Capítulo 3

Diseños estáticos

En este capítulo, centraremos la atención en los experimentos diseñados, en los que supondremos que el investigador posee total control sobre qué valores tomarán los factores a lo largo de la experimentación[1]*. Puntualizamos también los aspectos principales a tener en cuenta a la hora de elaborar 'buenos diseños', discutiendo las extensiones de los diseños para datos normales cuando se consideran otras distribuciones de la respuesta. Conduciremos los principios de los diseños de modo que desemboquen en aquellos casos en que la respuesta tenga una distribución binaria. El enfoque que daremos será más del tipo "estático" que "dinámico", dejando este segundo aspecto para el capítulo siguiente.*

3.1. Introducción

Al iniciar un estudio en el que intervengan datos, entre los cuales se presuma que existe una relación funcional que pueda vincular una cierta variable respuesta con un conjunto de variables explicativas, es posible enfocar el mismo al menos desde dos puntos de vista:

a. *Estudios observacionales:* en este tipo de estudios, las observaciones se realizan sin seguir algún patrón específico de medición de las cantidades que intervienen. En particular, esto cobra importancia en cuanto a los valores que puedan tomar los factores de variabilidad, los cuales en muchas ocasiones simplemente "toman los valores que tocan", sin seguir un criterio específico. Un ejemplo de esto es cuando se buscan correlaciones entre datos de una tabla de registro de lluvias y algún conjunto de factores meteorológicos particular (p. ej.: presión atmosférica, temperatura de bulbo seco, etc.).

[1] En este trabajo no abordaremos las connotaciones del tipo ético y moral de las que serían susceptibles estudios sociales o biomédicos, puesto que excede los alcances de los objetivos propuestos.

b. *Diseños experimentales:* en esta clase de experimentos, y partiendo de consideraciones estadísticas precisas, los valores que toman los factores son controlados por el experimentador, de forma tal que los resultados a los que se llega al analizar los datos contendrán un mayor volumen de información acerca de la respuesta y su relación con los factores. Cabe remarcar que no siempre es posible tener bajo total control a los niveles de los factores bajo estudio, ya sea por motivos morales, éticos, técnicos o económicos. La calidad de información que puede derivarse de este tipo de estudios es significativamente mayor que si no se hubieran diseñado aquellos. Citando a Cox y a Reid[2], diremos que *"el objetivo del diseño es hacer que tanto el análisis como la interpretación de los resultados sean de la forma más simple y clara posible"*.

En la práctica, y dadas estas limitaciones de los experimentos diseñados, muchas veces no quedará más remedio que trabajar con estudios observacionales. Sin embargo, al trabajar con datos observados del pasado, el experimentador se expone a ciertos riesgos que pueden llevar a considerar más adelante conclusiones confusas o incluso erróneas. A este respecto, pueden verse una serie de recomendaciones que se sugieren en Prat *et al.* (1997 y 2004), capítulo 7.

A modo de síntesis, podemos indicar que a la hora de estudiar un conjunto de características variables en una muestra de individuos[3], éstas podrán ya venir medidas en aquéllos, en cuyo caso hablaremos de *estudios observacionales*, o bien será el experimentador quien decidirá de qué forma y con qué criterios se le asignan valores a dichas características, que conducirá a los *diseños experimentales*.

3.2. El enfoque estático del problema

Como hemos dicho anteriormente, a medida que vamos experimentando con los niveles de las variables de un sistema que nos interesa estudiar y conocer un poco más, el enfoque secuencial de avance del conocimiento irá atravesando etapas a medida que vayamos incorporando nueva información de las etapas anteriores. Cada etapa que se vaya atravesando tendrá una configuración especial en cuanto a *diseño* se refiera, ya que los niveles que tomarán los factores estarán definidos de antemano por el experimentador. En palabras de Myers[4], el problema de cada etapa dentro del diseño será el de *"decidir qué combinación de niveles deberían utilizarse en cada etapa, de modo*

[2] *Vid.* Cox y Reid (2000), p. 6.

[3] Esto da una idea del concepto de *unidad experimental*, en cuanto a individuos provenientes de una población determinada y de los que se disponen medidas de alguna característica variable entre ellos.

[4] *Vid.* Myers (1976), p. 107.

que las estimaciones de los parámetros resulte con máxima precisión", lo cual complementa adecuadamente el punto de vista aportado por Cox y Reid que citamos en la subsección anterior.

Al quedar definidas las condiciones sobre las cuales se realizarán las observaciones de la respuesta, diremos que el diseño —en cuanto a etapas en la sucesión de avance del conocimiento que se tenga de un sistema— proporcionará un enfoque del tipo *estático*, el cual puede asociarse a un fotógrafo que prepara el objeto del cual tomará imágenes, las que luego revelará y analizará, lo cual constituye una segunda fase del estudio, que será la del *análisis* e *interpretación* de los resultados.

Pero la sola utilización de este enfoque no permitirá conocer con suficiente eficiencia si existen otras condiciones —también *estáticas* cada una de ellas— mejores que las actuales que lleven a poder observar más claramente si el diseño admite "mejores" va-lores aún para los factores, definiendo ellos mediante nuevas condiciones experimentales. Con esto, intentamos destacar que el proceso de avance secuencial del conocimiento del sistema deberá estar enmarcado dentro de un enfoque del tipo *dinámico*, encadenando sucesivos enfoques de naturaleza *estática*. Dentro de este mismo contexto secuencial de experimentación, sería deseable que esta integración de enfoques tuviese al menos las siguientes características:

- que esté formado por una sucesión de etapas, que tendrán cada una un enfoque estático, obviamente,
- que se trate de atravesar un número relativamente no muy grande de etapas hasta llegar a cumplir los objetivos del estudio,
- que para pasar de una etapa anterior a la siguiente, se tenga en cuenta la mayor cantidad de información obtenida en la etapa anterior.

La consideración de ciertos tipos adecuados de diseños de experimentos, como así también la utilización de criterios estadísticos para decidir las condiciones de experimentación de las etapas sucesivas, conforman la columna vertebral de la *MSR*. Como indicamos anteriormente, cuando la respuesta sigue una distribución normal, el problema ya ha sido resuelto, mientras que para datos binarios, no hemos encontrado todavía una presentación clara del problema y sus alternativas de solución de forma completa, comenzando por la primera etapa y llegando a la última, de modo análogo a la teoría clásica de *MSR*.

