Instituto Tecnológico Autónomo de México

Diseño y análisis de experimentos, diseños definitivos y detección de factores

Tesis

que para obtener el título de

Licenciado en Matemáticas Aplicadas

presenta

Pablo Freyria Dueñas

Asesor: Dr. Ernesto Juvenal Barrios Zamudio

México, D.F. 2019

"Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Diseño y análisis de experimentos, diseños definitivos y detección de factores**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación".

Pablo Freyria Dueñas

Fecha

Firma

Agradecimientos

Quisiera aprovechar este espacio para agradecer y dar crédito a todas las personas que hicieron posible este trabajo mediante su apoyo, enseñanzas, amistad y cariño. Sin poder asignar un orden de importancia, las mencionaré simplemente en el orden que me aparezcan a la mente en este momento.

Mis primeros agradecimientos son a mi familia que siempre han estado y estarán en momentos de decisiones o difíciles, además claro en celebraciones, aventuras o simplemente descanso. En particular, quiero agradecer a mi papa por transmitirme su pasión por la ciencia, que estoy seguro fue lo que me llevo a estudiar esta carrera, y su espíritu de inventor; a mi mama por su dedicación y apoyo constante a los demás. A mi hermana Ana por su empatía y admiro su disciplina para perseguir sus metas, y a mi hermano Santiago agradezco que me ha enseñado que la vida es una gran aventura y manteniendo buena energía vendrán cosas buenas.

Quiero agradecer también a mis amigos de Puebla, a Michel, Terry, Jonathan, Joc, Georges, Carlos, Sam e Ilias. Aunque hemos tomado diferentes caminos, pareciera que no ha pasado el tiempo, agradezco el apoyo, críticas y desde luego, las buenas platicas hasta la madrugada. Hemos compartido muchas historias que en gran parte me hacen quien soy ahora y estoy seguro qué compartiremos muchas más.

De igual manera, agradezco a mis amigos de la carrera: al conocidísimo Loko, a Joaquin(g), Isaac, Pato, Willy y Chisco, por sus ideas y por estar en todas esas horas de estudio y convivencia.

Mis agradecimientos van también a los roomies, Barrerra y Zubi a quienes aprecio mucho y ha sido muy divertido platicar, bromear y convivir con ustedes. Siempre han estado para un consejo, una inquietud o simplemente a ver que se nos ocurre.

A mis profesores de la carrera a quienes les agradezco no solo mi formación académica, sino que también siempre mantuvieron una puerta abierta y pasión por sus materias. En particular, quiero agradecer a Ernesto J. Barrios, quien además de ser muy paciente como asesor de este trabajo, me enseñó que la estadística es un tema fascinante de las matemáticas, sobre el cuál me gustaría ahondar en un futuro.

Por último, a mis compañeros y amigos de LSC: Dani, Marco, Deya y Caro por compartir sus ideas y formas distintas de pensar, además de mantener una buena actitud siempre. Y a Caro y Javier por ver por nuestro crecimiento profesional y buscar nuestro mejor trabajo.

Índice general

Capítulo 1

Introducción

R.A. Fisher menciona en [Fisher, 1935, cap. 1] que las conclusiones científicas derivadas de evidencia experimental pueden ser refutadas con dos tipos de argumentos: el primero atacará a la interpretación del experimento y argumentará que los resultados obtenidos no son congruentes con la conclusión, o bien, que pudieron ser obtenidos aunque ésta fuera falsa; el segundo argumento criticará el diseño del experimento y dirá que su estructura lógica sesgó los resultados y por lo tanto la conclusión.

Por esta razón, el conocimiento generado de manera inductiva, es decir, a través de experimentación, requiere simultáneamente de experimentos bien diseñados y del uso adecuado de procedimientos estadísticos que permitan obtener conclusiones válidas.

Por su naturaleza, el aprendizaje a partir de datos es un proceso iterativo: a partir de una conjetura (modelo), se diseña un experimento cuyos resultados la confirman (deducción), o en caso de haber una discrepancia, sirven para formular otra conjetura (inducción) [Box et al., 2005, cap. 1]. De manera que, el experimento tendrá etapas distintas con objetivos distintos: en las primeras etapas, la prioridad será detectar los factores activos y encontrar una dirección para seguir investigando, mientras que en etapas más avanzadas se buscará caracterizar con mayor precisión y certidumbre la superficie de respuesta.

Una característica de los estudios *experimentales* es que el investigador tiene control sobre el sistema que estudia y puede medir la respuesta en configuraciones predeterminadas. Esto se diferencia de un *estudio observacional* en el que el investigador trabaja con datos dados cuyo origen y configuración no controla, por lo que puede haber un factor no considerado o latente que invalide la inferencia que se quiera realizar. Además, dichos datos pudieran no cubrir toda la región de interés o no presentarse en configuraciones que permitan separar los efectos de factores distintos que ocurran en patrones similares, por ejemplo, en estudios macroeconómicos.

El objetivo de este trabajo es recopilar y exponer la teoría del diseño de experimentos: se mostrarán distintos diseños que han sido propuestos en la literatura clásica y las técnicas de análisis adecuadas a los datos generados por los experimentos realizados según cada diseño. Éstas técnicas incluyen: la estimación de los efectos de cada factor e interacción en la respuesta, así como la detección de factores activos.

Los diseños de experimentos propuestos en este trabajo estarán enfocados a etapas tempranas de investigación, donde el objetivo del investigador sea realizar pocos ensayos y poder detectar los factores relevantes en la respuesta, descartando el resto y pudiendo obtener un modelo, aunque sea sencillo, que le permita guiar su investigación futura y obtener más detalle en la superficie de respuesta. A este enfoque, cuyo objetivo principar es detectar a los factores activos en la respuesta, se le conoce como *cribado o detección de factores* (del inglés *factor screening*).

A diferencia de los modelos lineales "clásicos" de regresión, se considerará a las variables (en este contexto llamadas factores) como categóricas y no se buscará establecer una relación entre un cambio en el valor de los factores y la respuesta, sino que se estimará la respuesta en cada valor de los factores, obteniendo así una superficie de respuesta y no una expresión (o modelo) que permita generalizar a valores no observados.

La organización general del trabajo es presentar en el capítulo 2 el análisis estadístico sobre los resultados experimentales que permitirá estimar el efecto de cada factor e interacción. Los temas abarcados incluyen: la estimación de los efectos y su comparación contra el ruido

aleatorio del experimento, verificación de los supuestos y propiedades de los diseños de experimentos. En este capítulo, el diseño de experimento utilizado consiste en dividir las variables de interés en distintos valores, llamados niveles, y evaluar la respuesta en todas las combinaciones de éstos.

En los capítulos 3 y 4 se presentarán clases de diseños que permitirán estimar una superficie de respuesta lineal y cuadrática respectivamente. En el proceso iterativo de experimentación servirá para definir la región futura de interés y en general se va a asumir que los factores tendrán efectos simples antes que complejos[1]. El objetivo principal de éste capítulo es minimizar el número de ensayos necesarios para estimar la superficie de respuesta y estudiar la confusión entre los estimadores de los efectos que surge al reducir el número de ensayos.

En el capítulo 5 se presenta una clase de diseños de experimentos propuesta por Jones y Nachtsheim en el 2011, que nombraron de "diseño definitivo", ya que requieren un número menor de ensayos que las clases de diseños mostradas en los capítulos anteriores y permitirán estudiar un comportamiento cuadrático por factor en la respuesta, y donde los estimadores de efectos lineales son independientes de los estimadores de los efectos de interacción, permitiendo identificar a los factores activos de manera aislada. En éste capítulo se mostrarán propiedades de ésta clase de diseños de experimentos, así como su construcción.

En el capítulo 6 se describirán 4 métodos de detección de factores: 2 bajo el enfoque frecuentista de la estadística y los otros 2 bajo el enfoque bayesiano. En el capítulo 7 se presenta un ejemplo de estas 4 técnicas aplicadas a un experimento realizado de acuerdo a los diseños definitivos descritos en el capítulo 5. En el capítulo 8 están contenidas las conclusiones y relevancia del trabajo, en las que se recopila el desarrollo de los temas expuestos y se comparan los métodos utilizados en el capítulo 7.

[1]Por ejemplo, si un factor influye de manera cuadrática en la respuesta, también deberá influir de manera lineal, pues se asume que es poco probable haber escogido los puntos del experimento exactamente simétricos al vértice de la respuesta cuadrática tal que la pendiente de la respuesta entre ellos sea cero, pero que varíe en puntos intermedios

Finalmente, en la sección de apéndices se encuentran los detalles más técnicos de los temas desarrollados y se muestran matrices de diseño asociadas a ciertas clases específicas de diseños de experimentos.

Capítulo 2

Análisis de modelos estadísticos

En este capítulo se presentará el análisis que se realizará a los resultados experimentales, esto incluye el tipo de modelo a ajustar y los supuestos realizados (tanto del experimento como del fenómeno). Se presentará primero el caso de un factor para mostrar la idea general del análisis y se continuará con el modelo de 2 factores finalizando con la generalización a k factores, la descripción del análisis concluirá con un ejemplo. Por último, se describirán los cambios en el análisis en diseños por bloques y se definirán criterios de optimalidad que son comúnmente utilizados para evaluar distintos diseños de experimentos.

2.1. Análisis del modelo lineal

2.1.1. Definición del modelo

El diseño del experimento depende del modelo que se intentará ajustar, por lo que se deberá proponer antes de realizar los ensayos. En este trabajo, se supondrá que el fenómeno puede ser descrito mediante

un modelo lineal en los parámetros y que los factores son de naturaleza cuantitativa, es decir variables continúas. Se utilizará el *modelo de efectos fijos*, que en el caso de dos parámetros, se define de la siguiente manera:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \tag{2.1}$$

Donde y_{ijk} es la respuesta de la repetición k cuando el primer factor está en el nivel i (cuyo efecto se denota por τ_i) y el segundo en el nivel j (β_j), el término $(\tau\beta)_{ij}$ representa el efecto de interacción entre ambos factores, μ es la media general de la respuesta Y ϵ_{ijk} es el error del modelo (causado por variabilidad desconocida e incontrolable del fenómeno).

Los efectos se expresarán como desviaciones de la media por lo que la expresión (2.1) se debe complementar con las *restricciones de estimabilidad*, que tienen este nombre porque serán necesarias para estimar el efecto en la respuesta cada nivel de los factores:

$$\Sigma_i^a \tau_i = 0$$
$$\Sigma_i^b \beta_j = 0$$
$$\Sigma_i^a (\tau_i \beta_j) = \Sigma_j^n (\tau_i \beta_j) = 0$$

A diferencia de modelos clásicos de regresión, en que se asume una relación de la respuesta con el factor, en este trabajo no se asume ningun comportamiento especifico del factor con la respuesta, sino que se estima la respuesta en cada nivel factor. El efecto de cada nivel podrá luego ser graficado para observar el tipo de relación y poder luego proponer modelos del tipo $y = a + b\tau$ o $y = a + b\tau + c\tau^2$ por ejemplo.

Para que el modelo no esté sesgado desde su construcción y se pueda considerar representativo del fenómeno y las conclusiones derivadas sean válidas, se harán los siguientes supuestos:

i) Los errores ϵ_{ijk} son independientes, idénticamente distribuidos (i.i.d) con media cero y varianza finita[1].

[1]Frecuentemente se hará el supuesto adicional $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$

ii) Los niveles de cada factor fueron fijados antes del experimento y no son resultado de una muestra aleatoria de un conjunto con un mayor número de niveles.

iii) El orden de los ensayos se determina de manera aleatoria, esto para evitar efectos sistemáticos subyacentes no considerados explícitamente en el modelo.

Debido a que los niveles de los factores están fijos, el análisis del modelo se centrará en obtener conclusiones sobre la respuesta en cada nivel de los factores y las conclusiones son sólo validas para los niveles estudiados. Si el modelo fuera de *efectos aleatorios*, los niveles a estudiar serían una muestra aleatoria de un conjunto con un mayor número de niveles y el objetivo del estudio sería extender las conclusiones a los demás factores, por lo que el análisis se centra en la variabilidad de la respuesta en los distintos niveles [Montgomery, 2001, cap. 3].

2.1.2. Análisis del modelo con un factor

El modelo lineal de un solo factor con a niveles y n repeticiones en cada nivel es:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \qquad \begin{matrix} i = 1, \ldots, a \\ j = 1, \ldots, n \end{matrix} \tag{2.2}$$

Se buscará estimar el efecto del factor en cada nivel y probar si el factor influye en la respuesta de manera significativa, es decir, que el cambio en la respuesta correspondiente a cambios en el nivel del factor se pueda diferenciar del ruido causado por el error aleatorio ϵ_{ij}. Esta idea se formaliza mediante una prueba de hipótesis en la que el factor no será significativo si el efecto en todos sus niveles es el mismo, es decir:

$$H_0 := \tau_1 = \tau_2 = \cdots = \tau_a \quad vs. \quad H_1 := \tau_i \neq \tau_j \text{ para algún } i \neq j \tag{2.3}$$

Bajo el supuesto que μ es la media general, si todos los efectos fueran iguales, estarían completamente considerados en μ y la prueba de hipótesis (2.3) es equivalente a:

$$H_0 := \tau_1 = \tau_2 = \cdots = \tau_a = 0 \quad vs. \quad H_1 := \exists\ i \text{ tal que } \tau_i \neq 0 \tag{2.4}$$

Estas pruebas requerirán de la estimación τ_i y de conocer la distribución de dichos estimadores, la cual depende de la forma del modelo y de la distribución de los errores, ya que son la única fuente de variabilidad.

Ecuaciones normales y estimadores

El uso de un modelo lineal permite utilizar el método de mínimos cuadrados para calcular los estimadores de los efectos y que estos tengan la propiedad de ser insesgados y de varianza mínima [Draper and Smith, 1998, cap. 5] y la respuesta estimada será ortogonal a los residuos (demostración en el apéndice A.1), además con el supuesto que $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, este estimador coincidirá con el de máxima verosimilitud.

El modelo descrito por la ecuación (2.2) sirve para mostrar la relación funcional de los factores con la respuesta, sin embargo no resulta muy útil para escribir todos los resultados del experimento en una ecuación. En su lugar se utilizará la forma matricial (2.5), que al ser una forma compacta, permite la manipulación simultanea de todos los datos con operaciones sencillas.

$$Y = X\beta + \mathcal{E}, \quad X = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ & & \vdots & & \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ & & \vdots & & \\ 1 & 0 & 1 & \dots & 0 \\ & & \vdots & & \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_a \end{bmatrix} \tag{2.5}$$

Donde Y es el vector con las na respuestas observadas y X es una matriz de na filas y $a+1$ columnas, conocida como la **matriz del modelo**[2], y β es el vector con los p parámetros que determinan el modelo $(a+1)$.

[2]Se diferencia la **matriz del modelo** de la *matriz del diseño o experimento* en

Con está notación, se pueden escribir a los residuos como función de los parámetros de la siguiente forma:

$$f(\beta) = (Y - \hat{Y})^T(Y - \hat{Y}) = (Y - X\beta)^T(Y - X\beta)$$

Cuyo mínimo se encuentra al igualar las derivadas parciales a cero:

$$\begin{aligned} &\left.\frac{\partial f}{\partial \beta}\right|_{\beta=\hat{\beta}} = 0 \\ &\Rightarrow 2X^TX\hat{\beta} - 2X^TY = 0 \\ &\Rightarrow X^TX\hat{\beta} = X^TY \end{aligned} \tag{2.6}$$

A la ecuación (2.6) se le conoce como las ecuaciones normales y la unicidad de la solución dependerá de que X sea de rango completo, sin embargo la matriz (2.5) no lo es, pues la suma de todas sus columnas es igual a dos veces la primera y será necesario añadir *restricciones de estimabilidad* mencionadas anteriormente.

Las ecuaciones normales en su forma matricial son abstractas y no ayudan a proponer una expresión general para los estimadores. Montgomery et al. [Montgomery, 2001, cap. 3] proponen 3 reglas para escribir las ecuaciones normales en términos de la respuesta y los parámetros. Estas reglas se basan en que la matriz X describe solamente si un parámetro está presente en la respuesta o no, por lo que X^TX y X^TY tienen una forma predeterminada y simplemente definen sumas sobre ciertos renglones. Las reglas son:

i) Hay una ecuación normal para cada parámetro del modelo que se va a estimar

ii) El lado derecho de cualquier ecuación es simplemente la suma de todos los ensayos que contienen el parámetro asociado a esa ecuación

iii) El lado izquierdo de cualquier ecuación es la suma de todos los parámetros del modelo multiplicado por el número de veces que aparece

---

que la matriz del diseño muestra en que configuración se realizó cada ensayo, pero no contiene a todos los parámetros del modelo, por ejemplo: la interacción $\tau_i\beta_j$, se encontrará en la matriz del modelo, pero no en la de diseño

Aplicando estas reglas, las ecuaciones normales en términos de los parámetros y asociadas a cada estimador son:

$$\hat{\mu} \leftarrow (na)\mu + n\tau_1 + \cdots + n\tau_a = \sum_{i,j} y_{ij} \tag{2.7}$$

$$\hat{\tau_i} \leftarrow (n)\mu + n\tau_1 = \sum_{j} y_{ij} \tag{2.8}$$

Anteriormente, se mencionó que X deberá ser de rango completo para que la solución a (2.6) sea única. En las ecuaciones (2.7) y (2.8), es fácil observar que la suma de todas las ecuaciones relacionadas a los estimadores de τ_i es igual a la ecuación del estimador de μ, por lo que es un sistema de ecuaciones no linealmente independiente y no proviene de una matriz de rango completo.

Para romper la dependencia lineal, se agrega la restricción de estimabilidad que los efectos son desviaciones de la respuesta general: i.e. $\sum_{i=1}^{a} \tau_i = 0$.

En las ecuaciones normales de (2.6), esta restricción se incorpora al agregar a la matriz X en (2.5) un renglón con cero en la primer columna y unos en todas las demás y agregando de igual manera al vector Y un cero. En las ecuaciones normales descritas como en (2.7) y (2.8), simplemente se agrega la restricción de estimabilidad al sistema, obteniendo los siguientes estimadores:

$$\hat{\mu} = \sum_{i,j} y_{ij}/na \tag{2.9}$$

$$\hat{\tau_i} = \sum_{j} y_{ij}/n - \hat{\mu} \tag{2.10}$$

El supuesto de normalidad en los errores implica que la respuesta sigue una distribución normal y los estimadores, siendo combinación lineal de la respuesta, seguirán también una distribución normal con parámetros: [Rice, 2007, cap. 14]

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E[Y] \\ &= (X^T X)^{-1} X^T X \beta = \beta \end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= Var((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^T\sigma^2\mathbb{I}X(X^TX)^{-1}) \\ &= \sigma^2(X^TX)^{-1} \end{aligned} \tag{2.11}$$

En la última ecuación se asume que los errores son independientes e idénticamente distribuidos y que X ha sido modificada con las restricciones de estimabilidad para que X^TX sea invertible. Al igual que en las ecuaciones normales, la última ecuación resulta poco fácil de interpretar, por lo que será más útil utilizar los estimadores de (2.9) y (2.10) en el cálculo de la varianza:

$$\mathrm{Var}(\hat{\mu}) = Var\left(\sum_{i,j} y_{ij}/na\right) = \sigma^2/na$$

$$\mathrm{Var}(\hat{\tau_i}) = Var\left(\sum_{j} y_{ij}/n - \hat{\mu}\right) = \sigma^2/n + \sigma^2/na = \sigma^2(a+1)/na$$

Normalmente la varianza será desconocida, por lo que hará falta un estimador de esta. Se asumirá que la diferencia entre la respuesta estimada y la observada proviene solamente de ϵ, por lo que se utilizará a la expresión (2.12) como estimador de la varianza.

$$\hat{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.12}$$

Donde N es el número total de ensayos, p el número total de parámetros (columnas en la matriz del modelo) y la suma se realiza sobre todos los ensayos. En el apéndice A.2 se demuestra que este estimador es insesgado, además, es un término que aparece naturalmente al descomponer la variabilidad de la respuesta en la explicada por el modelo y la residual, denotado por $\mathrm{MS_E}$ en el análisis de la varianza, que se verá más adelante.

Conocer la distribución de los estimadores permitirá realizar pruebas de hipótesis y calcular intervalos de confianza para cada efecto (y con algún ajuste, por ejemplo: el método de Bonferroni [Rice, 2007, cap. 11] se podrán realizar pruebas para un conjunto de estimadores). La prueba de hipótesis (2.4) se aplica a todo el modelo para probar si algún nivel

del factor es distinto de 0 y poder concluir que el modelo explica a los datos mejor que la pura aleatoriedad de los errores. Esta prueba se realiza frecuentemente con la ayuda de una tabla ANOVA (Tabla 2.1), en la que se compara la variabilidad explicada por el modelo contra la causada por los errores.

Análisis de la variabilidad del modelo

Para probar si el modelo explica satisfactoriamente la variabilidad de los datos, se realiza la prueba de hipótesis (2.4), en la que se compara si algún nivel del factor tiene un efecto significativo[3] en la respuesta o si la variabilidad se puede atribuir simplemente al error del modelo. La variabilidad se medirá respecto a la media de la siguiente manera:

$$\mathrm{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2 \tag{2.13}$$

Donde la notación con punto significa "sobre todos los valores del indice", de manera que $\bar{y}_{\cdot\cdot}$ es el promedio de la respuesta sobre todos los niveles y sobre todas las repeticiones. Siguiendo esta notación, es posible reescribir la ecuación pasada como:

$$\begin{aligned}
\mathrm{SS}_T &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2 + 2\sum_{i=1}^{a}(n\bar{y}_{i\cdot}^2 - n\bar{y}_{i\cdot}^2 - n\bar{y}_{i\cdot}\bar{y}_{\cdot\cdot} + n\bar{y}_{i\cdot}\bar{y}_{\cdot\cdot}) \\
&\quad + \sum_{i=1}^{a} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{a} n(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2
\end{aligned}$$

[3]Significativo en el contexto estadístico: que la probabilidad del estimador obtenido, *dada* la hipótesis nula, sea lo suficientemente pequeña para que el investigador decida que probablemente la hipótesis nula sea falsa

$$= \sum_{i=1}^{a} n\hat{\tau}_i^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 = \mathrm{SS}_N + \mathrm{SS}_E$$

El primer término puede ser interpretado como la suma de cuadrados de los efectos del factor A (como cambia la respuesta en el nivel i en comparación a la media general) y el segundo término como la suma de cuadrados correspondiente a los residuos en cada nivel (suma el error dentro de cada nivel). Esto indica que la variabilidad total es la suma de la variabilidad causada por el cambio en los niveles del factor y la variabilidad de la respuesta dentro de cada nivel. El objetivo del modelo es explicar la variabilidad de la respuesta a través del factor, por lo que se buscará que SS_N sea más grande que SS_E.

Pruebas de hipótesis sobre los efectos

Para realizar la prueba de hipótesis (2.4) y concluir si gran parte de la variabilidad de la respuesta puede ser explicada mediante el factor, se utilizará la razón entre SS_N y SS_E y una distribución de referencia. Existen (al menos) dos distribuciones de referencia que pueden ser utilizadas: la primera se conoce como **prueba de aleatorización**, en la cual se asume que los factores no tienen efecto sobre la respuesta y que son indistinguibles, de manera que se calcula SS_N y SS_E para todas las combinaciones de respuesta y niveles del factor. Esta tiene la ventaja que no requiere el supuesto de normalidad en los errores y la desventaja que se basa en combinaciones, por lo que tiene una complejidad computacional factorial. El segundo método requiere del supuesto de normalidad en los errores, utiliza una distribución conocida y se ha probado que incluso sin el supuesto de normalidad éste método se puede ver como una aproximación a la prueba de aleatorización [Montgomery, 2001, cap. 3].

Ambas pruebas hacen uso del cociente entre los **cuadrados medios**, que se obtienen al dividir la suma de cuadrados por sus respectivos **grados de libertad**. Los grados de libertad se refieren al número de datos independientes utilizados, por ejemplo SS_N tiene $a - 1$ grados de libertad, pues dada $\bar{y}_{\cdot\cdot}$ implicará que una vez observadas las primer

$(a-1)$ $\bar{y}_{i\cdot}$, se determine $\bar{y}_{a\cdot}$.

La información sobre la variabilidad de la respuesta se puede resumir en la siguiente tabla de "análisis de la varianza" también conocida como tabla ANOVA (acrónimo en inglés ANalysis Of VAriance). El estadístico de prueba (en ambos métodos) se denota por F_0.

Tabla 2.1: Tabla ANOVA

Tipo de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrados medios	F_0
Entre niveles	SS_N	$a-1$	$\mathrm{MS}_N = \frac{\mathrm{SS}_N}{a-1}$	$F_0 = \frac{\mathrm{MS}_N}{\mathrm{MS}_E}$
Error	SS_E	$an-a$	$\mathrm{MS}_E = \frac{\mathrm{SS}_E}{na-a}$	
Total	SS_T	$an-1$		

En las pruebas de aleatorización, la distribución de referencia se obtiene al calcular F_0 para todas las permutaciones posibles sobre las respuestas que se asignan al nivel i (bajo la hipótesis nula todos los efectos son iguales, por lo que el ensayo y_{ij} es indistinguible del ensayo y_{kj}) y si se tienen a factores y n repeticiones, será necesario calcular F_0 para las $(an)!/(a \cdot n!)$ permutaciones y la significancia de la prueba será el porcentaje de permutaciones cuyo estadístico de prueba resultó mayor o igual que F_0 observado.

El otro método es conocido como pruebas F, ya que el supuesto de normalidad de los errores implica que SS_N/σ^2 y SS_E/σ^2 siguen una distribución Ji-Cuadrada con $a-1$ y $a(n-1)$ grados de libertad respectivamente [Rice, 2007, cap. 6].

Además, aunque se tenga la relación $\mathrm{SS}_T = \mathrm{SS}_N + \mathrm{SS}_E$, el teorema de Cochran (2.1.1) [Montgomery, 2001, cap. 3] garantiza que $\mathrm{SS}_N + \mathrm{SS}_E$ son independientes, por lo que el cociente F_0 seguirá una distribución $F_{a-1,a(n-1)}$[Rice, 2007, cap. 6].

