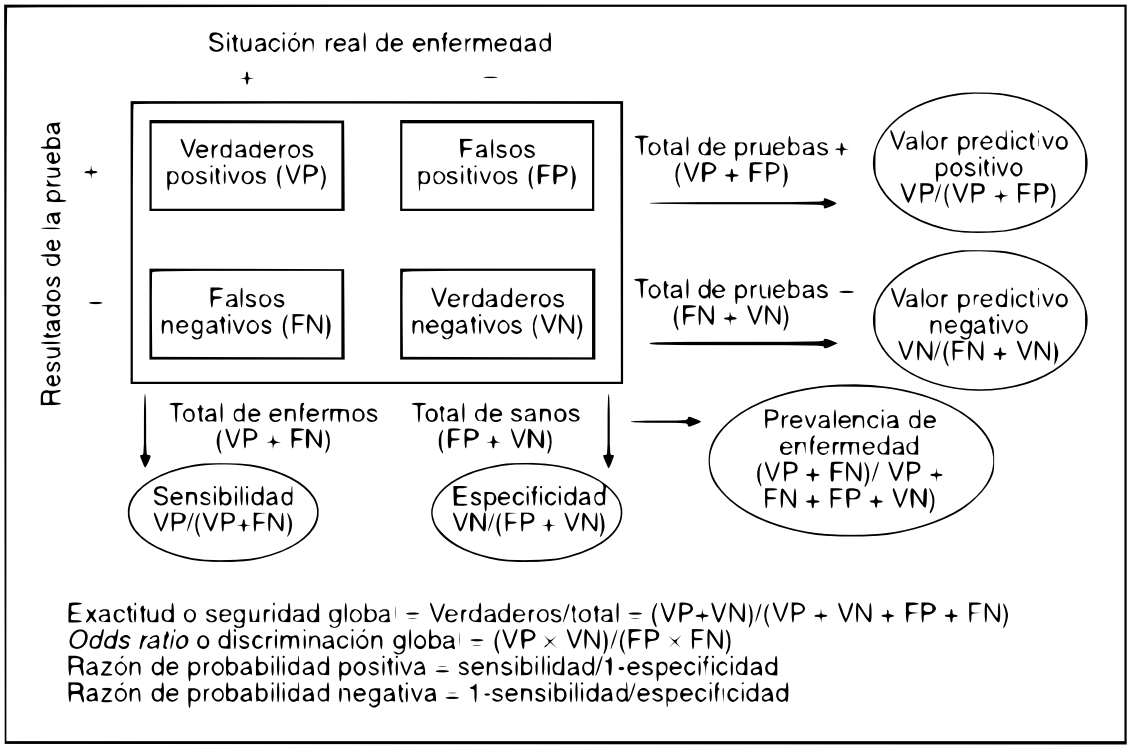

Figura 2.1.: Tabla de confusión. Tomado de Escrig-Sos et al. (2006)

Según Lalkhen and McCluskey (2008):

- **Verdadero Positivo:** el paciente tiene la enfermedad y el resultado de la prueba es positivo.
- **Falso Positivo:** el paciente no tiene la enfermedad pero el resultado de la prueba es positivo.

- **Verdadero Negativo:** el paciente no tiene la enfermedad y el resultado de la prueba es negativo.
- **Falso Negativo:** el paciente tiene la enfermedad pero el resultado de la prueba es negativo.
- **Sensibilidad:** probabilidad de tomar la decisión correcta cuando una persona tiene la enfermedad. Se estima como:

$$Sensibilidad = \frac{VP}{VP + FN} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

- **Especificidad:** probabilidad de tomar la decisión correcta cuando una persona no tiene la enfermedad. Se estima como:

$$Especificidad = \frac{VN}{VN + FP} = \frac{\text{Verdaderos Negativos}}{\text{Verdaderos Negativos} + \text{Falsos Positivos}}$$

2.2.1. Intervalos de confianza para la sensibilidad y la especificidad

La precisión de la sensibilidad y especificidad se determina por medio de intervalos de confianza, colocando la estimación en un rango de valores coherentes con los datos. Cuanto más estrecho sea el intervalo de confianza, más precisa será la estimación.
Se debe informar la precisión de cualquier estimación de sensibilidad y especificidad, sin importar el valor, para evitar engaños sobre los resultados. Se recomienda que todas las estimaciones de sensibilidad y especificidad se informen con intervalos de confianza del 95 %. La fórmula para intervalos de confianza al 95 % está dada por (Hess et al., 2012):

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde $\hat{p}$ es la estimación de la sensibilidad o de la especificidad, n es el número de muestras de verdaderos positivos (para la sensibilidad) o el número de muestras de verdaderos negativos (para la especificidad). La fórmula es apropiada siempre que $n\hat{p}$ y $n(1-\hat{p})$ no sean menores que 5.

2.3. Metodología GSK

La metodología GSK fue desarrollada por Grizzle, Starmer y Koch (Grizzle et al., 1969). Esta metodología se basa en aplicaciones del modelo lineal general (la base tanto para el análisis de regresión como para el análisis de varianza para datos continuos) a datos categóricos (Forthofer, 2012) y permite la estimación de parámetros en modelos lineales generalizados,

como es el caso de modelos con variable respuesta de distribución multinomial. Es una metodología flexible y poderosa para aplicar en diversas situaciones relacionadas con el manejo de tablas de contingencia.

La metodología GSK puede definirse en tres pasos: 1) La definición de la variable dependiente, la cual no se refiere en sí a individuos sino a probabilidades o funciones de probabilidad, 2) La definición del modelo el cual depende del esquema de muestreo que se asume, si solo se tiene una variable respuesta, o si hay variables independientes o fijas, las cuales definirán estratos y se deberá considerar la construcción de una matriz de diseño, 3) La estimación y validación del modelo donde se debe asegurar un tamaño de muestra para poder garantizar resultados asintóticos (Correa, 2016). La justificación teórica del método se basa en lo descrito por Neyman (1949) y Wald (1943).

El modelamiento de variables categóricas utilizando la metodología GSK puede utilizar funciones de la variable dependiente del modelo, cuya variabilidad queremos explicar, llamada función de respuesta. Para el caso de variables respuesta con distribución multinomial la función de respuesta se basa en las probabilidades de la variable respuesta o en las tasas de ocurrencia. Los objetivos son que esta función consiga linealizar la relación entre el predictor lineal y la esperanza de la variable respuesta, facilite la estimación de los parámetros ampliando un espacio paramétrico restringido hacia los reales, y debe ser interpretable.

Poblaciones (factores)	1	2	...	J	Total
1	n_{11}	n_{12}	...	n_{1J}	n_1
2	n_{21}	n_{22}	...	n_{2J}	n_2
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	...	n_{IJ}	n_I

Tabla 2.1.: Distribución de Frecuencias

Poblaciones (factores)	1	2	...	J	Total
1	π_{11}	π_{12}	...	π_{1J}	π_1
2	π_{21}	π_{22}	...	π_{2J}	π_2
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	...	π_{IJ}	π_I

Tabla 2.2.: Tabla de Probabilidades

Para la formación de las funciones de la variable respuesta se parte de la tabla de frecuencias de un conjunto de datos hipotéticos, donde se tienen J categorías con I factores o poblaciones diferentes que permiten obtener la Tabla 2.1, donde cada celda de la tabla corresponde

al conteo de datos pertenecientes a la i-ésima población en la j-ésima categoría de respuesta.

A partir de la tabla de frecuencias se genera la Tabla 2.2, en la cual las celdas representan las proporciones en lugar de los conteos, donde π_{ij} es la probabilidad de que un sujeto de la i-ésima población tenga la j-ésima categoría. Así se tienen variables respuesta multinomiales a partir de las probabilidades de cada celda de la tabla de contingencia.

Sea $\boldsymbol{\pi'} = [\boldsymbol{\pi'_1}, \ldots, \boldsymbol{\pi'_I}]$ donde $\boldsymbol{\pi'_i} = [\pi_{i1}, \pi_{i2}, \ldots, \pi_{iJ}]$ es un vector que denota la distribución de Y en el nivel i, se tiene así que a partir de las probabilidades de cada celda de la tabla de contingencia se define a $f(\boldsymbol{\pi})$ como una función de los elementos de $\boldsymbol{\pi}$, estos elementos tienen derivadas continuas hasta el segundo orden con respecto a π_{ij}, y $f(\boldsymbol{\pi})$ es un vector con u funciones de respuesta, $u \leq (J-1)\, I$. Así se tiene $[f(\boldsymbol{\pi})]' = [f_1(\boldsymbol{\pi}), \ldots, f_u(\boldsymbol{\pi})]$ Grizzle et al. (1969).

A partir de las respuestas observadas, se tiene el vector $\hat{\boldsymbol{\pi}} = [\hat{\boldsymbol{\pi}}_{\mathbf{1}}, \ldots, \hat{\boldsymbol{\pi}}_{\boldsymbol{i}}]$ donde $\hat{\boldsymbol{\pi}}'_{\boldsymbol{i}} = (\hat{\pi}_{i1}, \ldots, \hat{\pi}_{iJ})$. Se denota $\hat{\pi}_{ij} = \frac{n_{ij}}{n_i}$ la proporción muestral es decir la frecuencia observada de la i-ésima subpoblación perteneciente a la j-ésima categoría sobre n_i que es el total de observaciones de la i-ésima subpoblación y $E[\hat{\pi}_{ij}] = \pi_{ij}$. Se tiene así que $[f(\hat{\boldsymbol{\pi}})]' = [f_1(\hat{\boldsymbol{\pi}}), \ldots, f_u(\hat{\boldsymbol{\pi}})]$ denota el vector de funciones respuesta muestrales.

La matriz de varianzas y covarinzas de $\boldsymbol{\pi}$ es $\boldsymbol{V}$ una matriz $IJ \times IJ$:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_1} & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \boldsymbol{V_2} & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{V_i} & \ldots & 0 \\ \vdots & \vdots & \ldots & 0 & \ddots & 0 \\ 0 & 0 & \ldots & 0 & \ldots & \boldsymbol{V_I} \end{bmatrix} \tag{2.1}$$

$\boldsymbol{V}(\hat{\boldsymbol{\pi}}_{\boldsymbol{i}})$ es el estimado de $\boldsymbol{V}(\boldsymbol{\pi}_{\boldsymbol{i}})$:

$$\boldsymbol{V_i} = \boldsymbol{V}(\boldsymbol{\pi_i}) = \frac{1}{n_i} \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \ldots & -\pi_{i1}\pi_{iJ} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \ldots & -\pi_{i2}\pi_{iJ} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_{iJ}\pi_{i1} & -\pi_{iJ}\pi_{i2} & \ldots & \pi_{iJ}(1-\pi_{iJ}) \end{bmatrix} \tag{2.2}$$

$\boldsymbol{V}(\hat{\boldsymbol{\pi}})$ es una matriz bloque diagonal con $\boldsymbol{V}(\hat{\boldsymbol{\pi}}_{\boldsymbol{i}})$ en la diagonal principal.

La matriz de covarianza asintótica de $f(\hat{\boldsymbol{\pi}})$ depende de la matriz $\boldsymbol{H}$ de dimensiones $u \times IJ$:

$$H = \left[\frac{\partial f_m(\boldsymbol{\pi})}{\partial \pi_{ij}} \middle| \pi_{ij} = \hat{\pi}_{ij} \right]' \tag{2.3}$$

para $m = 1, \ldots, u$ con m la m-ésima función f construida y todas la IJ combinaciones (i, j). La matriz de covarianzas asintótica de $f(\hat{\boldsymbol{\pi}})$ es $\boldsymbol{V_f} = \boldsymbol{HVH'}$, y la versión muestral de esta matriz de covarianzas es $\hat{\boldsymbol{V}}_{\boldsymbol{f}}$, la cual se obtiene substituyendo las proporciones muestrales en la matriz $\boldsymbol{V}$ y $\boldsymbol{H}$ (Agresti, 1996a)(Grizzle et al., 1969).

Se asume que $f(\boldsymbol{\pi}) = \boldsymbol{X\beta}$, donde $\boldsymbol{X}$ es una matriz de diseño conocida $u \times v$ y $\boldsymbol{\beta}$ es el vector de parámetros $v \times 1$(Agresti, 1996a). Si el modelo hipotético ajusta los datos, la mejor estimación asintótica normal de $\boldsymbol{\beta}$ está dada por $\boldsymbol{b}$ cuando éste es el vector que minimiza

$$\left(\hat{\boldsymbol{f}} - \boldsymbol{X\beta}\right)' \hat{\boldsymbol{V}}_{\hat{\boldsymbol{f}}}^{-\mathbf{1}} \left(\hat{\boldsymbol{f}} - \boldsymbol{X\beta}\right)$$

El estimado de $\boldsymbol{\beta}$ es

$$\boldsymbol{b} = \left(\boldsymbol{X'}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{f}}}^{-\mathbf{1}}\boldsymbol{X}\right)^{-1} \boldsymbol{X'}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{f}}}^{-\mathbf{1}}\hat{\boldsymbol{f}}$$

La prueba de bondad de ajuste del modelo usa el termino de residual

$$\hat{\boldsymbol{f}}'\hat{\boldsymbol{V}}_{\hat{\boldsymbol{f}}}^{-\mathbf{1}}\boldsymbol{f} - \boldsymbol{b}'(\boldsymbol{X}'\hat{\boldsymbol{V}}_{\hat{\boldsymbol{f}}}^{-\mathbf{1}}\boldsymbol{X})\boldsymbol{b},$$

el cual compara las funciones respuestas muestrales con los valores del modelo predicho. Bajo la hipótesis nula $H_0 : f(\boldsymbol{\pi}) - \boldsymbol{X\beta} = 0$ el estadístico es asintóticamente χ^2 con $u \times v$ grados de libertad (Agresti, 1996a).

Las hipótesis sobre los contrastes y otros efectos de las variables explicativas tiene la forma $H_0 : \boldsymbol{C\beta} = 0$, la cual es producida por métodos convencionales de regresión múltiple ponderada donde C es una matriz (d x v) de constantes arbitrarias de rango completo $d \leq v$. El estadístico está dado por

$$W_c = \boldsymbol{b'C'}[\boldsymbol{C}(\boldsymbol{X'}\hat{\boldsymbol{V}}_{\hat{\boldsymbol{f}}}^{-\mathbf{1}}\boldsymbol{X}^{-\mathbf{1}}\boldsymbol{C'}]^{-1}\boldsymbol{Cb},$$

el cual tiene asintóticamente una distribución χ^2 con d grados de libertad si H_0 es cierta (Agresti, 1996a)

2.4. Distancia de Kullback-Leibler

Sea X una variable aleatoria que toma valores continuos en el intervalo Ω y sean $f(x)$ y $g(x)$ las densidades de probabilidad de dos procesos aleatorios. Se define la entropía relativa de g respecto a f, el número de Kullback-Leibler o la discriminación entre los dos procesos como:

$$H(f||g) = \int_{x\in\Omega} f(x) \log \frac{f(x)}{g(x)} dx$$

La entropía relativa entre dos distribuciones de probabilidad es no negativa, siendo nula unicamente si las dos distribuciones son idénticas. De este modo, se puede considerar como una medida de la divergencia entre dos distribuciones de probabilidad (Ramırez et al., 2002).

La divergencia de Kullback-Leibler (KL) es una medida en estadística que cuantifica en bits qué tan cerca está una distribución de probabilidad $p(x)$ de una distribución modelo (o candidata) $q(x)$. Si las distribuciones de probabilidad $p(x)$ y $q(x)$ pertenecen a una variable aleatoria discreta su divergencia KL se define como

$$D(p||q) = \sum_{x\in\chi} p(x) \log \frac{p(x)}{q(x)}$$

La entropía relativa siempre es no negativa y es cero si y sólo si $p = q$. Sin embargo, no es una verdadera distancia entre distribuciones ya que no es simétrica y no satisface la desigualdad triangular (Cover and Thomas, 2006).

3. Estimación simultánea de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables

En diferentes ramas de las ciencias se proponen pruebas que permiten identificar la presencia o ausencia de una condición de interés. Sin embargo, estas pruebas deben ser contrastadas con la realidad del suceso con el fin de corroborar los resultados y así determinar si la prueba puede considerarse como fiable. Para este fin se emplean dos medidas conocidas como sensibilidad y especificidad. Estas son consideradas bajo condiciones específicas, teniendo en cuenta el hecho de que la prueba puede ser influenciada por variables externas. En este capítulo, se plantea la estimación de ambas medidas utilizando la metodología GSK.

3.1. Organización de los datos

En la métodología GSK las observaciones objeto de estudio son clasificadas en tablas de contingencia 2×2 por subpoblaciones. En nuestro caso, las tablas de contingencia son tablas de confusión, es decir, se relaciona el resultado de la prueba (positivo o negativo) con la condición o realidad del suceso (positivo o negativo) como se muestra en la tabla 3.1, mientras que las subpoblaciones son la combinación entre los niveles pertenecientes a las m variables regresoras. Cada variable X_k con $k = 1, 2, \ldots, m$, tiene un número determinado P_k de niveles, los cuales se representan por ${C_1}^{(k)}, {C_2}^{(k)}, \ldots, C^{(k)}_{P_k-1}, C^{(k)}_{P_k}$, donde el superíndice k hace referencia a la covariable a la que pertenece y en total se generan $I = P_1 \times P_2 \times \cdots \times P_k$ combinaciones que corresponden a las diferentes subpoblaciones como se muestra en la Tabla 3.2.

Los datos se estructuran en una tabla por subpoblaciones, categorías y el total de observaciones (3.2). Las categorías son la combinación entre una prueba positiva (P^+) o negativa (P^-) con la condición positiva (C^+) o negativa (C^-), para un total de 4 categorías que se representan como: P^+C^+, P^+C^-, P^-C^+, P^-C^-. A cada categoría le corresponde un número de individuos $n^{(i)}_{lq}$ (previamente clasificados en la tabla 3.1) donde el subíndice l hace referencia a la condición positiva ($l = 1$) o negativa ($l = 2$), el subíndice q hace referencia a la prueba

		Prueba		
		+	-	
Condición	+	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$n_{1+}^{(i)}$
	-	$n_{21}^{(i)}$	$n_{22}^{(i)}$	$n_{2+}^{(i)}$
		$n_{+1}^{(i)}$	$n_{+2}^{(i)}$	n_i

Tabla 3.1.: Tabla de confusión para la i-ésima subpoblación

positiva ($q = 1$) o negativa ($q = 2$) y el superíndice hace referencia a la subpoblación a la que pertenece con $i = 1, 2, \ldots, I$.

Subpoblación				Categoría				
X_1	...	X_{k-1}	X_k	P^+C^+	P^-C^+	P^+C^-	P^-C^-	Total
$C_1^{(1)}$	...	$C_1^{(k-1)}$	$C_1^{(k)}$	$n_{11}^{(1)}$	$n_{12}^{(1)}$	$n_{21}^{(1)}$	$n_{22}^{(1)}$	n_1
$C_1^{(1)}$	...	$C_1^{(k-1)}$	$C_2^{(k)}$	$n_{11}^{(2)}$	$n_{12}^{(2)}$	$n_{21}^{(2)}$	$n_{22}^{(2)}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_1^{(1)}$	...	$C_1^{(k-1)}$	$C_{P_k}^{(k)}$	$n_{11}^{(P_k)}$	$n_{12}^{(P_k)}$	$n_{21}^{(P_k)}$	$n_{22}^{(P_k)}$	n_{P_k}
$C_1^{(1)}$	...	$C_2^{(k-1)}$	$C_1^{(k)}$	$n_{11}^{(P_{k+1})}$	$n_{12}^{(P_{k+1})}$	$n_{21}^{(P_{k+1})}$	$n_{22}^{(P_{k+1})}$	$n_{P_{k+1}}$
$C_1^{(1)}$	...	$C_2^{(k-1)}$	$C_2^{(k)}$	$n_{11}^{(P_{k+2})}$	$n_{12}^{(P_{k+2})}$	$n_{21}^{(P_{k+2})}$	$n_{22}^{(P_{k+2})}$	$n_{P_{k+2}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_1^{(1)}$	...	$C_2^{(k-1)}$	$C_{P_k}^{(k)}$	$n_{11}^{(2P_k)}$	$n_{12}^{(2P_k)}$	$n_{21}^{(2P_k)}$	$n_{22}^{(2P_k)}$	n_{2P_k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_1^{(1)}$	...	$C_{P_{k-1}}^{(k-1)}$	$C_1^{(k)}$	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$n_{21}^{(i)}$	$n_{22}^{(i)}$	n_i
$C_1^{(1)}$	...	$C_{P_{k-1}}^{(k-1)}$	$C_2^{(k)}$	$n_{11}^{(i+1)}$	$n_{12}^{(i+1)}$	$n_{21}^{(i+1)}$	$n_{22}^{(i+1)}$	n_{i+1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_{P_1}^{(1)}$	...	$C_{P_{k-1}}^{(k-1)}$	$C_{P_k}^{(k)}$	$n_{11}^{(I)}$	$n_{12}^{(I)}$	$n_{21}^{(I)}$	$n_{22}^{(I)}$	n_I

Tabla 3.2.: Tabla de distribución de frecuencias teórica

De manera equivalente la tabla de contingencia 3.2 puede reescribirse como la tabla 3.3 resumiendo las subpoblaciones.

En correspondencia con la tabla de frecuencias de la tabla 3.3, se encuentra la tabla de probabilidades que se muestra en la tabla 3.4, donde π_{ij} es la probabilidad de que un sujeto de la i-ésima subpoblación tenga el atributo 1 o 2 de respuesta. Las probabilidades en cada fila suman 1. Esto es el resultado de ver cada nivel de las combinaciones de factores, como una subpoblación distinta (es decir, los n_i's se han fijado de antemano) y de calcular las probabilidades dentro de cada subpoblación.

Para el tipo de muestreo poblacional, donde se tienen 2 categorías de una variable respuesta y varios factores que conforman I subpoblaciones, se cumplen las siguientes restricciones:

Subpoblación	P^+C^+	P^-C^+	P^+C^-	P^-C^-	Total
1	$n_{11}^{(1)}$	$n_{12}^{(1)}$	$n_{21}^{(1)}$	$n_{22}^{(1)}$	n_1
2	$n_{11}^{(2)}$	$n_{12}^{(2)}$	$n_{21}^{(2)}$	$n_{22}^{(2)}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$n_{21}^{(i)}$	$n_{22}^{(i)}$	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$n_{11}^{(I)}$	$n_{12}^{(I)}$	$n_{21}^{(I)}$	$n_{22}^{(I)}$	n_I

Tabla 3.3.: Tabla de muestras

Subpoblación	P^+C^+	P^-C^+	P^+C^-	P^-C^-	Total
1	$\pi_{11}^{(1)}$	$\pi_{12}^{(1)}$	$\pi_{21}^{(1)}$	$\pi_{22}^{(1)}$	1
2	$\pi_{11}^{(2)}$	$\pi_{12}^{(2)}$	$\pi_{21}^{(2)}$	$\pi_{22}^{(2)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\pi_{11}^{(i)}$	$\pi_{12}^{(i)}$	$\pi_{21}^{(i)}$	$\pi_{22}^{(i)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$\pi_{11}^{(I)}$	$\pi_{12}^{(I)}$	$\pi_{21}^{(I)}$	$\pi_{22}^{(I)}$	1

Tabla 3.4.: Distribución de probabilidades para I subpoblaciones generadas.

$$\pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{21}^{(i)}, \pi_{22}^{(i)} > 0 \tag{3.1}$$

$$\pi_{11}^{(i)} + \pi_{12}^{(i)} + \pi_{21}^{(i)} + \pi_{22}^{(i)} = 1 \tag{3.2}$$

A la tabla 3.3 se le ajusta un modelo multinomial para la i-ésima subpoblación con función de masa de probabilidad

$$P\left(n_{11}^{(i)}, n_{12}^{(i)}, n_{21}^{(i)}, n_{22}^{(i)} | \pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{21}^{(i)}, \pi_{22}^{(i)}\right) = \frac{n_i}{n_{11}^{(i)}! n_{12}^{(i)}! n_{21}^{(i)}! n_{22}^{(i)}!} {\pi_{11}^{(i)}}^{n_{11}^{(i)}} {\pi_{12}^{(i)}}^{n_{12}^{(i)}} {\pi_{21}^{(i)}}^{n_{21}^{(i)}} {\pi_{22}^{(i)}}^{n_{22}^{(i)}} \tag{3.3}$$

donde
n_i: es el número de sujetos de la i-ésima subpoblación. No es una variable aleatoria.
$n_{lq}^{(i)}$: es el número de individuos cuya condición (l) es positiva ($l = 1$) o negativa ($l = 2$), cuya prueba (q) es positiva ($q = 1$) o negativa ($q = 2$) y pertenece a la i-ésima subpoblación con $i = 1, 2, \ldots, I$. Es una variable aleatoria.
$\pi_{lq}^{(i)}$: probabilidad de que un individuo tenga condición (l) positiva ($l = 1$) o negativa ($l = 2$), y cuya prueba (q) sea positiva ($q = 1$) o negativa ($q = 2$) en la i-ésima subpoblación con

$i = 1, 2, \ldots, I$.

