UNIVERSIDAD DE SONORA

DIVISIÓN DE CIENCIAS EXACTAS Y NATURALES

DEPARTAMENTO DE MATEMÁTICAS

A BER DE MIS HIJOS
RÁ MI GRANDEZA

Aplicación de la Metodología de Superficies de Respuesta para el Mejoramiento de la Calidad del Aceite de Soya

TESIS

Que para obtener el Título de:

LICENCIADO EN MATEMÁTICAS

Presenta:

Alejandra Siqueiros Tarazón

Hermosillo, Sonora **Octubre de 2004**

DEDICATORIA

A DIOS

Por estar siempre conmigo y por colmarme a mí y a mis seres queridos con salud y amor; también por darme la bendición de estar viva y así poder culminar uno más de mis proyectos.

A MIS PADRES

Quienes me han brindado su amor, tiempo, comprensión, desvelos y mucho más. Los llevo siempre en mi corazón.

A MIS HERMANOS

Martha Cecilia, Ivonne y Ricardo, por su preocupación hacia mí y porque siempre me alentaron a seguir adelante y también por la confianza que depositaron en mí.

A MIS SOBRINOS

Said Isaac, Jesús Eduardo, Jesús Antonio, Dante Ricardo y Juan Pablo, pues con su ternura y cariño me dan fuerza para seguir luchando.

A MIS FAMILIARES Y AMIGOS

Por creer en mi, por estar conmigo en las diferentes etapas de mi vida, por el apoyo moral que siempre me han dado y por tantas cosas más.

A JAIME

Por su amor, su apoyo incondicional y por alentarme en todo momento para la realización de mis metas. Siempre será una parte importante de mi vida.

CONTENIDO

CONTENIDO (cont...)

CAPÍTULO 1

INTRODUCCIÓN

En la actualidad existen muchísimos estudios que han demostrado de diversas maneras los beneficios nutritivos de la soya. Investigaciones médicas recientes han encontrado que el consumo de fríjol de soya puede bajar el nivel de colesterol, reducir el riesgo de cáncer, prevenir la osteoporosis y otras enfermedades crónicas [31]. Por esa razón muchos consumidores buscan comidas hechas con soya, y los investigadores e industrias se han preocupado por introducir nuevas y mejores maneras de incorporar la soya en nuestras dietas.

Las semillas de soya contienen un veinte por ciento de aceite, la mayor parte del cual se extrae cuando éstas se prensan, no contiene colesterol, y tiene bajo contenido en grasas saturadas; además, de contener una mezcla exclusiva de ácidos grasos específicos que reducen el riesgo de enfermedades cardíacas [15] y [24].

Un aspecto importante dentro de la calidad nutricional de los aceites es el contenido de vitaminas, como son los tocoferoles, los cuales son reducidos por efecto de las condiciones extremas durante parte de su procesamiento. Los tocoferoles son reconocidos por su eficiente inhibición de la oxidación lipídica en los alimentos y en los sistemas biológicos, por lo cual es recomendable mantenerlos en los alimentos y no buscar este efecto antioxidante con sustitutos extraños.

En México las pérdidas de tocoferoles se registran en las dos últimas etapas de la refinación, que son el blanqueo y la desodorización, y son de entre el 29% y 47% [25]. Esto nos lleva a promover un nuevo concepto de calidad en los aceites y grasas, es decir, que no sólo se consideren las características físicas y estabilidad oxidativa, sino que también se considere de manera prioritaria el no alterar la calidad nutricional de estos aceites durante su procesamiento.

Para poder lograr las condiciones óptimas de blanqueo, en el cual intervienen factores como la temperatura, cantidad de absorbente, tiempo de contacto y presión, y obtener un aceite de soya blanqueado con un alto contenido de tocoferoles y de buena calidad, necesario para la siguiente etapa del proceso de la refinación que es la desodorización, se requiere diseñar un experimento con el cual se puedan encontrar estas condiciones. Por ello, se trabajó conjuntamente con el M.C. Jesús Ortega, investigador del Departamento de Investigaciones Científicas y Tecnológicas de la Universidad de Sonora y bajo la dirección de M.C. Gudelia Figueroa Preciado, para diseñar un experimento que permitiera estudiar y establecer estas condiciones. Por lo que, el presente trabajo justifica la metodología utilizada en el diseño de este experimento, que consistió en aplicar un diseño central compuesto, el cual es un tipo de diseño de superficie de

respuesta, que requiere ciertos conocimientos previos de diseño experimental y análisis de regresión, que se abordarán en los primeros capítulos de esta tesis. Hay que aclarar que el experimento se corrió con y sin presión, es decir, esta variable es de tipo cualitativo. Los resultados que se mostrarán en el capítulo de aplicación solamente corresponden al diseño realizado en ausencia de oxígeno.

El capítulo dos inicia con las ideas fundamentales del diseño experimental, como son los conceptos de tratamiento, unidad experimental, factores y sus niveles, error experimental, aleatorización, etcétera. Se explican los pasos básicos a seguir al diseñar un experimento, así como se presenta una clasificación de los diseños experimentales, basándose en varios puntos, como son: el objetivo del experimento, la característica de los factores y sus niveles, los efectos que interesa investigar, así como el costo y tiempo con el que se cuenta. Se parte del diseño más simple, que es el diseño completamente al azar o en una clasificación, los resultados de éste se agrupan en una tabla conocida como "andeva" o análisis de varianza, la cual explica y desglosa la variabilidad de las observaciones en el experimento en diferentes partes. Además en esté capítulo, se explican los supuestos que se deben cumplir en el modelo como son la normalidad, homogeneidad de varianzas e independencia. Después de esto se introducen los diseños factoriales, comenzando primeramente con diseños de dos factores y el cómo medir sus efectos, para después pasar a tres factores y su tabla de análisis de varianza correspondiente, con las pruebas de hipótesis respectivas. Dado que el diseño utilizado en el problema del aceite de soya contiene tres factores a dos niveles, se estudia con detalle este tipo de diseño y se explican además las características de los diseños, que, por diversas circunstancias, no pueden incluir todas las corridas de un diseño factorial completo, y que se conocen como diseños factoriales fraccionarios.

En el capítulo tres se aborda el tema de regresión lineal, iniciando con el modelo de regresión lineal simple, para el cual se explican las características, se obtienen los estimadores de los parámetros, se establecen los supuestos que se deben cumplir y se introduce el análisis de varianza para este modelo de regresión, con sus pruebas de hipótesis para la pendiente e intersección, así como pruebas para el coeficiente de correlación lineal. Se explica la importancia de realizar el análisis de residuales y las propiedades que éstos deben de cumplir, entre ellos, cómo verificar que no exista autocorrelación en los errores, lo cual se puede efectuar con la prueba de Durbin-Watson. Como los modelos pueden, en ocasiones, mostrar falta de ajuste, se explica también cómo efectuar esta prueba. Se ejemplifica lo anterior con el problema del aceite de soya, tomando solamente la variable temperatura como variable independiente y la retención de tocoferol como variable dependiente y efectuando todo el análisis expuesto con anterioridad. Se continúa después con el modelo de regresión lineal múltiple, utilizando ahora la notación matricial, que permite obtener más fácilmente los estimadores de los parámetros del modelo y establecer las propiedades de éstos. Se desglosa el análisis de varianza para este modelo y se explican sus pruebas de hipótesis respectivas. En la siguiente sección, con el fin de ilustrar un ejemplo de regresión lineal múltiple, se toma nuevamente como respuesta la retención del tocoferol y

como variables independientes la temperatura y el tiempo. Se hacen los cálculos respectivos para llegar al modelo propuesto, para el cual se realizan pruebas de hipótesis para sus parámetros, prueba de falta de ajuste, de autocorrelación de los errores, etc.

El capítulo cuatro se centra en exponer en qué consiste la metodología de superficie de respuesta, técnica que permite encontrar las mejores condiciones de operabilidad de un proceso, optimizando su(s) respuesta(s). Se ilustra con un diagrama el procedimiento secuencial en que ésta se puede llevar a cabo. Se continúa con el método de escalamiento ascendente y después se introducen los diseños para estimar superficies de respuesta de segundo orden, entre ellos el diseño central compuesto en general, para el cual se explican sus propiedades, y después se abordan diseños similares pero utilizando dos y tres factores exclusivamente. Enseguida se exponen los diseños Box-Behnken y sus propiedades. Se describen después los criterios más comunes para determinar cómo seleccionar el modelo de segundo orden más adecuado; se expone luego una técnica que permite analizar la superficie de respuesta y determinar si ésta contiene un máximo, mínimo, o punto silla, la cual se conoce como análisis canónico y se finaliza explicando el método gráfico para análisis de superficies de respuesta.

El capítulo cinco inicia con una exposición de la importancia de aplicar la metodología de superficie de respuesta en la industria. Se detallan los pasos a seguir en el procesamiento del aceite de soya, partiendo del desgomado. Se enuncian los factores que afectan el proceso de blanqueo, en el cual se centró este experimento, y por qué se escogieron los niveles que se utilizaron en estos factores. Se realiza la aplicación de esta metodología de superficie de respuesta a los datos obtenidos del experimento completo, esto es, se analiza la influencia de los tres factores: temperatura, tiempo y cantidad de tierras en las respuestas estudiadas, que fueron retención de tocoferol e índice de peróxido, considerando, como se dijo antes, la ausencia de oxígeno. Se realizan después las pruebas de hipótesis para los parámetros del modelo y se incluyen, en el modelo final, solamente las que resultan significativas. Se verifican los supuestos del modelo y se proponen valores posibles para los factores en estudio que maximicen la retención de tocoferol y minimicen el índice de peróxido. Se efectúan nuevos cálculos incluyendo solamente como factores la temperatura y el tiempo, pues la cantidad de tierras parece no ser significativa, y se explica la solución obtenida. Se realiza después un análisis similar para la respuesta índice de peróxido y finalmente se propone, utilizando el método gráfico, una solución óptima simultánea para las dos respuestas.

CAPÍTULO 2

PRINCIPIOS DE DISEÑO EXPERIMENTAL

2.1 ¿Qué es un Diseño Experimental?

Se entiende por diseño experimental la planeación de una serie de experimentos donde se varían los valores de las variables de entrada de un proceso o sistema y se miden los valores de la variable respuesta o variable de salida, generalmente con el fin de optimizarla en algún sentido. También se llama diseño experimental al resultado de dicha planeación.

Podemos ilustrar lo anterior con la figura 2.1, donde $X_1, X_2, X_3, \ldots$ son las variables independientes o variables de entrada con las que se va a trabajar en el diseño, $Z_1, Z_2, Z_3 \ldots$, pueden ser otras variables que intervienen en el proceso y que son factores no controlables (o quizá factores que no interesa controlar) o ruido, y Y es la variable de respuesta o variable dependiente.

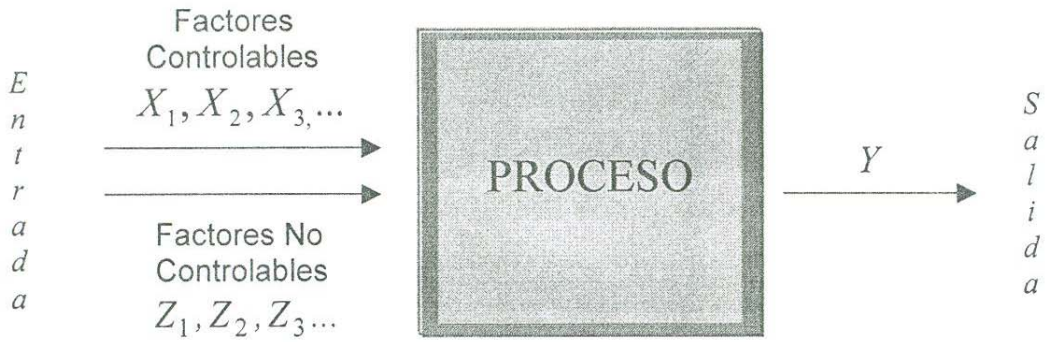


Figura 2.1

El diseño experimental es utilizado ampliamente para la mejora en el rendimiento de los procesos industriales, así como para el desarrollo de nuevos productos, obteniendo de esta manera un ahorro en tiempos y costos de operación. Aporta además un conocimiento profundo de los procesos, generando herramientas eficaces en el manejo de los mismos.

El *diseño estadístico de experimentos* es el proceso de planear un experimento para obtener datos apropiados que puedan ser analizados mediante métodos estadísticos, con objeto de producir conclusiones válidas y objetivas.

Generalizando, podemos aplicar el diseño de experimentos para:

- Determinar qué variables tienen mayor influencia en los valores de respuesta Y.
- Determinar el mejor valor de las variables $X's$, que permitan obtener un valor cercano al valor de respuesta deseado.
- Determinar el o los mejores valores de las variables independientes, con los cuales la variable respuesta tenga menor variabilidad.

"*Para que un experimento se realice en la forma más eficiente, es necesario emplear métodos científicos en su planeación*"[1]. Se requiere entonces de un enfoque estadístico en el diseño de experimentos para obtener conclusiones significativas a partir de los datos. La metodología estadística es el único enfoque objetivo para analizar un problema que involucre datos sujetos a errores experimentales. Así, se puede decir que hay dos aspectos básicos en cualquier problema experimental: el diseño del experimento y el análisis estadístico de los datos.

2.2 Conceptos Básicos en el Diseño de Experimentos

Para comprender los términos utilizados dentro del diseño experimental se deben definir algunos conceptos como los siguientes:

Tratamiento. Es el conjunto de circunstancias que se crean específicamente para el experimento en respuesta a las hipótesis de investigación.

Unidad Experimental. Es la unidad física o el sujeto expuesto al tratamiento, independientemente de otras unidades. La unidad experimental constituye una réplica simple del tratamiento.

Factores y niveles. Un factor es una variable independiente, que puede variar a voluntad del experimentador. El término niveles, hace referencia a las distintas clases, dosis o cantidades de un factor. Un nivel puede ser entonces, una clase, estado o cantidad particular de un factor. Por ejemplo, si se comparan varias razas de cerdos, el factor es la raza y las diferentes razas corresponden a los niveles del factor raza. Si se estudia el efecto de una dieta, sobre la ganancia de peso en pollos, el factor es la dieta y las diferentes dosis o cantidades suministradas, son los niveles.