Note que la suma de los grados de libertad de $\mathrm{SS}_N/\sigma^2 + \mathrm{SS}_E/\sigma^2$ es igual a los grados de libertad de SS_T/σ^2; el cual a su vez, es equivalen-

Teorema 2.1.1 (Teorema de Cochran)**.**
Sean $Z_i \sim N(0,1)$ *independientes para* $i = 1, \ldots, n$ *y*

$$\sum_{i=1}^{n} Z_i^2 = Q_1 + Q_2 + \cdots + Q_s$$

Con $s \leq n$*. Entonces* $Q_1, \ldots, Q_s$ *son variables aleatorias independientes Ji-Cuadradas con* ν_i *grados de libertad si y solo si*

$$n = \nu_1 + \nu_2 + \cdots + \nu_s$$

te a la suma de $an - 1$ variables aleatorias normales estándar, por lo que se cumplen las condiciones del teorema y el cociente F_0 sigue una distribución $F_{a-1,a(n-1)}$.

Finalmente, para que el modelo sea válido será importante revisar que se cumplan los supuestos, sobretodo que los residuos no tengan ningún comportamiento sistemático, que sean independientes e idénticamente distribuidos normal (o aproximadamente normal), sobre esto se hablará en la sección de verificación visual de supuestos.

2.1.3. Análisis del modelo con 2 factores

En esta sección se generalizarán los métodos expuestos en la sección pasada a modelos que incluyan solo 2 factores. La principal diferencia contra los modelos de un factor será que, al incluir el segundo factor, se deberá considerar el efecto de la interacción entre ambos, tanto en los efectos, como en la descomposición de la varianza. El modelo de dos factores está expresado por:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \qquad \begin{matrix} i = 1, \ldots, a \\ j = 1, \ldots, b \\ k = 1, \ldots, n \end{matrix}$$

Y las ecuaciones normales se pueden deducir utilizando las reglas propuestas por Montgomery y mencionadas en la sección anterior, para

obtener el siguiente sistema:

$$\hat{\mu} \leftarrow abn\mu + bn\sum_{i}^{a}\tau_i + an\sum_{j}^{b}\beta_j + n\sum_{i,j}(\tau\beta)_{ij} = y_{...} \tag{2.14}$$

$$\hat{\tau_i} \leftarrow bn\mu + bn\tau_i + n\sum_{j}^{b}\beta_j + n\sum_{j}(\tau\beta)_{ij} = y_{i..} \tag{2.15}$$

$$\hat{\beta_j} \leftarrow an\mu + n\sum_{i=1}^{a}\tau_i + an\beta_j + n\sum_{i}(\tau\beta)_{ij} = y_{.j.} \tag{2.16}$$

$$\hat{\tau\beta}_{ij} \leftarrow n\mu + n\tau_i + n\beta_j + n(\tau\beta)_{ij} = y_{ij.} \tag{2.17}$$

Y al igual que el caso con un solo factor, las ecuaciones no son linealmente independientes: al sumar sobre todas las i o sobre todas las j en la ecuación (2.17), se obtiene la ecuación (2.16) o (2.15) respectivamente; y al sumar sobre ambos indices, se obtiene la ecuación (2.14). Al igual que en el caso de un factor es necesario definir restricciones extras para que la solución sea única; utilizando la misma idea que μ deberá captar el efecto en común de todos los niveles, las restricciones de estimabilidad son:

$$\sum_{i}\tau_i = 0 \qquad \sum_{j}\beta_j = 0$$
$$\sum_{i}(\tau\beta)_{ij} = \sum_{j}(\tau\beta)_{ij} = 0$$

Con las que se puede obtener la solución de las ecuaciones normales directamente al sustituir las restricciones de estimabilidad en la ecuación de $\hat{\mu}$ y sustituyendo "hacia abajo":

$$\hat{\mu} = \bar{y}_{...}$$
$$\hat{\tau_i} = \bar{y}_{i..} - \hat{\mu}$$
$$\hat{\beta_j} = \bar{y}_{.j.} - \hat{\mu}$$
$$\hat{\tau\beta}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

Una propiedad de estos estimadores es que la media de las repeticiones es igual al valor estimado de la respuesta en cada configuración:

$$\bar{y}_{ij.} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) \tag{2.18}$$

$$= \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \widehat{(\tau\beta)}_{ij} = \hat{y}_{ij}$$

Para descomponer la variabilidad total en la causada por cada efecto, se utilizará como paso intermedio que la variabilidad de cada ensayo se puede descomponer en la desviación de la configuración a respecto a la media general más la variabilidad de la observación respecto a la media de la configuración.

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{ij.} - \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

Con lo que se puede escribir la desviación de la respuesta sobre la media como:

$$SS_T = \sum_{i,j,k} (y_{ijk} - \bar{y}_{...})^2 \tag{2.19}$$

$$= \sum_{i,j,k} [(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})]^2$$

$$= bn \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + an \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 + n \sum_{i,j} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$+ \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij.})^2$$

$$= SS_A + SS_B + SS_{AB} + SS_E \tag{2.20}$$

Pues cada uno de los 4 en 2 combinaciones[4] (= 6) de términos cruzados son cero[5]. El hecho que los términos cruzados sean igual a cero se debe a que el experimento es *balanceado*, es decir con el mismo número de repeticiones sobre cada configuración, lo que implica que los estimadores de los efectos factores e interacciones sean ortogonales.

La suma total de la variabilidad puede ser descompuesta en la variabilidad explicada por el factor A, por el factor B, por la interacción

[4]En adelante, se denotará como nCk al coeficiente binomial "n en k", el número de formas en que se pueden seleccionar k elementos de un conjunto de n

[5]La demostración es directa, pero una manera rápida de convencerse que es cierto es que al sumar sobre todos los indices y cambiando el orden de la sumas, los términos cruzados se pueden expandir en la suma y resta de $\bar{y}^2_{...}$, $\bar{y}^2_{i..}$, $\bar{y}^2_{.j.}$ o $\bar{y}^2_{ij.}$, dependiendo el producto

AB y por el error aleatorio. La Tabla 2.2 muestra la tabla ANOVA correspondiente al análisis del modelo con dos factores.

Tabla 2.2: Tabla ANOVA para 2 factores

Tipo de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico de prueba
Factor A	SS_A	$a-1$	$\frac{\mathrm{SS}_A}{a-1}$	$\frac{\mathrm{MS}_A}{\mathrm{MS}_E}$
Factor B	SS_B	$b-1$	$\frac{\mathrm{SS}_B}{b-1}$	$\frac{\mathrm{MS}_B}{\mathrm{MS}_E}$
Inter. AB	SS_{AB}	$(a-1)(b-1)$	$\frac{\mathrm{SS}_{AB}}{(a-1)(b-1)}$	$\frac{\mathrm{MS}_{AB}}{\mathrm{MS}_E}$
Error	SS_E	$ab(n-1)$	$\frac{\mathrm{SS}_E}{nab-ab}$	
Total	SS_T	$abn-1$		

Note que por la ecuación (2.20) y la suma de grados de libertad de cada término, el teorema de Cochran (2.1.1) garantizará que todos los términos de la descomposición de la suma de cuadrados son independientes, en particular SS_E será independiente de SS_A, SS_B y SS_{AB}, por lo que el estadístico de prueba de cada efecto seguirá una distribución F con los grados de libertad correspondientes a su renglón.

2.1.4. Generalización a k factores

Para generalizar los estimadores a k factores, se utilizarán las reglas propuestas por Montgomery para escribir las ecuaciones normales y se cambiará un poco la notación. Los factores se denotarán por $A_1, A_2, \ldots, A_k$ y cada factor tendrá $a_1, a_2, \ldots, a_k$ niveles, de manera que, el efecto de una interacción de orden l se va a denotar por:

$$\tau_{\mathcal{I}}^{\mathcal{J}} \quad \begin{cases} \mathcal{J} \subseteq \{A_1, \ldots, A_k\} & , |\mathcal{J}| = l \\ \mathcal{I} \in \{1, \ldots, |\mathcal{J}[1]|\} \times \cdots \times \{1, \ldots, |\mathcal{J}[l]|\} \end{cases}$$

Es decir, el conjunto de factores activos en la interacción se colocará en el superíndice y el conjunto con los niveles en los que se está evaluando

cada factor se colocará en el subíndice. En esta notación, $\mathcal{J}[i]$ representa al i-ésimo factor de la interacción y $|\mathcal{J}[i]|$ representa el número de niveles de dicho factor. Esta notación es simplemente una forma compacta de escribir que la interacción contiene l factores y que están siendo evaluados en los niveles dados en $\mathcal{I}$. Como caso particular se definirá que $\mathcal{J} = \emptyset$ representará al parámetro μ.

Para obtener el estimador de cada interacción se utilizarán las reglas de Montgomery para obtener las ecuaciones normales correspondientes. La tercera regla de Montgomery dice que el lado izquierdo de las ecuaciones normales (2.6) es la suma de los distintos parámetros, multiplicados por el número de veces que aparecen.

A su vez, estos parámetros se pueden separar en fijos y libres con base en el parámetro que corresponde a esa ecuación normal; en el ejemplo de dos factores (sean τ_i y β_j), en la ecuación normal correspondiente a τ_i, μ y τ_i están fijos y el subíndice "j" (β_j y $(\tau\beta)_{ij}$) está libre, por lo que se suma sobre todos sus valores (véase (2.15)) y las restricciones que se añaden para garantizar solución única, igualan dichas sumas a cero.

Las restricciones de estimabilidad añadidas implican que las ecuaciones normales solo contienen los términos con índices fijos y pueden ser descritas por la siguiente ecuación (2.21):

$$\tau_{\mathcal{I}}^{\mathcal{J}} \leftarrow \sum_{i=0}^{l} \sum_{\substack{S \subseteq \mathcal{J} \\ |S|=i}} N_S \tau_{\mathcal{I}_S}^{S} = y_{\mathcal{I}\cdot} \tag{2.21}$$

Donde $\mathcal{J}$ es el conjunto con los l factores que definen la interacción, $\mathcal{I}_S$ son los indices en los que se encuentran los factores de S y $y_{\mathcal{I}\cdot}$ denota que se suma la respuesta sobre todos los subíndices que no están fijados por $\mathcal{I}$, N_S es la cardinalidad de esa suma (i.e $\prod a_i$ tal que $a_i \notin S$).

Escribir las ecuaciones de esta manera permite calcular los estimadores recursivamente:

$$N\hat{\mu} = y_{\cdot}$$

$$\hat{\tau}_{\mathcal{I}}^{\mathcal{J}} = \bar{y}_{\mathcal{I}\cdot} - \sum_{i=0}^{l-1} \sum_{\substack{S \subseteq \mathcal{J} \\ |S|=i}} \hat{\tau}_{\mathcal{I}_S}^{S}$$

Y por último, se puede resolver esta relación en términos de y solamente:

$$\hat{\tau}_{\mathcal{I}}^{\mathcal{J}} = \sum_{i=l}^{0} \sum_{\substack{S \subseteq \mathcal{J} \\ |S|=i}} (-1)^{l-i}\, \bar{y}_{\mathcal{I}_S\cdot}$$

Se puede verificar directamente que el modelo completo con los estimadores calculados de esta manera tienen la propiedad que la respuesta estimada de una configuración es igual al promedio de las respuestas observadas en esa configuración (i.e de las repeticiones como en (2.18)).

Y la tabla ANOVA se construye de igual manera que en (2.2), con los efectos principales con $a_i - 1$ grados de libertad, los de segundo orden con $(a_i - 1)(a_j - 1)$, los de tercer orden con $(a_i - 1)(a_j - 1)(a_k - 1)$ y así sucesivamente. Sin embargo, es poco común realizar estas tablas para todos los efectos principales e interacciones, por lo que se prefiere utilizar métodos gráficos para seleccionar a los efectos significativos y continuar la investigación con un menor número de factores.

Además, es posible construir una tabla ANOVA que separe la variabilidad en la explicada por el modelo completo (evitando la separación en cada tipo de efecto y permitiendo quitar interacciones de orden grande) y la variabilidad no explicada de la siguiente manera:

$$\begin{aligned} \mathrm{SS}_T &= (Y - \bar{Y})^T (Y - \bar{Y}) = (Y - \hat{Y} + \hat{Y} - \bar{Y})^T (Y - \hat{Y} + \hat{Y} - \bar{Y}) \\ &= (Y - \hat{Y})^T (Y - \hat{Y}) - 2(Y - \hat{Y})^T (\hat{Y} - \bar{Y}) \\ &+ (\hat{Y} - \bar{Y})^T (\hat{Y} - \bar{Y}) \\ &= (Y - \hat{Y})^T (Y - \hat{Y}) + (\hat{Y} - \bar{Y})^T (\hat{Y} - \bar{Y}) \\ &= SS_E + SS_M \end{aligned} \tag{2.22}$$

Donde SS_M representa la variabilidad explicada por *todo* el modelo, sin importar que factores e interacciones se incluyeron, con $p - 1$ grados de

libertad, donde p es el número de parámetros i.e. columnas en la matriz del modelo, incluyendo la de la media y la suma de cuadrados de error tiene $n-p$ grados de libertad. Note entonces que MS_E es el estimador de la varianza descrito en (2.12). La demostración que el producto cruzado es igual a 0 se encuentra en el apéndice A.3.

Supuesto sobre los efectos activos

En la introducción se mencionó que uno de los objetivos principales del diseño de experimentos es poder identificar entre efectos activos e inactivos. Sin embargo, intentar considerar todas las interacciones implica estimar un gran número de parámetros, que además, crece rapidamente (2^k). Por esta razón, es común realizar los siguientes dos supuestos para reducir la complejidad del modelo.

El primero se conoce como el **principio de pocos efectos**, en el que se asume que solo unos pocos efectos serán relevantes y que los sistemas estarán dominados por interacciones de orden bajo [Montgomery, 2001, cap. 6].

El segundo se conoce como *principio de herencia o jerárquico* a la idea que si en un modelo existen interacciones de orden mayor, entonces todos los efectos de orden menor de esos factores deben estar también incluidos en el modelo (e.g. si la interacción ABC está presente, entonces el modelo deberá incluir los efectos de A,B,C,AB,AC y BC). Cox explica en [Cox and N.Reid, 2000, cap. 5] que un modelo con interacción sin alguno de los efectos principales es artificial. Por ejemplo, un modelo en el que se suponga que la interacción AB está activa y que el factor B no tiene efecto en la respuesta, implicaría que B está activo solo en combinación de algunos niveles de A y que para cada nivel de B, la suma de estos efectos sea cero, lo cual se asume como poco realista.

2.2. Verificación visual de supuestos

Para que las pruebas F sean válidas se debe verificar que los supuestos sobre los errores se cumplan, lo que se hará a través de los residuos, que son la forma de observar los errores. La verificación visual consiste en realizar distintas graficas y que no sea posible observar un patrón en los residuos que sesge los estimadores.

Se analizarán dos tipos de violación de supuestos: en el primero, la forma funcional del modelo es incorrecta, los ensayos no son independientes, la varianza no es igual en todos los ensayos [6] o que existe algún otro factor que genera ruido sistemáticamente; en el segundo sólo se verificará que la distribución de los errores sea aproximadamente normal.

2.2.1. Forma correcta del modelo

Para verificar que no hayan violaciones graves en la forma de los errores, se acostumbra graficar los residuos observados contra la respuesta estimada, contra el orden de los ensayos y contra los factores. Cabe resaltar que las verificaciones gráficas no sirven para comprobar que los supuestos se cumplen, simplemente verifican que no hay desviaciones graves. A continuación se presentarán gráficas creadas a partir de datos simulados, que ejemplificarán violaciones a los supuestos.

Idealmente los residuos deberán seguir un comportamiento aleatorio en el que no se detecte ningún patrón y que estén todos dentro de una banda de ancho constante, como se puede observar en la Figura 2.1.

Graficar los residuos contra la respuesta estimada, ayudará a detectar un comportamiento no considerado en el modelo (e.g. cuadrático) y cuando la varianza esta correlacionada con la magnitud de la respuesta (e.g. el error de medición es un porcentaje del valor observado). Los ejemplos de ambos casos se pueden observar en la Figura 2.2 y en la Figura 2.3 respectivamente.

[6] A este problema se le conoce como *heteroscedasticidad*

Figura 2.1: Ejemplo del comportamiento adecuado en los residuos.

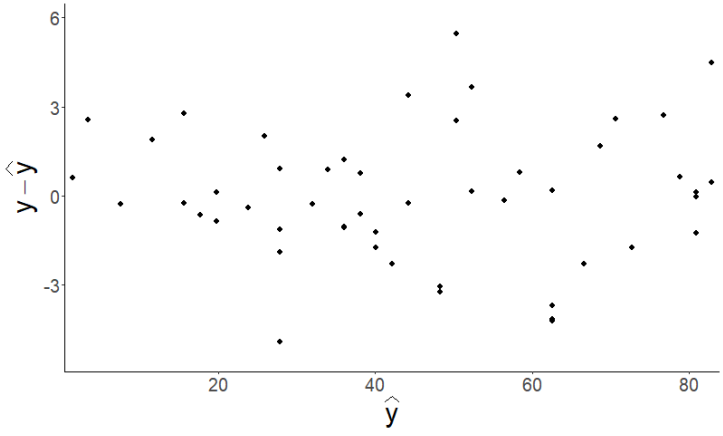


Figura 2.2: Comportamiento cuadrático.

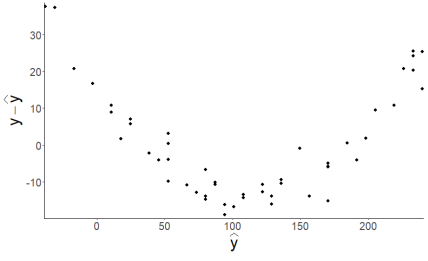
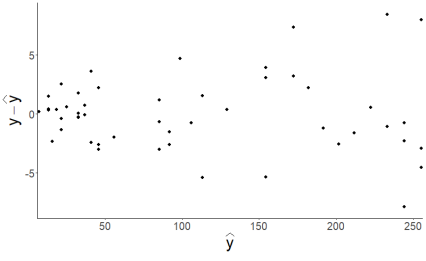
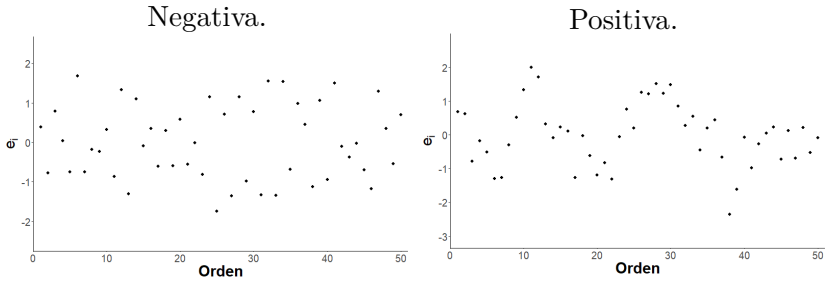


Figura 2.3: Varianza creciente con valores mayores



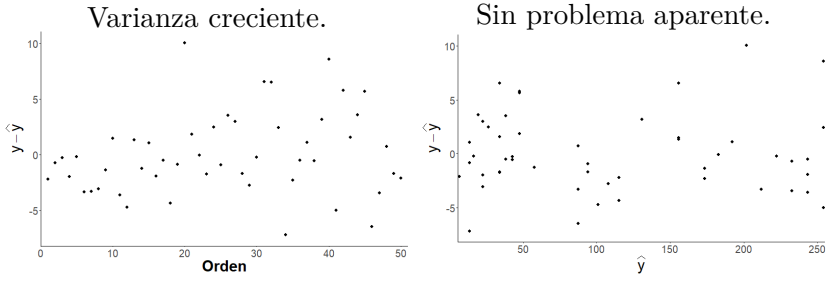
Las gráficas contra el orden de los ensayos serán útiles para detectar condiciones del experimento que afectan la respuesta, por ejemplo que la máquina acumule producto y que después de una producción pequeña se siga una grande (correlación negativa) o que la temperatura exterior afecte la respuesta, causando que en los ensayos secuenciales se observen consistentemente excesos o faltas respecto la media (correlación positiva); por último, podría ser que la máquina se caliente o que el experimentador vaya perdiendo el estado de alerta y la varianza del error aumente con el tiempo. Las gráficas que comparan los residuos contra el orden de ensayos se muestran en la Figura 2.4 y en la Figura 2.5.

Figura 2.4: Autocorrelaciones de los residuos.



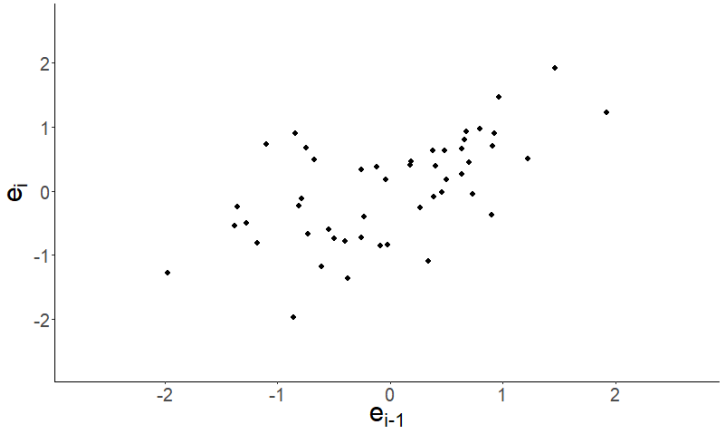
Para enfatizar que las gráficas respecto al orden de los ensayos proveen información diferente que contra la respuesta estimada, se muestra el problema de varianza no constante en el tiempo, graficado contra $\hat{y}$, donde no se alcanza a percibir el problema y contra su orden de ensayo, donde es más evidente. La autocorrelación no siempre existe con el en-

Figura 2.5: Varianza no constante en el tiempo.



sayo anterior, por ejemplo: si los ensayos ocurrieran por estaciones o por turno laboral. Por esta razón, también se grafican los residuos contra los observados i ensayos antes, donde i es el periodo en que el problema sugiera pueda haber auto correlación. En la Figura 2.6 se observa un ejemplo, donde se eligió $i = 1$, o sea el periodo anterior.

Figura 2.6: Correlación positiva con un periodo de retraso.



2.2.2. Distribución normal de los errores

El supuesto de normalidad en los errores puede ser verificado mediante las *gráficas normales* y las *gráficas cuantil-cuantil*, este tipo de gráficas también servirá para detectar visualmente qué efectos son significativos, pues bajo la hipótesis nula los estimadores de éstos siguen una distribución normal con media 0 y solo hará falta ver los que se desvíen demasiado.

Las gráficas normales se construyen a partir de la distribución empírica, calculada por: [Draper and Smith, 1998, ap. 2A]

$$F_{Emp}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(e_i) - 0.5}{n}$$

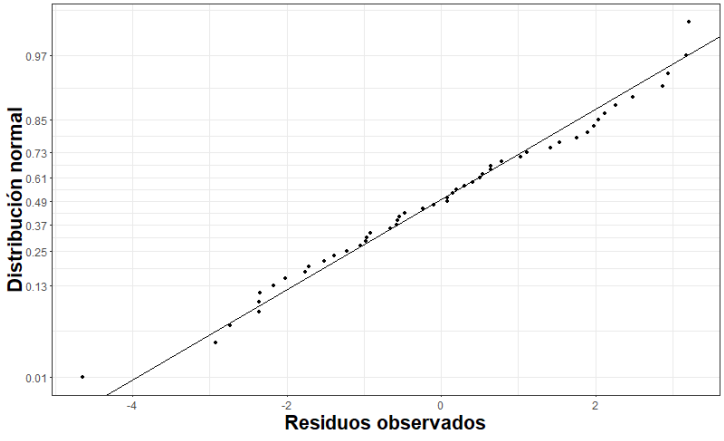
Donde $\mathbb{I}$ es la función indicadora, que toma el valor de 1 si el argumento está dentro del intervalo especificado en subíndice y 0 si está afuera.

Y se grafican las coordenadas $(e_i, F_{Emp}(e_i))$ en un plano donde la escala de las ordenadas no es lineal, sino que se utiliza a la distribución normal (la línea recta de pendiente 1 que pasa por el $(0, 0.5)$ representa a la distribución normal estándar)[7], de manera que, si los residuos siguen

[7] Antes, era común comprar un plano con dicha escala marcada, este era conocido

una distribución aproximadamente normal se encontrarán sobre una linea recta de pendiente $m \approx 1/\hat{\sigma}$. La Figura 2.7 muestra un ejemplo de esto cuando los errores siguen una distribución normal con varianza 4. En el ejemplo los puntos están razonablemente sobre la línea recta. Las

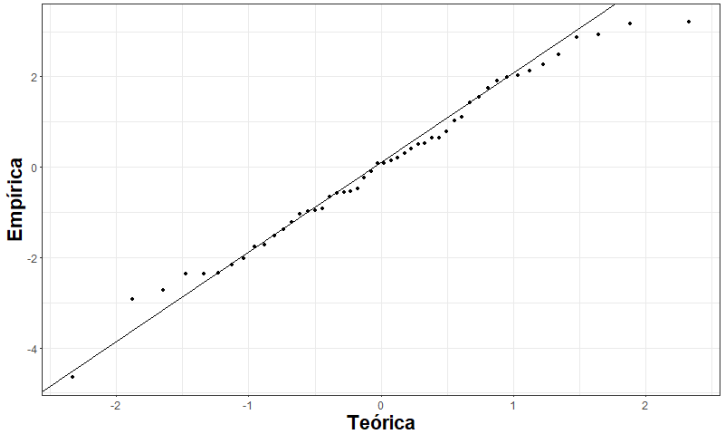
Figura 2.7: Gráfica normal



gráficas cuantil-cuantil sirven para comparar si dos conjuntos de datos de igual tamaño provienen de la misma distribución. Para construirlas se calcula la distribución empírica para cada conjunto de datos y se grafican las parejas de ensayos cuyo valor de la distribución empírica sea igual. Es decir, sean X y Y dos conjuntos de datos, se calcula la distribución empírica $F_X(X_i)$ y $F_Y(Y_i)$ y se grafican las coordenadas (X_i, Y_i) tal que $F_X(X_i) = F_Y(Y_i)$. En este tipo de gráficas, si los datos provienen de una misma distribución, los puntos estarán sobre la identidad; si la distribución tiene la misma forma, pero varia en su centro y dispersión, los puntos se encontrarán en una recta con una pendiente que refleje la razón de las dispersiones y pasará cerca de (μ_X, μ_Y). La Figura 2.8 muestra un ejemplo para dos poblaciones normales de distinta varianza.

como “papel normal”

Figura 2.8: Gráfica normal



2.2.3. Ejemplo

A continuación, se mostrará un ejemplo simulado para ejemplificar un problema que se resuelve con experimentación. Como el número de efectos a estimar crece rápidamente se utilizará un modelo con 2 factores y 3 niveles, dando un total de 9 configuraciones y 16 efectos a estimar, se utilizarán 4 repeticiones por configuración y en la simulación los errores siguen una distribución $\mathcal{N}(0, 4)$ y son independientes. Los datos se simularán a partir de la ecuación (2.23).

$$\begin{aligned} y =& 20 - 4\tau_1 + 4\tau_3 - 2\beta_1 + 4\beta_2 - 2\beta_3 - 1(\tau\beta)_{11} + (\tau\beta)_{12} \\ &+ 1(\tau\beta)_{21} - 3(\tau\beta)_{22} + 2(\tau\beta)_{23} + 2(\tau\beta)_{32} - 2(\tau\beta)_{33} + \epsilon \end{aligned} \tag{2.23}$$

En esta ecuación se deben de entender a las expresiones τ, β y $\tau\beta$ como funciones indicadoras, pues los modelos que se utilizarán en este trabajo no tienen coeficientes, simplemente se busca estimar el efecto que se tiene en la respuesta cuando los factores toman ciertos valores predeterminados. Note que en la ecuación del ejemplo (2.23) se cumplen las restricciones de estimabilidad mencionadas.