Se cumple además $n_{11}^{(i)} + n_{12}^{(i)} + n_{21}^{(i)} + n_{22}^{(i)} = n_i$
Como el esquema de muestreo es multinomial independiente o muestreo producto de multinomiales, la función de probabilidad conjunta para todo el conjunto de datos es el producto de funciones multinomiales y tienen la forma:

$$P\left(n_{11}^{(i)}, n_{12}^{(i)}, n_{21}^{(i)}, n_{22}^{(i)} | \pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{21}^{(i)}, \pi_{22}^{(i)}\right) = \prod_{i=1}^{I} \frac{n_i}{n_{11}^{(i)}! n_{12}^{(i)}! n_{21}^{(i)}! n_{22}^{(i)}!} {\pi_{11}^{(i)}}^{n_{11}^{(i)}} {\pi_{12}^{(i)}}^{n_{12}^{(i)}} {\pi_{21}^{(i)}}^{n_{21}^{(i)}} {\pi_{22}^{(i)}}^{n_{22}^{(i)}} \tag{3.4}$$

$$i = 1, 2, \ldots, I$$

3.2. Definición de la función respuesta

Nuestro interés está en modelar la sensibilidad y la especificidad de forma simultánea. La metodología GSK permite hacerlo a través de la formación de dos funciones f_1 y f_2 a partir de la tabla de distribución de probabilidad generada para cada estrato o subpoblación y las k covariables que se distribuyen según se muestra en la Tabla 3.4 que relacionan las variables binarias *Prueba* y *condición*.
Las funciones con respecto a la sensibilidad y la especificidad pueden ser planteadas especialmente de tres formas:

1. De forma directa, de manera que los valores obtenidos en la variable respuesta se interpretan como la sensibilidad y la especificidad para cada subpoblación.
   La ventaja de trabajar con las funciones expresadas de forma directa es la sencillez en su interpretación. La desventaja es que si las probabilidades estimadas tienen valores muy cercanos a 0 o 1 pueden implicar estimaciones erróneas en la sensibilidad y especificidad.

2. Expresarlas de manera logarítmica. Los valores obtenidos en la variable respuesta son el logaritmo de la sensibilidad y la especificidad para cada subpoblación.
   La ventaja de trabajar con las funciones expresadas de forma logarítmica es que da la posibilidad de trabajar con muestras de tamaño más pequeño y aún así garantizar buenas estimaciones.

3. Expresarlas en terminos del logit. Los valores obtenidos en la variable respuesta son el logit de la sensibilidad y la especificidad para cada subpoblación.
   Tiene como ventaja garantizar una buena estimación aún con eventos raros que pueden causar sesgo y probabilidades muy cercanas a 0. Como desventaja está el grado de complejidad en su interpretación.

3.2.1. Funciones definidas de forma directa

Las funciones f_1 y f_2 para la i-ésima población se plantean como sigue:

$$f_1{}^{(i)} = \text{sensibilidad} = \frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)} + \pi_{12}^{(i)}} \tag{3.5}$$

$$f_2{}^{(i)} = \text{especificidad} = \frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)} + \pi_{22}^{(i)}} \tag{3.6}$$

Formación de las funciones

A partir de la tabla 3.4 se obtiene el vector $\boldsymbol{\pi}$ donde cada 4 componentes corresponden a una de las I subpoblaciones organizado en orden ascendente desde la primera subpoblación hasta la última.

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{12}^{(1)} \\ \pi_{21}^{(1)} \\ \pi_{22}^{(1)} \\ \vdots \\ \pi_{11}^{(i)} \\ \pi_{12}^{(i)} \\ \pi_{21}^{(i)} \\ \pi_{22}^{(i)} \\ \vdots \\ \pi_{11}^{(I)} \\ \pi_{12}^{(I)} \\ \pi_{21}^{(I)} \\ \pi_{22}^{(I)} \end{bmatrix} \tag{3.7}$$

Este vector permite bajo ciertas operaciones generar la función respuesta.

Primeramente se formará un vector $\boldsymbol{A\pi}$ como sigue:

$$\boldsymbol{A\pi} = \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{11}^{(1)} + \pi_{12}^{(1)} \\ \pi_{22}^{(1)} \\ \pi_{21}^{(1)} + \pi_{22}^{(1)} \\ \vdots \\ \pi_{11}^{(i)} \\ \pi_{11}^{(1)} + \pi_{12}^{(i)} \\ \pi_{22}^{(i)} \\ \pi_{21}^{(i)} + \pi_{22}^{(i)} \\ \vdots \\ \pi_{11}^{(I)} \\ \pi_{11}^{(I)} + \pi_{12}^{(I)} \\ \pi_{22}^{(I)} \\ \pi_{21}^{(I)} + \pi_{22}^{(I)} \end{bmatrix} \tag{3.8}$$

Para que se obtenga el vector (3.8) la matriz $\boldsymbol{A}$ se construye como:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 1 \end{bmatrix} \tag{3.9}$$

donde la matriz $\boldsymbol{A}$ es una matriz diagonal de dimensiones $4I \times 4I$ donde en su diagonal contiene al bloque (3.10).

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{3.10}$$

De manera que

$$\boldsymbol{A\pi} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{12}^{(1)} \\ \pi_{21}^{(1)} \\ \pi_{22}^{(1)} \\ \vdots \\ \pi_{11}^{(i)} \\ \pi_{12}^{(i)} \\ \pi_{21}^{(i)} \\ \pi_{22}^{(i)} \\ \vdots \\ \pi_{11}^{(I)} \\ \pi_{12}^{(I)} \\ \pi_{21}^{(I)} \\ \pi_{22}^{(I)} \end{bmatrix} = \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{11}^{(1)} + \pi_{12}^{(1)} \\ \pi_{22}^{(1)} \\ \pi_{21}^{(1)} + \pi_{22}^{(1)} \\ \vdots \\ \pi_{11}^{(i)} \\ \pi_{11}^{(1)} + \pi_{12}^{(i)} \\ \pi_{22}^{(i)} \\ \pi_{21}^{(i)} + \pi_{22}^{(i)} \\ \vdots \\ \pi_{11}^{(I)} \\ \pi_{11}^{(I)} + \pi_{12}^{(I)} \\ \pi_{22}^{(I)} \\ \pi_{21}^{(I)} + \pi_{22}^{(I)} \end{bmatrix}$$

Una vez definida la matriz $\boldsymbol{A}$ pasamos a tomar el logaritmo natural de (3.8) componente a componente:

$$\begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)} + \pi_{22}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)} + \pi_{22}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{11}^{(I)} + \pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)} + \pi_{22}^{(I)}\right) \end{bmatrix} \tag{3.11}$$

Multiplicamos una matriz $\boldsymbol{K}$ de dimensiones $2I \times 4I$ por el vector (3.11) de dimensiones $4I \times 1$:

$$\boldsymbol{K} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -1 \end{bmatrix} \tag{3.12}$$

Notese que $\boldsymbol{K}$ es una matriz donde en la diagonal tiene el bloque que se muestra en (3.13) para cada subpoblación i.

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{3.13}$$

Luego,

$$\boldsymbol{K} * \begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)} + \pi_{22}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)} + \pi_{22}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{11}^{(I)} + \pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)} + \pi_{22}^{(I)}\right) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)} + \pi_{22}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)} + \pi_{22}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{11}^{(I)} + \pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)} + \pi_{22}^{(I)}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \ln\left(\frac{\pi_{11}^{(1)}}{\pi_{11}^{(1)}+\pi_{12}^{(1)}}\right) \\ \ln\left(\frac{\pi_{22}^{(1)}}{\pi_{21}^{(1)}+\pi_{22}^{(1)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{(i)}}{\pi_{11}^{(1)}+\pi_{12}^{(i)}}\right) \\ \ln\left(\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{(I)}}{\pi_{11}^{(I)}+\pi_{12}^{(I)}}\right) \\ \ln\left(\frac{\pi_{22}^{(I)}}{\pi_{21}^{(I)}+\pi_{22}^{(I)}}\right) \end{bmatrix} = \begin{bmatrix} \ln\left(\pi_1^{(1)}\right) \\ \ln\left(\pi_2^{(1)}\right) \\ \vdots \\ \ln\left(\pi_1^{(i)}\right) \\ \ln\left(\pi_2^{(i)}\right) \\ \vdots \\ \ln\left(\pi_1^{(I)}\right) \\ \ln\left(\pi_2^{(I)}\right) \end{bmatrix}$$

con $\pi_1^{(i)} = \frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)}+\pi_{12}^{(i)}}$ y $\pi_2^{(i)} = \frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}$.

Por último, se exponencia componente a componente el vector $\boldsymbol{K}\ln(\boldsymbol{A\pi})$, y se multiplica por una matriz identidad $\boldsymbol{Q}$ de dimensiones $2I \times 2I$ tal que premultiplicando a $\boldsymbol{K}\ln(\boldsymbol{A\pi})$ de dimensiones $2I \times 1$ genera el vector respuesta $\boldsymbol{f}$:

$$\begin{bmatrix} e^{\ln\left(\pi_1^{(1)}\right)} \\ e^{\ln\left(\pi_2^{(1)}\right)} \\ \vdots \\ e^{\ln\left(\pi_1^{(i)}\right)} \\ e^{\ln\left(\pi_2^{(i)}\right)} \\ \vdots \\ e^{\ln\left(\pi_1^{(I)}\right)} \\ e^{\ln\left(\pi_2^{(I)}\right)} \end{bmatrix} = \begin{bmatrix} \pi_1^{(1)} \\ \pi_2^{(1)} \\ \vdots \\ \pi_1^{(i)} \\ \pi_2^{(i)} \\ \vdots \\ \pi_1^{(I)} \\ \pi_2^{(I)} \end{bmatrix} \tag{3.14}$$

$$\boldsymbol{Q} * \begin{bmatrix} \pi_1^{(1)} \\ \pi_2^{(1)} \\ \vdots \\ \pi_1^{(i)} \\ \pi_2^{(i)} \\ \vdots \\ \pi_1^{(I)} \\ \pi_2^{(I)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_1^{(1)} \\ \pi_2^{(1)} \\ \vdots \\ \pi_1^{(i)} \\ \pi_2^{(i)} \\ \vdots \\ \pi_1^{(I)} \\ \pi_2^{(I)} \end{bmatrix} \tag{3.15}$$

$$= \begin{bmatrix} \pi_1^{(1)} \\ \pi_2^{(1)} \\ \vdots \\ \pi_1^{(i)} \\ \pi_2^{(i)} \\ \vdots \\ \pi_1^{(I)} \\ \pi_2^{(I)} \end{bmatrix} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ \vdots \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_1^{(I)} \\ f_2^{(I)} \end{bmatrix} = \mathbf{f}_{\text{exp}} \tag{3.16}$$

De manera que $\mathbf{f}_{\text{exp}}$ se obtiene operacionalmente como:

$$\mathbf{f}_{\text{exp}} = \boldsymbol{Q} \exp[\boldsymbol{K} \ln(\boldsymbol{A}\boldsymbol{\pi})]$$

donde los elementos impares de $\mathbf{f}_{\text{exp}}$ corresponden a la sensibilidad y los elementos pares corresponden a la especificidad por subpoblación.

3.2.2. Funciones definidas de forma logarítmica

Las funciones f_1 y f_2 para la i-ésima población se proponen como sigue:

$$f_1{}^{(i)} = \ln(\text{sensibilidad}) = \ln\left(\frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)} + \pi_{12}^{(i)}}\right) \tag{3.17}$$

$$f_2{}^{(i)} = \ln(\text{especificidad}) = \ln\left(\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)} + \pi_{22}^{(i)}}\right) \tag{3.18}$$

Formación de las funciones

A partir de la tabla 3.4, se obtiene el vector $\boldsymbol{\pi}$ donde cada 4 componentes corresponden a una de las subpoblaciones de I, organizado en orden ascendente desde la primera subpoblación hasta la última.

$$\boldsymbol{\pi}' = \left[\begin{array}{cccccccccccccccc} \pi_{11}^{(1)} & \pi_{12}^{(1)} & \pi_{21}^{(1)} & \pi_{22}^{(1)} & \dots & \pi_{11}^{(i)} & \pi_{12}^{(i)} & \pi_{21}^{(i)} & \pi_{22}^{(i)} & \dots & \pi_{11}^{(I)} & \pi_{12}^{(I)} & \pi_{21}^{(I)} & \pi_{22}^{(I)} \end{array}\right]$$

Este vector permite bajo ciertas operaciones generar la función respuesta.
Primeramente se formará el vector $\boldsymbol{A\pi}$ como se plantea en (3.8) con matriz $\boldsymbol{A}$ de dimensiones $4I \times I$ dada por (3.9)
Una vez definida la matriz $\boldsymbol{A}$ pasamos a tomar el logaritmo natural de $\boldsymbol{A\pi}$ componente a componente como en (3.11).
Por último, multiplicamos la matriz $\boldsymbol{K}$ (3.12) de dimensiones $2I \times 4I$ por la matriz (3.11) de dimensiones $4I \times 1$ tal que genera el vector respuesta f. Desarrollando se obtiene:

$$\boldsymbol{K} * \begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)} + \pi_{22}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{11}^{(1)} + \pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)} + \pi_{22}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{11}^{(I)} + \pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)} + \pi_{22}^{(I)}\right) \end{bmatrix} = \begin{bmatrix} \ln\left(\frac{\pi_{11}^{(1)}}{\pi_{11}^{(1)}+\pi_{12}^{(1)}}\right) \\ \ln\left(\frac{\pi_{22}^{(1)}}{\pi_{21}^{(1)}+\pi_{22}^{(1)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{(i)}}{\pi_{11}^{(1)}+\pi_{12}^{(i)}}\right) \\ \ln\left(\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{(I)}}{\pi_{11}^{(I)}+\pi_{12}^{(I)}}\right) \\ \ln\left(\frac{\pi_{22}^{(I)}}{\pi_{21}^{(I)}+\pi_{22}^{(I)}}\right) \end{bmatrix} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ \vdots \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_1^{(I)} \\ f_2^{(I)} \end{bmatrix} = \mathbf{f}_{\ln} \tag{3.19}$$

De manera que $\mathbf{f}_{\ln}$ se obtiene operacionalmente como:

$$\mathbf{f}_{\ln} = \boldsymbol{K} \ln(\boldsymbol{A}\boldsymbol{\pi})$$

donde los elementos impares de $\mathbf{f}_{\ln}$ corresponden a la sensibilidad y los elementos pares corresponden a la especificidad por subpoblación.

3.2.3. Funciones definidas en forma de logit

Las funciones f_1 y f_2 para la i-ésima población son:

$$f_1{}^{(i)} = \text{logit(sensibilidad)} = \ln\left(\frac{\pi_1^{(i)}}{1-\pi_1^{(i)}}\right) \tag{3.20}$$

donde

$$\pi_1^{(i)} = \frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)} + \pi_{12}^{(i)}} \tag{3.21}$$

$$\frac{\pi_1^{(i)}}{1-\pi_1^{(i)}} = \frac{\frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)}+\pi_{12}^{(i)}}}{1-\frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)}+\pi_{12}^{(i)}}} = \frac{\frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)}+\pi_{12}^{(i)}}}{\frac{\pi_{12}^{(i)}}{\pi_{11}^{(i)}+\pi_{12}^{(i)}}} = \frac{\pi_{11}^{(i)}}{\pi_{12}^{(i)}}$$

$$f_2{}^{(i)} = \text{logit(especificidad)} = \ln\left(\frac{\pi_2^{(i)}}{1-\pi_2^{(i)}}\right) \tag{3.22}$$

donde

$$\pi_2^{(i)} = \frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)} + \pi_{22}^{(i)}} \tag{3.23}$$

$$\frac{\pi_2^{(i)}}{1-\pi_2^{(i)}} = \frac{\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}}{1-\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}} = \frac{\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}}{\frac{\pi_{21}^{(i)}}{\pi_{21}^{(i)}+\pi_{22}^{(i)}}} = \frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}}$$

Formación de las funciones

A partir de la tabla 3.4, se obtiene el vector $\boldsymbol{\pi}$ donde cada 4 componentes corresponden a una de las subpoblaciones de I, organizado en orden ascendente desde la primera subpoblación hasta la última.

$$\boldsymbol{\pi}' = \begin{bmatrix} \pi_{11}^{(1)} & \pi_{12}^{(1)} & \pi_{21}^{(1)} & \pi_{22}^{(1)} & \dots & \pi_{11}^{(i)} & \pi_{12}^{(i)} & \pi_{21}^{(i)} & \pi_{22}^{(i)} & \dots & \pi_{11}^{(I)} & \pi_{12}^{(I)} & \pi_{21}^{(I)} & \pi_{22}^{(I)} \end{bmatrix}$$

Este vector permite bajo ciertas operaciones generar la función respuesta.
Primeramente se formará un vector $A\pi$ como sigue:

$$\boldsymbol{A\pi} = \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{12}^{(1)} \\ \pi_{22}^{(1)} \\ \pi_{21}^{(1)} \\ \vdots \\ \pi_{11}^{(i)} \\ \pi_{12}^{(i)} \\ \pi_{22}^{(i)} \\ \pi_{21}^{(i)} \\ \vdots \\ \pi_{11}^{(I)} \\ \pi_{12}^{(I)} \\ \pi_{22}^{(I)} \\ \pi_{21}^{(I)} \end{bmatrix} \tag{3.24}$$

Para que se obtenga el vector anterior la matriz $\boldsymbol{A}$ se construye como:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \end{bmatrix} \tag{3.25}$$

Luego, la matriz $\boldsymbol{A}$ tiene dimensiones $4I \times I$ compuesta por bloques con la siguiente estructura para cada subpoblación i:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Una vez definida la matriz $\boldsymbol{A}$ pasamos a tomar el logaritmo natural de $\boldsymbol{A\pi}$ (3.24) componente a componente:

$$\begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)}\right) \end{bmatrix} \tag{3.26}$$

Por último, multiplicamos una matriz $\boldsymbol{K}$ de dimensiones $2I \times 4I$ por la matriz (3.26) de dimensiones $4I \times 1$ tal que genera el vector respuesta $\boldsymbol{f}_{\text{logit}}$. Nótese que $\boldsymbol{K}$ es una matriz definida como en (3.12).

$$\begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)}\right) \end{bmatrix}$$

Desarrollando se obtiene:

$$\boldsymbol{K} * \begin{bmatrix} \ln\left(\pi_{11}^{(1)}\right) \\ \ln\left(\pi_{12}^{(1)}\right) \\ \ln\left(\pi_{22}^{(1)}\right) \\ \ln\left(\pi_{21}^{(1)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(i)}\right) \\ \ln\left(\pi_{12}^{(i)}\right) \\ \ln\left(\pi_{22}^{(i)}\right) \\ \ln\left(\pi_{21}^{(i)}\right) \\ \vdots \\ \ln\left(\pi_{11}^{(I)}\right) \\ \ln\left(\pi_{12}^{(I)}\right) \\ \ln\left(\pi_{22}^{(I)}\right) \\ \ln\left(\pi_{21}^{(I)}\right) \end{bmatrix} = \begin{bmatrix} \ln\left(\frac{\pi_{11}^{(1)}}{\pi_{12}^{(1)}}\right) \\ \ln\left(\frac{\pi_{22}^{(1)}}{\pi_{21}^{(1)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{(i)}}{\pi_{12}^{(i)}}\right) \\ \ln\left(\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{11}^{(I)}}{\pi_{12}^{(I)}}\right) \\ \ln\left(\frac{\pi_{22}^{(I)}}{\pi_{21}^{(I)}}\right) \end{bmatrix} = \begin{bmatrix} \ln\left(\frac{\pi_{1}^{(1)}}{1-\pi_{1}^{(1)}}\right) \\ \ln\left(\frac{\pi_{2}^{(1)}}{1-\pi_{2}^{(1)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{1}^{(i)}}{1-\pi_{1}^{(i)}}\right) \\ \ln\left(\frac{\pi_{2}^{(i)}}{1-\pi_{2}^{(i)}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{1}^{(I)}}{1-\pi_{1}^{(I)}}\right) \\ \ln\left(\frac{\pi_{2}^{(I)}}{1-\pi_{2}^{(I)}}\right) \end{bmatrix} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ \vdots \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_1^{(I)} \\ f_2^{(I)} \end{bmatrix} = \mathbf{f}_{\text{logit}} \qquad (3.27)$$

De manera que $\mathbf{f}_{\text{logit}}$ se obtiene operacionalmente como:

$$\mathbf{f}_{\text{logit}} = \boldsymbol{K} \ln(\boldsymbol{A}\boldsymbol{\pi})$$

donde los elementos impares de $\mathbf{f}_{\text{logit}}$ corresponden al logit de la sensibilidad y los elementos pares corresponden al logit de la especificidad por subpoblación.

3.3. Matrices de varianzas y covarianzas

Para obtener las funciones $\mathbf{f}_{\text{exp}}$, $\mathbf{f}_{\ln}$ y $\mathbf{f}_{\text{logit}}$, se hace uso de datos muestrales lo cual acarrea el uso de un estimador para cada una de las cuatro probabilidades planteadas por subpoblación. El estimador de máxima verosimilitud para cada una de ellas es:

$$\hat{\pi}_{11}^{(i)} = \frac{n_{11}^{(i)}}{n_i}, \quad \hat{\pi}_{12}^{(i)} = \frac{n_{12}^{(i)}}{n_i}, \quad \hat{\pi}_{21}^{(i)} = \frac{n_{21}^{(i)}}{n_i} \quad y \quad \hat{\pi}_{22}^{(i)} = \frac{n_{22}^{(i)}}{n_i}$$

Con $n_{11}^{(i)}, n_{12}^{(i)}, n_{21}^{(i)}$ y $n_{22}^{(i)}$ frecuencias observadas en la i-ésima subpoblación y n_i el total de observaciones de la i-ésima subpoblación Grizzle et al. (1969).

Haciendo uso de los estimadores anteriores se puede construir $\hat{\pi}$ como el estimador de máxima verosimilitud para π:

$$\begin{aligned}
\hat{\boldsymbol{\pi}}' &= \left[\begin{array}{cccccccccccccc} \hat{\pi}_{11}^{(1)} & \hat{\pi}_{12}^{(1)} & \hat{\pi}_{21}^{(1)} & \hat{\pi}_{22}^{(1)} & \cdots & \hat{\pi}_{11}^{(i)} & \hat{\pi}_{12}^{(i)} & \hat{\pi}_{21}^{(i)} & \hat{\pi}_{22}^{(i)} & \cdots & \hat{\pi}_{11}^{(I)} & \hat{\pi}_{12}^{(I)} & \hat{\pi}_{21}^{(I)} & \hat{\pi}_{22}^{(I)} \end{array} \right] \\
&= \left[\begin{array}{cccccccccccccc} \frac{n_{11}^{(1)}}{n_1} & \frac{n_{12}^{(1)}}{n_1} & \frac{n_{21}^{(1)}}{n_1} & \frac{n_{22}^{(1)}}{n_1} & \cdots & \frac{n_{11}^{(i)}}{n_i} & \frac{n_{12}^{(i)}}{n_i} & \frac{n_{21}^{(i)}}{n_i} & \frac{n_{22}^{(i)}}{n_i} & \cdots & \frac{n_{11}^{(I)}}{n_I} & \frac{n_{12}^{(I)}}{n_I} & \frac{n_{21}^{(I)}}{n_I} & \frac{n_{22}^{(I)}}{n_I} \end{array} \right]
\end{aligned}$$

Por el teorema central del límite multivariado se puede demostar que $\hat{\boldsymbol{\pi}}$ sigue asintóticamente una distribución normal $AN(\boldsymbol{\pi}, \boldsymbol{\Sigma}_{\boldsymbol{\pi}})$. Por lo cual los valores esperados de los estimadores para la i-ésima subpoblación son:

$$E\left(\hat{\pi}_{11}^{(i)}\right) = \pi_{11}^{(i)}, \quad E\left(\hat{\pi}_{12}^{(i)}\right) = \pi_{12}^{(i)}, \quad E\left(\hat{\pi}_{21}^{(i)}\right) = \pi_{21}^{(i)} \quad y \quad E\left(\hat{\pi}_{22}^{(i)}\right) = \pi_{22}^{(i)}$$

En la metodología GSK se requiere estimar las varianzas y covarianzas de $\hat{\boldsymbol{\pi}}$ para la estimación de los parámetros del modelo. Adicionalmente el análisis de las matrices de varianzas y covarianzas para las funciones planteadas en forma exponencial ($\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\text{exp}}}$), en forma logarítmica ($\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\ln}}$) y en forma de logit ($\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\text{logit}}}$) nos permitirá observar la relación que existe entre la sensibilidad y la especifidad para cada subpoblación.