En este capítulo, hablaremos resumidamente sobre los distintos modelos estáticos que disponemos para analizar datos binarios, es decir, describiremos las formas básicas de diseñar experimentos, y cuyas conclusiones intentaremos entrelazar luego en

el siguiente capítulo, en los que consideraremos los diseños encadenados entre sí, que conformarán el *enfoque dinámico* del que hablamos anteriormente. Mediante el estudio de lo que ocurre entre las etapas, es decir, entre los distintos diseños que vayan ensayándose, pretendemos hacer un aporte que permita establecer algunas bases sobre las cuales poder construir la *MSR* para datos binarios, que es el propósito último de esta tesis.

3.3. El modelo logístico como herramienta

La forma más general de especificar un predictor lineal para el modelo logístico puede expresarse de acuerdo con la expresión (2.15):

$$g\left[\pi\left(\mathbf{x},\boldsymbol{\beta}\right)\right]=\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x}.$$

Una vez consideradas estas formas de expresar el predictor lineal, una manera de facilitar el cálculo de los coeficientes y de aprovechar las bondades y alcances que ofrecen los modelos lineales, es a través de la transformación logit de la probabilidad de éxito [expresión (2.17)], es decir:

$$\text{logit}\left[\pi\left(\mathbf{x},\boldsymbol{\beta}\right)\right]=\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x}.$$

De este modo, y anticipándonos a lo que será el enfoque dinámico de este estudio, diremos que mediante el uso de esta transformación será posible aproximar la superficie de respuesta verdadera $\pi\left(\mathbf{x},\boldsymbol{\beta}\right)$ mediante funciones lineales en sus parámetros, que en nuestro caso será mediante la transformación logit. La forma con que analizaremos este tipo de modelos será a través de las herramientas y posibilidades que nos brindan los *MLG*.

Si bien el modelo logístico se transformará en el corazón operativo del análisis de los resultados de los diseños factoriales de primero y segundo orden para los *MDB*[5], no daremos mayores detalles sobre el cálculo en este capítulo, como así tampoco de propiedades, análisis e interpretación, ya que los mismos se encuentran extensamente detallados en la bibliografía disponible[6], sin prejuicio que hagamos referencia a aspectos particulares de interés especial. En el **Apéndice B** comentaremos algunos de estos aspectos.

[5] No obstante que existen otros modelos disponibles, en esta tesis hemos elegido el enfoque que ofrece la transformación logit de la probabilidad de éxito.

[6] *Vid.* p. ej.: Collett (2003); Hosmer y Lemeshow (2000); Kleinbaum (1994); etc.

3.4. Aspectos sobre el problema general del diseño

3.4.1. Generalidades

A la hora de pensar en cómo obtener datos y qué hacer con ellos, de modo que a partir de ellos pueda reducirse el desconocimiento sobre un cierto sistema, el experimentador distingue —cuanto menos— dos grandes capítulos en esta ruta de trabajo que ha decidido transitar. En una primera etapa, llamada muchas veces *diseño estadístico de experimentos*[7], o sencillamente "diseño", se intentará definir cómo deberían recogerse los datos de forma tal que la información contenida en ellos pueda llevar a que las etapas siguientes de lectura de la misma resulte lo más sencilla posible. Una nota importante que se suele agregar acerca del diseño es que la información contenida en los datos se establece cuando el experimento es llevado a cabo y, en general, no será posible disponer de otra información que la que no se encuentre presente en los datos.

Y en la segunda, *el análisis estadístico*[8], lo que se pretende es extraer dicha información, utilizar herramientas eficientes para analizarla y presentarla, y proporcionar criterios de decisión adecuados que permitan tomar decisiones con cierto soporte objetivo, cual es, el estadístico. No obstante la importancia que tiene el análisis, en esta tesis haremos hincapié más en el diseño que en el análisis, constantando que los diseños considerados conduzcan a un análisis satisfactorio y lo menos dificultoso posible.

En general, diseñar un experimento que nos permita conocer mejor el funcionamiento un sistema implica muchas veces pensar cuál será el modelo ajustado que podamos obtener de los datos. De este modo, la calidad del modelo ajustado dependerá mucho de la forma con que se han elegido y utilizado los datos provenientes de la observación del sistema en funcionamiento. Hablar de un modelo de "buena calidad" es sinónimo de asociarle a éste, cuanto menos, las siguientes tres cualidades básicas:

a. que el modelo sea capaz de explicar la relación entre la respuesta y los factores de modo claro,

b. que el modelo sea capaz de predecir razonablemente el valor que toma la respuesta para un cierto valor del vector de variables explicativas, y

c. que el modelo sea capaz de detectar valores anómalos o raros.

En síntesis, podemos considerar que un modelo de calidad es aquél que *explica bien* y *predice bien*, de la forma más sencilla posible.

[7] *Vid.* Bates y Watts (1988), p. 123.
[8] *Ídem* anterior.

Un primer paso obligado y necesario dentro del diseño será definir qué factores deberán elegirse —de entre un conjunto seguramente más numeroso— para explicar la relación funcional, lo cual impondrá también conocer qué niveles se les asignará a cada uno de los factores. Se dice que una vez que se han elegido qué factores serán tenidos en cuenta y qué niveles tomarán los mismos, "queda definida la región de experimentación"[9] del experimento, valga la redundancia. Luego, y una vez establecidos cuáles son los factores que realmente tienen influencia en el valor esperado de la respuesta bajo estudio, el mismo experimentador podrá obtener una gran ventaja tanto técnica como económica si con anterioridad a la observación de las respuestas se plantea qué valores asignará a los factores, es decir, qué niveles se considerarán de los mismos para realizar las mediciones de la respuesta bajo estudio. En particular, si se utilizaran modelos polinómicos de aproximación de la respuesta, una elección racional de estos niveles proporcionará seguramente una cierta ganancia en eficiencia al estimar los coeficientes del modelo que si no se hubiera elegido con algún criterio específico.

De esta forma, queda definido el problema de los diseños, que repetimos: *"decidir qué combinación de niveles deberían utilizarse en cada etapa de la experimentación, de modo que las estimaciones de los parámetros resulten con máxima precisión"*[10].

3.4.2. Propiedades deseables para los diseños

A la hora de pensar en qué diseño resultará más el adecuado, queda claro que éste podrá variar según el objetivo que se esté persiguiendo en el estudio. En Box y Draper (1975), por ejemplo, encontramos una lista de 14 propiedades deseables que se esperan de un "buen diseño" para los problemas de *MSR*[11]. Según esto, un buen diseño debería:

1. Generar una distribución satisfactoria de la información a lo largo de la región de interés.
2. Asegurar que los valores ajustados en un cierto $\mathbf{x}$, $\widehat{y}(\mathbf{x})$, se encuentren tan cerca como sea posible del verdadero valor en el mismo punto, $y(\mathbf{x})$.
3. Proporcionar una buena forma de detección de la falta de ajuste.
4. Permitir que puedan ser estimadas las transformaciones, si las hubiere.
5. Permitir que los experimentos puedan ser realizados en bloques.

