TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa

Título: Combinación óptima del número de participantes y el número de medidas repetidas en estudios longitudinales con exposición variable en el tiempo

Autor: Jose Barrera Gómez

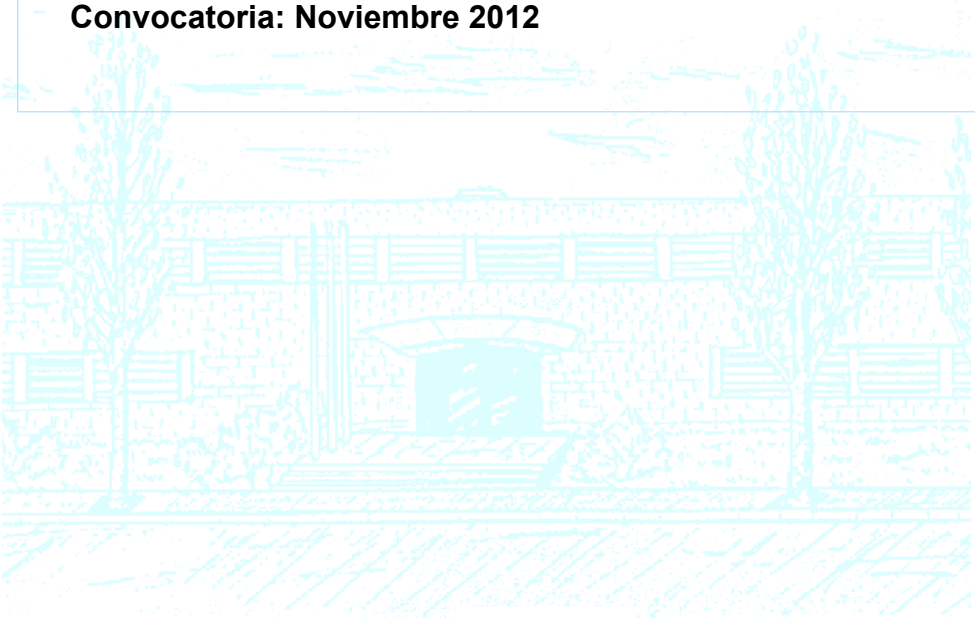
Director: Xavier Basagaña Flores

Institución: Centro de Investigación en Epidemiología Ambiental (CREAL)

Ponente: Josep A. Sánchez Espigares

Departamento: Estadística e Investigación Operativa

Convocatoria: Noviembre 2012



UPC

Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

UNIVERSITAT DE BARCELONA

Combinación óptima del número de participantes y el número de medidas repetidas en estudios longitudinales con exposición variable en el tiempo

Tesis para el
Máster Interuniversitario en Estadística e Investigación Operativa

Universitat Politècnica de Catalunya - Universitat de Barcelona

Autor:
Jose Barrera Gómez
Centro de Investigación en Epidemiología Ambiental (CREAL)

Director:
Xavier Basagaña Flores
Centro de Investigación en Epidemiología Ambiental (CREAL)

Ponente:
Josep Anton Sánchez Espigares
Departamento de Estadística e Investigación Operativa,
Universitat Politècnica de Catalunya

Septiembre de 2012

A Sara, a Marcel y a Laura.

Agradecimientos

Quiero dar las gracias a las personas que, de un modo u otro, me han ayudado en la realización de este trabajo.

A Klaus Langohr y a Xavi Duran que, coincidiendo con ellos en el máster (como alumno del primero y compañero del segundo), me hicieron llegar una oferta de trabajo en el CREAL que supuso el origen de mi nueva carrera profesional como estadístico.

A Xavier Basagaña, por aceptar dirigir este trabajo y también por encontrar siempre un momento para atenderme y ayudarme a seguir aprendiendo día a día.

A Josep A., por aceptar la ponencia de este trabajo y por sus valiosos comentarios que lo han mejorado.

A Juan Ramón González, Alejandro Cáceres y Mikel Esnaola, por su ayuda en la compilación del paquete R, y a Mercè Medina, Josep Maria Antó y Jan-Paul Zock, por su permiso para usar los datos de EPIASLI en el ejemplo ilustrativo.

A todas mis compañeras y a todos mis compañeros de la *República Independent de Planta Baixa*, por el excelente ambiente de trabajo que se disfruta en nuestra sala. Y, claro, por los *vermutillos*.

A mis padres, por la educación que me han brindado y por entender tan bien lo feliz que me hace eso de "estar siempre con los numeritos", a pesar de su preocupación por la precariedad que envuelve la carrera científica en este país.

A Laura, por la paciencia que tiene cuando quiere reclamar mi atención y me ve concentrado ante la pantalla del portátil. Y a Marcel, por interesarse en mis gráficos sobre *Odds ratios*, pidiendo que le explique su significado.

Y, muy especialmente, a Sara, por su interés y apoyo en todo lo que hago, incluyendo la redacción de este documento, y por darme un plus de motivación en cada paso que intento dar.

Abstract

In the context of observational longitudinal studies, we obtained the optimal values of the number of participants and the number of repeated measurements that maximize the power to detect the hypothesized effect, given the total cost of the study, or that minimize the total cost of the study while achieving a required power. We considered a continuous response variable, whose covariance structure was assumed to be exponentially damped, and a binary time-varying exposure, both for an acute and transient exposure effect through a constant mean difference between exposed and non-exposed, and for a cumulative exposure effect through a linearly divergent difference. We derived closed-form expressions for the solution to the problem in the particular case in which the covariance structure of the response is assumed to be compound symmetry. Results showed the relevance of the exposure intraclass correlation regarding the determination of the optimal combination of the number of participants and the number of repeated measurements, and therefore the optimized power or cost. Thus, the assumption of a time-invariant exposure when the exposure is actually time-varying leads to biased results in calculations of sample size and power or cost. We also analyzed the sensitivity of results to allowing dropout, changing the response covariance structure and allowing a time-varying exposure prevalence. As an example, we illustrated some of these results applying our methods in a real study. In addition we provide a software to perform all the calculations required to get the optimal combination of the number of participants and the number of repeated measurements.

Keywords: optimal design; longitudinal study; sample size; intraclass correlation.

MSC2010: 26A06, 62-04, 62-09, 62F99, 62J05, 62K05, 62P10, 68N19.

Resumen

En el contexto de los estudios longitudinales observacionales, hemos obtenido los valores óptimos para el número de participantes y el número de medidas repetidas que maximizan la potencia para detectar el hipotético efecto, sin rebasar un presupuesto económico determinado, o que minimizan el coste total del estudio, alcanzando una potencia determinada. Hemos considerado una variable respuesta continua, cuya estructura de covarianza se supone exponencialmente amortiguada, y una exposición binaria variable en el tiempo, tanto para un efecto agudo y transitorio de la exposición, a través de una diferencia constante en la media entre expuestos y no expuestos, como para un efecto acumulativo, a través de una diferencia linealmente divergente. Hemos derivado expresiones cerradas para la solución del problema en el caso particular en que se supone que la estructura de covarianza de la respuesta es de simetría compuesta. Los resultados mostraron la importancia de la correlación intraclase de la exposición sobre la determinación de la combinación óptima del número de participantes y el número de medidas repetidas y, por tanto, sobre la potencia o el coste optimizados. Así, la incorrecta asunción de una exposición constante conduce a resultados sesgados en el cálculo del tamaño de la muestra y de la potencia o el coste del estudio. También analizamos la sensibilidad de los resultados a la presencia de datos faltantes debido al abandono de participantes durante el seguimiento, al cambio de la estructura de covarianza de la respuesta y a una prevalencia de la exposición variable en el tiempo. Como ejemplo, ilustramos algunos de estos resultados aplicando nuestros métodos en un estudio real. Además, proporcionamos un software para realizar todos los cálculos necesarios para obtener la combinación óptima del número de participantes y el número de medidas repetidas.

Palabras clave: diseño óptimo; estudio longitudinal; tamaño de muestra; correlación intraclase.

MSC2010: 26A06, 62-04, 62-09, 62F99, 62J05, 62K05, 62P10, 68N19.

Índice general

1. Introducción **1**

1.1. Presentación 1

1.2. Motivación 2

2. El problema de asignación óptima **7**

2.1. Planteamiento del problema 7

2.2. Parametrización y modelos 8

2.3. Estimación de $\tilde{\sigma}^2$ 12

2.4. Adaptación a datos faltantes por abandono durante el seguimiento . 15

2.5. Función a optimizar 16

2.6. Resolución 16

3. Resultados **19**

3.1. Escenarios básicos 19

3.2. Desviaciones de los escenarios básicos 23

3.2.1. Efecto de variar la estructura de covarianza de la respuesta . 23

3.2.2. Efecto de la pérdida de participantes durante el seguimiento . 24

3.2.3. Efecto de una prevalencia de la exposición variable en el tiempo 25

4. Ejemplo ilustrativo **28**

5. Implementación de la metodología en R **32**

5.1. Primeros pasos 32

5.1.1. Función `OA()` 32

5.1.2. Funciones `plotExposedPeriods()` y `plotExposedPeriodsInt()` 34

5.2. Ejemplos de diseño de un estudio longitudinal 36

5.2.1. Estudio 1. Maximización de la potencia 36

5.2.2. Estudio 2. Minimización del coste 39

5.3. Caso particular: diseño de un estudio transversal 39

5.3.1. Estudio 3. Coste de un estudio transversal 39

5.3.2. Estudio 4. Potencia de un estudio transversal 41

5.4. Reproducción de los resultados obtenidos en el capítulo 4 41

6. Discusión **44**

Bibliografía **47**

A. Derivación de $\tilde{\sigma}^2$ en los escenarios básicos **51**

A.1. Caso particular de exposición constante 51

A.1.1. Patrón de respuesta CMD 51

A.1.2. Patrón de respuesta LDD 54

A.2. Caso general de exposición variable en el tiempo 56

A.2.1. Patrón de respuesta CMD 56

A.2.2. Patrón de respuesta LDD 59

B. Adaptación de la función coste a datos faltantes por abandono de participantes durante el seguimiento **62**

C. Obtención de $r_{\mathbf{opt}}$ bajo el patrón de respuesta CMD en el escenario básico **64**

C.1. Exposición constante ($\rho_e = 1$) 64

C.2. Exposición variable en el tiempo ($\rho_e < 1$) 65

C.2.1. Caso $\kappa = 1$ 65

C.2.2. Caso $\kappa > 1$ 65

C.3. Resumen de los resultados 69

D. Obtención de $r_{\mathbf{opt}}$ bajo el patrón de respuesta LDD en el escenario básico **70**

D.1. Exposición constante ($\rho_e = 1$) 70

D.2. Exposición variable en el tiempo ($\rho_e < 1$) 72

D.2.1. Caso $\kappa = 1$ 72

D.2.2. Caso $\kappa > 1$ 72

D.3. Resumen de los resultados 74

E. Publicación de los resultados en congresos, revistas y conferencias **75**

Índice de figuras

1.1. Ejemplo de patrones de exposición intra-individuo variable en el tiempo: diseño *crossover* y estudio logitudinal observacional. 6

2.1. Ilustración de los modelos CMD y LDD. 9
2.2. Distribución del número de períodos bajo exposición por participante, E_i.. Ilustración para $r = 3$, $p_e = 1/4$, asumida constante, y diferentes valores de ρ_e. 14

3.1. Representación gráfica de r_{opt} bajo el modelo CMD en el escenario básico y en el caso de exposición constante. 21
3.2. Número óptimo de medidas repetidas bajo el patrón de respuesta CMD en el escenario básico, en función de los parámetros ρ_e, ρ y κ. 22
3.3. Número óptimo de medidas repetidas bajo el patrón de respuesta LDD en el escenario básico, en función de los parámetros ρ_e, ρ y κ. . 23

4.1. Evolución temporal de la prevalencia de la exposición, para uso de aspirador y de aerosoles, en el ejemplo ilustrativo. 29
4.2. Distribución del número de días bajo exposición por individuo, para uso de aspirador y de aerosoles, en el ejemplo ilustrativo. 30
4.3. Diseño óptimo basado en un estudio sobre salud respiratoria (ejemplo ilustrativo). 31

5.1. Ejemplo del resultado de la función `plotExposedPeriods()`. 35
5.2. Ejemplo del resultado de la función `plot()` al maximizar la potencia. 37
5.3. Ejemplo del resultado de la función `plot()` al minimizar el coste. . . 40

Índice de tablas

1.1. Programas de investigación en el CREAL 5

2.1. Parámetros necesarios para calcular el diseño óptimo. 17
2.2. Descripción de los escenarios básicos. 17

3.1. Expresiones de $\tilde{\sigma}^2$ bajo los patrones de respuesta CMD y LDD en los escenarios básicos 20
3.2. Efecto de θ en r_{opt} bajo el patrón de respuesta LDD 25
3.3. Efecto de una prevalencia de la exposición variable en el tiempo en el número óptimo de medidas repetidas para el patrón de respuesta CMD en el escenario básico 26
3.4. Efecto de una prevalencia de la exposición variable en el tiempo en el número óptimo de medidas repetidas para el patrón de respuesta LDD en el escenario básico 27

4.1. Diseño óptimo basado en un estudio sobre salud respiratoria (ejemplo ilustrativo). 31

C.1. Número total de medidas óptimo bajo el patrón de respuesta CMD en el escenario básico. 69

D.1. Número total de medidas óptimo bajo el patrón de respuesta LDD en el escenario básico. 74

Capítulo 1

Introducción

1.1. Presentación

Los objetivos principales del trabajo científico que se desarrolla en el Centro de Investigación en Epidemiología Ambiental (CREAL)[1] son la identificación de los determinantes ambientales de la salud y la ayuda a su prevención y control. En este contexto, la mayoría de estudios epidemiológicos desarrollados en el CREAL tratan de evaluar estadísticamente la posible asociación entre factores ambientales, conocidos como variables de exposición o exposiciones, y una determinada variable de interés relacionada con el estado de salud, conocida como variable respuesta. La Tabla 1.1 muestra los diferentes programas en que se estructura la investigación del CREAL y su caracterización en base a la exposición y a la respuesta de interés. La estadística juega un papel primordial en todos los estudios desarrollados en el centro hasta el punto que un 15 % de aproximadamente la centena de investigadores del centro son estadísticos y resulta prácticamente imposible encontrar un artículo publicado por afiliados al CREAL en el que no figure ningún estadístico como coautor. Lo anterior, junto con la diversidad de procedimientos y métodos involucrados, hacen del CREAL un lugar muy interesante tanto para el inicio como para el desarrollo de una carrera profesional en el ámbito de la bioestadística. Concretamente, a nivel personal y desde mi llegada al centro en octubre de 2009, he tenido la oportunidad de participar, dirigido por mi supervisor, el Dr. Xavier Basagaña, en diversos estudios epidemiológicos [1–3] y problemas metodológicos [4–6], algunos de los cuales han derivado en la creación de paquetes de R para la implementación de la metodología desarrollada [4, 5].

Desde mi punto de vista, cualquiera de los trabajos que he citado anteriormente resultaría adecuado como Trabajo de Fin de Máster pero decidí elegir para ello el artículo titulado *Optimal combination of number of participants and number of*

[1] `http://www.creal.cat`

repeated measurements in longitudinal studies with time-varying exposure [5] porque se trata del primer artículo que he escrito, dirigido por mi supervisor, y esto le da un carácter especial para mí. Además, el trabajo combina tres aspectos de la estadística característicos del trabajo que desarrollo en el CREAL: el desarrollo metodológico, su implementación en R y su aplicación a datos reales.

1.2. Motivación

En la fase del diseño de un estudio en el que se desea valorar estadísticamente una determinada hipótesis científica, adquieren especial relevancia dos aspectos: la cantidad de información que se obtiene sobre la población en estudio, habitualmente resumida en forma de base de datos, y el modelo teórico que intenta imitar la realidad y que, ajustado a partir de los datos obtenidos, sirve de herramienta de decisión sobre la certeza de la hipótesis formulada. Idealmente, se desea que el modelo detecte la hipótesis planteada cuando esta es cierta pero eso no siempre ocurre debido tanto a la definición del modelo (relacionada con su capacidad para "reproducir la realidad") como a la calidad de los datos (relacionada con, por ejemplo, errores de medida o la no consideración de aspectos de la población relacionados con la naturaleza de la hipótesis estudiada), y a su cantidad, es decir, al tamaño muestral. Entendemos por potencia del estudio la probabilidad de que sus resultados nos inclinen a aceptar la hipótesis analizada cuando es cierta; es decir, de detectar un determinado efecto real. Si, fijado el modelo, se llevase a cabo la recogida de los datos repetidas veces, bajo las mismas condiciones, se tendría que la potencia del estudio para detectar un determinado efecto aumentaría estadísticamente con el tamaño de la muestra. Esta relación entre potencia y tamaño muestral se cuantifica mediante fórmulas que dependen del modelo y del tamaño del efecto a detectar, y permiten calcular el tamaño de muestra necesario para alcanzar una determinada potencia de manera que, en la fase del diseño del estudio, se sepa la cantidad de información necesaria. Simétricamente, si el tamaño muestral está fijado en tal fase del estudio (por ejemplo, por motivos económicos), puede resultar interesante calcular la potencia que el estudio alcanzará (según el modelo) de manera que se pueda llegar a valorar su viabilidad. Así, los cálculos de potencia y tamaño muestral resultan un aspecto muy importante a la hora de diseñar el estudio.

Cuando el estudio se basa en una única medida sobre cada uno de los individuos que integran la muestra, el tamaño de ésta se caracteriza por el número de individuos que la integran, N. En cambio, si el estudio contempla la repetición de la medida inicial r veces sobre cada individuo (es decir, el número total de medidas por individuo es $r + 1$), el tamaño de la muestra se caracteriza por la combinación (N, r). Por ejemplo, en un posible estudio longitudinal sobre neurodesarrollo infan-

til, podría tomarse información sobre un conjunto de $N = 400$ recién nacidos y, tras esta medida inicial, tomar nuevas medidas a los 2 años de vida, a los 4 y a los 6, siendo $r = 3$.

Las fórmulas para el tamaño de la muestra y la potencia en el contexto de estudios longitudinales que comparan dos grupos (por ejemplo, expuestos y no expuestos) han sido ampliamente estudiadas, pero esencialmente en el contexto de los experimentos. Por lo tanto, suponen que la exposición se asigna por diseño y, en la mayoría de los casos, se considera que es invariante en el tiempo, como típicamente ocurre en los ensayos clínicos (por ejemplo, en la comparación entre un grupo de tratamiento y otro placebo) [7–18]. Algunos estudios han considerado el caso de las exposiciones que varían con el tiempo pero de una manera controlada por el investigador, como en los ensayos con diseño *crossover* [19–21]. Sin embargo, en estudios observacionales longitudinales, la exposición no está controlada por el investigador y por lo general estos estudios involucran exposiciones variables en el tiempo que pueden implicar tanto un gran número de patrones de exposición observados como una alta variabilidad en el número de períodos expuestos por los participantes. En efecto, en la Tabla 1.1 se puede observar que las exposiciones estudiadas en el CREAL corresponden a potenciales factores de riesgo para la salud; es decir, si existe una asociación entre la exposición y la variable de salud estudiadas, se espera que sea perjudicial. Por este motivo, la práctica totalidad de los estudios epidemiológicos desarrollados en el CREAL (y, en general, en los centros de investigación epidemiológica) son observacionales, entendiendo por ello que el equipo investigador que desarrolla el estudio no puede ni debe, por motivos éticos, controlar la distribución de la exposición analizada entre los individuos estudiados. Esta variabilidad de la exposición debe tenerse en cuenta al diseñar el estudio, como indican dos documentos publicados recientemente que muestran que resulta fundamental caracterizar la variación de la exposición intra-individuo con el fin de obtener los cálculos correctos de potencia y tamaño de la muestra en el contexto de los estudios longitudinales observacionales [22, 23]. A modo de ejemplo, la Figura 1.1 ilustra posibles patrones de exposición en dos tipos de estudios diferentes con exposición variable en el tiempo: un diseño *crossover* y un estudio longitudinal observacional.

Por otro lado, en el contexto de los estudios epidemiológicos longitudinales observacionales, en los que generalmente la exposición puede variar a lo largo del tiempo de seguimiento, puede resultar de interés decidir el número de participantes y el número de medidas repetidas cuando ninguno de estos dos parámetros del estudio se han fijado *a priori*. Una manera óptima de resolver tal decisión consiste en seleccionar aquella combinación del número de participantes y el número de medidas repetidas tal que maximice la potencia del estudio, sin superar un presupuesto financiero determinado, o bien minimice el coste económico del estudio, condicionado

a obtener una determinada potencia requerida. Algunos autores han estudiado este problema pero sin considerar una exposición variable en el tiempo [24–33]. Tal como ocurre en el problema de diseño sin restricciones, se espera que el diseño óptimo en el problema con la restricción financiera sea también sensible al grado de variación de la exposición intra-individuo.

Así, en este trabajo hemos estudiado la combinación óptima del número de participantes y el número de medidas repetidas en estudios observacionales longitudinales que maximiza la potencia para detectar un efecto hipotético sin exceder un presupuesto financiero fijado o que minimiza el coste económico del estudio garantizando una determinada potencia. Hemos considerado una variable respuesta continua con una estructura de covarianza amortiguada exponencialmente (DEX, *damped exponential*) así como una exposición binaria que puede variar en el tiempo. Hemos analizado dos patrones de respuesta diferentes bajo la hipótesis alternativa, uno asumiendo un efecto agudo y transitorio de la exposición y el otro, un efecto acumulativo. Hemos explorado el efecto de diversos parámetros sobre el diseño óptimo, incluyendo el porcentaje de individuos perdidos durante el seguimiento. Hemos ilustrado nuestra metodología usando datos de un estudio sobre los efectos respiratorios del uso de productos de limpieza [34]. Además, facilitamos la aplicación de la metodología desarrollada a través de la creación de un paquete de R que permite realizar todos los cálculos para obtener el diseño óptimo del estudio.

Tabla 1.1: Programas de investigación en el CREAL, caracterizados por las exposiciones y las variables respuesta de interés.

Programa	Exposiciones	Variables respuesta
Respiratorio	Factores ambientales, laborales y genéticos	Enfermedades respiratorias (asma y EPOC[†])
Cáncer	Factores ambientales y genéticos	Vejiga urinaria, mama, colon, leucemia, linfomas,...
Salud infantil	Partículas finas, organocloratos, mercurio,...	Crecimiento[§], reproducción, sistema neuroconductual
Contaminación atmosférica	Tráfico, industria y otras	Sistema cardiorrespiratorio y del neurodesarrollo
Contaminación del agua	Productos desinfectantes de agua potable y piscinas	Cáncer, enfermedades respiratorias, trastornos reproductivos,...
Radiaciones	Radiaciones ionizantes (☢) y no ionizantes (📡)	Tumores y otras

[†] Enfermedad Pulmonar Obstructiva Crónica.
[§] Intrauterino, postnatal e infantil.

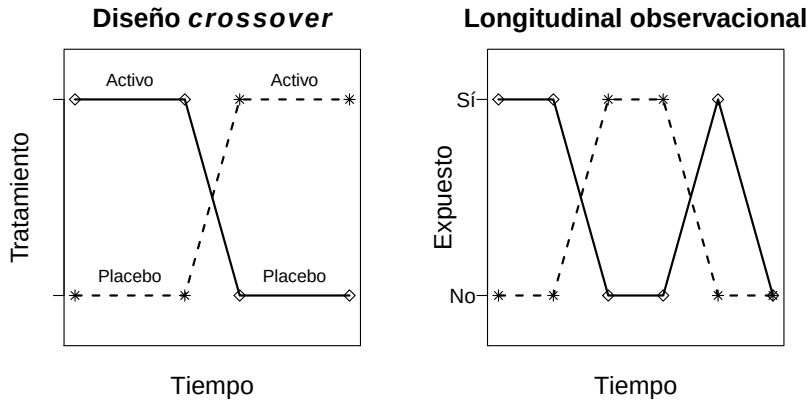


Figura 1.1: Ejemplo de patrones de exposición variable en el tiempo según el tipo de estudio. En un diseño *crossover* como el de la figura izquierda existen solo dos patrones de exposición, correspondientes a dos grupos de seguimiento. En el primer grupo (línea discontinua), los individuos reciben el tratamiento placebo durante una primera etapa del estudio mientras el otro grupo (línea continua) recibe el tratamiento activo. Después de una etapa de cruzado, los tratamientos se intercambian en los grupos. El tratamiento intra-individuo varía en el tiempo pero de una manera controlada por el investigador. Bajo este diseño, todos los individuos presentan el tratamiento activo durante el mismo período de tiempo. En canvio, en un estudio longitudinal observacional (figura de la derecha), cada individuo puede presentar variabilidad en su estado de exposición a lo largo del tiempo de manera no controlada por el investigador y pudiendo dar lugar a un elevado número de patrones de exposición diferentes (en la figura se han representado dos posibles patrones de exposición). Supongamos, por ejemplo, que la figura de la derecha representa la trayectoria de exposición diaria de dos trabajadores de la limpieza y que el estado de exposición refleja si el trabajador estuvo expuesto cada día a un determinado producto de limpieza. Entonces, el individuo cuya trayectoria de exposición está representada por la línea discontinua ha estado expuesto en las medidas tercera y cuarta, y no lo ha estado en las medidas primera, segunda, quinta y sexta. En cambio, el individuo cuya trayectoria de exposición está representada por la línea continua ha estado expuesto en las medidas primera, segunda y quinta, y no lo ha estado en las medidas tercera y cuarta. Además, estos dos individuos no han estado expuestos el mismo número de períodos.