Para la i-ésima subpoblación de una muestra aleatoria de tamaño n_i, la cual es multinomial, sea

$$\hat{\boldsymbol{\pi}}^{(i)} = \begin{bmatrix} \frac{n_{11}^{(i)}}{n_i} \\ \frac{n_{12}^{(i)}}{n_i} \\ \frac{n_{21}^{(i)}}{n_i} \\ \frac{n_{22}^{(i)}}{n_i} \end{bmatrix} \tag{3.28}$$

se tiene entonces que las varianzas de los estimadores en la i-ésima subpoblación son:

$$\text{Var}\left(\hat{\pi}_{11}^{(i)}\right) = \frac{1}{n_i}\pi_{11}^{(i)}\left(1-\pi_{11}^{(i)}\right), \quad \text{Var}\left(\hat{\pi}_{12}^{(i)}\right) = \frac{1}{n_i}\pi_{12}^{(i)}\left(1-\pi_{12}^{(i)}\right), \quad \text{Var}\left(\hat{\pi}_{21}^{(i)}\right) = \frac{1}{n_i}\pi_{21}^{(i)}\left(1-\pi_{21}^{(i)}\right)$$

$$\text{y} \quad \text{Var}\left(\hat{\pi}_{22}^{(i)}\right) = \frac{1}{n_i}\pi_{22}^{(i)}\left(1-\pi_{22}^{(i)}\right) \tag{3.29}$$

y las covarianzas están dadas por:

$$\text{Cov}\left(\hat{\pi}_{11}^{(i)}, \hat{\pi}_{12}^{(i)}\right) = \frac{1}{n_i}\left(-\pi_{11}^{(i)}\pi_{12}^{(i)}\right), \quad \text{Cov}\left(\hat{\pi}_{11}^{(i)}, \hat{\pi}_{21}^{(i)}\right) = \frac{1}{n_i}\left(-\pi_{11}^{(i)}\pi_{21}^{(i)}\right),$$
$$\text{Cov}\left(\hat{\pi}_{11}^{(i)}, \hat{\pi}_{22}^{(i)}\right) = \frac{1}{n_i}\left(-\pi_{11}^{(i)}\pi_{22}^{(i)}\right) \tag{3.30}$$

Luego, la matriz de varianzas y covarianzas estimada de $\hat{\boldsymbol{\pi}}^{(i)}$ se construye como :

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}^{(i)}} = \frac{1}{n_i}\begin{bmatrix} \pi_{11}^{(i)}\left(1-\pi_{11}^{(i)}\right) & -\pi_{11}^{(i)}\pi_{12}^{(i)} & -\pi_{11}^{(i)}\pi_{21}^{(i)} & -\pi_{11}^{(i)}\pi_{22}^{(i)} \\ -\pi_{11}^{(i)}\pi_{12}^{(i)} & \pi_{12}^{(i)}\left(1-\pi_{12}^{(i)}\right) & -\pi_{12}^{(i)}\pi_{21}^{(i)} & -\pi_{12}^{(i)}\pi_{22}^{(i)} \\ -\pi_{11}^{(i)}\pi_{21}^{(i)} & -\pi_{12}^{(i)}\pi_{21}^{(i)} & \pi_{21}^{(i)}\left(1-\pi_{21}^{(i)}\right) & -\pi_{21}^{(i)}\pi_{22}^{(i)} \\ -\pi_{11}^{(i)}\pi_{22}^{(i)} & -\pi_{12}^{(i)}\pi_{22}^{(i)} & -\pi_{21}^{(i)}\pi_{22}^{(i)} & \pi_{22}^{(i)}\left(1-\pi_{22}^{(i)}\right) \end{bmatrix} \tag{3.31}$$

En el caso de la distribución multinomial, la matriz de covarianzas no es de rango completo y frecuentemente requiere ser trabajada con una inversa generalizada (Tanabe and Sagae, 1992).

Considerando lo anterior, la matriz de varianzas y covarianzas para $\hat{\boldsymbol{\pi}}$ está dada por $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}}$; una matriz diagonal por bloques de dimensión $4I \times 4I$ la cual está dada por:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}(1)} & 0 & \dots & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}(2)} & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}(i)} & \dots & 0 \\ \vdots & \vdots & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \dots & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}(I)} \end{bmatrix} \tag{3.32}$$

Para el caso donde la sensibilidad y la especificidad se plantean de forma directa, teniendo en cuenta que asintóticamente $\hat{\boldsymbol{\pi}}$ se distribuye multinomial con media $\boldsymbol{\pi}$ y matriz de varianzas y covarianzas dada por $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}}$, entonces $\hat{\boldsymbol{f}}_{\mathbf{exp}} = \boldsymbol{Q}\exp[\boldsymbol{K}\ln(\boldsymbol{A}\hat{\boldsymbol{\pi}})]$ se distribuye asintóticamente multinormal con matriz de varianzas y covarianzas dada por $\boldsymbol{\Sigma}_{\hat{\mathbf{f}}_{\exp}} = \boldsymbol{Q}\boldsymbol{D}_{\mathbf{ln}}\boldsymbol{\Sigma}_{\mathbf{ln}}\boldsymbol{D}_{\mathbf{ln}}\boldsymbol{Q}'$ (Forthofer and Koch, 1972), donde $\boldsymbol{D}_{\mathbf{ln}}$ es una matriz diagonal de dimensiones $2I \times 2I$, cuya diagonal principal está compuesta por el i-ésimo elemento del vector $\boldsymbol{f}_{\mathbf{exp}} = \exp(\boldsymbol{f}_{\mathbf{ln}})$ (3.14) donde $\boldsymbol{f}_{\mathbf{ln}} = \boldsymbol{K}\ln(\boldsymbol{A}\boldsymbol{\pi})$. Sea $f_{\exp}(i)$ el i-ésimo elemento de $\boldsymbol{f}_{\mathbf{exp}}$:

$$\boldsymbol{D}_{\mathbf{ln}} = \begin{bmatrix} f_{\exp}(1) & 0 & \dots & 0 \\ 0 & f_{\exp}(2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_{\exp}(2I) \end{bmatrix} \tag{3.33}$$

Además

$$\boldsymbol{\Sigma}_{\mathbf{ln}} = \boldsymbol{K}\boldsymbol{D}_{\mathbf{lineal}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{lineal}}\boldsymbol{D}_{\mathbf{lineal}}^{-\mathbf{1}}\boldsymbol{K}' \tag{3.34}$$

donde $\boldsymbol{D}_{\mathbf{lineal}}$ es una matriz diagonal $4I \times 4I$ con su diagonal principal compuesta por la multiplicación entre el vector $\boldsymbol{\pi}$ y $\boldsymbol{a'}_{\boldsymbol{i}}$, que es la i-ésima fila de $\boldsymbol{A}$:

$$\boldsymbol{D}_{\mathbf{lineal}} = \begin{bmatrix} a'_1\pi & 0 & \dots & 0 \\ 0 & a'_2\pi & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a'_{4I}\pi \end{bmatrix} \tag{3.35}$$

Por otro lado $\boldsymbol{\Sigma}_{\mathbf{lineal}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}}\boldsymbol{A}'$ y las matrices $\boldsymbol{A}$, $\boldsymbol{K}$ y $\boldsymbol{Q}$ son como se plantean en las ecuaciones (3.9), (3.12) y (3.15) respectivamente.

La matriz de varianzas y covarianzas estimada del vector de probabilidades estimadas para la i-ésima población es:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}^{(i)}} = \frac{1}{n_i}\begin{bmatrix} \hat{\pi}_{11}^{(i)}\left(1-\hat{\pi}_{11}^{(i)}\right) & -\hat{\pi}_{11}^{(i)}\hat{\pi}_{12}^{(i)} & -\hat{\pi}_{11}^{(i)}\hat{\pi}_{21}^{(i)} & -\hat{\pi}_{11}^{(i)}\hat{\pi}_{22}^{(i)} \\ -\hat{\pi}_{11}^{(i)}\hat{\pi}_{12}^{(i)} & \hat{\pi}_{12}^{(i)}\left(1-\hat{\pi}_{12}^{(i)}\right) & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{21}^{(i)} & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{22}^{(i)} \\ -\hat{\pi}_{11}^{(i)}\hat{\pi}_{21}^{(i)} & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{21}^{(i)} & \hat{\pi}_{21}^{(i)}\left(1-\hat{\pi}_{21}^{(i)}\right) & -\hat{\pi}_{21}^{(i)}\hat{\pi}_{22}^{(i)} \\ -\hat{\pi}_{11}^{(i)}\hat{\pi}_{22}^{(i)} & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{22}^{(i)} & -\hat{\pi}_{21}^{(i)}\hat{\pi}_{22}^{(i)} & \hat{\pi}_{22}^{(i)}\left(1-\hat{\pi}_{22}^{(i)}\right) \end{bmatrix} \tag{3.36}$$

La matriz de covarianzas estimada para las I subpoblaciones será:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}^{(1)}} & 0 & \dots & 0 & \dots & 0 \\ 0 & \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}^{(2)}} & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}^{(i)}} & \dots & 0 \\ \vdots & \vdots & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \dots & \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}^{(I)}} \end{bmatrix} \tag{3.37}$$

La matrices diagonales $\boldsymbol{D}_{\mathbf{ln}}$ y $\boldsymbol{D}_{\mathbf{lineal}}$ estimadas son:

$$\hat{\boldsymbol{D}}_{\mathbf{ln}} = \begin{bmatrix} \hat{f}_{\exp}(1) & 0 & \dots & 0 \\ 0 & \hat{f}_{\exp}(2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{f}_{\exp}(2I) \end{bmatrix} \tag{3.38}$$

$$\hat{\boldsymbol{D}}_{\mathbf{lineal}} = \begin{bmatrix} a'_1\hat{\pi} & 0 & \dots & 0 \\ 0 & a'_2\hat{\pi} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a'_{4I}\hat{\pi} \end{bmatrix} \tag{3.39}$$

Entonces la matriz estimada $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{exp}}}$ de varianzas y covarianza de $\hat{\boldsymbol{f}}$ es:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{exp}}} = \boldsymbol{Q}\hat{\boldsymbol{D}}_{\mathbf{ln}}\hat{\boldsymbol{\Sigma}}_{\mathbf{ln}}\hat{\boldsymbol{D}}_{\mathbf{ln}}\boldsymbol{Q'} \tag{3.40}$$

con

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{ln}} = \boldsymbol{K}\hat{\boldsymbol{D}}^{-\mathbf{1}}_{\mathbf{lineal}}\hat{\boldsymbol{\Sigma}}_{\mathbf{lineal}}\hat{\boldsymbol{D}}^{-\mathbf{1}}_{\mathbf{lineal}}\boldsymbol{K'} \tag{3.41}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{lineal}} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}}\boldsymbol{A'} \tag{3.42}$$

Por otra parte para los casos donde la sensibilidad y la especificidad se plantean de forma logaritmica y logit, teniendo en cuenta que asintóticamente $\hat{\boldsymbol{\pi}}$ se distribuye multinomial con media $\boldsymbol{\pi}$ y matriz de varianzas y covarianzas dada por $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}}$, entonces $\hat{\mathbf{f}}_{\text{logit}} = \hat{\mathbf{f}}_{\text{ln}} = \boldsymbol{K}\ln(\boldsymbol{A}\hat{\boldsymbol{\pi}})$

se distribuye asintóticamente multinormal con matriz de varianzas y covarianzas dada por $\boldsymbol{\Sigma}_{\hat{\mathbf{f}}_{\text{logit}}} = \boldsymbol{\Sigma}_{\hat{\mathbf{f}}_{\mathbf{ln}}} = \boldsymbol{K}\boldsymbol{D}_{\boldsymbol{lineal}}^{-\mathbf{1}}\boldsymbol{A}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}}\boldsymbol{A}'\boldsymbol{D}_{\boldsymbol{lineal}}^{-\mathbf{1}}\boldsymbol{K}'$ Grizzle et al. (1969), donde $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\pi}}}$ y $\boldsymbol{D}_{\boldsymbol{lineal}}$ son matrices como se planteó anteriormente en (3.32) y (3.35) respectivamente.
Entonces la matriz estimada $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{ln}}}$ ($\boldsymbol{\Sigma}_{\hat{\mathbf{f}}_{\text{logit}}}$) de varianzas y covarianza de $\hat{\mathbf{f}}_{\text{ln}}$ ($\hat{\mathbf{f}}_{\text{logit}}$) es:

$$\boldsymbol{\Sigma}_{\hat{\mathbf{f}}_{\text{logit}}} = \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{ln}}} = \boldsymbol{K}\hat{\boldsymbol{D}}_{\boldsymbol{lineal}}^{-\mathbf{1}}\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}}\boldsymbol{A}'\hat{\boldsymbol{D}}_{\boldsymbol{lineal}}^{-\mathbf{1}}\boldsymbol{K}' \tag{3.43}$$

donde $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\pi}}}$ y $\hat{\boldsymbol{D}}_{\boldsymbol{lineal}}$ vienen dadas por (3.37) y (3.39) respectivamente. Nótese que aunque $\hat{\mathbf{f}}_{\text{logit}}$, $\hat{\mathbf{f}}_{\text{ln}}$ y $\boldsymbol{\Sigma}_{\hat{\mathbf{f}}_{\text{logit}}}$, $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{ln}}}$ tienen la misma estructura, se diferencia por el valor de las matrices A y K, de manera que son como se muestran en (3.9) y (3.12) para el caso en que se planteen de forma logarítmica y, (3.25) y (3.12) para el caso en que se plantee en forma de logit la sensibilidad y la especificidad.

3.4. Modelo lineal bajo la metodología GSK

El modelo lineal paramétrico que relaciona la sensibilidad y la especificidad con un conjunto de covariables para cada subpoblación es:

$$\hat{\boldsymbol{f}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.44}$$

donde $\hat{\boldsymbol{f}}$ es la función respuesta muestral.

El modelo que se propone tiene dos respuestas, una que corresponde a la sensibilidad y otra a la especificidad con lo cual $\hat{\boldsymbol{f}}$ viene dado por:

$$\hat{\boldsymbol{f}} = \begin{bmatrix} \hat{f}_1^{(1)} \\ \hat{f}_2^{(1)} \\ \hat{f}_1^{(2)} \\ \hat{f}_2^{(2)} \\ \vdots \\ \hat{f}_1^{(i)} \\ \hat{f}_2^{(i)} \\ \vdots \\ \hat{f}_1^{(I)} \\ \hat{f}_2^{(I)} \end{bmatrix}$$

donde $\hat{f}_1^{(i)}$ y $\hat{f}_2^{(i)}$ son las funciones respuestas con respecto a la sensibilidad y la especificidad respectivamente, generadas para la i-ésima subpoblación que van a ser modeladas linealmente. Cabe resaltar que $\hat{f}_1^{(i)}$ y $\hat{f}_2^{(i)}$ pueden variar conjuntamente dependiendo de la forma en que desee obtenerse la sensibilidad y la especificidad, es decir, pueden plantearse en forma

directa, logarítmica o aplicando el logit como se mencionó al inicio de la sección 3.2 y como se plantean en las subsecciones 3.2.1, 3.2.2 y 3.2.3.

$\boldsymbol{X}$ es una matriz de diseño, $\boldsymbol{\beta}$ es un vector de parámetros desconocido y $\boldsymbol{\epsilon}$ un vector que se asume que se distribuye asintóticamente normal, con media 0 y $\mathrm{Var}\,(\epsilon) = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\epsilon}}}$, esto es, $(\boldsymbol{\epsilon} \sim \mathrm{AN}(\mathbf{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\epsilon}}}))$.

A continuación se presenta la definición de la matriz de diseño y de los demás elementos del modelo.

3.4.1. Definición de la matriz de diseño

Matriz de diseño para un modelo naive

Los coeficientes generados en el modelo naive son válidos tanto para $f_1^{(i)}$ como para $f_2^{(i)}$ y por tanto genera las mismas probabilidades para ambas respuestas. Se expondrá la forma en la que está construida la matriz de diseño para este modelo con el fin de posteriormente utilizarla para la formación de un modelo general, que será el que se utilizará finalmente.

Para el modelo (3.44) la matriz de diseño $\boldsymbol{X}$ tiene dimensiones $2I \times \left(1 + P_1^{(1)} - 1 + \cdots + P_k^{(k)} - 1\right)$, $\boldsymbol{\beta}$ es un vector de parámetros desconocidos de dimensión $\left(1 + P_1^{(1)} - 1 + \cdots + P_k^{(k)} - 1\right) \times 1$ y $\boldsymbol{\epsilon}$ se asume se distribuye aproximadamente normal. La matriz de diseño puede contener información de la subpoblación expresado en términos de las variables categóricas asociadas con la subpoblación.

Para $(P_k - 1)$ niveles de la variable X_k con $k = 1, 2, \ldots, m$ definimos una variable indicadora, es decir una variable que toma el valor 1 para denotar la presencia de un atributo cualitativo y usa el valor 0 para denotar la ausencia de este atributo, es decir si el nivel de la categoría considerada es o no observada en una subpoblación (Dutta, 1982). La variable indicadora se representará como $z_r^{(k)}$ donde r es el r-ésimo nivel de una variable X_k con $r = 1, \ldots, P_k - 1$ y k se refiere a la k-ésima varible X tal que:

$$z_r^{(k)} = \begin{cases} 0 & \text{el nivel } r \text{ para la covariable } k \text{ no es observado} \\ \\ 1 & \text{el nivel } r \text{ para la covariable } k \text{ es observado} \end{cases}$$

La matriz de diseño teniendo en cuenta las variables indicadores relacionadas con los niveles de las covariables tiene la forma:

	$z_1^{(1)}$	$z_2^{(1)}$	...	$z_{P_1-1}^{(1)}$	...	$z_1^{(k-1)}$	$z_2^{(k-1)}$	...	$z_{P_{k-1}-1}^{(k-1)}$	$z_1^{(k)}$	$z_2^{(k)}$	...	$z_{P_k-1}^{(k)}$
1	0	0	...	0	...	0	0	...	0	0	0	...	0
1	0	0	...	0	...	0	0	...	0	0	0	...	0
1	0	0	...	0	...	0	1	...	0	0	1	...	0
1	0	0	...	0	...	0	1	...	0	0	1	...	0
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	0	...	0	...	0	0	...	0	0	0	...	0
1	0	0	...	0	...	0	0	...	0	0	0	...	0
1	0	0	...	0	...	0	1	...	0	1	0	...	0
1	0	0	...	0	...	0	1	...	0	1	0	...	0
1	0	0	...	0	...	0	0	...	0	0	1	...	0
1	0	0	...	0	...	0	0	...	0	0	1	...	0
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	0	...	0	...	0	1	...	0	0	0	...	0
1	0	0	...	0	...	0	1	...	0	0	0	...	0
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	0	...	0	...	0	0	...	0	1	0	...	0
1	0	0	...	0	...	0	0	...	0	1	0	...	0
1	0	0	...	0	...	0	0	...	0	0	1	...	0
1	0	0	...	0	...	0	0	...	0	0	1	...	0
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	...	0	...	0	0	...	0	0	0	...	0
1	1	0	...	0	...	0	0	...	0	0	0	...	0

Tabla 3.5.: Elementos de la matriz de diseño $\boldsymbol{X}$ para un modelo naive

Al mismo tiempo los coeficientes del modelo para la variable indicadora $z_r^{(k)}$ están dados por $\beta_r^{(k)}$, donde k hace referencia a la covariable X_k y r a la r-ésima categoría de esta covariable y el primer componente del vetor es β_0 . Se obtiene el vector de parámetros $\boldsymbol{\beta}$

$$\boldsymbol{\beta'} = \left[\beta_0, \beta_1^{(1)}, \beta_2^{(1)}, \ldots, \beta_{P_1-1}^{(1)}, \ldots, \beta_1^{(k-1)}, \beta_2^{(k-1)}, \ldots, \beta_{P_{k-1}-1}^{(k-1)}, \ldots, \beta_1^{(k)}, \beta_2^{(k)}, \ldots, \beta_{P_k-1}^{(k)}\right]$$

El vector de error $\boldsymbol{\epsilon}$ es

$$\boldsymbol{\epsilon'} = \left[\epsilon_1^{(1)}, \epsilon_2^{(1)}, \ldots, \epsilon_1^{(i)}, \epsilon_2^{(i)}, \ldots, \epsilon_1^{(I)}, \epsilon_2^{(I)}\right]$$

Teniendo en cuenta lo anterior, el modelo para la i-ésima población sería:

$$\begin{aligned}\hat{f}_1^{(i)} =&\beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_2^{(1)} z_2^{(1)} + \cdots + \beta_{P_1-1}^{(1)} z_{P_1-1}^{(1)} + \cdots + \beta_1^{(k-1)} z_1^{(k-1)} + \beta_2^{(k-1)} z_2^{(k-1)} \\ &+ \cdots + \beta_{P_{k-1}-1}^{(k-1)} z_{P_{k-1}-1}^{(k-1)} + \cdots + \beta_1^{(k)} z_1^{(k)} + \beta_2^{(k)} z_2^{(k)} + \cdots + \beta_{P_{k-1}-1}^{(k)} z_{P_{k-1}-1}^{(k)} + \epsilon_1^{(i)}\end{aligned}$$

$$\begin{aligned}\hat{f}_2^{(i)} =&\beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_2^{(1)} z_2^{(1)} + \cdots + \beta_{P_1-1}^{(1)} z_{P_1-1}^{(1)} + \cdots + \beta_1^{(k-1)} z_1^{(k-1)} + \beta_2^{(k-1)} z_2^{(k-1)} \\ &+ \cdots + \beta_{P_{k-1}-1}^{(k-1)} z_{P_{k-1}-1}^{(k-1)} + \cdots + \beta_1^{(k)} z_1^{(k)} + \beta_2^{(k)} z_2^{(k)} + \cdots + \beta_{P_{k-1}-1}^{(k)} z_{P_{k-1}-1}^{(k)} + \epsilon_2^{(i)}\end{aligned}$$

Matriz de Diseño para el Modelo General

Para generalizar el modelo naive y generar diferentes probabilidades para las respuestas f_1 y f_2 se quiere determinar las probabilidades de manera independiente para cada respuesta planteada, se debe considerar el modelo Henao and Correa (2018):

$$\hat{\boldsymbol{f}} = [\boldsymbol{O} \circ \boldsymbol{X}]\, \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.45}$$

donde $\boldsymbol{O}$ es una matriz que tiene las mismas dimensiones que la matriz de diseño construida para el modelo naive, cuyas columnas son iguales con respecto a sus elementos contenidos y tienen la siguiente estructura:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & \cdots & 0 & 1 \end{bmatrix}$$

Entonces la matriz $\boldsymbol{O}$ tendría la forma:

$$\boldsymbol{O} = \begin{bmatrix} 0 & 0 & \ldots & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 & \ldots & 1 \\ 0 & 0 & \ldots & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 & \ldots & 1 \end{bmatrix}$$

Esto permite hacer referencia a la primera o a la segunda respuesta. Al realizar el producto

Hadamard entre la matriz $\boldsymbol{O}$ y la matriz $\boldsymbol{X}$, se produce la matriz de diseño para el modelo general. Es importante aclarar que debido al surgimiento de nuevos párametros relacionados con f_2 se debe introducir una notación para distinguirlos, siendo así $z_r^{(k,2)}$ la notación establecida, donde k hace referencia a la covariable X_k, r a la r-ésima categoría de esta covariable y el número $'2$ hace referencia a su relación con f_2.