[9] *Vid.* Khuri y Cornell (1996), p. 47.
[10] *Vid.* Myers (1976).
[11] Cuanto menos desde el punto de vista clásico, es decir, para el modelo lineal y normal.

6. Permitir ampliar el orden de los experimentos, de modo que faciliten un encadenamiento secuencial.
7. Proporcionar una fuente de estimación del error.
8. Ser insensible ante valores anómalos y frente al apartamiento de las hipótesis habituales del modelo lineal normal.
9. Requerir un número mínimo de condiciones experimentales.
10. Proveer formas simples de patrones de datos que permitan una cierta apreciación visual.
11. Conllevar cálculos simples.
12. Comportarse satisfactoriamente cuando aparezcan errores al establecer los valores de las x.
13. No requerir la consideración de un número grande de variables, de modo que hagan impracticable el experimento.
14. Proporcionar una herramienta de verificación de la constancia de la varianza.

Si bien queda claro que esta enumeración fue realizada para datos con distribución normal, la misma resulta útil como punto de partida para extenderlo a otras distribuciones, considerando o modificando el punto que corresponda cuando el caso lo amerite.

3.4.3. Aspectos más formales

En símbolos, definir un diseño implicará la especificación de las columnas y las filas de la matriz de datos, $\mathbf{X}$, de modo que queden definidas tanto las k variables que se considerarán como sus niveles. El problema en la elección de la matriz de datos, $\mathbf{X}$, $\dim(\mathbf{X}) = n \times p$, es considerado desde al menos dos puntos de vista particulares[12]:

a. *Diseños para la estimación de los parámetros:*

Cuando la forma de la verdadera relación funcional $f(\mathbf{x}, \boldsymbol{\beta})$ se supone conocida, el objetivo del diseño será que el mismo sea capaz de proporcionar buenas estimaciones del vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$; y

b. *Diseños para la exploración de la superficie:*

Cuando la forma de la verdadera relación funcional se presume desconocida, el objetivo del diseño[13] será aproximar $f(\mathbf{x}, \boldsymbol{\beta})$ dentro de una cierta región de interés

[12] *Vid.* Box y Draper (1959).
[13] *Vid.*, p. ej., Fang *et al.* (2006), p. 11.

perteneciente al espacio de k variables mediante una cierta función $g(\mathbf{x}, \boldsymbol{\beta})$.

Al respecto de esta segunda clase de diseños, en Box y Draper (1959) aparece una lista más abreviada —y tal vez más general— de los requerimientos que deberían tenerse en cuenta para un buen diseño para la exploración de superficies:

1. El diseño debería permitir la aproximación polinómica gradual de la respuesta verdadera, tan bien como sea posible dentro de la región de interés.
2. Debería permitir una verificación acerca del grado de adecuación de esta aproximación polinómica.
3. No debería contener un número excesivamente grande de puntos experimentales.
4. Debería permitir a sí mismo la separación en bloques.
5. Debería formar parte de un encadenamiento de diseños satisfactorios de mayor orden, en caso necesario.

Asimismo, las consecuencias de un diseño pobremente realizado llevarán a que el valor esperado de la superficie de respuesta verdadera, $E(\mathbf{y} \mid \mathbf{x})$, difiera de la ajustada $\widehat{y}(\mathbf{x})$, con lo que aparecerán dos grandes fuentes de "errores": uno debido al sesgo y otro debido a la dispersión. La misma referencia continúa la discusión más al detalle de cada una de las propiedades citadas en diferentes casos de estudio.

3.4.4. Calidad de los diseños para modelos lineales

Cuando se parte de la hipótesis que un modelo de naturaleza lineal con errores $DII\mathbf{N}(0, \sigma^2\mathbf{I})$ resulta ser la más adecuada para ajustar un modelo a datos con igual distribución, gran parte de la calidad explicativa del modelo ajustado estará dada por el vector estimado de los coeficientes del modelo, $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$, cuya distribución estará dada por:

$$\widehat{\boldsymbol{\beta}} \sim \mathbf{N}\left[\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]$$

Dado que la estimación es insesgada, interesará centrar la atención en la matriz —simétrica— de varianzas y covarianzas, $\mathbf{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$, cuya forma es la siguiente:

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \begin{bmatrix} V(\widehat{\beta}_0) & cov(\widehat{\beta}_0, \widehat{\beta}_1) & \cdots & cov(\widehat{\beta}_0, \widehat{\beta}_k) \\ & V(\widehat{\beta}_1) & \cdots & cov(\widehat{\beta}_1, \widehat{\beta}_k) \\ & & \cdots & \cdots \\ \text{sim.} & & & V(\widehat{\beta}_k) \end{bmatrix}$$

Puesto que lo deseable es que esta matriz $\mathbf{X}'\mathbf{X}$ sea lo más pequeña posible, el mejor diseño en cuanto a los coeficientes será aquél para el cual:

a. Los elementos de la diagonal de la matriz $\mathbf{X}'\mathbf{X}$ resulten los más grandes posibles[14]. Esto se puede traducir mediante una configuración de los puntos de tal modo que resulten lo más dispersos[15] posibles entre sí. Esta dispersión de $\mathbf{X}'\mathbf{X}$ tendrá repercusión en la matriz $\mathbf{V}(\widehat{\boldsymbol{\beta}})$, que resultará "tanto más pequeña" cuanto mayor sea aquella dispersión, puesto que $\mathbf{X}'\mathbf{X} = [\mathbf{V}(\widehat{\boldsymbol{\beta}})]^{-1}$.

b. Los elementos de fuera de la diagonal, nulos o lo más pequeños posibles en valor absoluto, lo que estará indicando la ortogonalidad entre las columnas de la matriz de datos, $\mathbf{X}$.

En cuanto a la capacidad predictiva del modelo, se comprueba que la varianza de predicción[16] está dada por:

$$V\left[\widehat{y}\left(\mathbf{x}_0\right)\right] = \mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0 \tag{3.1}$$

la cual depende también de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Por tanto, y observando ya sea la matriz de varianzas y covarianzas de los coeficientes estimados, como la expresión de la varianza de predicción, queda claro que ambas cantidades dependen de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$y no del nivel de la respuesta, es decir, de las y_i, como así tampoco del valor esperado de las mismas para distintos valores de $\mathbf{x}$, es decir, de $E(\mathbf{y} \mid \mathbf{x})$. Así, se sigue que un buen diseño que tenga en cuenta la estimación de los coeficientes y la variabilidad de la predicción, es deseable que éste dependa lo máximo posible de la matriz $\mathbf{X}$ y lo menos posible de los niveles de la respuesta observada.