En este caso se consideró que lo factores podían tomar los niveles $\tau_i \in \{0, 1, 3\}$ y $\beta_j \in \{1, 3, 5\}$. Los valores particulares servirán para

poder dar un configuración que produzca algún resultado deseado o que vaya "en la dirección correcta" y no se debe entender que éstos valores se multiplican por los coeficientes de (2.23). Note además que algunos efectos son cero ($\tau_2, (\tau\beta)_{13}, (\tau\beta)_{31}$). Los resultados de la simulación se muestran en la Tabla 2.3.

Tabla 2.3: Resultados de la simulación

τ	β	$\bar{y}$	y_1	y_2	y_3	y_4
0	1	12.06	10.44	13.90	12.56	11.33
0	3	19.09	17.63	18.81	18.36	21.54
0	5	14.37	14.76	13.69	15.80	13.24
1	1	18.21	19.14	14.88	19.75	19.08
1	3	22.09	20.56	23.19	21.40	23.20
1	5	17.16	13.67	20.42	17.94	16.62
3	1	23.63	24.06	24.30	24.25	21.90
3	3	29.66	29.90	31.12	29.91	27.70
3	5	19.61	16.92	18.05	18.55	24.93

Los estimadores de mínimos cuadrados que se obtuvieron se presentan en la Tabla 2.4 junto con los valores reales como referencia.

Tabla 2.4: Efectos reales y estimados

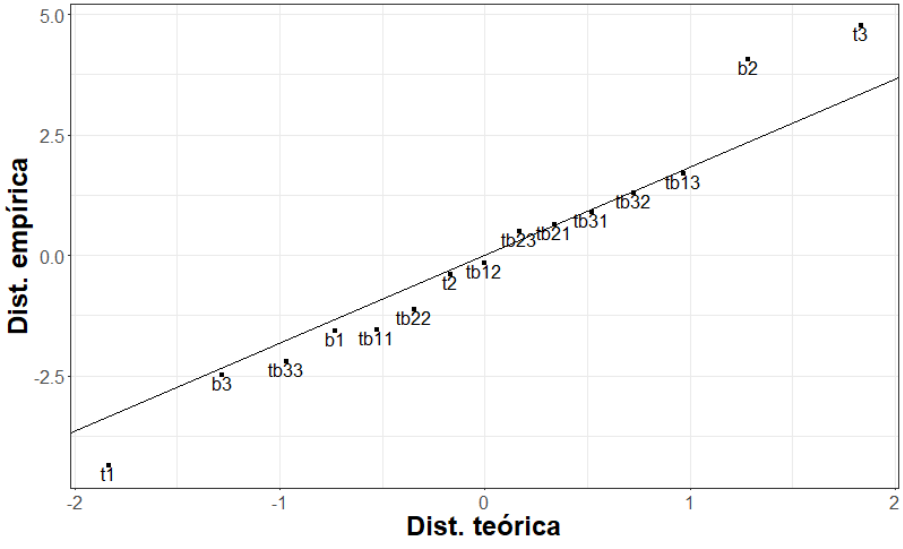
	μ	τ_1	τ_2	τ_3	β_1	β_2
Reales	20.00	-4.00	0.00	4.00	-2.00	4.00
Estimados	19.54	-4.37	-0.39	4.76	-1.58	4.07
	β_3	$(\tau\beta)_{11}$	$(\tau\beta)_{12}$	$(\tau\beta)_{13}$	$(\tau\beta)_{21}$	$(\tau\beta)_{22}$
	-2.00	-1.00	1.00	0.00	1.00	-3.00
	-2.49	-1.54	-0.16	1.69	0.63	-1.14
	$(\tau\beta)_{23}$	$(\tau\beta)_{31}$	$(\tau\beta)_{32}$	$(\tau\beta)_{33}$		
	2.0	0.0	2.00	-2.00		
	0.5	0.9	1.29	-2.19		

Algunos estimadores están bastante lejanos del valor real, por ejemplo $(\tau\beta)_{13}$, esto se explica en parte por que se eligió una varianza en los errores relativamente grande, casi en todos los casos mayor a la magnitud

del efecto. Además, el modelo que se buscó ajustar tiene más parámetros que los que "generaron" el fenómeno, por lo que están explicando a este conjunto particular de datos[8] y las restricciones que se imponen para garantizar la unicidad fuerzan los valores de algunos parámetros.

Para determinar que efectos son significativos se puede observar gráficamente los que presenten una desviación grande contra una distribución normal de media cero (con gráficas normales o cuantil cuantil) como en la Figura 2.9 y se puede realizar la prueba F como en se muestra en la Tabla 2.5.

Figura 2.9: Comparación de efectos contra una distribución normal



La Figura 2.9 se produce al asumir la hipótesis nula como cierta y suponer que todos los efectos provienen de la misma distribución de media cero y varianza σ^2. Una muestra de dicha distribución estará razonablemente sobre una línea recta de pendiente σ^2, de manera que si se traza una recta con esta pendiente que cruce las coordenadas $(0, 0)$,

[8] Al problema de tener parámetros de más que ajustan el modelo a los datos observados pero probablemente no serán muy buenos estimadores de datos futuros se le conoce como *sobreajuste* o *overfitting*

los efectos que posiblemente provengan de otra distribución podrán ser detectados al separarse de esta recta, en particular se supondrá que la desviación proviene de que la media es diferente de cero.

A primera vista se observa en la Figura 2.9 que los efectos τ_1, τ_3 y β_2 están alejados de la recta definida por la hipótesis nula, por lo que se tiene evidencia que sus efectos son distintos de cero; note que por definición del modelo son los únicos que cumplen $|\beta|/\sigma^2 \geq 1$, recalcando el punto que la varianza de los errores es relativamente grande. Para analizar si, considerando todos sus niveles, el factor τ y el factor β son significativos se realiza el análisis de la varianza, como se muestra en la Tabla 2.5.

Tabla 2.5: Tabla ANOVA

Término	SS	Grados de libertad	MS	F_0	Valor-p
τ	502.4	2	251.22	60.20	1.12e-10
β	303.0	2	151.52	36.31	2.22e-08
$(\tau\beta)$	58.0	4	14.50	3.47	2.05e-02
Residual	112.7	27	4.17		
Total	976.1	35	27.89		

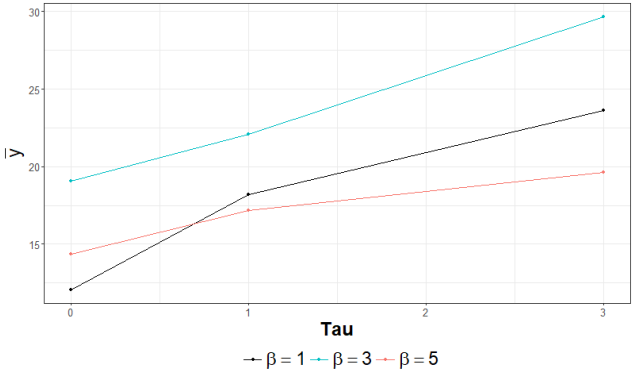
Visualmente, en la gráfica cuantil-cuantil, no se pudo confirmar que el efecto de la interacción es significativo, puees varios de los efectos se veían razonablemente sobre la línea recta[9]. Sin embargo, con el análisis de la varianza se podría decir que su suma de cuadrados tiene suficiente variabilidad para pensar que es distinta de cero. Una diferencia entre ambas pruebas es que la gráfica analiza *cada* efecto, mientra que la tabla de ANOVA responde la pregunta de si incluir el factor resultará en un modelo más preciso; además, aunque la prueba F concluya que el efecto

[9]El método gráfico de detectar errores depende mucho de una interpretación subjetiva de que es "cercano" a la recta. Esta subjetividad se ve de cierta manera reducida en las pruebas F al tener que preestablecer con que valor-p se rechazará la hipótesis nula

es distinto de cero, este puede ser lo suficientemente pequeño para que no haya interés práctico en incluirlo.

Las pruebas pasadas son útiles para responder preguntas *sobre el modelo*. Sin embargo, el interés real reside en poder contestar preguntas *sobre el fenómeno*; a continuación se muestran unas gráficas que podrían ser útiles para analizar el fenómeno y orientar futuras investigaciones. La primera de estas preguntas sobre el fenómeno es responder que comportamiento tiene la respuesta (hasta ahora resulta muy poco intuitivo), lo cuál se puede observar en la Figura 2.10.

Figura 2.10: Respuesta promedio por configuración.

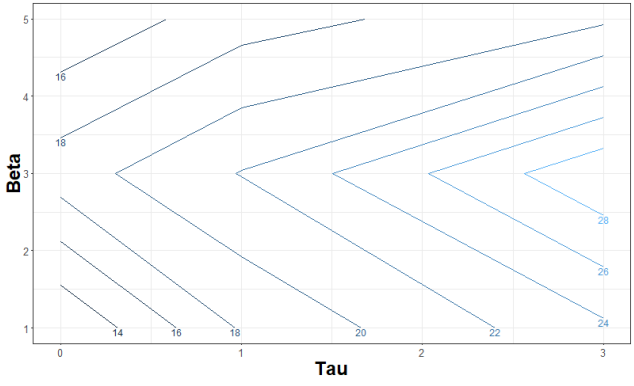


En esta figura se observa la respuesta promedio para los distintos niveles de τ y de β, la interacción entre ambos factores se detecta en el cambio de la pendiente que hay entre los distintos niveles de β, que siguieren que el modelo no es aditivo, es decir el cambio en la respuesta al variar τ va a estar también influenciado por el nivel de β; si la pendiente cambiará de signo se diría además que la interacción es *cualitativa* [Cox and N.Reid, 2000, cap. 5].

Por último, también resulta de gran utilidad graficar las curvas de nivel, que se obtienen al utilizar el valor medio de la respuesta en cada configuración (τ_i, β_j) para aproximar una superficie de respuesta, i.e. se asume que la respuesta entre dos puntos observados varía de manera líneal. En la Figura 2.11, se muestra un ejemplo de las curvas de nivel

obtenidas en este experimento, donde se puede apreciar que β sigue un comportamiento aproximadamente cuadrático y τ aproximadamente lineal, además si se quisiera encontrar las condiciones del experimento en que la respuesta sea, por ejemplo: 35, se buscaría experimentar al rededor de $\tau = 4$ y $\beta = 3$. En esta gráfica también se pueden apreciar los ligeros efectos de interacción, pues las curvas no son completamente paralelas, aunque parece que este efecto se puede descartar sin mayores implicaciones.

Figura 2.11: Curvas de nivel para la respuesta.



Una última nota sobre este ejemplo es que los datos fueron simulados por lo que: la forma funcional del modelo que se va a ajustar coincide exactamente con la forma de generar los datos. En una aplicación práctica se debería justificar de alguna manera porque se eligió ese tipo de modelo. Además, los errores se simularon como variables aleatorias independientes idénticamente distribuidas, por lo que las gráficas de análisis de residuos no muestran ningún comportamiento anormal.

2.3. Diseño de experimentos

Hasta ahora, se ha mencionado solamente el análisis del modelo lineal, en el que simplemente se estiman los efectos de k factores a partir de n ensayos preestablecidos. A continuación se hablará sobre la influen-

cia del diseño en evitar cierto tipo de error sistemático y de algunas medidas para comparar la varianza de los estimadores bajo distintos diseños con el mismo número de parámetros.

2.3.1. Diseños por bloque

Los diseños por bloque surgen cuando existe algún factor de ruido (del inglés *nuisance factor*) que es conocido y controlable, pero cuyo efecto en la respuesta no es de interés [Montgomery, 2001, cap. 4], el objetivo del bloqueo simplemente es crear unidades experimentales que sean homogéneas. Por ejemplo: se quiere comparar la durabilidad de dos tipos de zapato, para lo que se asigna aleatoriamente por pie un tipo de calzado, sin embargo cada sujeto tiene un nivel de actividad diferente lo que afectará la durabilidad; el diseño por bloques restringiría la aleatorización para que cada sujeto reciba los dos tipos de calzados (este tipo de bloqueo en particular se conoce como muestra pareada).

Si cada bloque contiene todas las configuraciones una vez (de manera que cada bloque es una repetición), se dice que el diseño es completo, en caso contrario se dice que el diseño es incompleto.

Diseño aleatorizado completo por bloques

La principal consecuencia del diseño de bloques, es que se impone una restricción a la aleatorización, ya que dentro de cada bloque están predefinas la configuraciones a observar, la aleatorización se aplica a un número menor de ensayos. En el caso de diseños completos se tiene que n ensayos se pueden hacer de $n!$ maneras diferentes, mientras que si hay b bloques, la restricción hace que solo haya $b \cdot (n/b)!$ permutaciones posibles, pues cada bloque contiene (n/b) configuraciones, que pueden ser permutadas entre sí y hay b bloques.

Utilizando el modelo de efectos visto antes y suponiendo que solo hay un factor en el modelo con a niveles y b bloques distintos, los diseños

por bloque se pueden modelar mediante:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \tag{2.24}$$

Donde τ es el efecto del factor y β es el efecto del bloque. Con los supuestos usuales en los errores y que los efectos del factor y de los bloques suman cero, la ecuación (2.24) es muy parecida a (2.1), solo se omite el término de la interacción, lo que resulta en que la respuesta solo tiene 2 subíndices (hay ab ensayos en lugar de abn). Así pensando que k solo toma el valor de 1, se puede utilizar la expresión (2.19) para obtener la descomposición de la suma de cuadrados.

$$\begin{aligned} \mathrm{SS}_T &= \sum_{i,j}(y_{ij} - \bar{y}_{..})^2 \\ &= b\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i,j}(\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \\ &= \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_E \end{aligned} \tag{2.25}$$

La diferencia con la expresión sin bloque (2.20) es que lo que antes era la suma de cuadrados por la interacción, ahora es la suma de cuadrados del error, por lo que SS_E tendrá $(a-1)(b-1)$ grados de libertad, en vez de $b(a-1)$ (comparando contra un experimento con un factor con a niveles y b repeticiones). Aunque se pierdan grados de libertad en el denominador de la distribución de referencia, este tipo de diseños es más eficiente para detectar efectos, pues introducir bloques reduce la suma de cuadrados del error en SS_B unidades. Este último punto enfatiza la filosofía del diseño por bloques, identificar una fuente de variabilidad controlable para quitarla de la variabilidad en la respuesta (que últimamente es la fuente de incertidumbre de un experimento).

De esta manera, SS_B es una medida de qué tanto se ganó al definir los bloques (por ejemplo, para decidir si los próximos experimentos deberán ser por bloque), sin embargo no está muy claro que se puedan hacer pruebas estadísticas sobre si captura parte significativa de la varianza: por un lado, se podría realizar una prueba F simplemente como aproximación a las pruebas aleatorizadas; por otro lado solo se tiene un ensayo de cada bloque, por lo que no es posible observar su variabilidad ni hacer cualquier tipo de inferencia basada en SS_B.

Diseños incompletos balanceados de bloque

En el segundo caso, el de diseños incompletos, hay más configuraciones posibles (a) que las que permite un bloque (k), un ejemplo de esto podría darse cuando el tamaño del bloque está restringido por un lote de material.

Cuando el tamaño de bloque no permite todas las configuraciones, pero todas son igual de importantes se utilizan los diseños incompletos balanceados, en los que cada par de configuraciones ocurren un mismo número de veces [Box et al., 2005, cap. 4], estos diseños se pueden construir al tomar a combinaciones en k (aCk) bloques y asignar cada combinación a un bloque, aunque frecuentemente son necesarios menos bloques[Montgomery, 2001, cap. 4]. Una implicación importante de un diseño no balanceado es sobre los estimadores de mínimos cuadrados para los efectos, lo que a su vez afecta la partición de la suma de cuadrados. Este estimador (ajustado por el efecto de los bloques) está dado por [Cox and N.Reid, 2000, cap. 4]:

$$\hat{\tau}_j = \frac{k}{\lambda a} Q_j, \qquad Q_j = \sum_{s=1}^{k} y_{js} - \frac{1}{k} \sum_{j=1}^{a} n_{ij} \sum_{s=1}^{k} y_{js} * n_{js} \tag{2.26}$$

Donde $n_{ij} = 1$ si el nivel i está en el bloque j y cero en cualquier otro caso. La constante λ describe el número de veces que cada par de configuraciones ocurre en el mismo bloque, esta constante toma el valor de [Cox and N.Reid, 2000, cap. 4]:

$$\lambda = \frac{n(k-1)}{a-1}$$

Esto se puede deducir de la siguiente manera: por un lado la configuración i aparece con otras $k-1$ configuraciones por bloque y se tienen n replicas del par de configuraciones; por otro lado esta configuración aparece en par con las otras $a-1$ configuraciones λ veces por bloque, de manera que $\lambda(a-1) = n(k-1)$.

Como los estimadores de los efectos fueron ajustados por el efecto de los bloques, como se definió en (2.26), la suma de cuadrados será

también diferente a la habitual (dada en (2.25)). En este caso la suma de cuadrados puede ser descompuesta en:

$$SS_T = SS_{A'} + SS_B + SS_E$$

Donde A' denota el estimador ajustado de los efectos de τ y cada término está dado por [Montgomery, 2001, cap. 4]:

$$\begin{aligned} SS_T &= \sum\sum y_{ij} - \frac{y_{..}^2}{N} \\ SS_{A'} &= \frac{k\sum Q_j^2}{\lambda a} \\ SS_B &= \frac{1}{k}\sum y_{.j} - \frac{y_{..}^2}{N} \\ SS_E &= SS_T - SS_{A'} - SS_B \end{aligned}$$

Con $N-1$, $a-1$, $b-1$ y $N-a-b+1$ grados de libertad para SS_T, $\mathrm{SS}_{A'}$, SS_B y SS_E respectivamente. Ya que se conoce la partición de la suma de cuadrados y los grados de libertad de cada término, se puede realizar el análisis de la varianza de forma normal (recordando que no tiene mucho sentido inferir sobre el efecto de los bloques).

2.3.2. Medidas de optimalidad

Para finalizar este capítulo sobre análisis de un modelos estadísticos, se presentarán tres medidas de optimalidad del diseño de un experimento, que buscarán estimadores con varianza pequeña. En (2.11) se vio que el diseño del experimento influye en la varianza a través de $(X^TX)^{-1}$, por lo que se medirá la optimalidad del diseño mediante funciones de los valores propios $\lambda_1, \lambda_2, \ldots, \lambda_p$ de la matriz $(X^TX)^{-1}$. En [Box and R.Draper, 2007, cap. 14] se mencionan tres medidas de optimalidad de la *teoría de diseño alfabético óptimo.*

- $D = |X^TX| = \prod \lambda_i^{-1}$
- $A = \sum \lambda_i = \sigma^{-2}\sum Var(\hat{\beta}_i)$
- $E = max\{\lambda_i\}$

Los diseños D-óptimos son aquellos que *maximizan* D, lo que a su vez minimiza el volumen de la transformación $(X^T X)^{-1}$, con lo que se minimizan los intervalos conjuntos de confianza de todos los estimadores. Los diseños A-óptimos, son proporcionales al promedio de la varianza de cada estimador, similar a los D-óptimos, minimizar A implica minimizar, en promedio, la varianza de los estimadores, sin importar su covarianza. Por último, los diseños E-óptimos buscarán minimizar E, de manera que se minimice la varianza máxima de los estimadores.

La selección de qué medida de optimalidad usar dependerá de los objetivos del experimento: por ejemplo, si se busca dar una estimación conjunta de los parámetros, los diseños D-óptimos podrán ser preferidos, mientras que si el objetivo es minimizar la varianza esperada de cada estimador, se buscarán diseños A-óptimos; y finalmente, si se quiere acotar la varianza máxima tolerada, se utilizarán diseños E-óptimos.

2.4. Conclusión

En este capítulo se mostró cómo se pueden estimar superficies de respuesta mediante ensayos en puntos específicos que son analizadas con un modelo lineal. También, se pudo mostrar que un acercamiento directo a este problema resulta poco eficiente; se requieren muchos ensayos, los estimadores y ecuaciones normales resultan en expresiones largas o complicadas y, el mayor problema, es que no resulta en mayor claridad sobre el comportamiento de los factores ya que se modela la respuesta en puntos específicos y no se intenta describir como función del factor.

Por último, se mencionaron dos conceptos del diseño de experimentos: el primero, que son los diseños por bloque, ayudan a separar efectos de factores de ruido controlables (i.e. cuyo efecto en la respuesta es potencialmente importante, pero no de interés); y el segundo, que son las medidas de optimalidad, que permitirán evaluar la región de la varianza de los estimadores, la cual determina directamente la confianza que se tendrá en los estimadores y por lo tanto en el modelo.

Capítulo 3

Diseños factoriales 2^k

En el capítulo pasado se expusieron técnicas utilizadas para estimar efectos principales y de interacción, así como para detectar a los factores que no son relevantes en un fenómeno. Sin embargo, utilizar varios niveles aumenta considerablemente el número de experimentos que se requieren. Los diseños factoriales 2^k ofrecen una solución a este problema, pues se reduce considerablemente el número de ensayos necesarios y sigue siendo posible identificar efectos e interacciones activas para hacer más especifica la investigación en iteraciones futuras.

La característica principal de estos diseños es que todos los factores tienen solo 2 niveles, por lo que solamente se podrán estimar comportamientos lineales.

En estos diseños, los factores están codificados de manera que sus niveles sean -1 o 1 y se utilizará como notación la letra minúscula en el subíndice del factor para indicar cuándo toma el valor de 1 y para denotar la suma sobre todas las repeticiones con esa configuración; si el factor está en su nivel bajo simplemente se omitirá su letra[1]. Por ejemplo: $y_a = a$ es la respuesta cuando el factor A toma el valor alto y B el bajo; $\bar{y}_{ab}$ representa la media cuando tanto A como B están en el valor alto.

[1]En el caso en que todos los factores estén en su nivel bajo se denotará por (1).

3.1. Diseños factoriales con 2 factores

Utilizar dos niveles introduce el concepto de *contrastes*[2],definidos como la diferencia en la respuesta entre el nivel alto y bajo del factor, esta definición aplica sobre los efectos y no sobre los estimadores. Al ser la diferencia contra la media, el efecto del factor será el contraste dividido entre 2. Además, la restricción $\sum \tau^A = 0$, implica que $\tau_1^A = -\tau_{-1}^A$, por lo que solo es necesario estimar los efectos cuando el factor esté en su nivel alto y se denotarán como $\tau_A := \tau_1^A$. Utilizando las expresiones del capítulo anterior y pasando a la nueva notación, el estimador del factor A en un modelo con dos factores está dado por:

$$\begin{aligned}\hat{\tau}_A &= (\bar{y}_{1..} - \bar{y}_{...}) = (ab + a)/2n - (ab + a + b + (1))/4n \\ &= (ab + a - b - (1))/4n)\end{aligned}$$

El estimador de B se calcula análogamente: solo se debe intercambiar a con b.

El estimador del efecto de AB está dado por:

$$\begin{aligned}\hat{\tau}_{AB} &= (\bar{y}_{11.} - \bar{y}_{1..} - \bar{y}_{.1.} + \bar{y}_{...}) \\ &= ab/n - (ab + a)/2n - (ab + b)/2n + (ab + a + b + (1))/4n \\ &= (ab - a - b + (1))/4n\end{aligned}$$

Una ventaja de reducir el número de niveles a 2 es que, al estimar el efecto del factor cuando toma el valor de 1 respecto a su efecto cuando toma el valor de cero, i.e. el efecto de la media, se está considerando al factor como variable y en realidad se estima el cambio en la respuesta por un incremento de una unidad en el factor. Es decir, implícitamente se está construyendo un modelo lineal y no sólo una estimación puntual en una configuración especifica de factores. Operacionalmente, este cambio permite poner directamente el valor del factor en la matriz del modelo, como se muestra en la Tabla 3.1. En esta tabla se utiliza que μ siempre está en su "nivel alto" y se denota por I.

[2]En general en [Cox and N.Reid, 2000, cap. 3] se define un contraste como una combinación lineal de los efectos tal que los coeficientes sumen cero

Tabla 3.1: Tabla de signos

Respuesta	I	A	B	AB
$y_{(1)}$	+	-	-	+
y_a	+	+	-	-
y_b	+	-	+	-
y_{ab}	+	+	+	+

Este tipo de tablas, conocidas como tablas de signos, son una forma de evitar poner todos los unos y conservar solo el signo del valor que toma cada factor, por ejemplo: $y_{(1)}$ es la respuesta cuando A y B están en su nivel bajo, es decir toman el valor de -1, y_{ab} es la respuesta cuando A y B están en su nivel alto, es decir con el valor de 1. Los signos en la columna de la interacción son el producto de las columnas de los factores involucrados, y son los valores que se considera qué toma AB, visto como variable, para estimar el efecto de la interacción[3].

Las tabla de signos se crean a partir del **orden estándar**[Box and R.Draper, 2007, cap. 4], que consiste en intercalar los signos de los efectos principales, todos empezarán con $-$, y el primer factor cambiará de signo cada renglón, el segundo factor cambiará de signos cada 2 renglones, el tercer factor cambiará de signos cada 4 renglones, y en general, la columna l cambiara de signos cada 2^{l-1} renglones.

Teniendo las columnas para los efectos principales, las columnas de interacción son el producto, elemento por elemento, de los factores involucrados [Cox and N.Reid, 2000, cap. 5]. El orden estándar es una manera de escribir las configuraciones que serán necesarias en el experimento, *no* es el orden en el que se deban de realizar; pues, como ya se mencionó el orden de los experimentos tiene que ser aleatorio para evitar efectos sistemáticos del experimento.