Capítulo 2

El problema de asignación óptima

2.1. Planteamiento del problema

Supongamos que nos planteamos llevar a cabo un estudio longitudinal observacional en el que la primera medida que se realizará a cada uno de los N participantes en el estudio tiene un coste económico c_1 que es κ veces más elevado que el de cada una de las medidas subsiguientes, es decir,

$$\kappa = \frac{\text{Coste de la } 1^{\underline{\text{a}}} \text{ medida } (c_1)}{\text{Coste de cada una de las medidas subsiguientes}} \geq 1.$$

El coste no menor (en general, mayor) de la medida inicial del estudio está justificado por el hecho de que en él se incluyen los gastos derivados del reclutamiento del participante. De esta manera, el coste económico total del reclutamiento de N participantes y de su seguimiento a través de $r+1$ medidas (la medida inicial más r repeticiones) es

$$\text{Coste} = C(N, r) = Nc_1 \left(1 + \frac{r}{\kappa}\right). \tag{2.1}$$

Por otro lado, la potencia de un estudio para la significación de un determinado coeficiente de regresión, β, mediante la prueba de Wald se puede expresar como

$$\text{Potencia} = \Phi\left(\frac{|\tilde{\beta}|\sqrt{N}}{\tilde{\sigma}} - z_{1-\alpha/2}\right), \tag{2.2}$$

donde $\tilde{\beta}$ es el valor de β bajo la hipótesis alternativa, $\tilde{\sigma}/\sqrt{N}$ es su error estándar, α es el nivel de significación y z_q y $\Phi(\cdot)$ son, respectivamente, el q-ésimo cuantil y la densidad acumulada de la distribución normal estándar.

Nuestro interés radica en encontrar la combinación óptima de N y r (N_{opt},

r_{opt}) que maximiza la potencia del estudio sin exceder un determinado presupuesto monetario, es decir,

$$\begin{array}{ll} \text{máx}_{N,r} & \text{Potencia} \\ \text{sujeto a} & \text{Coste} \ \leq \ \text{Presupuesto} \end{array} .$$

Se puede demostrar fácilmente que el valor de r que resuelve este problema es el mismo que resuelve el problema de minimizar el coste del estudio garantizando una determinada potencia, es decir,

$$\begin{array}{ll} \text{mín}_{N,r} & \text{Coste} \\ \text{sujeto a} & \text{Potencia} \ \geq \ \text{Potencia}_{\text{requerida}} \end{array} .$$

En efecto, teniendo en cuenta la expresión (2.1) para el coste, ambos problemas de optimización son equivalentes a resolver el siguiente problema de optimización sin restricciones:

$$\min_{r \in \mathbb{N}} (\kappa + r)\tilde{\sigma}^2, \tag{2.3}$$

cuya solución proporciona el valor de r_{opt}. Después de ello, el valor de N_{opt} se puede obtener de la restricción en el problema de optimización como se detallará más adelante.

La resolución del problema (2.3) requiere obtener previamente una expresión para $\tilde{\sigma}^2$ que dependerá de la modelización estadística.

2.2. Parametrización y modelos

Sin pérdida de generalidad, definimos la duración del estudio como la unidad temporal. Como es habitual en estudios de cohorte, asumimos que todas las medidas para todos los participantes son tomadas en el mismo conjunto de puntos temporales, $t = 0, 1/r, 2/r, \ldots, 1$, donde $1/r$ es el intervalo temporal entre dos medidas consecutivas.

Asumimos una estructura lineal para la media, $\mathbb{E}(\mathbf{Y}_i) = \mathbf{X}_i\mathbf{B}$ $(i = 1, \ldots, N)$, donde $\mathbf{Y}_i$ y $\mathbf{X}_i$ son la variable respuesta continua de interés y la matriz de covariables para el participante i, respectivamente, $\mathbf{B}$ es el vector de parámetros de regresión desconocidos, y $\text{Var}(\mathbf{Y}_i|\mathbf{X}_i) = \mathbf{\Sigma}$ $(i = 1, \ldots, N)$, donde $\mathbf{\Sigma}$ es la matriz $(r+1)\times(r+1)$ de covarianzas residual, supuesta igual para todos los participantes.

Para un efecto agudo y transitorio de la exposición, definimos el patrón de diferencia de medias constante para la respuesta (CMD, *constant mean difference*) como

$$\mathbb{E}(Y_{ij}|E_{ij}) = \beta_0 + \beta_1 t_j + \beta_2 E_{ij}, \tag{2.4}$$

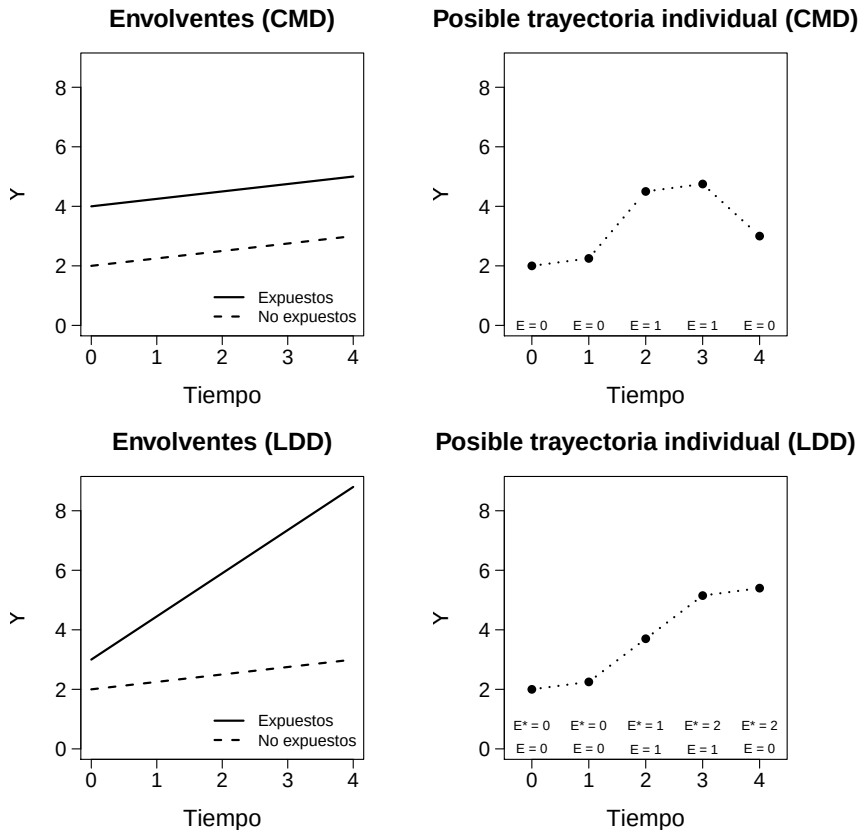


Figura 2.1: Ilustración de los modelos CMD y LDD. Las figuras de la izquierda representan las envolventes; es decir, las trayectorias correspondientes a los patrones de exposición nula (línea discontinua) y de máxima exposición (línea continua). Se observa como el modelo CMD asume que la diferencia entre ambas envolventes es una constante (el efecto de la exposición). En cambio, dichos patrones difieren linealmente bajo el modelo LDD. Las figuras de la derecha ilustran potenciales trayectorias de un individuo cuya variable de exposición se indica en el cuerpo del gráfico.

donde Y_{ij} y $E_{ij} \in \{0, 1\}$ son la respuesta y el estado de exposición, respectivamente, para el i-ésimo participante en la j-ésima medida, y $t_j = j/r,\ j = 0, 1, \ldots, r$ son los puntos temporales de medición. Bajo este modelo, asumimos que cuando el individuo deja de estar expuesto, la variable respuesta vuelve a su nivel normal, tal como ilustra la Figura 2.1. El parámetro de interés es β_2, que se puede interpretar como la diferencia esperada en la media de la variable respuesta, en cualquier punto temporal, entre los expuestos y los no expuestos. El valor mínimo de r en el modelo (2.4) es 0, correspondiendo a un estudio transversal.

Para un efecto acumulativo de la exposición, definimos el patrón de diferencia de medias linealmente divergente para la respuesta (LDD, *linearly divergent difference*)

como

$$\mathbb{E}(Y_{ij}|E_{ij}, t_j) = \beta_0 + \beta_1 E_{i0} + \beta_2 t_j + \beta_3 E^{\star}_{ij}, \tag{2.5}$$

donde $E^{\star}_{ij} = \frac{1}{r}\sum_{k=1}^{j} E_{ik}$ es la exposición acumulada para el i-ésimo individuo en la j-ésima medida, asumiendo $E^{\star}_{i0} = 0$ para todos los participantes. Bajo este modelo, asumimos que cuando el individuo deja de estar expuesto, la variable respuesta no recupera su nivel normal sino que queda afectada permanentemente por la exposición acumulada anteriormente, tal como ilustra la Figura 2.1. El parámetro de interés es β_3, que se puede interpretar como la diferencia esperada en la media de la variable respuesta al final del seguimiento entre el peor patrón de exposición posible (es decir, aquellos individuos expuestos en todas las medidas) y aquellos que no estuvieron expuestos en ningún momento durante el seguimiento. Para el caso particular de una exposición invariante en el tiempo, el modelo (2.5) es equivalente a un modelo con los efectos principales del tiempo y de la exposición, y su interacción. El valor mínimo de r en el modelo (2.5) es 1, dado que se necesitan al menos dos medidas para estimar la tasa de cambio temporal.

Para caracterizar la estructura de covarianza de la respuesta, consideraremos una amortiguación exponencial de la correlación entre medidas (DEX, *damped exponential structure*) [35], cuya matriz de covarianzas tiene elementos en la diagonal con valor σ^2, la varianza residual, y elementos $[j, j']$ fuera de la diagonal de valor $\sigma^2\rho^{|\frac{j'-j}{r}|^{\theta}}$ donde ρ es la correlación entre la primera y la última medida de la respuesta y $\theta \in [0, 1]$ es el parámetro de amortiguación:

$$\mathbf{\Sigma}_{\text{DEX}} = \sigma^2 \begin{pmatrix} 1 & \rho^{\left(\frac{1}{r}\right)^{\theta}} & \rho^{\left(\frac{2}{r}\right)^{\theta}} & \cdots & \rho^{\left(\frac{r-2}{r}\right)^{\theta}} & \rho^{\left(\frac{r-1}{r}\right)^{\theta}} & \rho \\ \rho^{\left(\frac{1}{r}\right)^{\theta}} & 1 & \rho^{\left(\frac{1}{r}\right)^{\theta}} & \cdots & \rho^{\left(\frac{r-3}{r}\right)^{\theta}} & \rho^{\left(\frac{r-2}{r}\right)^{\theta}} & \rho^{\left(\frac{r-1}{r}\right)^{\theta}} \\ \rho^{\left(\frac{2}{r}\right)^{\theta}} & \rho^{\left(\frac{1}{r}\right)^{\theta}} & 1 & \cdots & \rho^{\left(\frac{r-4}{r}\right)^{\theta}} & \rho^{\left(\frac{r-3}{r}\right)^{\theta}} & \rho^{\left(\frac{r-2}{r}\right)^{\theta}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \rho^{\left(\frac{r-2}{r}\right)^{\theta}} & \rho^{\left(\frac{r-3}{r}\right)^{\theta}} & \rho^{\left(\frac{r-4}{r}\right)^{\theta}} & \cdots & 1 & \rho^{\left(\frac{1}{r}\right)^{\theta}} & \rho^{\left(\frac{2}{r}\right)^{\theta}} \\ \rho^{\left(\frac{r-1}{r}\right)^{\theta}} & \rho^{\left(\frac{r-2}{r}\right)^{\theta}} & \rho^{\left(\frac{r-3}{r}\right)^{\theta}} & \cdots & \rho^{\left(\frac{1}{r}\right)^{\theta}} & 1 & \rho^{\left(\frac{1}{r}\right)^{\theta}} \\ \rho & \rho^{\left(\frac{r-1}{r}\right)^{\theta}} & \rho^{\left(\frac{r-2}{r}\right)^{\theta}} & \cdots & \rho^{\left(\frac{2}{r}\right)^{\theta}} & \rho^{\left(\frac{1}{r}\right)^{\theta}} & 1 \end{pmatrix}.$$

Consideraremos dos casos particulares de esta estructura, la de simetría com-

puesta (CS, *compound symmetry*), cuando $\theta = 0$,

$$\mathbf{\Sigma}_{\text{CS}} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho & \rho & \rho \\ \rho & 1 & \rho & \cdots & \rho & \rho & \rho \\ \rho & \rho & 1 & \cdots & \rho & \rho & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \rho & \rho & \rho & \cdots & 1 & \rho & \rho \\ \rho & \rho & \rho & \cdots & \rho & 1 & \rho \\ \rho & \rho & \rho & \cdots & \rho & \rho & 1 \end{pmatrix}$$

y la autorregresiva de primer orden (AR(1), *first order autoregressive*), cuando $\theta = 1$,

$$\mathbf{\Sigma}_{\text{AR(1)}} = \sigma^2 \begin{pmatrix} 1 & \rho^{\frac{1}{r}} & \rho^{\frac{2}{r}} & \cdots & \rho^{\frac{r-2}{r}} & \rho^{\frac{r-1}{r}} & \rho \\ \rho^{\frac{1}{r}} & 1 & \rho^{\frac{1}{r}} & \cdots & \rho^{\frac{r-3}{r}} & \rho^{\frac{r-2}{r}} & \rho^{\frac{r-1}{r}} \\ \rho^{\frac{2}{r}} & \rho^{\frac{1}{r}} & 1 & \cdots & \rho^{\frac{r-4}{r}} & \rho^{\frac{r-3}{r}} & \rho^{\frac{r-2}{r}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \rho^{\frac{r-2}{r}} & \rho^{\frac{r-3}{r}} & \rho^{\frac{r-4}{r}} & \cdots & 1 & \rho^{\frac{1}{r}} & \rho^{\frac{2}{r}} \\ \rho^{\frac{r-1}{r}} & \rho^{\frac{r-2}{r}} & \rho^{\frac{r-3}{r}} & \cdots & \rho^{\frac{1}{r}} & 1 & \rho^{\frac{1}{r}} \\ \rho & \rho^{\frac{r-1}{r}} & \rho^{\frac{r-2}{r}} & \cdots & \rho^{\frac{2}{r}} & \rho^{\frac{1}{r}} & 1 \end{pmatrix}.$$

Algunos ejemplos de la estructura DEX son los siguientes, todos ellos con valores fijados $r = 4$, $\sigma^2 = 4$ y $\rho = 0{,}5$, y con valores de θ iguales a 0 (CS), 0,3, 0,7 y 1 (AR(1)):

$$\mathbf{\Sigma}_{\text{CS}} = \begin{pmatrix} 4 & 2 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 & 2 \\ 2 & 2 & 4 & 2 & 2 \\ 2 & 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 2 & 4 \end{pmatrix},$$

$$\mathbf{\Sigma}_{\theta=0,3} \approx \begin{pmatrix} 4 & 2{,}53 & 2{,}28 & 2{,}12 & 2{,}00 \\ 2{,}53 & 4 & 2{,}53 & 2{,}28 & 2{,}12 \\ 2{,}28 & 2{,}53 & 4 & 2{,}53 & 2{,}28 \\ 2{,}12 & 2{,}28 & 2{,}53 & 4 & 2{,}53 \\ 2{,}00 & 2{,}12 & 2{,}28 & 2{,}53 & 4 \end{pmatrix},$$

$$\mathbf{\Sigma}_{\theta=0,7} \approx \begin{pmatrix} 4 & 3{,}08 & 2{,}61 & 2{,}27 & 2{,}00 \\ 3{,}08 & 4 & 3{,}08 & 2{,}61 & 2{,}27 \\ 2{,}61 & 3{,}08 & 4 & 3{,}08 & 2{,}61 \\ 2{,}27 & 2{,}61 & 3{,}08 & 4 & 3{,}08 \\ 2{,}00 & 2{,}27 & 2{,}61 & 3{,}08 & 4 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{\mathrm{AR}(1)} \approx \begin{pmatrix} 4 & 3{,}36 & 2{,}83 & 2{,}38 & 2{,}00 \\ 3{,}36 & 4 & 3{,}36 & 2{,}83 & 2{,}38 \\ 2{,}83 & 3{,}36 & 4 & 3{,}36 & 2{,}83 \\ 2{,}38 & 2{,}83 & 3{,}36 & 4 & 3{,}36 \\ 2{,}00 & 2{,}38 & 2{,}83 & 3{,}36 & 4 \end{pmatrix}.$$

2.3. Estimación de $\tilde{\sigma}^2$

Los cálculos se basarán en la asunción de que el análisis de los datos se realizará mediante mínimos cuadrados generalizados (GLS, *generalized least squares*), para tener en cuenta la correlación entre observaciones de un mismo individuo.

Dado que la matriz de covariables $\mathbf{X}_i$ es desconocida *a priori*, usaremos, siguiendo a Whittemore [36] y Shieh [37], la varianza asintótica de la estimación GLS de $\mathbf{B}$ que es $\frac{1}{N}\boldsymbol{\Sigma}_B$, donde, siguiendo a Yi [16],

$$\boldsymbol{\Sigma}_B = \left\{\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]\right\}^{-1}, \tag{2.6}$$

y $\frac{1}{N}\boldsymbol{\Sigma}_B$ puede ser especificada completamente por los momentos de primer y segundo orden de la distribución de las covariables dado que $\boldsymbol{\Sigma}$ es independiente de las covariables en nuestro contexto [38].

Estamos interesados en el $[m,m]$-ésimo elemento de $\boldsymbol{\Sigma}_B$, donde m es la posición de β en el vector de coeficientes de regresión $\mathbf{B}$ o, equivalentemente, la posición de la columna asociada al parámetro β de interés en la matriz de diseño, $\mathbf{X}_i$,

$$\tilde{\sigma}^2 = \boldsymbol{\Sigma}_{B[m,m]}. \tag{2.7}$$

Por otro lado, también es necesario conocer el vector de prevalencias[1] de la exposición, $\boldsymbol{p}_e = (p_{e0}, \ldots, p_{er})$. Asumiremos que la prevalencia de la exposición puede variar linealmente desde p_{e0} en la medida inicial hasta p_{er} en la medida al final del seguimiento. Así, consideraremos una tendencia lineal parametrizada como

$$p_{ej} = \frac{1+\gamma j/r}{1+\gamma/2}\bar{p}_e, \quad j = 0, 1, \ldots, r\,, \tag{2.8}$$

donde $\bar{p}_e = \frac{p_{e0}+p_{er}}{2}$ es la prevalencia media de la exposición a lo largo de las $r+1$ medidas y $\gamma = \frac{p_{er}-p_{e0}}{p_{e0}}$ es el cambio relativo en la prevalencia de la exposición entre la primera medida y la última. El caso particular en que la prevalencia de la exposición

[1] En el ámbito de la bioestadística, se utiliza el término prevalencia para denotar la proporción de unidades de estudio que presentan una determinada característica. Por ejemplo, si en una determinada población, en un determinado momento, el 28 % de los individuos fuma, podemos decir que la prevalencia de fumadores en tal población en ese momento es 0,28.

es constante corresponde a $\gamma = 0$.

La presencia del cálculo de la esperanza en la expresión (2.6) implica la necesidad de caracterizar la estructura de covarianza de la exposición, $\mathbf{\Sigma}_E$, bajo los modelos (2.4) y (2.5). Puede probarse que conocer la correlación intraclase de la exposición permite determinar el valor exacto de $\tilde{\sigma}^2$ si la estructura de covarianza de la respuesta, $\mathbf{\Sigma}$, es CS y el patrón de respuesta es CMD (Apéndice A) o una buena aproximación en caso contrario [22, 23], sin la necesidad de caracterizar la matriz $\mathbf{\Sigma}_E$ completamente. La correlación intraclase de la exposición se define como

$$\rho_e = \frac{\text{sum}(\mathbf{\Sigma}_E) - \text{Tr}(\mathbf{\Sigma}_E)}{r\text{Tr}(\mathbf{\Sigma}_E)}, \tag{2.9}$$

donde sum() y Tr() denotan la suma de los elementos y la traza de una matriz respectivamente [39]. El parámetro ρ_e puede interpretarse como una medida de la variación intra-individuo de la exposición. Cuando ρ_e toma su valor máximo, 1, cada uno de los individuos presenta una exposición constante (es decir, o está expuesto durante todo el seguimiento o no lo está en ningún momento del mismo). Opuestamente, cuando ρ_e toma su valor mínimo, la variabilidad intra-individuo de la exposición es máxima [39]. El límite superior de ρ_e es menor que 1 cuando la prevalencia de la exposición varía en el tiempo [22] y 1 en caso contrario. Para variables binarias, como es nuestro caso, el límite inferior de ρ_e es

$$-\frac{1}{r} + \frac{\text{frac}[(r+1)\bar{p}_e]\,\{1 - \text{frac}[(r+1)\bar{p}_e]\}}{r(r+1)\bar{p}_e(1-\bar{p}_e)}$$

donde frac(x) denota la parte decimal de x [40]. En el caso particular en que $\rho_e = -1/r$, la exposición intra-individuo presentaría una variabilidad máxima, alternando, cada uno de los individuos, estados de exposición con estados de no exposición y todos los individuos tenderían a presentar el mismo número de períodos expuestos. Este caso particular sería análogo a un estudio con diseño *crossover*. En el extremo opuesto, en el caso particular en que $\rho_e = 1$, no existe variabilidad intra-individuo en la exposición de manera que sólo existen dos patrones posibles de exposición: individuos no expuestos en ninguna medida e individuos expuestos en todas las medidas. Este caso sería equivalente a comparar dos grupos paralelos. En general, en estudios observacionales tendremos valores intermedios de ρ_e. Para el caso particular en que la estructura de covarianza de la exposición E, $\mathbf{\Sigma}_E$, es CS y la prevalencia de la exposición es constante, ρ_e coincide con el elemento común fuera de la diagonal de la matriz de correlación de E.

A modo de herramienta para decidir el valor apropiado de ρ_e, puede resultar útil explorar la distribución del número de períodos bajo exposición por participante, una vez fijados los valores de ρ_e, del número de medidas repetidas, r, de la prevalencia

de la exposición, asumida constante, $p_{e_j} = p_e$, $\forall j = 0, \ldots, r$, y asumiendo que $\mathbf{\Sigma}_E$ tiene estructura CS. Por ejemplo, la Figura 2.2 muestra tal distribución para el caso en que $r = 3$ y $p_e = 1/4$, para $\rho_e = -1/3, 0, 1/2$ y 1. Nuestro paquete de R puede ser empleado para crear este tipo de gráficas (ver capítulo 5, sección 5.1.2). También puede tenerse en cuenta la naturaleza de la exposición a la hora de fijar el valor de ρ_e. Por ejemplo, en un estudio en que se considera el consumo de tabaco como exposición de interés, cabe esperar poca variabilidad intra-individuo en la exposición ya que los fumadores tenderán a seguir siéndolo durante el seguimiento igual que los no fumadores también tenderán a mantener su estado. Así, en un caso como este, cabría esperar valores de ρ_e cercanos a 1. En cambio, ciertas actividades laborales relacionadas con la actividad física variada (operarios de construcción, mantenimiento, stock de almacén, etc) están asociadas a ciertas actividades de riesgo para lesiones cervicales (posturas, movimiento de objetos pesados, etc) que pueden variar en el día a día en función del tipo de actividad concreta que deba realizarse, de manera que algunas exposiciones pueden mostrar una elevada variabilidad a lo largo del tiempo, esperándose entonces un valor de ρ_e más cercano a 0.

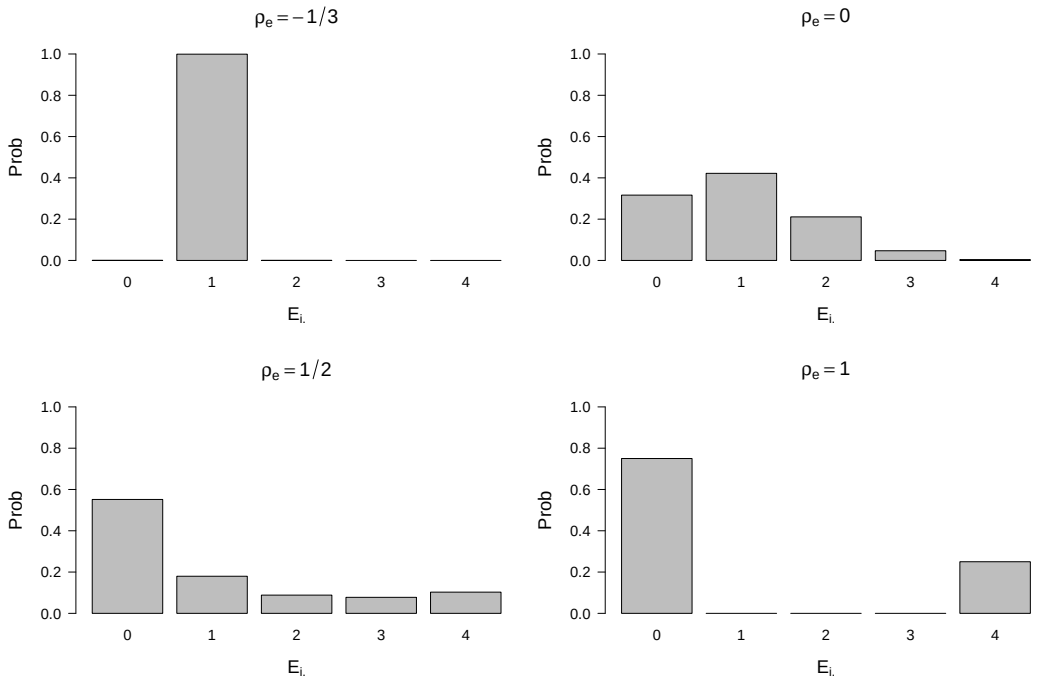


Figura 2.2: Distribución del número de períodos bajo exposición, $E_{i\cdot}$, para $r = 3$, $p_e = 1/4$, asumida constante, y diferentes valores de ρ_e, asumiendo estructura CS para la covarianza de la exposición con parámetro ρ_e, y sin abandono durante el seguimiento.