Como puede verse entonces, se pueden tener factores cualitativos o cuantitativos. Los factores cualitativos son aquellos para cuyos niveles no puede establecerse una noción de distancia. Cada nivel puede considerarse una clase y éstas pueden

[1] Douglas Montgomery (2002), Diseño y Análisis de Experimentos, Editorial Limusa, Segunda Edición.

estar ordenadas o no, por ejemplo, los tipos de raza, diferentes métodos de aplicación, tipos de máquinas, operadores, etc. En nuestro ejemplo el factor cualitativo fue la presión, pues el experimento se corrió con o sin presión. Por otro lado, los factores cuantitativos son aquellos donde los diferentes niveles se expresan en valores numéricos definidos, que corresponden a determinadas cantidades de las variables bajo estudio. Por ejemplo temperatura, dosis, tiempo, etcétera.

Factores Controlables y no Controlables. Los factores controlables son aquellos factores o variables de entrada a las que se les puede asignar ciertos valores o niveles de operación, esto es, son manipulables. Los factores no controlables, también conocidos como factores de ruido, son aquellos que durante la operación del proceso quedan fuera del control del diseñador, como por ejemplo factores ambientales, ánimo de los operadores, calidad u homogeneidad del producto o materia prima recibida, etcétera.

Error Experimental. Este describe la variación observada entre idénticas unidades experimentales, tratadas independientemente. Se puede deber a varias causas, entre ellas: la variación natural entre unidades experimentales, la variabilidad al medir la respuesta, falta de habilidad al reproducir exactamente las mismas condiciones de una unidad a otra, la existencia de interacción entre unidades experimentales y los tratamientos, o bien algún otro factor extraño.

Error Aleatorio. Es la variabilidad observada que no se puede explicar por los factores estudiados y resulta del efecto de los factores no estudiados y del error experimental.

Variable respuesta. Es la característica o variable de salida cuyo valor interesa medir.

2.3 Principios Básicos del Diseño Experimental

Expuestos ya los conceptos anteriores, se puede decir que un *diseño de experimentos* es el arreglo de las unidades experimentales usadas, con el fin de controlar el error experimental y al mismo tiempo asignar el diseño de los tratamientos en el experimento.

A efecto de poder dar un enfoque estadístico al diseño se deben respetar tres principios básicos en el diseño de experimentos:

- Replicación, o repetición de ensayos.
- Aleatorización de las corridas.
- Bloqueo.

Análisis. Se debe tener idea de diseño de experimentos para poder seleccionar el mejor diseño y realizar el análisis de varianza más adecuado, pues se tiene que describir de la mejor manera posible el comportamiento de los datos.

Interpretación. Después del análisis estadístico se deben explicar estos resultados en términos del problema planteado, verificar las conjeturas iniciales, seleccionar el mejor tratamiento, y deducir los nuevos conocimientos encontrados sobre este proceso.

Los cuales podemos definir como:
Replicación. Consiste en correr más de una vez un tratamiento o combinación específica de factores. El efectuar réplicas nos permite estimar la variabilidad natural o error aleatorio, aumentando así la confiabilidad en las mediciones.

Aleatorización. Es la piedra angular que fundamenta el uso de los métodos estadísticos en el diseño de experimentos. Se entiende por aleatorización el hecho de que tanto la asignación del material como el orden en que se realizan las pruebas individuales o ensayos se determinan aleatoriamente. Al aleatorizar adecuadamente el experimento se pueden cancelar los efectos de factores extraños que pudieran estar presentes.

Bloqueo. Un bloque es una parte del material experimental que es más homogénea que el total del material. Al realizarse un análisis por bloques se hacen las comparaciones entre las condiciones de interés del experimentador dentro de cada bloque. La formación de bloques es necesaria para eliminar la variabilidad transmitida por factores perturbadores, es decir aquellos factores que pueden influir en la respuesta pero en los que no hay un interés específico.

2.4 Pasos a seguir para diseñar un experimento.

Para poder diseñar bien un experimento es necesario comprender totalmente el problema que se desea estudiar, elegir las variables más apropiadas y sus niveles de uso, definir la(s) respuesta(s) a evaluar, el diseño experimental a utilizar, realizar el experimento, analizar los datos y obtener las conclusiones correspondientes. Todas estas actividades podemos resumirlas en: *planeación, análisis e interpretación.*

Planeación. En la elección de las variables a utilizar durante el experimento, juega un papel de gran importancia la experiencia previa del experimentador, así como el nivel de conocimientos del problema específico. La elección inapropiada de los niveles de las variables se traduce en la obtención de respuestas fuera de niveles operables. Por ejemplo, la elección de niveles inapropiados de temperatura en el proceso de desodorización nos daría una respuesta tal vez fuera del rango que buscamos.

Análisis. Se debe tener idea de diseño de experimentos para poder seleccionar el mejor diseño y realizar el análisis de varianza más adecuado, pues se tiene que describir de la mejor manera posible el comportamiento de los datos.

Interpretación. Después del análisis estadístico se deben explicar estos resultados en términos del problema planteado, verificar las conjeturas iniciales, seleccionar el mejor tratamiento, y deducir los nuevos conocimientos encontrados sobre este proceso.

2.5 Clasificación de los diseños

Existen varios aspectos que pueden influir en la selección de un diseño experimental, y el modificar alguno(s) conduce generalmente a cambiar el diseño. Estos aspectos son básicamente los siguientes:

1. *El objetivo del experimento.* Es necesario comprender totalmente el problema que se desea estudiar y tener claro el objetivo principal y los objetivos específicos.

2. *El número de factores a controlar.* Es necesario investigar previamente cuál o cuáles factores son los que conviene incluir en el experimento. Si son varios se puede partir de diseños fraccionarios para dilucidar cuál o cuales son los más importantes.

3. *El número de niveles que se prueban en cada factor.* La elección inapropiada de los niveles de las variables se traduce en la obtención de respuestas fuera de los niveles esperados

4. *Los efectos que interesa investigar.* Es importante conocer cuál o cuáles efectos son los más importantes, pues si solamente se incluye una parte de éstos se puede reducir notablemente el diseño.

5. *El costo del experimento, tiempo y precisión deseada.* La consideración de estos aspectos en la selección y planeación del diseño pueden hacer la diferencia entre la selección de un diseño u otro.

El objetivo del experimento se ha utilizado como un criterio general de clasificación de los diseños experimentales, mientras que los otros cuatro aspectos son útiles para subclasificarlos. En estos sentidos, los diseños se pueden clasificar como:

- *Diseños para comparar dos o más tratamientos.*
  - Diseño completamente al azar
  - Diseño de bloques completos al azar
  - Diseño en cuadrados latinos y grecolatinos

- *Diseños para estudiar el efecto de varios factores sobre la(s) respuesta(s).*
  - Diseños factoriales 2^k
  - Diseños factoriales 3^k
  - Diseños factoriales fraccionados 2^{k-p}

- *Diseños para determinar el punto óptimo de operación del proceso.*
  - Diseños para modelos de primer orden:
    - Diseños factoriales 2^k y 2^{k-p}
    - Diseño de Plakett – Burman
    - Diseño Simples

- Diseños para modelos de segundo orden:
  - Diseño central compuesto
  - Diseño Box – Behnken
  - Diseños factoriales 3^k y 3^{k-p}

o *Diseños de mezclas.*
- Diseño de lattice simples
- Diseño simples con centroide
- Diseño con restricciones
- Diseño axial

o *Diseños robustos.*
- Diseños ortogonales
- Diseños con arreglos interno y externo.

En esta tesis se trabajará principalmente con algunos aspectos de las primeras tres clasificaciones.

2.6 Análisis de Varianza en una Clasificación

Probablemente uno de los primeros análisis estadísticos que realiza el investigador, es la comparación de dos medias. Esta situación se plantea cuando se están comparando dos grupos (normalmente dos tratamientos) con relación a una variable cuantitativa. Cuando se generaliza este caso a más tratamientos se utiliza el análisis de varianza en una clasificación o en un solo sentido o dirección. Se comparan entonces, tres o más muestras independientes cuya clasificación viene dada por la variable llamada *factor*. La base de este procedimiento consiste en particionar la variabilidad total en dos componentes que son: la variabilidad entre los promedios de los tratamientos y el gran promedio, y la variabilidad de las observaciones dentro de los tratamientos y el promedio de los tratamientos. De ahí el nombre *análisis de varianza.*

Es de mucha utilidad escribir las observaciones de un experimento con un modelo, por ejemplo un modelo lineal aditivo que puede ser representado por:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad (2.1)$$

con $i = 1,2,...,a$ y $j = 1,2,..., n_i$

que debe cumplir con los supuestos de normalidad, varianza constante e independencia y donde

Y_{ij} representa la j-ésima observación del i-ésimo tratamiento

μ es la media global o general.

τ_i es el efecto del i-ésimo tratamiento

ε_{ij} es el error experimental, de la j-ésima observación en el i-ésimo tratamiento.

Un requerimiento es que el tipo de experimento que se está modelando se lleve a cabo en orden aleatorio, por eso se le conoce también como *diseño completamente aleatorizado.*

Al efectuar un análisis de varianza, interesa probar la igualdad de las a medias de los tratamientos, esto es,

$$E(Y_{ij}) = \mu + \tau_i = \mu_i \qquad i = 1,\ldots, a \quad j = 1,2,\ldots, n$$

y las hipótesis planteadas son:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$$

$$H_1 : \mu_i \neq \mu_j \text{ para al menos un par } (i, j), \qquad i \neq j$$

Definiendo

$y_{i.} = \sum_{j=1}^{n_i} y_{ij}$ total de las observaciones en el *i-ésimo* tratamiento. (2.2)

$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$ gran total de todas las observaciones, (2.3)

$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$ promedio de todas las observaciones, (2.4)

donde $N = \sum_{i=1}^{a} n_i$ es el número total de observaciones.

La tabla 2.1 del análisis de varianza nos resume la variabilidad de las observaciones en el experimento. La variación correspondiente a los tratamientos explica la variabilidad entre las medias de los tratamientos, y la debida al error, que se estima con $s^2 = SC_E/(N-a)$ y que se conoce como cuadrado medio del error, estima la varianza del error experimental.

La tabla de análisis de varianza para comparar varios tratamientos con un solo factor, en un diseño completamente aleatorizado, se muestra como ya se mencionó, en la tabla 2.1.

ANÁLISIS DE VARIANZA (ANDEVA)

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	F_0
Tratamientos	$a-1$	$SC_{Trat}=\sum_{i=1}^{a}\frac{y_{i.}^2}{n_i}-\frac{y_{..}^2}{N}$	$CM_{Trat}=\frac{SC_{Trat}}{a-1}$	$F_0=\frac{CM_{Trat}}{CM_E}$
Error	$N-a$	$SC_E=SCT-SCTrat$	$CM_E=\frac{SC_E}{N-a}$	
Total	$N-1$	$SC_T=\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-\frac{y_{..}^2}{N}$		

Tabla 2.1

De la teoría estadística se sabe que la suma de los cuadrados de variables aleatorias distribuidas normalmente, siguen una distribución ji-cuadrada. El estadístico:

$$F_0=\frac{CM_{Trat}}{CM_E} \tag{2.5}$$

es el cociente de dos variables aleatorias que siguen distribuciones ji-cuadrada y a las que se dividió entre sus respectivos grados de libertad. Bajo el supuesto de que no existe diferencia entre las medias de los tratamientos, se conoce que este estadístico sigue una distribución F, con $a-1$ y $N-a$ grados de libertad (Ver [17] pp. 279-288). Entonces la hipótesis nula

$$H_0:\mu_1=\mu_2=...=\mu_a$$

será rechazada, con un nivel de significancia α, cuando

$$F_0>F_{1-\alpha,a-1,N-a}$$

donde $F_{1-\alpha,a-1,N-a}$ es el valor en la distribución F con $a-1$ y $N-a$ grados de libertad, que deja a su izquierda un área de $1-\alpha$.

2.6.1 Verificación de los supuestos del modelo.

Para que los resultados de un análisis de varianza sean válidos es necesario que los supuestos del modelo se cumplan. Hay varios supuestos como *normalidad, varianza constante e independencia,* que deben cumplirse para poder analizar los datos y probar las hipótesis en un análisis de varianza.

Se mostrará cómo verificar estos supuestos para el caso que se está tratando, que es el de un diseño completamente aleatorizado, el cual, como ya se mencionó, puede ser modelado de la siguiente manera:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

donde $i = 1,\ldots, a \quad j = 1,2,\ldots, n_i$.

La respuesta predicha se puede escribir como

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\tau}_i \tag{2.6}$$

donde $\hat{\mu}$ es la media global estimada y $\hat{\tau}_i$ es el efecto estimado del tratamiento i. Si se estima la media global con $\bar{y}_{..}$ y el efecto del tratamiento con $\bar{y}_{i.} - \bar{y}_{..}$, la respuesta predicha se puede rescribir

$$\hat{y}_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) = \bar{y}_{i.} \tag{2.7}$$

Entonces, el residual asociado a la observación y_{ij} se calcula como la diferencia entre ésta y el valor predicho por el modelo. Esto es,

$$e_{ij} = y_{ij} - \hat{y}_{ij} \tag{2.8}$$

Para verificar los supuestos del modelo en término de los residuales, se deberá cumplir que

- Los e_{ij} siguen una distribución normal con media cero.
- Los e_{ij} son independientes entre sí.
- Los tratamientos tienen varianza constante σ^2.

A continuación se muestra como se podrían verificar estos supuestos.

Supuesto de Normalidad. Se pueden graficar los residuales en papel de probabilidad normal, que es un formato para hacer una gráfica del tipo XY, donde una escala es lineal y la otra logarítmica. Si los residuos siguen una distribución normal la gráfica obtenida se asemeja a una línea recta; cuando no es así, el supuesto de normalidad no se cumple. Para hacer esta gráfica:

- Se ordenan los n residuales de menor a mayor y se les asigna un rango de 1 a n, el cual denotamos por r_i.
- Se calcula su posición de graficación en base a su rango y al total de observaciones, con la siguiente fórmula: $(i-0.5)/n, \; i = 1,2,\ldots, n$.
- Se dibujan los puntos $(r_i, (i-0.5)/n)$, r_i sobre la escala lineal y $(i-0.5)/n$, sobre la escala logarítmica.

Hay que señalar que la interpretación de esta gráfica es subjetiva aunque en muchas ocasiones suficiente para llegar a una conclusión sobre el supuesto de normalidad. También se pueden efectuar pruebas de hipótesis como la de Shapiro Wilks que se explica en el apéndice C.