La utilidad de esta tabla es que permite leer directamente los estimadores para el efecto de cada factor (solo falta dividir entre el número de ensayos involucrados, i.e. $n2^k$), por ejemplo: el estimador del efecto

[3]Recordando la diferencia entre matriz de diseño y matriz del modelo, ésta tabla describe a la matriz del modelo

AB es $(ab - a - b + (1))/n2^k$.

Además las 4 columnas de la tabla son ortogonales, por lo que los estimadores no están correlacionados entre ellos (las restricciones de estimabilidad que se añadieron en el capítulo pasado están implícitas en la definición del contraste). Más adelante, cuando se reduzca el número de ensayos, se tratará el caso en que hay correlación entre los estimadores, lo que implica que no se pueda distinguir el efecto entre factores o interacciones, a esto se le conoce como *confusión*.

3.2. Generalización a k factores

Con el orden estándar resulta sencillo generalizar el diseño a k factores, solo es necesario seguir la regla y multiplicar las columnas para las interacciones. Por ejemplo, la tabla Tabla 3.2 define la matriz del modelo para un experimento de 3 factores.

Tabla 3.2: Tabla de signos para 3 factores

Respuesta	I	A	B	C	AB	AC	BC	ABC
1. $(y_{(1)})$	+	-	-	-	+	+	+	-
2. (y_a)	+	+	-	-	-	-	+	+
3. (y_b)	+	-	+	-	-	+	-	+
4. (y_{ab})	+	+	+	-	+	-	-	-
5. (y_c)	+	-	-	+	+	-	-	+
6. (y_{ac})	+	+	-	+	-	+	-	-
7. (y_{bc})	+	-	+	+	-	-	+	-
8. (y_{abc})	+	+	+	+	+	+	+	+

Si k fuera lo suficientemente grande, construir una tabla de signos ya no sería práctico, por lo que Montgomery propone en [Montgomery, 2001, cap. 6] una expresión para los estimadores:

$$\hat{\tau}_{A_1 A_2 \ldots A_k} = (a_1 \pm 1)(a_2 \pm 1) \ldots (a_k \pm 1)/2^k$$

Donde se utiliza el signo negativo en cada paréntesis si el factor está siendo incluido en el efecto y positivo si no, por ejemplo, para 3 factores

el efecto de A se puede estimar como:

$$\begin{aligned}\hat{\tau}_A &= (a-1)(b+1)(c+1)/2^3 \\ &= (abc+ab+ac-bc-b-c+a-(1))/2^3\end{aligned}$$

Que coincide con el estimador que se obtendría multiplicando los signos de la columna A por la respuesta y sumando, como se muestra en la tabla.

3.3. Diseños factoriales fraccionados

Aunque restringir el número de niveles a solo dos por factor reduce considerablemente el número de ensayos necesarios, cada vez que añade un factor se requerirá el doble de ensayos para poder estimar el efecto todas las interacciones, ya que habrá el doble de interacciones posibles[4] y con un número relativamente pequeño de factores, el experimento se volvería poco práctico. Por esta razón surgen los *diseños factoriales fraccionados*, en los que se realizarán 2^{k-l} ensayos (donde l se conoce como la fracción del experimento).

Al omitir ensayos se introduce el concepto de **confusión**, pues se tendrán más efectos por estimar que ensayos, por lo que los estimadores dejarán de ser linealmente independientes y habrá estimadores que contengan información de otros efectos. Por ejemplo, si se tuvieran 4 factores y solo se pudieran realizar 8 ensayos, se tendría un diseño 2^{4-1} (llamado media fracción de un diseño 2^4); para mantener los estimadores de los efectos principales no correlacionados, se utiliza la matriz del modelo para 3 factores (2^3 ensayos) y se utiliza la columna de alguna interacción cualquiera[5] como los signos que tendrá el cuarto factor, un ejemplo de esta clase de diseños se presenta en la tabla Tabla 3.3.

[4] Es fácil ver que el número de interacciones se duplica: si con k factores hubiera n interacciones posibles, añadir un factor es equivalente a pensar que las n interacciones que habían son las que existen que no involucran al factor adicional, por lo que habrá otras n que lo involucren

[5] Como se menciona un poco más adelante, la elección de la columna de interacción definirá el patrón de confusión del diseño

Tabla 3.3: Matriz de diseño 2^{4-1}

Respuesta	I	A	B	C	AB	AC	BC	ABC D
$y_{(1)}$	+	-	-	-	+	+	+	-
y_{ad}	+	+	-	-	-	-	+	+
y_{bd}	+	-	+	-	-	+	-	+
y_{ab}	+	+	+	-	+	-	-	-
y_{cd}	+	-	-	+	+	-	-	+
y_{ac}	+	+	-	+	-	+	-	-
y_{bc}	+	-	+	+	-	-	+	-
y_{abcd}	+	+	+	+	+	+	+	+

Por razones de legibilidad, se omitieron las interacciones que involucran al factor D, pero con el método de multiplicar columnas, es fácil calcularlas.

Un diseño fraccionado se describe mediante su **relación generadora**[Box et al., 2005, cap. 6], la cual está definida por la columna que se utilizó para definir las configuraciones del factor adicional, en este caso, $ABC = D$. Debido a la multiplicación de columnas[6], esta relación tiene muchas formas equivalentes: por ejemplo, el lector puede verificar directamente con las columnas de la tabla que si se multiplica ambos lados de la relación por B se obtendría $AC = DB$. Es por esto, que se usará como relación generadora aquella que sea igual a la identidad y que tenga el menor número de factores posibles, en este caso es $ABCD = I$.

Las expresiones equivalentes a la relación generadora definen las *relaciones de alias*, que indican los estimadores que están siendo confundidos. En este caso, los efectos principales están siendo confundidos con las interacciones de tercer orden (que bajo el principio de pocos efectos podrían fácilmente ser ignoradas) y los efectos de segundo orden están confundidos entre ellos, como se muestra en la Tabla 3.4.

Este diseño se llama de **resolución IV**, debido a que la relación

[6]Note que las columnas forman un grupo bajo la multiplicación y que cada columna es su propia inversa

Tabla 3.4: Confusión de efectos en un diseño 2^{4-1}

Relación generadora	Efectos principales y de 3^{er} orden	Efectos de 2^{do} orden
$ABCD = I$	$A = BCD$	$AB = CD$
	$B = ACD$	$AC = BD$
	$C = ABD$	$AD = BC$
	$D = ABC$	

generadora tiene 4 letras asociados a la identidad [Box et al., 2005, cap. 6]. Por otro lado, si se hubiera escogido la columna BC para determinar las configuraciones del factor D, el patrón de confusión sería $BC = D \Rightarrow BCD = I$ y el diseño sería de resolución III. Se observa que una característica de los diseños de resolución III es que los efectos principales están confundidos con las interacciones de segundo orden.

La notación que se utiliza para describir al experimento es: los factores y la fracción como superíndice y la resolución como subínidice; por ejemplo la Tabla 3.3 muestra un diseño 2^{4-1}_{IV}. En general, se busca confundir efectos principales con los efectos de orden mayor que, bajo el principio de pocos efectos, se pueden asumir como ausentes; de manera que se preferirá un diseño 2^{4-1}_{IV} a uno 2^{4-1}_{III}.

No todas las fracciones son posibles, por ejemplo, un diseño 2^{4-2} no es posible, pues se buscaría estimar 4 efectos principales y la media con solo $2^2 = 4$ columnas distintas (y por lo tanto también sólo 4 observaciones[7]), lo que implicaría que dos efectos principales están siendo confundidos (e.g. no se puede distinguir B de C y realmente serían 3 factores); el número máximo de factores que se pueden estimar es $n - 1$, donde n es el número de ensayos que se realizan, a este tipo de diseños se les conoce como *diseños saturados* [Box and R.Draper, 2007, cap. 5].

Para definir fracciones más reducidas, hay que notar que un diseño 2^{k-l} tiene l relaciones generadoras, es decir, iguales a la identidad. El producto de dichas relaciones generadoras será también igual a la identidad, por lo que será posible obtener hasta 2^l expresiones iguales a la

[7]Las tablas de signos tienen el mismo número de renglones que de columnas

identidad. Por último, al multiplicar cada una de estas 2^l igualdades por algún efecto determinado, se obtendrá su relación alias.

La resolución es el número de letras de la "palabra" más corta de todas las relaciones generadoras[Box et al., 2005, cap. 6]. Por ejemplo, en un diseño 2^{5-2} en el que los patrones de confusión son $ABC = D$ y $BC = E$, las relaciones generadoras son $I = ABCD = BCE = AB^2C^2DE = ADE$, por lo que este diseño es de resolución III y la relación alias para el factor A es $A \rightarrow A + BCD + ABCE + DE$.

Otro concepto importante en el diseño de experimentos es el de **proyectabilidad**, que se refiere al número de factores que pueden ser estimados con un diseño factorial completo a partir de un diseño fraccional, en general se tiene que la proyectabilidad es uno menos que la resolución $(P = R - 1)$[Box et al., 2005, cap. 6]. Por ejemplo, en el diseño visto 2^{4-1}_{IV}, cualquier subconjunto de 3 factores puede ser estimado como si el experimento hubiera sido el diseño factorial completo.

3.4. Uso del reflejo para evitar confusión

La confusión que se genera al fraccionar un diseño dificulta la conclusión del experimento, por lo que se podría realizar otro experimento fraccionado que aclare patrones de confusión del primer diseño. El reflejo o doblez (del inglés *foldover*) sirve para aclarar patrones de confusión al invertir el signo de los efectos confundidos para combinar después los estimadores de ambos diseños.

Al igual que en la elección del patrón de confusión, hay varios factores sobre los que se puede "hacer el doblez". Por ejemplo, en el diseño 2^{5-2} con el patrón de confusión mencionado arriba $(I = ABCD = BCE = ADE)$, si se quisiera "desconfundir" el factor E mediante el reflejo, se utilizaría el patrón de confusión $I = -ADE = -BCE = ABCD$, con lo que en cada experimento se tendrían los estimadores de la Tabla 3.5 (omitiendo interacciones de orden mayor a 2).

Con lo que el efecto de E puede ser calculado como $(E + E')/2$ y las interacciones con el factor E también pueden ser calculadas combinando

Tabla 3.5: Contrastes para un experimento 2^{5-2} y su reflejo

Experimento Original	Experimento reflejado
$A \to A + DE$	$A' \to A - DE$
$B \to B + CE$	$B' \to B - BE$
$C \to C + BE$	$C' \to C - CE$
$D \to D + AE$	$D' \to D - DE$
$E \to E + BC + AD$	$E' \to E - BC - AD$

la información de ambos experimentos, por ejemplo: $\hat{DE} \to (A - A')/2$.

Note que realizar un doblez duplica el número de ensayos y sería equivalente a realizar un experimento $2^{k-(l-1)}$ desde el principio, con el que se podría obtener una resolución mayor. Un patrón de doblez particular es sobre todos los factores, lo que elimina la confusión entre efectos principales e interacciones de orden impar, pero confundiendo a las interacciones de orden par, pasando de un diseño de resolución III a uno de resolución IV.

3.5. Ejemplo

En esta sección se presentará un ejemplo, con datos de [Box et al., 2005, cap. 6] sobre la respuesta (el porcentaje de reacción) en un reactor químico, en el que se pueden modificar las siguientes 5 variables continuas, cuyo valor en el nivel bajo y alto se muestra entre corchete después de las unidades[8]:

A) Tasa de alimentación(L/min) [10,15]
B) Catalizador (%) [1,2]
C) Tasa de agitación (rpm) [100,120]
D) Temperatura ($°C$) [140,180]
E) Concentración [3,4]

El objetivo del investigador en este ejemplo sería optimizar el reactor y

[8]Los datos para en análisis se encontraron en el software estadístico R y no fue necesario simularlos

maximizar el porcentaje de reacción. Los resultados del experimento y del cálculo de los estimadores de los efectos e interacciones se encuentran en la Tabla 3.6, donde se omitió la columna asociada a la media, denotada anteriormente por I, ya que toda la columna tiene como valor un signo positivo.

En la Tabla 3.6 se incluyeron también los estimadores de la media fracción del experimento correspondiente a un diseño fraccionado 2^{5-1} y que fue definido por la relación $I = ABCDE$, los cuales se compararán después con los estimadores obtenidos con el diseño completo.

Es de notarse, que al comparar los resultados de los estimadores de la Tabla 3.6 con los presentados en el libro, todos excepto el de la media y la interacción $ABCDE$ son la mitad que los efectos presentados ahí. Esto se debe a que en el libro se reportan los estimadores de los contrastes de los efectos, es decir, el cambio en la respuesta de pasar del nivel bajo al alto de cada efecto; a comparación, en esta tabla se reportan los estimadores del cambio en la respuesta por unidad de cambio en cada efecto.

Si se quisiera obtener los resultados reportados por el libro, en lugar de calcular el promedio de la respuesta multiplicado por el signo de la columna de la interacción, se deberán calcular restando al promedio de la respuesta cuando la interacción está en el nivel alto el promedio de la respuesta cuando la interacción está en el nivel bajo ($\bar{y}$(interacción en nivel alto) $-$ $\bar{y}$(interacción en nivel bajo)).

A primera vista, hay muchos estimadores cercanos al cero, por lo que se realiza una gráfica normal para detectar aquellos que probablemente no provengan de una distribución normal de media cero y así identificarlos como efectos presentes en el fenómeno estudiado (ver Figura 3.1).

Gráficamente, se observa que los efectos activos son el B, D, E, BD y DE.

Para analizar el mismo experimento bajo un diseño fraccionado, definido por el patrón de confusión $I = ABCDE$ (resultando en un diseño 2_V^{5-1}) y cuyas relaciones de confusión se pueden observar en la Tabla 3.7.

Tabla 3.6: Resultados del diseño 2^5 y de la media fracción

Matriz de diseño y respuesta						Resultado del análisis		
A	B	C	D	E	y	Efecto	Est.	Est. 2^{5-1}
-1	-1	-1	-1	-1	61	μ	65.5000	62.25
1	-1	-1	-1	-1	53	A	-0.6875	-1
-1	1	-1	-1	-1	63	B	9.7500	10.25
1	1	-1	-1	-1	61	AB	0.6875	0.75
-1	-1	1	-1	-1	53	C	-0.3125	0
1	-1	1	-1	-1	56	AC	-0.3750	0.25
-1	1	1	-1	-1	54	BC	0.4375	0.75
1	1	1	-1	-1	61	ABC	0.7500	
-1	-1	-1	1	-1	69	D	5.3750	6.125
1	-1	-1	1	-1	61	AD	-0.4375	0.375
-1	1	-1	1	-1	94	BD	6.6250	5.375
1	1	-1	1	-1	93	ABD	0.6875	
-1	-1	1	1	-1	66	CD	1.0625	0.125
1	-1	1	1	-1	60	ACD	-0.3750	
-1	1	1	1	-1	95	BCD	0.5625	
1	1	1	1	-1	98	ABCD	0.0000	
-1	-1	-1	-1	1	56	E	-3.1250	-3.125
1	-1	-1	-1	1	63	AE	-0.0625	0.625
-1	1	-1	-1	1	70	BE	1.0000	0.625
1	1	-1	-1	1	65	ABE	-0.9375	
-1	-1	1	-1	1	59	CE	1.125	
1	-1	1	-1	1	55	ACE	-1.2500	
-1	1	1	-1	1	67	BCE	0.0625	
1	1	1	-1	1	65	ABCE	0.7500	
-1	-1	-1	1	1	44	DE	-5.5000	-4.75
1	-1	-1	1	1	45	ADE	0.3125	
-1	1	-1	1	1	78	BDE	-0.1250	
1	1	-1	1	1	77	ABDE	0.3125	
-1	-1	1	1	1	49	CDE	0.0625	
1	-1	1	1	1	42	ACDE	0.5000	
-1	1	1	1	1	81	BCDE	-0.3125	
1	1	1	1	1	82	ABCDE	-0.2500	

Con los 16 ensayos sólo se pueden estimar 16 efectos y se seleccionaron los de menor orden. En la Figura 3.2 se muestran los estimadores de los efectos en el diseño fraccionado en una gráfica normal, es de notarse que la conclusión sobre qué efectos están activos no difiere de la

Figura 3.1: Gráfica normal para los efectos

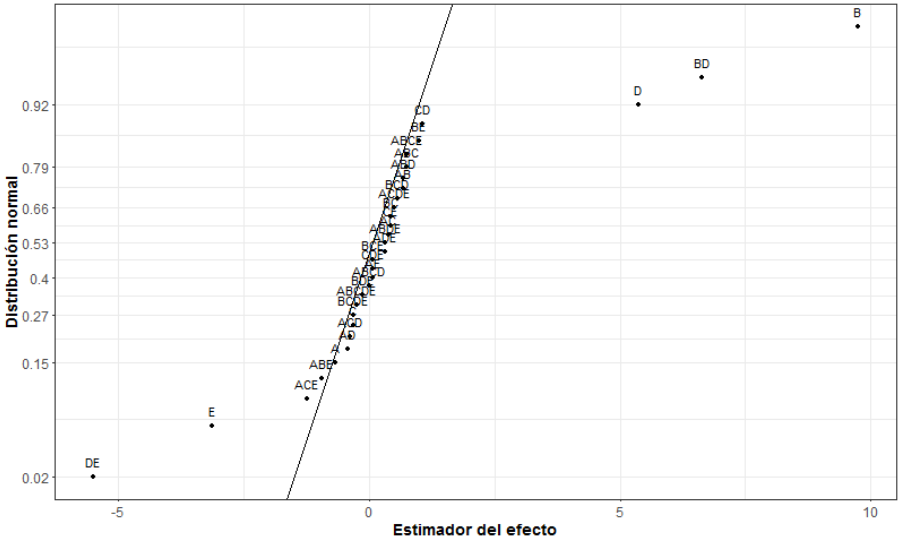


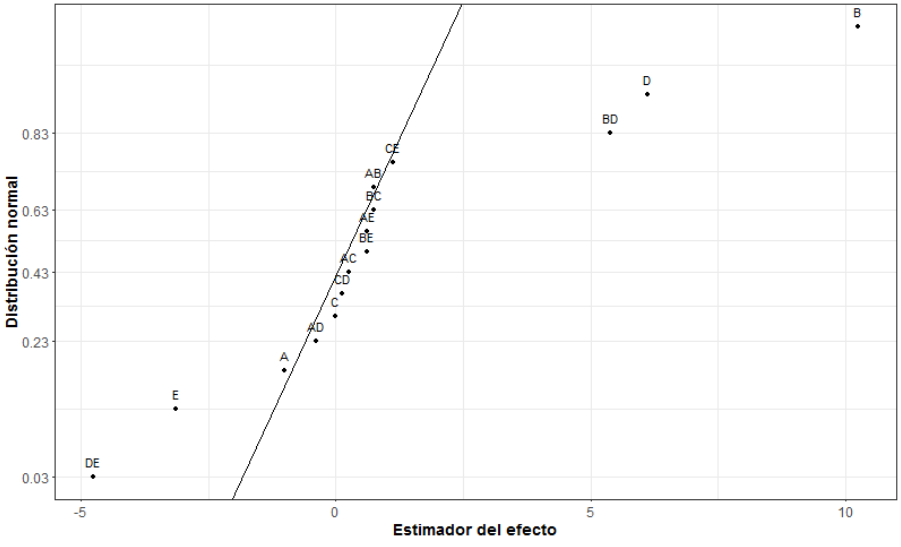
Tabla 3.7: Confusión de efectos en el diseño 2_V^{5-1}

Efecto	Confusión	Efecto	Confusión
mu	ABCDE	AD	BCE
A	BCDE	AE	BCD
B	ACDE	BC	ADE
C	ABDE	BD	ACE
D	ABCE	BE	ACD
E	ABCD	CD	ABE
AB	CDE	CE	ABD
AC	BDE	DE	ABC

obtenida con el diseño completo de 32 ensayos.

Además, al contar con ambos diseños, es posible observar directamente la confusión. Por ejemplo, en el diseño fraccionado, el patrón de confusión de A es $A = BCDE$ y el estimador de A es -1. Si observamos el diseño completo, se observa que el estimador del efecto de A es -0.6875 y el de $BCDE$ es -0.3125, es decir el estimador del efecto de

Figura 3.2: Gráfica normal para los estimadores de la media fracción



A más el estimador del efecto de $BCDE$ en el diseño completo es igual al estimador del efecto de A en el diseño fraccionado, donde el efecto de A es igual al de $BCDE$.

Note además que los 16 ensayos que no se utilizaron corresponden al experimento reflejado sobre E y que este reflejo bastaría para eliminar toda confusión.

En ambos experimentos, se identificarían a los mismos efectos como significativos, lo cuales fueron asociados a tres factores: B,D y E, por lo que se podría pensar que el experimento fue realizado para 3 factores con 4 y 2 replicas para la media fracción y el diseño completo respectivamente. Al contar con repeticiones en la misma configuración de factores activos, se podría estimar la varianza y realizar pruebas F como se describió en el capítulo pasado.

3.6. Diseños de Placket-Burman

Aunque los diseños factoriales fraccionados ayudan a disminuir el número de ensayos necesarios para realizar un experimento sobre k factores y obtener una aproximación razonable de la respuesta; el número posible de factores a investigar está limitado por el número de ensayos que continúa creciendo en potencias de 2. Robin Plackett y Peter Burman publicaron en 1946 [Box and R.Draper, 2007, cap. 5] una clase de diseños en que el número de ensayos crece en múltiplos de 4. Estos diseños son de resolución III y si n, el número de ensayos, es potencia de 2 coinciden con los diseños factoriales fraccionados presentados con anterioridad [Montgomery, 2001, cap. 8].

Para construir la matriz de diseño de Placket Burman (PB_n) solo es necesario conocer el primer renglón y los renglones subsecuentes se construyen al recorrer los signos una posición a la derecha, pasando el signo de la última posición a la primera y poniendo hasta el final un renglón de signos negativos. Existen tres casos para los que no es suficiente conocer el primer renglón de la matriz de diseño y en los que se debe conocer un bloque más grande, estos casos son $n = 28, 40$ y 56

Para $n = 28$, se utilizan tres bloques de signos de 9×9: A, B, C que se permutan de la misma manera que los renglones para obtener una matriz de 27×27 a la que se agrega un último renglón de signos negativos.

Para $n = 40$ y 56 se "dobla" el diseño de $n = 20$ y 28 de la siguiente manera: sea A la matriz de diseño de $n = 20$ y 28, entonces la matriz de $n = 40$ y 56 es

$$\begin{bmatrix} A & A & I \\ A & -A & -I \end{bmatrix}$$

Algunos diseños de este tipo se inlcuyen en la Tabla 3.8 y el resto de los diseños, desde $n = 4$ hasta $n = 100$, omitiendo $n = 92$ pueden ser encontrados al final de [Plackett and Burman, 1946].

Por ejemplo, el diseño PB_{12} se construye a partir de su patrón ge-

Tabla 3.8: Tabla del primer renglón de los diseños Placket-Burman

$n = 8$	+	+	+	-	+	-	-																															
$n = 12$	+	+	-	+	+	+	-	-	-	+	-																										
$n = 16$	+	+	+	+	-	+	-	+	+	-	-	+	-	-	-																						
$n = 20$	+	+	-	-	+	+	+	+	-	+	-	+	-	-	-	-	+	+	-																		
$n = 24$	+	+	+	+	+	-	+	-	+	+	-	-	+	+	-	-	+	-	+	-	-	-	-														
$n = 32$	-	-	-	-	+	-	+	-	+	+	+	-	+	+	-	-	-	+	+	+	+	+	-	-	+	+	-	+	-	-	+						

nerador [9].

$$n = 12: \quad ++-+++---+-$$

Lo que genera el diseño completo mostrado en la Tabla 3.9.

Tabla 3.9: Matriz de diseño PB_{12}

Ensayo	A	B	C	D	E	F	G	H	I	J	K
1	+	+	-	+	+	+	-	-	-	+	-
2	-	+	+	-	+	+	+	-	-	-	+
3	+	-	+	+	-	+	+	+	-	-	-
4	-	+	-	+	+	-	+	+	+	-	-
5	-	-	+	-	+	+	-	+	+	+	-
6	-	-	-	+	-	+	+	-	+	+	+
7	+	-	-	-	+	-	+	+	-	+	+
8	+	+	-	-	-	+	-	+	+	-	+
9	+	+	+	-	-	-	+	-	+	+	-
10	-	+	+	+	-	-	-	+	-	+	+
11	+	-	+	+	+	-	-	-	+	-	+
12	-	-	-	-	-	-	-	-	-	-	-

Si se quisieran analizar menos factores, simplemente no se consideran las columnas correspondientes.

Analisis de diseños Plackett Burman

Los diseños Plackett Burman tienen un patrón de confusión más complicado que los diseños fraccionados vistos anteriormente. En los diseños

[9] Obtenido de [Box and R.Draper, 2007, cap. 5]

Plackett Burman, los efectos principales están parcialmente confundidos (i.e. el valor esperado de los estimadores de los efectos principales será el efecto del factor más, o menos, una fracción del efecto de alguna otra interacción de factores)[Box et al., 2005, cap. 7].

Es fácil ver que los efectos principales no están confundidos con las interacciones que incluyen al factor de la siguiente manera: sea $y = X_p\beta_p + X_s\beta_s$ el modelo propuesto, donde X_p es la matriz de diseño (solo efectos principales) y X_s es la matriz relacionada a los efectos de interacción, entonces:

$$\begin{aligned} \mathrm{E}[\hat{\beta}_p] &= \beta_p + (X_p^T X_p)^{-1} X_p^T X_s \beta_s \\ &= \beta_p + \frac{1}{n} X_p^T X_s \beta_s \end{aligned} \tag{3.1}$$

Donde cada columna de interacción en X_s es la multiplicación, elemento por elemento, de las columnas de los efectos principales que componen la interacción, de manera que la relación de alias para el efecto $\hat{\beta}_i$ con la interacción $\beta_i\beta_k$ es:

$$\begin{aligned} \mathrm{E}[\hat{\beta}_i] &= \beta_i + \frac{1}{n} \sum_k x_{ki} x_{ki} x_{kj} \\ &= \beta_i + \frac{1}{n} \sum x_{kj} = \beta_i \end{aligned}$$

Ya que, por construcción cada columna de la matriz X_p tiene el mismo número de signos positivos que negativos.

Para probar que la confusión es parcial entre los efectos principales y la interacción que no incluye al factor, será suficiente un ejemplo, pues ya se mencionó que estos diseños coinciden con los fraccionados para n potencia de 2, por lo que no siempre será parcial la confusión. Utilizando el diseño PB_{12} y considerando solo los primeros 4 factores (así hay 6 interacciones en vez de $12C2 = 66$), se puede observar en la Tabla 3.10 el patrón de alias, donde el coeficiente de confusión es menor a uno.