Para algunas estructuras de covarianza de $\mathbf{\Sigma}$, tales como CS, existe una expresión cerrada para (2.7), como se expone en la sección 3.1, mientras que para otras, como

DEX (incluyendo AR(1)), debe ser calculada numéricamente.

2.4. Adaptación a datos faltantes por abandono durante el seguimiento

En estudios epidemiológicos longitudinales no es infrecuente la pérdida de seguimiento por abandono de algún individuo, por lo que resulta más realista tener en cuenta esta posibilidad en nuestro problema. Así, consideraremos la posible existencia de un patrón monótono de abandono; es decir, la pérdida de las medidas sobre un determinado individuo en una determinada medición implica la pérdida de todas sus medidas subsiguientes. Asumimos que no hay datos faltantes en la primera medida y que cada individuo que no ha salido del estudio en una determinada medición tiene asociada una probabilidad π_m de abandonar el estudio antes de la siguiente medición. Así, existen $r+1$ patrones de pérdida posible. Representando cada uno de estos patrones por una cadena de unos (presencia) y ceros (ausencia) de longitud $r+1$, el patrón g-ésimo es $(1, \overset{(g)}{\ldots}, 1, 0, \overset{(r+1-g)}{\ldots}, 0)$ con probabilidad

$$\pi_g = \begin{cases} \pi_m(1-\pi_m)^{g-1} & , \ g = 1, 2, \ldots, r \\ (1-\pi_m)^r & , \ g = r+1 \end{cases} . \tag{2.10}$$

Si $r = 0$, solo habrá una medida y, por tanto, $\pi_m = 0$.

La experiencia de algunos investigadores les puede permitir una estimación aproximada *a priori* de la fracción de individuos para los cuales no se completa el seguimiento, π_M, mientras que la probabilidad π_m puede resultar más complicada de estimar. Por este motivo, consideraremos el parámetro π_M que se relaciona con π_m según la expresión

$$\pi_M = 1 - (1-\pi_m)^r.$$

En presencia de la estructura de datos faltantes considerada, el coste total del estudio cambia de manera que la función (2.1) puede ser adaptada mediante el cálculo del coste esperado sobre los $r+1$ patrones de datos faltantes (Apéndice B), que queda

$$\text{Coste} = C(N, r) = Nc_1 \left\{ 1 + \frac{(1-\pi_m)[1-(1-\pi_m)^r]}{\pi_m \kappa} \right\} . \tag{2.11}$$

Asimismo, la expresión (2.6) también se ve afectada por la estructura de datos faltantes de manera que, ponderando por los diferentes patrones de abandono, queda,

$$\boldsymbol{\Sigma}_B = \left\{ \mathbb{E}_X \left[\sum_{g=1}^{r+1} \mathbf{X}_{ig}' \boldsymbol{\Sigma}_g^{-1} \mathbf{X}_{ig} \pi_g \right] \right\}^{-1} , \tag{2.12}$$

donde π_g es definida en (2.10).

2.5. Función a optimizar

Combinando las expresiones (2.3), (2.7) y (2.12), la función a optimizar se puede expresar como

$$\min_{r\in\mathbb{N}} \left\{\kappa + \frac{1-\pi_m}{\pi_m}\left[1-(1-\pi_m)^r\right]\right\} \left\{\mathbb{E}_X\left[\sum_{g=1}^{r+1} \mathbf{X}_{ig}'\boldsymbol{\Sigma}_g^{-1}\mathbf{X}_{ig}\pi_g\right]\right\}^{-1}_{[m,m]}, \tag{2.13}$$

cuya resolución proporcionará una solución para r_{opt}.

Una vez obtenido el valor de r_{opt}, el número óptimo de participantes, N_{opt}, se obtiene como se detalla a continuación. Si el investigador está interesado en maximizar la potencia del estudio sin exceder un presupuesto dado, entonces debe proporcionar el coste de la primera medida (c_1) y N_{opt} se obtiene de la expresión (2.11) como

$$N_{\text{opt}} = \lfloor \tilde{N} \rfloor, \text{ donde } \quad \tilde{N} = \begin{cases} \dfrac{\text{Presupuesto}}{c_1\left(1+\frac{r_{\text{opt}}}{\kappa}\right)} & , \text{ si } \pi_M = 0 \\ \dfrac{\text{Presupuesto}}{c_1\left\{1+\frac{\pi_M}{\kappa\left[(1-\pi_M)^{-\frac{1}{r_{\text{opt}}}}-1\right]}\right\}} & , \text{ si } \pi_M > 0 \end{cases} \tag{2.14}$$

y la función suelo $\lfloor x \rfloor$ denota el mayor número entero no mayor que x.

Si el investigador está interesado en lograr una determinada potencia minimizando el coste del estudio, entonces se requiere el valor de dicha potencia y el efecto esperado bajo la hipótesis alternativa, $\tilde{\beta}$, de manera que N_{opt} se obtiene de la expresión (2.2) como

$$N_{\text{opt}} = \lceil \tilde{N} \rceil, \text{ donde } \quad \tilde{N} = \frac{\tilde{\sigma}^2_{\text{opt}}}{|\tilde{\beta}|^2}\left[z_{1-\alpha/2} + \Phi^{-1}(\text{Potencia})\right]^2 \tag{2.15}$$

y $\tilde{\sigma}^2_{\text{opt}}$ es el resultado de calcular la expresión (2.7) para $r = r_{\text{opt}}$ y la función techo $\lceil x \rceil$ denota el menor número entero no menor que x.

Todos los parámetros que el investigador debe proporcionar para calcular la combinación óptima del número de participantes y el número de medidas repetidas se describen en la Tabla 2.1.

2.6. Resolución

La optimización (2.13) no tiene por resultado, en general, una expresión analítica cerrada y debe resolverse numéricamente, salvo para unos casos particulares

Tabla 2.1: Parámetros necesarios para calcular la asignación óptima, $(r_{\text{opt}}, N_{\text{opt}})$.

Parámetro	Descripción
Patrón	Patrón asumido para Y (CMD o LDD)
σ^2	Elemento de la diagonal de la matriz DEX de covarianzas de Y
ρ	Correlación entre las medidas inicial y final de Y
θ	Parámetro de amortiguamiento de la matriz DEX de covarianzas de Y
ρ_e	Correlación intraclase de la exposición E
p_{e0}	Prevalencia de la exposición E en la medida inicial (en $t = 0$)
p_{er}	Prevalencia de la exposición E en la medida al final del seguimiento
π_M	Proporción de individuos perdidos al final del seguimiento
κ	Ratio entre el coste de la medida inicial y el de cada una de las subsiguientes
c_1	Coste de la medida inicial (incluyendo el reclutamiento) por un participante
Presupuesto[†]	Presupuesto financiero total (reclutamiento y seguimiento)
Potencia[§]	Potencia requerida del estudio
$\alpha^{\S}$	Nivel de significación
$\tilde{\beta}^{\S}$	Tamaño del efecto bajo la hipótesis alternativa

[†] Si interesa maximizar la potencia del estudio sin exceder un presupuesto determinado.
[§] Si interesa minimizar el coste del estudio logrando una determinada potencia.

que llamaremos escenarios básicos. Estos escenarios básicos para los cuales podemos obtener una expresión analítica cerrada para r_{opt} (y, por tanto, para N_{opt}) se caracterizan por las asunciones indicadas en la Tabla 2.2.

Tabla 2.2: Asunciones bajo los escenarios básicos, necesarias para la existencia de una expresión analítica cerrada para el número óptimo de medidas repetidas, r_{opt}.

Asunción	Parametrización
Ausencia de datos faltantes	$\pi_M = 0$
La estructura de correlación de la respuesta es $\text{CS}(\sigma, \rho)$	$\theta = 0$
La prevalencia de la exposición es constante	$p_{e_j} = p_e,\ \forall j = 0, \ldots, r$
La estructura de covarianza de la exposición es CS[†]	Parámetro de correlación ρ_e

[†] Necesario solo si el patrón de la respuesta es LDD y la exposición varía en el tiempo.

Tales escenarios pueden resultar poco realistas en algunos estudios epidemiológicos observacionales. En efecto, es razonable esperar que, en general, la correlación entre dos medidas de la respuesta en un mismo individuo pueda decaer con la distancia temporal entre dichas medidas por lo que conviene poder considerar $\theta > 0$. Por otro lado, no es infrecuente que exista abandono del estudio por parte de algún participante, con lo que cabe prever que pueda ser $\pi_M > 0$. Además, el hecho de considerar una exposición no controlada por el investigador implica que en algunos estudios podría darse una prevalencia de la exposición variable en el tiempo. No existe una expresión cerrada para la solución del problema (2.3) bajo ningún escenario diferente al básico, por lo que, en tales casos, debemos recurrir a una resolución numérica. Para ello, una posible estrategia consiste en fijar un valor máximo admitido para el número de medidas repetidas, $r_{\text{máx}}$, y, de entre todos los valores naturales no superiores a $r_{\text{máx}}$, tomar como óptimo aquél que optimiza la función objetivo. Para

llevar a cabo este procedimiento, se creó y utilizó un paquete de R con el que, una vez introducidos los valores de los parámetros necesarios de la Tabla 2.1, se obtiene la combinación $(N_{\text{opt}}, r_{\text{opt}})$. Este paquete se describe en el capítulo 5.

Capítulo 3

Resultados

3.1. Escenarios básicos

Como hemos comentado anteriormente, bajo los escenarios básicos caracterizados por las asunciones descritas en la Tabla 2.2, existe una expresión analítica cerrada para r_{opt}. En efecto, para estos escenarios básicos, la expresión de $\tilde{\sigma}^2$ se muestra en la Tabla 3.1 (su derivación de muestra en el Apéndice A). Substituyendo las expresiones de $\tilde{\sigma}^2$ de la Tabla 3.1 en (2.3) y despreciando la constante multiplicativa K, obtenemos las correspondientes funciones de r a minimizar que contienen los parámetros ρ_e, ρ y κ. Así, en estos escenarios básicos, el valor de r_{opt} depende solo de la correlación intraclase de la exposición, de la correlación de la respuesta y del cociente entre el coste de la primera medida y el de cada una de las medidas subsiguientes.

Para el patrón de respuesta CMD en el escenario básico, la solución al problema (2.3) (derivada en el Apéndice C) es

$$r_{\text{opt}} = \begin{cases} 0\ , & \text{si } (\rho_e \geq \rho \text{ y } \kappa = 1) \text{ o } (\rho_e > \rho \text{ y } \kappa \leq \kappa_0) \\ r_c\ , & \text{si } \rho_e > \rho \text{ y } \kappa \in (\kappa_0, \kappa_c) \\ \infty\ , & \text{para otras combinaciones de } (\kappa, \rho, \rho_e) \end{cases} \tag{3.1}$$

donde

$$r_c := \frac{\kappa - 1 - (\kappa_c - 1)\rho + \sqrt{(1-\rho)(\kappa_c - 1)(\kappa - 1)[1 - \rho + \rho(\kappa_c - \kappa)]}}{\rho(\kappa_c - \kappa)}\ ,$$

$$\kappa_0 := \frac{1-\rho}{1 - \rho_e + (1-\rho)^2} \qquad \text{y} \qquad \kappa_c := \frac{\rho_e(1-\rho)}{\rho(1-\rho_e)}.$$

A efectos prácticos, el caso $r_{\text{opt}} = \infty$ se interpreta como que deberíamos tomar tantas medidas repetidas como fuese posible, en función de los aspectos logísticos del estudio.

Tabla 3.1: Expresiones de $\tilde{\sigma}^2$ bajo patrones de respuesta CMD y LDD en los escenarios básicos; es decir, asumiendo CS(σ, ρ) para la estructura de covarianza de la respuesta, ausencia de datos faltantes ($\pi_M = 0$) y prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 1, \ldots, r+1$). Los resultados para CMD son válidos para cualquier estructura de correlación de la exposición mientras que los resultados para LDD se han obtenido bajo la asunción de estructura CS para la correlación de la exposición.

Patrón de respuesta	$\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}$
CMD[†]	$\dfrac{K(\rho r+1)}{(r+1)[\rho(1-\rho_e)r+1-\rho]}, \quad K := \dfrac{\sigma^2(1-\rho)}{p_e(1-p_e)}$
LDD[§]	$\dfrac{12K(\rho r+1)\,r}{(r+1)\left\{\rho\rho_e r^2+[2\rho+\rho_e+3(1-\rho)\rho_e(1-\rho_e)]r+2[1+(1-\rho_e)(2-\rho)]\right\}}$
Caso particular de exposición invariante en el tiempo ($\rho_e = 1$):	
CMD	$\dfrac{K(\rho r+1)}{(1-\rho)(r+1)}$
LDD	$\dfrac{12Kr}{(r+1)(r+2)}$

[†] Para cualquier estructura de correlación de la exposición.
[§] Asumiendo CS para la estructura de correlación de la exposición.

En el caso particular de una exposición invariante en el tiempo ($\rho_e = 1$), la expresión (3.1) se reduce a

$$r_{\text{opt}} = \begin{cases} 0, & \text{si} \quad \kappa \leq \frac{1}{1-\rho} \\ \sqrt{\frac{1-\rho}{\rho}(\kappa-1)} - 1, & \text{si} \quad \kappa > \frac{1}{1-\rho} \end{cases} . \tag{3.2}$$

El resultado (3.2) fue dado, en el contexto particular de los ensayos por conglomerados aleatorizados, por Cochran [24] y Raudenbush [25]. La Figura 3.1 muestra la representación gráfica de la fórmula (3.2). Según esta expresión, lo óptimo bajo el patrón de respuesta CMD es tomar solo una medida (es decir, no realizar medidas repetidas) si κ no es mayor que el umbral $\frac{1}{1-\rho}$, que aumenta al hacerlo. En otro caso, lo óptimo es tomar un número de medidas repetidas que aumenta al hacerlo κ o al disminuir ρ. En general, en el caso de exposición invariante en el tiempo, valores de κ y ρ cercanos a 1 favorecen la realización de un estudio transversal mientras que un elevado coste relativo de la medida inicial y una baja correlación de la respuesta suponen factores a favor de tomar medidas repetidas. Cuando se generaliza la solución al contexto de una exposición variable en el tiempo ($\rho_e < 1$), la relación entre el número óptimo de medidas y los parámetros ρ_e, ρ y κ no es trivial teniendo en cuenta la complejidad de las expresiones en (3.1). La Figura 3.2 ilustra más claramente esta relación. La correlación intraclase de la exposición ρ_e afecta severamente

al número óptimo de medidas. Al comparar el valor de r_{opt} para una combinación dada de κ y ρ, y diferentes valores de ρ_e en la Figura 3.2, se puede apreciar fácilmente que, en la mayoría de casos, se recomienda más medidas repetidas a medida que aumenta la variación intra-individuo de la exposición (es decir, disminuye ρ_e). En muchas situaciones, lo óptimo es tomar tantas medidas repetidas como sea posible (matemáticamente, infinitas, correspondientes a aquellos puntos sin etiquetar en el gráfico), situación que deviene más común a medida que aumentan κ o ρ. Este comportamiento se acentúa al decrecer ρ_e, hasta el punto en que si $\rho_e = 0{,}1$ entonces $r_{\text{opt}} = \infty$ para cualquier combinación de $\kappa \geqslant 2$ y $\rho \geqslant 0{,}1$.

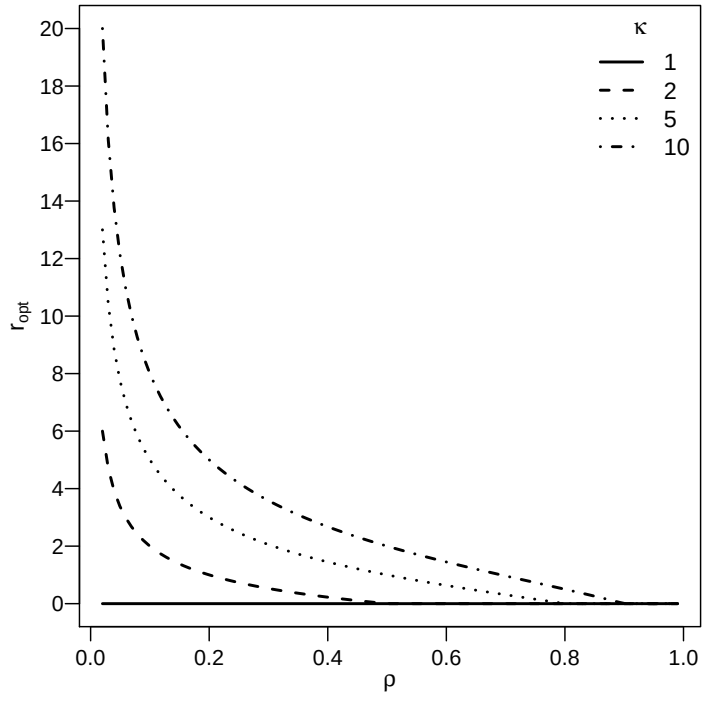


Figura 3.1: Representación gráfica de r_{opt} bajo el modelo CMD en el escenario básico y en el caso de exposición constante, correspondiente a la expresión (3.2).

Para el patrón de respuesta LDD en el escenario básico, la solución al problema (2.3) (derivada en el Apéndice D) es

$$r_{\text{opt}} = \begin{cases} 1\ , & \text{si } \kappa \leqslant \kappa^\star := 5 + \frac{6(1-\rho_e)[2+(1-\rho)\rho_e]}{(1+\rho)\rho_e} \\ \infty\ , & \text{si } \kappa > \kappa^\star \end{cases} . \tag{3.3}$$

El umbral $\kappa^\star$, que toma el valor 5 en el caso particular de una exposición inva-

riante en el tiempo, aumenta al disminuir ρ_e o ρ, tal como muestra la Figura 3.3, donde cada número en el área del gráfico representa el valor de $\kappa^\star$ para una combinación específica de ρ_e y ρ. Así, si κ no es mayor que 5, el número óptimo de medidas repetidas es 1 independientemente de los valores de ρ_e y ρ. En comparación con el caso particular de exposición invariante en el tiempo, cuando la exposición varía en el tiempo resulta más difícil justificar tomar más de una medida repetida, ya que la ratio de costes necesaria para decidirlo, $\kappa^\star$, es más elevada.

Las expresiones (3.1) y (D.12) proporcionan, en general, un valor de r_{opt} no entero. El valor efectivo de r_{opt} es aquel que, en combinación con la aproximación entera del valor para N_{opt} dado por (2.14) o (2.15), proporciona la máxima potencia o el mínimo coste, respectivamente. El paquete de R que proporcionamos, descrito en el capítulo 5, realiza esta corrección.

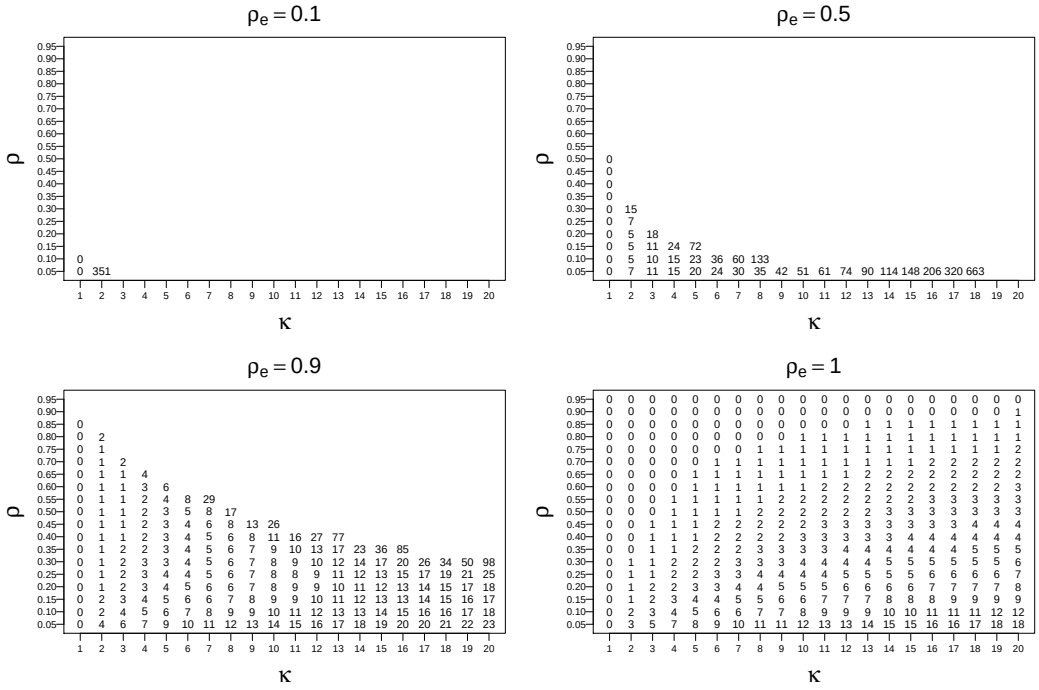


Figura 3.2: Cada número en el área del gráfico indica el número óptimo de medidas repetidas, r_{opt}, bajo el patrón de respuesta CMD en el escenario básico, que asume estructura de covarianza CS(σ, ρ) para la respuesta ($\theta = 0$), ausencia de datos faltantes ($\pi_M = 0$) y prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 0, \ldots, r$). Los puntos sin etiqueta corresponden a aquellos casos en que debemos tomar tantas medidas como sea posible (matemáticamente, infinitas). Los resultados que se muestran corresponden a valores de la ratio entre el coste económico de la primera medida y una de las subsiguientes, $\kappa = 1, 2, \ldots, 20$, valores de $\rho = 0{,}05, 0{,}10, \ldots 0{,}95$ y valores de la correlación intraclase $\rho_e = 0{,}1; 0{,}5; 0{,}9$ así como el caso particular de exposición invariante en el tiempo, $\rho_e = 1$.

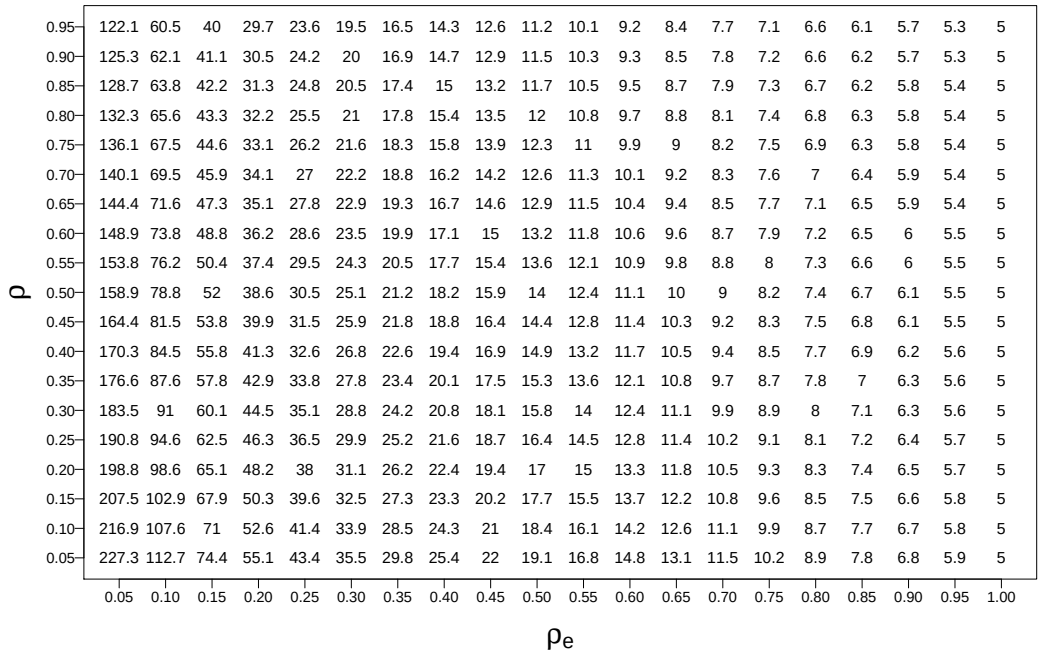


Figura 3.3: Umbral para la ratio de costes de la primera medida respecto a una de las subsiguientes ($\kappa^\star$) a partir del cual conviene tomar tantas medidas como sea posible bajo el patrón de respuesta LDD en el escenario básico. Por otro lado, si la ratio de costes es menor que $\kappa^\star$, lo óptimo es tomar una única medida repetida ($r_{\text{opt}} = 1$). El escenario básico para el patrón de respuesta LDD asume una estructura de covarianza $\text{CS}(\sigma, \rho)$ para la respuesta ($\theta = 0$), ausencia de datos faltantes ($\pi_M = 0$), prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 0, \ldots, r$) y estructura de correlación de la exposición $\text{CS}(\rho_e)$. La expresión para el umbral es $\kappa^\star := 5 + \frac{6(1-\rho_e)[2+(1-\rho)\rho_e]}{(1+\rho)\rho_e}$. La columna con valor constante de $\kappa^\star = 5$ corresponde a una exposición constante ($\rho_e = 1$).