Homogeneidad de varianzas. Una manera de verificar que la varianza es constante o igual en todos los tratamientos, es trazando una gráfica con los valores predichos (generalmente en el eje X), contra los residuales (en el eje Y). Si estos puntos se distribuyen aleatoriamente en una banda horizontal, esto es sin patrón específico alguno, se puede decir que los tratamientos tienen igual varianza. En el capítulo tres se ilustran con detalle estas gráficas. Los paquetes estadísticos nos proporcionan generalmente estas gráficas y en ocasiones pruebas de homogeneidad de varianzas. En el apéndice C se anexa una de estas pruebas, conocida como prueba de Bartlett. Aunque no se explicarán en este apartado, cabe mencionar que existen transformaciones para estabilizar la varianza y son simplemente transformaciones que cambian la escala de las observaciones con el fin de que cumplan con los supuestos del modelo.

Independencia. La suposición de independencia en los residuos se puede verificar al graficar el orden en que se recolectaron los datos y los residuales correspondientes. No debe mostrar una tendencia o patrón, pues de lo contrario nos indica que existe una correlación entre los errores y por lo tanto el supuesto de independencia no se cumple, lo cual indica generalmente deficiencias en la planeación y ejecución del experimento. Existen además pruebas específicas como la de Durbin-Watson que se explica en el capítulo tres, que permiten diagnosticar la presencia de correlación (autocorrelación) de los residuales ordenados en el tiempo.

2.7 Experimentos Factoriales

En muchas ocasiones el experimento contempla dos o más factores y los diseños conocidos como diseños factoriales son generalmente los más adecuados en estos casos. El experimento factorial, en su concepto, está restringido a un tipo especial de diseño de tratamientos que abarca todos los tratamientos posibles que resultan de combinar cada uno de los diferentes niveles de cada factor a estudiar, lo que quiere decir, que cada tratamiento es una combinación de un nivel de cada uno de los factores involucrados en la investigación.

Entre las ventajas de utilizar experimentos factoriales, se pueden mencionar:

- Los diseños pueden aumentarse para formar diseños compuestos en el caso de que se requiera una exploración más completa.
- Cuando el número de combinaciones de tratamientos es muy grande, se pueden utilizar fracciones del diseño factorial, que son muy útiles en las primeras etapas de una investigación.

- Estos diseños pueden combinarse con diseños por bloques, cuando hay restricciones en la aleatorización.
- El cálculo de los efectos en un diseño factorial es muy sencillo.

Aunque hay que mencionar que el utilizar estos diseños tiene también sus desventajas, entre las cuales están:

- A medida que se incrementa el número de factores o de niveles, el diseño factorial se hace impráctico, debido a las limitaciones de material experimental o de recursos
- Cuando aumenta el número de factores, se dificulta la interpretación de interacciones de orden superior.

A pesar de estas desventajas, el procedimiento factorial es de innegable importancia, y puede ser aplicado a muy variadas situaciones.

2.7.1 Factoriales a Dos Niveles

Estos diseños se utilizan en experimentos en los que intervienen k factores, y cada uno de ellos tiene dos niveles, los cuales pueden ser cuantitativos (temperatura, presión, tiempo, etc...) o cualitativos (máquinas, operadores, proveedores, etcétera). Una réplica completa de tal diseño requiere que se recopilen 2^k observaciones y se conoce como *diseño factorial* 2^k. Este diseño es particularmente útil en las primeras fases del trabajo experimental, cuando hay muchos factores por investigar. Con este diseño se hace menor número de corridas que con las que pueden estudiarse k factores en un diseño factorial completo, ya que cada factor tiene sólo dos niveles. Debe suponerse que la respuesta es aproximadamente lineal en el intervalo de los niveles elegidos de los factores.

Si existen dos factores A y B con a niveles para A y b niveles para B, entonces cada corrida contiene ab combinaciones de los tratamientos. Cuando los factores A y B, están cada uno a dos niveles $A<->$, $A<+>$, $B<->$, $B<+>$, el factorial se indicaría como un 2^2 y constaría de cuatro posibles combinaciones de tratamientos, representadas como:

$A<->, B<->$	ó	1
$A<->, B<+>$	ó	b
$A<+>, B<->$	ó	a
$A<+>, B<+>$	ó	ab

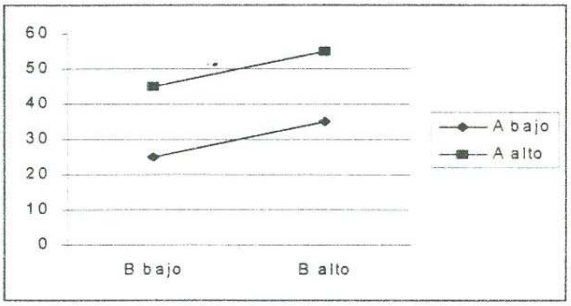
El análisis factorial implica análisis de los efectos de los factores a fin de constatar si el efecto de un factor es independiente del otro factor o si por el contrario existe interacción. Para comprender mejor esto, se debe aclarar cómo medir el efecto en los experimentos factoriales. Esto es, conocer lo siguiente:

Efecto Principal. Es igual a la respuesta promedio observada en el nivel alto de un factor, menos la respuesta promedio en el nivel bajo de éste.

Efecto de Interacción. Cuando el efecto de un factor depende del nivel en el que está otro de los factores, se dice que estos factores interactúan significativamente. Se pueden ilustrar gráficamente estos efectos. Así vemos que en la gráfica 2.1 no existe interacción, pero en la gráfica 2.2 sí hay interacción entre los factores A y B, como se muestra a continuación.

En la gráfica 2.1 el efecto de la interacción AB se mide de la siguiente manera: con el nivel bajo de A el efecto de B es $(35-25)=10$, y con el nivel alto de A, el efecto de B es $(55-45)=10$, la interacción AB es la magnitud de esta diferencia promedio, o sea

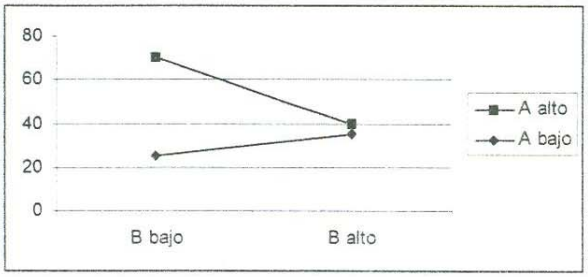
$$AB=(10-10)/2=0,$$



Gráfica 2.1

por lo que se concluye que no existe interacción.
En cambio, en el siguiente caso, para el nivel bajo de A, el efecto de B es $(35-25)=10$ y para el nivel alto de A se tiene $(40-70)=-30$; entonces la interacción

$$AB=(-30-10)/2=-20$$



Gráfica 2.2

es diferente de cero, por lo que se concluye que sí existe interacción entre A y B.

En el análisis de los resultados de un diseño factorial 2^2, como el mostrado anteriormente, es necesario estimar los efectos principales y las interacciones. De no detectarse interacción, los efectos principales serán entonces las mejores estimaciones de los efectos y sobre los cuales se basarían las interpretaciones de la investigación. En caso contrario es necesario examinar e interpretar la naturaleza de la interacción. Cuando una interacción es grande, los efectos principales correspondientes tienen muy poco significado práctico. El conocimiento de la interacción AB es generalmente más útil que el conocimiento del efecto principal de A, puesto que el efecto del factor A depende de los niveles del factor B.

Muchas veces se utiliza el método de analizar los factores uno a la vez, cuando lo más apropiado es hacerlo simultáneamente, ya que si estudiamos los factores por separado nos dará un resultado que puede no corresponder a situaciones reales del experimento, con lo cual se podría generar confusión y tal vez gastos innecesarios.

2.7.2 Diseños factoriales con tres factores.

Cuando se tienen tres factores, que podemos denotar por A, B y C, los cuales tienen a, b, y c niveles, respectivamente, entonces el arreglo factorial completo tendrá $a \times b \times c$ tratamientos.

En un diseño factorial 2^3, el comportamiento de la variable respuesta se puede describir mediante el siguiente modelo

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \qquad (2.9)$$

$$i = 1,2,\ldots,a; \quad j = 1,2,\ldots,b; \quad k = 1,2,\ldots,c; \quad l = 1,2,\ldots,n$$

donde

- μ es la media global
- α_i es el efecto del i-ésimo nivel del factor A
- β_j es el efecto del j-ésimo nivel del factor B
- γ_k es el efecto del k-ésimo nivel del factor C
- $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, y $(\beta\gamma)_{jk}$, representan los efectos de las interacciones dobles, en los niveles ij, ik, y jk respectivamente.
- $(\alpha\beta\gamma)_{ijk}$ es el efecto de interacción triple en la combinación ijk.

Por último,

- ε_{ijkl} representa el error aleatorio.

Para este diseño se tienen siete efectos de interés y para cada uno de ellos se puede plantear una hipótesis nula del tipo

$$H_0 : Efecto\ A = 0$$

con su correspondiente hipótesis alternativa

$$H_1 : Efecto\ A \neq 0$$

La hipótesis nula será rechazada cuando la probabilidad $P(F > F_0)$ calculada, correspondiente al efecto, sea menor que el nivel de significancía especificado.

La tabla de análisis de varianza correspondiente se muestra en la tabla 2.2:

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F_0
A	$a-1$	SC_A	CM_A	CM_A / CM_E
B	$b-1$	SC_B	CM_B	CMB / CM_E
C	$c-1$	SC_C	CM_C	CM_C / CM_E
AB	$(a-1)(b-1)$	SC_{AB}	CM_{AB}	CM_{AB} / CM_E
AC	$(a-1)(c-1)$	SC_{AC}	CM_{AC}	CM_{AC} / CM_E
BC	$(b-1)(c-1)$	SC_{BC}	CM_{BC}	CM_{BC} / CM_E
ABC	$(a-1)(b-1)(c-1)$	SC_{ABC}	CM_{ABC}	CM_{ABC} / CM_E
Error	$abc(n-1)$	SC_E	CM_E	
Total	$abcn-1$	SC_T		

Tabla 2.2

Las sumas de cuadrados indicadas en la tabla se calculan de la siguiente forma:

$$SC_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} y_{ijkl}^2 - \left(\frac{y_{....}^2}{abcn}\right), \qquad SC_A = \sum_{i=1}^{a}\frac{y_{i...}^2}{bcn} - \left(\frac{y_{....}^2}{abcn}\right),$$

$$SC_B = \sum_{j=1}^{b}\frac{y_{.j..}^2}{acn} - \left(\frac{y_{....}^2}{abcn}\right), \qquad SC_C = \sum_{k=1}^{c}\frac{y_{..k.}^2}{abn} - \left(\frac{y_{....}^2}{abcn}\right),$$

$$SC_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{y_{ij..}^2}{cn} - \left(\frac{y_{....}^2}{abcn}\right) - SC_A - SC_B,$$

$$SC_{AC} = \sum_{i=1}^{a}\sum_{k=1}^{c}\frac{y_{i.k.}^2}{bn} - \left(\frac{y_{....}^2}{abcn}\right) - SC_A - SC_C,$$

$$SC_{BC} = \sum_{j=1}^{b}\sum_{k=1}^{c}\frac{y_{.jk.}^2}{an} - \left(\frac{y_{....}^2}{abcn}\right) - SC_B - SC_C,$$

$$SC_{ABC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{y_{ijk.}^2}{n} - \left(\frac{y_{....}^2}{abcn}\right) - SC_A - SC_B - SC_C - SC_{AB} - SC_{AC} - SC_{BC},$$

$$SC_E = SC_T - \left(\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{y_{ijk.}^2}{n} - \frac{y_{....}^2}{abcn}\right)$$

2.7.3 Diseños factoriales 2^3.

Cuando en particular se tiene el caso de tres factores, cada uno de ellos a dos niveles, se puede representar geométricamente el diseño por medio de un cubo donde se observan las ocho combinaciones de tratamientos. En estos casos se utiliza la notación "+" y "-" para representar los niveles alto y bajo de los factores, o bien las letras itálicas minúsculas a, b, y c indicarán el nivel alto del factor. (ver las figuras 2.2 y 2.3) :

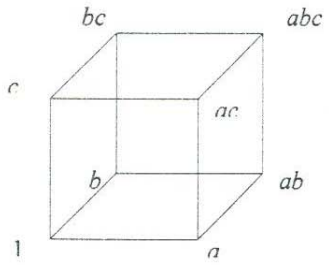


Figura 2.2

En los diseños factoriales a dos niveles es muy sencillo calcular la estimación de los efectos y las sumas de cuadrados. Por ejemplo, el efecto promedio de A se calcula como:

$$A = \frac{1}{4n}\left[a - (1) + ab - b + ac - c + abc - bc\right]$$

esto es, el efecto de A es el promedio de las cuatro corridas donde A está en el nivel alto $(\bar{y}_{A^+})$ menos el promedio de las cuatro corridas donde A está en el nivel bajo $(\bar{y}_{A^-})$. Esto se muestra en la figura 2.3:

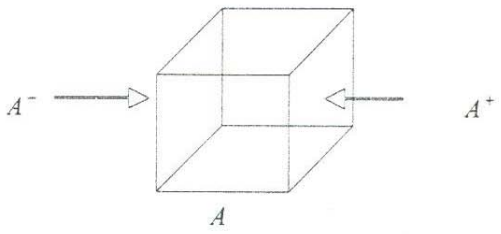


Figura 2.3

Es muy apropiado presentar la información como se muestra en la tabla 2.3, donde se pueden enlistar en la primera columna las diferentes combinaciones de tratamientos y en el primer renglón los efectos factoriales. Al observar las columnas de los efectos principales se puede ver que se utiliza la notación (+) para indicar que el tratamiento contiene la letra del efecto, y las columnas de las interacciones se obtienen multiplicando, según sea la interacción, las columnas correspondientes a los efectos principales. Con esta tabla es muy fácil calcular los efectos principales que se mostraron anteriormente.