Haciendo uso de la nueva notación la matriz de diseño generalizada tiene la forma:

β_0	$z_1^{(1)}$	$\dots$	$z_{P_1-1}^{(1)}$	$\dots$	$z_1^{(k)}$	$z_2^{(k)}$	$\dots$	$z_{P_k-1}^{(k)}$	$\beta_0^{(,2)}$	$z_1^{(1,2)}$	$\dots$	$z_{P_1-1}^{(1,2)}$	$\dots$	$z_1^{(k,2)}$	$z_2^{(k,2)}$	$\dots$	$z_{P_k-1}^{(k,2)}$
1	0	$\dots$	0	$\dots$	0	0	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	0	$\dots$	0	1	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	1	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	1	$\dots$	0	1	0	$\dots$	0	$\dots$	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	$\dots$	0	$\dots$	0	0	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	0	$\dots$	0	1	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	1	0	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	1	0	$\dots$	0	1	0	$\dots$	0	$\dots$	1	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	1	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	1	$\dots$	0	1	0	$\dots$	0	$\dots$	0	1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	$\dots$	1	$\dots$	0	0	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	1	$\dots$	0	0	$\dots$	0	1	0	$\dots$	1	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	0	$\dots$	1	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	0	$\dots$	0	$\dots$	0	0	$\dots$	1	1	0	$\dots$	0	$\dots$	0	0	$\dots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	$\dots$	0	$\dots$	0	0	$\dots$	0	0	0	$\dots$	0	$\dots$	0	0	$\dots$	0
1	1	$\dots$	0	$\dots$	0	0	$\dots$	0	1	1	$\dots$	0	$\dots$	0	0	$\dots$	0

Al mismo tiempo los coeficientes del modelo para la variable indicadora $z_r^{(k,2)}$ están dados por $\beta_r^{(k,2)}$, donde k hace referencia a la covariable X_k y r a la r-ésima categoría de esta covariable y el primer componente del vetor es $\beta_0^{(,2)}$. Se obtiene el vector de parámetros $\boldsymbol{\beta}$

$$\boldsymbol{\beta'} = \Big[\beta_0, \beta_1^{(1)}, \beta_2^{(1)}, \dots, \beta_{P_1-1}^{(1)}, \dots, \beta_1^{(k-1)}, \beta_2^{(k-1)}, \dots, \beta_{P_{k-1}-1}^{(k-1)}, \dots, \beta_1^{(k)}, \beta_2^{(k)}, \dots, \beta_{P_k-1}^{(k)},$$
$$\beta_0^{(,2)}, \beta_1^{(1,2)}, \beta_2^{(1,2)}, \dots, \beta_{P_1-1}^{(1,2)}, \dots, \beta_1^{(k-1,2)}, \beta_2^{(k-1,2)}, \dots, \beta_{P_{k-1}-1}^{(k-1,2)}, \dots, \beta_1^{(k,2)}, \beta_2^{(k,2)}, \dots, \beta_{P_k-1}^{(k,2)}\Big]$$

El vector de error $\boldsymbol{\epsilon}$ es

$$\boldsymbol{\epsilon'} = \left[\epsilon_1^{(1)}, \epsilon_2^{(1)}, \dots, \epsilon_1^{(i)}, \epsilon_2^{(i)}, \dots, \epsilon_1^{(I)}, \epsilon_2^{(I)}\right]$$

Teniendo en cuenta lo anterior el modelo para la i-ésima población sería:

$$\begin{aligned}\hat{f}_1^{(i)} =&\beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_2^{(1)} z_2^{(1)} + \cdots + \beta_{P_1-1}^{(1)} z_{P_1-1}^{(1)} + \cdots + \beta_1^{(k-1)} z_1^{(k-1)} + \beta_2^{(k-1)} z_2^{(k-1)} \\ &+ \cdots + \beta_{P_{k-1}-1}^{(k-1)} z_{P_{k-1}-1}^{(k-1)} + \cdots + \beta_1^{(k)} z_1^{(k)} + \beta_2^{(k)} z_2^{(k)} + \cdots + \beta_{P_{k-1}-1}^{(k)} z_{P_{k-1}-1}^{(k)} + \epsilon_1^{(i)}\end{aligned}$$

$$\begin{aligned}\hat{f}_2^{(i)} =& \left(\beta_0 + \beta_0^{(,2)}\right) + \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) z_1^{(1,2)} + \left(\beta_2^{(1)} + \beta_2^{(1,2)}\right) z_2^{(1,2)} + \cdots + \left(\beta_{P_1-1}^{(1)} + \beta_{P_1-1}^{(1,2)}\right) z_{P_1-1}^{(1)} + \\ &\cdots + \left(\beta_1^{(k-1)} + \beta_1^{(k-1,2)}\right) z_1^{(k-1,2)} + \left(\beta_2^{(k-1)} + \beta_2^{(k-1,2)}\right) z_2^{(k-1,2)} + \cdots \\ &+ \left(\beta_{P_{k-1}-1}^{(k-1)} + \beta_{P_{k-1}-1}^{(k-1,2)}\right) z_{P_{k-1}-1}^{(k-1,2)} + \cdots + \left(\beta_1^{(k)} + \beta_1^{(k,2)}\right) z_1^{(k,2)} + \left(\beta_2^{(k)} + \beta_2^{(k,2)}\right) z_2^{(k,2)} + \cdots \\ &+ \left(\beta_{P_{k-1}-1}^{(k)} + \beta_{P_{k-1}-1}^{(k,2)}\right) z_{P_{k-1}-1}^{(k,2)} + \epsilon_2^{(i)}\end{aligned}$$

3.4.2. Estimación de parámetros del modelo

Para el modelo $\hat{\boldsymbol{f}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, dónde Var $(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\epsilon}}}$ y $\boldsymbol{\epsilon} \sim \text{AN}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\epsilon}}}\right)$ [1] el estimador via mínimos cuadrados ponderados de $\boldsymbol{\beta}$, el cual es un estimador BAN (best asymptotic normal) es:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}}^{-1}\hat{\boldsymbol{f}} \tag{3.46}$$

$\hat{\boldsymbol{\beta}}$ es un estimador asintóticamente insesgado de $\boldsymbol{\beta}$ y es el valor que minimiza la ecuación:

$$S\left(\boldsymbol{\beta}\right) = \left(\hat{\boldsymbol{f}} - \boldsymbol{X}\boldsymbol{\beta}\right)' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}}^{-1} \left(\hat{\boldsymbol{f}} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{3.47}$$

La matriz de varianzas y covarianzas de $\hat{\boldsymbol{\beta}}$ está dada por $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$, y su estimación está dada por:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}}^{-1}\boldsymbol{X}\right)^{-1} \tag{3.48}$$

La estimacion de $\hat{\boldsymbol{f}}$ denotada $\hat{\boldsymbol{f}}*$ está dada por:

$$\hat{\boldsymbol{f}}* = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{3.49}$$

Entonces la matriz estimada de varianzas y covarianzas de $\hat{\boldsymbol{f}}*$ denotada por $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}*}$ es

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}*} = \boldsymbol{X}\left(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{f}}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \tag{3.50}$$

[1] AN asintóticamente normal

Estos resultados son asintóticos (Agresti (1996a), Grizzle et al. (1969) y Rao et al. (2008)).

Cabe resaltar que las matrices de varianzas y covarianzas anteriormente mencionadas para la estimación de los parámetros del modelo lineal varían dependiendo de la forma en que desee obtenerse la sensibilidad y la especificidad, es decir, en forma directa, logarítmica o aplicando el logit como se plantearon en las subsecciones 3.2.1, 3.2.2 y 3.2.3.

3.5. Residuales de la sensibilidad y la especificidad

Para determinar la adecuación del ajuste del modelo se hace un análisis de los residuos. Para el modelo GSK, se puede definir residuos en términos de la función respuesta y de la tabla de conteo.

3.5.1. Residuales para la función respuesta

Los residuales se definen como la diferencia entre los valores observados de $\boldsymbol{f}$ denotados $\hat{\boldsymbol{f}}$ y los ajustados denotados $\hat{\boldsymbol{f}}*$:

$$\hat{\boldsymbol{e}} = \hat{\boldsymbol{f}} - \hat{\boldsymbol{f}}* \tag{3.51}$$

A partir del modelo GSK se pueden obtener los residuales con respecto a la sensibilidad y a la especificidad por subpoblaciones.

- Para la sensibilidad:

  Para los residuales de la sensibilidad se tiene que $\hat{\boldsymbol{e}}_{\boldsymbol{2i-1}}$ es el $(2i-1)$-ésimo elemento del vector $\hat{\boldsymbol{e}}$ y corresponde al residual para la i-ésima subpoblación:

  $$\hat{\boldsymbol{e}}_{\boldsymbol{2i-1}} = \hat{f}_{2i-1} - \hat{f}*_{2i-1} \tag{3.52}$$

  donde, $\hat{f}_{2i-1}$ es el $(2i-1)$-ésimo elemento observado y $\hat{f}*_{2i-1}$ corresponde al $(2i-1)$-ésimo elemento del vector predicho por el modelo.

  Los residuales estandarizados y pseudoestandarizados están dados por:

  $$r_{2i-1} = \frac{\hat{f}_{2i-1} - \hat{f}*_{2i-1}}{s.e\left(\hat{f}*_{2i-1}\right)} \tag{3.53}$$

$$r.p_{2i-1}^{(i)} = \frac{\hat{f}_{2i-1} - \hat{f}*_{2i-1}}{\sqrt{d_{2i-1}}} \tag{3.54}$$

donde d_{2i-1} es el $(2i-1)$-ésimo elemento de la diagonal principal de la matriz $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{e}}}$.

- Para la especificidad:
  Para los residuales de la especificidad se tiene que $\hat{\boldsymbol{e}}_{\boldsymbol{2i}}$ es el $(2i)$-ésimo elemento del vector $\hat{\boldsymbol{e}}$ y corresponde al residual para la i-ésima subpoblación:

$$\hat{\boldsymbol{e}}_{\boldsymbol{2i}} = \hat{f}_{2i} - \hat{f}*_{2i} \tag{3.55}$$

donde, $\hat{f}_{2i}$ es el $(2i)$-ésimo elemento observado y $\hat{f}*_{2i}$ corresponde al $(2i)$-ésimo elemento del vector predicho por el modelo.

Los residuales estandarizados y pseudoestandarizados están dados por:

$$r_{2i} = \frac{\hat{f}_{2i} - \hat{f}*_{2i}}{s.e\left(\hat{f}*_{2i}\right)} \tag{3.56}$$

$$r.p_{2i}^{(i)} = \frac{\hat{f}_{2i} - \hat{f}*_{2i}}{\sqrt{d_{2i}}} \tag{3.57}$$

donde $s.e$ es el error estándar y d_{2i} es el $(2i)$-ésimo elemento de la diagonal principal de la matriz $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{e}}}$.

4. Ilustración con datos pseudo-reales

En este capítulo se expondrá la obtención de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables empleando datos pseudo reales.

4.1. Ejemplo alfa-fetoproteína

Consideramos una nueva prueba para detectar la alfa-fetoproteína como predictor de un defecto del cierre del tubo neural para una población de 10 000 mujeres de alto riesgo y 100 000 de bajo riesgo como se menciona en el ejemplo de Henquin (2013). Y a su vez proponemos una población de 10 000 hombres de alto riesgo y 100 000 de bajo riesgo para realizar la ilustración con tamaños de muestra iguales, e introducir la covariable sexo. Los individuos se clasificaron como sanos o enfermos (presenta defecto en el tubo neuronal) como se muestra en la siguiente tabla:

			EVENTOS AL NACER			
Subpoblación	**Sexo**	**Riesgo**	**alfa-fetoproteína**	**Defecto del tubo neuronal**	**Sano**	**Total**
1	Mujer	Alto	Anormal	87	18	
			Normal	13	9882	10 000
2	Mujer	Bajo	Anormal	128	179	
			Normal	19	99 674	100 000
3	Hombre	Alto	Anormal	75	116	
			Normal	86	9723	10 000
4	Hombre	Bajo	Anormal	108	267	
			Normal	61	99 564	100 000

Tabla 4.1.: Tabla de contingencia clasificación aplicando el método gold estándar

En este caso, se desea conocer la sensibilidad y la especificidad de la prueba en presencia de las covariables sexo y riesgo. La variable sexo tiene 2 niveles: hombre y mujer, mientras que la variable riesgo tiene 2 niveles: alto y bajo. Con lo cual, se obtienen 4 subpoblaciones: mujer con alto riesgo, mujer con bajo riesgo, hombre con alto riesgo y hombre con bajo riesgo correspondientes a las posibles combinaciones de los diferentes niveles de las variables.

La tabla 4.1 se puede reescribir como la tabla de contingencia 4.2 donde cada subpoblación se reorganiza para las diferentes categorías (P^+C^+, P^-C^+, P^+C^-, P^-C^-). Cabe resaltar

que C^+ representa a las personas con defecto del tubo neural, C^- a las personas sanas , P^+ representa resultados anormales de la alfa-fetoproteína según la nueva prueba y P^- resultados normales de la alfa-fetoproteína según la nueva prueba. Luego, la combinación entre P^+, P^- con C^+ y C^- significaría:

P^+C^+: la prueba arroja resultados anormales de la alfa-fetoproteína en personas con defecto del tubo neural.
P^-C^+: la prueba arroja resultados normales de la alfa-fetoproteína en personas con defecto del tubo neural.
P^+C^-: la prueba arroja resultados anormales de la alfa-fetoproteína en personas sanas.
P^-C^-: la prueba arroja resultados normales de la alfa-fetoproteína en personas sanas.

Subpoblación	**Sexo**	**Riesgo**	$\boldsymbol{P^+C^+}$	$\boldsymbol{P^-C^+}$	$\boldsymbol{P^+C^-}$	$\boldsymbol{P^-C^-}$	**Total**
1	Mujer	Alto	87	13	18	9882	10 000
2	Mujer	Bajo	128	19	179	99674	100 000
3	Hombre	Alto	75	86	116	9723	10 000
4	Hombre	Bajo	108	61	267	99564	100 000

Tabla 4.2.: Tabla de contingencia por subpoblaciones y categorías

La distribución de probabilidad de la tabla 4.2 viene dada por:

Subpoblación	$\boldsymbol{P^+C^+}$	$\boldsymbol{P^-C^+}$	$\boldsymbol{P^+C^-}$	$\boldsymbol{P^-C^-}$	**Total**
1	0.00870	0.00130	0.00180	0.98820	1
2	0.00128	0.00019	0.00179	0.99674	1
3	0.00750	0.00860	0.01160	0.97230	1
4	0.00087	0.00013	0.00018	0.09882	1

Tabla 4.3.: Tabla distribución de probabilidad

4.1.1. Sensibilidad y especificidad directas

Se trabajarán las funciones de forma directa, por lo cual definimos:

$$f_1{}^{(i)} = \text{sensibilidad} = \frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)} + \pi_{12}^{(i)}} \quad ; \qquad i = 1, 2, 3, 4 \tag{4.1}$$

$$f_2{}^{(i)} = \text{especificidad} = \frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)} + \pi_{22}^{(i)}} \quad ; \qquad i = 1, 2, 3, 4 \tag{4.2}$$

Para su construcción hallamos los elementos:

$$\boldsymbol{\pi'} = \begin{bmatrix} \pi_{11}^{(1)} & \pi_{12}^{(1)} & \pi_{21}^{(1)} & \pi_{22}^{(1)} & \pi_{11}^{(2)} & \pi_{12}^{(2)} & \pi_{21}^{(2)} & \pi_{22}^{(2)} & \dots & \pi_{11}^{(4)} & \pi_{12}^{(4)} & \pi_{21}^{(4)} & \pi_{22}^{(4)} \end{bmatrix} \tag{4.3}$$

Donde $\boldsymbol{\pi}$ es un vector columna de dimensiones 16×1 y $\pi_{11}^{(1)} = 0.00870$ representa la probabilidad teórica de que a una mujer con alto riesgo que tengan defecto del tubo neural la prueba le arroje resultados anormales de alfa-fetoproteína.

$$\boldsymbol{A\pi} = \begin{bmatrix} \pi_{11}^{(1)} \\ \pi_{11}^{(1)} + \pi_{12}^{(1)} \\ \pi_{22}^{(1)} \\ \pi_{21}^{(1)} + \pi_{22}^{(1)} \\ \pi_{11}^{(2)} \\ \pi_{11}^{(1)} + \pi_{12}^{(i)} \\ \pi_{22}^{(2)} \\ \pi_{21}^{(2)} + \pi_{22}^{(i)} \\ \vdots \\ \pi_{11}^{(4)} \\ \pi_{11}^{(4)} + \pi_{12}^{(4)} \\ \pi_{22}^{(4)} \\ \pi_{21}^{(4)} + \pi_{22}^{(4)} \end{bmatrix} \tag{4.4}$$

Para que se obtenga (4.4) la matriz $\boldsymbol{A}$ de dimensiones 16×16 se construye como:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.5}$$

Se considera un nuevo vector con los logaritmos naturales de los componentes del vector $\boldsymbol{A\pi}$ de dimensiones 16×1 y luego se premultiplica por una matriz $\boldsymbol{K}$ de dimensiones 8×16, donde:

$$K = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Por último, se exponencia componente a componente el vector $\boldsymbol{K}\ln(\boldsymbol{A\pi})$, y se multiplica por una matriz identidad $\boldsymbol{Q}$ de dimensiones 8×8. Con las operaciones anteriores se obtiene el vector respuesta $\mathbf{f}_{\text{exp}}$:

$$\mathbf{f}_{\text{exp}} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ f_1^{(2)} \\ f_2^{(2)} \\ f_1^{(3)} \\ f_2^{(3)} \\ f_1^{(4)} \\ f_2^{(4)} \end{bmatrix} = \begin{bmatrix} 0.87 \\ 0.9981818 \\ 0.8707483 \\ 0.9982074 \\ 0.4658385 \\ 0.9882102 \\ 0.87 \\ 0.9981818 \end{bmatrix} \tag{4.6}$$

Adicionalmente, haciendo los cálculos correspondientes la matriz estimada de varianzas y covarianza de $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{exp}}}$ es:

$$\begin{bmatrix} 0.3367 & 0.3185 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3185 & 0.3575 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.2345 & 0.0325 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0325 & 0.0101 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0026 & 0.0033 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0033 & 0.0193 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.6221 & 0.5638 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5638 & 1.3331 \end{bmatrix} \tag{4.7}$$

De la matriz de varianzas y covarianzas podemos hacer un análisis con respecto a la relación entre la sensibilidad (se) y la especificidad (sp) calculando la correlación ($\rho_{se,sp}^{(i)}$) entre ambas

en cada subpoblación (i).
Para la primera subpoblación de mujeres con alto riesgo se tiene que la correlación

$$\rho_{se,sp}^{(1)} = \frac{0.3185}{\sqrt{(0.3367 * 0.3575)}} = 0.918 \quad ,$$

lo cual indica una muy fuerte relación lineal entre la sensibilidad y la especificidad, mientras que para las subpoblaciones 2 y 4 la correlación es $\rho_{se,sp}^{(2)} = 0.669$ y $\rho_{se,sp}^{(4)} = 0.619$ respectivamente, lo cual indica una relación lineal moderada entre la sensibilidad y la especificidad. Por último, para la tercera subpoblación la correlación $\rho_{se,sp}^{(3)} = 0.474$ sugiriendo una relación lineal débil entre la sensibilidad y la especificidad.

Para la construcción del modelo lineal general bajo la metodología GSK, hallamos la matriz de diseño $\boldsymbol{O} \circ \boldsymbol{X}$ y la estimación de $\boldsymbol{\beta}$. La matriz de diseño es:

β_0	$z_1^{(1)}$	$z_1^{(2)}$	$\beta_0^{(,2)}$	$z_1^{(1,2)}$	$z_1^{(2,2)}$
1	0	0	0	0	0
1	0	0	1	0	0
1	0	1	0	0	0
1	0	1	1	0	1
1	1	0	0	0	0
1	1	0	1	1	0
1	1	1	0	0	0
1	1	1	1	1	1

Tabla 4.4.: Elementos de la matriz de diseño $\boldsymbol{X}$ para un modelo general en el ejemplo

donde para f_1, la sensibilidad: $z_1^{(1)}$ = Mujer , $z_2^{(1)}$ = Hombre , este último es el nivel de referencia para la primera variable, $z_1^{(2)}$ = Riesgo alto y $z_2^{(2)}$ = Riesgo bajo, el cual es el nivel de referencia para la segunda variable.
Y con respecto a f_2, la especificidad: $z_1^{(1,2)}$ = Mujer , $z_2^{(1,2)}$ = Hombre , este último es el nivel de referencia para la primera variable, $z_1^{(2,2)}$ = Riesgo alto y $z_2^{(2,2)}$ = Riesgo bajo, el cual es el nivel de referencia para la segunda variable.

El vector de parametros $\boldsymbol{\beta}$ estaría dado por :

$$\boldsymbol{\beta'} = \left[\beta_0\, \beta_1^{(1)}, \beta_1^{(2)}, \beta_0^{(,2)}, \beta_1^{(1,2)}, \beta_1^{(2,2)}\right]$$

El vector de error $\boldsymbol{\epsilon}$ es

$$\boldsymbol{\epsilon'} = \left[\epsilon_1^{(1)}, \epsilon_2^{(1)}, \epsilon_1^{(2)}, \epsilon_2^{(2)}, \epsilon_1^{(3)}, \epsilon_2^{(4)}, \epsilon_1^{(4)}, \epsilon_2^{(4)}\right]$$

El modelo de la sensibilidad para la i-ésima población es:

$$\hat{f}_1^{(i)} = \beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_1^{(2)} z_1^{(2)} + \epsilon_1^{(i)}$$

El modelo de la especificidad para la i-ésima población es:

$$\hat{f}_2^{(i)} \quad = \quad \left(\beta_0 + \beta_0^{(,2)}\right) \quad + \quad \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) z_1^{(1,2)} \quad + \quad \left(\beta_1^{(2)} + \beta_1^{(2,2)}\right) z_1^{(2,2)} \quad + \quad \epsilon_2^{(i)}$$

Utilizando (3.46) y las matrices de varianzas y covarianzas necesarias y correspondientes al caso directo de las sensibilidad y la especificidad se tiene como resultado:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.7741 \\ -0.3078 \\ 0.2193 \\ 0.1500 \\ 0.3681 \\ -0.1298 \end{bmatrix}$$

De lo anterior $\hat{\boldsymbol{f}}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$ es igual a:

$$\hat{\boldsymbol{f}}^* = \begin{bmatrix} 0.7741 \\ 0.9241 \\ 0.9934 \\ 1.0136 \\ 0.4663 \\ 0.9844 \\ 0.6856 \\ 1.0739 \end{bmatrix}$$

Finalmente, obtendríamos que para las mujeres con alto riesgo la sensibilidad y la especificidad son 0.7741 y 0.9241 respectivamente, para las mujeres con bajo riesgo la sensibilidad y la especificidad son 0.9934 y 1.0136 respectivamente, para los hombres con alto riesgo la sensibilidad y la especificidad son 0.4663 y 0.9844 respectivamente y por último para los hombres

con bajo riesgo la sensibilidad y la especificidad son 0.6856 y 1.0739 respectivamente.

Es importante resaltar que para la segunda y cuarta subpoblación la especificidad da un valor un poco mayor a 1. Esto se debe a que la estimación de la sensibilidad y la especificidad se han hecho de forma directa, que tiene como desventaja que en caso de que las probabilidades muestrales (tabla 3.4 o en el ejemplo tabla 4.3) sean muy cercanas a 1 o 0, es decir a los extremos, la estimación puede presentar inconsistencias. Para superar lo expuesto anteriormente se sugiere utilizar el modelo logit de la sensibilidad y la especificidad.