3.4.5. Sobre la ortogonalidad en general

La alternativa que se utiliza para mejorar las propiedades del diseño[17] es la de la *ortogonalidad*: al imponérsele a las columnas —o filas— de la matriz de datos que sean ortogonales entre sí, el aspecto de esta matriz es el siguiente:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & \cdots & 0 \\ & n & \cdots & 0 \\ & & \cdots & \cdots \\ \text{sim.} & & & n \end{bmatrix} = n\,\mathbf{I}_k$$

[14] Esto se evidencia con mayor claridad si expresamos lo anterior mediante: $diag\{\mathbf{X}'\mathbf{X}\} = \{n, \sum_{i=1}^n x_{i1}^2, \sum_{i=1}^n x_{i2}^2, \ldots, \sum_{i=1}^n x_{ik}^2\}$

[15] Aquí nos referimos a dispersión en cuanto a que la mayor región dentro del espacio de los factores nos proporcionará mayor información.

[16] *Vid.*, p. ej., Myers (1990).

[17] Una primera observación acerca de $(\mathbf{X}'\mathbf{X})^{-1}$ es que, al ser justamente una matriz, no existe una métrica ordenada en el espacio al cual pertenece. Esto es: dadas dos matrices, no es posible indicar dentro de su espacio cuál de las dos es mayor o menor que la otra. Por tanto, pretender que la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ resulte "pequeña" no tiene mucho sentido en el espacio de $k \times k$ dimensiones.

$$\therefore \ \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 & \cdots & 0 \\ & \frac{1}{n} & \cdots & 0 \\ & & \cdots & \cdots \\ \text{sim.} & & & \frac{1}{n} \end{bmatrix} = \frac{1}{n}\,\mathbf{I}_k,$$

en donde $\mathbf{I}_k$ es la matriz identidad de $k \times k$, con unos en su diagonal y ceros fuera de ella. Al imponerle entonces la ortogonalidad a la matriz de diseños se logra al menos lo siguiente:

- $cov(\widehat{\beta}_i, \widehat{\beta}_j) = 0$, es decir, los elementos fuera de la diagonal resultan nulos;
- al tratarse de una matriz diagonal, cada $\widehat{\beta}_j$ es asignable solamente a un factor x_j, y que las estimaciones de dos coeficientes cualesquiera resultan independientes entre sí, quedando el diseño a cubierto de la multicolinealidad;
- las varianzas de las $\widehat{\beta}_j$ tienen siempre el mismo valor, $\frac{\sigma^2}{n}$, puesto que $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2 \frac{1}{n}\mathbf{I} = \frac{\sigma^2}{n}\mathbf{I}$. [Ver expresión (3.2)].

Un ejemplo: la ortogonalidad en diseños a 2 niveles

Los diseños a 2 niveles, tanto factoriales completos como fraccionales, poseen la propiedad de ortogonalidad cuando lo que pretenden diseñar son modelos lineales, que podrán incluir también términos de interacciones lineales en el sentido en que $\frac{\partial}{\partial \beta_j}[f(\mathbf{x}, \boldsymbol{\beta})]$ no depende de β_j, para todo j.

Por ejemplo, si se piensa que la relación funcional entre la respuesta y un conjunto 3 de factores siguen una relación lineal, puede pensarse en diseñar un factorial estándar completo a dos niveles, es decir un 2^3, requiriendo un mínimo de $n = 8$ observaciones de la respuesta. Si en el modelo se tienen en consideración los términos de interacción de segundo y tercer orden —lo que lleva a considerar $p = 8$ parámetros— la i-ésima observación de la respuesta $(i = 1, \ldots, n), n > 8$, puede escribirse como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3} + \beta_{123} x_{i1} x_{i2} x_{i3} + \varepsilon_i,$$

o bien matricialmente como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

La matriz de diseño correspondiente al modelo, $\mathbf{X}$, $\dim(\mathbf{X}) = 8 \times 8$, se puede escribir como:

$$\mathbf{X} = \left[\mathbf{1}, \mathbf{x}_1', \mathbf{x}_2', \mathbf{x}_3', (\mathbf{x}_1\mathbf{x}_2)', (\mathbf{x}_1\mathbf{x}_3)', (\mathbf{x}_2\mathbf{x}_3)', (\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3)'\right]',$$

en la que cada coordenada de este vector corresponde a las siguientes columnas —linealmente independientes entre sí— de la matriz estándar de diseño:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Esta configuración, y en ausencia de réplicas de la respuesta, hace que la matriz de diseño sea ortogonal, con lo cual el producto de ésta por su transpuesta da como resultado la siguiente matriz diagonal:

$$\mathbf{X}'\mathbf{X} = diag\left\{2^3\right\} = diag\left\{8\right\} = 8\mathbf{I}$$

Se demuestra[18] que este resultado puede generalizarse para k factores, con lo que se obtiene de forma análoga:

$$\mathbf{X}'\mathbf{X} = diag\left\{2^k\right\} = 2^k\mathbf{I},$$

en donde $\mathbf{I}$ es la matriz identidad de orden 2^k o menor[19]. Puesto que se trata de una matriz diagonal, su inversa, $(\mathbf{X}'\mathbf{X})^{-1}$, también será diagonal[20]:

$$(\mathbf{X}'\mathbf{X})^{-1} = diag\left\{\frac{1}{2^k}\right\} = \frac{1}{2^k}\mathbf{I}_{2^k},$$

es decir, una matriz simétrica con $\frac{1}{2^k}$ en su diagonal y ceros fuera de ella.

Tanto el vector estimado de coeficientes como la matriz de varianzas y covarianzas de éstos es función de la matriz $\mathbf{X}$. Ambas cantidades, si se tiene en cuenta la ortogonalidad, pueden escribirse como:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{2^k}\mathbf{X}'\mathbf{y}$$

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\sigma^2}{2^k}\mathbf{I}_{2^k} \tag{3.2}$$

[18] *Vid.* Myers *et al.* (2002), por ejemplo.

[19] Esto es así puesto que pueden formarse matrices identidad de orden 2^k u órdenes menores, dependiendo del modelo que se considere.

[20] *Vid.* p.ej. Lipschutz (1992).

Por lo tanto, siguiendo esta forma de construir los diseños factoriales a dos niveles, para el modelo lineal normal las estimaciones de los parámetros son independientes entre sí y la matriz de varianzas y covarianzas no depende de los parámetros, además de tener éstos la propiedad **BLUE** (del acrónimo inglés "**B**est **L**inear **U**nbiased **E**stimators")[21].