Tabla 3.10: Patrón de alias PB_{12} considerando 4 factores

Estimador	Relación de alias
$A \rightarrow$	$A + \frac{1}{3}(-BC - BD + CD)$
$B \rightarrow$	$B + \frac{1}{3}(-AC - AD - CD)$
$C \rightarrow$	$C + \frac{1}{3}(-AB + AD - BD)$
$D \rightarrow$	$D + \frac{1}{3}(-AB + AC - BC)$

3.7. Conclusión

El objetivo de este capítulo es mostrar diseños que redujeran el número de ensayos necesarios para un número creciente de factores considerados; empezando al restringir el número de niveles permitidos por factor (sacrificando la capacidad de estimar curvatura),limitando luego el número de ensayos realizados (a costo de introducir confusión entre los estimadores).

El diseño de experimentos no se agota con estas técnicas. Quedan dos principales problemas por resolver: el primero concierne a poder estimar comportamientos cuadráticos en la respuesta respecto a los efecto sin incrementar de manera exponencial el número de combinaciones entre los 3 niveles; y el segundo consiste en obtener patrones de confusión más adecuados, por ejemplo, donde los efectos principales no tengan confusión y el número de ensayos requeridos tenga un crecimiento lineal.

Estos problemas serán abordados en los siguientes dos capítulos: en el primero se hablará de dos clases de diseños con los que es posible estimar curvatura y en el segundo se hablará de una clase de diseños en la que los efectos principales están sin confusión y que el número de ensayos requeridos crece en orden lineal conforme el número de factores.

Capítulo 4

Estimación de curvatura en la respuesta

En el capítulo anterior se describió una clase de diseños de experimentos que restringe el número de niveles en los factores a 2, con esta simplificación es posible estimar solo efectos lineales y encontrar una dirección en donde continuar el experimento para acercarse a los resultados deseados.

Una vez que se conoce la dirección en que se encuentra la respuesta objetivo, es recomendable conocer también la curvatura, lo que requerirá de un tercer nivel. La manera más obvia de incluirlo es utilizar las ideas del capítulo anterior y crear diseños factoriales 3^k, con el inconveniente que el tamaño de los experimentos crecería rápidamente y los haría poco prácticos.

En este capítulo, se verán dos tipos de diseños que permiten estimar la curvatura a partir de los diseños factoriales 2^k, manteniendo el número de ensayos necesarios relativamente bajo.

A partir de este capitulo se cambiará de notación respecto a la utilizada anteriormente, los factores serán denotados por x y se identificarán por un subíndice de 1 a k, la interacción del factor i con el factor j se denotará por $x_i x_j$ y la media se seguirá denotando por μ. Bajo está

nueva notación los modelo de segundo orden que se buscarán ajustar en este tipo de diseño son, para 2 factores, expresados por:

$$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

Donde x_i puede tomar el valor de $-1, 0$ o 1 y β_i denota el efecto de cada factor o interacción.

Antes de realizar ensayos adicionales para ajustar un modelo cuadrático es importante verificar que dicho esfuerzo sea necesario. Box y Hunter [Box et al., 2005, cap. 11] proponen realizar n_o ensayos centrales (todos los factores en el nivel 0) y calcular

$$q = \bar{y_p} - \bar{y_o} \tag{4.1}$$

Donde y_p representa la respuesta en el perímetro de la región experimental, es decir, $\bar{y_p}$ es el promedio de la respuesta cuando al menos un factor está en sus valores más extremos, por lo que si la respuesta fuera aproximadamente lineal, se esperaría que $\bar{y_p}$ fuera cercano al promedio de la respuesta en el centro $\bar{y_o}$ y por lo tanto que q (4.1) fuera cercana a 0. Para medir que quiere decir cercano del cero, en [Box et al., 2005, cap.11] se sugiere utilizar a la desviación estándar de q, dada por $s(q) = s\sqrt{1/n_p + 1/n_o}$, donde n_p son el número de ensayos del experimento y s es un estimado previo de la desviación estándar de la respuesta.

4.1. Diseños centrales compuestos

Los diseños centrales compuestos se crean a partir de los diseños factoriales 2^k al agregar un punto central y puntos axiales. Los puntos axiales se definirán a una distancia α del centro del experimento. Por ejemplo los puntos axiales sobre el primer factor de un diseño con 3 factores están dados por $(\pm\alpha, 0, 0)$, es decir con los otros dos factores en su nivel 0. Se sugiere basar la selección de α en el número de factores [Box et al., 2005, cap. 11], por ejemplo $\alpha = \sqrt{k}$ coloca a todos los puntos equidistantes al centro del experimento.

En la siguientes figuras (Figura 4.1 y Figura 4.2)se muestran gráficamente los diseños para 2 y 3 factores; los puntos del diseño factorial se encuentran en negro y los puntos axiales (la estrella) de la parte central compuesta en gris. En el lenguaje de los capítulos anteriores, la Figura 4.2 sería equivalente a añadir a un diseño 2^2 cinco ensayos, tres de ellos con el factor A en los niveles $-\sqrt{2}, 0, \sqrt{2}$ con el factor B en el nivel 0 y los otros dos con el factor B en los niveles $-\sqrt{2}, \sqrt{2}$ y con el factor A en 0.

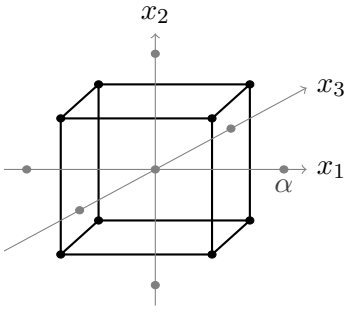


Figura 4.1: Diseño 3 factores

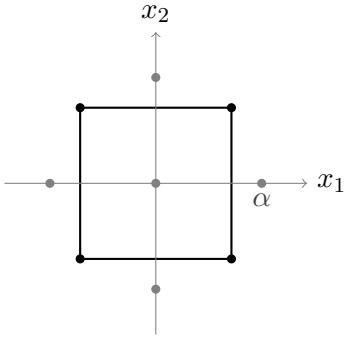


Figura 4.2: Diseño 2 factores

4.2. Diseños de Box-Behnken

Los diseños centrales compuestos de segundo orden normalmente requieren de 5 niveles[1], pues se debe definir α. Además, estos diseños requieren ensayos en los vértices de las configuraciones, lo cual no siempre es siempre posible (por ejemplo, configurar la máquina con presión y temperatura altas).

En [Box and Behnken, 1960] se proponen una nueva clase de diseños que es económica en el número de ensayos, requiere de solo tres niveles y evita configuraciones en los vértices. Estos se forman al combinar diseños factoriales 2^k con diseños incompletos de bloque, los autores proponen

[1]Si se define $\alpha = 1$, se mantendría el número de niveles en 3

las matrices de diseño que deberán se utilizadas en [Box and Behnken, 1960] y se pueden encontrar algunas de ellas en el apéndice B. En la Tabla 4.1 se muestra un ejemplo para 4 factores, basado en 6 bloques de tamaño 2.

Figura 4.3: Matriz conceptual de diseño Box-Behnken para $k = 4$

Diseño incompleto de bloques

Bloque	x_1	x_2	x_3	x_4
1	⋆	⋆		
2			⋆	⋆
3	⋆			⋆
4		⋆	⋆	
5		⋆		⋆
6	⋆		⋆	

Diseño factorial 2^2

x_1	x_2
-1	-1
1	-1
-1	1
1	1

En cada renglón la primer estrella se reemplaza por la primer columna del diseño factorial 2^2 y la segunda por la segunda columna. Por último, se deben agregar n_0 puntos centrales, en este diseño particular $n_0 = 3$, uno por cada par de bloques [Box and Behnken, 1960], resultando en la matriz:

Tabla 4.1: Matriz de diseño Box-Behnken para $k = 4$

bloque 1, 2					bloque 3, 4					bloque 5, 6				
n	x_1	x_2	x_3	x_4	n	x_1	x_2	x_3	x_4	n	x_1	x_2	x_3	x_4
1	-	-	0	0	10	-	0	0	-	19	-	0	-	0
2	+	-	0	0	11	+	0	0	-	20	+	0	-	0
3	-	+	0	0	12	-	0	0	+	21	-	0	+	0
4	+	+	0	0	13	+	0	0	+	22	+	0	+	0
5	0	0	-	-	14	0	-	-	0	23	0	-	0	-
6	0	0	+	-	15	0	+	-	0	24	0	+	0	-
7	0	0	-	+	16	0	-	+	0	25	0	-	0	+
8	0	0	+	+	17	0	+	+	0	26	0	+	0	+
9	0	0	0	0	18	0	0	0	0	27	0	0	0	0

4.3. Conclusión

El diseño de experimentos no se limita solamente a restringir el número de ensayos para se puedan estimar modelos lineales solamente, sino que es una disciplina que sirve para definir las configuraciones de los ensayos requeridos de "forma inteligente", permitiendo incluso evaluar primero si la sospecha de una curvatura importante justifica esfuerzo adicional en un mayor número de ensayos para poder estimarla.

Sin embargo, estas no son las únicas clases de diseños de experimentos que permiten estimar curvatura en la superficie de respuesta. En el siguiente capítulo se expondrá otra clase de diseños, la cual fue propuesta más recientemente y presenta propiedades favorables en el orden de número de ensayos necesarios como función del número de factores a evaluar, así como en su patrón de confusión.

Capítulo 5

Diseño definitivo de experimentos

En este capítulo, objetivo final de este trabajo, se describirá el *diseño definitivo de experimentos*, descrito por B. Jones y C. Nachtsheim en su artículo "*A Class of Three-Level Designs for Definitive Screening in the Presence of Second Order Effects*", publicado en 2011 Jones and Nachtsheim [2011]. Al igual que los diseños presentados antes, éstos están pensados para etapas tempranas del experimento donde se asume el principio de pocos efectos y el jerarquico, además se asume que el objetivo principal es identificar los efectos principales y de segundo orden activos.

5.1. Introducción

Esta clase de diseños surge como una alternativa a los diseños factoriales fraccionados de resolución III que, al confundir los efectos principales con las interacciones de segundo orden, el investigador no puede identificar inequívocamente el efecto principal de un factor y requerirá realizar más ensayos para quitar la confusión de la interacción.

Si se utilizara un diseño de resolución IV, requeriría el doble de ensayos, y aunque los efectos principales están libres de confusión, las interacciones de segundo orden siguen completamente confundidas. Por último, los diseños vistos con anterioridad que cuentan con solo 2 niveles por factor son incapaces de detectar curvatura causada por el efecto cuadrático puro.

Utilizando la notación descrita en el capítulo pasado, el ensayo i de un modelo con k factores será descrito con una ecuación lineal en los parámetros con efectos principales, cuadráticos y de segundo orden de la siguiente manera:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \beta_{jl} x_{ji} x_{li} + \sum_{j=1}^{k} \beta_{jj} x_{ji}^2 + \epsilon_i$$

Los autores del artículo original destacan las siguientes 6 propiedades de esta clase de diseños; las propiedades i) al v) son la razón por que fueron llamados "de cribado definitivo" (del inglés *definitive screening*), ya que con un número pequeño de ensayos es posible identificar efectos principales y cuadráticos activos, y en presencia del principio de pocos efectos, también se detectarán las interacciones activas de segundo orden.

i) El número de ensayos es solamente uno más que el doble del número de factores

ii) A diferencia de los diseños de resolución III, los efectos principales son completamente independientes de interacciones de segundo orden en diseños que contengan un número par de factores

iii) A diferencia de los diseños de resolución IV las interacciones de segundo orden no están completamente confundidas entre ellas, aunque si están correlacionadas

iv) A diferencia de los diseños de resolución III, IV y V con un punto central añadido, todos los efectos cuadráticos son estimables en modelos que no incluyan interacciones

v) Los efectos cuadráticos son ortogonales a los efectos principales y,

aunque correlacionados, no están completamente confundidos con interacciones de segundo orden

vi) Los diseños definitivos son capaces de estimar modelos cuadráticos completos en experimentos con 6 a 12 factores, en los que solo 3 resulten activos. Dichos diseños tendrán además una eficiencia estadística alta[1]

La última ventaja sobre los diseños de dos niveles es que normalmente un investigador espera una relación no lineal entre la respuesta y los factores, por lo que preferirá un modelo que permita no solo detectar curvatura, sino también identificarla con los factores que la puedan estar generando.

5.2. Construcción y propiedades

Los diseños definitivos requieren $2k+1$ ensayos donde k es el número de factores, cada uno restringido a contener exactamente 3 niveles. La matriz de diseño se construye con k pares de filas en la que una fila es el reflejo de la otra (i.e. todos los signos al revés) y es tal que todos los factores tienen el nivel 0 exactamente tres veces, las primeras 2 en un par de filas y la última en el ensayo de punto central. La estructura general se puede observar en la Tabla 5.1, recordando que, para realizar el experimento, el orden de los ensayos deberá ser aleatorizado.

La elección del signo en la Tabla 5.1 sucede solo en las filas impares, pues las filas pares son simplemente su reflejo y se elegirán tal que maximicen $|X^TX|$. Se utilizará la medida de *D-eficiencia* para comparar el impacto de esta restricción sobre las filas pares contra el diseño *D-óptimo* [2] de $2k$ ensayos con un ensayo en punto central añadido.

Definición 5.2.1. En un experimento con k factores, se conoce como

[1] D-eficiencia, véase la definición 5.2.1
[2] Medidas de optimalidad, Cap.2

Tabla 5.1: Estructura general del diseño

Par de filas	Ensayo	Factor			
		x_1	x_2	...	x_k
1	1	0	± 1	...	± 1
	2	0	∓ 1	...	∓ 1
2	3	± 1	0	...	± 1
	4	∓ 1	0	...	∓ 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
k	$2k-1$	± 1	± 1	...	0
	$2k$	∓ 1	∓ 1	...	0
Centro	$2k+1$	0	0	0	0

D − eficiencia de un diseño d_1 respecto a un diseño d_2 a:

$$D_e(d_1, d_2) = \left(\frac{|X(d_1)^T X(d_1)|}{|X(d_2)^T X(d_2)|} \right)^{(1/k)}$$

En el artículo se muestra una tabla con la D-eficiencia calculada para 6 a 12 factores Tabla 5.2 y se observa que la restricción es tolerable sobre la eficiencia del diseño, pues en todos los casos se encuentra entre el 84 % y el 90 %.

Tabla 5.2: D-efficiencia de los diseños defintivos

Número de factores	D-efficiencia (%)
6	85.5
7	84.1
8	88.8
9	86.8
10	90.9
11	89.1
12	89.8

El algoritmo para la construcción de la matriz de diseño es relativamente sencillo: la idea es inicializar la matriz en valores aleatorios entre -1 y 1 e iterar sobre las posiciones que no están fijas, observando el cambio en el determinante de X^TX si la posición (y su reflejo) tuviera el valor de -1 o de 1, repitiendo este proceso hasta que no haya mejoría o se supere un número máximo de iteraciones. El pseudocódigo 1 muestra con más detalle este proceso.

Se sugieren dos estrategias en la implementación del algoritmo para evitar máximos locales; la primera es que se deberá correr varias veces el algoritmo con distintos puntos iniciales; y la segunda es que la matriz inicial se deberá definir con valores en el intervalo continuo $(-1, 1)$ en vez de con números del conjunto $\{-1, 1\}$ [Jones and Nachtsheim, 2011].

En el artículo original se incluyen los diseños construidos con este método para $k = 4$ hasta $k = 12$ factores, mismos que se incluyen en el apéndice C.4 y se resalta que las columnas de todos los diseños con un número par de factores son ortogonales [3], excepto para $k = 12$; y aunque no se incluyen los diseños en el artículo, los autores originales calcularon y publicaron hasta $k = 30$ y solo se encontró a $k = 12$ como diseño no ortogonal.

5.3. Confusión del diseño

En esta clase de diseños de experimentos, se tiene un número menor de ensayos que de posibles efectos a estimar ($2k + 1$ contra $k + k(k + 1)/2$), por lo que la matriz X^TX, cuando se incluyan a todos los efectos principales, cuadráticos y de interacción no será invertible. Esto implica que no se puede estimar el sesgo de los estimadores de los efectos hasta definir un modelo estimable. Sin embargo, los estimadores de los efectos principales son independientes de los efectos de segundo orden.

Por esta razón, la confusión de los efectos principales se seguirá mi-

[3] Si el número de factores fuera impar, es imposible tener dos columnas ortogonales, pues el producto punto de cada par de renglones, de los que hay uno por factor, es 2 o -2, y la correlación entre dos columnas será al menos $2/2k$

Algoritmo 1 Construcción de la matriz X de diseño definitivos

$X_{ij} \leftarrow rand(-1, 1)$ ▷ $i \in \{1, 3, \ldots, 2k-1\}$
$X_{ij} \leftarrow 0$ ▷ $j = (i+1)/2, i \in \{1, 3, \ldots, 2k-1\}$
$X_{ij} \leftarrow -X_{ij}$ ▷ $i \in \{2, 4, \ldots, 2k\}$
$X_{i,2k+1} \leftarrow 0$
$T \leftarrow X$
$cond \leftarrow$ TRUE
$bestDet \leftarrow det(X^T X)$
while cond && $k < maxiter$ **do**
  $cond \leftarrow$ FALSE
  for $i \in \{1, 3, \ldots, 2k-1\}$,$j \in \{1, \ldots, p\}$,$j \neq (i+1)/2$ **do**
    $T_{ij} \leftarrow -1, \quad T_{i+1,j} \leftarrow 1$
    $det1 \leftarrow det(T^T T)$
    $T_{ij} \leftarrow 1, \quad T_{i+1,j} \leftarrow -1$
    $det2 \leftarrow det(T^T T)$
    if $det1 > bestDet \quad || \quad det2 > bestDet$ **then**
      $cond \leftarrow$ TRUE
      if $det1 > det2$ **then**
        $X_{ij} \leftarrow -1$
        $X_{i+1,j} \leftarrow 1$
        $det \leftarrow det1$
      else
        $X_{ij} \leftarrow 1$
        $X_{i+1,j} \leftarrow -1$
        $det \leftarrow det2$
      end if
      $T \leftarrow X, \quad bestDet \leftarrow det$
    end if
  end for
  $k \leftarrow k + 1$
end while

diendo en relación a la matriz $X^T X$ y la confusión de los efectos de segundo orden se medirá utilizando la formula de la correlación empírica, definida como el producto vectorial de desviaciones respecto a la

media de las columnas analizadas. Por diseño, la media de las columnas de los efectos principales es cero por lo que en estos casos el producto vectorial de las desviaciones respecto a la media está también definido por X^TX. Aunque las columnas de los efectos a analizar no sean variables aleatorias, se acostumbra llamar a esa medida de confusión correlación, lo cual se hará también en este trabajo.

La propiedad mencionada de independencia entre los efectos principales del modelo respecto a los efectos de segundo orden se observa en las ecuaciones normales, donde se desacoplan los estimadores de ambas partes:

$$y = X_1\beta_1 + X_2\beta_2 = \begin{bmatrix} X_1 \mid X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} X_1^T y \\ X_2^T y \end{bmatrix} = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Donde β_1 es el vector de los k efectos principales, β_2 es el vector con los k efectos cuadráticos y las $(k-1)k/2$ interacciones de segundo orden.

Para probar que las ecuaciones normales se desacoplan, basta probar que $X_1^TX_2 = 0$, lo que se hace a detalle en el apéndice C.1; la idea de la demostración es que en las columnas de interacciones, cada par de renglones toma el valor de -1 o de 1 y en la columna de efectos principales cada par de renglones tiene signos contrarios, por lo que la suma dentro del producto punto de cada par de renglones es cero.

Esto implica que las ecuaciones normales del modelo se pueden separar en las ecuaciones normales de los efectos principales y en las ecuaciones de los efectos de segundo orden:

$$X_1^T y = X_1^T X_1 \beta_1$$
$$X_2^T y = X_2^T X_2 \beta_2$$

Con la advertencia que $X_2^TX_2$ no es invertible, pues no es de rango completo ya que tiene $2k$ renglones "independientes" y $k(k+1)/2$ columnas. Sin embargo, $X_1^TX_1$ si es invertible, por lo que los estimadores de los efectos principales se pueden calcular independientemente de los

términos de segundo orden, esta propiedad se puede observar en la matriz alias (5.1) y en el mosaico de correlaciones en la Figura 5.2 y la Figura 5.3

$$\begin{aligned} E[\hat{\beta}_1] &= E[(X_1^T X_1)^{-1} X_1^T y] = E[(X_1^T X_1)^{-1} X_1^T (X_1\beta_1 + X_2\beta_2 + \epsilon)] \\ &= \beta_1 + A\beta_2 = \beta_1 + 0 \quad \text{, donde } A = (X_1^T X_1)^{-1} X_1^T X_2 \end{aligned} \tag{5.1}$$

La expresión (5.1) prueba que el estimador de los efectos principales es insesgado e independiente de los efectos de interacción y cuadráticos.

La confusión entre efectos de segundo orden se analizará mediante la correlación empírica, cuya expresión se muestra en la expresión (5.2). Se analizará por separado los siguientes 5 tipos de correlación empírica entre los efectos:

i) Entre dos efectos cuadráticos (qq,ss)
ii) Entre un efecto cuadrático y una interacción que incluya al factor del término cuadrático (qq,qs)
iii) Entre un efecto cuadrático y una interacción que no incluya al factor (qq,st)
iv) Entre dos interacciones que tengan algún factor en común entre ellas (st,sv)
v) Entre interacciones que no tengan ningún factor en común (st,uv)

Se detonará por $r^c_{st,uv}$ a la correlación entre la columna que involucra a la interacción ST y a la UV, donde S, T, U y V son los factores involucrados, por ejemplo. la interacción entre el efecto cuadrático del factor Q y el efecto de interacción del factor Q y S se denotará por $r_{qq,qs}$.

El cálculo de la correlación empírica se hace respecto a desviaciones de la media de cada columna, mediante la expresión:

$$\begin{aligned} r^c_{st,uv} &= \frac{\sum (X_{i,st} - \bar{X}_{st})(X_{i,uv} - \bar{X}_{uv})}{SS_{st} SS_{uv}} \\ SS^2_{st} &= \sum (X_{i,st} - \bar{X}_{st})^2 \end{aligned} \tag{5.2}$$

Entre dos efectos cuadráticos, se puede obtener una expresión general para la correlación empírica dada por (5.3). Esta se deduce al notar que las columnas contienen $2k + 1 - 2 - 1$ unos y 3 ceros y que el producto

punto de dos columnas de efectos cuadráticos es igual a $2k+1-2-2-1$ (detalles en el apéndice C.2).

$$r^c_{qq,ss}(k) = \frac{1}{3} - \frac{1}{k-1} \tag{5.3}$$

En el caso de un efecto cuadrático y una interacción, con y sin factor común, se va a suponer en el cálculo de una expresión general que el número de factores es par y que la columna de la interacción suma cero. Los autores del artículo orígnal mencionan que este último supuesto se ha observado en estos diseños cuando los efectos principales son ortogonales[4] [Jones and Nachtsheim, 2011], sin embargo no se garantiza que se cumpla siempre. Al igual que en la expresión (5.3), los detalles se desarrollan en el apéndice C.2 y la expresión de ambos tipo de correlaciones se muestra en la expresión (5.4) y (5.5).

$$r^c_{qq,qs} = 0 \tag{5.4}$$

$$r^c_{qq,st} = \pm\sqrt{\frac{2k+1}{3(k-1)(k-2)}} \tag{5.5}$$

El cálculo de una expresión cerrada para la correlación entre dos interacciones tiene la dificultad que en el producto del numerador, el primer término es $\sum X_{i,st}X_{uv}$ y no es posible utilizar alguna propiedad general como en los casos anteriores. Por esta razón $r^c_{st,uv}$ deberá ser calculado para cada diseño en especifico.

Con base en estas expresiones es posible obtener dos conclusiones: la primera es que la correlación entre efectos cuadráticos empieza en 0 y se acerca a $1/3$ a medida que k crece; y que la correlación entre el efecto cuadrático y una interacción que no incluya al factor empezará en $\sqrt{2}/2 \approx 0.71$ ($k = 4$), pero descenderá a medida que k crezca. Ambas conclusiones pueden apreciarse en la Figura 5.1.

Además de las gráficas en que se ve el valor de la correlación empírica por número de factores, se incluyeron dos mosaicos (Figura 5.2 y

[4]En este trabajo se encontró que esto no era cierto en todos los diseños, por ejemplo en un diseño con 16 factores, la suma de la columna de la interacción AH es 4

Figura 5.1: Valor absoluto de la correlación con efectos cuadráticos

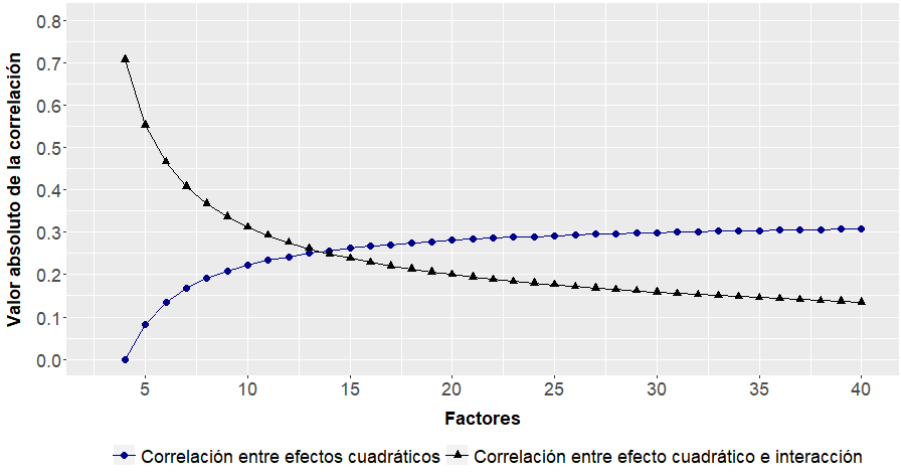


Figura 5.3) que muestran el valor absoluto de correlaciones entre todos los efectos para $k = 6$ y $k = 7$. Estos mosaicos fueron calculados con la formula de la correlación empírica (5.2), que como ya se mencionó es aplicable también a los efectos principales, pues su media es cero.

En los mosaicos se pueden apreciar las diferencias cualitativas que hay en las correlaciones de ambos diseños, uno con número par y el otro con número impar de factores. La parte inferior de cada mosaico describe el mapa de calor con la escala utilizada para el valor absoluto de las correlaciones.