Combinación De tratamientos	*Efecto factorial*							
	I	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
c	+	-	-	+	+	-	-	+
abc	+	+	+	+	+	+	+	+
ab	+	+	+	-	+	-	-	-
ac	+	+	-	+	-	+	-	-
bc	+	-	+	+	-	-	+	-
(1)	+	-	-	-	+	+	+	-

Tabla 2.3

El cálculo de las interacciones es similar, por ejemplo la interacción AB se calcularía

$$AB = \frac{[abc - bc + ab - b - ac + c - a + 1]}{4n}$$

y la interacción triple será

$$ABC = \frac{[abc - bc - ac + c - ab + b + a - (1)]}{4n}$$

En las fórmulas anteriores para calcular los efectos, lo expresado entre corchetes es lo que se conoce con el nombre de *contraste* y a partir de éstos es muy sencillo calcular las sumas de cuadrados para cada uno de los factores y sus interacciones cuando se tiene un diseño factorial a dos niveles. Por ejemplo, en un diseño 2^3 con n repeticiones, las sumas de cuadrados se calculan

$$SC = \frac{(Contraste)^2}{8n} \qquad (2.10)$$

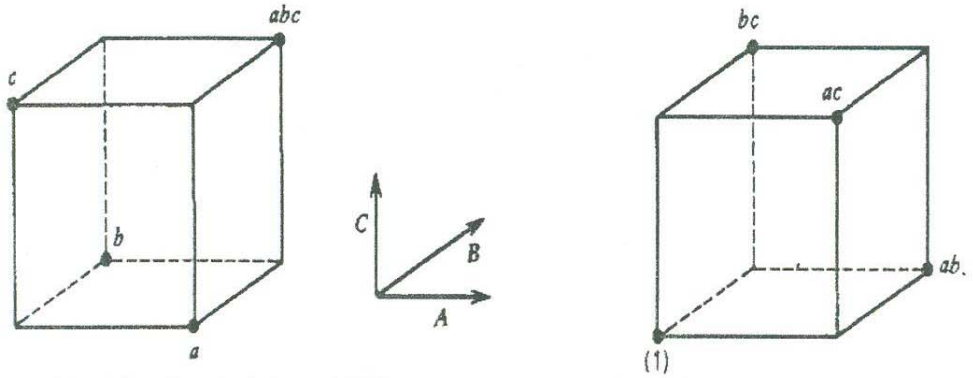
2.7.4 Factoriales Fraccionarios a Dos Niveles

A medida que el número de factores en un diseño factorial 2^k aumenta, el número de combinaciones de tratamientos sobrepasan generalmente los recursos de la mayoría de los experimentadores. En algunos diseños la información de los efectos principales y las interacciones de menor orden puede obtenerse realizando sólo una fracción del experimento factorial completo, cuando algunas interacciones de orden superior son despreciables. A estos diseños se les conoce como *diseños factoriales fraccionarios* y se encuentran entre los más usados para el diseño de productos y procesos, así como para detección y solución de problemas. Este tipo de diseño se practica mucho en los experimentos cribados, esto es, aquellos experimentos donde se consideran muchos factores con el fin de identificar aquellos que tienen efectos importantes. Cuando los hay, se pueden investigan con mayor detalle en experimentos posteriores.

Aunque en el problema de aplicación que manejaremos en esta tesis, no se utiliza un factorial fraccionario, es de importancia mencionarlos ya que son ampliamente utilizados. Por ejemplo, para el caso de un factorial 2^3 se puede usar una fracción un medio de este diseño. Esta fracción puede formarse considerando solamente las combinaciones que tienen signo positivo en la columna ABC de la tabla 2.3. De esta manera a ABC se le llama el *generador* de esta fracción. Como la columna I contiene solo valores positivos, se tiene que

$$I = ABC$$

la cual se conoce como *relación de definición* del diseño. También se podían haber tomado los de signo negativo en la columna ABC, y se tendría que $I = -ABC$, como se muestra en la figura 2.4.



(a) (a) Fracción principal, $I = +ABC$ (b) Fracción alterna, $I = -ABC$

Figura 2.4

En la fracción un medio de un diseño 2^3, o sea en un diseño 2^{3-1}, se tienen tres grados de libertad que pueden usarse para estimar los efectos principales, los cuales, utilizando la tabla 2.3, pueden calcularse como:

$$\ell_A = \frac{(a - b - c + abc)}{2}$$

$$\ell_B = \frac{(-a + b - c + abc)}{2}$$

$$\ell_C = \frac{(-a - b + c + abc)}{2}$$

Se puede observar, entonces, que si queremos calcular los efectos de las interacciones dobles, éstos serán:

$$\ell_{BC} = \frac{(a - b - c + abc)}{2}$$

$$\ell_{AC} = \frac{(-a + b - c + abc)}{2}$$

$$\ell_{AB} = \frac{(-a - b + c + abc)}{2}.$$

Se tiene, por lo tanto, que $\ell_A = \ell_{BC}$, $\ell_B = \ell_{AC}$ y $\ell_C = \ell_{AB}$; entonces es imposible diferenciar entre A y BC, entre B y AC, y entre C y AB. Cuando dos o más efectos tienen esta propiedad se les denomina *alias*.

Cuando en un diseño se tienen alias, se maneja el concepto de *resolución del diseño* para poder conocer la estructura de estos alias. Los más importantes son los diseños de resolución III, IV y V, que se definen a continuación.

Diseño de Resolución III. Es aquel en que ningún efecto principal se confunde con otro efecto principal, pero los efectos principales se confunden con interacciones de dos factores y éstas algunas veces con otras interacciones de dos factores.

Diseño de Resolución IV. Ningún efecto principal se confunde con otro efecto principal o con interacciones de dos factores, pero las interacciones de dos factores son alias entre sí.

Diseño de Resolución V. Son diseños en el que ningún efecto principal ni interacción de dos factores se confunde con otro efecto principal o interacción de dos factores, pero éstas últimas se confunden con las interacciones de tres factores.

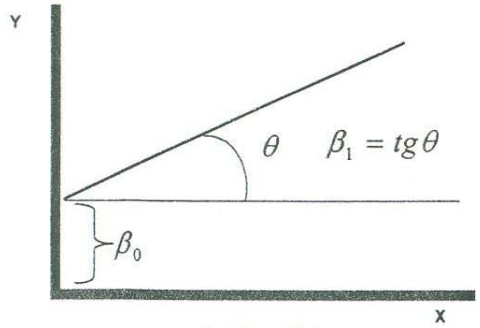
Es de gran importancia conocer y manejar estos conceptos pues en la práctica se encuentra con frecuencia esta clasificación que puede ser de gran influencia en la selección del diseño a utilizar.

CAPÍTULO 3

ANÁLISIS DE REGRESIÓN LINEAL

3.1 Introducción al Análisis de Regresión Lineal

El objetivo de efectuar un análisis de regresión es investigar la relación estadística que existe entre una variable *dependiente* Y con una o más variables *independientes* $X_1, X_2, ..., X_n$. Para poder realizar esto, se postula una relación funcional entre estas variables. Debido a su simplicidad analítica, la forma funcional que más se utiliza en la práctica es la relación *lineal*. Cuando solo existe una variable independiente, esto se reduce a modelar los datos con una línea recta y se le conoce como regresión lineal simple, como se puede apreciar en la gráfica 3.1:



Gráfica 3.1

$$Y = \beta_0 + \beta_1 X, \qquad (3.1)$$

En este caso los coeficientes β_0 y β_1 son parámetros que definen la posición e inclinación de la recta, así β_0 es la ordenada en el origen y β_1 es la pendiente, que nos indica cuánto aumenta Y por cada aumento de una unidad en X.

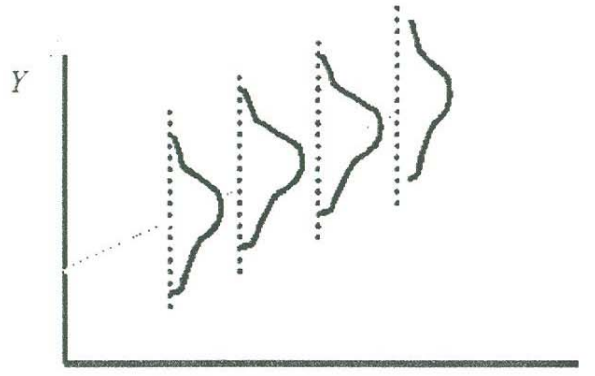
Un modelo de regresión lineal entre dos variables, de la forma $Y = \beta_0 + \beta_1 X,$ es un modelo probabilístico, que puede también escribirse de la siguiente manera:

$$Y / X = \beta_0 + \beta_1 X + \varepsilon \qquad (3.2)$$

A la variable Y se la denomina variable *dependiente* y a X *variable independiente.*

En este modelo de regresión lineal se asume que:

a. X no es una variable aleatoria.
b. Para cada valor X_i de X existe una variable aleatoria Y / X_i cuya media está dada por el modelo.
c. Todas las variables Y / X_i se distribuyen normalmente y se suponen independientes, con igual varianza; esto se muestra en la gráfica 3.2.



Gráfica 3.2

A partir de una muestra aleatoria, la teoría estadística permite:

a. Estimar los coeficientes β_0, β_1 del modelo de regresión lineal simple, para lo cual se pueden utilizar algunos métodos, entre los se pueden mencionar, por ejemplo, mínimos cuadrados y máxima verosimilitud.

b. Estimar la varianza de las variables Y/X_i, conocida como *cuadrado medio del error* y representada por s^2 o CM_E. A su raíz cuadrada se le conoce como *error estándar de la estimación.*

c. Conocer la distribución muestral de los estimadores utilizados, esto contempla calcular su error estándar y valor esperado, lo que permite hacer una estimación de los parámetros, ya sea por intervalos de confianza o por contrastes de hipótesis.

Para el modelo de regresión lineal simple propuesto anteriormente $Y = \beta_0 + \beta_1 X + \varepsilon$, su modelo estimado es $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, donde los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ son aquellos que minimizan $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Para esto se define $S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, (3.3)

donde $$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

el cual debe minimizarse. Para ello se toman las derivadas parciales de S con respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$, se igualan a cero, y se resuelven para $\hat{\beta}_0$ y $\hat{\beta}_1$.

Así,

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

de donde se obtiene

$$-\sum y_i + \sum \hat{\beta}_0 + \sum \hat{\beta}_1 x_i = 0 \quad \text{y}$$

$$\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

$$-\sum x_i y_i + \sum \hat{\beta}_0 x_i + \sum \hat{\beta}_1 x_i^2 = 0$$

Así, simplificando quedan las llamadas ecuaciones normales, que son

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i$$

y

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

Es usual utilizar la notación siguiente:

$$S_{xy} = \sum_{j=1}^{n} y_j (x_j - \bar{x}) = \sum_{j=1}^{n} x_j y_j - \frac{\left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)}{n} \tag{3.4}$$

$$S_{xx} = \sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{\left(\sum_{j=1}^{n} x_j\right)^2}{n} \tag{3.5}$$

$$S_{yy} = \sum_{j=1}^{n} (y_j - \bar{y})^2 = \sum_{j=1}^{n} y_j^2 - \frac{\left(\sum_{j=1}^{n} y_j\right)^2}{n} \tag{3.6}$$

De esta manera se tiene que:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.7}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{3.8}$$

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ mostrados anteriormente satisfacen las ecuaciones normales y son llamados estimadores de mínimos cuadrados, pues minimizan S. Estos mismos estimadores se obtienen por el método de máxima verosimilitud, lo cual se muestra en el apéndice C. En notación matricial se puede expresar lo anterior como,

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

y

$$\begin{bmatrix} \hat{\beta}^*_o \\ \hat{\beta}^*_1 \end{bmatrix} = \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n\sum x_i y_i \end{bmatrix}$$

donde $\hat{\beta}^*_0$ es una estimación de β_0 y $\hat{\beta}^*_1$ es una estimación de β_1.

3.2 Análisis de Varianza para Regresión Lineal

Se puede utilizar el método de análisis de varianza para probar si el modelo de regresión es significativo. Este se efectúa particionando la variabilidad total de la siguiente manera:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (3.9)$$

lo cual se obtiene partiendo de la identidad

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

se eleva al cuadrado en ambos lados de la ecuación, y se suma para las n observaciones. La ecuación (3.9) se puede escribir como:

$$SC_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SC_R + SC_E$$

donde $$SC_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2;$$

sustituyendo $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_i x_i$ se puede llegar a que

$$SC_E = SC_T - \hat{\beta}_1 S_{xy}.$$

De ahí

$$SC_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}. \tag{3.10}$$

Para la regresión se tiene un grado de libertad, por lo que:

$$CM_R = SC_R \tag{3.11}$$

y

$$CM_E = SC_E /(n-2) \tag{3.12}$$

Las hipótesis a probar en un análisis de varianza son:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

y la tabla de análisis de varianza se forma como lo muestra la tabla 3.1:

ANDEVA PARA REGRESIÓN

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Cuadrado medio*	F_0
Regresión	$SC_R = \hat{\beta}_1 S_{xy}$	1	CM_R	CM_R / CM_E
Error o residual	$SC_E = SC_T - \hat{\beta}_1 S_{xy}$	$n - 2$	CM_E	
Total	SC_T	$n-1$		

Tabla 3.1

La hipótesis nula H_0, se rechazará si $P(F^1_{n-2} > F_0) < \alpha$, donde $\alpha = P(rechazar\ H_0 / H_0\ es\ \ cierta)$ y $F_0 = CM_R / CM_E$ es el estadístico que se utilizará en la prueba, el cual sigue una distribución F con 1 y $n-2$ grados de libertad, como se demuestra en el apéndice C. Cuando se rechaza H_0, se concluye que la prueba es significativa para el nivel de significancia establecido.

3.3 Prueba de Hipótesis de la Pendiente y de la Ordenada al Origen.

Con frecuencia se desea probar la significancia de la pendiente y de la ordenada al origen en el caso de una regresión lineal simple. Estas pruebas necesitan el supuesto de que los errores estén distribuidos en forma normal e independiente, con media cero y varianza constante σ^2. Las hipótesis, por ejemplo para el caso de la pendiente son

$$H_0 : \beta_1 = \beta_1^*$$
$$H_1 : \beta_1 \neq \beta_1^*$$

donde β_1^* es una constante. El estadístico de prueba para este se puede formar como:

$$T_0 = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{CM_E / S_{xx}}} \tag{3.13}$$

ya que $E(\hat{\beta}_1) = \beta_1$, y $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, como se demuestra en el apéndice C. Se toma además CM_E como un estimador de σ^2. Se rechaza la hipótesis nula cuando $|T_0| > T_{1-\alpha/2,n-2}$.

Por otra parte, para el caso de querer probar

$$H_0 : \beta_0 = \beta_0^*$$
$$H_1 : \beta_0 \neq \beta_0^*$$

el estadístico a utilizar sería:

$$T_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{CM_E\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \tag{3.14}$$

donde también β_0^* es una constante. En el apéndice C, se demuestra también que $E(\hat{\beta}_0) = \beta_0$, y $V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$. Se rechazará la hipótesis nula, al nivel de significancia α, cuando $|T_0| > T_{1-\alpha/2,n-2}$.

3.4 Adecuación del Modelo de Regresión.

Una vez ajustada la recta de regresión a las observaciones, es importante disponer de una técnica que mida la bondad de ajuste realizado y permita decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos, transformaciones, etcétera. A continuación se muestran algunas maneras de realizar esto.