3.2. Desviaciones de los escenarios básicos

Si no puede asumirse alguna de las condiciones descritas en la Tabla 2.2, el valor de r_{opt} debe determinarse numéricamente. En tal caso, dado el elevado número de parámetros que intervienen en el problema (Tabla 2.1), resulta difícil evaluar cómo estos afectan a su solución. Por este motivo, se optó por evaluar las desviaciones en los resultados tomando como partida los escenarios básicos y variando, de uno en uno, los parámetros adicionales que intervienen en los escenarios no básicos, análisis que se describe a continuación. En todo caso, el paquete R proporcionado permite comparar escenarios que difieren en los valores de más de un parámetro.

3.2.1. Efecto de variar la estructura de covarianza de la respuesta

Los resultados bajo los escenarios básicos en la sección 3.1 fueron obtenidos asumiendo una estructura de covarianza de la respuesta $\text{CS}(\sigma,\ \rho)$, que corresponde al

caso particular de una estructura de covarianza DEX con parámetro de amortiguamiento $\theta = 0$. En esta sección, examinamos cómo cambian los resultados para $\theta > 0$, en cuyo caso el problema de optimización debe resolverse numéricamente. Puede probarse fácilmente que r_{opt} no depende del valor de σ, de manera que puede evaluarse la función objetivo para una combinación específica de los parámetros θ, ρ_e, ρ y κ y un rango de valores $r = 0, 1, \ldots, r_{\text{máx}}$, y entonces encontrar el mínimo en (2.13) condicionado a $r_{\text{opt}} \leqslant r_{\text{máx}}$. El valor de $r_{\text{máx}}$ fue fijado en 20 y 30 para los patrones de respuesta CMD y LDD, respectivamente. El valor máximo explorado de κ fue 20.

Para el patrón de respuesta CMD, al variar θ desde 0 (CS) hasta 1 (AR(1)), si la exposición es constante ($\rho_e = 1$), las diferencias en los resultados para r_{opt} aperecen solo para valores pequeños de ρ o costes de la primera medida elevados ($\kappa \geqslant 10$). Estas diferencias no fueron mayores a 3 unidades. Por ejemplo, para estructura de covarianza de la respuesta AR(1), el número óptimo de medidas repetidas varía solo entre 0 y 2 para $\kappa \leqslant 20$ mientras que para una estructura de covarianza de la respuesta CS, varía entre 0 y 4. Si la exposición es variable en el tiempo ($\rho_e < 1$), el efecto de θ sobre r_{opt} depende complejamente de la combinación de los valores de los parámetros ρ_e, ρ y κ.

Para el patrón de respuesta LDD, la exploración, limitada a $r_{\text{máx}} = 30$ y $\kappa \leqslant 20$, mostró que valores positivos de θ rompen la dicotomía $1 - \infty$ que se da en los resultados para el número óptimo de medidas repetidas bajo el escenario básico (CS, $\theta = 0$). A pesar de ello, el valor 1 fue obtenido para la mayoría de combinaciones de valores de los parámetros θ, ρ_e, ρ y κ, pero se obtuvieron valores mayores en una pequeña proporción de casos. Concretamente, si $\theta \geqslant 0{,}5$, el número óptimo de medidas repetidas fue 1 independientemente de los valores de los parámetros restantes. El número óptimo de medidas repetidas también fue 1 para todo $\kappa \leqslant 9$. En general, aumentar ρ_e o κ, o disminuir ρ tiende a aumentar el valor de r_{opt}. A modo de ilustración, la Tabla 3.2 muestra el valor mínimo aproximado de κ para el cual el número óptimo de medidas repetidas es mayor que 1 así como el correspondiente valor máximo del número óptimo de medidas repetidas obtenido en la exploración dentro del intervalo $\rho \in (0{,}05, 0{,}95)$. El valor máximo de r_{opt} fue alcanzado sistemáticamente para el valor máximo explorado de κ y para el valor mínimo explorado de ρ.

3.2.2. Efecto de la pérdida de participantes durante el seguimiento

En esta sección analizamos el efecto potencial de posibles datos faltantes debidos a una pérdida por abandono de participantes a lo largo del seguimiento, respecto a los resultados obtenidos bajo los escenarios básicos. Para hacerlo, hemos explorado una parrilla de valores para la proporción de participantes perdidos al final del estudio,

Tabla 3.2: Efecto de θ en r_{opt} bajo el patrón de respuesta LDD asumiendo ausencia de datos faltantes, prevalencia de la exposición constante, estructura de covarianza de la exposición CS(ρ_e) y estructura de covarianza de la respuesta DEX(θ, ρ). La exploración fue restringida a $\kappa \leqslant 20$ y $\rho \in (0{,}05, 0{,}95)$. Cada par representa (máx(r_{opt}), $\kappa_{\text{mín}}$), donde $\kappa_{\text{mín}}$ es una aproximación del valor mínimo de κ para el cual r_{opt} es mayor que 1 (a lo largo del rango de ρ) y máx(r_{opt}) es el valor máximo de r_{opt}, que fue alcanzado siempre para el máximo valor explorado de κ (20) y el mínimo valor explorado de ρ (0,05).

	Correlación intraclase de la exposición (ρ_e)						
θ	0.1	0.5	0.6	0.7	0.8	0.9	1
0 (CS)	$(\infty, 60{,}5)$	$(\infty, 11{,}2)$	$(\infty, 9{,}2)$	$(\infty, 7{,}7)$	$(\infty, 6{,}6)$	$(\infty, 5{,}7)$	$(\infty, 5)$
0.10	$*$	$*$	(18, 19)	(21, 14)	(23, 12)	(26, 10)	(30, 8)
0.20	$*$	$*$	$*$	$*$	(14, 16)	(15, 13)	(17, 10)
0.25	$*$	$*$	$*$	$*$	(11, 19)	(12, 15)	(14, 11)
0.30	$*$	$*$	$*$	$*$	$*$	(10, 17)	(12, 13)
0.40	$*$	$*$	$*$	$*$	$*$	$*$	(8, 16)
$\geqslant 0{,}50$	$*$	$*$	$*$	$*$	$*$	$*$	$*$

$* r_{\text{opt}} = 1$ para cualquier combinación de $\kappa \leqslant 20$, $\rho_e \geqslant 0{,}1$ y $\rho \in (0{,}05, 0{,}95)$.

π_M, entre 0 (que corresponde a los escenarios básicos) y 0,65. La exploración fue restringida a $r \leqslant 30$ con θ fijado en 0 (estructura de covarianza de la respuesta CS).

Para ambos patrones de respuesta CMD y LDD, los resultados para r_{opt} fueron los mismos que los obtenidos cuando $\pi_M = 0$, excepto para un pequeño porcentaje de combinaciones de los valores de ρ_e, κ y ρ. Concretamente, para el patrón de respuesta CMD, incrementos de más de 3 unidades en r_{opt} fueron observados en solo el 5 % de los escenarios explorados, correspondientes a valores elevados de π_M ($\geqslant 0{,}5$). En unos pocos escenarios, caracterizados por un valor elevado de r_{opt} (de 24 a 30) y $\kappa = 1$, se detectó un ligero decremento de r_{opt} (de 1 a 3 unidades) para valores elevados de π_M (0,6) Para el patrón de respuesta LDD, un aumento de π_M favorece $r_{\text{opt}} = 1$ para valores elevados de π_M en el caso de una exposición constante mientras que favorece un aumento de r_{opt} para valores pequeños de ρ y elevados de κ en el caso de una exposición variable en el tiempo. Además, apenas de observaron cambios cuando θ se fijó en 1 (estructura de covarianza de la respuesta AR(1)).

Por otro lado, y teniendo en cuenta (2.11), debe tenerse presente que la presencia de datos faltantes podría afectar a N_{opt} incluso en aquellos casos en que no se observasen cambios en r_{opt}.

3.2.3. Efecto de una prevalencia de la exposición variable en el tiempo

Para explorar el efecto de una prevalencia de la exposición variable en el tiempo hemos considerado una posible variación lineal de dicha prevalencia según (2.8). Hemos considerado una parrilla de valores para la prevalencia media de la exposición, $\bar{p}_e$, entre 0,05 y 0,95 y para el cambio relativo en dicha prevalencia, γ, entre $-0{,}95$

(es decir, $p_{er} = p_{e0}/20$) y 20 (es decir, $p_{er} = 21p_{e0}$), para una tendencia lineal de la prevalencia. La exploración fue restringida a $r \leqslant 30$, $\theta = 0$ y $\kappa \leqslant 20$.

Para el patrón de respuesta CMD, la exploración reveló que un incremento de la diferencia de la prevalencia de la exposición entre el principio y el final del seguimiento (es decir, un aumento de $|\gamma|$) tiende a incrementar el valor de r_{opt}. Estos cambios, prácticamente despreciables en la mayoría de casos, se tornan más importantes para cambios extremos en la prevalencia de la exposición y valores elevados de la prevalencia media, tal como se ilustra en la Tabla 3.3.

Tabla 3.3: Efecto de una prevalencia de la exposición variable en el tiempo en r_{opt} para el patrón de respuesta CMD asumiendo estructura de covarianza de la respuesta CS(ρ), ausencia de datos faltantes y estructura de covarianza de la exposición CS(ρ_e). La exploración fue realizada para todas las combinaciones de los valores de los parámetros $\rho_e = 0{,}3, 0{,}8$, $\rho = 0{,}2, 0{,}8$, $\kappa = 1, 2, 5, 8$, $\bar{p}_e = 0{,}1, 0{,}3, 0{,}5$ y $p_{er}/p_{e0} = 1/20, 1, 21$ (es decir, $\gamma = -0{,}95, 0, 20$, respectivamente). Los números en el cuerpo de la tabla corresponden a r_{opt}.

		ρ_e :	0,3			0,8					
		ρ :	0,2		0.8	0,2				0,8	
p_{er}/p_0	$\bar{p}_e$	κ :	1	$\geqslant 2$	$\geqslant 1$	1	2	5	8	1	$\geqslant 2$
	0,1		0	30	30	0	0	7	17	1	30
1/20	0,3		0	30	30	0	0	8	16	0	30
	0,5		30	30	30	4	5	10	19	6	30
1	$\star$		0	30	30	0	2	6	11	0	30
	0,1		2	30	30	1	2	7	17	1	30
21	0,3		8	30	30	2	3	8	16	2	30
	0,5		30	30	30	4	5	10	19	6	30

$\star$: r_{opt} no depende de $\bar{p}_e$ si la prevalencia de la exposición es constante.

Para el patrón de respuesta LDD, una prevalencia de la exposición variable en el tiempo rompe la dicotomía $1 - \infty$ para r_{opt}. Como en el caso del patrón de respuesta CMD, solo aparecieron cambios para grandes diferencias en la prevalencia de la exposición entre la primera medida y la última. Para prevalencias de la exposición fuertemente crecientes en el tiempo, casi no se detectaron cambios para valores de $\kappa \leqslant 5$ mientras que para $\kappa \geqslant 8$ y prevalencia media no superior a 0,3, el efecto es cambiar r_{opt} de ∞ a 1. Para prevalencias de la exposición fuertemente decrecientes en el tiempo, r_{opt} esencialmente tiende a aumentar unas pocas unidades (de 1 a 4) en aquellos casos en que r_{opt} es 1 bajo una prevalencia de la exposición constante, como se muestra en la Tabla 3.4.

Tabla 3.4: Efecto de una prevalencia de la exposición variable en el tiempo r_{opt} para el patrón de respuesta LDD asumiendo estructura de covarianza de la respuesta CS(ρ), ausencia de datos faltantes y estructura de covarianza de la exposición CS(ρ_e). La exploración fue realizada para todas las combinaciones de los valores de los parámetros $\rho_e = 0{,}3, 0{,}8$, $\rho = 0{,}2, 0{,}8$, $\kappa = 1, 2, 5, 8$, $\bar{p}_e = 0{,}1, 0{,}3, 0{,}5$ y $p_{er}/p_{e0} = 1/20, 1, 21$ (es decir, $\gamma = -0{,}95, 0, 20$, respectivamente). Los números en el cuerpo de la tabla corresponden a r_{opt}.

		$\rho_e = 0{,}3$						$\rho_e = 0{,}8$							
		$\rho = 0{,}2$			$\rho = 0{,}8$			$\rho = 0{,}2$				$\rho = 0{,}8$			
p_{er}/p_0	$\bar{p}_e$	$\kappa \leqslant 2$	$\kappa = 5$	$\kappa = 8$	$\kappa \leqslant 2$	$\kappa = 5$	$\kappa = 8$	$\kappa = 1$	$\kappa = 2$	$\kappa = 5$	$\kappa = 8$	$\kappa = 1$	$\kappa = 2$	$\kappa = 5$	$\kappa = 8$
1/20	0,1	2	← – 3 – →		2	3	4	3	4	← – 30 – →		3	5	← – 30 – →	
	0,3	← – 2 – →		3	2	← – 3 – →		2	3	← – 30 – →		3	4	← – 30 – →	
	0,5	← – – – 2 – – – →			← – – – 2 – – – →			← – 2 – →		26	30	2	3	← – 30 – →	
1	⋆	← – 1 – →													30
21	0,1	← – 1 – →													
	0,3	← – 1 – →													
	0,5	← – – – – – – – – – – – – – – – – – 2 – – – – – – – – – – – – – – – – →						← – 30 – →				2	3	← – 30 – →	

⋆: r_{opt} no depende de $\bar{p}_e$ si la prevalencia de la exposición es constante.

Capítulo 4

Ejemplo ilustrativo

En este capítulo aplicamos la metodología desarrollada a los datos de un estudio sobre salud respiratoria en un conjunto de trabajadoras de la limpieza para proporcionar un diseño óptimo para un hipotético nuevo estudio sobre el mismo tema. En el estudio basado en estos datos, Medina-Ramón *et al.* [34] testaron la hipótesis de que la exposición a productos de limpieza irritantes puede agravar una enfermedad de obstrucción pulmonar preexistente en trabajadoras de la limpieza doméstica. Para ello, se evaluó un efecto agudo y transitorio de la exposición transitoria a productos y tareas de limpieza sobre el flujo espiratorio máximo, PEF (*peak expiratory flow*), medido en L/min (litros por minuto), sobre un grupo de $N = 43$ limpiadoras domésticas (de 31 a 66 años de edad) que presentaban una historia reciente de enfermedad por obstrucción pulmonar (asma y/o bronquitis crónica). El seguimiento se llevó a cabo durante 15 días tomando medidas diariamente (es decir, $r = 14$). Al final de la jornada correspondiente a cada uno de esos días, a las trabajadoras se les medía su función pulmonar mediante espirometría. Además, las trabajadoras anotaron diariamente si habían utilizado determinados productos de limpieza y si habían realizado determinadas tareas relacionadas con la limpieza. El estudio fue observacional y, por consiguiente, las exposiciones (a productos y tareas de limpieza) no fueron asignadas por diseño sino que las limpiadoras llevaron a cabo sus tareas de limpieza y usaron los productos que su trabajo diario requería, mostrando todas estas exposiciones variaciones diarias intra sujeto. En este ejemplo, nos centraremos en las dos exposiciones que mostraron el mayor y el menor valor de ρ_e, que fueron los usos de aspirador y de aerosoles ambientadores, respectivamente. El uso de aspirador mostró $\rho_e = 0{,}13$ con una prevalencia media $\bar{p}_e = 0{,}37$ mientras que el uso de aerosoles ambientadores mostró $\rho_e = 0{,}60$ y $\bar{p}_e = 0{,}17$. La Figura 4.2 muestra la distribución del número de días bajo exposición por trabajadora, para ambas exposiciones. Por otro lado, y como cabía esperar en este caso y tal como muestra la Figura 4.1, la prevalencia de la exposición no mostró tendencia temporal en ninguna de las dos exposiciones de

manera que es razonable asumir una prevalencia de exposición constante. En cuanto a la pérdida de trabajadoras en el seguimiento, treinta y una de las participantes en el estudio proporcionaron información completa al final del seguimiento con lo que fijamos $\pi_M = 0{,}28$. La varianza residual y el parámetro de amortiguamiento de la covarianza de la respuesta se tomaron del estudio, fijando $\sigma^2 = 0{,}43$ y $\theta = 0{,}12$. Tomamos un valor bajo (0,3) y otro más elevado (0,7) para ρ. Por lo que se refiere al efecto hipotético a detectar, se fijó en una diferencia del 10 % en el valor medio esperado de la respuesta entre trabajadoras expuestas y no expuestas, asumiendo un efecto agudo y transitorio (es decir, un patrón de respuesta CMD) que resultó en el valor $\tilde{\beta} = -0{,}39$ L/min. El objetivo del hipotético nuevo estudio era minimizar su coste total, fijando una potencia requerida de, al menos, el 90 %. Se asumió que el coste de la primera medida, fijado como unidad monetaria, era el doble del de cada una de las medidas subsiguientes (es decir, $\kappa = 2$). Al tener que resolver el problema de diseño óptimo numéricamente, restringimos el número máximo de medidas repetidas a 20. Todos los cálculos fueron realizados fijando un nivel de significación $\alpha = 0{,}05$.

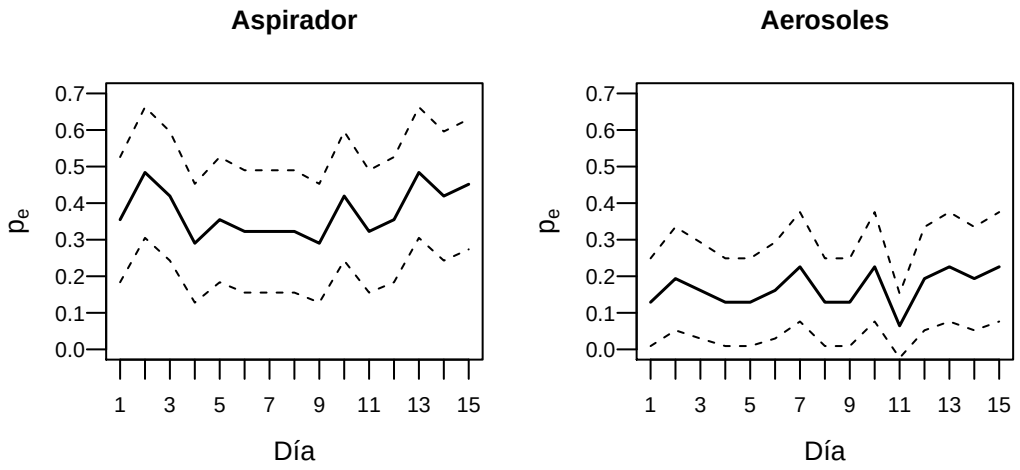


Figura 4.1: Evolución temporal de la prevalencia de la exposición, para uso de aspirador y de aerosoles. La línea sólida indica la prevalencia diaria de la exposición y las líneas discontinuas indican los límites de su intervalo de confianza al 95 %.

Los resultados, mostrados en la Tabla 4.1, revelaron una notable discrepancia tanto en el número óptimo de medidas repetidas como en el número óptimo de participantes entre las asunciones de exposición constante (es decir, asumiendo $\rho_e = 1$) y una exposición variable en el tiempo (es decir, utilizando el valor observado de ρ_e). Así, usando el valor observado de ρ_e, el diseño óptimo es tomar un elevado número de medidas tanto para el uso de aspirador (entre 15 y 18) como para el uso

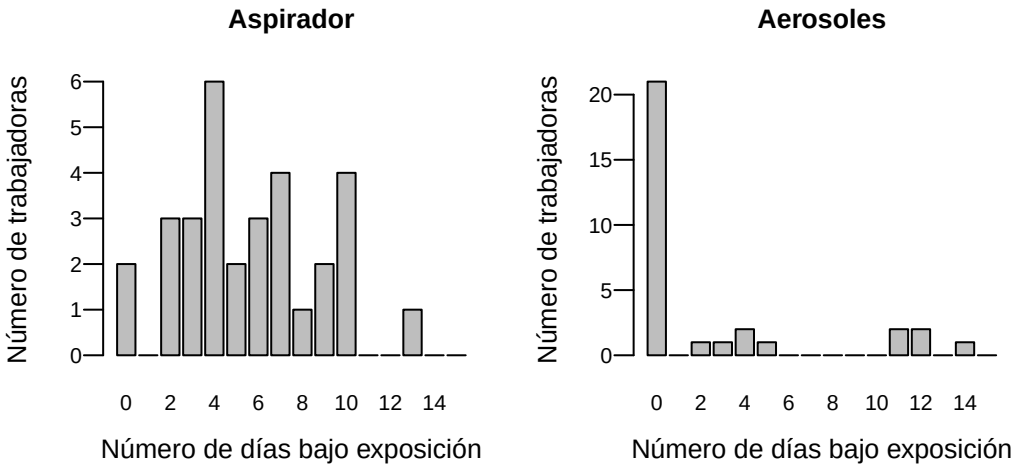


Figura 4.2: Distribución del número de días bajo exposición por individuo, para uso de aspirador y de aerosoles. La línea sólida indica la prevalencia diaria de la exposición y las líneas discontinuas indican los límites de su intervalo de confianza al 95 %.

de aerosoles (entre 19 y 20) mientras que asumiendo $\rho_e = 1$, lo óptimo sería realizar un estudio transversal (si $\rho = 0{,}7$) o un estudio longitudinal con solo dos medidas (si $\rho = 0{,}3$). Por tanto, el uso incorrecto de las fórmulas para el caso en que $\rho_e = 1$, cuando la exposición es en realidad variable en el tiempo, puede llevar a discrepancias no solo en r_{opt} y N_{opt} sino también en el coste final del estudio. Así, diseñar el estudio a partir de las fórmulas para el caso de exposición constante lleva a resultados con un incremento innecesario del coste de entre el 140 % y el 480 % para el uso de aspirador, y de entre el 30 % y el 190 %, para el uso de aerosoles, respecto al coste óptimo obtenido al usar las fórmulas correctas. Estos resultados apuntan a la importancia de considerar las fórmulas que hemos derivado en este trabajo cuando existe variabilidad intra-individuo en la exposición en lugar de utilizar incorrectamente las fórmulas existentes basadas en una exposición constante.

Para explorar cómo afecta r al resultado del diseño, se ha creado la Figura 4.3, que muestra como, para valores elevados de r, el investigador puede reducir (incrementar) el número de medidas repetidas por participante, a cambio de incrementar (reducir) el número de participantes en el estudio, sin un incremento significativo en su coste.

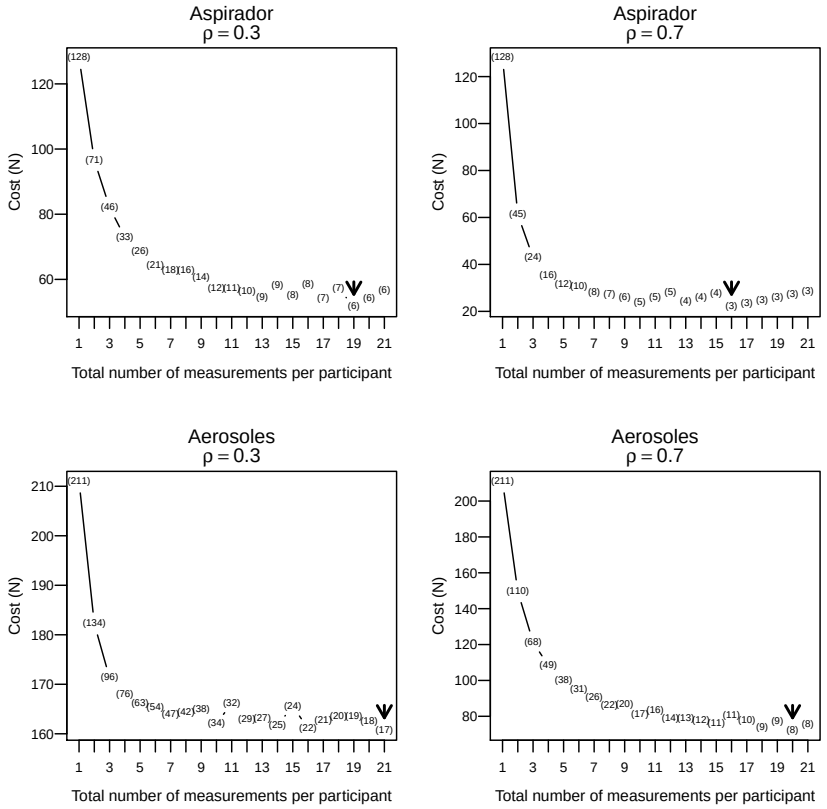


Figura 4.3: Coste minimizado y número de participantes (entre paréntesis) en función del número total de medidas por participante. Las flechas apuntan al diseño óptimo.

Tabla 4.1: Combinación óptima del número de medidas repetidas y el número de participantes para alcanzar una potencia de, al menos, el 90 %, y correspondiente coste mínimo necesario. El problema fue restringido a $r_{\text{opt}} \leq 20$. Basados en los datos del ejemplo, se utilizaron los valores siguientes de los parámetros. Para uso del aspirador, $\rho_e = 0{,}13$ y $\bar{p}_e = 0{,}37$, y para uso de aerosoles ambientadores $\rho_e = 0{,}60$ y $\bar{p}_e = 0{,}17$. Se asumió una prevalencia de la exposición constante. Para todos los cálculos, $\sigma^2 = 0{,}43$, $\theta = 0{,}12$, $\pi_M = 0{,}28$, $c_1 = 1$, $\kappa = 2$, $\alpha = 0{,}05$. Los cálculos fueron realizados para detectar una diferencia del 10 % entre expuestas y no expuestas bajo el patrón de respuesta CMD.