Recapitulando conceptos, y frente a la pregunta fundamental del diseño sobre en qué parte de la región de experimentación se establecen los valores de $\mathbf{x}$, siempre que el modelo sea lineal y normal, cuanto "más ortogonales" sean las columnas de la matriz de diseño, tanto más convenientes y deseables serán las propiedades estadísticas de los coeficientes del modelo.

3.4.6. Calidad de los diseños para modelos no lineales

En el caso en que el experimentador tenga evidencias que el proceso bajo estudio siga alguna forma no lineal determinada, de valor esperado $f(\mathbf{x}, \boldsymbol{\beta})$, la idea de matriz de diseño tiene más sentido asimilarla como matriz de derivadas de $f(\mathbf{x}, \boldsymbol{\beta})$ por los motivos comentados en el capítulo anterior. Al extenderse a un conjunto de n puntos muestrales, dicha matriz será: $\mathbf{F} = \{f_{ij}\} = \{\frac{\partial}{\partial \beta_j}[f(\mathbf{x}_i, \boldsymbol{\beta})]\}$, para $i = 1, \ldots, n$ y $j = 1, \ldots, p$.

Partiendo de la expresión que permite determinar la matriz asintótica de varianzas y covarianzas de los coeficientes, es decir $\mathbf{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{F}'\mathbf{W}\mathbf{F})^{-1}$, aparecen inmediatamente las siguientes características:

- Las estimaciones máximo verosímiles de los coeficientes serán asintóticamente insesgadas y con varianza asintóticamente mínima.
- Dado que existirá al menos una variable j tal que la derivada parcial de $f(\mathbf{x}, \boldsymbol{\beta})$ con respecto a su respectivo coeficiente sea función de éste, habrá al menos una componente de la matriz $\mathbf{F}$ que será a su vez función de aquél. En otras palabras, al menos una de las componentes de esta matriz $\mathbf{F}$ será no lineal.

Por lo precitado, diremos que debido a que la matriz de varianzas y covarianzas de los estimadores de los coeficientes del predictor lineal $\eta = \mathbf{x}'\boldsymbol{\beta}$, el modelo logístico —de naturaleza típicamente no normal y no lineal— agregará una mayor dificultad

[21] Esta es una propiedad general del modelo lineal. En otras palabras, la consideración de términos lineales en el modelo hasta un orden igual o inferior al número de factores, lleva a una disposición ortogonal de la matriz de diseño y diagonal de la matriz de varianzas y covarianzas de los parámetros del modelo. Si, por otro lado, la componente aleatoria siguiese una distribución $DII\mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_{2^k})$, entonces las estimaciones de los parámetros tendrán además mínima varianza y, en particular, coincidirán además con las estimaciones máximo-verosímiles de los mismos.

a la hora de considerar el problema del diseño de los experimentos si los comparamos con los modelos lineales normales. La razón principal está en que la matriz de varianzas y covarianzas depende de los valores inciales que se consideren para $\boldsymbol{\beta}$, que son desconocidos. A partir de estas consideraciones, no resultará trivial responder, al menos, a la pregunta: "*¿en qué zona de la región de experimentación debemos considerar los niveles de los factores de tal forma que nos conduzca a obtener la mayor información sobre el valor esperado de la respuesta?*". La respuesta a esta pregunta dependerá claramente de la forma no lineal que se suponga para el modelo estudiado.

Para terminar este punto, comentaremos algunos otros aspectos que el experimentador deberá tener en cuenta en el diseño de modelos no lineales:

- *Puntos de soporte:* la cantidad de puntos a considerar debe ser mayor que el número de parámetros considerados para el modelo;
- *Réplicas:* para cada punto a considerar, deberá tenerse al menos una medición del valor de la respuesta, cuanto menos en el sentido de "réplica genuina";
- *Secuencialidad:* la forma más eficiente de ganar información sobre el sistema no será sino la forma secuencial, ubicando nuevos puntos en cada etapa del experimento dentro de la zona no lineal.

Sin el cumplimiento de estos puntos no se podrán realizar pruebas de falta de ajuste del modelo, ni validarlo de modo razonable, lo cual llevará a un conocimiento más bien pobre de cómo se relacionan la respuesta con los factores, lo que a su vez conducirá a predicciones inadecuadas.

3.4.7. Comentarios sobre los diseños óptimos

En términos generales, es frecuente no asimilar a la métrica de la matriz $(\mathbf{F}'\mathbf{F})^{-1}$ dentro de un espacio matricial sino hacerlo en una sola dimensión, la real. Para ello, se realizan transformaciones que convierten a esta última matriz en un número real, de modo que el mismo tenga una métrica ordenada. Un ejemplo de ello es el cálculo del determinante de esta matriz, que conducirá a un valor real puntual cuando sea posible. De este modo, es posible definir una cierta clase de diseños muy utilizados, que emplean este tipo de métricas, llamados *diseños alfabéticamente óptimos*, los cuales no abordaremos sino someramente en este trabajo.

En particular para modelos lineales, interesarán aquellos diseños en donde los valores de la matriz de diseño haga que los datos se tomen lo más espaciadamente posible —además de la ortogonalidad— de forma tal que cuanto más separados estén, por simple geometría, las estimaciones de los parámetros de las formas rectilíneas (rectas,

planos, hiperplanos, etc.) que pasen por ellos serán más exactas, beneficiándose el experimentador por las bondades que ello trae en materia de varianza y de predicción.

Para el caso de los diseños correspondientes a modelos no lineales los estimadores de las β, la matriz de varianzas y covarianzas de éstas y la varianza de predicción no dependen exclusivamente de la matriz de diseño $\mathbf{X}$ sino que también lo harán de los valores que tome el vector de respuestas, $\mathbf{y}$. Por lo tanto, parece claro entender que con antelación a la realización del experimento, será necesario conocer tanto $\mathbf{V}(\widehat{\boldsymbol{\beta}})$ como $V\left[\widehat{y}\left(\mathbf{x}\right)\right]$, que tampoco se conocen. La estrategia, por lo tanto, deberá ser necesariamente secuencial y dependerá fuertemente de la forma de cómo estará definida la parte sistemática del modelo, $f(\mathbf{x}, \boldsymbol{\beta})$, que evidenciará la no linealidad del mismo.

Frente a la cuestión de encontrar cuáles serán las x que minimicen las matrices de varianzas y covarianzas de los coeficientes, $\mathbf{V}(\widehat{\boldsymbol{\beta}})$, y la de predicción, $V\left[\widehat{y}\left(\mathbf{x}\right)\right]$ (expresión 3.1), actualmente hay 3 líneas de investigación, que resumiremos a continuación[22]:

- *Diseños localmente óptimos.* Del capítulo anterior, el determinante $|\mathbf{F}'\mathbf{F}|$ depende del vector $\boldsymbol{\beta}$. Se supone de entrada un cierto valor $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ para el vector de parámetros y se escoge el mejor diseño para ese valor. A menudo[23], aunque no siempre, se puede demostrar que dado un valor para el parámetro, el diseño óptimo requiere p puntos diferentes, siendo p el número de parámetros del modelo. El trabajo seminal que dio solidez a estos enfoques fue el artículo de Box y Lucas (1959), y actualmente hay disponibles varias obras dedicadas a estudiar los diseños óptimos.