3.4.1 Coeficiente de Determinación y de Correlación.

Es común utilizar el *coeficiente de determinación* como una medida de la bondad del ajuste. Este se define como:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SC_R}{SC_T}, \tag{3.15}$$

Como $SC_R \leq SC_T$, se verifica que $0 \leq R^2 \leq 1$.

comprobar si se cumplen o no los supuestos del modelo. Un residual es la diferencia entre el valor estimado por la línea de regresión y el valor observado, es decir:

$$e_i = y_i - \hat{y}_i \quad i = 1,2,\ldots,n \tag{3.19}$$

Puede considerarse al residual como la desviación entre los datos y el ajuste; también mide la variabilidad de la variable respuesta, que no es explicada por el

Y el *coeficiente de determinación ajustado* se calcula: $R^2_{Ajust} = 1 - \dfrac{SC_E/(n-2)}{SC_T/(n-1)}$,

El coeficiente de determinación mide el porcentaje de variabilidad total, de la variable dependiente Y respecto a su media, que puede ser explicado por el modelo de regresión propuesto.

Por otra parte, dadas dos variables X e Y, una medida de la relación lineal que hay entre ambas variables es el coeficiente de correlación definido por:

$$\rho = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{3.16}$$

donde $\sigma(X)$ representa la desviación estándar de la variable X y $\sigma(Y)$ es la desviación estándar de la variable Y. Un buen estimador de este parámetro es el coeficiente de correlación lineal muestral o *coeficiente de correlación de Pearson*, definido por

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{3.17}$$

Este coeficiente es una medida de la bondad del ajuste de la recta de regresión, y su valor está en el intervalo [-1,1]. Es recomendable, efectuar una prueba de hipótesis para el coeficiente de correlación, y para ello las hipótesis a plantear son:

$$H_o : \rho = 0$$
$$H_1 : \rho \neq 0$$

Para hacer inferencias acerca de ρ utilizando r, se requiere conocer la distribución muestral de r, que es bastante compleja, por lo que se tomará el resultado de la teoría estadística de que, cuando $\rho = 0$ y (X,Y) es normal bivariada, el estadístico

$$T_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3.18}$$

sigue una distribución *t* de Student con $n-2$ grados de libertad y puede ser utilizado para probar $H_o : \rho = 0$. La hipótesis nula será rechazada, al nivel de significancia α, cuando $|T_0| > T_{1-\alpha/2,n-2}$.

3.4.2 Análisis de los residuales

También es importante estudiar el análisis de residuales, pues este permite comprobar si se cumplen o no los supuestos del modelo. Un residual es la diferencia entre el valor estimado por la línea de regresión y el valor observado, es decir: $e_i = y_i - \hat{y}_i \quad i = 1,2,\ldots,n$ (3.19)

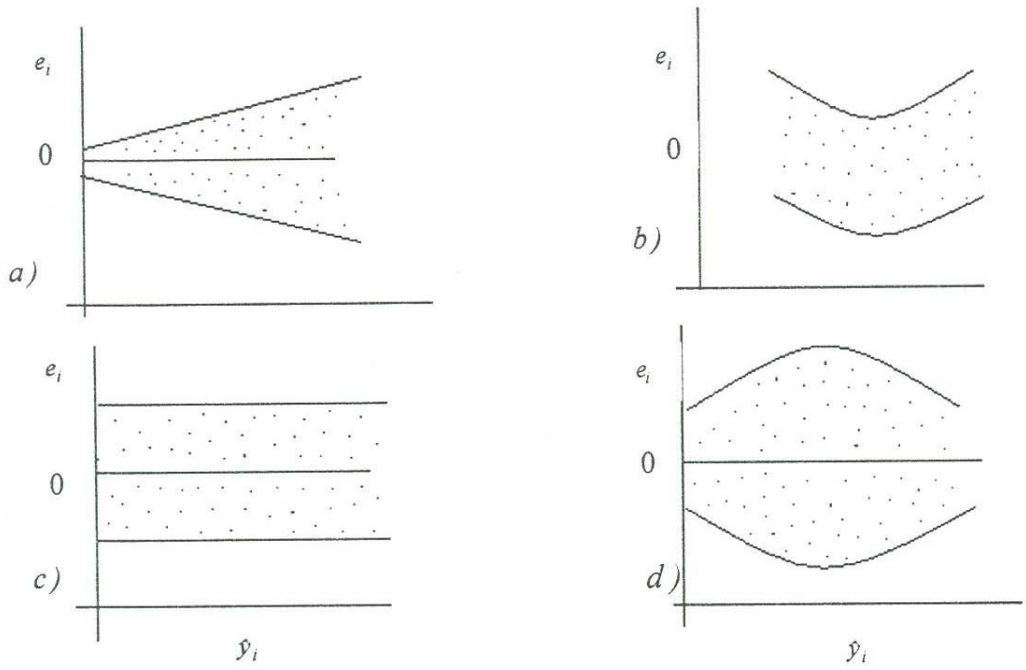
Puede considerarse al residual como la desviación entre los datos y el ajuste; también mide la variabilidad de la variable respuesta, que no es explicada por el

análisis de regresión. El análisis gráfico de residuales permite comparar si las suposiciones del modelo de regresión se cumplen.

En este análisis podemos detectar:

a. La normalidad de los errores.
b. Valores anormales en la distribución de errores.
c. Varianza constante
d. Independencia de los errores.

En las gráficas 3.3 se muestran diversos comportamientos de los residuales



Gráfica 3.3

La gráfica 3.3 *a)* en forma de embudo abierto, nos indica que la varianza de los errores es una función creciente de la respuesta. La gráfica 3.3 *b)* indica no linealidad, esto es, se puede necesitar algún término cuadrático en el modelo. La gráfica 3.3 *c)* muestra los residuales encerrados en una banda horizontal y con un comportamiento aleatorio, lo que significa que no hay defectos en el modelo. Por último la gráfica 3.3 *d)* es un tipo de gráfica que se presenta generalmente cuando la respuesta es una proporción entre cero y uno.

Hay varias propiedades importantes en los residuales: su media es cero y su varianza promedio aproximada se estima con:

$$CM_E = \frac{\sum_{i=1}^{n}(e_i - \bar{e}_i)^2}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1} = \frac{SC_E}{n-k-1} \tag{3.20}$$

donde k es el número de parámetros estimados. Los residuales no son independientes pues tienen $n-k-1$ grados de libertad asociados a ellos, o sea dependen del número de parámetros a estimar. Esto no tiene importancia cuando el número de observaciones es grande con respecto a los parámetros a estimar.

Se acostumbra usar generalmente el *residual estandarizado*, el cual se obtiene al dividir el residual entre la desviación estándar de éstos:

$$e_i(est) = \frac{e_i}{\sqrt{CME}} \qquad i = 1,2,\ldots,n \tag{3.21}$$

Estos residuales tienen media cero y varianza aproximadamente unitaria, por lo tanto, un residual estandarizado mayor que tres, denota un valor atípico potencial.

Los residuales pueden presentar la característica de estar autocorrelacionados. Una prueba muy usual para detectar cierto tipo de correlación es la llamada *Prueba de Durbin-Watson*, que mide el grado de autocorrelación entre el residuo correspondiente a cada observación y el anterior [13].

Este estadístico d se calcula con la siguiente fórmula:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{3.22}$$

donde las e_t, $t = 1,\ldots,n$ son los residuales calculados por mínimos cuadrados. Durbin y Watson [13] demuestran que si d se encuentra entre dos cotas, digamos d_L y d_U, no se puede llegar a una conclusión determinante, si es menor que d_L se rechazará la hipótesis nula y si es mayor que d_U, no hay evidencia para rechazar ésta; donde las hipótesis están dadas por:

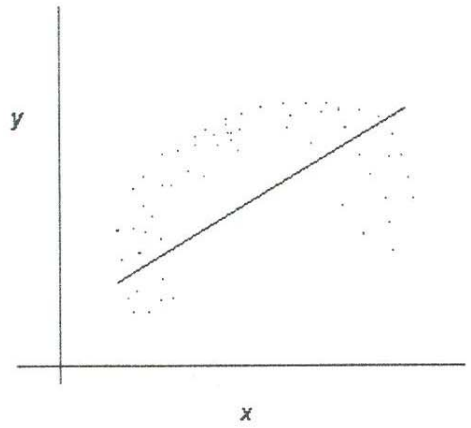
$$H_0 : \rho_r = 0$$
$$H_1 : \rho_r \neq 0$$

Estos valores de d_L y d_U se muestran en la tabla A.4 del apéndice A, y dependen del número de observaciones n y cuántas variables independientes hay en el modelo. En la tabla incluida se denota por k al número de variables predictoras.

3.4.3 Prueba de falta de ajuste

En los modelos de regresión lineal se desea saber si los datos se ajustan al modelo propuesto, pues si esto pasa, entonces el modelo se podría considerar correcto. Para comprobar si el modelo es adecuado se puede efectuar una prueba de hipótesis conocida como *prueba de falta de ajuste.*

En la gráfica 3.4 se muestra una nube de puntos con falta de ajuste lineal.



Gráfica 3.4

Para esta prueba podemos considerar que se tienen n_i observaciones de la respuesta al i-ésimo nivel de la variable independiente. Sea Y_{ij} la j-ésima observación de la respuesta en X_i, donde $i = 1,2,...,m$ y $j = 1,2,...,n_i$. En total se tendrán $n = \sum_{i=1}^{m} n_i$ observaciones.

Para la aplicación de esta prueba se requiere que los supuestos de normalidad, independencia y homogeneidad de varianza, sobre la variable error, se cumplan, y además deberán existir varias observaciones de la variable respuesta Y, para al menos uno de los niveles de X, esto con el fin de calcular SC_{Ep}, que se conoce como *suma de cuadrados del error puro*. Esta prueba consiste en descomponer:

$$y_{ij} - \hat{y}_{ij} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

donde $\bar{y}_i$ es el promedio de las n_i observaciones en X_i. Elevando al cuadrado en ambos lados de la ecuación y sumando sobre la i y las j se obtendrá:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2 \qquad (3.23)$$

esto es

$$SC_E = SC_{Ep} + SC_{FA}$$

donde SC_E es la suma de cuadrados del error o residual, y SC_{FA} se conoce como suma de cuadrados debido a la falta de ajuste.

Y

$$\bar{y}_i = media\ de\ las\ respuestas\ en\ el\ valor\ x_i\ de\ X$$
$$y_{ij} = respuesta\ observada$$
$$\hat{y}_i = valor\ ajustado\ por\ el\ \text{modelo}$$

Entonces la suma de cuadrados de la falta de ajuste es:

$$SC_{FA} = SC_E - SC_{Ep},$$

donde hay $n_e = \sum_{i=1}^{m}(n_i - 1) = n - m$ grados de libertad asociados con la suma de cuadrados del error puro y $(n-2)$ grados de libertad para la suma de cuadrados del error o residual, por lo tanto $(n-2)-(n-m) = m-2$ son los grados de libertad de la suma de cuadrados de la falta de ajuste.

Las hipótesis a plantear son las siguientes:

$$H_0: El\ modelo\ de\ regresión\ se\ ajusta\ adecuadamente\ a\ los\ datos$$
$$H_1: El\ modelo\ de\ regresión\ no\ se\ ajusta\ adecuadamente\ a\ los\ datos$$

ó sus hipótesis equivalentes:

$$H_0: No\ hay\ falta\ de\ ajuste$$
$$H_1: Hay\ falta\ de\ ajuste$$

En el caso, por ejemplo, de regresión lineal simple, hipótesis equivalentes serían:

$$H_0 : E(Y_i) = \beta_0 + \beta_1 X_i$$
$$H_1 : E(Y_i) \neq \beta_0 + \beta_1 X_i$$

El estadístico de prueba para verificar la falta de ajuste es:

$$F_0 = \frac{SC_{FA}/(m-2)}{SC_{Ep}/(n-m)} = \frac{CM_{FA}}{CM_{Ep}} \qquad (3.24)$$

con el cual se rechazará la hipótesis H_0 de adecuación del modelo, si $F_0 > F_{\alpha,m-2,n-m}$. Esto es, si F_0 es significativa, indica que el modelo aparentemente

es inadecuado y se tendrá que buscar otro que resulte más apropiado. Por otra parte si F_0 no es significativa, aparentemente no existe razón para dudar de la adecuación del modelo. Los cuadrados medios de la falta de ajuste y del error puro pueden tomarse como estimados de σ^2. Esta prueba de falta de ajuste suele utilizarse en diseño experimental cuando los niveles del factor de estudio son cuantitativos.

3.5 Ejemplo

En el siguiente ejemplo, se toman las mediciones de retención de tocoferol en el aceite de soya como respuesta y se considera la temperatura como variable independiente. Los valores para esta última son solamente cinco, que corresponderán, como se verá después, a los valores establecidos por el diseño seleccionado en la aplicación de la metodología de superficie de respuesta. La tabla 3.2 muestra, además de esta variable, las mediciones de retención de tocoferol calculadas en el experimento, así como los residuales y valores estimados al ajustar el modelo.

Orden Corridas	*Temperatura* (°C)	*Retención de Tocoferol* (%)	*Residual*	*Valor ajustado*
19	90	86.95	-2.53068	89.48068
8	130	82.45	3.252676	79.19732
20	90	93.01	3.529324	89.48068
12	130	87.94	8.742676	79.19732
3	90	85.05	-4.43068	89.48068
6	130	75.6	-3.59732	79.19732
1	90	89.41	-0.07068	89.48068
10	130	66.65	-12.5473	79.19732
9	77	83.4	-9.58623	92.98623
2	143	66.49	-9.20177	75.69177
4	110	86.84	2.501	84.339
5	110	85.83	1.491	84.339
16	110	83.25	-1.089	84.339
14	110	83.52	-0.819	84.339
15	110	90.31	5.971	84.339
7	110	87.87	3.531	84.339
17	110	87.41	3.071	84.339
11	110	88.33	3.991	84.339
13	110	88.16	3.821	84.339
18	110	88.31	3.971	84.339

Tabla 3.2

Las hipótesis a probar son:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Es decir, se quiere probar si β_1 tiene efecto sobre la respuesta, para lo cual se utilizan las ecuaciones (3.9) a la (3.14).