Exposición	**Correlación de la respuesta, ρ**	**Covarianza de la exposición**	**Óptimo** r_{opt}	N_{opt}	**Coste**[†]
Aspirador	0,3	CS(ρ_e)	18	6	51,6
		t.i.e.[§]	1	92	125,1
	0,7	CS(ρ_e)	15	3	22,0
		t.i.e.[§]	0	128	128,0
Aerosoles	0,3	CS(ρ_e)	20	17	160,7
		t.i.e.[§]	1	152	206,7
	0,7	CS(ρ_e)	19	8	72,2
		t.i.e.[§]	0	211	211,0

[†] Tomando el coste de la primera medida como unidad monetaria.
[§] Usando las fórmulas que asumen exposición constante (i.e. $\rho_e = 1$).

Capítulo 5

Implementación de la metodología en R

En este capítulo ilustramos el uso del paquete optimalAllocation de R con algunos ejemplos, incluyendo la reproducción de los resultados mostrados en el capítulo 4. El paquete puede descargarse en `http://www.mat.uab.cat/~jbarrera/Software.html`.

5.1. Primeros pasos

Comenzamos la sesión de R cargando el paquete:

```
> library(optimalAllocation)
```

El contenido de este capítulo es parte de la viñeta (en inglés) del paquete, a la que puede accederse mediante:

```
> vignette("optimalAllocation")
```

5.1.1. Función `OA()`

La función principal del paquete es `OA()` y realiza los cálculos para el diseño del estudio. Podemos obtener información sobre esta función:

```
> ?OA
```

Los parámetros de entrada para la función `OA()` son:

- `target`: el objetivo del diseño. Puede tomar dos valores: `"maxPower"`, si se desea maximizar la potencia del estudio, o `"minCost"`, si se desea minimizar su coste total.

- `pattern`: el patrón de respuesta asumido bajo la hipótesis alternativa. Puede tomar dos valores: `"CMD"`, si se asume un efecto agudo y transitorio de la exposición, o `"LDD"`, si se asume un efecto acumulativo de la exposición; es decir, una interacción exposición-tiempo.

- `rMax`: el valor máximo a evaluar para el número de medidas repetidas. Debe tenerse en cuenta que el coste computacional aumenta exponencialmente con `rMax`.

- `theta`: el parámetro de amortiguamiento, θ, en la estructura DEX para la matriz de covarianzas de la respuesta. La estructura es CS si `theta` $= 0$ y AR(1) si `theta` $= 1$.

- `rho`: la correlación en la respuesta entre la primera medida y la correspondiente al final del seguimiento, ρ.

- `sigma2`: la varianza residual de la respuesta, σ^2.

- `rhoe`: la correlación intraclase de la exposición, ρ_e, definida según (2.9). Para decidir el valor de ρ_e en la fase del diseño del estudio, puede resultar útil explorar la distribución del número de períodos bajo exposición por participante, una vez fijados los valores de ρ_e, la prevalencia de la exposición, p_e, asumida constante, y el número de medidas repetidas, r, y asumiendo estructura CS para la covarianza de la exposición, $\mathbf{\Sigma}_E$. A tal efecto, el paquete incluye las funciones `plotExposedPeriods()` y `plotExposedPeriodsInt()` (sección 5.1.2).

- `pe0`: la prevalencia de la exposición en la primera medida.

- `per`: la prevalencia de la exposición al final del seguimiento. Se asume que la prevalencia de la exposición varía linealmente desde `pe0` hasta `per`. Si `per` es igual a `pe0`, se asume que la prevalencia de la exposición es constante.

- `piM`: la fracción de individuos que se ha perdido durante el seguimiento. El patrón de pérdida se asume monótono; es decir, la pérdida de un individuo implica la pérdida de todas sus medidas subsiguientes. Se asume que no hay datos faltantes en la primera medida.

- `kappa`: el cociente entre el coste de la primera medida (incluyendo el reclutamiento) y el de cada una de las medidas subsiguientes, κ.

- `budget`: el presupuesto total disponible para el estudio, en el caso en que el objetivo del diseño sea maximizar la potencia (`target = "maxPower"`).

- `c1`: el coste de la primera medida por participante, incluyendo el reclutamiento.

- `reqPower`: la potencia requerida del estudio, en el caso en que el objetivo del diseño sea minimizar el coste (`target = "minCost"`).
- `beta`: la medida del efecto esperado bajo la hipótesis alternativa, $\tilde{\beta}$. Si `pattern = "CMD"`, `beta` puede interpretarse como la diferencia esperada en la media de la variable respuesta, en cualquier punto temporal del estudio, entre los expuestos y los no expuestos. Si `pattern = "LDD"`, `beta` puede interpretarse como la diferencia esperada en la media de la variable respuesta, al final del seguimiento, entre el peor patrón de exposición (es decir, estar expuesto en todas las medidas) y los no expuestos.
- `alpha`: el nivel de significación, α.

5.1.2. Funciones `plotExposedPeriods()` y `plotExposedPeriodsInt()`

La función `plotExposedPeriods()` regresenta gráficamente la distribución del número de períodos bajo exposición por participante, una vez fijados los valores de ρ_e, la prevalencia de la exposición, p_e, asumida constante, y el número de medidas repetidas, r, y asumiendo estructura CS para $\mathbf{\Sigma}_E$ [41]. Tal distribución es equivalente a la distribución de la suma de $r+1$ variables Bernoulli(p_e) no independientes con correlación ρ_e entre cada posible pareja de ellas. El gráfico obtenido es una herramienta para decidir el valor de la correlación intraclase de la exposición.

Podemos obtener información sobre la función `plotExposedPeriods()`:

```
> ?plotExposedPeriods
```

Los parámetros de entrada para la función `plotExposedPeriods()` son:

- `r`: el número de medidas repetidas, r; es decir, el número total de medidas por participante es `r` + 1.
- `pe`: la prevalencia de la exposición, asumida constante.
- `rhoe`: la correlación intraclase de la exposición.
- `eps`: la precisión en los resultados (error relativo). El valor por defecto es 0,001.
- `maxIter`: el número máximo de iteraciones para el cálculo de la distribución. El valor por defecto es 1.000.

Por ejemplo, fijando el número de medidas repetidas en 3 y asumiendo una prevalencia de exposición constante de 0,2, podemos explorar la distribución del número de períodos expuestos por participante para diversos valores de la correlación intraclase de la exposición con el código:

```
> rhoes <- c(-0.2, 0, 0.5, 0.9)
> par(las = 1, mfrow = c(2, 2))
> for (i in 1:4)
+  plotExposedPeriods(r = 3, pe = 0.2, rhoe = rhoes[i])
```

que proporciona la Figura 5.1. Así, el valor de ρ_e puede fijarse en aquel valor que proporciona una distribución razonable del número de períodos expuestos por participante en la población bajo estudio.

La función `plotExposedPeriodsInt()` es una versión interactiva de la función `plotExposedPeriods()` que actualiza dinámicamente el gráfico de la distribución del número de períodos expuestos cuando el usuario modifica el valor de ρ_e mediante una barra de deslizamiento.

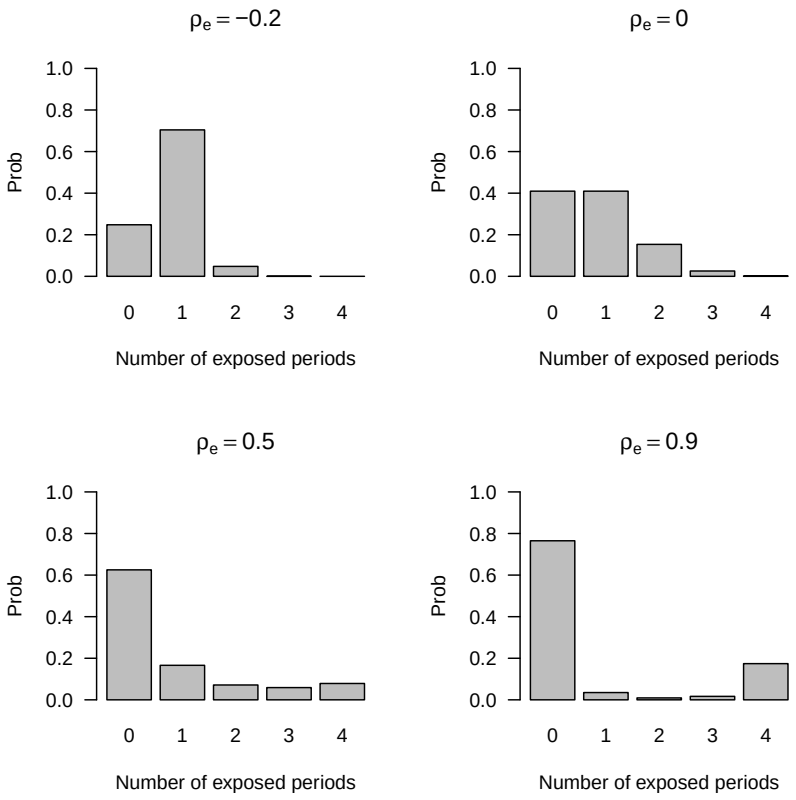


Figura 5.1: Distribución del número de períodos expuestos por participante para diversos valores de la correlación intraclase de la exposición. El número de medidas repetidas se fijó en 3 y se asumió una prevalencia de la exposición constante e igual a 0,2. Se asumió una estructura CS para la covarianza de la exposición.

5.2. Ejemplos de diseño de un estudio longitudinal

5.2.1. Estudio 1. Maximización de la potencia

Supongamos que estamos interesados en maximizar la potencia de un estudio longitudinal, asumiendo el patrón de respuesta CMD, sin exceder un presupuesto de 40 unidades monetarias, siendo la unidad monetaria el coste de la primera medida por participante. El coste de la primera medida es $\kappa = 3$ veces el coste de cada una de las subsiguientes. La estructura de covarianza de la respuesta es DEX($\sigma = 1, \rho = 0{,}7, \theta = 0{,}5$). La correlación intraclase de la exposición es $\rho_e = 0{,}2$. Se espera que uno de cada cinco participantes abandonará el estudio antes de concluirlo; es decir, $\pi_M = 0{,}2$. Se asume que la prevalencia de la exposición aumenta linealmente desde $p_{e0} = 0{,}2$ en la primera medida hasta $p_{er} = 0{,}3$ al final del seguimiento. El tamaño del efecto que se desea detectar es $\tilde{\beta} = -0{,}3$ y el nivel se significación se fija en $\alpha = 0{,}05$. El número máximo de medidas repetidas permitido es $r_{\max} = 20$. Así, podemos realizar los cálculos para el estudio y almacenar los resultados en el objeto `study1`:

```
> study1 <- OA(target = "maxPower", pattern = "CMD", rMax = 20,
+              theta = 0.5, rho = 0.7, sigma2 = 1, rhoe = 0.2,
+              pe0 = 0.2, per = 0.3, piM = 0.2, kappa = 3,
+              budget = 40, c1 = 1, beta = -0.3, alpha = 0.05)
> study1

  Results subject to r not greater than 20:
-----------------------------------------------------
Optimal total number of measurements (r+1): 20
Optimal number of participants (N)        : 6
Maximized power                           : 0.9670238
```

Así, condicionado a un máximo de 20 medidas repetidas, lo óptimo es realizar un estudio longitudinal con $N_{\text{opt}} = 6$ participantes y tomando un total de $r_{\text{opt}} + 1 = 20$ medidas por participante. La potencia maximizada de tal estudio es del 97 %.

Con la función `plot()` puede obtenerse una representación gráfica de los resultados. Concretamente,

```
> plot(study1)
```

genera la Figura 5.2, que muestra que la estrategia óptima es tomar tantas medidas como sea posible, así como las consecuencias, en cuanto a número de participantes y potencia, derivadas de variar el número de medidas repetidas desde su valor óptimo.

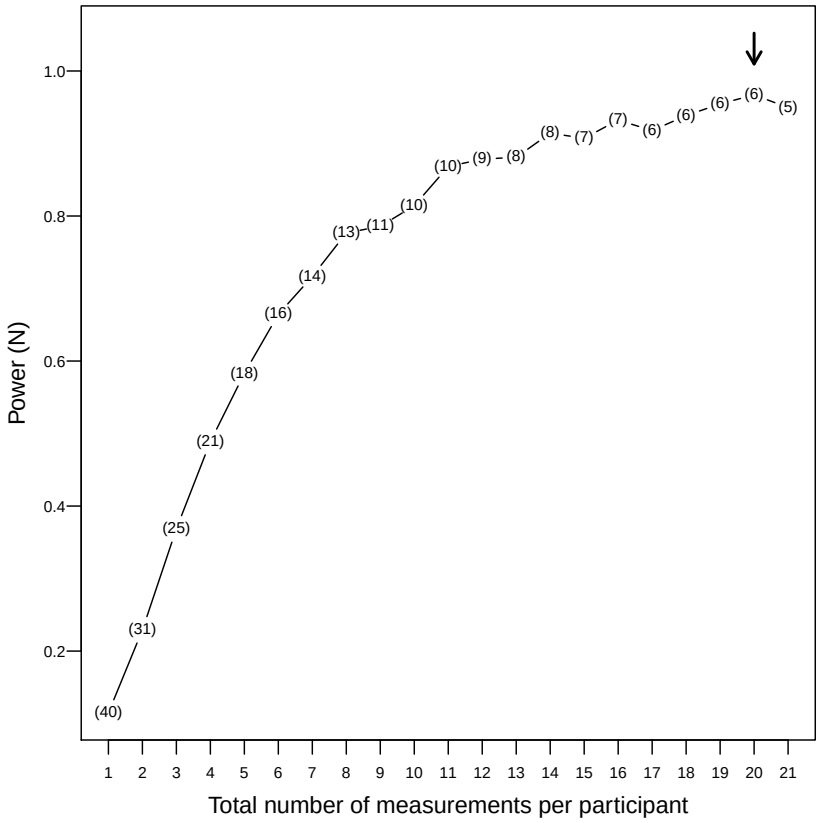


Figura 5.2: Potencia maximizada, y número de participantes (entre paréntesis) en función del número total de medidas por participante. La flecha apunta al diseño óptimo.

Con la función `summary()` se obtienen resultados adicionales asociados al diseño óptimo, incluyendo la estimación del error estándar de $\tilde{\beta}$, `sdBeta` y el presupuesto necesario, `cost`:

```
> summary(study1)

$roptreal
[1] 20

$ropt
[1] 19

$Nopt
[1] 6

$maxPower
```

```
[1] 0.9670238

$cost
[1] 39.85918

$sdBeta
[1] 0.07897415

$parameters
    target pattern rMax theta rho sigma2 rhoe pe0 per piM kappa
1 maxPower     CMD   20   0.5 0.7      1  0.2 0.2 0.3 0.2     3
  budget c1 beta alpha
1     40  1 -0.3  0.05

$f
    r  N     power
1   0 40 0.1148722
2   1 31 0.2291079
3   2 25 0.3686537
4   3 21 0.4887750
5   4 18 0.5829804
6   5 16 0.6643834
7   6 14 0.7166546
8   7 13 0.7766321
9   8 11 0.7857411
10  9 10 0.8143079
11 10 10 0.8682019
12 11  9 0.8784733
13 12  8 0.8809511
14 13  8 0.9142015
15 14  7 0.9072734
16 15  7 0.9322918
17 16  6 0.9178524
18 17  6 0.9384925
19 18  6 0.9546310
20 19  6 0.9670238
21 20  5 0.9495415
```

5.2.2. Estudio 2. Minimización del coste

Supongamos que estamos interesados en minimizar el coste de un estudio longitudinal, asumiendo el patrón de respuesta LDD, y alcanzando una potencia de, al menos, 0,8. El coste de la primera medida es $c_1 = 50$ unidades monetarias, que es $\kappa = 3$ veces el coste de cada una de las medidas subsiguientes. La estructura de covarianza de la respuesta es CS($\sigma = 1, \rho = 0{,}6$). La correlación intraclase de la exposición es $\rho_e = 0{,}6$. Se espera que uno de cada cinco participantes abandonará el estudio antes de concluirlo; es decir, $\pi_M = 0{,}2$. La prevalencia de la exposición se supone constante, igual a 0,2. El tamaño del efecto que se desea detectar es $\tilde{\beta} = 0{,}8$ y el nivel se significación se fija en $\alpha = 0{,}05$. El número máximo de medidas repetidas permitido es $r_{\max} = 20$. Así, podemos realizar los cálculos para el estudio y almacenar los resultados en el objeto `study2`:

```
> study2 <- OA(target = "minCost", pattern = "LDD", rMax = 20,
+              theta = 0, rho = 0.6, sigma2 = 1, rhoe = 0.6,
+              pe0 = 0.2, per = 0.2, piM = 0.2, kappa = 3,
+              reqPower = 0.8, c1 = 50, beta = 0.8, alpha = 0.05)
> study2

  Results subject to r not greater than 20:
-----------------------------------------------------
Optimal total number of measurements (r+1): 2
Optimal number of participants (N)        : 66
Minimized cost                            : 4180
```

Así, lo óptimo es realizar un estudio longitudinal con $N_{\text{opt}} = 66$ participantes y tomando un total de $r_{\text{opt}} + 1 = 2$ medidas. El coste minimizado de tal estudio es de 4180 unidades monetarias. La Figura 5.3 se ha obtenido usando la función `plot()` como en la sección previa.

5.3. Caso particular: diseño de un estudio transversal

La función `OA()` puede utilizarse también para el diseño de un estudio transversal. En este caso, es necesario fijar `pattern = "CMD"` y `rMax = 0`, tal como se ilustra en los dos ejemplos siguientes.

5.3.1. Estudio 3. Coste de un estudio transversal

Supongamos que estamos interesados en calcular el coste de un estudio transversal alcanzando una potencia de, al menos, 0,9 para detectar un efecto de tamaño

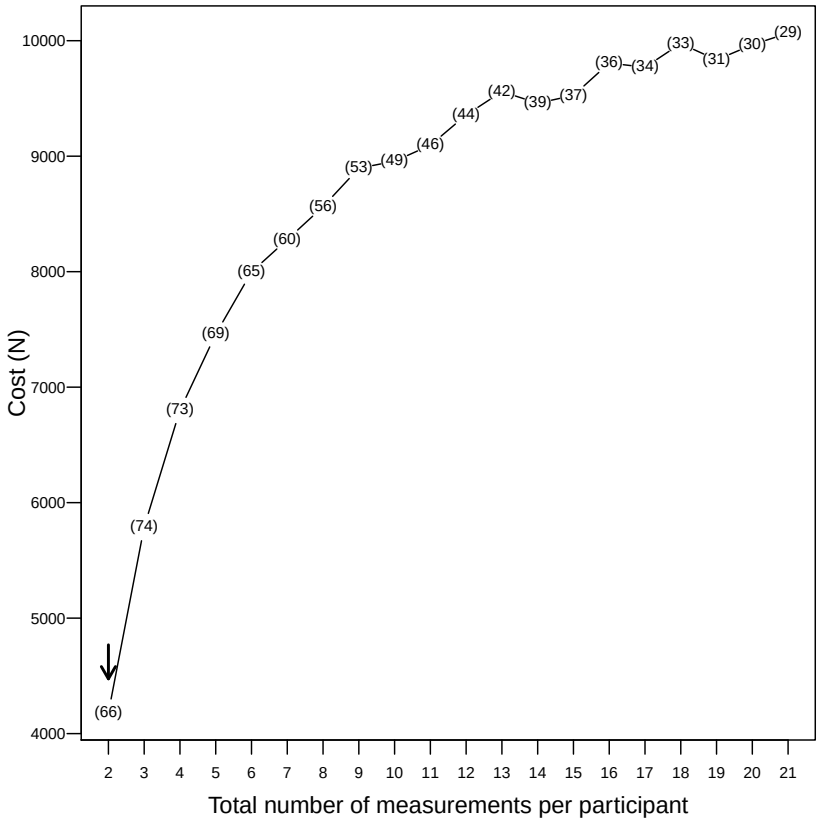


Figura 5.3: Coste minimizado y número de participantes (entre paréntesis) en función del número total de medidas por participante. La flecha apunta al diseño óptimo.

$\beta = -0{,}3$ con un nivel de significación $\alpha = 0{,}05$. El coste de la única medida por participante es $c_1 = 25$ unidades monetarias. La proporción de expuestos se asume igual a 0,3 y la varianza residual se estima en $\sigma^2 = 1$. Entonces, los cálculos para el estudio son:

```
> study3 <- OA(target = "minCost", pattern = "CMD", rMax = 0,
+              sigma2 = 1, pe0 = 0.3, reqPower = 0.9, c1 = 25,
+              beta = -0.3, alpha = 0.05)
> study3

  Results subject to a cross-sectional design:
-----------------------------------------------------
Number of participants (N): 556
Cost                      : 13900
```

Así, el número de participantes necesarios es $N = 556$ y el coste total es de 13900 unidades monetarias.

5.3.2. Estudio 4. Potencia de un estudio transversal

Supongamos que estamos interesados en calcular la potencia de un estudio transversal para detectar un efecto de tamaño $\beta = -0{,}3$ con un nivel de significación $\alpha = 0{,}05$. El presupuesto total disponible para el estudio es de 10.000 unidades monetarias y el coste de la única medida por participante es $c_1 = 25$ unidades monetarias. La proporción de expuestos se asume igual a 0,2 y la varianza residual se estima en $\sigma^2 = 1$. Entonces, los cálculos son:

```
> study4 <- OA(target = "maxPower", pattern = "CMD", rMax = 0,
+              sigma2 = 1, pe0 = 0.2, budget = 10000, c1 = 25,
+              beta = -0.3, alpha = 0.05)
> study4

  Results subject to a cross-sectional design:
-----------------------------------------------------
Number of participants (N): 400
Power                     : 0.6700445
```

Así, el número de participantes necesarios es $N = 400$ y la potencia alcanzada es 0,67.

5.4. Reproducción de los resultados obtenidos en el capítulo 4

Teniendo en cuenta los valores de los parámetros descritos en el ejemplo ilustrativo del capítulo 4, todos los cálculos necesarios para reproducir los resultados que allá se muestran pueden realizar con el código siguiente:

```
> # Creación de los escenarios:
>
> res <- expand.grid(Exposure = c("Aspirador", "Aerosoles"),
+                    rho = c(0.3, 0.7))
> res$pe0 <- 0.37
> res$pe0[res$Exposure == "Aerosoles"] <- 0.17
> res$per <- res$pe0
> res$rhoe <- 0.13
> res$rhoe[res$Exposure == "Aerosoles"] <- 0.60
> res$TimeVaryingExposure <- TRUE
> aux <- res
> aux$TimeVaryingExposure <- FALSE
```

```
> aux$rhoe <- 1
> res <- rbind(res, aux)
> res$r <- NA
> res$N <- NA
> res$cost <- NA
> # Sorting the table:
> ord <- order(res$Exposure, res$rho, 1 - res$TimeVaryingExposure)
> res <- res[ord, ]
> rownames(res) <- NULL
> res
> # Diseño óptimo en cada escenario:
>
> studies <- list()
> for (i in 1:nrow(res))
+  {
+   studies[[i]] <- OA(target = "minCost", pattern = "CMD", rMax = 20,
+                      theta = 0.12, rho = res$rho[i], sigma2 = 0.43,
+                      rhoe = res$rhoe[i], pe0 = res$pe0[i],
+                      per = res$per[i], piM = 0.28, kappa = 2,
+                      reqPower = 0.9, c1 = 1, beta = -0.39,
+                      alpha = 0.05)
+   res$r[i] <- studies[[i]]$ropt
+   res$N[i] <- studies[[i]]$Nopt
+   res$cost[i] <- round(studies[[i]]$minCost, 1)
+  }
> # Results:
>
> studyDesigns <- res[, -c(3:5)]
```

que permite reproducir la Tabla 4.1:

```
> studyDesigns

   Exposure rho TimeVaryingExposure  r   N  cost
1 Aspirador 0.3                TRUE 18   6  51.6
2 Aspirador 0.3               FALSE  1  92 125.1
3 Aspirador 0.7                TRUE 15   3  22.0
4 Aspirador 0.7               FALSE  0 128 128.0
5 Aerosoles 0.3                TRUE 20  17 160.7
6 Aerosoles 0.3               FALSE  1 152 206.7
```

```
7 Aerosoles 0.7                TRUE 19   8  72.2
8 Aerosoles 0.7               FALSE  0 211 211.0
```

Por otro lado, la Figura 4.3 se reproduce con el código siguiente:

```
> for (i in c(1,3,5,7))
+  {
+   plot(studies[[i]], Ncex = 0.5)
+   mtext(text = res$Exposure[i], side = 3, line = 1, cex = 0.8)
+   mtext(text = bquote(paste(rho, sep = "") ==. (res$rho[i])),
+         side = 3, line = 0, cex = 0.8)
+  }
```

Capítulo 6

Discusión

En este trabajo hemos encontrado la combinación óptima del número de participantes y el número de medidas repetidas cuando existen restricciones económicas en el diseño del estudio. Examinamos una variedad de situaciones en el contexto de los estudios longitudinales observacionales en los que la exposición intra-participante puede variar a lo largo del tiempo de una manera no controlada por el investigador. En nuestro estudio se ha revelado que el grado de variación intra-individuo de la exposición, concretamente medido por su correlación intraclase, juega un papel esencial en el resultado del diseño óptimo. Este diseño óptimo puede expresarse con fórmulas cerradas para unos escenarios básicos mientras que, para escenarios más generales, el diseño óptimo debe encontrarse numéricamente para lo cual proveemos el paquete `R` que hemos creado, llamado optimalAllocation.