- *Diseños bayesianos.* La idea básica es suponer una distribución para los parámetros, $\pi(\beta)$, y se escoge un cierto vector de puntos, $\mathbf{x}$, que maximicen el valor de ciertas transformaciones[24] de $\mathbf{F}'\mathbf{F}$, luego de lo cual se hace un promedio ponderado del tipo $\int_{\Omega} |\mathbf{F}'\mathbf{F}| \ \pi(\beta)\ d\beta$. En este tipo de diseños requiere un mayor número de puntos de soporte que los diseños localmente óptimos.

- *Diseños minimax.* En estos diseños, se escoge un vector de puntos, $\mathbf{x}$, de tal modo que resulten el óptimo para el peor valor posible del parámetro, requiriendo el conocimiento de toda la región que define el espacio de parámetros. Así, si dicha región es $\Re$, se buscará encontrar los valores de β tales que se minimice $\min\limits_{\beta \in \Re} |\mathbf{F}'\mathbf{F}|$.

Al plantearnos el problema de obtener diseños óptimos, y puesto que la matriz de varianzas y covarianzas de los estimadores de los parámetros del modelo depende de la matriz de derivadas $\mathbf{F}$, el foco de este estudio suele estar puesto en la matriz $\mathbf{F}'\mathbf{F}$,

[22] Ginebra (2001).
[23] Por ejemplo, para diseños D-óptimos.
[24] Sin embargo, la transformación utilizada no tiene que ser necesariamente el determinante.

que en el modelo lineal normal con las hipótesis clásicas suele escribirse como $\mathbf{X}'\mathbf{X}$. En Box y Draper (1987), pp. 489 *et seq.* puede encontrarse una síntesis muy adecuada sobre ésta y otras implicaciones de dicha matriz, como así también los criterios más utilizados en este contexto, que son los llamados *diseños alfabéticamente óptimos*[25].

3.4.8. Diseños óptimos para *MDB*

Resulta bastante conocida la utilidad de emplear diseños ortogonales cuando el modelo propuesto se supone de naturaleza lineal y normal. Como hemos comentado anteriormente, uno de los motivos es que la matriz de varianzas y covarianzas de los parámetros estimados del modelo, $\mathbf{V}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, resulta diagonal. Sin embargo, la utilización de diseños ortogonales cuando los datos no son de naturaleza lineal ni normal no es un aspecto que esté del todo claro tanto como lo está en el modelo normal y lineal, puesto que lo que realmente interesa no es que el diseño sea ortogonal sino que la matriz $\mathbf{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{F}'\mathbf{W}\mathbf{F})^{-1}$ resulte diagonal[26]. Puede verse que la diagonalidad queda "obstaculizada" por la presencia de la matriz de "pesos" $\mathbf{W}$, que es equivalente a indicar que la falta de constancia de la varianza de la respuesta para distintos niveles de los factores, ofrece una dificultad para encontrar diseños óptimos en datos provenientes de naturaleza binaria, como así también en otros modelos con esta misma inconstancia.

Para el caso particular en que cada observación de la respuesta en cada condición experimental $\mathbf{x}_i$ proviniese de una distribución binomial independiente de parámetro π_i e índice m_i, la matriz de pesos de los coeficientes será diagonal y se calculará a partir de la ecuación (2.12), quedando una expresión de la forma[27]:

$$\mathbf{W} = diag\left\{V\left(y_i\right)\right\} = diag\left\{m_i\ \pi\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\left[1 - \pi\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right]\right\}$$

Desarrollando la expresión del interior de las llaves:

$$V\left(y_i\right) = \left[\frac{m_i}{1+\exp(\eta_i)}\right]\left[1 - \frac{1}{1+\exp(\eta_i)}\right] = \frac{m_i\ \exp(\eta_i)}{\left[1+\exp(\eta_i)\right]^2},$$

en donde para abreviar notación, utilizamos $\eta_i = \beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i$. Por lo tanto, la matriz diagonal $\mathbf{W}$ se podrá escribir como:

$$\mathbf{W} = diag\left\{\frac{m_i\ \exp(\beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i)}{\left[1+\exp(\beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i)\right]^2}\right\} \tag{3.3}$$

La naturaleza lineal del predictor $\beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i$, para cada una de las n condiciones experimentales llevará a considerar directamente la matriz de diseño $\mathbf{X}$ en lugar

[25] *Vid.* Sitter y Wu (1993); Myers *et al.* (1994) y Sitter y Fainaru (1999), entre otros.

[26] En otros términos, cuando el modelo no sigue una distribución normal, **no** se verifica la siguiente expresión: $\mathbf{X}$ ortogonal $\Longrightarrow \mathbf{V}(\widehat{\beta})$ diagonal, debido a que la presencia de la matriz $\mathbf{W}$ en $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ no garantiza su diagonalidad.

[27] *Vid.* p. ej. Myers *et al.* (2002), pp. 164 *et seq.* o McCullagh y Nelder (1989), pp. 116 *et seq.*

de la de derivadas por los motivos indicados en el capítulo anterior. Por otro lado, ya que para el caso específico de la distribución binomial se tiene que el parámetro de dispersión ϕ toma el valor unidad[28], la expresión de la matriz asintótica de varianzas y covarianzas de los parámetros del modelo logístico tomará la forma:

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} = \left(\mathbf{X}'\left[diag\left\{\frac{m_i\ \exp(\eta_i)}{[1+\exp(\eta_i)]^2}\right\}\right]\mathbf{X}\right)^{-1} \tag{3.4}$$

Observando que la expresión anterior es función del vector de parámetros verdaderos, $\boldsymbol{\beta}$, si se reemplaza el mismo por su estimación $\widehat{\boldsymbol{\beta}}$, esto conduce a considerar la matriz diagonal *estimada*, $\widehat{\mathbf{W}}$, lo cual lleva finalmente a la matriz asintótica *estimada* de los estimadores, denotada como $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})$, cuya expresión se deduce de la última:

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1} = \left(\mathbf{X}'\left[diag\left\{\frac{m_i\ \exp(\widehat{\eta}_i)}{[1+\exp(\widehat{\eta}_i)]^2}\right\}\right]\mathbf{X}\right)^{-1} \tag{3.5}$$

para $i = 1, \ldots, n$ puntos de diseño, cada una de los cuales definido por su respectivo vector de factores $\mathbf{x}_i = (x_1, \ldots, x_k)'$.