4.1.2. Sensibilidad y especificidad en forma logarítmica

Se trabajarán las funciones de forma logarítmica, por lo cual definimos:

$$f_1{}^{(i)} = \ln(\text{sensibilidad}) = \ln\left(\frac{\pi_{11}^{(i)}}{\pi_{11}^{(i)} + \pi_{12}^{(i)}}\right) \tag{4.8}$$

$$f_2{}^{(i)} = \ln(\text{especificidad}) = \ln\left(\frac{\pi_{22}^{(i)}}{\pi_{21}^{(i)} + \pi_{22}^{(i)}}\right) \tag{4.9}$$

Para su construcción hallamos los elementos:

$$\boldsymbol{\pi'} = \left[\begin{array}{cccccccccccccc} \pi_{11}^{(1)} & \pi_{12}^{(1)} & \pi_{21}^{(1)} & \pi_{22}^{(1)} & \pi_{11}^{(2)} & \pi_{12}^{(2)} & \pi_{21}^{(2)} & \pi_{22}^{(2)} & \dots & \pi_{11}^{(4)} & \pi_{12}^{(4)} & \pi_{21}^{(4)} & \pi_{22}^{(4)} \end{array}\right] \tag{4.10}$$

Donde $\boldsymbol{\pi}$ es un vector columna de dimensiones 16×1 y $\pi_{11}^{(1)} = 0.00870$ representa la probabilidad teórica de que a una mujer con alto riesgo que tengan defecto del tubo neural la prueba le arroje resultados anormales de alfa-fetoproteína.

La matriz $\boldsymbol{A}$ de dimensiones 16×16 se construye como:

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix} \tag{4.11}$$

De manera que el vector $\boldsymbol{A\pi}$ (4.4) sería:

$$\boldsymbol{A\pi} = \begin{bmatrix}
0.00870 \\
0.01000 \\
0.98820 \\
0.99000 \\
0.00128 \\
0.00147 \\
0.99674 \\
0.99853 \\
0.00750 \\
0.01610 \\
0.97230 \\
0.98390 \\
0.00087 \\
0.00100 \\
0.09882 \\
0.09900
\end{bmatrix} \tag{4.12}$$

Se considera un nuevo vector con los logaritmos naturales de los componentes del vector $\boldsymbol{A\pi}$ de dimensiones 16×1 y luego se premultiplica por una matriz $\boldsymbol{K}$ de dimensiones 8×16, donde:

$$K = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Con las operaciones anteriores se obtiene el vector respuesta $\mathbf{f}_{\log}$:

$$\mathbf{f}_{\log} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ f_1^{(2)} \\ f_2^{(2)} \\ f_1^{(3)} \\ f_2^{(3)} \\ f_1^{(4)} \\ f_2^{(4)} \end{bmatrix} = \begin{bmatrix} -0.139262067 \\ -0.001819837 \\ -0.138402323 \\ -0.001794244 \\ -0.763916251 \\ -0.011859867 \\ -0.139262067 \\ -0.001819837 \end{bmatrix} \tag{4.13}$$

Adicionalmente, haciendo los cálculos correspondientes la matriz estimada de varianzas y covarianza de $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\log}}$ es:

$$\begin{bmatrix} 0.4449 & 0.3667 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3667 & 0.3588 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.3093 & 0.0374 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0374 & 0.0101 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0118 & 0.0072 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0072 & 0.0197 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.8219 & 0.6492 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.6492 & 1.3379 \end{bmatrix} \tag{4.14}$$

De la matriz de varianzas y covarianzas podemos hacer un análisis con respecto a la relación entre la sensibilidad (se) y la especificidad (sp) calculando la correlación ($\rho_{se,sp}^{(i)}$) entre ambas en cada subpoblación (i).

Para la primera subpoblación de mujeres con alto riesgo se tiene que la correlación

$$\rho_{se,sp}^{(1)} = \frac{0.3667}{\sqrt{(0.4449 * 0.3588)}} = 0.918 \quad ,$$

lo cual indica una muy fuerte relación lineal entre la sensibilidad y la especificidad, mientras que para las subpoblaciones 2 y 4 la correlación es $\rho^{(2)}_{se,sp} = 0.669$ y $\rho^{(4)}_{se,sp} = 0.619$ respectivamente, lo cual indica una relación lineal moderada entre la sensibilidad y la especificidad. Por último, para la tercera subpoblación la correlación $\rho^{(3)}_{se,sp} = 0.474$ sugiriendo una relación lineal débil entre la sensibilidad y la especificidad.

Para la construcción del modelo lineal general bajo la metodología GSK, hallamos la matriz de diseño $\boldsymbol{O} \circ \boldsymbol{X}$ y la estimación de $\boldsymbol{\beta}$. La matriz de diseño es:

β_0	$z_1^{(1)}$	$z_1^{(2)}$	$\beta_0^{(,2)}$	$z_1^{(1,2)}$	$z_1^{(2,2)}$
1	0	0	0	0	0
1	0	0	1	0	0
1	0	1	0	0	0
1	0	1	1	0	1
1	1	0	0	0	0
1	1	0	1	1	0
1	1	1	0	0	0
1	1	1	1	1	1

Tabla 4.5.: Elementos de la matriz de diseño $\boldsymbol{X}$ para un modelo general en el ejemplo

donde para f_1, la sensibilidad: $z_1^{(1)}$ = Mujer , $z_2^{(1)}$ = Hombre , este último es el nivel de referencia para la primera variable, $z_1^{(2)}$ = Riesgo alto y $z_2^{(2)}$ = Riesgo bajo, el cual es el nivel de referencia para la segunda variable.
Y con respecto a f_2, la especificidad: $z_1^{(1,2)}$ = Mujer , $z_2^{(1,2)}$ = Hombre , este último es el nivel de referencia para la primera variable, $z_1^{(2,2)}$ = Riesgo alto y $z_2^{(2,2)}$ = Riesgo bajo, el cual es el nivel de referencia para la segunda variable.

El vector de parametros $\boldsymbol{\beta}$ estaría dado por :

$$\boldsymbol{\beta'} = \left[\beta_0\, \beta_1^{(1)}, \beta_1^{(2)}, \beta_0^{(,2)}, \beta_1^{(1,2)}, \beta_1^{(2,2)}\right]$$

El vector de error $\boldsymbol{\epsilon}$ es

$$\boldsymbol{\epsilon'} = \left[\epsilon_1^{(1)}, \epsilon_2^{(1)}, \epsilon_1^{(2)}, \epsilon_2^{(2)}, \epsilon_1^{(3)}, \epsilon_2^{(4)}, \epsilon_1^{(4)}, \epsilon_2^{(4)}\right]$$

El modelo de la sensibilidad para la i-ésima población es:

$$\hat{f}_1^{(i)} = \beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_1^{(2)} z_1^{(2)} + \epsilon_1^{(i)}$$

El modelo de la especificidad para la i-ésima población es:

$$\hat{f}_2^{(i)} \quad = \quad \left(\beta_0 + \beta_0^{(,2)}\right) \quad + \quad \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) z_1^{(1,2)} \quad + \quad \left(\beta_1^{(2)} + \beta_1^{(2,2)}\right) z_1^{(2,2)} \quad + \quad \epsilon_2^{(i)}$$

Utilizando (3.46) y las matrices de varianzas y covarianzas necesarias y correspondientes al caso del logaritmo de las sensibilidad y la especificidad se tiene como resultado:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.2862 \\ -0.4728 \\ 0.3369 \\ 0.1858 \\ 0.5582 \\ -0.2176 \end{bmatrix}$$

De lo anterior $\hat{\boldsymbol{f}}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$ es igual a:

$$\hat{\boldsymbol{f}}^* = \begin{bmatrix} -0.2862 \\ -0.1004 \\ 0.0507 \\ 0.0189 \\ -0.7590 \\ -0.01498 \\ -0.4221 \\ 0.1043 \end{bmatrix}$$

La función entrega el logaritmo de la sensibilidad y la especificidad por subpoblación, sin embargo podemos tomar el exponencial componente a componente para obtener la sensibilidad y la especificidad. Se obtendría:

$$\begin{bmatrix} 0.7511 \\ 0.9044 \\ 1.05197 \\ 1.0190 \\ 0.4681 \\ 0.9851 \\ 0.6557 \\ 1.1099 \end{bmatrix}$$

Finalmente, obtendríamos que para las mujeres con alto riesgo la sensibilidad y la especificidad son 0.7511 y 0.9044 respectivamente, para las mujeres con bajo riesgo la sensibilidad y la especificidad son 1.05197 y 1.0190 respectivamente, para los hombres con alto riesgo la sensibilidad y la especificidad son 0.4681 y 0.9851 respectivamente y por último para los hombres con bajo riesgo la sensibilidad y la especificidad son 0.6557 y 1.1099 respectivamente.
Es importante resaltar que para la segunda y cuarta subpoblación la especificidad da un valor un poco mayor a 1. Esto se debe a que la estimación de la sensibilidad y la especificidad se han hecho de forma logaritmica, que tiene como desventaja que en caso de que las probabilidades muestrales (tabla 3.4 o en el ejemplo tabla 4.3) sean muy cercanas a 1 o 0, es decir a los extremos, la estimación puede presentar inconsistencias. Para superar lo expuesto anteriormente se sugiere utilizar el modelo logit de la sensibilidad y la especificidad.

4.1.3. Sensibilidad y especificidad en forma logit

Se trabajarán las funciones de forma directa, por lo cual definimos:

$$f_1{}^{(i)} = \text{logit(sensibilidad)} = \ln\left(\frac{\pi_1^{(i)}}{1-\pi_1^{(i)}}\right) \tag{4.15}$$

$$f_2{}^{(i)} = \text{logit(especificidad)} = \ln\left(\frac{\pi_2^{(i)}}{1-\pi_2^{(i)}}\right) \tag{4.16}$$

Para su construcción hallamos los elementos:

$$\boldsymbol{\pi'} = \left[\begin{array}{cccccccccccccc} \pi_{11}^{(1)} & \pi_{12}^{(1)} & \pi_{21}^{(1)} & \pi_{22}^{(1)} & \pi_{11}^{(2)} & \pi_{12}^{(2)} & \pi_{21}^{(2)} & \pi_{22}^{(2)} & \dots & \pi_{11}^{(4)} & \pi_{12}^{(4)} & \pi_{21}^{(4)} & \pi_{22}^{(4)} \end{array}\right] \tag{4.17}$$

Donde $\boldsymbol{\pi}$ es un vector columna de dimensiones 16×1 y $\pi_{11}^{(1)} = 0.00870$ representa la probabilidad teórica de que a una mujer con alto riesgo que tengan defecto del tubo neural la prueba le arroje resultados anormales de alfa-fetoproteína.
La matriz $\boldsymbol{A}$ de dimensiones 16×16 se construye como:

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix} \tag{4.18}$$

De manera que el vector $\boldsymbol{A\pi}$ (4.4) sería:

$$\boldsymbol{A\pi} = \begin{bmatrix}
\pi_{11}^{(1)} \\ \pi_{12}^{(1)} \\ \pi_{22}^{(1)} \\ \pi_{21}^{(1)} \\
\pi_{11}^{(2)} \\ \pi_{12}^{(2)} \\ \pi_{22}^{(2)} \\ \pi_{21}^{(2)} \\
\pi_{11}^{(3)} \\ \pi_{12}^{(3)} \\ \pi_{22}^{(3)} \\ \pi_{21}^{(3)} \\
\pi_{11}^{(4)} \\ \pi_{12}^{(4)} \\ \pi_{22}^{(4)} \\ \pi_{21}^{(4)}
\end{bmatrix} = \begin{bmatrix}
0.00870 \\ 0.00130 \\ 0.98820 \\ 0.00180 \\
0.00128 \\ 0.00019 \\ 0.99674 \\ 0.00179 \\
0.00750 \\ 0.00860 \\ 0.97230 \\ 0.01160 \\
0.00087 \\ 0.00013 \\ 0.09882 \\ 0.00018
\end{bmatrix} \tag{4.19}$$

Se considera un nuevo vector con los logaritmos naturales de los componentes del vector $\boldsymbol{A\pi}$ de dimensiones 16×1 y luego se premultiplica por una matriz $\boldsymbol{K}$ de dimensiones 8×16, donde:

$$K = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Con las operaciones anteriores se obtiene el vector respuesta $\mathbf{f}_{\text{logit}}$:

$$\mathbf{f}_{\text{logit}} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ f_1^{(2)} \\ f_2^{(2)} \\ f_1^{(3)} \\ f_2^{(3)} \\ f_1^{(4)} \\ f_2^{(4)} \end{bmatrix} = \begin{bmatrix} 1.9009588 \\ 6.3080984 \\ 1.9075913 \\ 6.3222743 \\ -0.1368592 \\ 4.4286593 \\ 1.9009588 \\ 6.3080984 \end{bmatrix} \tag{4.20}$$

Adicionalmente, haciendo los cálculos correspondientes la matriz estimada de varianzas y covarianza de $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{f}}_{\mathbf{log}}}$ es:

$$\begin{bmatrix} 0.0884 & 7.5e^{-18} & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 0.3601002 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0604 & 0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0 & 0.0102 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.02496 & 0 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0 & 0.0202 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.3751 & 0.1909 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1909 & 1.3428 \end{bmatrix} \tag{4.21}$$

De la matriz de varianzas y covarianzas podemos hacer un análisis con respecto a la relación entre la sensibilidad (se) y la especificidad (sp) calculando la correlación ($\rho_{se,sp}^{(i)}$) entre ambas en cada subpoblación (i).

Para la primera subpoblación de mujeres con alto riesgo se tiene que la correlación

$$\rho_{se,sp}^{(1)} = \frac{0}{\sqrt{(0.0884 * 0.3601)}} = 0 \quad ,$$

lo cual indica que no existe relación lineal entre la sensibilidad y la especificidad, al igual que para las subpoblaciones 2 y 3 la correlación es $\rho^{(2)}_{se,sp} = 0$ y $\rho^{(4)}_{se,sp} = 0$. Por último, para la cuarta subpoblación la correlación $\rho^{(3)}_{se,sp} = 0.269$ sugiriendo una relación lineal débil entre la sensibilidad y la especificidad.

Para la construcción del modelo lineal general bajo la metodología GSK, hallamos la matriz de diseño $\boldsymbol{O} \circ \boldsymbol{X}$ y la estimación de $\boldsymbol{\beta}$. La matriz de diseño es:

β_0	$z_1^{(1)}$	$z_1^{(2)}$	$\beta_0^{(,2)}$	$z_1^{(1,2)}$	$z_1^{(2,2)}$
1	0	0	0	0	0
1	0	0	1	0	0
1	0	1	0	0	0
1	0	1	1	0	1
1	1	0	0	0	0
1	1	0	1	1	0
1	1	1	0	0	0
1	1	1	1	1	1

Tabla 4.6.: Elementos de la matriz de diseño $\boldsymbol{X}$ para un modelo general en el ejemplo

donde para f_1, la sensibilidad: $z_1^{(1)}$ = Mujer , $z_2^{(1)}$ = Hombre , este último es el nivel de referencia para la primera variable, $z_1^{(2)}$ = Riesgo alto y $z_2^{(2)}$ = Riesgo bajo, el cual es el nivel de referencia para la segunda variable.
Y con respecto a f_2, la especificidad: $z_1^{(1,2)}$ = Mujer , $z_2^{(1,2)}$ = Hombre , este último es el nivel de referencia para la primera variable, $z_1^{(2,2)}$ = Riesgo alto y $z_2^{(2,2)}$ = Riesgo bajo, el cual es el nivel de referencia para la segunda variable.

El vector de parametros $\boldsymbol{\beta}$ estaría dado por :

$$\boldsymbol{\beta'} = \left[\beta_0\, \beta_1^{(1)}, \beta_1^{(2)}, \beta_0^{(,2)}, \beta_1^{(1,2)}, \beta_1^{(2,2)}\right]$$

El vector de error $\boldsymbol{\epsilon}$ es

$$\boldsymbol{\epsilon'} = \left[\epsilon_1^{(1)}, \epsilon_2^{(1)}, \epsilon_1^{(2)}, \epsilon_2^{(2)}, \epsilon_1^{(3)}, \epsilon_2^{(4)}, \epsilon_1^{(4)}, \epsilon_2^{(4)}\right]$$

El modelo de la sensibilidad para la i-ésima población es:

$$\hat{f}_1^{(i)} = \beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_1^{(2)} z_1^{(2)} + \epsilon_1^{(i)}$$

El modelo de la especificidad para la i-ésima población es:

$$\hat{f}_2^{(i)} \quad = \quad \left(\beta_0 + \beta_0^{(,2)}\right) \quad + \quad \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) z_1^{(1,2)} \quad + \quad \left(\beta_1^{(2)} + \beta_1^{(2,2)}\right) z_1^{(2,2)} \quad + \quad \epsilon_2^{(i)}$$

Utilizando (3.46) y las matrices de varianzas y covarianzas necesarias y correspondientes al caso del logaritmo de las sensibilidad y la especificidad se tiene como resultado:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1.5952 \\ -1.6457 \\ 0.5215 \\ 4.4626 \\ 0.0307 \\ -0.2499 \end{bmatrix}$$

De lo anterior $\hat{\boldsymbol{f}}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$ es igual a:

$$\hat{\boldsymbol{f}}^* = \begin{bmatrix} 1.5952 \\ 6.0577 \\ 2.1166 \\ 6.3293 \\ -0.0505 \\ 4.4427 \\ 0.471 \\ 4.7143 \end{bmatrix}$$

La función entrega el logit de la sensibilidad y la especificidad por subpoblación, sin embargo podemos tomar el exponencial componente a componente y dividirlo entre uno más el exponencial componente a componente para obtener la sensibilidad y la especificidad. Los valores correspondientes serían:

$$\begin{bmatrix} 0.8313 \\ 0.9977 \\ 0.8925 \\ 0.9982 \\ 0.4874 \\ 0.9884 \\ 0.6156 \\ 0.9911 \end{bmatrix}$$

Finalmente, obtendríamos que para las mujeres con alto riesgo la sensibilidad y la especificidad son 0.8313 y 0.9977 respectivamente, para las mujeres con bajo riesgo la sensibilidad y la especificidad son 0.8925 y 0.9982 respectivamente, para los hombres con alto riesgo la sensibilidad y la especificidad son 0.4874 y 0.9884 respectivamente y por último para los hombres con bajo riesgo la sensibilidad y la especificidad son 0.6156 y 0.9911 respectivamente.
Es importante resaltar que a diferencia de los casos anteriores realizados de forma directa y logaritmica, con el logit en ninguna subpoblación la especificidad o sensibilidad da un valor mayor a 1. Esto se debe a que la estimación de la sensibilidad y la especificidad se han hecho de forma logit, que tiene como ventaja que en caso de que las probabilidades muestrales sean muy cercanas a 1 o 0, es decir a los extremos, no presenta inconsistencias al realizar la estimación.

4.2. Interpretación de los Parámetros

El modelo se reescribió en términos de variables indicadoras $Z_r^{(k)}$ para la sensibilidad y en términos de variables indicadoras $Z_r^{(k,2)}$ para la especificidad trabajando individualmente los niveles de cada covariable, estableciendo niveles de referencia y tomando a β_0 y $\beta_0 + \beta_0^{(,2)}$ como el promedio de la sensibilidad y la especificidad bajo los niveles de referencia.

Cuando la sensibilidad y la especificidad se plantean de forma directa, la interpretación de los parámetros es la usual, es decir si el signo es positivo tendrá una influencia positiva sobre la sensibilidad (especificidad) o si por el contrario el signo es negativo tendrá una influencia negativa sobre la sensibilidad (especificidad), de manera que cuando la variable indicadora $Z_r^{(k)}$ (o $Z_r^{(k,2)}$) toma el valor de 1 entonces la variable respuesta incrementa en $\beta_0 + \beta_r^{(k)}$ (o $\beta_0 + \beta_0^{(,2)} + \beta_r^{(k)} + \beta_r^{(k,2)}$) unidades. De manera que el valor del parámetro $\beta_r^{(k)}$ ($\beta_r^{(k)} + \beta_r^{(k,2)}$) asociado a la variable indicadora cuyo valor es 1 aporta a la sensibilidad (especificidad) el valor de la diferencia entre el nivel de referencia y la variable indicadora.

Cuando las observaciones muestrales se refieren a una sola variable indicadora los parámetros que no están relacionados con estas variables se convierten en cero. En el ejemplo, se tomaron

las variables Género (Hombre, mujer) y Riesgo (Alto, bajo) con niveles de referencia hombre y bajo.

Luego se tiene para la sensibilidad:

$$\begin{array}{rcll} Z_1^{(1)} & = & 1 & \text{Es una mujer} \\ & = & 0 & \text{Otro} \\ Z_1^{(2)} & = & 1 & \text{Riesgo Alto} \\ & = & 0 & \text{Otro} \end{array}$$

Para la especificidad

$$\begin{array}{rcll} Z_1^{(1,2)} & = & 1 & \text{Es una mujer} \\ & = & 0 & \text{Otro} \\ Z_1^{(2,2)} & = & 1 & \text{Riesgo Alto} \\ & = & 0 & \text{Otro} \end{array}$$

Para el modelo más general, las respuestas f_1 para la sensibilidad y f_2 para la especificidad en el que se determinan las probabilidades de manera independiente para cada una de las respuestas planteadas, están dadas por:

$$\hat{f}_1 = \beta_0 + \beta_1^{(1)} Z_1^{(1)} + \beta_1^{(2)} Z_1^{(2)} + \epsilon_1$$

$$\hat{f}_2 \quad = \quad \left(\beta_0 + \beta_0^{(,2)}\right) \quad + \quad \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) Z_1^{(1,2)} \quad + \quad \left(\beta_1^{(2)} + \beta_1^{(2,2)}\right) Z_1^{(2,2)} \quad + \quad \epsilon_2$$

Las funciones se puede interpretar como a continuación, nótese que este análisis supone la aditividad lineal de los coeficientes de las variables indicadoras. Para la sensibilidad se tiene:

$$\begin{array}{cl} \text{Hombre + Bajo riesgo} & E\left[f\left(\pi_1 | X\right)\right] = \beta_0 \\ \text{Bajo riesgo + Mujer} & E\left[f\left(\pi_1 | X\right)\right] = \beta_0 + \beta_1^{(1)} \\ \text{Hombre + Alto riesgo} & E\left[f\left(\pi_1 | X\right)\right] = \beta_0 + \beta_1^{(2)} \\ \text{Mujer + Alto riesgo} & E\left[f\left(\pi_1 | X\right)\right] = \beta_0 + \beta_1^{(1)} + \beta_1^{(2)} \end{array}$$

donde

$$\pi_1 = \frac{\pi_{11}}{\pi_{11} + \pi_{12}}$$

Para la especificidad se tiene:

$$\begin{array}{cc}
\text{Hombre + Bajo riesgo} & E\left[f\left(\pi_2|\, X\right)\right] = \beta_0 + \beta_0^{(,2)} \\
\text{Bajo riesgo + Mujer} & E\left[f\left(\pi_2|\, X\right)\right] = \left(\beta_0 + \beta_0^{(,2)}\right) + \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) \\
\text{Hombre + Alto riesgo} & E\left[f\left(\pi_2|\, X\right)\right] = \left(\beta_0 + \beta_0^{(,2)}\right) + \left(\beta_1^{(2)} + \beta_1^{(2,2)}\right) \\
\text{Mujer + Alto riesgo} & E\left[f\left(\pi_2|\, X\right)\right] = \left(\beta_0 + \beta_0^{(,2)}\right) + \left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) + \left(\beta_1^{(2)} + \beta_1^{(2,2)}\right)
\end{array}$$

donde

$$\pi_2 = \frac{\pi_{22}}{\pi_{21} + \pi_{22}}$$

4.3. Inferencia sobre el modelo

La significancia estadística de los diversos parámetros en β puede probarse mediante procedimientos estándar de regresión múltiple. Esto se hace escribiendo hipótesis en la forma (Johnson and Koch, 1970):

$$H_0 : \boldsymbol{C\beta} = 0$$

donde C es una matriz de hipótesis de coeficientes y es de rango completo.