La diferencia más evidente entre ambos diseños es que los efectos principales no están libres de confusión cuando el número de factores es impar. Además se observa que la correlación entre efectos cuadráticos y una interacción que incluya al factor no es cero, y que un mismo tipo de correlación tiene diferentes valores, razón por la cual no fue posible encontrar una expresión general. Se alcanza a notar que la correlación más alta sucede entre interacciones que no comparten factor común.

A raíz de la diferencia entre la estructura de correlación entre 6 y 7 factores, se decidió que sería útil observar en una gráfica el promedio

Figura 5.2: Correlación de efectos para el modelo con 6 factores

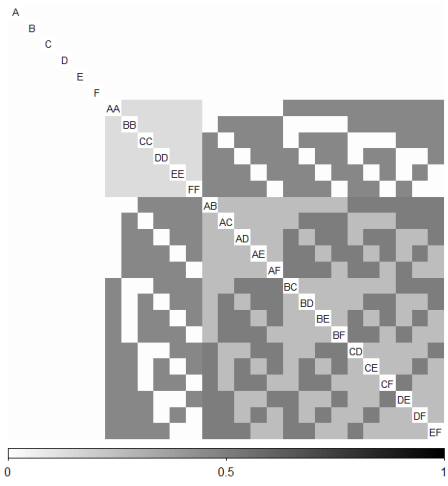
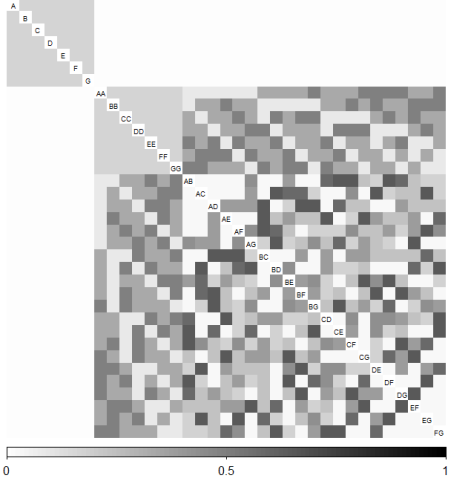
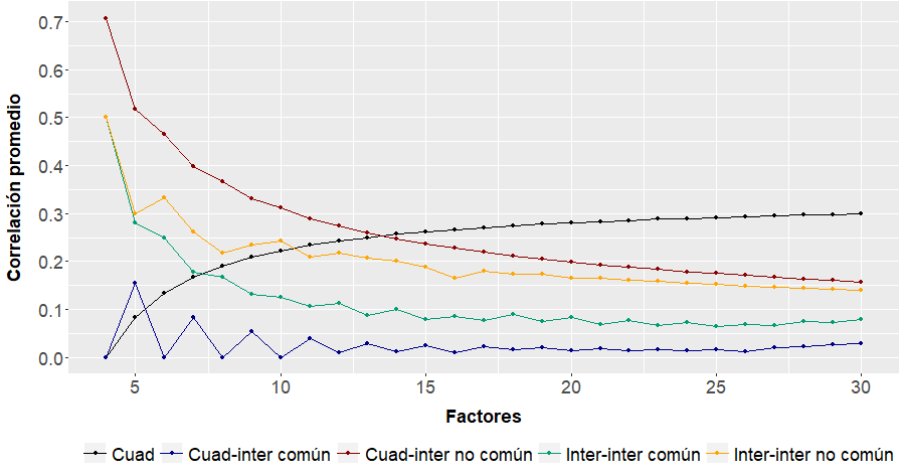


Figura 5.3: Correlación de efectos para el modelo con 7 factores



del valor absoluto (Figura 5.4) y en otra gráfica el máximo del valor absoluto (Figura 5.5) de los 5 tipos de correlaciones para los factores $k = 4$ a $k = 30$, con los diseños calculados en el artículo original.

Figura 5.4: Promedio del valor absoluto de la correlación por tipo de efecto

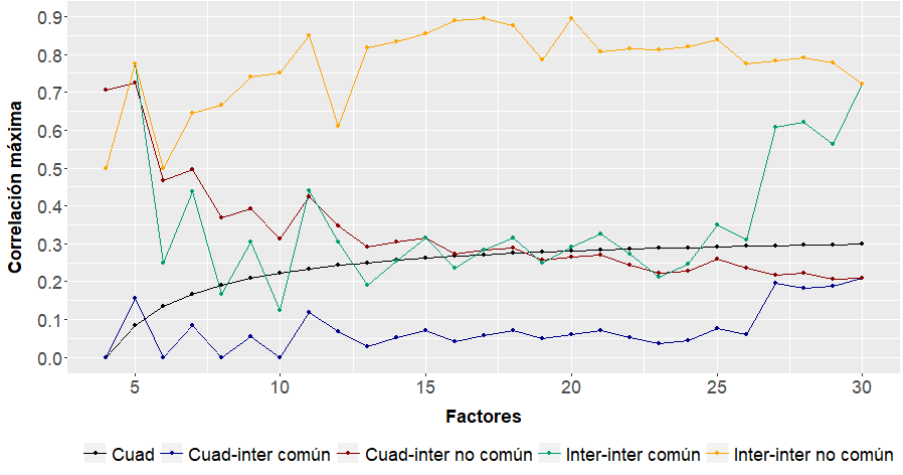


En la Figura 5.4, aquella con los promedios del valor absoluto de las correlaciones, se observa que la correlación calculada coincide con la teórica excepto entre un efecto cuadrático y una interacción con factor común, la diferencia se debe a que las sumas de las columnas de las interacciones de los diseños utilizados no suman cero, lo cual era un supuesto en la expresión analítica (5.4).

Se observa además que la correlación empírica entre efectos disminuye, lo cual se debe principalmente a que el número de efectos aumenta en orden cuadrático a medida que aumenta el número de factores, reduciendo la participación en el promedio de los valores extremos de correlación

En la Figura 5.5, aquella con los valores máximos de las correlaciones empíricas, se observa que la correlación en diseños con número par de factores se comporta diferente que en los diseños con número impar. Además, con esta gráfica se puede confirmar que las correlaciones promedio disminuyen por el número de efectos, ya que las correlaciones máximas se mantienen en una banda relativamente constante.

Figura 5.5: Máximo del valor absoluto de la correlación por tipo de efecto



5.4. Proyectabilidad

La proyectabilidad en el diseño de experimentos, donde el objetivo es seleccionar factores activos, es un concepto importante que permite, una vez seleccionados 2 o 3 factores relevantes[5], utilizar los ensayos observados con variaciones en los factores no seleccionados como réplicas de los factores activos.

Al proyectar este tipo de diseño a dos factores, se observa que resulta en un diseño con un punto central, cuatro puntos en las orillas (un factor en el centro y el otro en ± 1) y $2k - 4$ puntos en las esquinas (ambos factores en ± 1). Esta proyección sobre dos factores será balanceada solo si k es par, lo cual se prueba en el apéndice C.3, además, si las dos columnas suman cero y son ortogonales, la conversa también es cierta.

En la proyección a tres factores se obtiene un diseño con un punto central y sólo es posible proyectar a 6 puntos en las orillas del cubo, por

[5]Más de tres factores podrían resultar relevantes, sin embargo, solo se mostrarán las propiedades proyectivas a dos y tres factores

lo que no se proyecta completamente a un diseño factorial 3^3 (no están las 12 orillas), los otros $2k - 6$ puntos se encontrarán en los vértices.

Este tipo de diseños no proyectan a diseños factoriales, por lo que se midió la pérdida de información mediante la D-eficiencia (comparada contra diseños D-óptimos de tres factores y $2k$ ensayos más un punto central) y se encontró que la D-eficiencia promedio (note que hay $kC3$ maneras distintas de proyectar el diseño) para $k = 6$ hasta $k = 12$ factores resultó siempre por arriba de 90 % e incluso llegó a 97 % para 8 factores [Jones and Nachtsheim, 2011].

Por último, como ya se mencionó en el punto (vi) de las propiedades de está clase de diseños, a partir de 6 factores (13 ensayos) se pueden estimar todos los parámetros de un modelo cuadrático completo donde hay hasta 3 factores activos (1 media + 3 principales + 3 cuadráticos + 3 interacciones = 10 parámetros).

Las propiedades proyectivas descritas en esta sección recalcan que además de ser posible estimar el modelo completo, el diseño original es bastante eficiente al proyectarse respecto un diseño factorial con un menor número de factores y mayor número de réplicas en cada configuración.

5.5. Conclusión

En conclusión, la clase de diseños definitivos de experimentos propuesta por Jones y Nachtsheim Jones and Nachtsheim [2011] puede ser de gran utilidad para identificar factores activos en un fenómeno desconocido o del que se tiene poca información, pues analíticamente tiene buenas propiedades, tales como crecimiento lineal en el orden de ensayos necesarios por número de factores a probar y baja correlación entre efectos, particularmente dejando los principales sin correlación con cualquier otro efecto[6].

[6]Aunque se quieran probar un número impar de factores con "peor" patrón de correlación, el crecimiento pequeño en el número de ensayos hace que simplemente se pueda utilizar el diseño con un factor más

Capítulo 6

Cribado de factores

En este capítulo se describirán cuatro métodos de cribado de factores, y en general, de efectos. Se supondrá que el experimento ya fue realizado con alguno de los diseños descritos en los capítulos anteriores y que el investigador conoce la respuesta obtenida, asi como las configuraciones y por tanto el número máximo de factores activos. Sin embargo, el investigador deseará estimar un modelo que aproxime a la respuesta, es decir deberá buscar en todas las combinaciones de números de factores activos e interacciones entre ellos aquel que mejor aproxime la respuesta sin tratarse de un problema de sobreajuste.

El diseño de experimentos y el conocimiento que tenga el investigador sobre el patrón de confusión, le ayudará a reducir el número de modelos a evaluar; por ejemplo, si los efectos principales no tienen confusión con los efectos de interacción y cuadráticos, podrá en un primer paso estimar cuales son los factores activos y una vez reducido este número calcular todos los modelos potenciales que incluyan a los efectos de interacción y cuadráticos. Esto presenta una gran ventaja computacional, pues ignorando de momento el principio de herencia, el número de modelos potenciales crece exponencialmente respecto al número de efectos: si hay n efectos, entonces hay 2^n modelos potenciales[1].

[1]Cada efecto tiene dos opciones: estar activo o no

Se presentarán dos métodos frecuentistas y dos bayesianos. Dentro de los métodos frecuentistas, el primero, propuesto por el presente autor, está inspirado en la selección de hiperparámetros usada comúnmente en aprendizaje de máquina y el segundo utiliza el *criterio de información de Akaike corregido* en que se castiga el uso de parámetros adicionales. De los métodos bayesianos descritos a continuación, el primero evalúa qué factores están activos al sumar la probabilidad de todos los modelos que los contengan *dados* los datos y el segundo construye modelos de manera secuencial basados en los factores con alta probabilidad de estar activos en la iteración pasada.

6.1. Métodos frecuentistas

6.1.1. Evaluación de hiperparámetros

En este método se buscará primero encontrar cual es el número de efectos o parámetros a incluir en el modelo. Se esperan dos fénomenos al aumentar el número de efectos activos en el modelo: el primero, que las medidas de ajuste a los datos aumentarán, es decir el error cuadrático medio del mejor modelo será menor mientras más efectos se agreguen; y el segundo es que el mejor modelo puede ser sobre especifico, lo que se observará en que el error cuadrático medio tendrá una mayor variación entre los modelos con el mismo número de parámetros.

Se sugiere el siguiente método para seleccionar a los efectos activos y evitar modelos con problemas de sobrajuste:

i) Calcular todos los modelos con un número fijo de efectos y estimar la varianza del error cuadrático medio y su valor mínimo

ii) Repetir el paso anterior para todos los números posibles de efectos considerados en el modelo

iii) Graficar el error cuadrático medio mínimo y su varianza contra el número de efectos incluidos

iv) Identificar cualitativamente, el número de efectos a considerar tal

que si se agregara uno adicional, la reducción en el error cuadrático medio sea poco importante y haya un incremento en la varianza

v) Finalmente, seleccionar el modelo con menor error cuadrático medio en que se consideren el número de efectos seleccionado en el paso anterior

Con este método el investigador seleccionará el mejor modelo con cierto número fijo de efectos y se cuidará que esta selección del número de efectos no resulte en modelos de sobre ajustados al buscar que los modelos considerados tengan un error cuadrático medio similar. La principal desventaja es que esto se hará de manera subjetiva sin demostración de optimalidad.

6.1.2. Criterio de información de Akaike corregido

El criterio de información de Akaike fue inicialmente propuesto por Hirotugu Akaike en 1973 como un método para discriminar modelos estadísticos con un número distinto de parámetros. Dicho criterio está basado en estimar el valor esperado de la pérdida de información resultante de utilizar un modelo para representar un fenómeno real; se calcula con la siguiente expresión Murvich and Tsai:

$$AIC = n * (ln(2\pi\hat{\sigma^2}) + 1) + 2(p + 1) \tag{6.1}$$

Donde n es el número de observaciones, $\hat{\sigma^2}$ es el estimador de la varianza, el cuál si X es la matriz del diseño está relacionado a $(X^T X^{-1})$, y p es el número de parámetros, el cuál es igual al número de columnas en la matriz del diseño.

En Murvich and Tsai se menciona que el estimador (6.1) es sesgado y se propone el *criterio de información de Akaike corregido*, el cuál se define con la siguiente expresión:

$$AIC_c = n * ln(2\pi\hat{\sigma^2}) + \frac{1 + p/n}{1 - (p + 2)/n} \tag{6.2}$$

En las expresiones (6.1) y (6.2) se observan dos elementos: el primero esta relacionado con la función de verosimilitud y el segundo con el número

de parámetros utilizados, de manera que, el criterio evalúa el ajuste del modelo a los datos, pero penaliza aquellos modelos que incluyen muchos parámetros y tienen un mayor riesgo de estar sobre ajustandose a las observaciones.

En ambos casos se seleccionará el modelo con la menor pérdida de información, es decir aquel modelo cuyo valor dado por (6.2) sea mínimo.

6.2. Métodos bayesianos

6.2.1. Cribado bayesiano de factores

En [Box and Meyer, 1993] se propone el cribado bayesiano de factores como una herramienta para situaciones en que la estructura de confusión es difícil de entender. En esta metodología se evalúa si un factor es activo a través de la probabilidad de que un modelo en el que se encuentre activo sea "correcto" dados los datos. Es decir, se define la probabilidad de que el factor j sea activo como:

$$P_j = \sum_{M_i \text{ tal que } x_j \in M_i} P(M_i|y)$$

Donde M_i denota un modelo definido por los efectos considerados, por ejemplo $y = \mu + \epsilon$ es un modelo sin efectos; $y = \mu + \beta_1 x_1 + \beta_{1,1} x_1^2$ es un modelo que considera el efecto principal y cuadrático del factor x_1.

Conceptualmente en esta metodología se proponen todos los modelos de interés[2]y a cada modelo se le asigna una creencia, mediante una probabilidad *a priori*, o previa, de que el modelo sea correcto, así como se le asigna a sus parámetros una distribución previa. Finalmente se calcula la probabilidad de que el modelo sea correcto *dados* los ensayos[3]. Se nota que el objetivo de este método es identificar los factores que

[2]O que se crean posibles, por ejemplo, excluyendo los que no cumplan el principio de jerárquico o de herencia, por considerarse poco probables en la vida real

[3]A este problema se le conoce como la *probabilidad inversa*, pues el objetivo es determinar la probabilidad de causas dados los efectos

podrían estar activos en el modelo, por lo que lo importante no es la magnitud de las probabilidades en su distribución final, o *a posteriori*, sino su relación entre los efectos, lo que permitirá detectar como activos a factores cuya probabilidad calculada sea menor a 0.5, pero que resalte de la de los otros factores.

En el artículo revisado para este trabajo [Box and Meyer, 1993], se propone una distribución mínimo informativa[4] para la media general y la varianza tal que $f(\beta_0, \sigma) \propto 1/\sigma$.

Los parámetros sobre los efectos de los factores se asumirá que siguen una distribución normal $N(0, \gamma^2\sigma^2)$, este supuesto se justifica con las siguientes razones: i) la media cero indica incertidumbre sobre la dirección del efecto; ii) la varianza en diferente escala permite distinguir el efecto de los parámetros del ruido aleatorio y se escoge la misma escala para todos los efectos pues tienen por lo general el mismo orden de magnitud; y iii) la distribución propuesta es la conjugada de la verosimilitud, lo que facilita cálculos y actualizaciones. De esta manera, el método consiste en:

Calcular para cada modelo M_i:

$$p(M_i|y) \propto p(y|M_i)p(M_i)$$

Donde la constante de proporcionalidad es simplemente $p(y)$ y computacionalmente, se calcula forzando que la suma de la probabilidad sobre todos los modelos sea 1. Para calcular $p(M_i)$ se asume que la probabilidad previa de que un factor sea activo es igual para todos los factores y que son independientes. Así, sea $\pi = p(x_i)$ la probabilidad de que el factor x_i sea activo, entonces

$$p(M_i) = \pi^{f_i}(1-\pi)^{k-f_i}$$

Donde k es el número total de factores y f_i el número de factores que se suponen activos en el modelo M_i.

[4]Una distribución mínimo informativa previa es tal que sus parámetros no aparecen en la probabilidad *a posteriori*

Para calcular $p(y|M_i)$ utilizará que la distribución de la respuesta es conocida una vez fijados los valores de los parámetros, por lo que se puede condicionar por un valor fijo y aplicar el teorema de la probabilidad total:

$$p(y|M_i) = \int_0^\infty \int_{\mathcal{R}^{t_i+1}} f(y|M_i, \beta, \sigma) f(\beta, \sigma|M_i) d\beta d\sigma$$

Donde β es el vector de f_i+1 parámetros y ya se mencionó que $f(\beta_0, \sigma|M_i)$ está dada por una función mínimo informativa y los demás efectos se asumen independientes y que siguen una distribución normal.

Es fácil darse cuenta que lo más complicado del método es calcular esta última expresión, sin embargo, en Box and Meyer [1993] se muestra el resultado:

$$f(y|M_i) \propto \gamma^{-t_i} |\Gamma_i + X_i^T X_i|^{-1/2} (S(\hat{\beta}_i) + \hat{\beta}_i{}^T \Gamma_i \hat{\beta}_i)^{-(n-1)/2}$$

Donde

$$\Gamma_i = \frac{1}{\gamma} \begin{bmatrix} 0 & 0 \\ 0 & I_i \end{bmatrix}$$

$$\hat{\beta}_i = (\Gamma_i + X_i^T X_i)^{-1} X_i^T y$$

$$S(\hat{\beta}_i) = (y - X\hat{\beta}_i)^T (y - X\hat{\beta}_i)$$

Y dividiendo cada ensayo por la probabilidad del modelo con 0 factores[5], se obtiene que la probabilidad final del modelo M_i dado el vector de datos observados y es:

$$f(M_i|y) \propto \left(\frac{\pi}{1-\pi}\right)^{f_i} \gamma^{-t_i} \frac{|X_0^T X_0|^{1/2}}{|\Gamma_i + X_i^T X_i|^{1/2}} \left(\frac{S(\hat{\beta}_i) + \hat{\beta}_i{}^T \Gamma_i \hat{\beta}_i)}{S(\hat{\beta}_0)}\right)^{-(n-1)/2} \tag{6.3}$$

Por último se calculará la probabilidad de que un factor x_i esté activo como la suma de las probabilidades dados los datos observados de todos los modelos M_j que lo contengan y se compararán las probabilidades de todos los factores que se evaluaron. Se seleccionarán como activos a aquellos factores cuya probabilidad sea notablemente mayor que los demás y no respecto a un umbral mínimo de aceptación.

[5] Al final todas las probabilidades se escalan para que la suma sea igual a 1 y en el artículo se menciona que se evitan problemas de punto flotante

6.2.2. Análisis secuencial bayesiano

Por su parte, Victor Aguirre, propone en Aguirre [2016] un método bayesiano secuencial de cribado de factores. Su modelo propuesto evita la compejidad de tener que enlistar todos modelos posibles y calcular la probabilidad final de cada modelo dadas las observaciones y en su lugar se propone calcular la probabilidad de que los parámetros sean distintos de cero en modelos cuya complejidad crece de manera secuencial al incorporar los efectos que se detectan como distintos de cero en cada paso. Se definirá que un efecto es distinto de cero cuando su momio[6] de ser positivo o su momio de ser negativo (que es el recíproco del momio que el efecto sea positivo) supera cierto umbral fijado. De manera que momios cercanos a uno implicarían una probabilidad similar entre que el efectos sea positivo y negativo, con lo que se asumirá que el efecto es cero. El esbozo de la elección secuencial de modelos es el siguiente:

i) Considere un modelo con todos los efectos principales y elija de manera laxa aquellos que puedan ser significativos, por ejemplo con momio de 4

ii) Después, considere un modelo con los efectos de interacción y cuadráticos de los factores que resultaron significativos en el paso anterior. Seleccionando de manera estricta, por ejemplo, con un momio de 20 a los efectos significativos

iii) Considere ahora un modelo que incluya a todos los efectos principales y a los efectos de segundo orden considerados como significativos en el paso anterior. De nuevo, seleccione como significativos aquellos efectos cuyos momios sean superiores a 20

iv) Considere un modelo con todos los efectos que resultaron significativos en el paso anterior, incluya todos los efectos de interacción y cuadráticos de los factores que resultaron como significativos en ese paso. Elija aquellos efectos con momios superiores 20 como significativos

[6]Sea A un evento, por ejemplo que el efecto del factor x_i sea positivo, el momio de A está definido por $P(A)/(1 - P(A))$

v) Finalmente, considere el modelo con los efectos que consistentemente mostraron significancia en los pasos anteriores

En este método los momios de un parámetro se calculan con base en su distribución dados los datos, la cual se calcula con la formula de Bayes (6.4)

$$f(\theta|y) \propto f(y|\theta)\pi(\theta) \tag{6.4}$$

Donde $\pi(\theta)$ es la distribución previa (*a priori*) del parámetro y $f(y|\theta)$ es la función de verosimilitud de los datos dado el parámetro.

En el artículo, se propone este método para modelos lineales generalizados, los cuales se mencionan brevemente en el apéndice D, y no se propone una forma cerrada para calcular la expresión anterior.

En su lugar se propone un método de muestreo MCMC, en particular Metropolis-Hastings, con el que se simula el parámetro de acuerdo a su distribución previa y se acepta el valor simulado de acuerdo a su distribución final (*a posteriori*) de manera que la distribución de la muestra simulada aproxima la distribución final. En este caso particular, el parámetro es un vector de parámetros y realmente se simula elemento por elemento (i.e. con los parámetros de la iteración anterior a medida que va avanzando la iteración); el pseudocódigo utilizado en este trabajo de titulación se observa en el algoritmo 2.

Este método de muestreo empieza en un punto completamente aleatorio y las muestras están correlacionadas (el valor de la muestra i depende de la muestra $i-1$ y en el criterio de aceptación), por lo que se descartan las primeras n simulaciones y para la muestra final se considera 1 de cada m simulaciones. Por ejemplo, si $n = 1000$ y $m = 50$, realizar $21,000$ simulaciones resultará en una muestra final de 400 simulaciones.

En el ejemplo que se discute a continuación se puede observar la aplicación de cuatro distintos métodos de cribado de factores, el primero buscando seleccionar a los factores activos de manera frecuentista y los últimos dos de manera bayesiana con los métodos descritos anteriormente.

Algoritmo 2 Método de muestreo para calcular la distribución a posterior

Definir:
$\underline{\beta} = (\beta_0, \ldots, \beta_k)$ ▷ Vector de parámetros a calcular
$\pi(\beta_i)$ ▷ Distribución a priori de cada parámetro
$f(y|\underline{\beta})$ ▷ Función de verosimilitud
N ▷ Número de simulaciones
$\underline{\beta}^{(0)} \sim \pi(\underline{\beta})$ ▷ Se realiza una simulación
$p_0(\underline{\beta} \leftarrow \pi(\underline{\beta}^{(0)})$ ▷ Como vector con k *entradas*
$P(\underline{\beta})) \leftarrow f(y|\underline{\beta}(0)) * p_0(\underline{\beta})$
for $i = 1, \ldots, N$ **do**
    $\underline{\beta}^{(i)} \leftarrow \underline{\beta}^{(i-1)}$
    for $j = 1, \ldots, k$ **do**
        $\underline{\beta}_j^{(i)} \sim \pi(\beta_j)$
        $p_0(\beta_j) \leftarrow \pi(\beta_j)$
        $p_1(\beta_j) \leftarrow f(y|\underline{\beta}(i)) * p_0(\beta_j)$
        if $u < p_1(\beta_j)/P(\underline{\beta}))_j$ **then** ▷ $u \sim Unif(0,1)$
            $P(\underline{\beta}))_j \leftarrow p_1(\beta_j)$
        else
            $\underline{\beta}_j^{(i)} \leftarrow \underline{\beta}_j^{(i-1)}$
        end if
    end for
end for

6.3. Conclusión

El diseño de experimentos es utilizado para obtener propiedades deseables en la estimación de los efectos y por lo tanto para determinar si están activos. Sin embargo, no deja una metodología fija para seleccionar cual es el modelo en que se estimarán los parámetros y el investigador deberá probar distintas estrategias para seleccionar el modelo óptimo a partir del universo de modelos posibles, cuya cardinalidad crece de manera exponencial con el número de factores.

Dichas estrategias pueden ser de tipo frecuentistas o bayesianas, y

existen ya varias metodologías propuestas, definidas y estudiadas por académicos relevantes en el área de experimentación estadística que pueden ser utilizadas.

Capítulo 7

Ejemplo de cribado de factores

En este capítulo se replicará el ejemplo propuesto en el artículo original [Jones and Nachtsheim, 2011] para comparar resultados de los distintos métodos de cribado de factores propuestos en el capítulo anterior.

El ejemplo fue simulado a partir de un modelo descrito por la siguiente expresión:

$$y = 20 + 4x_1 + 3x_2 - 2x_3 - x_4 + 5x_2x_3 + 6x_1^2 + \epsilon \tag{7.1}$$

Donde x_i son los 4 factores activos y ϵ se asumió con una distribución normal con media 0 y varianza 1 independiente para cada ensayo. Se simularon 13 respuestas bajo el diseño experimental de la clase de diseños definitivos de experimentos con 6 factores.