De esta forma, deducimos que los diseños correspondientes *MDB* requerirán el conocimiento de valores iniciales para el vector de coeficientes. Dado que esta cuestión excede los alcances del problema del diseño en sí mismo, solamente tomaremos en consideración los resultados más relevantes, sin desarrollarlos con específica profundidad. Estos detalles se encuentran abordados en profundidad en referencias tales como Begg y Kalish (1984); Wu (1985); Minkin (1987); Sitter y Wu (1993); Myers *et al.* (1994); Sitter y Fainaru (1997), entre varios otros.

3.4.9. Ejemplo: diseños con un solo factor

Cuando en un experimento interviene un solo factor de variabilidad, se tendrá que $k = 1$ variables. Inmediatamente surge que: $p = k+1 = 2$, con lo cual el modelo tendrá dos parámetros, digamos $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. De este modo, se tendrá una "expectation function" cuya forma general será: $f(x, \boldsymbol{\beta}) = f(x, \beta_0, \beta_1)$, que será por lo general desconocida antes de comenzar el experimento.

Como mencionáramos anteriormente, será necesario contar como mínimo con $n \geq p = 2$ puntos de diseño con los cuales estimar ambos parámetros, a partir de los cuales se construirá un modelo de aproximación $g(x, \widehat{\boldsymbol{\beta}})$ con el cual conocer mejor cómo se relaciona la respuesta con el factor considerado. La distribución que tengan los datos podrá dar una idea inicial acerca de la forma que tenga el valor esperado de $f(x, \boldsymbol{\beta})$, la cual derivará en distintos enfoques de acuerdo a su naturaleza lineal o no lineal, básicamente.

[28] Comentaremos este aspecto en el **Apéndice B**.

3.4.10. Modelos no lineales

En las figuras 3.1 y 3.2 pueden verse dos gráficos representativos en los cuales la relación entre la respuesta y el factor de variabilidad es del tipo no lineal. En la primera figura, la representación corresponde a una típica "sigmoide" obtenida al modelar la probabilidad de éxito $\pi(x, \boldsymbol{\beta})$ mediante el modelo logístico.

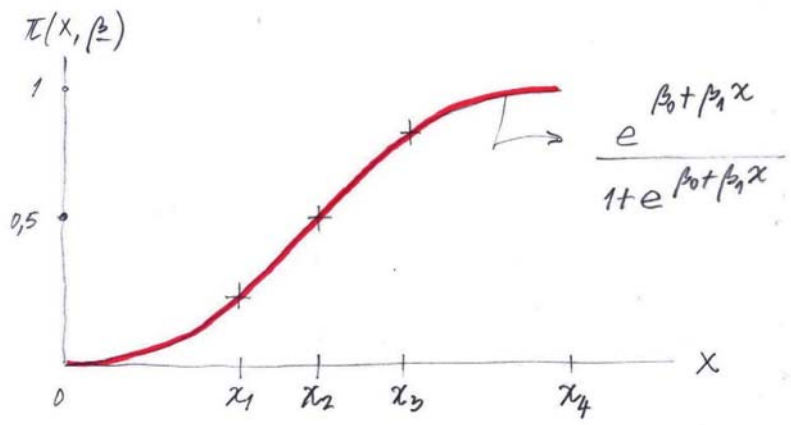

Figura 3.1: Relación entre el valor esperado de la respuesta —sigmoide— y un factor de variabilidad, mediante el modelo logístico.

La figura 3.2 es una representación gráfica exagerada de uno de los problemas que tienen los modelos no lineales en cuanto al diseño. Parece bastante claro el dibujo como para pensar que el lugar geométrico de $E(y \mid x)$ ya no resulta relativamente tan "simple" como en el caso anterior, sino que presenta evidentes zonas irregulares.

En particular, observamos que la curva $f(x, \boldsymbol{\beta})$ de la segunda figura tiene dentro del eje x tres zonas bastante bien definidas: (a) la zona comprendida entre x_1 y x_2, de aspecto lineal; (b) la comprendida entre x_2 y x_3, o *zona de transición*, que es evidentemente no lineal, y (c) la zona comprendida entre x_3 y x_4, que también pareciera ser lineal. El problema clave en este sentido, es que a priori de la realización del experimento *no se sabe cuál es la forma de* $f(x, \boldsymbol{\beta})$, y el gran desafío del experimentador consistirá, entre otras cosas, en decidir cómo distribuirá sus n puntos disponibles de forma tal que pueda obtener luego la máxima información posible sobre esta zona irregular especialmente. Sin adentrarnos demasiado en la resolución completa, indicaremos aquí algunos aspectos sobre el diseño que serán generalizables sin mayor dificultad a la hora de tener en cuenta diseños con más factores.

Si se siguiera la misma lógica que en el caso de los diseños para modelos lineales, definir los dos niveles del factor alrededor de los extremos de la zona de transición, x_2

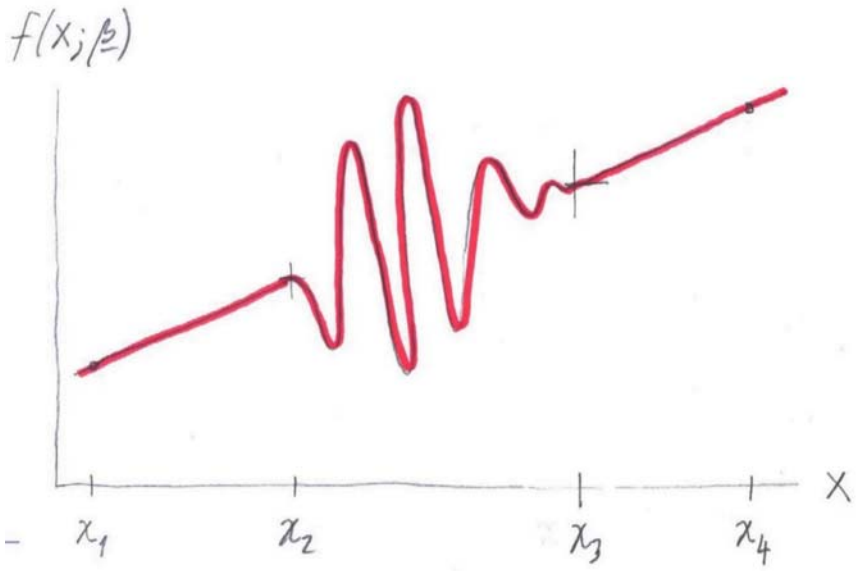

Figura 3.2: Representación genérica de una función exageradamente no lineal en sus parámetros.

y x_3, no resultará de mucha utilidad para poder averiguar qué aspecto tiene $f(x, \boldsymbol{\beta})$ en la zona no lineal, ya que lo que se estará haciendo es ajustar una recta entre los puntos (x_A, y_A) y (x_B, y_B). A partir de esto, diríamos que la zona comprendida entre ambos puntos es lineal, lo cual sería incorrecto.