1. **Prueba Global Sobre el Modelo**

   La prueba global para el modelo tiene en cuenta todos los coeficientes que componen a β excepto a β_0 y $\beta_0^{(,2)}$. Se plantean las siguientes pruebas de hipótesis:

   - Para la sensibilidad:

     $$H_0 : \quad \beta_1^{(1)} = \beta_2^{(1)} = \cdots = \beta_{P_1-1}^{(1)} = \cdots = \beta_1^{(k)} = \beta_2^{(k)} = \cdots = \beta_{P_k-1}^{(k)} = 0$$
     $$H_1 : \text{Al menos uno de los parámetros}$$
     $$\beta_1^{(1)}, \beta_2^{(1)}, \ldots, \beta_{P_1-1}^{(1)}, \ldots, \beta_1^{(k)}, \beta_2^{(k)}, \ldots, \beta_{P_k-1}^{(k)} \neq 0$$

   - Para la especificidad:

     $$H_0 : \quad \beta_1^{(1,2)} = \beta_2^{(1,2)} = \cdots = \beta_{P_1-1}^{(1,2)} = \cdots = \beta_1^{(k,2)} = \beta_2^{(k,2)} = \cdots = \beta_{P_k-1}^{(k,2)} = 0$$
     $$H_1 : \text{Al menos uno de los parámetros}$$
     $$\beta_1^{(1,2)}, \beta_2^{(1,2)}, \ldots, \beta_{P_1-1}^{(1,2)}, \ldots, \beta_1^{(k,2)}, \beta_2^{(k,2)}, \ldots, \beta_{P_k-1}^{(k,2)} \neq 0$$

Bajo el supuesto de H_0 verdadera, ninguna variable explica la sensibilidad (especificidad), es decir, la sensibilidad (especificidad) es igual entre las subpoblaciones.

El estadístico de prueba bajo H_0 verdadera está dado por:

$$\left(C\hat{\beta}\right)' \left(C\hat{\Sigma}_{\hat{\beta}}C'\right)^{-1} C\hat{\beta} \sim \chi^2_{(g)}$$

donde g es el rango de la matriz C.

En el caso del ejemplo utilizado se plantearía como sigue:

- Para la sensibilidad:

$$H_0: \quad \beta_1^{(1)} = \beta_1^{(2)} = 0$$
$$H_1: \text{Al menos uno de los parámetros} \quad \beta_1^{(1)}, \beta_1^{(2)} \neq 0$$

- Para la especificidad:

$$H_0: \quad \beta_1^{(1,2)} = \beta_1^{(2,2)} = 0$$
$$H_1: \text{Al menos uno de los parámetros } \beta_1^{(1,2)}, \beta_1^{(2,2)} \neq 0$$

Para el caso del modelo completo en el ejemplo utilizado la matriz C sería:

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Para el caso en que solo se quiera evaluar la sensibilidad la matriz C es:

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Luego $C\beta$ es

$$C\beta = \begin{bmatrix} \beta_1^{(1)} \\ \beta_1^{(2)} \end{bmatrix} \tag{4.22}$$

Para el caso en que solo se quiera evaluar especificidad la matriz C sería:

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Luego $C\beta$ es

$$C\beta = \begin{bmatrix} \beta_1^{(1,2)} \\ \beta_1^{(2,2)} \end{bmatrix} \tag{4.23}$$

2. **Pruebas a Coeficientes individuales**

Para probar la significancia de cualquier coeficiente relacionado con la sensibilidad se plantea la prueba de hipótesis:

$$\begin{array}{lll} H_0: & \beta_r^{(k)} = 0 & r = 1, 2, \ldots, P_k - 1, \quad k = 1, 2, \ldots, m \\ H_1: & \beta_r^{(k)} \neq 0 & \end{array}$$

El estadístico de prueba es:

$$\left(C\hat{\beta}\right)' \left(C\hat{\Sigma}_{\hat{\beta}}C'\right)^{-1} \left(C\hat{\beta}\right) \sim \chi^2_{(1)}$$

donde C es un vector fila asociado a cada coeficiente.

En el ejemplo, el vector C sería:

$$\beta_1^{(1)} \qquad C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\beta_1^{(2)} \qquad C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Para el caso de la especificidad la prueba de hipótesis se plantea como :

$$\begin{aligned} H_0: &\quad \beta_r^{(k)} + \beta_r^{(k,2)} = 0 \qquad r = 1, 2, \ldots, P_k - 1, \quad k = 1, 2, \ldots, m \\ H_1: &\quad \beta_r^{(k)} + \beta_r^{(k,2)} \neq 0 \end{aligned}$$

El estadístico de prueba es:

$$\left(C\hat{\beta}\right)' \left(C\hat{\Sigma}_{\hat{\beta}}C'\right)^{-1} \left(C\hat{\beta}\right) \sim \chi^2_{(1)} \tag{4.24}$$

donde C es un vector asociado al par de coeficientes establecidos en la hipótesis nula. De modo que para el ejemplo, se definiría como:

$$\left(\beta_1^{(1)} + \beta_1^{(1,2)}\right) \qquad C = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\left(\beta_1^{(2)} + \beta_1^{(2,2)}\right) \qquad C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

3. **Pruebas individuales sobre las variables**

Contrastamos conjuntamente el grupo de los parámetros que pertenecen a cada una de las categorías pertenecientes a una variable en especifico exceptuando la de referencia, con el fin de determinar si el conjunto de parámetros sirve o no a los objetivos del modelo, con lo cual se puede concluir si la variable aporta a la sensibilidad y a la especificidad.

Para cada variable X_k de la sensibilidad con $k = 1, 2, \ldots m$ se plantean las hipótesis como:

$$\begin{aligned} H_0: &\quad \beta_1^{(k)} = \beta_2^{(k)} = \beta_3^{(k)} = \cdots = \beta_{P_k-1}^{(k)} = 0 \\ H_1: &\text{ Al menos uno de los parámetros} \quad \beta_1^{(k)}, \beta_2^{(k)}, \beta_3^{(k)}, \ldots, \beta_{P_k-1}^{(k)} \neq 0 \end{aligned}$$

El estadístico de prueba está dado por:

$$\chi^2 = \left(C\hat{\beta}\right)' \left(C\hat{\hat{\Sigma}}_{\hat{\beta}}C'\right)^{-1} C\hat{\beta} \sim \chi^2_{(g)}$$

Donde C es una matriz asociada a los coeficientes establecidos en la hipótesis nula. En la r-ésima fila se pondría 1 en la posición correspondiente a $\beta_r^{(k)}$ con $r = 1, 2, \ldots, P_k - 1$ y cero en las demás celdas de la fila.

Para cada variable X_k de la especificidad con $k = 1, 2, \ldots m$ se plantean las hipótesis:

$$\begin{aligned} H_0 :&\quad \beta_1^{(k)} + \beta_1^{(k,2)} = \beta_2^{(k)} + \beta_2^{(k,2)} = \cdots = \beta_r^{(k)} + \beta_r^{(k,2)} = 0 \\ H_1 : \text{Al menos uno} &\quad \beta_1^{(k)} + \beta_1^{(k,2)}, \quad \beta_2^{(k)} + \beta_2^{(k,2)}, \ldots, \quad \beta_r^{(k)} + \beta_r^{(k,2)} \neq 0 \end{aligned}$$

El estadístico de prueba esta dado por:

$$\chi^2 = \left(C\hat{\beta}\right)' \left(C\hat{\hat{\Sigma}}_{\hat{\beta}}C'\right)^{-1} C\hat{\beta} \sim \chi^2_{(g)}$$

Donde C es una matriz asociada a los coeficientes establecidos en la hipótesis nula. En la r-ésima fila se pondrían 1 en la posición correspondiente a $\beta_r^{(k)}$, 1 en la posición correspondiente a $\beta_r^{(k,2)}$ con $r = 1, 2, \ldots, P_k - 1$, y cero en las demás celdas de la fila.

4. **Pruebas de Contraste sobre las Variables**

Esta prueba nos ayuda a concluir si todas las categorías pertenecientes a una variable en especifico exceptuando la de referencia tienen el mismo nivel de influencia o por el contrario si influyen en mayor o menor medida en la sensibilidad y la especificidad.

Para cada variable X_k de la sensibilidad con $k = 1, 2, \ldots m$ se plantean las hipótesis como:

$$\begin{aligned} H_0 :&\quad \beta_1^{(k)} = \beta_2^{(k)} = \beta_3^{(k)} = \cdots = \beta_{P_k-1}^{(k)} \\ H_1 : \text{Al menos uno de los parámetros} &\quad \beta_a^{(k)} \neq \beta_b^{(k)} \\ \text{con} &\quad a \neq b, \quad a, b \in r, \quad r = 1, 2, \ldots, P_k - 1 \end{aligned}$$

El estadístico de prueba esta dado por:

$$\chi^2 = \left(C\hat{\beta}\right)' \left(C\hat{\hat{\Sigma}}_{\hat{\beta}}C'\right)^{-1} C\hat{\beta} \sim \chi^2_{(g)}$$

Donde C es una matriz asociada a los coeficientes establecidos en la hipótesis nula. Está matriz puede ser planteada como sigue: en todas las filas de la columna que corresponde a $\beta_1^{(k)}$ se pone 1, y en la $(d-1)$-ésima fila se coloca -1 en la columna que corresponda a $\beta_d^{(k)}$ con $d = 2, \ldots, P_k - 1$ y cero en las demás celdas. Nótese que para cualquier par de parámetros $\beta_a^{(k)} = \beta_b^{(k)}$ implica $\beta_a^{(k)} - \beta_b^{(k)} = 0$, de ahí que se ponga 1 en una posición y -1 en las demás posiciones correspondientes.

Para cada variable X_k de la especificidad con $k = 1, 2, \ldots m$ las hipótesis se plantean como:

$$
\begin{aligned}
H_0 :&\quad \beta_1^{(k)} + \beta_1^{(k,2)} = \beta_2^{(k)} + \beta_2^{(k,2)} = \beta_3^{(k)} + \beta_3^{(k,2)} = \cdots = \beta_{P_k-1}^{(k)} + \beta_{P_k-1}^{(k,2)} \\
H_1 :&\text{ Al menos uno de los parámetros} \quad \beta_a^{(k)} + \beta_a^{(k,2)} \neq \beta_b^{(k)} + \beta_b^{(k,2)} \\
&\text{con} \quad a \neq b, \quad a, b \in r, \quad r = 1, 2, \ldots, P_k - 1
\end{aligned}
$$

El estadístico de prueba esta dado por:

$$\chi^2 = \left(C\hat{\beta}\right)' \left(C\hat{\hat{\Sigma}}_{\hat{\beta}}C'\right)^{-1} C\hat{\beta} \sim \chi^2_{(g)}$$

Donde C es una matriz asociada a los coeficientes establecidos en la hipótesis nula. Está matriz puede ser planteada como sigue: en todas las filas de la columna que corresponde a $\beta_1^{(k)}$ y en todas las filas de la columna que corresponde a $\beta_1^{(k,2)}$ se pone 1, y en la $(d-1)$-ésima fila se coloca el valor -1 en las columnas que correspondan a $\beta_d^{(k)}$, $\beta_d^{(k,2)}$ con $d = 2, \ldots, P_k - 1$ y cero en las demás celdas. Nótese que para cualquier grupo de parámetros $\beta_1^{(k)} + \beta_1^{(k,2)} = \beta_2^{(k)} + \beta_2^{(k,2)}$ implica $\beta_1^{(k)} + \beta_1^{(k,2)} - \left(\beta_2^{(k)} + \beta_2^{(k,2)}\right) = 0$, de ahí que se ponga 1 en dos posiciones y -1 en las otras dos posiciones correspondientes.

Cabe aclarar que para las pruebas individuales sobre las variables y las pruebas de contraste sobre las variables no se aplicaron en el ejemplo debido a que las variables solo tenían 2 niveles, donde uno de ellos era el de referencia.

5. Estudio de simulación

En esta sección se realizó un estudio de simulación para analizar el comportamiento del tamaño de muestra mínimo bajo el cual se alcanza la normalidad univariada y multivariada de los parámetros estimados utilizando la metodología GSK cuando la función respuesta se plantea de manera directa. Adicionalmente se indaga si dicho comportamiento se ve afectado por la escogencia de las probabilidades de la distribución o por el valor que se le asigna a la sensibilidad y a la especificidad. Para esto, se crearon diferentes escenarios para observar bajo qué condiciones se obtiene un menor tamaño de muestra. Las simulaciones se programaron y se realizaron utilizando el lenguaje de programación estadístico R.

5.1. Metodología

Para el estudio de simulación se consideraron 2 grandes escenarios, donde dentro del último se analizaron otros 2 escenarios para un total de 4 escenarios. Al interior de los escenarios se generaron muestras aleatorias para cada una de las subpoblaciones con distribución multinomial; se cambiaron los parámetros de la distribución p_{11}, p_{12}, p_{21}, p_{22}, y se hicieron variaciones en el tamaño de muestra n, conservándose igual entre las subpoblaciones.
En todos los escenarios se hizo una corrección de 0.5 en caso de que el número generado por la distribución fuera cero. Esta corrección es comúnmente utilizada en análisis de tablas de contingencia (Upton (1992), Hanley (1983), Agresti (1996b)).

5.2. Escenarios del estudio de simulación

A continuación se muestran los escenarios considerados en la simulación.

5.2.1. Escenario 1

: En este escenario se consideró un modelo con 4 subpoblaciones conformadas por 2 variables donde cada una cuenta con 2 niveles. Cada subpoblación tiene la misma sensibilidad y especificidad dadas por: $Se = 5/7 \approx 0.71$, $Sp = 2/3 \approx 0.67$. Y para la generación de las observaciones de la tabla para cada subpoblación se fijaron las probabilidades $\pi_{11} = 0.5$, $\pi_{12} = 0.2$, $\pi_{21} = 0.1$, $\pi_{22} = 0.2$.

5.2.2. Escenario 2

: En este escenario se consideró un modelo con 4 subpoblaciones conformadas por 2 variables donde cada una cuenta con 2 niveles. Cada subpoblación tiene la misma sensibilidad y especificidad dadas por: $Se = 0.8$, $Sp = 0.6$. Y para la generación de las observaciones de la tabla para cada subpoblación se fijaron las probabilidades $\pi_{11} = 0.4$, $\pi_{12} = 0.1$, $\pi_{21} = 0.2$, $\pi_{22} = 0.3$.

5.2.3. Escenario 3

En este escenario para las observaciones de la tabla para cada subpoblación se consideró fijar probabilidades parecidas o cercanas a las planteadas en el escenario 2 que generaran los mismos valores de sensibilidad y especificidad, $Se = 0.8$, $Sp = 0.6$, estas fueron:
$\pi_{11} = \frac{7}{15} \approx 0.47$, $\pi_{12} = \frac{7}{60} \approx 0.12$, $\pi_{21} = \frac{1}{6} \approx 0.17$, $\pi_{22} = \frac{1}{4} = 0.25$.

5.2.4. Escenario 4

En este escenario se consideró para las observaciones de la tabla, para cada subpoblación, fijar probabilidades diferentes o lejanas a las planteadas en el escenario 2 que generaran los mismos valores de sensibilidad y especificidad, $Se = 0.8$, $Sp = 0.6$, estas fueron: $\pi_{11} = \frac{2}{3} \approx 0.67$, $\pi_{12} = \frac{1}{6} \approx 0.17$, $\pi_{21} = \frac{1}{15} \approx 0.07$, $\pi_{22} = \frac{1}{10} = 0.1$.

5.3. Proceso de simulación

En cada escenario el proceso de simulación utilizado para evaluar la estimación de la sensibilidad y la especificidad, y la distribución de los parámetros del modelo consiste en los siguientes pasos:

1. Se especificaron las probabilidades teóricas p_{11}, p_{12}, p_{21}, p_{22} para la generación de las tablas 2×2 teóricas que conforman las subpoblaciones con una especificidad y una sensibilidad predefinidas.

2. Se define el tamaño muestral n para la tabla y se generan las observaciones usando la función `rmultinom()`.

3. Se estima la sensibilidad y la especificidad vía GSK, y se hayan los parámetros del modelo.

4. Se repiten los pasos 2 y 3 para $N = 1000$ veces.

5. Se calcula la media y la varianza de los parámetros del modelo, de la sensibilidad y de la especificidad.

Para que el proceso fuera reproducible los resultados expuestos fueron los obtenidos fijando una semilla, utilizando la función `seed.seed(2022)`.

5.4. Resultados

En esta sección presentaremos los resultados de los cuatro escenarios del estudio de simulación. Para cada uno de ellos se analizaron las densidades univariadas de los parámetros del modelo y se respaldaron los resultados con la prueba Shapiro-Wilk univariada utilizando la función `shapiro.test` y Shapiro-Wilk multivariada con la función `mvShapiro.Test`. Se analizaron las gráficas de contorno de la densidad bivariada de la sensibilidad y la especificidad utilizando el paquete **ash** de David Scott (2015) donde el número de clases o intervalos se obtuvieron utilizando el método de Sturges siguiendo lo propuesto por Rizzo (2019).

5.4.1. Resultados escenario 1

En la tabla 5.1 se muestra la media y la varianza de los parámetros del modelo, β_0, $\beta_1^{(1)}$, $\beta_1^{(2)}$, $\beta_0^{(,2)}$, $\beta_1^{(1,2)}$, $\beta_1^{(2,2)}$, para diferentes tamaños de muestra. Podemos observar que a medida que se aumenta el tamaño de muestra n los valores de la media de los parámetros se acercan más a cero. Este resultado es esperado debido a que todas las subpoblaciones se plantearon con la misma sensibilidad y especificidad, por lo que se esperaría que ninguna covariable tuviera un aporte significativo en ellas.

La figura 5.1 expone las densidades de los parámetros del modelo para diferentes tamaños de muestra. A medida que el tamaño muestral aumenta las densidades tienden a una normal. Para hallar el mínimo tamaño tal que tanto los parámetros individuales como agrupados siguieran una distribución normal se hizo uso del test de Shapiro-Wilk univariado y multivariado. El último tamaño de muestra $n = 230$ corresponde al valor mínimo donde en ambos tests dio como resultado normalidad.

La figura 5.2 ilustra las densidades bivariadas de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra. A medida que el tamaño de muestra aumenta los rangos del gráfico de contorno se estrechan en torno a los valores medios (Tabla 5.2) que se aproximan a los valores reales de la sensibilidad y especificidad dados por $Se = 5/7 \approx 0.71$, $Sp = 2/3 \approx 0.67$, y las varianzas tienden cada vez más a cero (Tabla 5.2). En ninguna de las gráficas se nota una correlación positiva o negativa.

Es de anotar que algunos de los rangos de la gráfica de contorno son inferiores a 0 o superiores a 1, esto reafirma lo expuesto en la sección 3.2 respecto a la repercusión que tiene trabajar con la función respuesta de manera directa (3.2.1). Cuando existen probabilidades cercanas a 0 o a 1 la estimación puede generar inconsistencias.

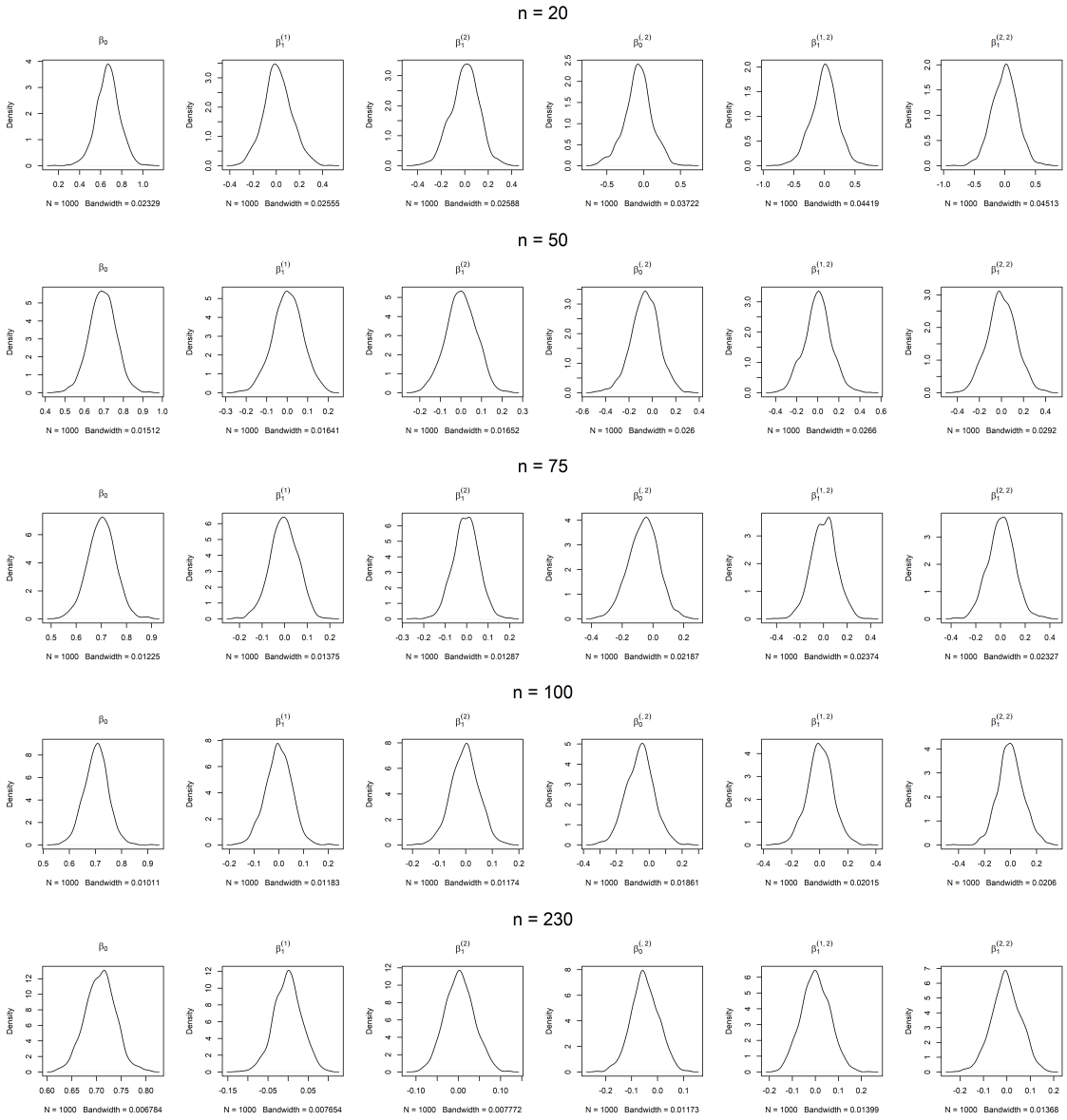

Figura 5.1.: Densidades univariadas de los parámetros estimados del modelo para distintos tamaños de muestras, escenario 1

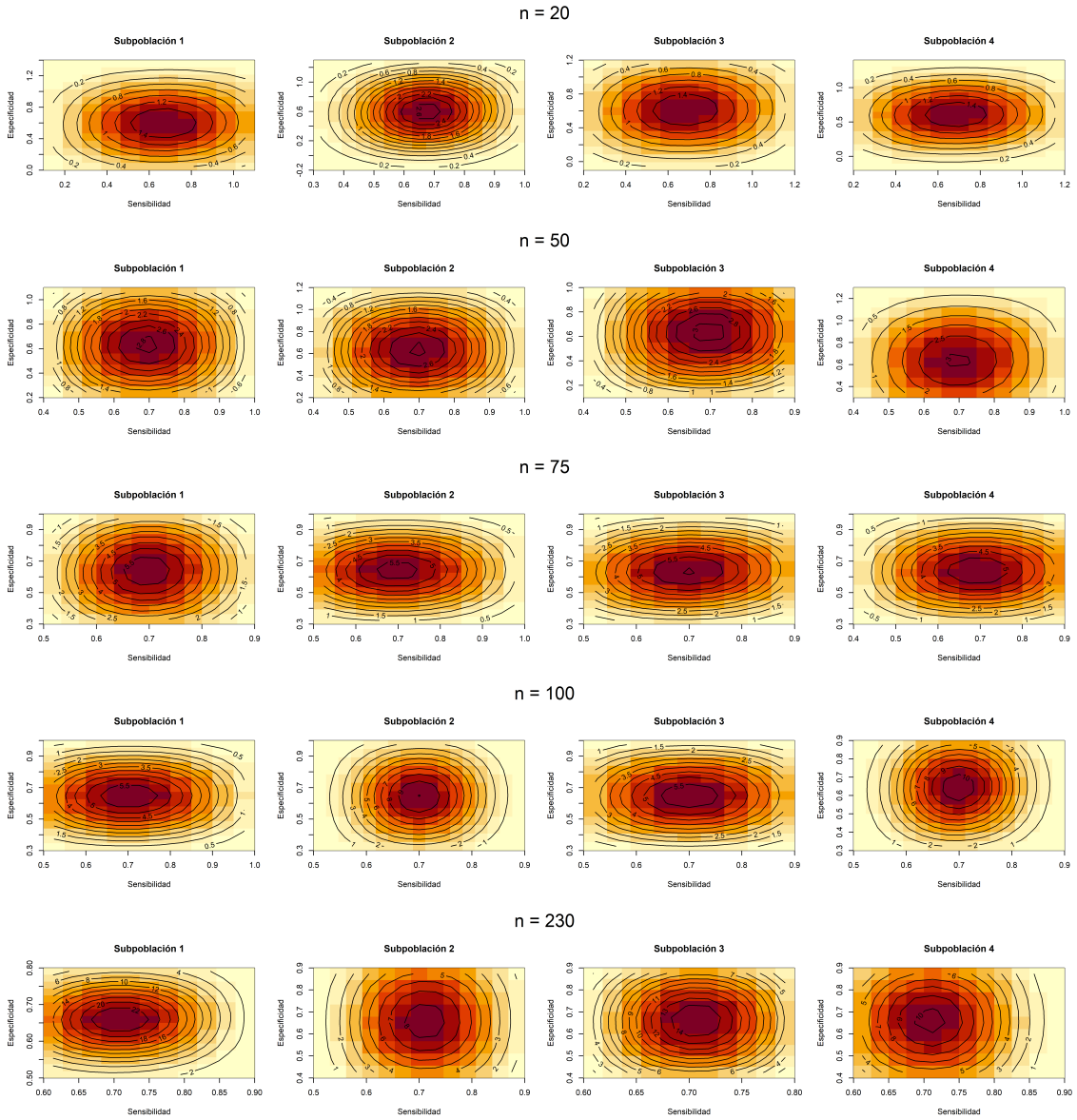

Figura 5.2.: Gráficas de contornos de la densidad bivariada de la sensibilidad y la especificidad para distintos tamaños de muestras, escenario 1