Y se obtiene:

$$SC_T = SC_R + SC_E = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 947.68$$

$$SC_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy} = 361.04$$

$$SC_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 586.638$$

$$CM_R = SC_R = 361.044, \qquad CM_E = SC_E/(n-2) = 586.638/18 = 32.59$$

ANDEVA PARA REGRESIÓN

Fuente de Variación	*Suma de Cuadrados*	*Grados de Libertad*	*Media de Cuadrados*	F_0
Regresión	361.04	1	361.044	11.07
Error o residual	586.638	18	32.59	
Total	947.68	19		

Tabla 3.3

y $P(F_{18}^1 > 11.08) = .0037$ por lo tanto el modelo es significativo, o sea la temperatura tiene una influencia significativa sobre la retención de tocoferol.

ANÁLISIS DE REGRESIÓN PARA EL MODELO $Y = \beta_0 + \beta_1 X_1$

Parámetro	Estimación	Error estándar	Estadístico t	Valor p
β_0	112.61	8.59	13.11	<.0001
β_1	-0.26	0.077	-3.33	0.0037

Tabla 3.4

Efectuando las pruebas hipótesis para los parámetros del modelo, usando los estadísticos mostrados en las ecuaciones (3.13) y (3.14), se tiene que:

$H_0 : \beta_0 = 0$
$H_1 : \beta_0 \neq 0$

H_0 se rechaza para el nivel de significancía $\alpha = 0.01$

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

H_0 se rechaza para el nivel de significancía $\alpha = 0.01$

Resultados que también se muestran en la tabla 3.4. Así tenemos que el modelo propuesto será: $\hat{Y} = 112.61 - 0.26X_1$

Por otra parte tenemos que:

$$R^2 = \frac{361.04}{947.68} = 0.381$$

y $r = -0.6173$ pues la pendiente de la recta es negativa.

Al efectuar una prueba para el coeficiente de correlación lineal:

$$H_o : \rho = 0$$
$$H_1 : \rho < 0$$

Aplicando el estadístico de prueba de la ecuación (3.18), este toma el valor de:

$$T_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.6173\sqrt{18}}{\sqrt{1-(0.6173)^2}} = -3.33$$

Comparando este valor con $-t_{0.99,18} = -2.552$, se rechaza la hipótesis nula para $\alpha = 0.01$ y se concluye que la correlación es significativa.

Para la prueba de falta de ajuste calculamos,

$$SC_{Ep} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 = 332.79$$

$$SC_{FA} = 586.54 - 332.79 = 253.84$$

las hipótesis a plantear son:

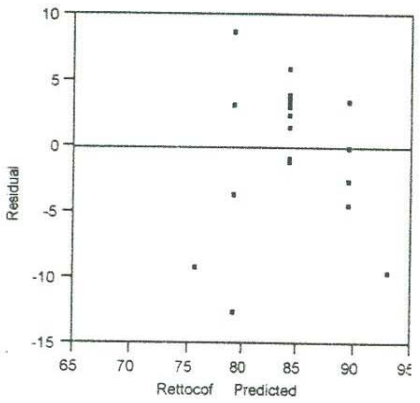
H_0: *El modelo de regresión se ajusta adecuadamente a los datos*
H_1: *El modelo de regresión no se ajusta adecuadamente a los datos*

y el estadístico de prueba que utilizaremos es:

$$F_0 = \frac{SC_{FA}/(m-2)}{SC_{Ep}/(n-m)} = \frac{253.84/3}{332.79/15} = \frac{CM_{FA}}{CM_{Ep}} = \frac{84.61}{22.19} = 3.81$$

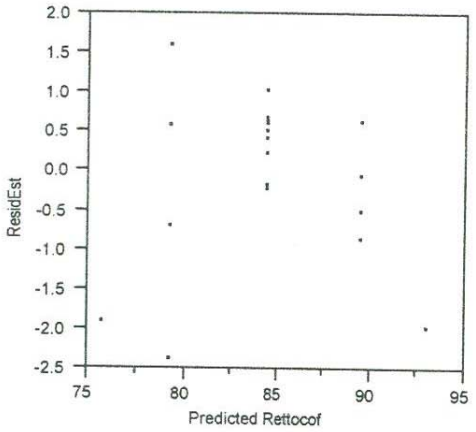
Como $P(F_{15}^{3} > 3.81) = .0326$, para $\alpha = 0.01$ no se rechaza la hipótesis nula de que el modelo de regresión se ajusta adecuadamente a los datos.

La gráfica de residuales, obtenida en JMP IN, es la gráfica 3.5:



Gráfica 3.5

y la gráfica de residuales estandarizados es la gráfica 3.6:



Gráfica 3.6

Se puede observar que ninguno de los residuales estandarizados está fuera del intervalo de entre menos tres y tres desviaciones estándar.

Por otra parte, en la prueba para verificar si existe autocorrelación en los errores, se plantean las siguientes hipótesis:

$$H_0 : \rho_r = 0$$
$$H_1 : \rho_r \neq 0$$

y como ya se explicó el estadístico de prueba d mostrado en la ecuación (3.22), permitirá tomar una decisión acerca de este parámetro.

Para facilitar estos cálculos se muestra la tabla 3.5 obtenida con JMP IN:

e_t	e_t^2	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$
-0.07068	0.00499566		
-9.20177	84.6725711	9.13109	83.3768046
-4.43068	19.6309253	-4.77109	22.7632998
2.501	6.255001	-6.93168	48.0481876
1.491	2.223081	1.01	1.0201
-3.59732	12.9407112	5.08832	25.8910004
3.531	12.467961	-7.12832	50.812946
3.252676	10.5799012	0.278324	0.07746425
-9.58623	91.8958056	12.838906	164.837507
-12.5473	157.434737	2.96107	8.76793554
3.991	15.928081	-16.5383	273.515367
8.742676	76.4343836	-4.751676	22.5784248
3.821	14.600041	4.921676	24.2228946
-0.819	0.670761	4.64	21.5296
5.971	35.652841	-6.79	46.1041
-1.089	1.185921	7.06	49.8436
3.071	9.431041	-4.16	17.3056
3.971	15.768841	-0.9	0.81
-2.53068	6.40434126	6.50168	42.2718428
3.529324	12.4561279	-6.060004	36.7236485
	$\sum_{t=1}^{n} e_t^2 = 586.63807$		$\sum_{t=2}^{n} (e_t - e_{t-1})^2 = 940.50032$

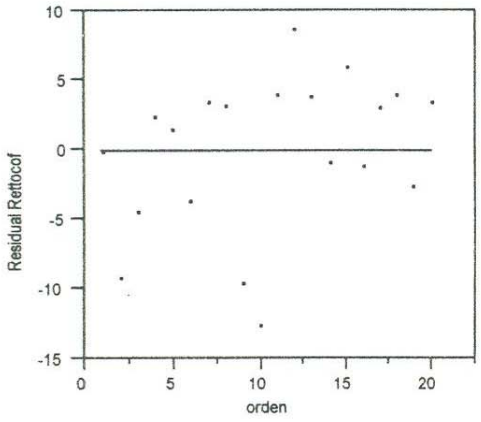
Tabla 3.5

De la cual podemos obtener:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{940.50032}{586.63807} = 1.6032037$$

Recordemos que si d se encuentra entre dos cotas, digamos d_L y d_U, no se puede llegar a una conclusión determinante, si d es menor que d_L se rechazará la hipótesis nula y si es mayor que d_U, no hay evidencia para rechazar ésta.
Para nuestro ejemplo estos valores de d_L y d_U para $n = 20$ observaciones, $k = 1$ variable predictora y $\alpha = 0.05$, son $d_L = 1.20$ y $d_U = 1.41$, entonces $d > d_U$ por lo que concluimos que no existe autocorrelación entre los residuales, pues no hay evidencia para rechazar la hipótesis nula.

La gráfica de residuales con respecto al orden de las corridas se muestra en la gráfica 3.7.



Gráfica 3.7

3.6 Regresión Lineal Múltiple

En muchos problemas de aplicación, la respuesta medida depende de más de una variable independiente, por lo que se necesitan modelos que permitan incluir dos o más de estas variables. Entre los modelos que permiten resolver este tipo de problemas se encuentra la regresión lineal múltiple, que se explica a continuación.

3.6.1 Modelo de Regresión Lineal Múltiple

Cuando se quiere relacionar la respuesta con más de una variable independiente o regresora, se tiene un modelo de regresión lineal múltiple, el cual podemos escribir como:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon \qquad (3.25)$$

El modelo de regresión lineal múltiple, en notación matricial se puede escribir como:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{3.26}$$

es decir:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

$$n \times 1 \qquad n \times (k+1) \qquad (k+1) \times 1 \qquad n \times 1$$

En este caso se necesita el vector $\boldsymbol{\beta}$ de estimadores de mínimos cuadrados que minimice

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta})$$

el cual, al desarrollar la ecuación anterior, podemos expresar como

$$S(\boldsymbol{\beta}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X\beta} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X\beta}$$

$$= \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X\beta}$$

Los estimadores de mínimos cuadrados se calcularán de la misma manera que se hizo para el caso de regresión lineal simple, o sea,

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

de donde se obtiene

$$\boldsymbol{X}^T \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{y}$$

que se conocen como ecuaciones normales de mínimos cuadrados. Multiplicando en ambos lados por la inversa de $\boldsymbol{X}^T \boldsymbol{X}$, se llega a que el estimador de $\boldsymbol{\beta}$ por el método de mínimos cuadrados es:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{3.27}$$

La inversa de la matriz $X^T X$ siempre existe si las variables regresoras son linealmente independientes, o sea, si ninguna columna de la matriz X puede expresarse como una combinación lineal no trivial de las otras columnas.

3.6.2 Propiedades de los estimadores de mínimos cuadrados

Para obtener las propiedades de $\hat{\beta}$, el estimador de mínimos cuadrados de β, recordemos que sobre los elementos del vector aleatorio ε se hacen tres supuestos:

- La esperanza es nula, $E(\varepsilon) = 0$.
- Tienen varianza común, $V(\varepsilon) = E(\varepsilon\varepsilon^T) = \sigma^2$.
- No hay relación entre ellos, esto es, $\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$

Tomando en cuenta esas propiedades se puede demostrar que el estimador $\hat{\beta}$ tiene la propiedad de ser insesgado, pues

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

debido a que $E(\varepsilon) = 0$ y $(X^T X)^{-1}(X^T X) = I$.
Por otra parte, la varianza de $\hat{\beta}$ puede ser calculada por medio de la matriz de covarianzas:

$$V(\hat{\beta}) = Cov(\hat{\beta}) \equiv E\left(\left(\hat{\beta} - E(\hat{\beta})\right)\left(\hat{\beta} - E(\hat{\beta})\right)^T\right)$$

que es una matriz simétrica cuyo i-ésimo elemento de la diagonal principal es la variancia de $\hat{\beta}_i$, y cuyos ij-ésimos elementos, son la covarianza entre las $\hat{\beta}_i$ y las $\hat{\beta}_j$.
La covarianza de la matriz de $\hat{\beta}$ es:

$$\begin{aligned} V(\hat{\beta}) = \text{cov}(\hat{\beta}) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T\right] = E(\hat{\beta}\hat{\beta}^T) - \beta\beta^T = E\left\{\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right]\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right]^T\right\} - \beta\beta^T \\ &= E\left\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \varepsilon)\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \varepsilon)\right]^T\right\} - \beta\beta^T \\ &= E\left\{\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\left[\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\right]^T\right\} - \beta\beta^T \\ &= \beta\beta^T + E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] - \beta\beta^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\varepsilon\varepsilon^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

entonces $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

3.6.3 Pruebas de Hipótesis en la Regresión Lineal Múltiple

Cuando se tienen calculados las estimaciones de los parámetros del modelo de regresión es importante saber si este modelo es adecuado y cuál o cuáles de las variables independientes son importantes o significativas. Para saber esto se efectúan pruebas de hipótesis para ver si hay una relación lineal entre la variable independiente y las variables regresoras, estableciendo en la hipótesis nula que los coeficientes de la ecuación de regresión son iguales a cero, frente a la hipótesis alternativa de que al menos uno de ellos es diferente de cero. Para esto se requieren la suposición de que los errores se distribuyen en forma normal, independientes, con media cero y varianza σ^2.

Las hipótesis que se plantean son:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \quad al \quad menos \quad para \quad una \quad j.$$

La prueba se efectúa generalizando el análisis de varianza que se utilizó en la regresión lineal simple y el rechazo de la hipótesis nula implica que al menos una de las variables independientes es significativa. La tabla de análisis de varianza que se utilizará se muestra a continuación:

ANÁLISIS DE VARIANZA EN REGRESIÓN LINEAL MÚLTIPLE

Fuente de Variación	*Suma de Cuadrados*	*Grados de libertad*	*Cuadrado Medio*	F_0	*Valor p*
Regresión	$SC_R = \hat{\beta}^T X^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	k	$CM_R = \frac{SC_R}{k}$	$F_0 = \frac{CM_R}{CM_E}$	$P(F > F_0)$
Error o residuo	$SC_E = y^T y - \hat{\beta}^T X^T y$	$n-k-1$	$CM_E = \frac{SC_E}{n-k-1}$		
Total	$SC_T = y^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	$n-1$			

Tabla 3.6

Se rechazará la hipótesis nula si $P(F > F_0) < \alpha$, el nivel de significancia asignado para la prueba.

En caso de ser así, es conveniente efectuar pruebas individuales sobre los coeficientes de la recta de regresión para probar su significancia.
En este caso las hipótesis a plantear son:

$$H_o : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Si no se rechaza H_o para la *j-ésima* variable independiente, entonces ésta se puede eliminar del modelo.

Es fácil demostrar, tomando en cuenta las propiedades del estimador $\hat{\beta}$, que el estadístico de prueba a utilizar es:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad (3.28)$$

donde $\hat{\sigma}^2 = CM_E$ y C_{jj} es el elemento de la diagonal de $(X^T X)^{-1}$, que corresponde a $\hat{\beta}_j$.
Y se rechazará la hipótesis nula cuando $|T_0| > T_{1-\alpha/2,n-k-1}$. Se resume lo anterior en la tabla 3.7.

ANÁLISIS DE REGRESIÓN PARA EL MODELO $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Parámetro	Estimación	Error estándar	Estadístico	Valor p
β_0	$\hat{\beta}_0$	$\sqrt{CM_E C_{11}}$	$\frac{\hat{\beta}_0}{\sqrt{CM_E C_{11}}}$	$P(T > \lvert t_0 \rvert)$
⋮	⋮	⋮	⋮	⋮
β_k	$\hat{\beta}_k$	$\sqrt{CM_E C_{k+1,k+1}}$	$\frac{\hat{\beta}_k}{\sqrt{CM_E C_{k+1,k+1}}}$	$P(T > \lvert t_0 \rvert)$

Tabla 3.7

3.6.3 Aplicación de Regresión Lineal Múltiple

A continuación se presentan las mismas mediciones de retención de tocoferol pero ahora tomando como variables independientes la temperatura y el tiempo, para ejemplificar la regresión lineal múltiple. Los valores asignados a esta nueva variable, se muestran en la tabla 3.8, así como los valores predichos por el modelo y el cálculo de sus residuales.