En el contexto de los modelos para efectos agudos y transitorios de la exposición, los resultados obtenidos ya eran conocidos para el caso particular de una exposición constante. En este caso, los resultados sugieren que cuando la medidas repetidas no son mucho más baratas que la primera (es decir, para costes de reclutamiento pequeños), lo óptimo es tomar solo una medida o unas pocas medidas repetidas. Sin embargo, en la generalización a una exposición variable en el tiempo, pueden requerirse más medidas repetidas incluso hasta el extremo de que lo óptimo sea tomar tantas medidas como sea posible cuando la variabilidad intra-individuo de la exposición es elevada.

En el contexto de los modelos para efectos acumulativos de la exposición, si la exposición es constante, lo óptimo es tomar solo una medida repetida en el caso en que la ratio entre el coste de la primera medida y el de las subsiguientes es menor que 5, y tomar tantas medidas como sea posible en caso contrario. Al generalizar a una exposición variable en el tiempo, se mantiene el patrón pero la ratio de costes umbral depende de la variabilidad intra-individuo, siendo en todo caso superior a 5, incrementando así el número de situaciones en que lo óptimo es tomar solo una

medida repetida.

Nuestros resultados extienden los obtenidos en trabajos previos sobre asignación óptima en estudios longitudinales [24–32] al contexto de una exposición variable en el tiempo. Algunos de los estudios previos utilizaron condiciones ligeramente diferentes tales como diferencias en la estructura de covarianza, en la función de coste o en la estructura de datos faltantes, aunque la metodología descrita aquí puede adaptarse fácilmente a esos cambios.

Los resultados generales encontrados que han sido descritos anteriormente corresponden a un escenario básico en que la respuesta tiene estructura de covarianza CS, no existen datos faltantes y la prevalencia de la exposición es constante pero también exploramos cómo esos resultados cambian cuando violamos esas asunciones. Nuestra exploración fue realizada violando las asunciones de una en una pero el paquete optimalAllocation que proveemos permite realizar comparaciones modificando simultáneamente más de una asunción.

En cuanto a la estructura de covarianza de la respuesta, exploramos desviaciones respecto a los resultados en el escenario básico considerando una estructura DEX en lugar de CS. Existen otras posibilidades; otros investigadores han empleado covarianzas resultantes de modelos mixtos con efectos aleatorios en la constante y en la pendiente [14, 22], o bien covarianzas que mezclaban un efecto aleatorio en la constante con errores autorregresivos [32]. Cuando la estructura de covarianza de la respuesta es DEX en lugar de CS, se observaron cambios. Bajo el modelo para un efecto agudo y transitorio de la exposición, el número óptimo de medidas repetidas aumenta al hacerlo el parámetro de amortiguamiento θ, amplificándose así las diferencias respecto al caso de exposición constante. Bajo el modelo para un efecto acumulativo de la exposición, el incremento del valor de θ rompe la dicotomía $1-\infty$ en el número óptimo de medidas repetidas que se observa cuando la respuesta tiene estructura de covarianza CS. En general, valores elevados de θ favorecen un menor número de medidas repetidas en comparación al caso de estructura CS para la covarianza de la respuesta.

En lo referente a posible abandono del estudio por parte de los participantes, nuestros resultados, basados en un patrón de pérdida monótono, mostraron que los resultados en r_{opt} prácticamente no se ven afectados por la pérdida de seguimiento excepto para unas pocas combinaciones de los parámetros involucrados en el diseño del estudio. Galbraith [31] también analizó el efecto de datos faltantes en la asignación óptima (N, r) para el patrón de respuesta bajo efecto acumulativo de la exposición pero en el contexto particular de una exposición constante. Examinaron diversos patrones de pérdida y también encontraron que la asignación óptima no se veía afectada en general por los niveles de deserción de los participantes. Los autores sugirieron calcular el tamaño muestral para una potencia del 90 % cuando de he-

cho se deseaba una potencia del 80 %, si la perdida total esperada de participantes no supera el 30 %. En nuestro caso, usamos el nivel exacto de la pérdida esperada de participantes directamente como un parámetro de entrada para los cálculos de diseño óptimo.

Respecto a la prevalencia de la exposición, su valor no afecta a r_{opt} si la exposición es constante. Sin embargo, no es así cuando la exposición varía en el tiempo, de manera que exploramos los cambios en r_{opt} para una tendencia lineal de esta prevalencia. Bajo el modelo para un efecto agudo y transitorio de la exposición, a mayor variabilidad de la prevalencia, mayor es el número óptimo de medidas. Bajo el modelo para un efecto acumulativo de la exposición, el efecto depende del signo de la tendencia de la prevalencia de la exposición. Así, si la prevalencia de la exposición aumenta en el tiempo, el número óptimo de medidas repetidas tiende a disminuir (cambiando siempre de ∞ a 1) mientras que si la prevalencia de la exposición disminuye en el tiempo, la dicotomía $1 - \infty$ se rompe sin patrones claros respecto al resto de los parámetros involucrados en el diseño del estudio.

Nuestra metodología es fácilmente adaptable a cambios en diversas asunciones como, por ejemplo, funciones de coste diferentes o estructuras de covarianza más complejas para la respuesta (como aquellas presentes en modelos mixtos con efectos aleatorios tanto en la constante como en el tiempo). También podría considerarse una exposición continua, en cuyo caso se necesitarían sus momentos de primer y segundo orden, como parámetros de entrada.

Cabe destacar que en este estudio se han asumido efectos lineales de la exposición sobre la respuesta. Tal asunción posibilita que lo óptimo pueda ser tomar solo dos medidas en diversos escenarios, situación que evidencia que, si se espera o se desea detectar un efecto de aceleración o de orden superior, el diseño analizado aquí sería inapropiado. Por otro lado, también podría ocurrir que otras motivaciones diferentes a la potencia y los costes favorecieses otros diseños. Por ejemplo, podrían considerarse estrategias de diseño que maximicen el compromiso de los participantes en el estudio.

En este trabajo hemos generalizado la búsqueda de la combinación óptima de participantes y número de medidas repetidas en estudios longitudinales observacionales en que la exposición varía en tiempo, incluso de una manera no controlada por el investigador. Hemos aplicado nuestra metodología a un ejemplo ilustrativo cuyos resultados mostraron que asumir erróneamente una exposición constante puede derivar en un diseño del estudio ineficiente. Para facilitar todos los cálculos requeridos en el estudio de diseño, hemos creado el paquete optimalAllocationde `R` que se encuentra disponible en `http://www.mat.uab.cat/~jbarrera/Software.html`.

El artículo [5], en el que se basa esta memoria, fue enviado a la revista *Statistics en Medicine* en junio de 2012 encontrándose actualmente bajo revisión por pares.

Bibliografía

[1] Basagaña X, Sartini C, Barrera-Gómez J, Dadvand P, Cunillera J, Ostro B, Sunyer J, Medina-Ramón M. Heat waves and cause-specific mortality at all ages. *Epidemiology*. 2011, 22(6), 765–772. DOI: 10.1097/EDE.0b013e31823031c5.

[2] Dadvand P, Basagaña X, Barrera-Gómez J, Diffey B, Nieuwenhuijsen M. Measurement errors in the assessment of exposure to solar ultraviolet radiation and its impact on risk estimates in epidemiological. *Photochemical & Photobiological Sciences*. 2011, 10(7), 1161–1168. DOI: 10.1039/C0PP00333F.

[3] Ostro B, Barrera-Gómez J, Ballester J, Basagaña X, Sunyer J. The impact of future summer temperature on public health in Barcelona and Catalonia, Spain. *International Journal of Biometeorology*. 2012 (e-publicación). DOI: 10.1007/s00484-012-0529-7.

[4] Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. A framework for multiple imputation in cluster analysis. *American Journal of Epidemiology*. 2012 (en prensa).

[5] Barrera-Gómez J, Spiegelman D, Basagaña X. Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure. (Enviado a *Statistics in Medicine*, actualmente en proceso de revisión por pares).

[6] Barrera-Gómez J, Basagaña X. Interpretación de modelos lineales con variables transformadas (enviado a *Revista Española de Salud Pública* en julio de 2012).

[7] Schlesselman JJ. Planning a longitudinal study. II. Frequency of measurement and study duration. *Journal of Chronic Diseases* 1973; **26**(9):561–570. DOI:10.1016/0021-9681(73)90061-1.

[8] Kirby AJ, Galai N, Muñoz A. Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials* 1994; **15**(3):165–172. DOI: 10.1016/0197-2456(94)90054-X.

[9] Frison LJ, Pocock SJ. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Statistics in Medicine* 1997; **16**(24):2855–2872. DOI: 10.1002/(SICI)1097-0258(19971230)16:24<2855::AID-SIM749>3.0.CO;2-Y.

[10] Dawson JD. Sample size calculations based on slopes and other summary statistics. *Biometrics* 1998; **54**(1):323–330. DOI:10.2307/2534019.

[11] Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine* 1998; **17**(14):1643–1658. DOI:10.1002/(SICI)1097-0258(19980730)17:14<1643::AID-SIM869>3.0.CO;2-3.

[12] Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics* 1999; **24**(1):70–93. DOI:10.2307/1165262.

[13] Schouten HJ. Planning group sizes in clinical trials with a continuous outcome and repeated measures. *Statistics in Medicine* 1999; **18**(3):255–264. DOI:10.1002/(SICI)1097-0258(19990215)18:3<255::AID-SIM16>3.0.CO;2-K

[14] Raudenbush SW, Xiao-Feng L. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods* 2001; **6**(4):387–401. DOI: 10.1037/1082-989X.6.4.387.

[15] Diggle P, Heagerty P, Liang KY, Zeger S. Analysis of Longitudinal Data (2nd edn.). *Oxford Statistical Science Series*, vol. 25. Oxford University Press: Oxford, New York, 2002.

[16] Yi Q, Panzarella T. Estimating sample size for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. *Controlled Clinical Trials* 2002; **23**(5):481–496. DOI:10.1016/S0197-2456(02)00223-4.

[17] Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine* 2003; **22**(8):1305–1315. DOI:10.1002/sim.1384.

[18] Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. Wiley Series in Probability and Statistics. Wiley-Interscience: Hoboken, NJ, 2004.

[19] Jones B, Kenward MG. Design and Analysis of Cross-over Trials (1st edn.). Monographs on Statistics and Applied Probability, vol. 34. Chapman & Hall: London, New York, 1989.

[20] Senn S. Cross-over trials in clinical research (2nd edn.). John Wiley, Chichester, Eng., New York, 2002.

[21] Julious SA. Sample sizes for clinical trials with normal data. *Statistics in Medicine* 2004; **23**(12):1921–1986. DOI:10.1002/sim.1783.

[22] Basagaña X, Spiegelman D. Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposición. *Statistics in Medicine* 2010; **29**(2):181–192. DOI: 10.1002/sim.3772.

[23] Basagaña, X., Liao, X., Spiegelman, D. Power and sample size calculations for longitudinal studies estimating a main effect of a time-varying exposición. *Statistical Methods in Medical Research* 2011; **20**(5):471–487. DOI:10.1177/0962280210371563.

[24] Cochran WG. *Sampling techniques* (2nd edn). Wiley, New York, 1977.

[25] Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods* 1997; **2**(2):173–185. DOI:10.1037/1082-989X.2.2.173.

[26] Bloch DA. Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Statistics in Medicine* 1986; **5**(6):663-667. DOI:10.1002/sim.4780050613.

[27] Snijders TAB, Bosker, RJ. Standard errors and sample sizes for two-level research. *Journal of Educational Statistics* 1993; **18**(3):237–259. DOI:10.2307/1165134.

[28] Allison DB, Allison RL, Faith MS, Paultre F, Pi-Sunyer FX. Power and money: designing statistically powerful studies while minimizing financial costs. *Psychological Methods* 1997; **2** (1):20-33. DOI:10.1037/1082-989X.2.1.20.

[29] Moerbeek M, Van Breukelen JP, Berger MPF. Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics* 2000; **25**(3):271–284. DOI:10.2307/1165206.

[30] Moerbeek M, Van Breukelen JP, Berger MPF. Optimal experimental designs for multilevel models with covariates. *Communications in Statistics - Theory and Methods* 2001; **30**(12):2683–2697. DOI:10.1081/STA-100108453.

[31] Galbraith S, Marschner IC. Guidelines for the design of clinical trials with longitudinal outcomes. *Controlled Clinical Trials* 2002; **23**(3):257–273. DOI:10.1016/S0197-2456(02)00205-2.

[32] Winkens B, Schouten HJ, van Breukelen GJ, Berger MP. Optimal number of repeated measures and group sizes in clinical trials with linearly divergent treatment effects. *Contemporary Clinical Trials* 2005; **27**(1):57–69. DOI: 10.1016/j.cct.2005.09.005.

[33] Zhang S, Ahn C. Adding subjects or adding measurements in repeated measurement studies under financial constraints. *Statistics in Biopharmaceutical Research* 2011; **3**(1):54–64. DOI:10.1198/sbr.2010.10022.

[34] Medina-Ramón M, Zock JP, Kogevinas M, Sunyer J, Basagaña X, Schwartz J, Burge PS, Moore V, Antó JM. Short-term respiratory effects of cleaning exposicións in female domestic cleaners. *European Respiratory Journal* 2006; **27**(6):1196–1203. DOI:10.1183/09031936.06.00085405.

[35] Muñoz A, Carey V, Schouten JP, Segal M, Rosner B. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* 1992; **48**(3):733–742. DOI:10.2307/2532340.

[36] Whittemore AS. Sample size for logistic regression with small response probability. *Journal of the American Statistical Association* 1981; **76**(373):27–32. DOI:10.2307/2287036.

[37] Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* 2000; **56**(4):1192–1196. DOI:10.1111/j.0006-341X.2000.01192.x.

[38] Tu, XM., Kowalski, J., Zhang, J., Lynch, KG., Crits-Christoph, P. Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine* 2004; **23** (18):2799–2815. DOI:10.1002/sim.1869.

[39] Kistner EO, Muller KE, Exact distributions of intraclass correlation and Cronbach's alpha with gaussian data and general covariance. *Psychometrika* 2004; **69**(3):459–474. DOI:10.1007/BF02295646.

[40] Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999; **55**(1):137–148. DOI:10.1111/j.0006-341X.1999.00137.x.

[41] Gange SJ. Generating Multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 1995; **49**(2): 134–138. DOI: 10.1080/00031305.1995.10476130.

Apéndice A

Derivación de $\tilde{\sigma}^2$ en los escenarios básicos

A.1. Caso particular de exposición constante

A.1.1. Patrón de respuesta CMD

Asumimos estructura de covarianza de la respuesta CS(σ, ρ), ausencia de datos faltantes y prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 0, \ldots, r$).

Del modelo (2.4) tenemos

$$\begin{aligned}
\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i &= \begin{pmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & s & \cdots & sk & \cdots & sr \\ E_i & E_i & \cdots & E_i & \cdots & E_i \end{pmatrix} \begin{pmatrix} \nu_{00} & \cdots & \nu_{0r} \\ \vdots & \ddots & \vdots \\ \nu_{r0} & \cdots & \nu_{rr} \end{pmatrix} \begin{pmatrix} 1 & 0 & E_i \\ 1 & s & E_i \\ \vdots & \vdots & \vdots \\ 1 & sj & E_i \\ \vdots & \vdots & \vdots \\ 1 & sr & E_i \end{pmatrix} \\
&= \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \\ 0 & \cdots & sk & \cdots & sr \\ E_i & \cdots & E_i & \cdots & E_i \end{pmatrix} \begin{pmatrix} \sum_{j=0}^r \nu_{0j} & s\sum_{j=0}^r j\nu_{0j} & E_i\sum_{j=0}^r \nu_{0j} \\ \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{mj} & s\sum_{j=0}^r j\nu_{mj} & E_i\sum_{j=0}^r \nu_{mj} \\ \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{rj} & s\sum_{j=0}^r j\nu_{rj} & E_i\sum_{j=0}^r \nu_{rj} \end{pmatrix} \\
&= \begin{pmatrix} \omega_{00} & s\omega_{10} & E_i\omega_{00} \\ s\omega_{10} & s^2\omega_{11} & sE_i\omega_{10} \\ E_i\omega_{00} & sE_i\omega_{10} & E_i^2\omega_{00} \end{pmatrix},
\end{aligned}$$

donde $s = \frac{1}{r}$ es el intervalo de tiempo entre dos medidas consecutivas en unidades

del tiempo total de seguimiento del estudio y

$$\omega_{pq} := \sum_{j=0}^{r}\sum_{k=0}^{r} j^p k^q \nu_{jk}.$$

Dado que $E_i \sim \text{Bernoulli}(p_e)$, tenemos

$$\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right] = \begin{pmatrix} \omega_{00} & s\omega_{10} & p_e\omega_{00} \\ s\omega_{10} & s^2\omega_{11} & sp_e\omega_{10} \\ p_e\omega_{00} & sp_e\omega_{10} & p_e\omega_{00} \end{pmatrix}$$

y

$$\det(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]) = \omega_{00}(\omega_{00}\omega_{11} - \omega_{10}^2)s^2p_e(1-p_e).$$

Nos interesa el elemento [3,3] de $\left(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]\right)^{-1}$ que es

$$\frac{\begin{vmatrix} \omega_{00} & s\omega_{10} \\ s\omega_{10} & s^2\omega_{11} \end{vmatrix}}{\det(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right])} = \frac{1}{\omega_{00}p_e(1-p_e)}$$

de manera que

$$\tilde{\sigma}^2 = \frac{1}{p_e(1-p_e)\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}}.$$

Si la estructura de covarianza de la respuesta es CS(σ, ρ), entonces

$$\boldsymbol{\Sigma}[j,k] = \begin{cases} \sigma^2 & , \quad j = k \\ \sigma^2\rho & , \quad j \neq k \end{cases}$$

y

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \sigma^{-2}\begin{pmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & \cdots & \rho & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}^{-1} \\ &= \frac{1}{\sigma^2(1-\rho)(\rho r+1)}\begin{pmatrix} \rho r+1-\rho & -\rho & \cdots & -\rho & -\rho \\ -\rho & \rho r+1-\rho & \cdots & -\rho & -\rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\rho & -\rho & \cdots & \rho r+1-\rho & -\rho \\ -\rho & -\rho & \cdots & -\rho & \rho r+1-\rho \end{pmatrix} \end{aligned}$$

por lo que

$$\omega_{00} = \sum_{j=0}^{r} \sum_{k=0}^{r} \nu_{jk} = \frac{r+1}{\sigma^2(\rho r + 1)}$$

y entonces

$$\tilde{\sigma}^2 = \frac{\sigma^2(\rho r + 1)}{p_e(1-p_e)(r+1)}.$$

A.1.2. Patrón de respuesta LDD

Asumimos que la estructura de covarianza de la respuesta es CS(σ, ρ), ausencia de datos faltantes y prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 0, \ldots, r$).

Del modelo (2.5) tenemos

$$\begin{aligned}
\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & s & \cdots & sr \\ E_i & E_i & \cdots & E_i \\ 0 & sE_i & \cdots & srE_i \end{pmatrix} \begin{pmatrix} \nu_{00} & \cdots & \nu_{0r} \\ \vdots & \ddots & \vdots \\ \nu_{r0} & \cdots & \nu_{rr} \end{pmatrix} \begin{pmatrix} 1 & 0 & E_i & 0 \\ 1 & s & E_i & sE_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & sj & E_i & sjE_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & sr & E_i & srE_i \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & s & \cdots & sr \\ E_i & E_i & \cdots & E_i \\ 0 & sE_i & \cdots & srE_i \end{pmatrix} \cdot \\
&\quad \begin{pmatrix} \sum_{j=0}^r \nu_{0j} & s\sum_{j=0}^r j\nu_{0j} & E_i\sum_{j=0}^r \nu_{0j} & sE_i\sum_{j=0}^r j\nu_{0j} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{mj} & s\sum_{j=0}^r j\nu_{mj} & E_i\sum_{j=0}^r \nu_{mj} & sE_i\sum_{j=0}^r j\nu_{mj} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{rj} & s\sum_{j=0}^r j\nu_{rj} & E_i\sum_{j=0}^r \nu_{rj} & sE_i\sum_{j=0}^r j\nu_{rj} \end{pmatrix} \\
&= \begin{pmatrix} \omega_{00} & s\omega_{10} & E_i\omega_{00} & sE_i\omega_{10} \\ s\omega_{10} & s^2\omega_{11} & sE_i\omega_{10} & s^2E_i\omega_{11} \\ E_i\omega_{00} & sE_i\omega_{10} & E_i^2\omega_{00} & sE_i^2\omega_{10} \\ sE_i\omega_{10} & s^2E_i\omega_{11} & sE_i^2\omega_{10} & s^2E_i^2\omega_{11} \end{pmatrix}.
\end{aligned}$$

Entonces

$$\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right] = \begin{pmatrix} \omega_{00} & s\omega_{10} & p_e\omega_{00} & sp_e\omega_{10} \\ s\omega_{10} & s^2\omega_{11} & sp_e\omega_{10} & s^2p_e\omega_{11} \\ p_e\omega_{00} & sp_e\omega_{10} & p_e\omega_{00} & sp_e\omega_{10} \\ sp_e\omega_{10} & s^2p_e\omega_{11} & sp_e\omega_{10} & s^2p_e\omega_{11} \end{pmatrix}$$

y

$$\det(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]) = \left[(\omega_{00}\omega_{11} - \omega_{10}^2)s^2p_e(1-p_e)\right]^2.$$

Nos interesa en elemento [4,4] de $\left(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]\right)^{-1}$ que es

$$\frac{\begin{vmatrix} \omega_{00} & s\omega_{10} & p_e\omega_{00} \\ s\omega_{10} & s^2\omega_{11} & sp_e\omega_{10} \\ p_e\omega_{00} & sp_e\omega_{10} & p_e\omega_{00} \end{vmatrix}}{\det(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right])} = \frac{\omega_{00}}{(\omega_{00}\omega_{11}-\omega_{10}^2)s^2p_e(1-p_e)}$$

de manera que

$$\tilde{\sigma}^2 = \frac{r^2\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}}{p_e(1-p_e)\left[\left(\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\right)\left(\sum_{j=0}^{r}\sum_{k=0}^{r}jk\nu_{jk}\right)-\left(\sum_{j=0}^{r}\sum_{k=0}^{r}j\nu_{jk}\right)^2\right]}.$$

Si la estructura de covarianza de la respuesta es CS(σ, ρ), entonces

$$\omega_{10} = \sum_{j=0}^{r}\sum_{k=0}^{r}j\nu_{jk} = \frac{r(r+1)}{2\sigma^2(\rho r+1)}$$

y

$$\omega_{11} = \sum_{j=0}^{r}\sum_{k=0}^{r}jk\nu_{jk} = \frac{r(r+1)(\rho r^2+(4-\rho)r+2)}{12\sigma^2(1-\rho)(\rho r+1)}$$

de manera que

$$\tilde{\sigma}^2 = \frac{12\sigma^2(1-\rho)r}{p_e(1-p_e)(r+1)(r+2)}.$$

A.2. Caso general de exposición variable en el tiempo

A.2.1. Patrón de respuesta CMD

Asumimos que la estructura de covarianza de la respuesta es CS(σ, ρ), ausencia de datos faltantes y prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 0, \ldots, r$). Consideramos una matriz de covarianzas general para la exposición:

$$\mathbf{\Sigma}_E[j,k] = \begin{cases} \text{Var}(E_{ij}) = p_e(1-p_e) & , \quad j = k \\ \sigma_{e_{jk}} := \text{Cov}(E_{ij}, E_{ik}) = \rho_{e_{jk}} p_e(1-p_e) & , \quad j \neq k \end{cases},$$

donde $\rho_{e_{jk}} = \text{Cor}(E_{ij}, E_{ik})$ es la correlación entre las medidas j-ésima y k-ésima de la exposición, asumida común a todos los individuos.