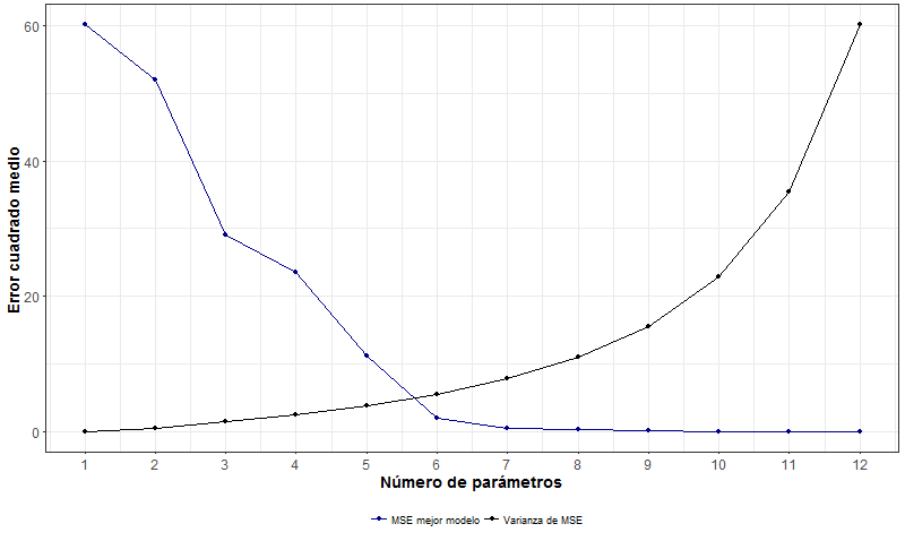
El objetivo de este ejemplo es mostrar cómo se podrían identificar a los factores activos. Se proponen 4 métodos: Los primeros dos siguen una estrategia frecuentista y se seleccionan a los factores activos mediante un criterio de mínimos cuadrados y utilizando el criterio de información de Akaike corregido; en los siguientes dos métodos se utiliza el cribado bayesiano de factores y el análisis secuencial bayesiano.

7.1. Métodos frecuentistas de cribado

7.1.1. Evaluación de hiperparámetros

En el primer análisis, se decidió calcular todos los modelos que tuvieran hasta 12 efectos activos a partir de los 6 factores que cumplieran el principio de herencia, resultando en $95,557$ posibles modelos[1]. Luego se estimaron los parámetros de cada modelo y se calcularon los errores medios cuadráticos y la varianza por número de efectos activos. Los resultados se pueden observar gráficamente en la Figura 7.1.

Figura 7.1: Error cuadrado medio por número de parámetros



En la gráfica, la escala de la varianza fue ajustada para hacerla comparable con el MSE mínimo, por lo que la intersección no tiene particular importancia. Sin embargo, se puede argumentar que el incremento en la varianza de 6 a 7 efectos es desproporcionado a la ganancia en MSE[2],

[1] En 7946 casos de 103523, la matriz del modelo no era de rango completo y se decidió omitirlos)

[2] Además que se forma "un codo" que sugiere que la ganancia ya es insignificante

por lo que se puede justificar elegir 6 efectos.

El modelo de 6 efectos elegido con este método basado en el error cuadrático medio, y con el que se debe seguir experimentando, está estimado por:

$$y = 19.8 + 3.9x_1 + 2.8x_2 - 1.5x_3 + 6.5x_1^2 + 5.3x_2x_3$$

Y en comparación con "el modelo verdadero", se obtuvieron todos los parámetros a excepción de x_4, cuyo efecto es el menor (y de la magnitud de la varianza). Si se hubiera escogido el modelo con 7 factores, se hubiera estimado el siguiente modelo:

$$y = 19.8 + 3.9x_1 + 2.8x_2 - 1.5x_3 - 1x_4 + 6.5x_1^2 + 5.3x_2x_3 \qquad (7.2)$$

El cuál coincide más cercanamente con el utilizado para generar los datos (7.1).

Cabe recalcar que en artículo original no se detectaron todos los efectos activos, siendo x_4 el efecto no detectado[3].

7.1.2. Criterio de información de Akaike corregido

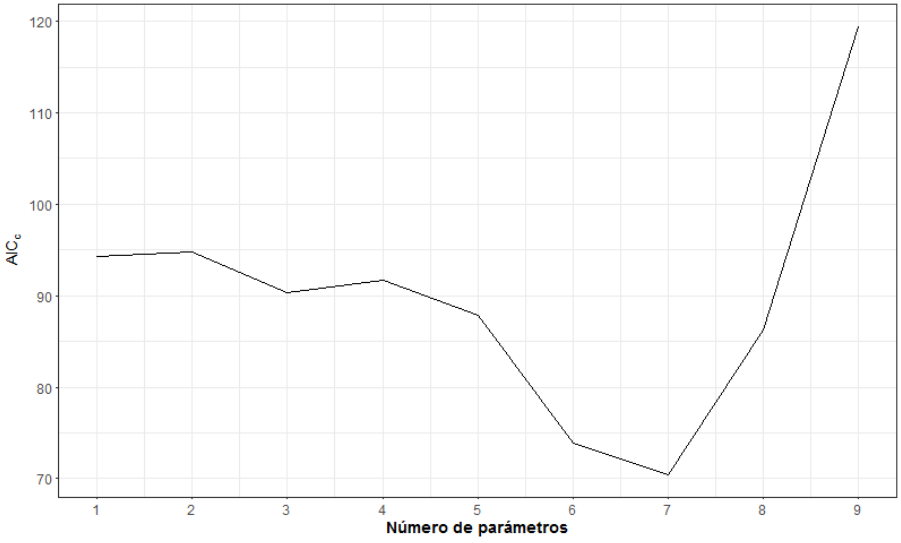
El segundo método consiste en utilizar el criterio de información de Akaike corregido para evaluar los mismos modelos que fueron identificados en el método anterior. El criterio de información de Akaike corregido está definido por [Murvich and Tsai]:

$$AIC_c = n\log(2\pi\hat{\sigma}^2) + n\frac{1 + p/n}{1 - (p+2)/n}$$

Donde p es el número de parámetros, n el de ensayos y $\hat{\sigma}^2$ el estimador de la varianza sin corregir por los grados de libertad. Los resultados de este criterio se observan en la Figura 7.2 Con este método, el mejor

[3]En el artículo se reporta que se utilizo el criterio de información de Akaike corregido para seleccionar modelos, lo que se hará en el siguiente método

Figura 7.2: Critero de información de Akaike



modelo es aquél con 7 parámetros, dado por:

$$y = 19.8 + 3.9x_1 + 2.8x_2 - 1.5x_3 - 1x_4 + 6.5x_1^2 + 5.3x_{23}$$

El cual identifica correctamente los factores activos e incluso estima con bastante precisión los coeficientes. Además es el mismo modelo que se estimó con el método anterior si se hubiera seleccionado 7 parámetros, sin embargo la selección del número óptimo de parámetros en el método anterior es subjetivo pues se elige visualmente.

7.2. Uso del diseño de experimentos

Como se observó en este ejemplo, el número de modelos que se deben evaluar, incluso considerando el principio de herencia es alto y resulta computacionalmente tardado e ineficiente estimarlos todos; para solo 6 factores fue necesario construir las combinaciones y evaluar $103,523$ modelos distintos. Sin embargo, al conocer la correlación que hay entre

los efectos en las observaciones, por ejemplo con el mosaico de correlaciones mostrado en la Figura 5.2, se puede reducir el número de modelos posibles.

Lo primero que se nota, es que los efectos principales no tienen correlación con los efectos cuadráticos y de interacción, por lo que la primer estrategia será estimar solo los efectos principales para luego estimar solo los efectos cuadráticos y de interacción. El primer modelo asi estimado está dado por:

$$y = 24.8 + 3.9x_1 + 2.8x_2 - 1.5x_3 - 1x_4 - 0.4x_5 + 0.1x_6$$

De donde se podría concluir que los factores x_5 y x_6 no están activos en la respuesta por lo que serán omitidos los efectos que incluyan a estos factores. Noté también que los efectos estimados para los efectos principales son iguales a los resultantes en las dos metodologías anteriores.

De esta manera, al eliminar los factores x_5 y x_6, el número de efectos posibles a detectar es $(4) + (4) * ((4) + 1)/2 = 14$. A continuación, se buscará estimar los efectos cuadráticos y de interacción, sin embargo la matriz del modelo que los incluya a todos es singular, por lo que se propone utilizar primero un modelo que solo incluya los términos cuadráticos. Los estimadores resultantes son:

$$y = 19.3 + 11.6x_1^2 + 0x_2^2 - 0.7x_3^2 - 5.2x_4^2$$

Sin embargo, estos estimadores ya tendrán confusión entre ellos y en mayor medida con los efectos de interacción que no incluyen al factor (véase la Figura 5.2), por lo que el estimador del efecto de x_1^2 será distinto si el modelo incluyera a los efectos de interacción que no incluyen al factor x_1. Con esto en mente se estimó el modelo que incluyera a los factores x_1^2, x_4^2 y x_{23}, es decir a los efectos cuadráticos que son notablemente distintos a x_2^2 y x_3^2 y al efecto de interacción que no incluye a los factores incluidos, resultando en:

$$y = 19.7 + 6.3x_1^2 + 0.3x_4^2 + 5.4x_{23}$$

Donde se observa que en realidad el efecto x_4^2 estaba es pequeño, y el detectado anteriormente era debido a la confusión con el efecto de interacción x_{23} y x_{12}, donde este último no se probó.

Por último, se propone entonces un modelo que incluya a lo efectos: x_1, x_2, x_3, x_4, x_1^2, x_{23} y x_{12}, resultando en:

$$y = 19.9 + 3.9x_1 + 2.8x_2 - 1.5x_3 - 1x_4 + 6.4x_1^2 - 0.1x_{12} + 5.3x_{23}$$

Y se observa que el efecto de interacción x_{12} es muy pequeño por lo que se excluye, resultando en el modelo final:

$$y = 19.8 + 3.9x_1 + 2.8x_2 - 1.5x_3 - 1x_4 + 6.5x_1^2 + 5.3x_{23}$$

El cual coincide con el modelo seleccionado utilizando el criterio de información de Akaike corregido, pero estimando solamente 5 modelos en lugar de los $103,523$ evaluados con los otros métodos, resaltando la importancia práctica de conocer el patrón de confusión de los efectos en el experimento y de poder detectar los factores activos en la respuesta.

7.3. Métodos bayesianos de cribado

7.3.1. Detección bayesiana de factores activos

En el tercer método, se utilizó el cribado bayesiano de factores, donde se calculó $p(M_i|y)$ para cada modelo posible[4] y finalmente se estimó la probabilidad de que el factor i estuviera presente como la suma de la probabilidad de todos los modelos que tuvieran al factor i activo.

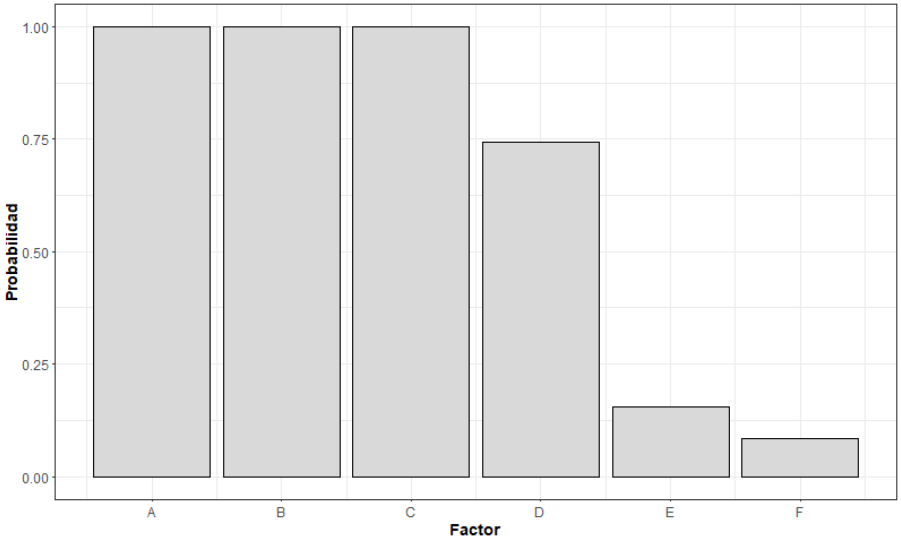
Se utilizó una probabilidad *a priori* de 0.25 para cada factor y que la distribución de los efectos tiene 3 veces mayor variabilidad que la del ruido aleatorio. Con estos parámetros se obtuvo la Figura 7.3.

Con lo que se seleccionarían como activos a los factores x_1, x_2, x_3, x_4 (A, B, C y D) tanto por valor de la probabilidad estimada, como por comportamiento, siendo $P(x_4)$ más de 3 veces mayor que $P(x_5)$.

Con este método se pudo identificar los factores activos, sin embargo no se tiene un modelo estimado. Para poder proponer un modelo para

[4]De los $103,523$ modelos posibles calculados en los métodos pasados

Figura 7.3: Nivel de creencia que un factor este presente



estimar, basta revisar en los pasos intermedios a los modelos que incluyan a los primeros 4 factores para seleccionar al de mayor probabilidad. El cuál identifica a los efectos: x_1, x_2, x_3, x_4, x_1^2, x_2x_3, por lo que el modelo estimado será (7.2), al igual que en los dos métodos anteriores.

7.3.2. Estrategia secuencial bayesiana

A continuación se muestran los resultados de las iteraciones realizadas de acuerdo al método propuesto en [Aguirre, 2016]. En el artículo citado se proponen las siguientes modificaciones al método descrito con anterioridad. La primera es que la probabilidad de que β_i sea positiva o negativa (utilizada en los momios) se calcularán de la siguiente manera (para evitar problemas de continuidad)Aguirre [2016]:

$$P(\beta_i > 0) = \frac{|\{\beta_{ij} \text{ tal que } \beta_{ij} > 0\}| + 1}{n + 2}$$

Donde n denota el numero de simulaciones consideradas y el indice j va de 1 a n. La probabilidad de que el efecto sea negativo se calcula

análogamente:

$$P(\beta_i < 0) = \frac{|\{\beta_{ij} \text{ tal que } \beta_{ij} < 0\}| + 1}{n + 2}$$

Así, en el primer paso, se realiza una iteración con todos los efectos principales, los resultados se encuentran en la Tabla 7.1.

Tabla 7.1: Resultado de la primera iteración

Efecto	Momio Positivo	Momio Negativo
μ	2001	0.00
A	2001	0.00
B	2001	0.00
C	0.03	37.5
D	0.11	9.48
E	0.38	2.65
F	1.48	0.67

Note que el momio negativo es igual a 1 entre el momio positivo; se decidió poner las dos columnas porque es más fácil identificar los momios grandes que indiquen que el efecto es distinto de cero y por lo tanto activo.

En este primer paso, el artículo original propone utilizar un umbral en los momios de 4 para aceptar un efecto como activo. En este caso se seleccionan los factores: A, B, C y D.

En la segunda iteración se sugiere probar con los efectos cuadráticos y de interacción de los factores activos. Sin embargo, esto resultaría en 15 efectos, que no pueden ser estimados con 13 ensayos [5]; de esta manera, se analizará primero un modelo solamente con los efectos cuadráticos, luego uno con los de interacción y por último uno que considere los efectos que resultaron significantes. El resultado del modelo con efectos cuadráticos se observa en la Tabla 7.2.

[5] Aunque en este método no se estiman los efectos directmente, pues se obtienen mediante simulación y no resolviendo un sistema y si es posible calcular la verosimilitud con más variables que ensayos. Sin embargo, sigue habiendo el mismo problema conceptual que tener más variables que observaciones resulta en multiplicidad de las soluciones

Tabla 7.2: Resultado de la segunda iteración

Efecto	Momio Positivo	Momio Negativo
μ	2001	0.00
A	2001	0.00
B	2001	0.00
C	0.02	43.49
D	0.09	10.92
AA	2001	0.00
BB	0.91	1.10
CC	1.95	0.51
DD	0.00	2001

En este paso, el articulo sugiere utilizar 20 como valor crítico con lo que se elegirán a los efectos : A, B, C y AA; el efecto DD fue omitido por el principio jerarquico, ya que D no mostró ser significativo. En el segundo paso intermedio, se probará el modelo con los efectos de interacción y ya es posible agregar AA sin sobresaturarlo, los resultados del segundo paso se muestran en la Tabla 7.3.

Tabla 7.3: Resultado de la segunda iteración

Efecto	Momio Positivo	Momio Negativo
μ	2001	0.00
A	2001	0.00
B	2001	0.00
C	0.00	1000
AB	0.54	1.84
AC	1.98	0.50
BC	2001	0.00
AA	2001	0.00

En el tercer paso, se considerarán todos los factores (efectos principales) y AA y BC como los efectos activos de segundo orden, cuyos resultados se muestran en la Tabla 7.4. En esta iteración se idetifican como activos a los efectos A, B, C, D, AA y BC, los cuales se probaran

Tabla 7.4: Resultado de la tercera iteración

Efecto	Momio Positivo	Momio Negativo
μ	2001	0.00
A	2001	0.00
B	2001	0.00
C	0.00	2001
D	0.00	499.5
E	0.08	11.75
F	2.09	0.48
AA	2001	0.00
BC	2001	0.00

en la siguiente iteración sin el ruido que pudo haber provocado los factores inactivos E y F. El resultado de esta última iteración se observa en la Tabla 7.5.

Tabla 7.5: Iteración final: efectos activos detectados

Efecto	Momio Positivo	Momio Negativo
μ	2001	0.00
A	2001	0.00
B	2001	0.00
C	0.00	2001
D	0.00	666.33
AA	2001	0.00
BC	2001	0.00

El objetivo de esta última iteración es confirmativo, pues ya se habían identificado los efectos activos en las iteraciones pasadas. En este caso, se identificaron correctamente los efectos activos y además se pueden observar el valor númerico que se encontró mediante simulación; se utilizó el promedio de las iteraciones guardadas como estimador bayesiano de los efecto. Todos los estimadores resultaron dentro de media desviación estándar del valor real.

	Factores						
	μ	A	B	C	D	AA	BC
Estimación	19.88	3.86	2.79	-1.49	-1.01	6.41	5.31
Real	20	4	3	-2	-1	6	5

7.4. Conclusión

En el ejemplo presentado en este capítulo se pudo observar, con cuatro tipos de análisis diferentes, dos frecuentistas y dos bayesianos, que es posible recuperar el modelo original cuando se utilizan los diseños de experimentos propuestos por en Jones and Nachtsheim [2011]. Sin embargo, el ejemplo era artificial y no se probó en este trabajo su utilidad en un ejemplo real donde el modelo propuesto tendrá mayores diferencias con el "modelo generador".

Por último, se encontró también que estimar los efectos activos en el modelo, aún con un número pequeño de factores a considerar y con el principio de herencia, resulta en un problema de optimización que resultará ineficiente abordar de manera directa, esto es, buscando en el espacio de todos los modelos posibles aquel que tenga mayor poder explicativo de la varianza de la respuesta sin sobre ajustarse a ella y se preferirá en su lugar utilizar métodos secuenciales en los que se pueda descartar a los factores activos antes de buscar estimar los efectos de interacción y cuadráticos.

Capítulo 8

Conclusión

El diseño de experimentos es una metodología central en el aprendizaje científico, pues permite identificar en qué combinaciones se deberán considerar las variables para maximizar la información que se obtiene de los ensayos, en el sentido que permitan estimar el efecto de cada factor e interacción para posteriormente identificar nivel de actividad de cada efecto. De esta manera, se puede comenzar con un modelo simple que contenga muchos factores y efectos simples para poder seleccionar los más relevantes y refinar el modelo para incluir comportamientos de curvatura y más niveles, llegando de manera iterativa a un modelo detallado.

En este trabajo, se desarrolla la teoría del diseño de experimentos, empezando por intentar estimar el modelo en distintos niveles que pudieran ser de interés, sin importar su cantidad, de los distintos factores. Después, se buscó estudiar el cambio en la respuesta en los niveles extremos de los factores, reduciendo la cantidad de datos necesarios para estudiar el fenómeno y se introdujo el principio de pocos efectos y el principio jerárquico o de herencia, que porporcionan una estructura al modelo que restringe el número de efectos activos. Luego, se introducen diseños específicos, pensados en superar las limitantes de los diseños anteriores, ya sea su crecimiento exponencial en número de ensayos necesarios o su incapacidad de detectar comportamientos cuadráticos.

Se presentaron los diseños definitivos de experimentos, que a diferencia de los diseños mostrados anteriormente, el número de ensayos crece en orden lineal con el número de factores y se mostraron técnicas de análisis que permitan identificar a los factores y efectos activos.

Por último, se mostraron dos metodologías de detección de factores. En la primera,dentro de una filosofía *frecuentista* o *clásica* de la estadística, supone que la incertidumbre de realizar inferencias proviene unicamente del error aleatorio y la selección del modelo está basada en minimizarlo y evaluar la probabilidad de que se haya obtenido un error igual o menor por pura coincidencia, ajustado o no por el número de parámetros utilizados. En la segunda, también conocida como *bayesiana*, la incertidumbre proviene también de los parámetros, por lo que es posible estimarlos con base en el nivel de confianza sobre su valor.

Es interesante que con ambos métodos se obtuvieron resultados similares y satisfactorios en el sentido que se pudieron identificar los efectos activos correctamente en el ejemplo analizado.

Considero que la relevancia de este trabajo reside en que se recolecta y se da una breve descripción sobre una metodología para optimizar los datos requeridos en un experimento para maximizar la información que se pueda extraer de ellos. Dicha disciplina es, a mi parecer, muy extensa, por lo que es de valor tener una recopilación de las metodologías más relevantes, los diseños resultantes y sus propiedades, así como futuras referencias de los temas no cubiertos en el presente trabajo. Además, el trabajo sirve para exponer y difundir una clase de diseños experimentales relativamente nueva, así como los métodos que pueden usarse para la detección de factores activos.

Apendices

Apéndice A

Propiedades de mínimos cuadrados

Este apéndice contiene demostraciones omitidas en el capítulo 2 sobre las propiedades de los estimadores de mínimos cuadrados.

A.1. La respuesta estimada es ortogonal a los residuos

Utilizando la notación matricial, se desea probar que $\hat{Y} = X\hat{\beta}$ es ortogonal a $Y - \hat{Y}$, lo cual se hará directamente.

Demostración.

$$\begin{aligned}
&\hat{\beta} = (X^T X)^{-1} X^T Y \\
&\Rightarrow \hat{Y} = X(X^T X)^{-1} X^T Y \\
&\Rightarrow \hat{Y}^T \hat{Y} = (Y^T X (X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T Y) \\
&\quad = (Y^T X (X^T X)^{-1})(X^T X)(X^T X)^{-1} X^T Y \\
&\quad = (Y^T X (X^T X)^{-1}) X^T Y) = Y^T X \hat{\beta}
\end{aligned}$$

$$\Rightarrow \hat{Y}^T\hat{Y} = Y^T\hat{Y}$$

$$\therefore (Y - \hat{Y})^T\hat{Y} = 0$$

□

A.2. Propiedades del estimador de la varianza

En esta parte del apéndice se demostrará que el estimador propuesto de la varianza en (2.12) es insesgado, la demostración se hará con la notación matricial descrita en el capítulo.

En primer lugar se probará un resultado sobre la matriz $X(X^TX)^{-1}X^T$, que será referida como A por simplicidad. Como siempre, $X \in \mathbb{R}^{nxp}$ y $rango(X) = p$.

$$E[\epsilon^T A\epsilon] = p\sigma^2 \tag{A.1}$$

Demostración. A es simétrica, por lo que existe Q ortogonal tal que:

$$Q^T AQ = \Lambda \tag{A.2}$$

Donde Λ es una matriz con los eigenvalores de A en la diagonal y ceros en los demás lados. Por definición de A, se tiene que $AX = X$, por lo que cada columna de X es un eigenvector de A cuyo eigenvalor es 1 y A tiene rango p, por lo que los otros $n - p$ eigenvalores son 0.

Ahora, sea $y = Q^T\epsilon$, se tiene que:

$$E[y] = Q^T E[\epsilon] = 0 \tag{A.3}$$

$$Var(y) = Q^T Var(\epsilon)Q = Q^T\sigma^2 Q = \sigma^2 \tag{A.4}$$

Juntando la descomposición de A mediante la matriz ortogonal Q y la transformación de y, se puede calcular directamente lo deseado:

$$E[\epsilon^T A\epsilon] = E[\epsilon^T Q\Lambda Q^T\epsilon] = E[y^T\Lambda y]$$

$$
\begin{aligned}
&= E[\sum_{i=1}^{p} y_i^2] = p(Var(y) + E[y]^2) \\
&= p\sigma^2
\end{aligned}
$$

□

Ahora, la prueba que se quería hacer en primer lugar: Sea $y = X\beta + \epsilon$ con $\epsilon \sim \mathcal{N}(0, \sigma^2)$, entonces $E[MS_E] = \sigma^2$

Demostración. En primer lugar, se tiene que:

$$
\begin{aligned}
(Y - \hat{Y})^T(Y - \hat{Y}) &= Y^TY - 2Y^T\hat{Y} + \hat{Y}^T\hat{Y} \\
&= Y^TY - 2Y^T\hat{Y} + \hat{Y}^TX\hat{\beta} \\
&= Y^TY - 2Y^T\hat{Y} + \hat{Y}^TX((X^TX)^{-1}X^T(X\beta + \epsilon)) \\
&= Y^TY - 2Y^T\hat{Y} + \hat{Y}^TX\beta + \hat{Y}^TX(X^TX)^{-1}X^T\epsilon \\
&= Y^TY - 2Y^T\hat{Y} + \hat{Y}^TY - \hat{Y}^T\epsilon + \hat{Y}^TX(X^TX)^{-1}X^T\epsilon \\
&= Y^TY - Y^T\hat{Y} - \hat{\beta}^TX^T\epsilon + \hat{\beta}^TX^TX(X^TX)^{-1}X^T\epsilon \\
&= Y^TY - Y^T\hat{Y} - \hat{\beta}^TX^T\epsilon + \hat{\beta}^TX^T\epsilon \\
&= Y^TY - Y^T\hat{Y}
\end{aligned}
$$

Ahora, el valor esperado del primer término es:

$$
\begin{aligned}
Y^TY &= (X\beta + \epsilon)^T(X\beta + \epsilon) \\
&= \beta^TX^TX\beta + 2\epsilon^TX\beta + \epsilon^T\epsilon \\
&\Rightarrow E[Y^TY] = \beta^TX^TX\beta + N\sigma^2
\end{aligned}
$$

Y el del segundo término es:

$$
\begin{aligned}
&Y^T\hat{Y} = (X\beta + \epsilon)^T(X\hat{\beta}) \\
&= \beta^TX^TX(X^TX)^{-1}X^T(X\beta + \epsilon) + \epsilon^T(X(X^TX)^{-1}X^T(X\beta + \epsilon)) \\
&= \beta^TX^TX\beta + \beta^TX^TX(X^TX)^{-1}X^T\epsilon + \epsilon^TX\beta + \epsilon^TX(X^TX)^{-1}X^T\epsilon \\
&\Rightarrow E[Y^T\hat{Y}] = \beta^TX^TX\beta + p\sigma^2
\end{aligned}
$$

Donde en el último renglón se utilizó el resultado anterior. Juntando ambos términos se obtiene finalmente:

$$E[(Y-\hat{Y})^T(Y-\hat{Y})] = N\sigma^2 - p\sigma^2 = (N-p)\sigma^2 \tag{A.5}$$

□

A.3. Los residuos son ortogonales a la variabilidad explicada

En esta parte del apéndice se probará que el producto cruzado de la descomposición de la suma de cuadrados en la explicada por el modelo y por los residuales es cero. En lugar de calcular el producto cruzado se calculará directamente la igualdad presentada en (2.22). Al igual que la prueba anterior, se utilizará la notación matricial, donde $\mathbb{1}$ representa el vector de unos

Demostración. En primer lugar, se observa que:

$$Y^TY = (Y-\hat{Y}+\hat{Y})^T(Y-\hat{Y}+\hat{Y}) = (Y-\hat{Y})^T(Y-\hat{Y}) + \hat{Y}^T\hat{Y} \tag{A.6}$$

Pues ya se probó que el producto $(Y-\hat{Y})^T\hat{Y} = 0$.