Orden Corridas	Temperatura (°C)	Tiempo (min.)	Retención (%) Tocoferol	Valor Ajustado	Residual
19	90	20	86.95	91.91066	-4.96066
8	130	20	82.45	81.62731	0.822693
20	90	20	93.01	91.91066	1.099342
12	130	20	87.94	81.62731	6.312693
3	90	60	85.05	87.05069	-2.00069
6	130	60	75.6	76.76734	-1.16734
1	90	60	89.41	87.05069	2.359307
10	130	60	66.65	76.76734	-10.1173
9	77	40	83.4	92.98623	-9.58623
2	143	40	66.49	75.69177	-9.20177
4	110	40	86.84	84.339	2.501
5	110	40	85.83	84.339	1.491
16	110	6.4	83.25	88.42573	-5.17573
14	110	74	83.52	80.25227	3.267727
15	110	40	90.31	84.339	5.971
7	110	40	87.87	84.339	3.531
17	110	40	87.41	84.339	3.071
11	110	40	88.33	84.339	3.991
13	110	40	88.16	84.339	3.821
18	110	40	88.31	84.339	3.971

Tabla 3.8

Utilizando la ecuación (3.26), para calcular los estimadores de los parámetros (3.27) ,se forman las matrices :

$$y = \begin{bmatrix} 86.95 \\ 82.45 \\ 93.01 \\ 87.94 \\ 85.05 \\ 75.6 \\ 89.41 \\ 66.65 \\ 83.4 \\ 66.49 \\ 86.84 \\ 85.83 \\ 83.25 \\ 83.52 \\ 90.31 \\ 87.87 \\ 87.41 \\ 88.33 \\ 88.16 \\ 88.31 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 90 & 20 \\ 1 & 130 & 20 \\ 1 & 90 & 20 \\ 1 & 130 & 20 \\ 1 & 90 & 60 \\ 1 & 130 & 60 \\ 1 & 90 & 60 \\ 1 & 130 & 60 \\ 1 & 77 & 40 \\ 1 & 143 & 40 \\ 1 & 110 & 40 \\ 1 & 110 & 40 \\ 1 & 110 & 6.4 \\ 1 & 110 & 74 \\ 1 & 110 & 40 \\ 1 & 110 & 40 \\ 1 & 110 & 40 \\ 1 & 110 & 40 \\ 1 & 110 & 40 \\ 1 & 110 & 40 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 90 & 130 & \cdots & 110 \\ 20 & 20 & \cdots & 40 \end{bmatrix} \begin{bmatrix} 1 & 90 & 20 \\ 1 & 130 & 20 \\ \vdots & \vdots & \vdots \\ 1 & 110 & 40 \end{bmatrix} = \begin{bmatrix} 20 & 2200 & 800.4 \\ 2200 & 247378 & 88044 \\ 800.4 & 88044 & 37517 \end{bmatrix}$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 2.5919 & -2.0454\times10^{-2} & -7.2963\times10^{-3} \\ -2.0454\times10^{-2} & 1.8594\times10^{-4} & 0 \\ -7.2963\times10^{-3} & 0 & 1.8232\times10^{-4} \end{bmatrix}$$

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} 2.5919 & -2.0454\times10^{-2} & -7.2963\times10^{-3} \\ -2.0454\times10^{-2} & 1.8594\times10^{-4} & 0 \\ -7.2963\times10^{-3} & 0 & 1.8232\times10^{-4} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 90 & 130 & \cdots & 110 \\ 20 & 20 & \cdots & 40 \end{bmatrix} \begin{bmatrix} 86.95 \\ 82.45 \\ \vdots \\ 88.31 \end{bmatrix}$$

$$= \begin{bmatrix} 117.69 \\ -0.26 \\ -0.12 \end{bmatrix}$$

Por lo tanto la ecuación encontrada por mínimos cuadrados es:

$$\hat{y} = 117.69 - 0.26 Temperatura - 0.12 Tiempo$$

Para obtener la tabla de análisis de varianza del modelo de regresión lineal múltiple, utilizamos las ecuaciones de la tabla 3.6:

$$SC_R = \hat{\beta}^T X^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 441.53512$$

$$SC_E = y^T y - \hat{\beta}^T X^T y = 506.14746$$

$$S_{yy} = y^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 947.68258$$

y así obtenemos

ANÁLISIS DE VARIANZA PARA EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Fuente de Variacion	*Suma de Cuadrados*	*Grados de libertad*	*Cuadrado Medio*	F_0	*Valor p*
Regresión	441.53512	2	220.768	7.4149	$P(F > F_0)$
Error o residuo	506.14746	17	29.773		<0.0048
Total	947.68258	19			

Tabla 3.9

Esta tabla de análisis de varianza nos permite probar:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_j \neq 0 \quad al \quad menos \quad para \quad una \quad j.$$

y vemos que $F_0 = \frac{CM_R}{CM_E} = \frac{220.768}{29.773} = 7.4149$, y como la probabilidad obtenida es pequeña, se concluye que la retención de tocoferoles se relaciona con la temperatura y/o el tiempo.

También podemos verificar la adecuación del modelo por medio del coeficiente de determinación, en este caso lo calculamos como:

$$R^2 = \frac{SC_R}{SC_T} = \frac{441.53512}{947.68258} = 0.46591$$

En ocasiones conviene trabajar con el *Coeficiente de Determinación Ajustado,* que se define como:

$$R^2_{Ajust} = 1 - \frac{SC_E/(n-k-1)}{SC_T/(n-1)}, \qquad (3.29)$$

puesto que el numerador es el cuadrado medio de los residuales y el denominador es una cantidad fija e independiente del número de variables que hay en el modelo, este coeficiente solamente aumentará cuando al agregar una nueva variable al modelo, ésta reduzca el cuadrado medio del residual.

En este caso

$$R^2_{Ajust} = 1 - \frac{29.773}{947.68258/19} = 0.403076.$$

Las pruebas sobre los coeficientes individuales de la regresión, como se explicó anteriormente, se plantean con las hipótesis

$$H_o : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

y el estadístico de prueba es el de la ecuación (3.28):

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}},$$

donde $\hat{\sigma}^2 = CM_E$ y C_{jj} es el elemento de la diagonal de $\left(X^T X\right)^{-1}$, que corresponde a $\hat{\beta}_j$.

Entonces, para probar las hipótesis:

$$H_o : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

se tiene que $$T_0 = \frac{-0.26}{\sqrt{(29.773)(1.8594 \times 10^{-4})}} = -3.49,$$

y se rechaza la hipótesis nula, pues la probabilidad asociada a este valor es menor de 0.003, y así concluimos que el efecto de la temperatura es significativo.

Para el caso de probar las hipótesis

$$H_o : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

el valor del estadístico es

$$T_0 = \frac{-0.12}{\sqrt{(29.773)(1.8232 \times 10^{-4})}} = -1.63.$$

Para un valor de $\alpha = 0.05$, se concluye que el efecto del tiempo sobre la retención de tocoferol no resulta significativo.

Se puede también efectuar esta prueba para el caso de β_0; aunque no se ilustra su cálculo en esta tesis, resulta significativa para $\alpha = 0.01$.

En cuanto a la prueba de falta de ajuste, tenemos que calcular la suma de cuadrados del error puro; de la ecuación (3.23), obtenemos

$$SC_{Ep} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 = 94.86365$$

y así

$$SC_{FA} = 506.14746 - 94.86365 = 411.28381$$

Como ya se vio anteriormente las hipótesis a plantear son:

H_0: *El modelo de regresión se ajusta adecuadamente a los datos*

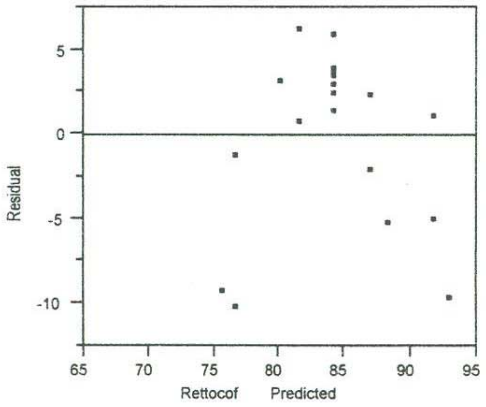
H_1: *El modelo de regresión no se ajusta adecuadamente a los datos*

el estadístico de prueba es:

$$F_0 = \frac{SC_{FA}/(m-3)}{SC_{Ep}/(n-m)} = \frac{411.28381/6}{94.86365/11} = \frac{CM_{FA}}{CM_{Ep}} = \frac{68.5473}{8.6240} = 7.9485$$

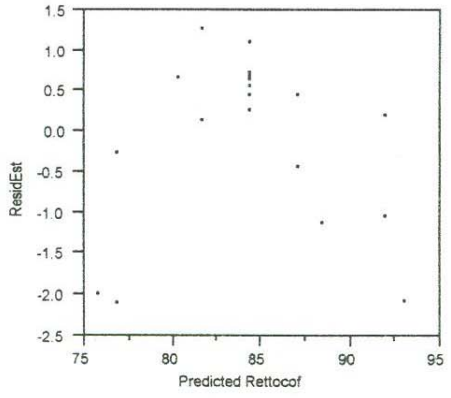
y $P(F_{15}^3 > 7.9485) = .0017$, por lo que al nivel de significancia de 0.01, se rechaza la hipótesis nula de que el modelo de regresión se ajusta adecuadamente a los datos.

La gráfica de residuales, obtenida en JMP IN, es la gráfica 3.8



Gráfica 3.8

y la gráfica de residuales estandarizados es la gráfica 3.9



Gráfica 3.9

Al igual que cuando se consideraba una variable independiente, se observa que ninguno de los residuales estandarizados está fuera de tres desviaciones estándar.La prueba para verificar la autocorrelación de los errores se efectúa planteando las siguientes hipótesis:

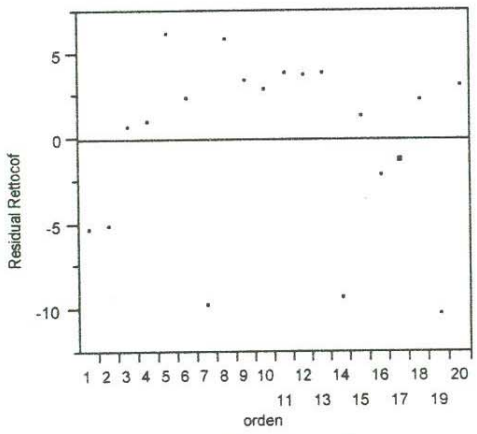
$$H_0 : \rho_r = 0$$

$$H_1 : \rho_r \neq 0$$

y como ya se calculó d en el ejemplo anterior, para este caso solamente diremos que

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = 1.7386.$$

Consultando la tabla de Durbin-Watson, los valores de d_L y d_U para $n = 20$ observaciones, $k = 2$ variables predictoras y un nivel de significancia de 0.01, son $d_L = 0.98$ y $d_U = 1.30$.



Gráfica 3.10

Entonces, como $d > d_U$ se puede decir que los residuales no están autocorrelacionados, pues no hay evidencia para rechazar el H_0.

CAPÍTULO 4

METODOLOGÍA DE SUPERFICIES DE RESPUESTA

4.1 La metodología de Superficies de Respuesta

El objetivo de la metodología de superficies de respuesta es optimizar una o más variables de interés, lo cual se logra al determinar sus mejores condiciones de operabilidad. Para ello se utiliza un conjunto de técnicas estadísticas que nos permiten analizar y modelar la forma en que la variable de interés es influenciada por otras. Se pueden distinguir tres aspectos claves en esta metodología: diseño, modelo y técnicas de optimización.

Se necesita, por lo tanto, conocer de diseño de experimentos para poder elegir el diseño más apropiado. Entre los diseños base más utilizados se encuentran los diseños factoriales completos o fraccionarios, estudiados en el capítulo dos. Para modelar una superficie de respuesta se necesitan también conocimientos de regresión lineal múltiple, como son los cubiertos en el capítulo tres y en cuanto a la optimización, son necesarios ciertos conocimientos de álgebra lineal y de cálculo. Para esto último no se incluye un capítulo en especial, sino que, en su momento, se detallan algunas de las herramientas utilizadas en el transcurso de este capítulo y otras se anexan en el apéndice C.

4.2 Diseños de Superficies de Respuesta

El grado de los modelos polinomiales usados generalmente en el análisis de superficies de respuesta permite la clasificación de los diseños; los más comúnmente usados son el modelo lineal o de primer orden y el modelo cuadrático o de segundo orden.

El modelo de primer orden, sin considerar interacciones, se puede escribir de la siguiente manera

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon \tag{4.1}$$

En caso de detectar curvatura en el sistema se empleará un modelo de segundo orden, de la forma

$$Y = \beta_o + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=2}^{k} \sum_{i=1}^{j-1} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_{ii}^2 + \varepsilon \tag{4.2}$$

Para estimar los parámetros de este modelo se puede utilizar el método de mínimos cuadrados. Si el modelo ajustado describe adecuadamente la función de respuesta, entonces el analizar este modelo es casi equivalente a analizar el sistema real. La metodología de superficies de respuesta es un procedimiento secuencial que se puede ilustrar con el diagrama de la figura 4.1.