Otro enfoque que puede darse sería el de disponer de información adicional que permita identificar las 3 zonas que definimos anteriormente, y pensar que la cantidad de puntos disponibles podría repartirse en tres tercios, asignando un tercio de los puntos para cada zona. La reflexión que surgiría de inmediato es que quizá ubicar un tercio de los puntos en ambas zonas lineales sea "malgastarlos", ya que por la naturaleza lineal de ambas zonas, harían falta al menos dos puntos solamente en cada una de ellas. Si admitimos por un momento que *una cuarta parte* de los puntos es una cantidad razonable para repartir en cada una de las zonas lineales, nos queda aún por resolver la siguiente incógnita: *¿en dónde sería conveniente colocar los* $\frac{n}{2}$ *puntos restantes de los que disponemos, a lo largo de la zona no lineal, como para poder obtener luego la mayor cantidad de información sobre ella?*.

Como comentario, diremos que si tenemos alguna evidencia de la existencia de esta zona no lineal, los puntos no deberán ser ubicados en los extremos de ella sino más bien en regiones intermedias dentro de la misma, con algún criterio que permita extraer la mayor cantidad de información posible. La habilidad del experimentador y toda otra información que pudiera conseguir sobre las características de esta zona, resultarán las

herramientas con las que contará para determinar cómo se comporta el valor esperado de la respuesta para el rango de x correspondiente.

Recapitulando ideas, los 3 criterios que comentamos anteriormente (diseños localmente óptimos, diseños bayesianos y diseños minimax) pretenden darle una solución relativamente aproximada a este problema, aunque para la solución definitiva seguramente será necesario pensar en disponer de más puntos de diseño, lo que se traduce en mayor esfuerzo experimental, tanto técnico como económico, puesto que cada modelo no lineal tiene prácticamente una solución *ad-hoc*.

3.4.11. Ejemplo: el diseño factorial 2^2

Dado que la matriz de varianzas y covarianzas de los parámetros resultará una herramienta muy útil a la hora de evaluar las distintas estrategias de diseño que describiremos más adelante, resulta interesante representar qué aspectos tendrán dichas matrices para diseños sencillos, como lo son los diseños factoriales a 2 niveles.

El esquema más simple de diseños factoriales para dos factores sea quizá el diseño 2^2, cada uno de los cuales, pudiendo adoptar dos niveles.

La forma habitual de expresar los valores que pudieran tomar los factores, expresados en unidades codificadas, es mediante la siguiente matriz de diseño $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

la cual puede ser reescrita considerando que se trata de un vector de 2 factores principales y uno de interacción:

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1\mathbf{x}_2)',$$

los cuales corresponden a las 4 condiciones experimentales que se ensayarán para obtener observaciones de la respuesta.

Si disponemos de cierta evidencia estadística según la cual la respuesta de interés sigue una distribución binomial, por las razones anteriormente expuestas podemos pensar que el modelo logístico puede servir para modelar los valores esperados de la respuesta en función de los factores considerados, de forma tal que para todo i se tendrá que:

$$\pi\left(\mathbf{x}_i, \boldsymbol{\beta}\right) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})} \tag{3.6}$$

Al considerar el "link" logit, y para la $i-$ésima condición experimental, esta expresión se transforma en:

$$\text{logit}[\pi\left(\mathbf{x}_i, \boldsymbol{\beta}\right)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \tag{3.7}$$

Para cada una de las condiciones experimentales podrá definirse la siguiente tabla:

CE	**1**	**A**	**B**	**C**	m_i	π_i	y_i	$\widehat{\pi}_i$
1	1	-1	-1	1	m_1	π_1	$y_1 \overset{\text{indep.}}{\sim} \text{binom}(m_1, \pi_1)$	$\frac{y_1}{m_1}$
2	1	1	-1	-1	m_2	π_2	$y_2 \overset{\text{indep.}}{\sim} \text{binom}(m_2, \pi_2)$	$\frac{y_2}{m_2}$
3	1	-1	1	-1	m_3	π_3	$y_3 \overset{\text{indep.}}{\sim} \text{binom}(m_3, \pi_3)$	$\frac{y_3}{m_3}$
4	1	1	1	1	m_4	π_4	$y_4 \overset{\text{indep.}}{\sim} \text{binom}(m_4, \pi_4)$	$\frac{y_4}{m_4}$

En la misma, mediante la notación $\mathbf{1}, \mathbf{A}, \mathbf{B}$ y $\mathbf{C}$ referimos al vector columna de unos, al primer factor (x_1), al segundo factor (x_2) y a la interacción de ambos factores (x_1x_2), respectivamente. El índice m_i refiere a la cantidad de repeticiones del experimento de Bernoulli ocurridos en la i-ésima condición experimental (CE), en donde cada uno de los cuales ocurre con una probabilidad de éxito igual a π_i. Así, para una CE determinada, se realizarán m_i observaciones binarias (0 ó 1), que llevarán a contar con y_i éxitos totales, $0 \leq y_i \leq m_i$. Las dos últimas columnas indican la distribución asociada al número de éxitos de cada CE —todas binomiales independientes— y la estimación de cada una de las correspondientes probabilidades de éxito, $\widehat{\pi}_i$, respectivamente.

La estimación máximo verosímil del vector de parámetros de (3.6) o de (3.7), se realiza mediante métodos iterativos, propios de los modelos no lineales. El estimador así obtenido resultará asintóticamente insesgado, con distribución asintóticamente normal y cuya variabilidad estará expresada mediante la matriz de varianzas y covarianzas que ya hemos comentado:

$$\mathbf{V}(\widehat{\boldsymbol{\beta}}) = [a\,(\phi)]^2 \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \tag{3.8}$$

que para el caso binomial, se tiene como propiedad que $a\,(\phi) = 1$. De acuerdo con la expresión (3.3), la última expresión tendrá por matriz $\mathbf{W}$ de "pesos" a la siguiente:

$$\mathbf{W} = diag\left\{\frac{m_i \ \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})}{[1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})]^2}\right\}$$

Estimando el vector de coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ mediante los algoritmos anteriormente indicados, se reemplazan en (3.8) y se llega a la matriz asintótica estimada de varianzas y covarianzas para este diseño. El programa **R** permite realizar esto sin mayor dificultad.