Y en segundo lugar:

$$\begin{aligned}(Y-\bar{Y})^T(Y-\bar{Y}) &= (Y-\frac{1}{n}Y^T\mathbb{1}\mathbb{1})^T(Y-\frac{1}{n}Y^T\mathbb{1}\mathbb{1}) \\ &= Y^TY - 2Y^T\mathbb{1}\mathbb{1}^T\frac{1}{n}Y^T\mathbb{1}\mathbb{1} + \frac{1}{n}Y^T\mathbb{1}\frac{1}{n}Y^T\mathbb{1}\mathbb{1}^T\mathbb{1} \\ &= Y^TY - 2\frac{1}{n}(Y^T\mathbb{1})^2 + \frac{1}{n^2}(Y^T\mathbb{1})^2n \\ &= Y^TY - \frac{1}{n}(Y^T\mathbb{1})^2 \end{aligned} \tag{A.7}$$

Y análogamente:

$$(\hat{Y}-\bar{Y})^T(\hat{Y}-\bar{Y}) = \hat{Y}^T\hat{Y} - \frac{1}{n}(Y^T\mathbb{1})^2 \tag{A.8}$$

$$\Rightarrow \hat{Y}^T\hat{Y} = (\hat{Y} - \bar{Y})^T(\hat{Y} - \bar{Y}) + \frac{1}{n}(Y^T\mathbb{1})^2 \tag{A.9}$$

Sustituyendo (A.6) en (A.9) y esta en el valor de la suma de cuadrados original, cuyo valor está dado por (A.8), se obtiene que:

$$\begin{aligned}(Y - \bar{Y})^T(Y - \bar{Y}) &= Y^TY - \frac{1}{n}(Y^T\mathbb{1})^2 \\ &= (Y - \hat{Y})^T(Y - \hat{Y}) + \hat{Y}^T\hat{Y} - \frac{1}{n}(Y^T\mathbb{1})^2 \\ &= (Y - \hat{Y})^T(Y - \hat{Y}) + (\hat{Y} - \bar{Y})^T(\hat{Y} - \bar{Y})\end{aligned}$$

Con lo que se probo directamente la igualdad en el capítulo, que era el objetivo principal de probar que el producto cruzado es cero (lo cual ya es evidente). □

Apéndice B

Diseños de Box-behnken

En este apéndice se muestran tablas de diseños de Box-Behnken para $k = 3, \ldots, 7$. Las tablas fueron obtenidas de [Box and R.Draper, 2007, cap. 15].

$k=3$				$k=4$					$k=5$					
A	B	C	n	A	B	C	D	n	A	B	C	D	E	n
$\star$	$\star$	0		$\star$	$\star$	0	0		$\star$	$\star$	0	0	0	
$\star$	$\star$	0	12	0	0	$\star$	$\star$	8	0	0	$\star$	$\star$	0	
$\star$	$\star$	0		0	0	0	0	1	0	0	$\star$	0	$\star$	20
0	0	0	5	$\star$	0	0	$\star$		$\star$	0	$\star$	0	0	
				0	$\star$	$\star$	0	8	0	0	0	$\star$	$\star$	
				0	0	0	0	1	0	0	0	0	0	3
				$\star$	0	$\star$	0		0	$\star$	$\star$	0	0	
				0	$\star$	0	$\star$	8	$\star$	0	0	$\star$	0	
				0	0	0	0	1	0	0	$\star$	0	$\star$	20
									$\star$	0	0	0	$\star$	
									0	$\star$	0	$\star$	0	
									0	0	0	0	0	3

Donde las estrellas denotan un diseño factorial determinado por el número de estrellas por columna por bloque. Por ejemplo, para $k = 6$,

$k=6$							$k=7$							
A	B	C	D	E	F	n	A	B	C	D	E	F	G	n
⋆	⋆	0	⋆	0	0		0	0	0	⋆	⋆	⋆	0	
0	⋆	⋆	0	⋆	0		⋆	0	0	0	0	⋆	⋆	
0	0	⋆	⋆	0	⋆	48	0	⋆	0	0	⋆	0	⋆	
⋆	0	0	⋆	⋆	0		⋆	⋆	0	⋆	0	0	0	56
0	⋆	0	0	⋆	⋆		0	0	⋆	⋆	0	0	⋆	
⋆	0	⋆	0	0	⋆		⋆	0	⋆	0	⋆	0	0	
0	0	0	0	0	0	6	0	⋆	⋆	0	0	⋆	0	
							0	0	0	0	0	0	0	6

Tabla B.1: Columnas para sustituir estrellas

x_1	x_2	x_3
-	-	-
+	-	-
-	+	-
+	+	-
-	-	+
+	-	+
-	+	+
+	+	+

se tienen 3 estrellas por columna que serán sustituidas por el diseño completo 2^3 dado en la Tabla B.1

Apéndice C

Detalles sobre el diseño definitivo de experimentos

En este apéndice se exponen algunas explicaciones y demostraciones que se omitieron en el capítulo 5. Al último se presentan los diseño obtenidos en [Jones and Nachtsheim, 2011].

C.1. Efectos principales independientes de términos de segundo orden

En esta parte se probará que $A \leftarrow X_1^T X_2$ es igual a cero, donde X_2 es la matriz que contiene las columnas de los efectos de segundo orden. Se probará que todos los elementos de la matriz A son cero y se denotara como $(A)_{p,st}$ al término de multiplicar la columna del factor p con la interacción st

Demostración. Para probar que $A_{p,st} = 0$ para toda $p, s, t \in \{1, 2, \ldots, k\}$, se utilizará que los renglones de los efectos principales se construyen por pares, donde el renglón par es el reflejo del impar.

Por construcción, cada par de renglones de la multiplicación de las dos columnas de la matriz X_1 ($A[(i : i+1), st], i \in \{1, 3, \ldots, 2k-1\}$) tiene tres posibilidades

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} * \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} * \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$
$$\begin{bmatrix} \bullet \\ \bullet \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Donde $\bullet$ denota -1 o 1, pero es indiferente pues se multiplica por el par de ceros. Como todos los pares de la columna de la interacción son el mismo número y cada par de renglones del factor principal está compuesto por -1 *y* 1, el resultado es que $\sum_i^{2k} X_{ip}X_{is}X_{it} = 0$, de hecho cada dos términos la suma es cero. Note que la observación de arriba sigue aplicando si el factor a multiplicar se encuentra en la interacción y aunque la matriz X_1 no sea ortogonal (como en el caso de $k = 12$), por lo que queda probado que la matriz alias de los efectos principales para este tipo de diseños es cero. □

C.2. Correlación entre la interacciones

En esta parte del apéndice se probarán las expresiones que describen la correlación entre las columnas con efectos de segundo orden.

La primera prueba es para la correlación entre dos efectos cuadráticos (5.3):

Demostración. En la prueba se utilizará que sin importar el factor, la columna de los efectos cuadráticos contiene $2m-2$ unos y 3 ceros, y que el producto de dos de estas columnas tendrá $2m-4$ ceros:

$$r^c_{qq,ss}(m) = \frac{\sum(X_{i,qq} - \bar{X}_{qq})(X_{i,ss} - \bar{X}_{ss})}{SS_{qq}SS_{ss}}$$

$$
\begin{aligned}
&= \sum \frac{X_{i,qq}X_{i,ss} - \bar{X}_{qq}X_{i,ss} - \bar{X}_{ss}X_{i,qq} + \bar{X}_{ss}\bar{X}_{qq}}{SS_{qq}^2} \\
&= \frac{2m-4-2(2m-2)^2/(2m+1)+(2m+1)(2m-2)^2/(2m+1)^2}{2m-2-(2m-2)^2/(2m+1)} \\
&= \frac{2m-8}{6m-6} = \frac{m-4}{3m-3} = \frac{m-1}{3m-3} - \frac{3}{3m-3} \\
&= \frac{1}{3} - \frac{1}{m-1}
\end{aligned}
$$

□

Para la prueba de la fórmula cerrada de la interacción entre un efecto cuadrático se asume que la columna de la interacción suma cero y que hay un número par de factores. La correlación será distinta si la interacción tiene un término en común con el efecto cuadrático o si no lo tiene. La prueba de la fórmula cuando hay un factor en común es:

Demostración.

$$
\begin{aligned}
r^c_{qq,qs}(m) &= \frac{\sum(X_{i,qq} - \bar{X}_{qq})(X_{i,qs} - 0)}{SS_{qq}SS_{qs}} \\
&= \sum \frac{X_{i,qq}X_{i,qs} - \bar{X}_{qq}X_{i,qs}}{SS_{qq}SS_{qs}} \\
&= \frac{0-0}{SS_{qq}SS_{qs}} = 0
\end{aligned}
$$

Ahora si la suma de la columna interacción fuera n

$$
\begin{aligned}
r^c_{qq,qs}(m) &= \frac{\sum(X_{i,qq} - \bar{X}_{qq})(X_{i,qs} - n)}{SS_{qq}SS_{qs}} \\
&= \sum \frac{X_{i,qq}X_{i,qs} - \bar{X}_{qq}X_{i,qs}}{SS_{qq}SS_{qs}} \\
&= \frac{n - n(2m-2)/(2m+1)}{\sqrt{(8m-2)/(2m+1)((2m-4)(2m+1)-n^2)/(2m+1)}} \\
&= \frac{n - n(2m-2)}{\sqrt{(8m-2)(2m-4)(2m+1)-n^2)}}
\end{aligned}
$$

Que es una expresión bastante más complicada que si se supone que la columna interacción suma cero, además que es un supuesto que se cumple en los diseños propuestos, simplemente no se garantiza que siempre se cumpla. □

La segunda correlación a probar es cuando no hay un factor en común (con los supuestos):

Demostración.

$$\begin{aligned}
r^c_{qq,st}(m) &= \frac{\sum(X_{i,qq}-\bar{X}_{qq})(X_{i,st}-0)}{SS_{qq}SS_{st}} \\
&= \frac{\sum X_{i,qq}X_{i,st}-\bar{X}_{qq}X_{i,st}-X_{i,qq}\bar{X}_{st}}{SS_{qq}SS_{st}} \\
&= \frac{\pm 2}{SS_{qq}SS_{st}} \\
SS^2_{qq} &= \sum(X_{i,qq}-\bar{X}_{qq})^2 = \sum X^2_{i,qq}-(2m+1)\bar{X}^2_{qq} \\
&= 2m-2-\frac{2m-2}{2m+1} = \frac{(2m-2)3}{2m+1} \\
SS^2_{st} &= \sum X^2_{i,st}-(2m+1)\bar{X}^2_{st} = 2m-4 \\
\Rightarrow r^c_{qq,st}(m) &= \frac{\pm 2}{\sqrt{(2m-2)3(2m-4)/(m+1)}} \\
&= \pm\sqrt{\frac{m+1}{3(m-1)(m-2)}}
\end{aligned}$$

Donde $\sum X_{i,qq}X_{i,st} = \pm 2$ por el supuesto que $\sum X_{i,st} = 0$ y que $X_{i,qq}$ contiene puros unos excepto un par de ceros que puede coincidir con un par de -1 o con un par de 1 de la columna $X_{i,st}$. Al igual que la correlación pasada, es posible encontrar una fórmula para el caso en que la suma de la columna interacción sea n, pero resulta muy larga y de poca relevancia (recordando que se ha observado que los supuestos se cumplen). □

C.3. Proyección a dos dimensiones con número par de factores activos

En esta sección se probará que las proyecciones de los diseños definitivos a dos dimensiones están balanceadas en el sentido que a cada vértice corresponde el mismo número de ensayos si las columnas son ortogonales.

Demostración. Primero hay que notar que cada columna tiene $2m-2$ elementos distintos de cero y que por definición suma cero, es decir, tiene el mismo número de unos y menos unos. Entonces sean:

$$\begin{aligned} k &\leftarrow \{j | x_{j1} = 1 \wedge x_{j2} \neq 0\} \\ l &\leftarrow \{j | x_{j1} = -1 \wedge x_{j2} \neq 0\} \\ p &\leftarrow \{j | x_{j2} = 1 \wedge x_{j1} \neq 0\} \\ q &\leftarrow \{j | x_{j2} = -1 \wedge x_{j1} \neq 0\} \end{aligned}$$

P.D: $|k \wedge p| = |k \wedge q| = |l \wedge p| = |l \wedge q|$ Como cada columna tiene el mismo número de unos y menos unos y el doblez del diseño implica que cada uno de estos conjuntos tiene el mismo número de elementos:

$$|k| = |l| = |p| = |q|$$

Que las columnas sean ortogonales implica que en el producto punto haya un número igual de unos (creados por 1×1 o por -1×-1) que de menos unos, sea este número a:

$$|k \wedge p| + |l \wedge q| = |k \wedge q| + |l \wedge p| = a \tag{C.1}$$

Además, es claro que $k \wedge l = p \wedge q = \emptyset$ y que $k = k \wedge (p \vee q)$, por lo que:

$$|k| = |k \wedge (p \vee q)| = |(k \wedge p) \vee (k \wedge q)| = |k \wedge p| + |k \wedge q| \tag{C.2}$$

$$|l| = |l \wedge p| + |l \wedge q| \tag{C.3}$$

$$|q| = |k \wedge q| + |l \wedge q| \tag{C.4}$$

Sumando las primeras dos de estas ecuaciones se obtiene:

$$\begin{aligned}
&|k| + |l| = |k \wedge p| + |k \wedge q| + |l \wedge p| + |l \wedge q = 2a \\
&\Rightarrow 2|k| = 2a \\
&\Rightarrow |k| = a
\end{aligned}$$

Por último sumando $|k|$ y su equivalente en ambos lados de la ecuación (C.1), se tiene

$$\begin{aligned}
&|k \wedge p| + |l \wedge q| + |k \wedge p| + |k \wedge q| = 2a \\
&\Rightarrow 2|k \wedge p| + |l \wedge q| + |k \wedge q| = 2a \\
&\Rightarrow 2|k \wedge p| + |q| = 2a \\
&\Rightarrow 2|k \wedge p| = a \\
&\therefore |k \wedge p| = a/2
\end{aligned}$$

Y sustituyendo el valor de $|k \wedge p|$ en C.1 y en la C.2 queda probada la igualdad y que las proyecciones a dos dimensiones resultan balanceadas □

C.4. Tablas de diseños definitivos

	$m=4$				$m=5$					$m=6$						$m=7$						
	A	B	C	D	A	B	C	D	E	A	B	C	D	E	F	A	B	C	D	E	F	G
1	0	+	-	-	0	+	+	-	-	0	+	-	-	-	-	0	+	-	+	-	+	-
2	0	-	+	+	0	-	-	+	+	0	-	+	+	+	+	0	-	+	-	+	-	+
3	-	0	-	+	+	0	-	-	+	+	0	-	+	+	-	-	0	+	-	+	+	-
4	+	0	+	-	-	0	+	+	-	-	0	+	-	-	+	+	0	-	+	-	-	+
5	-	-	0	-	+	-	0	+	-	-	-	0	+	-	-	+	-	0	+	+	+	+
6	+	+	0	+	-	+	0	-	+	+	+	0	-	+	+	-	+	0	-	-	-	-
7	-	+	+	0	+	-	+	0	+	-	+	+	0	+	-	+	-	-	0	+	-	-
8	+	-	-	0	-	+	-	0	-	+	-	-	0	-	+	-	+	+	0	-	+	+
9	0	0	0	0	+	+	+	+	0	+	-	+	-	0	-	-	-	+	+	0	-	-
10					-	-	-	-	0	-	+	-	+	0	+	+	+	-	-	0	+	+
11					0	0	0	0	0	+	+	+	+	-	0	-	+	-	+	+	0	+
12										-	-	-	-	+	0	+	-	+	-	-	0	-
13										0	0	0	0	0	0	+	+	+	+	+	-	0
14																-	-	-	-	-	+	0
15																0	0	0	0	0	0	0

	$m=8$								$m=9$									$m=10$									
	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H	J	A	B	C	D	E	F	G	H	J	K
1	0	-	+	+	-	+	+	+	0	+	+	+	+	+	+	+	+	0	+	+	-	+	+	+	+	-	+
2	0	+	-	-	+	-	-	-	0	-	-	-	-	-	-	-	-	0	-	-	+	-	-	-	-	+	-
3	-	0	-	+	+	+	+	-	+	0	+	-	+	-	-	+	-	+	0	-	+	+	-	+	+	-	-
4	+	0	+	-	-	-	-	+	-	0	-	+	-	+	+	-	+	-	0	+	-	-	+	-	-	+	+
5	-	-	0	+	+	-	-	+	-	+	0	-	+	-	+	-	-	-	+	0	-	-	-	+	-	-	-
6	+	+	0	-	-	+	+	-	+	-	0	+	-	+	-	+	+	+	-	0	+	+	+	-	+	+	+
7	+	-	+	0	+	+	-	-	-	-	+	0	+	-	-	-	+	-	+	+	0	+	-	-	+	+	-
8	-	+	-	0	-	-	+	+	+	+	-	0	-	+	+	+	-	+	-	-	0	-	+	+	-	-	+
9	-	-	+	-	0	-	+	-	+	-	+	-	0	+	+	-	-	-	-	-	-	0	+	+	+	+	-
10	+	+	-	+	0	+	-	+	-	+	-	+	0	-	-	+	+	+	+	+	+	0	-	-	-	-	+
11	+	-	-	-	+	0	+	+	-	-	-	-	+	0	+	+	+	-	+	-	+	+	0	+	-	+	+
12	-	+	+	+	-	0	-	-	+	+	+	+	-	0	-	-	-	+	-	+	-	-	0	-	+	-	-
13	-	+	+	-	+	+	0	+	+	+	-	-	+	+	0	-	+	+	+	-	-	-	-	0	+	+	+
14	+	-	-	+	-	-	0	-	-	-	+	+	-	-	0	+	-	-	-	+	+	+	+	0	-	-	-
15	+	+	+	+	+	-	+	0	-	-	-	+	+	+	-	0	-	+	+	+	+	-	+	+	0	+	-
16	-	-	-	-	-	+	-	0	+	+	+	-	-	-	+	0	+	-	-	-	-	+	-	-	0	-	+
17	0	0	0	0	0	0	0	0	-	+	+	-	-	+	-	+	0	+	+	-	-	+	+	-	-	0	-
18									+	-	-	+	+	-	+	-	0	-	-	+	+	-	-	+	+	0	+
19									0	0	0	0	0	0	0	0	0	+	-	+	-	+	-	+	-	+	0
20																		-	+	-	+	-	+	-	+	-	0
21																		0	0	0	0	0	0	0	0	0	0

	$m=11$											$m=12$											
	A	B	C	D	E	F	G	H	J	K	L	A	B	C	D	E	F	G	H	J	K	L	M
1	0	-	+	-	-	-	-	-	+	-	+	0	-	-	+	-	+	-	+	+	+	-	+
2	0	+	-	+	+	+	+	+	-	+	-	0	+	+	-	+	-	+	-	-	-	+	-
3	-	0	-	-	+	-	-	-	-	+	+	-	0	+	+	+	+	+	+	+	-	-	-
4	+	0	+	+	-	+	+	+	+	-	-	+	0	-	-	-	-	-	-	-	+	+	+
5	-	-	0	+	+	+	+	-	-	-	+	+	+	0	-	+	+	-	+	+	-	+	+
6	+	+	0	-	-	-	-	+	+	+	-	-	-	0	+	-	-	+	-	-	+	-	-
7	-	-	-	0	-	+	+	-	+	+	-	+	-	-	0	+	-	+	-	+	-	-	+
8	+	+	+	0	+	-	-	+	-	-	+	-	+	+	0	-	+	-	+	-	+	+	-
9	+	-	-	+	0	+	-	+	+	+	+	+	+	+	+	0	-	+	+	+	+	+	+
10	-	+	+	-	0	-	+	-	-	-	-	-	-	-	-	0	+	-	-	-	-	-	-
11	-	-	+	+	-	0	-	+	-	+	-	+	-	+	-	+	0	+	+	-	+	-	+
12	+	+	-	-	+	0	+	-	+	-	+	-	+	-	+	-	0	-	-	+	-	+	-
13	-	-	-	+	+	-	0	+	+	-	-	+	+	+	+	-	+	0	-	-	-	-	+
14	+	+	+	-	-	+	0	-	-	+	+	-	-	-	-	+	-	0	+	+	+	+	-
15	-	+	+	+	-	-	+	0	+	+	+	-	-	+	+	+	-	-	0	-	-	+	+
16	+	-	-	-	+	+	-	0	-	-	-	+	+	-	-	-	+	+	0	+	+	-	-
17	-	+	-	-	-	+	-	+	0	-	+	+	-	+	+	+	+	-	-	0	+	+	-
18	+	-	+	+	+	-	+	-	0	+	-	-	+	-	-	-	-	+	+	0	-	-	+
19	+	-	-	-	-	-	+	+	-	0	+	+	+	-	+	+	-	-	+	-	0	-	-
20	-	+	+	+	+	+	-	-	+	0	-	-	-	+	-	-	+	+	-	+	0	+	+
21	+	+	-	+	-	-	-	-	-	-	0	-	+	-	+	+	+	+	-	-	+	0	+
22	-	-	+	-	+	+	+	+	+	+	0	+	-	+	-	-	-	-	+	+	-	0	-
23	0	0	0	0	0	0	0	0	0	0	0	+	-	-	+	-	+	+	+	-	-	+	0
24												-	+	+	-	+	-	-	-	+	+	-	0
25												0	0	0	0	0	0	0	0	0	0	0	0

Apéndice D

Modelos lineales generalizados

La validez y el contenido de esta tesis están fuertemente basados en que la respuesta siga una distribución normal. Sin embargo, existen varias aplicaciones en las que no es el caso. Por ejemplo, si la respuesta es un simple "si" o "no", si se quiere medir una variable cuya respuesta pueda tomar solamente valores positivos o el fenómeno de interés es el número de veces que sucedió un evento, digamos, una falla, al momento de observación; en estos casos la variable de interés podrá ser modelada con una distribución binomial, gamma y Poisson, respectivamente.

La importancia de los **Modelos Lineales Generalizados** (GLM, por sus siglas en inglés) para estudiar fenómenos en los que no se puede suponer una distribución normal, ni se crea que hay suficientes datos para apelar a un resultado asintótico motiva que se mencionen en este trabajo; sin embargo, se abordarán de manera superficial, por lo que se decidió hacerlo en el apéndice.

Los modelos lineales generalizados surgen para poder utilizar la teoría desarrollada para modelos lineales en situaciones donde ese supuesto no es completamente válido[1]. Se supondrá que las respuestas y_i son inde-

[1]La respuesta puede no ser normal, pero se pedirá que pertenezca a la familia

pendientes y con media μ_i que puede ser estimada mediante el predictor lineal η:

$$\eta_i = X_i\beta + \epsilon_i \tag{D.1}$$

Este predictor se relaciona a la media mediante una función *enlace* (también conocida como *link*) : $\eta_i = g(\mu_i)$. A continuación, en la tabla Tabla D.1 se muestra, para los casos mencionados, la función enlace que se utiliza para calcular el predictor lineal.

Tabla D.1: Función enlace para GLM

Distribución	Enlace
Binomial	$\eta_i = ln(\frac{\mu_i}{1-\mu_i})$
Poisson	$\eta_i = ln(\mu_i)$
Gamma	$\eta_i = ln(\mu_i)$

De manera que los pasos a seguir son :

- Calcular el predictor lineal η al que se va a ajustar el modelo lineal
- Estimar los parámetros β con las ecuaciones normales
- Calcular la función enlace inversa, tal que $g^{-1}(X\hat{\beta}) = \mu_i$

exponencial [FRS, 1983, cap. 2]

Bibliografía

V. M. Aguirre. Bayesian analysis of definitive screening designs when the response is nonnormal. *Applied Stochastic Models in Business and Industry*, 2016.

G. Box and D. Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475, 1960.

G. Box and R. Meyer. Finding the active factors in fractionated screening experiments. *Journal of Quality Technology*, 25(2):94–105, 1993.

G. E. Box, J. S. Hunter, and W. G. Hunter. *Statistics for Experimenters*. John Wiley & Sons, Inc, second edition, 2005.

G. E. P. Box and N. R.Draper. *Response Surfaces, Mixtures, and Ridge Ana*. John Wiley & Sons, Inc, 2007.

D. Cox and N.Reid. *The Theory of the Design of Experiments*. Chapman & Hall /CRC, 2000.

N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, Inc, 1998.

R. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

P. M. . J. FRS. *Generalized Linear Models*. Chapman & Hall, second edition edition, 1983.

B. Jones and C. J. Nachtsheim. A class of three-level designs for definitive screening in the prescence of second-order effects. *Journal of Quality Technology*, 43(1), jan 2011.

D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, Inc, 2001.

C. M. Murvich and C.-L. Tsai. *R Package .ᴬICcmodavg". Junio 20 - 2017. Marc J. Mazerolle.*

R. L. Plackett and J. P. Burman. The design of optimum multifactorial experiments. *Biometrika*, 33(4):305–325, Junio 1946.

J. A. Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole, 2007.