Del modelo (2.4) tenemos

$$\begin{aligned} \mathbf{X}_i'\mathbf{\Sigma}^{-1}\mathbf{X}_i &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & s & \cdots & sr \\ E_{i0} & E_{i1} & \cdots & E_{ir} \end{pmatrix} \begin{pmatrix} \nu_{00} & \cdots & \nu_{0r} \\ \vdots & \ddots & \vdots \\ \nu_{r0} & \cdots & \nu_{rr} \end{pmatrix} \begin{pmatrix} 1 & 0 & E_{i0} \\ 1 & s & E_{i1} \\ \vdots & \vdots & \vdots \\ 1 & sj & E_{ij} \\ \vdots & \vdots & \vdots \\ 1 & sr & E_{ir} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \cdots & 1 \\ 0 & \cdots & sr \\ E_{i0} & \cdots & E_{ir} \end{pmatrix} \begin{pmatrix} \sum_{j=0}^r \nu_{0j} & s\sum_{j=0}^r j\nu_{0j} & \sum_{j=0}^r \nu_{0j}E_{ij} \\ \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{mj} & s\sum_{j=0}^r j\nu_{mj} & \sum_{j=0}^r \nu_{mj}E_{ij} \\ \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{rj} & s\sum_{j=0}^r j\nu_{rj} & \sum_{j=0}^r \nu_{rj}E_{ij} \end{pmatrix} \\ &= \begin{pmatrix} \omega_{00} & s\omega_{10} & \phi_0 \\ s\omega_{10} & s^2\omega_{11} & s\phi_1 \\ \phi_0 & s\phi_1 & \epsilon \end{pmatrix}, \end{aligned}$$

donde

$$\phi_q := \sum_{j=0}^r \sum_{k=0}^r k^q \nu_{jk} E_{ij},\ q = 0, 1; \qquad \epsilon := \sum_{j=0}^r \sum_{k=0}^r \nu_{jk} E_{ij} E_{ik}.$$

Ahora calculamos:

$$\begin{aligned} \mathbb{E}_X[\phi_q] &= \mathbb{E}_X\left[\sum_{j=0}^r \sum_{k=0}^r k^q \nu_{jk} E_{ij}\right] = \sum_{j=0}^r \sum_{k=0}^r k^q \nu_{jk} \mathbb{E}_X[E_{ij}] \\ &= p_e \sum_{j=0}^r \sum_{k=0}^r k^q \nu_{jk} = p_e \omega_{q0}, \quad q = 0, 1. \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_X[\epsilon] &= \mathbb{E}_X\left[\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}E_{ij}E_{ik}\right] = \sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\mathbb{E}_X\left[E_{ij}E_{ik}\right] \\
&= \sum_{j=0}^{r}\nu_{jj}\mathbb{E}_X\left[E_{ij}^2\right] + \sum_{j=0,\, j\neq k}^{r}\sum_{k=0}^{r}\nu_{jk}\mathbb{E}_X\left[E_{ij}E_{ik}\right] \\
&= \sum_{j=0}^{r}\nu_{jj}\mathbb{E}_X[E_{ij}] + \sum_{j=0,\, j\neq k}^{r}\sum_{k=0}^{r}\nu_{jk}\left\{\mathrm{Cov}(E_{ij},E_{ik}) + \mathbb{E}_X\left[E_{ij}\right]\mathbb{E}_X\left[E_{ik}\right]\right\} \\
&= p_e\sum_{j=0}^{r}\nu_{jj} + \sum_{j=0,\, j\neq k}^{r}\sum_{k=0}^{r}\nu_{jk}\left[\rho_{e_{jk}}p_e(1-p_e) + p_e^2\right] \\
&= p_e\sum_{j=0}^{r}\nu_{jj} + p_e(1-p_e)\sum_{j=0,\, j\neq k}^{r}\sum_{k=0}^{r}\nu_{jk}\rho_{e_{jk}} + p_e^2\sum_{j=0,\, j\neq k}^{r}\sum_{k=0}^{r}\nu_{jk} \\
&= p_e\sum_{j=0}^{r}\nu_{jj} + p_e(1-p_e)\left[\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\rho_{e_{jk}} - \sum_{j=0}^{r}\nu_{jj}\rho_{e_{jj}}\right] + p_e^2\left[\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk} - \sum_{j=0}^{r}\nu_{jj}\right] \\
&= p_e^2\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk} + p_e(1-p_e)\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\rho_{e_{jk}} = p_e^2\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk} + \sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\sigma^2_{e_{jk}} \\
&= p_e^2\,\omega_{00} + \alpha_0,
\end{aligned}
$$

donde

$$\alpha_0 := \sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\sigma^2_{e_{jk}}$$

y $\sigma^2_{e_{jk}}$ es el elemento $[j,k]$ de la matriz de covarianzas de la exposición $\mathbf{\Sigma}_E$.

Entonces,

$$\mathbb{E}_X\left[\mathbf{X}_i'\mathbf{\Sigma}^{-1}\mathbf{X}_i\right] = \begin{pmatrix} \omega_{00} & s\omega_{10} & p_e\omega_{00} \\ s\omega_{10} & s^2\omega_{11} & sp_e\omega_{10} \\ p_e\omega_{00} & sp_e\omega_{10} & p_e^2\,\omega_{00} + \alpha_0 \end{pmatrix}$$

y

$$\det(\mathbb{E}_X\left[\mathbf{X}_i'\mathbf{\Sigma}^{-1}\mathbf{X}_i\right]) = (\omega_{00}\omega_{11} - \omega_{10}^2)s^2\alpha_0\,.$$

Nos interesa el elemento [3,3] de $\left(\mathbb{E}_X\left[\mathbf{X}_i'\mathbf{\Sigma}^{-1}\mathbf{X}_i\right]\right)^{-1}$ que es

$$\frac{\begin{vmatrix} \omega_{00} & s\omega_{10} \\ s\omega_{10} & s^2\omega_{11} \end{vmatrix}}{\det(\mathbb{E}_X\left[\mathbf{X}_i'\mathbf{\Sigma}^{-1}\mathbf{X}_i\right])} = \frac{1}{\alpha_0}$$

y entonces

$$\tilde{\sigma}^2 = \frac{1}{\sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\sigma^2_{e_{jk}}}\,.$$

Si la estructura de covarianza de la respuesta es CS(σ, ρ), entonces

$$\nu_{jk} = \begin{cases} \frac{\rho r+1-\rho}{\sigma^2(1-\rho)(\rho r+1)} & , \quad j=k \\ \frac{-\rho}{\sigma^2(1-\rho)(\rho r+1)} & , \quad j\neq k \end{cases}$$

y

$$\begin{aligned}
\alpha_0 &= \sum_{j=0}^{r}\sum_{k=0}^{r}\nu_{jk}\sigma^2_{e_{jk}} = \sum_{j=0}^{r}\nu_{jj}\sigma^2_{e_{jj}} + \sum_{j=0,\,j\neq k}^{r}\sum_{k=0}^{r}\nu_{jk}\sigma^2_{e_{jk}} \\
&= \frac{1}{\sigma^2(1-\rho)(\rho r+1)}\left[p_e(1-p_e)\sum_{j=0}^{r}(\rho r+1-\rho) - \sum_{j=0,\,j\neq k}^{r}\sum_{k=0}^{r}\rho\sigma^2_{e_{jk}}\right] \\
&= \frac{1}{\sigma^2(1-\rho)(\rho r+1)}\left[p_e(1-p_e)(r+1)(\rho r+1-\rho) - \rho\sum_{j=0,\,j\neq k}^{r}\sum_{k=0}^{r}\sigma^2_{e_{jk}}\right] \\
&= \frac{1}{\sigma^2(1-\rho)(\rho r+1)}\left\{p_e(1-p_e)(r+1)(\rho r+1-\rho) - \rho\left[\sum_{j=0}^{r}\sum_{k=0}^{r}\sigma^2_{e_{jk}} - \sum_{j=0}^{r}\sigma^2_{e_{jj}}\right]\right\} \\
&= \frac{1}{\sigma^2(1-\rho)(\rho r+1)}\left\{p_e(1-p_e)(r+1)(\rho r+1-\rho) - \rho\left[\mathrm{sum}(\mathbf{\Sigma}_E) - (r+1)p_e(1-p_e)\right]\right\} \\
&= \frac{p_e(1-p_e)(r+1)(\rho r+1) - \rho\,\mathrm{sum}(\mathbf{\Sigma}_E)}{\sigma^2(1-\rho)(\rho r+1)} = \frac{p_e(1-p_e)(r+1)[\rho(1-\rho_e)r+1-\rho]}{\sigma^2(1-\rho)(\rho r+1)}
\end{aligned}$$

y entonces

$$\tilde{\sigma}^2 = \frac{\sigma^2(1-\rho)(\rho r+1)}{p_e(1-p_e)(r+1)[\rho(1-\rho_e)r+1-\rho]}\,,$$

donde la correlación intraclase de la exposición es

$$\rho_e = \frac{\mathrm{sum}(\mathbf{\Sigma}_E) - \mathrm{Tr}(\mathbf{\Sigma}_E)}{r\,\mathrm{Tr}(\mathbf{\Sigma}_E)}\,,$$

siendo sum() y Tr() la suma de los elementos y la traza de una matriz, respectivamente.

A.2.2. Patrón de respuesta LDD

Asumimos que la estructura de covarianza de la respuesta es CS(σ, ρ), ausencia de datos faltantes y prevalencia de la exposición constante ($p_{ej} = p_e,\ \forall j = 0, \ldots, r$). Consideramos una estructura CS para la matriz de covarianzas de la exposición:

$$\mathbf{\Sigma}_E[j,k] = \begin{cases} \text{Var}(E_{ij}) = p_e(1-p_e) & , \quad j = k \\ \sigma_{e_{jk}} := \text{Cov}(E_{ij}, E_{ik}) = \rho_e p_e(1-p_e) & , \quad j \neq k \end{cases},$$

donde $\rho_e = \text{Cor}(E_{ij}, E_{ik})$ es la correlación común de la exposición equivalente a su correlación intraclase.

Del modelo (2.5) tenemos

$$\begin{aligned} \mathbf{X}_i'\mathbf{\Sigma}^{-1}\mathbf{X}_i &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & s & \cdots & sr \\ E_{i0} & E_{i0} & \cdots & E_{i0} \\ 0 & E_{i1}^\star & \cdots & E_{ir}^\star \end{pmatrix} \begin{pmatrix} \nu_{00} & \cdots & \nu_{0r} \\ \vdots & \ddots & \vdots \\ \nu_{r0} & \cdots & \nu_{rr} \end{pmatrix} \begin{pmatrix} 1 & 0 & E_{i0} & 0 \\ 1 & s & E_{i0} & E_{i1}^\star \\ \vdots & \vdots & \vdots & \vdots \\ 1 & sr & E_{i0} & E_{ir}^\star \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & s & \cdots & sr \\ E_{i0} & E_{i0} & \cdots & E_{i0} \\ 0 & E_{i1}^\star & \cdots & E_{ir}^\star \end{pmatrix} \cdot \\ &\quad \begin{pmatrix} \sum_{j=0}^r \nu_{0j} & s\sum_{j=0}^r j\nu_{0j} & E_{i0}\sum_{j=0}^r \nu_{0j} & \sum_{j=0}^r \nu_{0j}E_{ij}^\star \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{mj} & s\sum_{j=0}^r j\nu_{mj} & E_{i0}\sum_{j=0}^r \nu_{mj} & \sum_{j=0}^r \nu_{mj}E_{ij}^\star \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=0}^r \nu_{rj} & s\sum_{j=0}^r j\nu_{rj} & E_{i0}\sum_{j=0}^r \nu_{rj} & \sum_{j=0}^r \nu_{rj}E_{ij}^\star \end{pmatrix} \\ &= \begin{pmatrix} \omega_{00} & s\omega_{10} & E_{i0}\omega_{00} & \eta_1 \\ s\omega_{10} & s^2\omega_{11} & sE_{i0}\omega_{10} & s\eta_2 \\ E_{i0}\omega_{00} & sE_{i0}\omega_{10} & E_{i0}^2\omega_{00} & E_{i0}\eta_1 \\ \eta_1 & s\eta_2 & E_{i0}\eta_1 & \eta_3 \end{pmatrix} \end{aligned}$$

donde

$$\eta_1 := \sum_{k=0}^r \sum_{j=1}^r \nu_{jk}E_{ij}^\star, \quad \eta_2 := \sum_{k=1}^r \sum_{j=1}^r k\nu_{jk}E_{ij}^\star, \quad \eta_3 := \sum_{k=1}^r \sum_{j=1}^r \nu_{jk}E_{ij}^\star E_{ik}^\star \,.$$

Ahora, para $j \geqslant 1$,

$$\mathbb{E}_X\left[E_{ij}^\star\right] = \mathbb{E}_X\left[s\sum_{m=1}^j E_{im}\right] = s\sum_{m=1}^j \mathbb{E}_X\left[E_{im}\right] = p_e sj\,,$$

$$\begin{aligned}
\mathbb{E}_X\left[{E_{ij}^{\star}}^2\right] &= s^2\mathbb{E}_X\left[\left(\sum_{m=1}^{j} E_{im}\right)^2\right] = s^2\mathbb{E}_X\left[\sum_{l=1}^{j}\sum_{m=1}^{j} E_{im}E_{il}\right] \\
&= s^2\left(j\mathbb{E}_X\left[E_{ij}^2\right] + (j^2-j)\mathbb{E}_X\left[E_{ij}E_{ik\neq j}\right]\right) \\
&= s^2 j\left\{\mathbb{E}_X\left[E_{ij}\right] + (j-1)\left[\rho_e \mathrm{Var}(E_{ij}) + \left(\mathbb{E}_X\left[E_{ij}\right]\right)^2\right]\right\} \\
&= p_e s^2 j\left[1+(j-1)\lambda\right],
\end{aligned}$$

donde

$$\lambda := \rho_e(1-p_e) + p_e$$

y

$$\begin{aligned}
\mathbb{E}_X\left[E_{ij}^{\star}E_{ik\neq j}^{\star}\right] &= s^2\mathbb{E}_X\left[\left(\sum_{l=1}^{\mathrm{mín}(j,k)} E_{il}\right)\left(\sum_{m=1}^{\mathrm{máx}(j,k)} E_{im}\right)\right] \\
&= s^2\mathbb{E}_X\left[\left(\sum_{l=1}^{\mathrm{mín}(j,k)} E_{il}\right)^2 + \left(\sum_{l=1}^{\mathrm{mín}(j,k)} E_{il}\right)\left(\sum_{m=1+\mathrm{mín}(j,k)}^{\mathrm{máx}(j,k)} E_{im}\right)\right] \\
&= \mathbb{E}_X\left[{E_{i\ \mathrm{mín}(j,k)}^{\star}}^2\right] + s^2|j-k|\,\mathrm{mín}(j,k)\mathbb{E}_X\left[E_{ij}E_{ik\neq j}\right] \\
&= p_e s^2\,\mathrm{mín}(j,k)\left[1+(\mathrm{máx}(j,k)-1)\,\lambda\right].
\end{aligned}$$

Entonces,

$$\mathbb{E}_X\left[\eta_1\right] = \sum_{k=0}^{r}\sum_{j=1}^{r}\nu_{jk}\mathbb{E}_X\left[E_{ij}^{\star}\right] = p_e s\sum_{k=0}^{r}\sum_{j=1}^{r} j\nu_{jk} = p_e s\,\omega_{10},$$

$$\mathbb{E}_X\left[\eta_2\right] = \sum_{k=1}^{r}\sum_{j=1}^{r} k\nu_{jk}\mathbb{E}_X\left[E_{ij}^{\star}\right] = p_e s\sum_{k=1}^{r}\sum_{j=1}^{r} jk\nu_{jk} = p_e s\,\omega_{11}$$

y

$$\mathbb{E}_X\left[\eta_3\right] = \sum_{k=1}^{r}\sum_{j=1}^{r}\nu_{jk}\mathbb{E}_X\left[E_{ij}^{\star}E_{ik}^{\star}\right] = p_e s^2[(1-\lambda)\psi_1 + \lambda\omega_{11}],$$

donde

$$\psi_1 := \sum_{k=1}^{r}\sum_{j=1}^{r}\nu_{jk}\,\mathrm{mín}(j,k).$$

También,

$$\begin{aligned}\mathbb{E}_X\left[E_{i0}\eta_1\right] &= \sum_{k=0}^{r}\sum_{j=1}^{r}\nu_{jk}\mathbb{E}_X\left[E_{i0}E_{ij}^{\star}\right] = s\sum_{k=0}^{r}\sum_{j=1}^{r}\nu_{jk}\mathbb{E}_X\left[E_{i0}\sum_{m=1}^{j}E_{im}\right]\\ &= s\sum_{k=0}^{r}\sum_{j=1}^{r}\nu_{jk}\sum_{m=1}^{j}\mathbb{E}_X\left[E_{i0}E_{im}\right] = s\sum_{k=0}^{r}\sum_{j=1}^{r}j\nu_{jk}\mathbb{E}_X\left[E_{i0}E_{im\neq 0}\right]\\ &= p_e s\lambda\omega_{10}\end{aligned}$$

Entonces

$$\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right] = \begin{pmatrix} \omega_{00} & s\omega_{10} & p_e\omega_{00} & p_e s\omega_{10}\\ s\omega_{10} & s^2\omega_{11} & p_e s\omega_{10} & p_e s^2\omega_{11}\\ p_e\omega_{00} & p_e s\omega_{10} & p_e\omega_{00} & p_e s\lambda\omega_{10}\\ p_e s\omega_{10} & p_e s^2\omega_{11} & p_e s\lambda\omega_{10} & p_e s^2[(1-\lambda)\psi_1+\lambda\omega_{11}]\end{pmatrix}$$

y

$$\det(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]) = s^4p_e^2(1-p_e)^2(\omega_{00}\omega_{11}-\omega_{10}^2)\left\{\omega_{00}\left[(1-\rho_e)\psi_1+\rho_e\omega_{11}\right]-\rho_e^2\omega_{10}^2\right\}.$$

Nos interesa el elemento [4,4] de $\left(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right]\right)^{-1}$ que es

$$\frac{\begin{vmatrix} \omega_{00} & s\omega_{10} & p_e\omega_{00}\\ s\omega_{10} & s^2\omega_{11} & p_e s\omega_{10}\\ p_e\omega_{00} & p_e s\omega_{10} & p_e\omega_{00}\end{vmatrix}}{\det(\mathbb{E}_X\left[\mathbf{X}_i'\boldsymbol{\Sigma}^{-1}\mathbf{X}_i\right])} = \frac{\omega_{00}}{s^2p_e(1-p_e)\left\{\omega_{00}\left[(1-\rho_e)\psi_1+\rho_e\omega_{11}\right]-\rho_e^2\omega_{10}^2\right\}}$$

y entonces

$$\tilde{\sigma}^2 = \frac{r^2\omega_{00}}{p_e(1-p_e)\left\{\omega_{00}\left[(1-\rho_e)\psi_1+\rho_e\omega_{11}\right]-\rho_e^2\omega_{10}^2\right\}}.$$

Si la estructura de covarianza de la respuesta es CS(σ, ρ), entonces

$$\begin{aligned}\psi_1 &= \sum_{k=1}^{r}\sum_{j=1}^{r}\nu_{jk}\,\text{mín}(j,k) = \frac{1}{\sigma^2(1-\rho)(\rho r+1)}\left[(\rho r+1-\rho)\sum_{j=1}^{r}j - 2\rho\sum_{k=1}^{r-1}\sum_{j=1}^{m}j\right]\\ &= \frac{r(r+1)(\rho r+3-\rho)}{6\sigma^2(1-\rho)(\rho r+1)}\end{aligned}$$

de manera que

$$\tilde{\sigma}^2 = \frac{12\sigma^2(1-\rho)(\rho r+1)r}{p_e(1-p_e)(r+1)\left\{\rho\rho_e r^2+[2\rho+\rho_e+3(1-\rho)\rho_e(1-\rho_e)]r+2[1+(1-\rho_e)(2-\rho)]\right\}}.$$

Apéndice B

Adaptación de la función coste a datos faltantes por abandono de participantes durante el seguimiento

Hemos considerado un abandono monótono (2.10); es decir, que la falta de una medición en un determinado individuo implica la pérdida de sus subsiguientes medidas. Asumimos que no hay datos faltantes en la primera medida y que cada individuo que ha permanecido en el seguimiento hasta una determinada medida tiene una probabilidad π_m de abandonar el estudio antes de la siguiente medición. Así, la probabilidad y el coste correspondientes a cada uno de los $r+1$ patrones de datos faltantes es

# de medidas antes de la pérdida	Probabilidad del patrón	Coste total
1	π_m	c_1
2	$(1-\pi_m)\,\pi_m$	$c_1\left(1+\frac{1}{\kappa}\right)$
3	$(1-\pi_m)^2\,\pi_m$	$c_1\left(1+\frac{2}{\kappa}\right)$
$\vdots$	$\vdots$	$\vdots$
j	$(1-\pi_m)^{j-1}\,\pi_m$	$c_1\left(1+\frac{j-1}{\kappa}\right)$
$\vdots$	$\vdots$	$\vdots$
r	$(1-\pi_m)^{r-1}\,\pi_m$	$c_1\left(1+\frac{r-1}{\kappa}\right)$
$r+1$	$(1-\pi_m)^r$	$c_1\left(1+\frac{r}{\kappa}\right)$

de manera que el valor esperado del coste total por N participantes es

$$\begin{aligned}
\mathbb{E}(\text{Coste}) &= Nc_1 \left[\pi_m + \pi_m \sum_{j=1}^{r-1} (1-\pi_m)^j \left(1 + \frac{j}{\kappa}\right) + (1-\pi_m)^r \left(1 + \frac{r}{\kappa}\right) \right] \\
&= Nc_1 \left[\pi_m + \pi_m \sum_{j=1}^{r-1} (1-\pi_m)^j + (1-\pi_m)^r + \frac{\pi_m}{\kappa} \sum_{j=1}^{r-1} j(1-\pi_m)^j + \frac{r}{\kappa}(1-\pi_m)^r \right] \\
&= Nc_1 \left\{ 1 + \frac{(1-\pi_m)[1-(1-\pi_m)^{r-1}(1+\pi_m(r-1))]}{\pi_m \kappa} + \frac{r}{\kappa}(1-\pi_m)^r \right\} \\
&= Nc_1 \left\{ 1 + \frac{(1-\pi_m)[1-(1-\pi_m)^r]}{\pi_m \kappa} \right\} \\
&= Nc_1 \left\{ 1 + \frac{\pi_M \left\{ [1-(1-\pi_M)^{1/r}]^{-1} - 1 \right\}}{\kappa} \right\},
\end{aligned}$$

donde

$$\pi_M = 1 - (1-\pi_m)^r$$

es la probabilidad de que un participante no acabe el seguimiento.

Además, como cabe esperar, si no existe posibilidad de datos faltantes, entonces

$$\begin{aligned}
\lim_{\pi_m \to 0} \mathbb{E}(\text{Coste}) &= Nc_1 \lim_{\pi_m \to 0} \left\{ 1 + \frac{(1-\pi_m)[1-(1-\pi_m)^r]}{\pi_m \kappa} \right\} \\
&= Nc_1 \lim_{\pi_m \to 0} \left\{ 1 + \frac{(1-\pi_m)[1-(1-\pi_m r)]}{\pi_m \kappa} \right\} \\
&= Nc_1 \lim_{\pi_m \to 0} \left[1 + \frac{(1-\pi_m)r}{\kappa} \right] = Nc_1 \left(1 + \frac{r}{\kappa}\right),
\end{aligned}$$

recuperándose la expresión (2.1).

Apéndice C

Obtención de r_{opt} bajo el patrón de respuesta CMD en el escenario básico

Sea r_{opt} el valor de r que minimiza la función

$$f(r, \kappa, \rho, \rho_e) = \frac{(\kappa + r)(1 + \rho\, r)}{(r+1)(1 - \rho + \rho(1 - \rho_e)r)}\,, \tag{C.1}$$

con $r \in \mathbb{N}$, $\kappa \geqslant 1$, $\rho \in (0, 1)$ y $\rho_e \in (0, 1]$.

C.1. Exposición constante ($\rho_e = 1$)

Si $\rho_e = 1$, (C.1) se reduce a

$$f(r, \kappa, \rho, \rho_e) = \frac{(\kappa + r)(1 + \rho\, r)}{(1 - \rho)(r + 1)}$$

y, equivalentemente, podemos minimizar la función

$$g(r, \kappa, \rho, \rho_e) = \frac{(\kappa + r)(1 + \rho\, r)}{r + 1}\,. \tag{C.2}$$

Si $\kappa = 1$, (C.2) se reduce a

$$g(r, \kappa, \rho, \rho_e) = 1 + \rho\, r$$

que es monótona creciente y entonces $r_{\text{opt}} = 0$.

Si $\kappa > 1$, el mínimo de (C.2) se encuentra en $r_0 = -1 + \sqrt{\frac{1-\rho}{\rho}(\kappa - 1)}$ que es positivo si $\kappa > \frac{1}{1-\rho}$.

Entonces, para una exposición constante:

$$r_{\text{opt}} = \begin{cases} 0 & , \quad \kappa \leqslant \frac{1}{1-\rho} \\ -1 + \sqrt{\frac{1-\rho}{\rho}(\kappa - 1)} & , \quad \kappa > \frac{1}{1-\rho} \end{cases} .$$

En general, r_0 es irracional. En tal caso, r_{opt} será $[r_0]$ o $[r_0] + 1$ ($[\cdot]$ es la función parte entera), dependiendo de cuál proporcione un menor valor de $g(r, \kappa, \rho, \rho_e)$. El paquete optimalAllocationproporciona efectivamente el óptimo.

C.2. Exposición variable en el tiempo ($\rho_e < 1$)

La forma de (C.1) cambia dependiendo de los valores de κ, ρ y ρ_e. Por tanto, consideraremos los diferentes escenarios.

C.2.1. Caso $\kappa = 1$

- Si $\kappa = 1$ y $\rho_e = \rho$, (C.1) se reduce a la constante $f(r, \kappa, \rho, \rho_e) = \frac{1}{1-\rho}$, independiente de r y por tanto, por simplicidad del diseño del estudio, fijamos $r_{\text{opt}} = 0$.

- Si $\kappa = 1$ y $\rho_e \neq \rho$, (C.1) se reduce a

$$f(r, \kappa, \rho, \rho_e) = \frac{1 + \rho\, r}{1 - \rho + \rho(1 - \rho_e) r}$$

y

$$\frac{\partial f(r, \kappa, \rho, \rho_e)}{\partial r} = \frac{\rho(\rho_e - \rho)}{[1 - \rho + \rho(1 - \rho_e) r]^2} ,$$

de manera que $f(r, \kappa, \rho, \rho_e)$ es monótona creciente si $\rho_e > \rho$ y monótona decreciente si $\rho_e < \rho$, para todo valor positivo de r. Así, $r_{\text{opt}} = 0$ si $\rho_e > \rho$ y $r_{\text{opt}} = +\infty$ si $\rho_e < \rho$.

C.2.2. Caso $\kappa > 1$

- Si $\rho_e = \rho$, (C.1) se reduce a

$$f(r, \kappa, \rho, \rho_e) = \frac{r + \kappa}{(1 - \rho)(r + 1)}$$

y

$$\frac{\partial f(r, \kappa, \rho, \rho_e)}{\partial r} = -\frac{\kappa - 1}{(1 - \rho)(r + 1)^2} < 0 \quad \forall r \neq -1 ,$$

de manera que $f(r, \kappa, \rho, \rho_e)$ es monótona decreciente y $r_{\text{opt}} = +\infty$.