Tamaño de muestra		β_0	$\beta_1^{(1)}$	$\beta_1^{(2)}$	$\beta_0^{(,2)}$	$\beta_1^{(1,2)}$	$\beta_1^{(2,2)}$
$n = 20$	**Media**	0.6690083	0.0139675	0.0017977	-0.0732061	-0.0031041	-0.0026737
	Varianza	0.0123083	0.014588	0.0142143	0.0348795	0.0460576	0.0427231
$n = 50$	**Media**	0.6927667	0.0021853	0.0012088	-0.0636074	0.0004391	0.0041833
	Varianza	0.0046149	0.0053564	0.0053373	0.0142063	0.0169197	0.0171744
$n = 75$	**Media**	0.7010017	-0.0033821	-0.0034639	-0.0626398	0.0034083	0.0062062
	Varianza	0.003038	0.0037011	0.0034997	0.0093598	0.011584	0.0116785
$n = 100$	**Media**	0.702605	-0.0008131	-0.0030723	-0.0563106	0.0020866	0.0006726
	Varianza	0.0021096	0.0027692	0.0027239	0.0070546	0.0084911	0.0094964
$n = 230$	**Media**	0.7076308	-0.0010342	0.0029367	-0.0487267	0.0014495	-0.0029421
	Varianza	0.0009296	0.0011822	0.0011818	0.0028132	0.0038273	0.003839

Tabla 5.1.: Media y varianza de los parámetros estimados del modelo para diferentes tamaños de muestra, escenario 1

Subpoblación		1		2		3		4	
Tamaño de muestra		Se	Sp	Se	Sp	Se	Sp	Se	Sp
$n = 20$	**Media**	0.6690083	0.5958023	0.670806	0.5949262	0.6829759	0.6066657	0.6847736	0.6057897
	Varianza	0.0123083	0.0276963	0.0107329	0.0282576	0.0117585	0.0281028	0.0124155	0.0285398
$n = 50$	**Media**	0.6927667	0.6291593	0.6939755	0.6345514	0.6949519	0.6317836	0.6961607	0.6371757
	Varianza	0.0046149	0.011212	0.0042029	0.0116814	0.0041663	0.0117321	0.0044617	0.0110567
$n = 75$	**Media**	0.7010017	0.6383619	0.6975378	0.6411042	0.6976196	0.6383881	0.6941556	0.6411304
	Varianza	0.003038	0.007503	0.0030075	0.0075105	0.0028771	0.0074125	0.0028123	0.0075146
$n = 100$	**Media**	0.702605	0.6462945	0.6995328	0.6438948	0.701792	0.647568	0.6987197	0.6451683
	Varianza	0.0021096	0.0058299	0.0020766	0.0057632	0.0022358	0.0058173	0.0022227	0.0051696
$n = 230$	**Media**	0.7076308	0.6589041	0.7105675	0.6588987	0.7065966	0.6593194	0.7095332	0.6593139
	Varianza	0.0009296	0.0022637	0.0009793	0.0023479	0.0008737	0.0023425	0.0008992	0.0024331

Tabla 5.2.: Media y varianza para las estimaciones de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra, escenario 1

5.4.2. Resultados escenario 2

En la tabla 5.3 se muestra la media y la varianza de los parámetros del modelo, $\beta_1^{(1)}$, $\beta_1^{(2)}$, $\beta_0^{(,2)}$, $\beta_1^{(1,2)}$, $\beta_1^{(2,2)}$, para diferentes tamaños de muestra. Podemos observar que a medida que se aumenta el tamaño de muestra n los valores de la media y la varianza de los parámetros se aproximan a cero.

La figura 5.3 expone las densidades de los parámetros del modelo para diferentes tamaños de muestra. A medida que el tamaño muestral aumenta las densidades tienden a una normal. Para hallar el mínimo tamaño tal que tanto los parámetros individuales como en su conjunto siguieran una distribución normal se hizo uso del test de Shapiro-Wilk univariado y multivariado. El último tamaño de muestra $n = 301$ corresponde al valor mínimo donde en ambos tests dio como resultado normalidad.

La figura 5.4 ilustra las densidades bivariadas de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra. A medida que el tamaño de muestra aumenta los rangos del gráfico de contorno se estrechan en torno a los valores medios (Tabla 5.4) que se aproximan a los valores reales de la sensibilidad y especificidad dados por $Se = 0.8$, $Sp = 0.6$, y las varianzas tienden cada vez más a cero (Tabla 5.4). En ninguna de las gráficas se nota una correlación positiva o negativa.

Es de anotar que para tamaños de muestra pequeños los rangos de la gráfica de contorno son

inferiores a 0 o superiores a 1, pero a medida que se aumenta el tamaño de muestra este error se corrige ya que los datos se agrupan en torno a la media como se mencionó anteriormente.

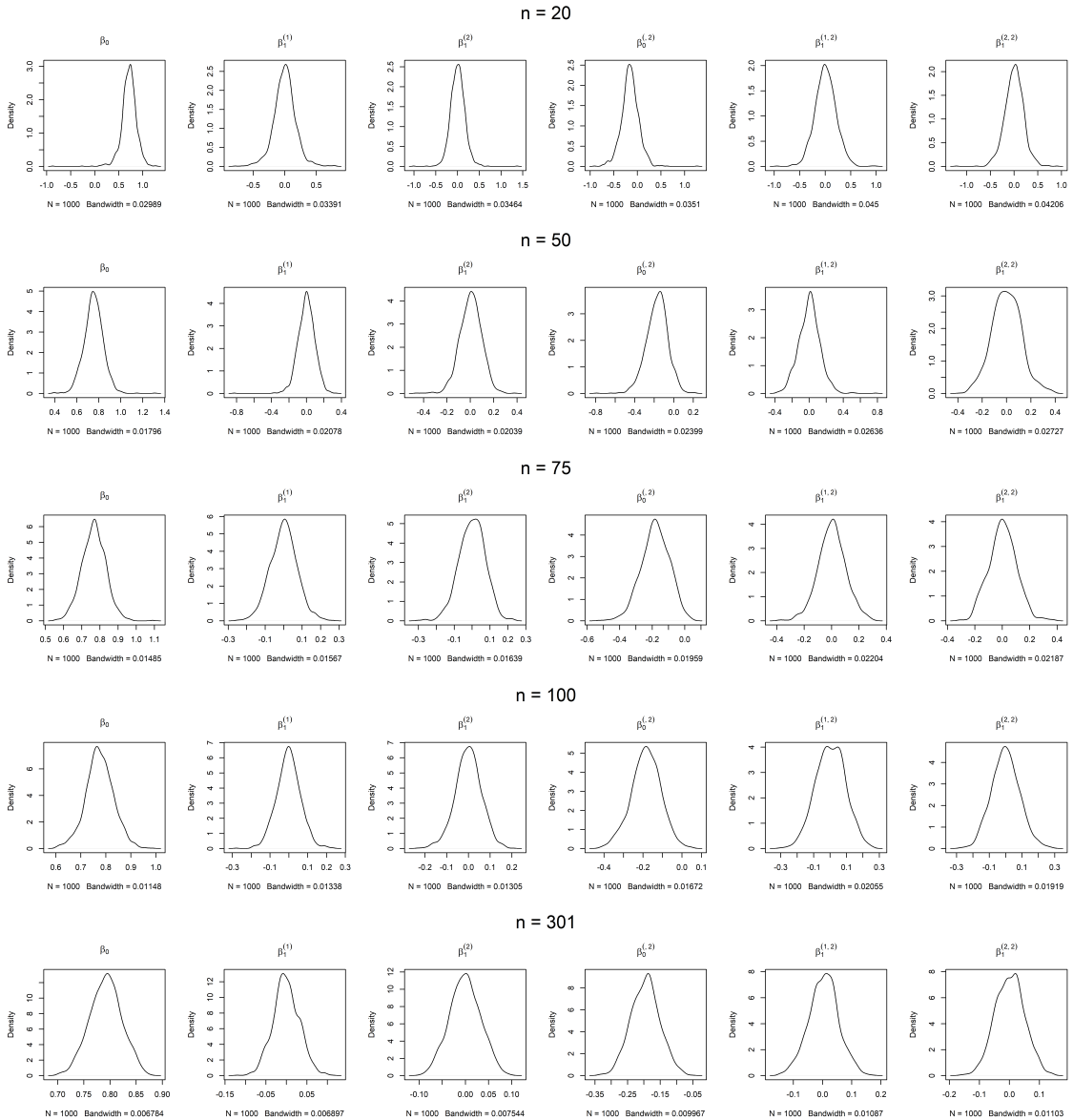

Figura 5.3.: Densidades univariadas de los parámetros estimados del modelo para distintos tamaños de muestras, escenario 2

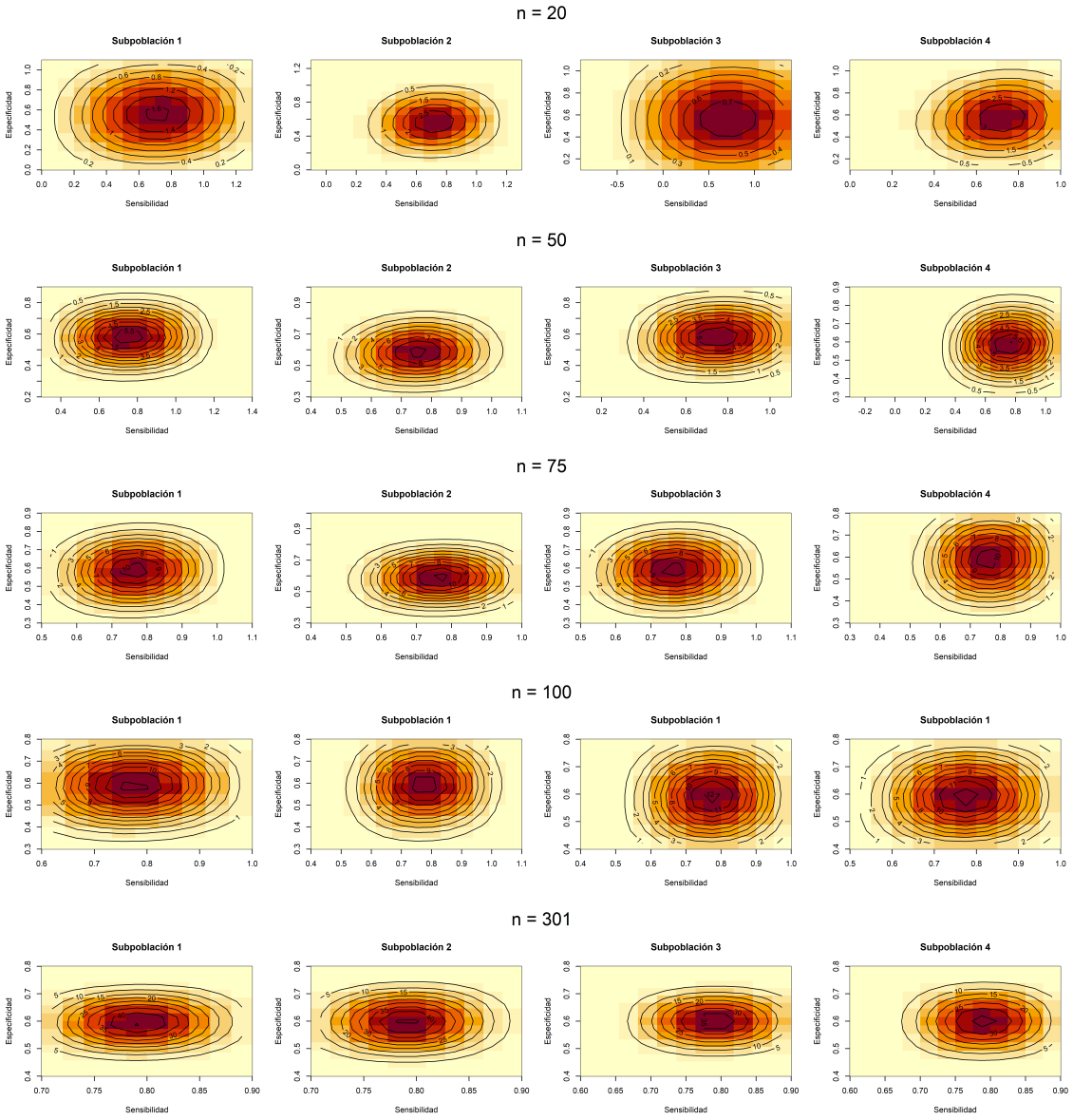

Figura 5.4.: Gráficas de contornos de la densidad bivariada de la sensibilidad y la especificidad para distintos tamaños de muestras, escenario 2

Tamaño de muestra		β_0	$\beta_1^{(1)}$	$\beta_1^{(2)}$	$\beta_0^{(,2)}$	$\beta_1^{(1,2)}$	$\beta_1^{(2,2)}$
$n = 20$	**Media**	0.7123401	-0.001015	0.0052836	-0.1546832	0.0067445	0.0018419
	Varianza	0.0259101	0.0309792	0.0284427	0.0343831	0.0432532	0.0400453
$n = 50$	**Media**	0.7535201	-0.001479	0.0029842	-0.1722049	0.0038286	0.0005614
	Varianza	0.0077311	0.0095691	0.0089062	0.0124505	0.0159521	0.0145932
$n = 75$	**Media**	0.7661981	-0.0001757	-0.0005808	-0.1811848	0.0028537	0.0033549
	Varianza	0.0043126	0.00528	0.0053446	0.007518	0.0102749	0.010232
$n = 100$	**Media**	0.7754847	-0.0019949	-0.0008767	-0.1847249	0.0029866	-0.0004762
	Varianza	0.0031147	0.0041481	0.0037292	0.0057836	0.0082599	0.0074299
$n = 301$	**Media**	0.7917129	-0.000577	-0.0009502	-0.1954957	0.0041229	0.0013556
	Varianza	0.0009293	0.0010422	0.0011136	0.0019603	0.0026115	0.0023813

Tabla 5.3.: Media y varianza de los parámetros estimados del modelo para diferentes tamaños de muestra, , escenario 2

Subpoblación		1		2		3		4	
Tamaño de muestra		Se	Sp	Se	Sp	Se	Sp	Se	Sp
$n = 20$	**Media**	0.7123401	0.5576569	0.7176237	0.5647825	0.7113251	0.5633864	0.7166087	0.570512
	Varianza	0.0259101	0.0175484	0.0223955	0.0189643	0.0244863	0.0187895	0.0255439	0.017275
$n = 50$	**Media**	0.7535201	0.5813152	0.7565043	0.5848608	0.7520411	0.5836648	0.7550253	0.5872104
	Varianza	0.0077311	0.0065114	0.0067561	0.0070942	0.007656	0.0071154	0.0087005	0.0065729
$n = 75$	**Media**	0.7661981	0.5850133	0.7656173	0.5877874	0.7660224	0.5876914	0.7654416	0.5904654
	Varianza	0.0043126	0.004826	0.0042707	0.0046156	0.0044292	0.00505	0.0044483	0.0043291
$n = 100$	**Media**	0.7754847	0.5907598	0.774608	0.5894069	0.7734898	0.5917515	0.7726131	0.5903986
	Varianza	0.0031147	0.0033809	0.0030287	0.0037026	0.0031998	0.0036186	0.0032498	0.0031398
$n = 301$	**Media**	0.7917129	0.5962172	0.7907627	0.5966226	0.7911359	0.5997631	0.7901857	0.6001685
	Varianza	0.0009293	0.0011554	0.0009019	0.001093	0.0008333	0.0011465	0.0008789	0.0011877

Tabla 5.4.: Media y varianza para las estimaciones de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra, escenario 2

5.4.3. Resultados escenario 3

En la tabla 5.5 se muestra la media y la varianza de los parámetros del modelo, $\beta_1^{(1)}$, $\beta_1^{(2)}$, $\beta_0^{(,2)}$, $\beta_1^{(1,2)}$, $\beta_1^{(2,2)}$, para diferentes tamaños de muestra. Podemos observar que a medida que se aumenta el tamaño de muestra n los valores de la media y la varianza de los parámetros se acercan más a cero.

La figura 5.5 expone las densidades de los parámetros del modelo para diferentes tamaños de muestra. A medida que el tamaño muestral aumenta las densidades tienden a una normal. Para hallar el mínimo tamaño tal que tanto los parámetros individuales como en su conjunto siguieran una distribución normal se hizo uso del test de Shapiro-Wilk univariado y multivariado. El último tamaño de muestra $n = 301$ corresponde al valor mínimo donde en ambos tests dio como resultado normalidad. Cabe destacar que a pesar de que las probabilidades son distintas a las del escenario 2 el modelo alcanza la normalidad univariada y multivariana con el mismo tamaño de muestra $n = 301$.

La figura 5.6 ilustra las densidades bivariadas de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra. A medida que el tamaño de muestra aumenta los rangos del gráfico de contorno se estrechan en torno a los valores medios (Tabla 5.6) que se aproximan a los valores reales de la sensibilidad y especificidad dados por $Se = 0.8$, $Sp = 0.6$, y las varianzas tienden cada vez más a cero (Tabla 5.6). En ninguna de las gráficas se nota una correlación positiva o negativa.

Es de anotar que para tamaños de muestra pequeños los rangos de la gráfica de contorno son inferiores a 0 o superiores a 1, pero a medida que se aumenta el tamaño de muestra este error se corrige ya que los datos se agrupan en torno a la media como se mencionó anteriormente.

Se cálculo la distancia de Kullback-Leibler con respecto a la distribución multinomial del escenario 2 que dio como resultado: 0.02032099.

Figura 5.5.: Densidades univariadas de los parámetros estimados del modelo para distintos tamaños de muestras, escenario 3

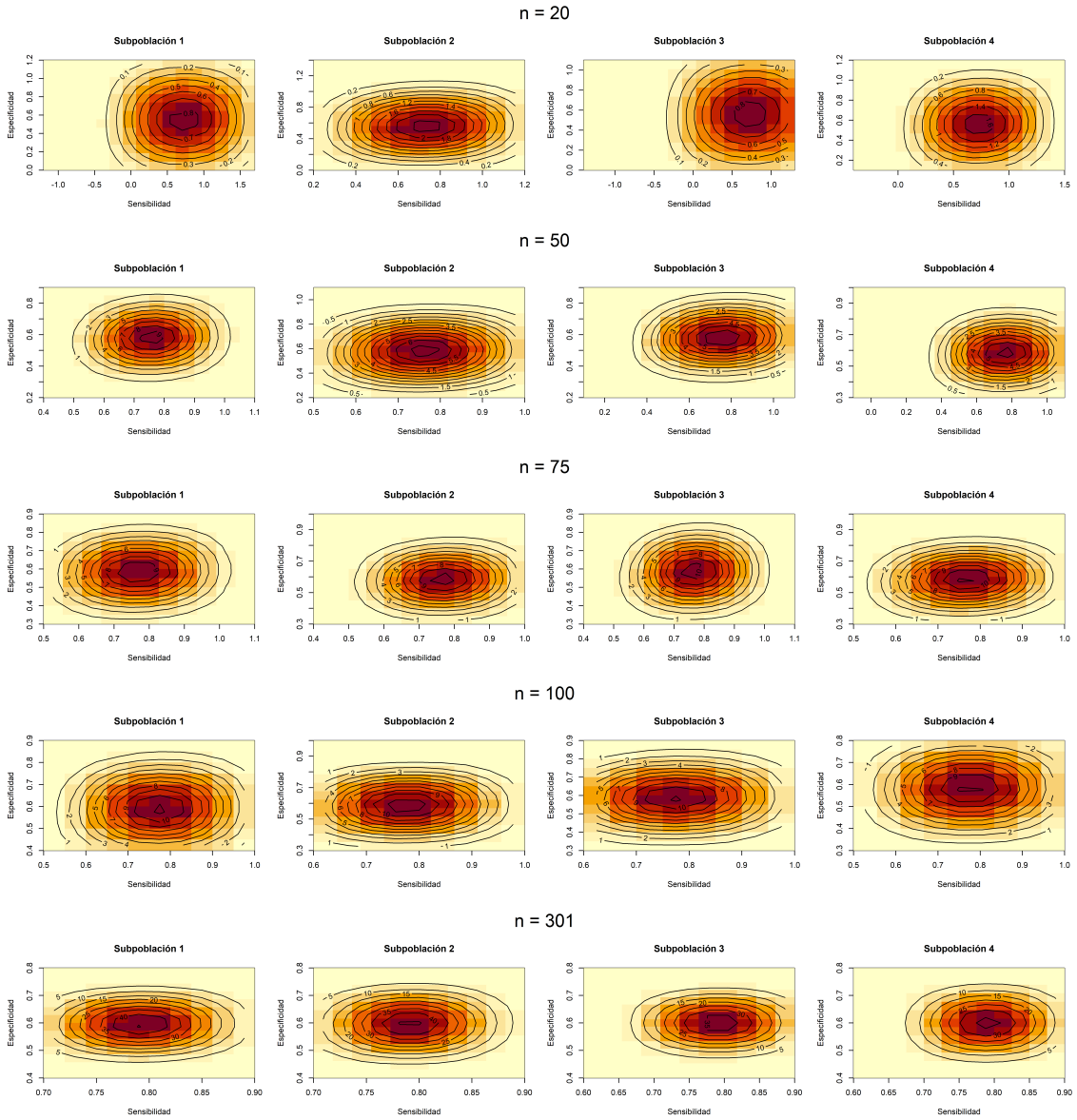

Figura 5.6.: Gráficas de contornos de la densidad bivariada de la sensibilidad y la especificidad para distintos tamaños de muestras, escenario 3

Tamaño de muestra		β_0	$\beta_1^{(1)}$	$\beta_1^{(2)}$	$\beta_0^{(,2)}$	$\beta_1^{(1,2)}$	$\beta_1^{(2,2)}$
$n = 20$	**Media**	0.7278065	-0.0043834	0.0037111	-0.1743926	0.0077292	0.002675
	Varianza	0.0221546	0.0230328	0.0239243	0.036494	0.0407763	0.0425882
$n = 50$	**Media**	0.7589532	-0.0006398	0.0020664	-0.1813746	0.0037748	0.0021212
	Varianza	0.0059235	0.0077027	0.0068864	0.0116835	0.015352	0.0141872
$n = 75$	**Media**	0.7676934	0.0007231	0.0006568	-0.1834477	-0.0003209	-0.0010231
	Varianza	0.0037295	0.004801	0.0042795	0.0069287	0.0099809	0.0100143
$n = 100$	**Media**	0.7740343	0.000475	0.0019464	-0.1832977	-0.0027473	-0.0054565
	Varianza	0.0027288	0.0032898	0.0031739	0.005863	0.0077015	0.0080913
$n = 301$	**Media**	0.7917129	-0.000577	-0.0009502	-0.1954957	0.0041229	0.0013556
	Varianza	0.0009293	0.0010422	0.0011136	0.0019603	0.0026115	0.0023813

Tabla 5.5.: Media y varianza de los parámetros estimados del modelo para diferentes tamaños de muestra, escenario 3

Subpoblación		1		2		3		4	
Tamaño de muestra		Se	Sp	Se	Sp	Se	Sp	Se	Sp
$n = 20$	**Media**	0.7278065	0.553414	0.7315176	0.5598001	0.7234231	0.5567597	0.7271342	0.5631458
	Varianza	0.0221546	0.0211994	0.0150698	0.0220937	0.0211919	0.022578	0.0199738	0.0209752
$n = 50$	**Media**	0.7589532	0.5775785	0.7610195	0.5817661	0.7583134	0.5807136	0.7603797	0.5849011
	Varianza	0.0059235	0.0076888	0.0053523	0.0084798	0.0059969	0.0083457	0.0068248	0.0077114
$n = 75$	**Media**	0.7676934	0.5842456	0.7683502	0.5838793	0.7684164	0.5846478	0.7690732	0.5842815
	Varianza	0.0037295	0.005176	0.0039013	0.0056962	0.0039194	0.0055597	0.0037349	0.0057016
$n = 100$	**Media**	0.7740343	0.5907366	0.7759807	0.5872265	0.7745093	0.5884643	0.7764557	0.5849542
	Varianza	0.0027288	0.0041569	0.0026007	0.0045968	0.0025446	0.004626	0.0026994	0.004119
$n = 301$	**Media**	0.7917129	0.5962172	0.7907627	0.5966226	0.7911359	0.5997631	0.7901857	0.6001685
	Varianza	0.0009293	0.0011554	0.0009019	0.001093	0.0008333	0.0011465	0.0008789	0.0011877

Tabla 5.6.: Media y varianza para las estimaciones de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra, escenario 3

5.4.4. Resultados escenario 4

En la tabla 5.7 se muestra la media y la varianza de los parámetros del modelo, $\beta_1^{(1)}$, $\beta_1^{(2)}$, $\beta_0^{(,2)}$, $\beta_1^{(1,2)}$, $\beta_1^{(2,2)}$, para diferentes tamaños de muestra. Podemos observar que a medida que se aumenta el tamaño de muestra n los valores de la media y la varianza de los parámetros se aproximan a cero.

La figura 5.7 expone las densidades de los parámetros del modelo para diferentes tamaños de muestra. A medida que el tamaño muestral aumenta las densidades tienden a una normal. Para hallar el mínimo tamaño tal que tanto los parámetros individuales como en su conjunto siguieran una distribución normal se hizo uso del test de Shapiro-Wilk univariado y multivariado. El último tamaño de muestra $n = 226$ corresponde al valor mínimo donde en ambos tests dio como resultado normalidad. Cabe resaltar que a pesar de que la sensibilidad y la especificidad de los escenarios 2 y 3 es la misma para este escenario y si bien las probabilidades son distintas a las del escenario 2 y 3 el modelo alcanza la normalidad univariada y multivariana con un tamaño de muestra distinto a los anteriores, he incluso un poco más bajo con $n = 226$.

La figura 5.8 ilustra las densidades bivariadas de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra. A medida que el tamaño de muestra aumenta los rangos del gráfico de contorno se estrechan en torno a los valores medios (Tabla 5.8) que se aproximan a los valores reales de la sensibilidad y especificidad dados por $Se = 0.8$,

$Sp = 0.6$, y las varianzas tienden cada vez más a cero (Tabla 5.8). En ninguna de las gráficas se nota una correlación positiva o negativa.

Es de anotar que para tamaños de muestra pequeños los rangos de la gráfica de contorno son inferiores a 0 o superiores a 1, pero a partir de $n = 75$ este error se corrige. Siendo este escenario el más rápido en reflejar dicha corrección.

Se cálculo la distancia de Kullback-Leibler con respecto a la distribución multinomial del escenario 2 que dio como resultado 0.4239985. Cabe destacar que esta distancia es mayor a la del escenario 3, con lo cual la distribución del escenario 3 se parece más a la del escenario 2, mientras que la del escenario 4 se parece en menor medida.

Tamaño de muestra		β_0	$\beta_1^{(1)}$	$\beta_1^{(2)}$	$\beta_0^{(,2)}$	$\beta_1^{(1,2)}$	$\beta_1^{(2,2)}$
$n = 20$	**Media**	0.7576863	0.0021922	-0.0000836	-0.2335939	0.0058321	0.0082404
	Varianza	0.0085181	0.0126886	0.0118253	0.0396892	0.0505359	0.0493113
$n = 50$	**Media**	0.7772041	0.0017404	0.0014675	-0.2202779	0.0003214	0.0034088
	Varianza	0.0030631	0.0043253	0.0037689	0.0211876	0.0276858	0.0291985
$n = 75$	**Media**	0.7816902	0.0005407	0.0021863	-0.2137744	0.0011244	0.0025835
	Varianza	0.0020398	0.0028486	0.0025123	0.0137195	0.0181951	0.0183655
$n = 100$	**Media**	0.7847791	0.0012151	0.0002842	-0.2112992	-0.0000164	0.0020772
	Varianza	0.0015286	0.0021531	0.0019137	0.0110711	0.0143012	0.0143639
$n = 226$	**Media**	0.7925741	0.00042	-0.0013632	-0.202612	-0.0027125	0.0031342
	Varianza	0.0006893	0.0008902	0.0008369	0.0047424	0.0062989	0.0060921

Tabla 5.7.: Media y varianza de los parámetros estimados del modelo para diferentes tamaños de muestra, escenario 4

Subpoblación		**1**		**2**		**3**		**4**	
Tamaño de muestra		**Se**	**Sp**	**Se**	**Sp**	**Se**	**Sp**	**Se**	**Sp**
$n = 20$	**Media**	0.7576863	0.5240925	0.7576028	0.5322493	0.7598785	0.5321167	0.7597949	0.5402736
	Varianza	0.0085181	0.0347325	0.008885	0.0304231	0.0096858	0.0350034	0.0099049	0.0334225
$n = 50$	**Media**	0.7772041	0.5569262	0.7786716	0.5618025	0.7789445	0.558988	0.780412	0.5638643
	Varianza	0.0030631	0.020612	0.0032586	0.0226062	0.0032598	0.0231586	0.0033379	0.0206099
$n = 75$	**Media**	0.7816902	0.5679158	0.7838765	0.5726856	0.7822309	0.5695808	0.7844172	0.5743507
	Varianza	0.0020398	0.0135231	0.0022183	0.0152072	0.0021351	0.0144293	0.002149	0.0132994
$n = 100$	**Media**	0.7847791	0.5734799	0.7850633	0.5758413	0.7859942	0.5746786	0.7862784	0.57704
	Varianza	0.0015286	0.0109686	0.0017799	0.01085	0.0015576	0.0112331	0.0016084	0.0092747
$n = 226$	**Media**	0.7925741	0.5899621	0.7912109	0.5917331	0.7929941	0.5876695	0.7916309	0.5894405
	Varianza	0.0006893	0.0043994	0.0007228	0.0044274	0.0007144	0.0042502	0.0006587	0.0045217

Tabla 5.8.: Media y varianza para las estimaciones de la sensibilidad y la especificidad por subpoblaciones para diferentes tamaños de muestra, escenario 4

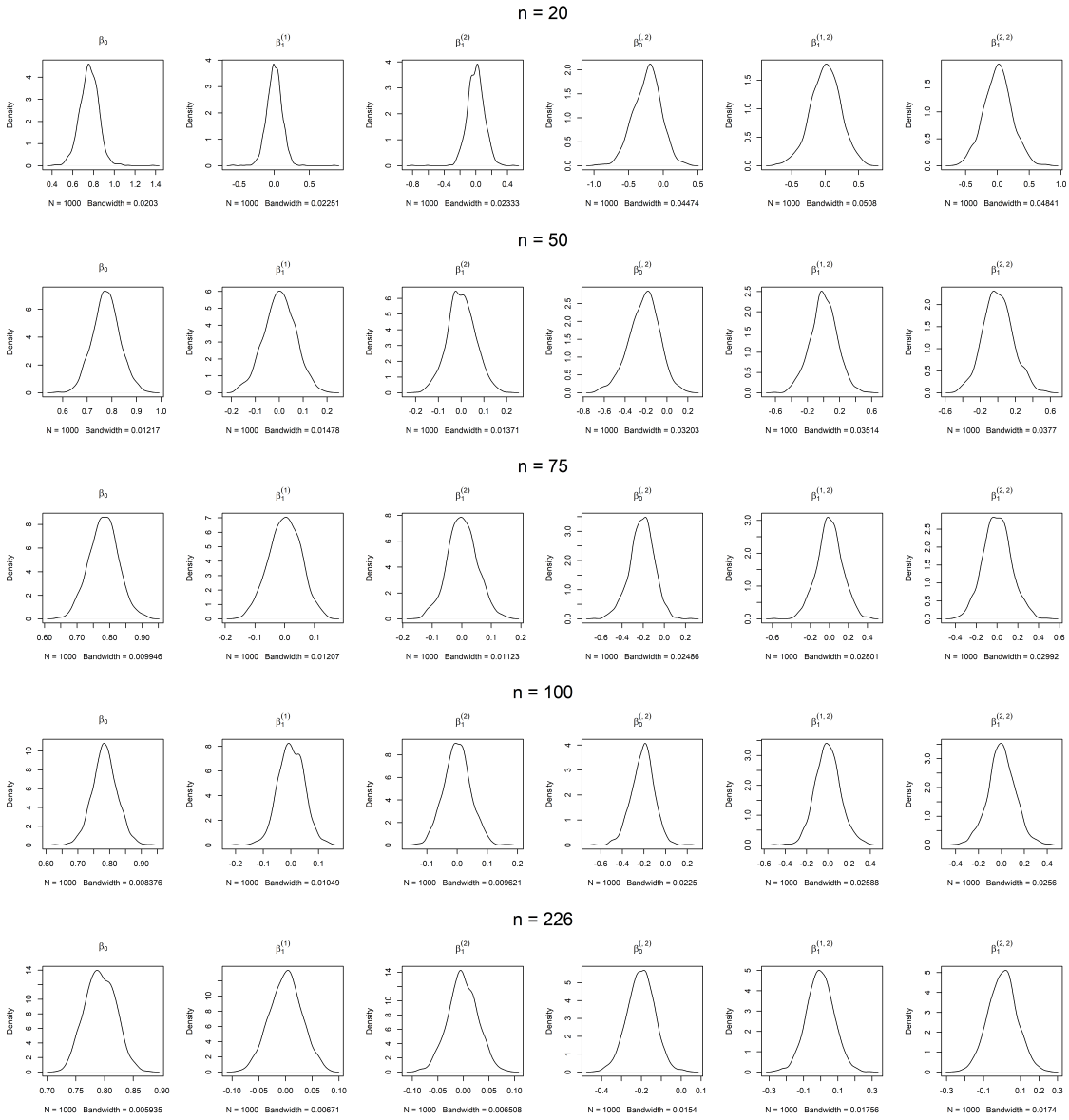

Figura 5.7.: Densidades univariadas de los parámetros estimados del modelo para distintos tamaños de muestras, escenario 4

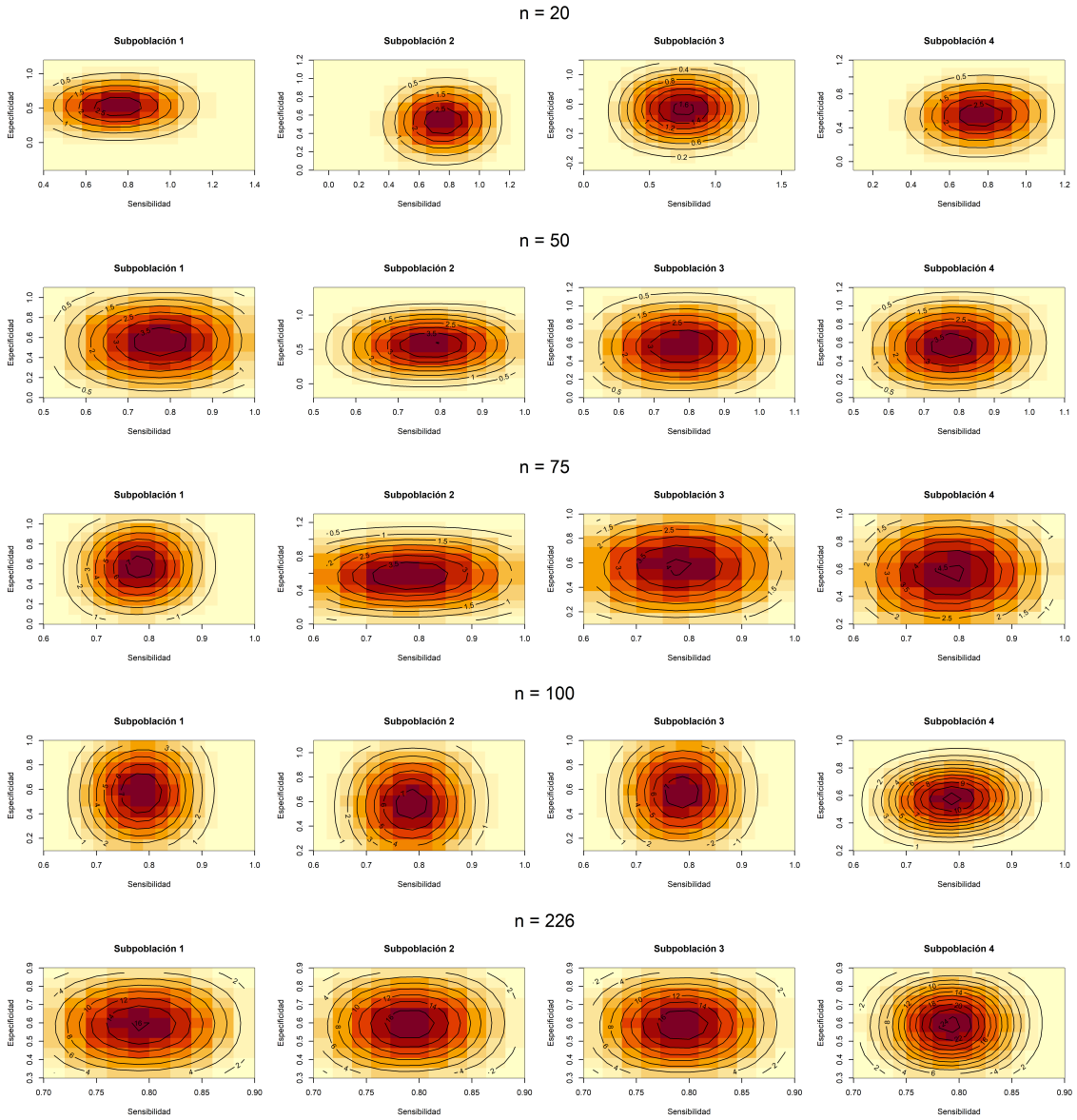

Figura 5.8.: Gráficas de contornos de la densidad bivariada de la sensibilidad y la especificidad para distintos tamaños de muestras, escenario 4

6. Conclusiones y recomendaciones

6.1. Conclusiones

Este trabajo presenta tres propuestas para la estimación de la sensibilidad y la especificidad teniendo en cuenta covariables utilizando la metodología GSK. La primera propuesta se basa en el cáculo directo de la sensibilidad y la especificidad haciendo uso de las probabilidad de la tabla de distribuciones, en la segunda se plantea el cálculo en términos del logaritmo de la sensibilidad y la especificidad y por último se presenta como respuesta el logit de la sensibilidad y la especificidad.

La principal ventaja de obtener la sensibilidad y la especificidad de forma directa es la facilidad en la interpretación de los coeficientes, mientras que tiene como desventaja el hecho de que al trabajar con probabilidades muy cercanas a 0 o 1 puede presentar una sobreestimación de la sensibilidad o especificidad, dando como resultado sensibilidades o especificidades mayores a 1. Ante este tipo de situaciones se recomienda utilizar el logit de la sensibilidad y la especificidad, ya que este tipo de transformaciones tiene un mejor comportamiento con respecto a los eventos raros que pueden causar sesgo y probabilidades muy cercanas a 0.

Por las propiedades que tienen cada una de las transformaciones sugeridas para la estimación de la sensibilidad y la especificidad se recomendaría primero usar el logit, como segunda opción el logaritmo y por último el cálculo de forma directa, aunque hay que tener en cuenta el grado de complejidad a la hora de la interpretación.

A pesar de obtener la sensibilidad y la especificidad simultáneamente del mismo conjunto de datos en el ejemplo que se desarrolló con datos psuedo reales se encontró que en general para las subpoblaciones la relación entre la sensibilidad y la especificidad es moderada y de tipo lineal.

Toda la metodología y propuestas planteadas para la estimación se hacen bajo el supuesto de tamaño de muestra grande, de manera más específica, de distribuciones normales de tipo asintótico.

Según el estudio de simulación realizado, el tamaño mínimo de muestra para obtener normalidad univariada y multivariada varía conforme a las probabilidades asignadas al modelo.

Adicionalmente, se compararon las densidades generadas por las distribuciones multinomiales empleadas en los escenarios 2, 3 y 4 con la distancia de Kullback-Leibler tomando como referencia el escenario 2; estos escenarios comparten la misma sensibilidad y especificidad. La distancia entre las densidades de los escenarios 2 y 3 fue más pequeña que la distancia entre las densidades de los escenarios 2 y 4, es decir, la distribución del escenario 3 se parece más a la del escenario 2, que la del escenario 4. Sin embargo el tamaño de muestra mínimo para los escenarios 2 y 3 fue de $n = 301$, y para el escenario 4 de $n = 226$, con lo cual se puede concluir que aunque la distancia Kullback-Leibler sea mayor, el tamaño de muestra mínimo puede ser menor, es decir, la cercanía entre las distribuciones multinomiales no parece tener un efecto directo en el tamaño de muestra mínimo.

6.2. Recomendaciones

- El código desarrollado en el software estadístico R puede mejorarse en su parte algorítmica y se puede crear una librería en R.
- Realizar diagnósticos del modelo.
- Estimar la sensibilidad y la especificidad utilizando aproximación Bayesiana en la metodología GSK.
- Realizar un estudio de simulación más exhaustivo para verificar el comportamiento del modelo variando el número de subpoblaciones, tomando probabilidades distintas para cada subpoblación, considerando casos donde los datos no provengan de una distribución multinomial, entre otros.

A. Código implementado en R para la estimación simultánea de la sensibilidad y la especificidad utilizando la metodología GSK en presencia de covariables

El código en R se encuentra alojado en:
https://rpubs.com/JessicaP/892479

Para mayor detalle del ejemplo con datos pseudo reales, los cálculos, estimaciones, valores de las matrices y código en R se puede encontrar en:

De manera directa: https://rpubs.com/JessicaP/861043
De manera logaritmica: https://rpubs.com/JessicaP/931253
De manera logit: https://rpubs.com/JessicaP/931259

El código para la simulación y mayor detalle de los resultados de cada uno de los escenarios se encuentra alojado en:

Escenario 1 : http://rpubs.com/JessicaP/892464
Escenario 2 : https://rpubs.com/JessicaP/892470
Escenario 3 : https://rpubs.com/JessicaP/892475
Escenario 4 : https://rpubs.com/JessicaP/892476

B. Shapiro-Wilk

En este apéndice se presenta información acerca de la prueba Shapiro Wilk univariada y multivariada a través del enlace https://rpubs.com/JessicaP/891652

Bibliografía

Agresti, A. (1996a). *Categorical data analysis*. New York: John Wiley & Sons.

Agresti, A. (1996b). *An Introduction to categorical data analysis*. John Wiley & Sons.

Alexander, L. K., Lopes, B., Ricchetti-Masterson, K., and Yeatts, K. B. (2014). Assessment of diagnostic and screening tests.

Correa, J. C. (2016). *Analysis of Contingency Tables Via GSK Using R*. Universidad Nacional de Colombia, sede Medellín.

Coughlin, S. S., Trock, B., Criqui, M. H., Pickle, L. W., Browner, D., and Tefft, M. C. (1992). The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *Journal of clinical epidemiology*, 45(1):1–7.

Cover, T. M. and Thomas, J. A. (2006). Elements of information theory 2nd edition (wiley series in telecommunications and signal processing). *Acessado em*.

Dutta, M. (1982). *Econometric methods*. Cinccinati: South-Western.

Engel, B., Swildens, B., Stegeman, A., Buist, W., and De Jong, M. (2006). Estimation of sensitivity and specificity of three conditionally dependent diagnostic tests in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4):360.

Escrig-Sos, J., Martínez-Ramos, D., and Miralles-Tena, J. M. (2006). Pruebas diagnósticas: nociones básicas para su correcta interpretación y uso. *Cirugía Española*, 79(5):267–273.

Forthofer, R. (2012). *Public program analysis: a new categorical data approach*. Springer Science & Business Media.

Forthofer, R. N. and Koch, G. G. (1972). An analysis for compounded logarithmic-exponential linear functions of categorical data. Technical report, North Carolina State University. Dept. of Statistics.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25(3):489–504.

Hanley, J. A., . L.-H. A. (1983). If nothing goes wrong, is everything all right?: interpreting zero numerators. *Jama*, 249(13):1743–1745.

Henao, K. J. and Correa, J. C. (2018). Regresión logística bivariable para tablas de contingencia usando metodología gsk. *Comunicaciones en Estadística*, 11(2):153–170.

Henquin, R. (2013). *Epidemiología y estadística para principiantes*. Corpus Editorial.

Hess, A., Shardell, M., Johnson, J., Thom, K., Strassle, P., Netzer, G., and Harris, A. (2012). Methods and recommendations for evaluating and reporting a new diagnostic test. *European journal of clinical microbiology & infectious diseases*, 31(9):2111–2116.

Johnson, W. D. and Koch, G. G. (1970). Analysis of qualitative data: Linear functions. *Health Services Research*, 5(4):358.

Lalkhen, A. G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain*, 8(6):221–223.

Neyman, J. (1949). Contribution to the theory of the x2 test. In *Proceedings of the Berkeley symposium on mathematical statistics and probability*, pages 239–273.

Ochoa, C. O. S., de Dios, J. G., and Álvarez, J. C. B. (2007). Evaluación de artículos científicos sobre pruebas diagnósticas. *Evidencias en pediatría*, 3(1):24.

Puggioni, G., Gelfand, A. E., and Elmore, J. G. (2008). Joint modeling of sensitivity and specificity. *Statistics in medicine*, 27(10):1745–1761.

Ramırez, J., Segura, J. C., Benıtez, C., de la Torre, A., and Rubio, A. J. (2002). Detector de actividad de voz basado en la distancia de kullback-leibler con aplicaci on a reconocimiento robusto de voz.

Rao, R., Toutenburg, H., Shalabh, and Heumann, C. (2008). *Linear Models and Generalizations: Least Squares and Alternatives.* Berlin Springer.

Rizzo, M. L. (2019). *Statistical computing with R*. Chapman and Hall/CRC.

Scott, D. (2015). *ash: David Scott's ASH Routines*. R package version 1.0-15.

Sharma, D., Yadav, U., and Sharma, P. (2009). The concept of sensitivity and specificity in relation to two types of errors and its application in medical research. *J Reliability Stat Stud*, 2:53–58.

Tanabe, K. and Sagae, M. (1992). An exact cholesky decomposition and the generalized inverse of the variance-covariance matrix of the multinomial distribution, with applications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):211– 219.

Tosteson, T. D., Titus-Ernstoff, L., Baron, J., and Karagas, M. R. (1994). A two-stage validation study for determining sensitivity and specificity. *Environmental health perspectives*, 102(suppl 8):11–14.

Upton, G. J. . (1992). Fisher's exact test. *Journal of Royal statistical Society. series A (Statistics in Society)*, 155(3):395–402.

Versi, E. (1992). "gold standardïs an appropriate term. *BMJ: British Medical Journal*, 305(6846):187.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Reports (1896-1970)*, pages 1432–1449.