- Si $\rho_e \neq \rho$, la estructura de $f(r, \kappa, \rho, \rho_e)$ varía dependiendo del valor del parámetro

$$\kappa^\star := \frac{1-\rho}{\rho(1-\rho_e)} > 0.$$

  - Si $\kappa = \kappa^\star > 1$ y $\rho_e \neq \rho$, (C.1) se reduce a

$$f(r; \rho, \kappa, \kappa^\star) = \frac{\rho\, r + 1}{\rho(1-\rho_e)(r+1)}$$

    y

$$\frac{\partial f(r; \rho, \kappa, \kappa^\star)}{\partial r} = -\frac{1-\rho}{\rho(1-\rho_e)(r+1)^2} < 0 \quad \forall r \neq -1\,,$$

    de manera que $f(r)$ es monótona decreciente y $r_{\text{opt}} = +\infty$.

  - Si $\kappa > 1$, $\kappa \neq \kappa^\star$ y $\rho_e \neq \rho$ (el caso general y más común), (C.1) puede expresarse como

$$f(r; \rho, \kappa, \kappa^\star) = \frac{\kappa^\star}{1-\rho} \cdot \frac{(r+\kappa)(1+\rho\, r)}{(r+1)(r+\kappa^\star)} \tag{C.3}$$

    y

$$\frac{\partial f(r;\rho,\kappa,\kappa^\star)}{\partial r} = \frac{\kappa^\star}{1-\rho} \cdot \frac{[\rho(\kappa^\star-\kappa)-(1-\rho)]r^2+2(\rho\kappa^\star-\kappa)r+[(\kappa^\star-\kappa)-\kappa\kappa^\star(1-\rho)]}{(r+1)^2(r+\kappa^\star)^2} \tag{C.4}$$

    y, por tanto, resolver

$$\frac{\partial f(r; \rho, \kappa, \kappa^\star)}{\partial r} = 0$$

    es equivalente a resolver la ecuación

$$[\rho(\kappa^\star - \kappa) - (1-\rho)]r^2 + 2(\rho\kappa^\star - \kappa)r + [(\kappa^\star - \kappa) - \kappa\kappa^\star(1-\rho)] = 0\,. \tag{C.5}$$

    Si $\rho(\kappa^\star - \kappa) - (1-\rho)$ es igual a cero (equivalentemente, $\kappa = \frac{\rho_e(1-\rho)}{\rho(1-\rho_e)}$), entonces

$$\frac{\partial f(r; \rho, \kappa, \kappa^\star)}{\partial r} = -(\kappa-1)\frac{2r + \kappa^\star(\kappa^\star+1)}{(r+1)^2(r+\kappa^\star)^2} < 0 \quad \forall r \notin \{-\kappa^\star, -1\}$$

    y $r_{\text{opt}} = +\infty$.

    Consideremos ahora $\rho(\kappa^\star - \kappa) - (1-\rho) \neq 0$. El discriminante de (C.5) puede expresarse como

$$\Delta = 4(1-\rho)(\kappa-1)(\kappa-\kappa^\star)(1-\rho\kappa^\star)$$

    y el signo de Δ es igual al signo de $(\kappa - \kappa^\star)(1-\rho\kappa^\star)$. Como $\kappa \neq \kappa^\star$ y $\rho\kappa^\star \neq 1$ (consecuencia de $\rho_e \neq \rho$), Δ no puede ser cero. Ahora, analizamos todas las combinaciones de signos de $(\kappa - \kappa^\star)$ y $(1-\rho\kappa^\star)$:

- $1 - \rho\kappa^\star < 0$ (equivalentemente, $\rho_e > \rho$) y $\kappa > \kappa^\star$:
  En este caso, $\Delta < 0$ y por tanto $f(r)$ es monótona $\forall r \notin \{-\kappa^\star, -1\}$. Entonces, dado que según (C.4),

$$\frac{\partial f(r;\rho,\kappa,\kappa^\star)}{\partial r}\Big|_{r=0} = -\frac{(\kappa - \kappa^\star) + \kappa\kappa^\star(1-\rho)}{(1-\rho)\kappa^\star} < 0\,,$$

  $f(r)$ es monótona decreciente $\forall r \notin \{-\kappa^\star, -1\}$ y $r_{\text{opt}} = +\infty$.

- $1 - \rho\kappa^\star > 0$ (equivalentemente, $\rho_e < \rho$) y $\kappa < \kappa^\star$:
  En este caso, $\Delta < 0$ y por tanto $f(r)$ es monótona $\forall r \notin \{-\kappa^\star, -1\}$. Entonces, dado que según (C.4),

$$\lim_{r\to+\infty} \frac{\partial f(r;\rho,\kappa,\kappa^\star)}{\partial r} = \lim_{r\to+\infty} \frac{\kappa^\star[\rho(\kappa^\star - \kappa) - (1-\rho)]}{(1-\rho)r^2}\,,$$

  $f(r)$ es monótona decreciente $\forall r \notin \{-\kappa^\star, -1\}$, y entonces $r_{\text{opt}} = +\infty$, dado que $\rho(\kappa^\star - \kappa) - (1-\rho) < 0$, como se muestra a continuación:

$$1 - \rho\kappa^\star > 0 \Rightarrow 1 - \rho\kappa^\star > \rho(1-\kappa) \Rightarrow \rho(\kappa^\star - \kappa) - (1-\rho) < 0\,.$$

- $1 - \rho\kappa^\star > 0$ (equivalentemente, $\rho_e < \rho$) y $\kappa > \kappa^\star$:
  En este caso, $\Delta > 0$ y las dos soluciones de (C.5) son

$$r_\pm = \frac{\pm\frac{\sqrt{\Delta}}{2} - (\kappa - \rho\kappa^\star)}{(1-\rho) + \rho(\kappa - \kappa^\star)}$$

  cuyo denominador es positivo y entonces $r_+ > r_-$. Además, como $\kappa > \kappa^\star$, entonces $r_- < 0$.
  De (C.4),

$$\frac{\partial^2 f(r;\rho,\kappa,\kappa^\star)}{\partial r^2}\Big|_{r=r_\pm} = \mp\frac{\kappa^\star\sqrt{\Delta}}{(1-\rho)(r_\pm + 1)^2(r_\pm + \kappa^\star)^2}\,, \tag{C.6}$$

  y existe un mínimo local en $r_- < 0$ y un máximo local en r_+, de manera que $f(r)$ es monótona decreciente para todo $r > r_+$. Podemos probar que $r_+ < 0$ y entonces $f(r)$ is decreciente en $[0, +\infty)$ y, por tanto, $r_{\text{opt}} = +\infty$. Para probar que $r_+ < 0$, solo necesitamos probar

que el numerador de r_+ es negativo:

$$\begin{aligned}\frac{\sqrt{\Delta}}{2} - (\kappa - \rho\kappa^\star) &= \sqrt{(1-\rho)(\kappa-1)(\kappa-\kappa^\star)(1-\rho\kappa^\star)} - (\kappa - \rho\kappa^\star) \\ &< \sqrt{(1-\rho)(\kappa-1)(\kappa-\rho\kappa^\star)(1-\rho\kappa^\star)} - (\kappa - \rho\kappa^\star) \\ &< \sqrt{(1-\rho)(\kappa-\rho\kappa^\star)(\kappa-\rho\kappa^\star)(1-\rho\kappa^\star)} - (\kappa - \rho\kappa^\star) \\ &= (\kappa - \rho\kappa^\star)\left[\sqrt{(1-\rho)(1-\rho\kappa^\star)} - 1\right] \\ &< (\kappa - \rho\kappa^\star)\,[1-1] = 0\,.\end{aligned}$$

- $1 - \rho\kappa^\star < 0$ (equivalentemente, $\rho_e > \rho$) y $\kappa < \kappa^\star$:
  En este caso, $\Delta > 0$ y las dos soluciones de (C.5) pueden expresarse como

$$r_\pm = \frac{\pm\frac{\sqrt{\Delta}}{2} + (\rho\kappa^\star - \kappa)}{(1-\rho) - \rho(\kappa^\star - \kappa)} \tag{C.7}$$

  y, de (C.6), existe un mínimo local en $r_- < 0$ y un máximo local en r_+, de manera que $f(r)$ es monótona decreciente para todo $r > r_+$. Pero ahora el denominador de (C.7) puede ser positivo o negativo, y debemos considerar las dos posibilidades.
  Si el denominador de (C.7) es positivo, entonces $r_+ > r_-$ y $\kappa^\star < \kappa + \frac{1-\rho}{\rho}$, de manera que

$$\begin{aligned}\frac{\sqrt{\Delta}}{2} + (\rho\kappa^\star - \kappa) &= \sqrt{(1-\rho)(\kappa-1)(\kappa^\star-\kappa)(\rho\kappa^\star-1)} + \rho\kappa^\star - \kappa \\ &< \sqrt{(1-\rho)(\kappa-1)\left(\tfrac{1-\rho}{\rho}\right)(\rho\kappa-\rho)} + \rho\kappa + 1 - \rho - \kappa \\ &= \sqrt{(1-\rho)^2(\kappa-1)^2} - (1-\rho)(k-1) = 0\end{aligned}$$

  y entonces $r_+ < 0$, de manera que $f(r)$ es monótona decreciente para todo r positivo, y $r_{\text{opt}} = +\infty$.
  Si el denominador de (C.7) es negativo, entonces $r_+ < r_-$ y $\kappa^\star > \kappa + \frac{1-\rho}{\rho}$. En este caso, el signo de r_- no es constante. Por ejemplo, si tenemos $\rho = 0{,}55$ y $\rho_e = 0{,}80$, si $\kappa = 1{,}3$, el mínimo local está en $r_- \approx -0{,}24$, y $r_{\text{opt}} = 0$ mientras que si $\kappa = 2{,}6$, el mínimo local está en $r_- \approx 4{,}08$, y $r_{\text{opt}} = 4$ o $r_{\text{opt}} = 5$. Entonces,

$$r_{\text{opt}} = \text{máx}\left(0, \frac{\sqrt{(1-\rho)(\kappa-1)(\kappa^\star-\kappa)(\rho\kappa^\star-1)} - \rho\kappa^\star + \kappa}{\rho(\kappa^\star-\kappa) - (1-\rho)}\right).$$

  Además, se puede probar que $\frac{\sqrt{(1-\rho)(\kappa-1)(\kappa^\star-\kappa)(\rho\kappa^\star-1)} - \rho\kappa^\star + \kappa}{\rho(\kappa^\star-\kappa)-(1-\rho)} < 0$ si

$\kappa \leqslant \frac{\rho\kappa^\star}{1+(1-\rho)\rho\kappa^\star}$ y, por tanto,

$$r_{\text{opt}} = \begin{cases} 0 & , \text{ si } \kappa \leqslant \frac{\rho\kappa^\star}{1+(1-\rho)\rho\kappa^\star} \\ \frac{\sqrt{(1-\rho)(\kappa-1)(\kappa^\star-\kappa)(\rho\kappa^\star-1)}-\rho\kappa^\star+\kappa}{\rho(\kappa^\star-\kappa)-(1-\rho)} & , \text{ en otro caso} \end{cases} .$$

C.3. Resumen de los resultados

Los resultados quedan resumidos en la Tabla C.1.

Tabla C.1: Número total de medidas óptimo bajo el patrón de respuesta CMD en el escenario básico.

Exposición variable en el tiempo ($\rho_e < 1$)		
$\boldsymbol{\rho_e}$	$\boldsymbol{\kappa}$	**Número total de medidas óptimo**
$\rho_e \geq \rho$	1	1
$\rho_e > \rho$	$\kappa \leq \kappa_0$	1
	$\kappa \in (\kappa_0, \kappa_c)$	$1 + r_0$
otras combinaciones de (κ, ρ, ρ_e)		$+\infty$

donde:

$\kappa_0 := \frac{1-\rho}{1-\rho_e+(1-\rho)^2}$, $\kappa_c := \frac{\rho_e(1-\rho)}{\rho(1-\rho_e)}$

$r_0 := \frac{\kappa-1-\rho(\kappa_c-1)+\sqrt{(1-\rho)(\kappa_c-1)(\kappa-1)[1-\rho+\rho(\kappa_c-\kappa)]}}{\rho(\kappa_c-\kappa)}$

Exposición constante ($\rho_e = 1$)	
$\boldsymbol{\kappa}$	**Número total de medidas óptimo**
$\kappa \leqslant \frac{1}{1-\rho}$	1
en otro caso	$\sqrt{\frac{1-\rho}{\rho}(\kappa-1)}$

Apéndice D

Obtención de r_{opt} bajo el patrón de respuesta LDD en el escenario básico

Sea r_{opt} el valor de r que minimiza la función

$$f(r,\kappa,\rho,\rho_e) = \frac{r(r+\kappa)(\rho r+1)}{(r+1)[\rho\rho_e r^2+[2\rho+\rho_e+3(1-\rho)\rho_e(1-\rho_e)]r+2[1+(2-\rho)(1-\rho_e)]]}\,, \tag{D.1}$$

con $r \in \mathbb{N}^+$, $\kappa \geqslant 1$, $\rho \in (0,1)$ y $\rho_e \in (0,1]$.

D.1. Exposición constante ($\rho_e = 1$)

Si $\rho_e = 1$, (D.1) se reduce a

$$f(r,\kappa) = \frac{r(r+\kappa)}{(r+1)(r+2)} \tag{D.2}$$

que no depende de ρ.

- Si $\kappa = 1$ o $\kappa = 2$, (D.2) se reduce a

  $$f(r,\kappa) = \frac{r}{r+3-\kappa}$$

  y

  $$\frac{\partial f(r,\kappa)}{\partial r} = \frac{3-\kappa}{(r+3-\kappa)^2} > 0 \quad \forall r \notin \{-2,-1\}\,,$$

  de manera que $f(r,\kappa)$ es monótona creciente y $r_{\text{opt}} = 1$.

- Si $\kappa \notin \{1,2\}$,

  $$\frac{\partial f(r,\kappa)}{\partial r} = \frac{(3-\kappa)r^2+4r+2\kappa}{(r+1)^2(r+2)^2} \tag{D.3}$$

y resolver

$$\frac{\partial f(r,\kappa)}{\partial r} = 0$$

equivale a resolver la ecuación

$$(3-\kappa)r^2 + 4r + 2\kappa = 0\,. \tag{D.4}$$

- Si $\kappa = 3$,

$$\frac{\partial f(r,\kappa)}{\partial r} = \frac{4r+2\kappa}{(r+1)^2(r+2)^2} > 0 \quad \forall r \geqslant 0\,,$$

de manera que $f(r,\kappa)$ es monótona creciente y $r_{\text{opt}} = 1$.

- Si $\kappa \neq 3$, la solución de (D.4) es

$$r_\pm = \frac{-2 \pm \sqrt{2(\kappa-1)(\kappa-2)}}{3-\kappa}\,. \tag{D.5}$$

Si $\kappa < 2$, el discriminante de (D.5) es negativo y por tanto $f(r,\kappa)$ es monótona. Y, dado que de (D.3) $\frac{\partial f(r,\kappa)}{\partial r}$ es positivo en $r = 0$, $f(r,\kappa)$ es monótona creciente y $r_{\text{opt}} = 1$.

Si $\kappa > 2$, ambos valores de (D.5) son reales.

Si $\kappa \in (2,3)$, de (D.5) tenemos

$$r_- < r_+ = \frac{-2+\sqrt{2(\kappa-1)(\kappa-2)}}{3-\kappa} < \frac{-2+\sqrt{2\cdot 2\cdot 1}}{3-\kappa} = 0\,,$$

de manera que $f(r,\kappa)$ es monótona creciente para valores no negativos de r y por tanto $r_{\text{opt}} = 1$.

Si $\kappa > 3$, $r_+ < 0$ y $r_- > 0$, de manera que solo necesitamos analizar r_-. Derivando (D.3) y dado que $r_\pm$ es solución de (D.4), tenemos

$$\left.\frac{\partial^2 f(r,\kappa)}{\partial r^2}\right|_{r=r_\pm} = \pm\frac{2\sqrt{2(\kappa-1)(\kappa-2)}}{(r_\pm+1)^2(r_\pm+2)^2} \tag{D.6}$$

que es negativo para $r = r_-$, presentando $f(r,\kappa)$ un máximo local en $r = r_-$ y $f(r,\kappa)$ es creciente para $0 \leqslant r < r_-$ y decreciente para $r > r_-$. Por otro lado, puede probarse fácilmente que r_- es decreciente en κ para $\kappa > 3$ y que $\lim_{\kappa\to+\infty} r_- = \sqrt{2}$, de manera que $r_- > 1$ para todo $\kappa > 3$. Por tanto, $r_{\text{opt}} = 1$ si $f(r=1,\kappa) \leqslant \lim_{r\to+\infty} f(r,\kappa)$ (equivalentemente, si $\kappa \leqslant 5$, de (D.2)) y $r_{\text{opt}} = +\infty$ en caso contrario.

En resumen, para una exposición constante:

$$r_{\text{opt}} = \begin{cases} 1 & , \quad \kappa \leqslant 5 \\ +\infty & , \quad \kappa > 5 \end{cases} .$$

D.2. Exposición variable en el tiempo ($\rho_e < 1$)

D.2.1. Caso $\kappa = 1$

En este caso, (D.1) se reduce a:

$$f(r;\kappa,\rho,\rho_e) = \tfrac{r(\rho\, r+1)}{\rho\rho_e r^2+[2\rho+\rho_e+3(1-\rho)\rho_e(1-\rho_e)]r+2[1+(2-\rho)(1-\rho_e)]} . \tag{D.7}$$

Nótese que al ser positivos los tres coeficientes del polinomio de grado 2 en r del denominador de $f(r)$, $f(r;\kappa,\rho,\rho_e)$ es continua para todo r positivo.

Por otro lado,

$$\frac{\partial f(r;\kappa,\rho,\rho_e)}{\partial r} = \tfrac{\rho[2\rho+3(1-\rho)\rho_e(1-\rho_e)]r^2+4\rho[1+(2-\rho)(1-\rho_e)]r+2[1+(2-\rho)(1-\rho_e)]}{[\rho\rho_e r^2+[2\rho+\rho_e+3(1-\rho)\rho_e(1-\rho_e)]r+2[1+(2-\rho)(1-\rho_e)]]^2} . \tag{D.8}$$

Nótese ahora que los seis coeficientes en la función racional (D.8) son positivos. Por tanto, $\frac{\partial f(r;\kappa,\rho,\rho_e)}{\partial r}$ es positivo para todo r positivo y así $f(r;\kappa,\rho,\rho_e)$ es monótona creciente para todo r positivo, siendo $r_{\text{opt}} = 1$.

D.2.2. Caso $\kappa > 1$

Consideraremos los dos casos complementarios y excluyentes: $\rho_e = \frac{2\rho(2-\rho)}{3(1-\rho)}$ y $\rho_e \neq \frac{2\rho(2-\rho)}{3(1-\rho)}$.

- Caso $\rho_e = \frac{2\rho(2-\rho)}{3(1-\rho)}$ (que implica $\rho < \frac{1}{2}$):

  En este caso, ignorando factores positivos constantes, (D.1) se reduce a

  $$f(r;\kappa,\rho,\rho_e) = \frac{r(r+\kappa)}{(r+1)(r+\alpha)}, \qquad \alpha := \tfrac{9-20\rho+11\rho^2-2\rho^3}{\rho(2-\rho)} \in (+2,+\infty) . \tag{D.9}$$

  Si $\kappa = \alpha$, $f(r;\kappa,\rho,\rho_e) = \frac{r}{r+1}$ que es monótona creciente para todo r positivo y $r_{\text{opt}} = 1$.

  Consideramos ahora $\kappa \neq \alpha$. Entonces,

  $$\frac{\partial f(r;\kappa,\rho,\rho_e)}{\partial r} = \frac{(\alpha+1-\kappa)r^2+2\alpha r+\alpha\kappa}{(r+1)^2(r+\alpha)^2} . \tag{D.10}$$

  Si $\kappa \leqslant \alpha+1$, entonces $f(r;\kappa,\rho,\rho_e)$ es monótona creciente para todo r positivo, y $r_{\text{opt}} = 1$.

Si $\kappa > \alpha + 1$, las raíces de (D.10) son

$$r_\pm = \frac{\alpha \pm \sqrt{\Delta}}{\kappa - (\alpha+1)}, \quad \text{donde} \quad \Delta = \alpha(\kappa-1)(\kappa-\alpha) > \alpha^2 \tag{D.11}$$

que implica $r_- < 0$ y solo necesitamos analizar r_+, donde hay un máximo local ya que

$$\left.\frac{\partial^2 f(r,\kappa,\rho,\rho_e)}{\partial r^2}\right|_{r=r_+} = \frac{-2\sqrt{\Delta}}{(r_+ + 1)^2(r_+ + \alpha)^2} < 0\,.$$

Puede probarse que $r_+ > 1$ y, por tanto, $f(r;\kappa,\rho,\rho_e)$ es creciente para $r \in [1, r_+)$ y decreciente para $r \in (r_+, +\infty)$. Así, $r_{\text{opt}} = 1$ si $f(r=1;\kappa,\rho,\rho_e) \leqslant \lim_{r\to+\infty} f(r;\kappa,\rho,\rho_e)$ (equivalentemente, si $\kappa \leqslant 2\alpha + 1$) y $r_{\text{opt}} = +\infty$ en caso contrario.

Resumiendo,

$$\rho_e = \tfrac{2\rho(2-\rho)}{3(1-\rho)} \quad \Rightarrow \quad r_{\text{opt}} = \begin{cases} 1 & , \quad \kappa \leqslant \kappa^* \\ +\infty & , \quad \kappa > \kappa^* \end{cases} \quad , \qquad \kappa^* := \tfrac{18-38\rho+21\rho^2-4\rho^3}{\rho(2-\rho)}\,.$$

- Caso $\rho_e \neq \frac{2\rho(2-\rho)}{3(1-\rho)}$ (el caso más general y común):

  En este caso, la optimización analítica de (D.1) es prácticamente imposible dado que requiere resolver una ecuación polinómica de grado 4 con expresiones complejas para sus coeficientes, que dependen de ρ, ρ_e y κ. Sobre una parrilla fina de valores de los parámetros $\rho \in (0,1)$, $\rho_e \in (0,1)$ y $\kappa \in (1, 10000)$, encontramos que siempre era $r_{\text{opt}} = 1$ o $r_{\text{opt}} = +\infty$. Asumiendo que el valor de r_{opt} puede ser solo 1 o $+\infty$, y dado que de (D.1), $f(r=1) = \frac{(\kappa+1)(1+\rho)}{6[2-(1-\rho)\rho_e^2]}$ y $\lim_{r\to+\infty} f(r) = \frac{1}{\rho_e}$, tenemos que

$$r_{\text{opt}} = \begin{cases} 1 & , \quad \kappa \leqslant \kappa^* \\ +\infty & , \quad \kappa > \kappa^* \end{cases} \quad , \qquad \kappa^\star := 5 + \tfrac{6(1-\rho_e)[2+(1-\rho)\rho_e]}{(1+\rho)\rho_e} \tag{D.12}$$

que es coherente con el resto de casos analizados en esta sección.

Además, dado que

$$\frac{\partial \kappa^*}{\partial \rho} = -\frac{12(1-\rho_e^2)}{(1+\rho)\rho_e} < 0 \quad \text{y} \quad \frac{\partial \kappa^*}{\partial \rho_e} = -\frac{6[2+(1-\rho_e)\rho^2]}{(1+\rho)\rho_e^2} < 0\,,$$

κ^* es decreciente en ρ y en ρ_e.

D.3. Resumen de los resultados

Los resultados se resumen en la Tabla D.1. En general, debemos tomar solo una medida si κ no es mayor que el umbral $\frac{6[2-(1-\rho)\rho_e^2]}{(1+\rho)\rho_e} - 1$, que es decreciente en ρ y en ρ_e.

Tabla D.1: Número total de medidas óptimo bajo el patrón de respuesta LDD en el escenario básico.

Exposición variable en el tiempo ($\rho_e < 1$)	
$\boldsymbol{\kappa}$	**Número total de medidas óptimo**
$\kappa \leqslant \frac{6[2-(1-\rho)\rho_e^2]}{(1+\rho)\rho_e} - 1$	2
en caso contrario	$+\infty$

Exposición constante ($\rho_e = 1$)	
$\boldsymbol{\kappa}$	**Número total de medidas óptimo**
$\kappa \leqslant 5$	2
en caso contrario	$+\infty$

Apéndice E

Publicación de los resultados en congresos, revistas y conferencias

Los resultados de este trabajo han sido divulgados, total o parcialmente, a través de las actividades siguientes:

- Participación en la **XIII Conferencia Española y III Encuentro Iberoamericano de Biometría - 2011**, celebrada en Barcelona, en septiembre de 2011:

  Presentación oral: *Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure.*

- Envío a la revista **Statistics in Medicine**, en junio de 2012:

  Artículo: *Optimal combination of number of participants and number of repeated measurements in longitudinal studies with time-varying exposure.*

- Participación en el ciclo de conferencias y talleres **Els dissabtes de les matemàtiques**, organizadas por la Universitat Autònoma de Barcelona, en abril de 2012:

  Conferencia y taller: *Matemàtiques a l'Epidemiologia*[1].

- Diseño y dirección del Trabajo de Fin de Grado titulado *Optimització de la potència en estudis longitudinals observacionals*, basado en resultados parciales del presente documento, y que actualmente está realizando una estudiante del Grado de Estadística Aplicada de la Universitat Autònoma de Barcelona.

[1] `http://www.uab.cat/servlet/Satellite/divulgacio/dissabtes-de-les-matematiques-1195630210586.html`