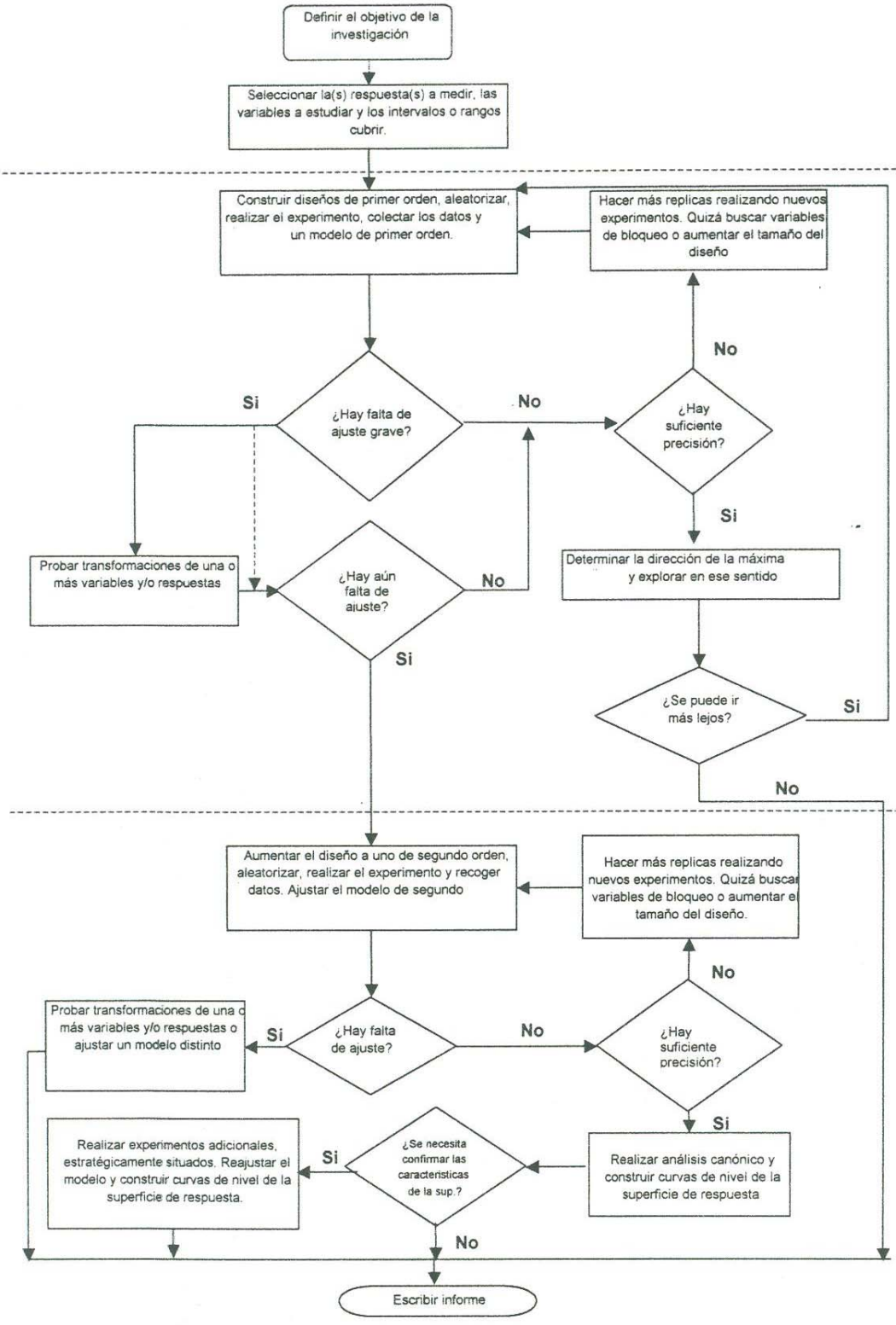

Definir el objetivo de la investigación
Seleccionar la(s) respuesta(s) a medir, las variables a estudiar y los intervalos o rangos cubrir.
Construir diseños de primer orden, aleatorizar, realizar el experimento, colectar los datos y un modelo de primer orden.
Hacer más replicas realizando nuevos experimentos. Quizá buscar variables de bloqueo o aumentar el tamaño del diseño
No
Si
¿Hay falta de ajuste grave?
No
¿Hay suficiente precisión?
Si
Probar transformaciones de una o más variables y/o respuestas
¿Hay aún falta de ajuste?
No
Si
Determinar la dirección de la máxima y explorar en ese sentido
¿Se puede ir más lejos?
Si
No
Aumentar el diseño a uno de segundo orden, aleatorizar, realizar el experimento y recoger datos. Ajustar el modelo de segundo
Hacer más replicas realizando nuevos experimentos. Quizá buscar variables de bloqueo o aumentar el tamaño del diseño.
No
Probar transformaciones de una o más variables y/o respuestas o ajustar un modelo distinto
Si
¿Hay falta de ajuste?
No
¿Hay suficiente precisión?
Si
Realizar experimentos adicionales, estratégicamente situados. Reajustar el modelo y construir curvas de nivel de la superficie de respuesta.
Si
¿Se necesita confirmar las características de la sup.?
Realizar análisis canónico y construir curvas de nivel de la superficie de respuesta
No
Escribir informe

Figura 4.1

4.3 Método de Escalamiento Ascendente

Este método es un procedimiento preliminar que se utiliza cuando el sistema está siendo investigado por primera vez y sirve para determinar la dirección de ascenso (o descenso) máximo a partir del centro del diseño inicial. Generalmente le sigue un ajuste de segundo orden, es decir, es la base de una experimentación secuencial.

En las etapas iniciales de un experimento es posible mejorar bastante la respuesta. Se debe seleccionar una región de operabilidad, que generalmente es una subregión de toda la región de interés, y que muchas veces va cambiando conforme la investigación progresa. En esta región seleccionada se corre un diseño de primer orden para explorar la región experimental determinada antes, ajustar este modelo y si existe falta de ajuste investigar las causas posibles. Para seguir con el proceso, una vez obtenido el modelo depurado y ajustado, se procede a determinar la dirección óptima de movimiento, a partir del centro del diseño. Se debe decidir una longitud de paso en unidades codificadas y reales. Generalmente se recomienda un paso unitario codificado en el factor con mayor influencia, aunque en la práctica el tamaño real del paso lo determina el experimentador con base en el conocimiento del proceso ([26], pp. 430-431). En la práctica se ha visto que por lo general se utilizan un mínimo de cuatro o cinco puntos en la dirección óptima. Se mide la respuesta obtenida en esos puntos hasta detectar un cambio en la tendencia ascendente dada por el plano. Parar detectar fácilmente este cambio se recomienda graficar un número suficiente de puntos o pasos contra el valor de la respuesta; cuando en esta gráfica se detecta un cambio en el comportamiento de la respuesta, entonces ya no se debe de seguir esa dirección de búsqueda y se recomienda determinar el centro de la nueva región experimental, que por lo general es el último punto de la dirección óptima con la que se mantuvo la tendencia ascendente (descendente) y así volver a iniciar el proceso hasta que la aproximación de primer orden sea inadecuada.

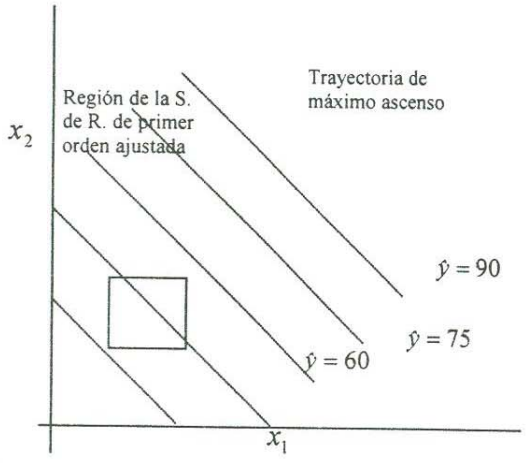


Figura 4.2

Para comprender cómo se sigue la ruta de la máxima pendiente en ascenso, necesitamos algunas herramientas del cálculo diferencial. Recordemos que si $f: U \subset R^3 \to R$ es diferenciable, el gradiente de f en (x, y, z) es el vector en el espacio dado por:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$$

el cual también se denota por $\nabla f(x, y, z)$ y donde U es un abierto en R^3. Asimismo, la derivada direccional de f en x, en la dirección del vector v, está dada por

$$Df(x)v = \frac{d}{dt} f(x + tv)\Big|_{t=0},$$

si es que existe.

Con frecuencia en esta definición se escoge a $\mathbf{v}$ como un vector unitario. En este caso nos movemos en la dirección de $\mathbf{v}$ con rapidez unitaria.

También es necesario recordar los siguientes teoremas cuyas demostraciones se anexan en el apéndice C:

Teorema 4.1

Si $f: R^3 \to R$ es diferenciable, entonces todas las derivadas direccionales existen y la derivada direccional en $\mathbf{x}$ en la dirección de $\mathbf{v}$ está dada por

$$Df(x)v = grad\, f(x) \bullet v = \nabla f(x) \bullet v = \left[\frac{\partial f}{\partial x}(x) \right] v_1 + \left[\frac{\partial f}{\partial y}(x) \right] v_2 + \left[\frac{\partial f}{\partial z}(x) \right] v_3$$

donde $\mathbf{v} = (v_1, v_2, v_3)$.

Teorema 4.2

Supongamos que $\nabla f(x) \neq 0$. Entonces $\nabla f(\mathbf{x})$ apunta en la dirección a lo largo de la cual f crece más rápido.

Utilizando lo anterior, para maximizar la respuesta, el movimiento desde el centro del diseño debe ser en la dirección del gradiente de la función de respuesta, esto es, en la dirección

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_k} \right).$$

Partiendo de que el modelo ajustado es un plano, la dirección óptima de movimiento es perpendicular a los contornos o curvas de nivel. Se recomienda, como ya se dijo, un paso de movimiento unitario en el factor de mayor influencia, con lo cual se asegura que los pasos en los otros factores serán de menor amplitud y proporcionales a sus coeficientes. Esta longitud de paso representa la mitad del rango experimental utilizado (es decir, nivel alto menos nivel bajo, entre dos).

Así, si el modelo es de la forma

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

y el factor con mayor influencia es x_2, se tomará en unidades codificadas el incremento $\Delta x_2 = 1$, y para el factor x_1 se tomará en unidades codificadas $\Delta x_1 = \frac{\hat{\beta}_1}{\hat{\beta}_2} \Delta x_2$. Similarmente se hará si se tienen más factores. Se generan entonces puntos en la dirección óptima, a partir del centro del diseño, hasta detectar un cambio en la variable respuesta. Cuando esto ocurre se tiene el centro de una nueva región de experimentación.

4.4 Diseños para estimar superficies de respuesta de segundo orden

Los diseños de segundo orden son aquellos que permiten estudiar los efectos de interacción y efectos cuadráticos, aparte de los efectos lineales. Se utilizan ante la necesidad de explorar una superficie más compleja o bien cuando se tiene identificada la región de respuesta óptima y se quiere caracterizar esa superficie de respuesta.

La selección de estos diseños depende de las características del problema, pero deben en general cumplir ciertos requerimientos como capacidad para realizar estimaciones eficientes de los coeficientes del modelo y medir tanto el error experimental como la posible presencia de falta de ajuste.

Un modelo de segundo orden podemos representarlo como (4.2):

$$Y = \beta_o + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=2}^{k} \sum_{i=1}^{j-1} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_{ii}^2 + \varepsilon$$

el cual tiene $p = (k+1)(k+2)/2$ términos, por lo que se requiere al menos ese número de puntos en el diseño. Estos diseños deben tener por lo menos tres niveles en cada uno de los factores para poder estimar la curvatura de la superficie de respuesta en cada uno de los factores. Existen algunas otras propiedades que en ocasiones se pueden conseguir y que se explican en la siguiente sección.

4.4.1 Diseño Central Compuesto

Cuando se busca un modelo de segundo orden que se ajuste a los datos, el diseño central compuesto, también llamado Box-Wilson es uno de los diseños más utilizados, por su gran flexibilidad. Se empieza con un diseño factorial o factorial fraccionario a dos niveles (conocido como la *porción factorial*), n_0 puntos centrales (que sirven para examinar la presencia de curvatura, dar información acerca de los efectos cuadráticos y estimar la magnitud del error experimental) y $2k$ *puntos*

axiales o puntos estrella, de la forma: ($-\alpha$, 0, ... , 0), (α, 0, ... , 0), (0, $-\alpha$, ... ,0), (0, α, ... ,0), ... (0, 0, ... , $-\alpha$), (0, 0, ... , α). Se tendrá por lo tanto, en total, $N = 2^k + 2k + n_0$ corridas.

La distancia de los puntos axiales al origen, que se denota por α, varía según las propiedades que se desean en el diseño. Las propiedades más buscadas, generalmente, son rotabilidad, ortogonalidad y precisión uniforme. A continuación se explica en qué consiste cada una de estas propiedades, sin demostración alguna, ya que ello se considera fuera del alcance de esta tesis y solamente se citan las referencias en cada caso.

Rotabilidad. Box y Hunter [6], establecen que un diseño experimental es rotable o girable si la varianza de la respuesta predicha $\hat{y}$, en algún punto x, es función sólo de la distancia al punto desde el centro del diseño y no es una función de la dirección. Un diseño central compuesto se hace rotable mediante una elección apropiada del espaciamiento axial α. Si el diseño es rotable, la desviación estándar (o la varianza) de la respuesta predicha $\hat{y}$ es constante en todos los puntos que están a la misma distancia del centro del diseño. Para lograr rotabilidad en un diseño factorial completo a dos niveles, se debe elegir un valor de $\alpha = (2^k)^{1/4}$. Si consideramos, por ejemplo, que la respuesta a medir es el rendimiento, la predicción de éste tendrá la misma precisión para todos los puntos que se encuentran a la misma distancia del centro del diseño. Esta propiedad de rotabilidad cobra importancia en la exploración de una superficie de respuesta, porque la precisión de la superficie estimada no depende de la orientación del diseño con respecto a la superficie de respuesta real o a la dirección de búsqueda de las condiciones óptimas. Dado que la finalidad del diseño de superficies de respuesta es la optimización y se desconoce la localización del óptimo antes de correr el experimento, tiene sentido usar un diseño que proporcione estimaciones igualmente precisas en todas direcciones. Cuando se tienen r_f réplicas del factorial 2^k y r_a réplicas de los puntos axiales, se debe escoger la distancia α como $\alpha = \left(r_f 2^k / r_a\right)^{1/4}$. Por otra parte, cuando se usa un factorial fraccionario 2^{k-p}, como base del diseño central compuesto, se escogerá $\alpha = \left(r_f 2^{k-p} / r_a\right)^{1/4}$.

Ortogonalidad. Otra propiedad, que también es de importancia considerar en los diseños, es la ortogonalidad. Para un modelo de primer orden, esta es la propiedad de diseño óptima, ya que minimiza la varianza de los coeficientes de regresión. Se considera que un diseño es ortogonal cuando los coeficientes estimados en el modelo ajustado, no están correlacionadas entre si. Una propiedad de los diseños centrales compuestos es que mediante la elección apropiada de los puntos centrales el diseño puede hacerse ortogonal, esto es, que las estimaciones de los parámetros para el modelo de segundo orden, están mínimamente correlacionados con las estimaciones de otros parámetros. Box y

Wilson [7] han demostrado que para dar la propiedad de ortogonalidad al diseño se elige el valor de α como:

$$\alpha = \left(\frac{\left(2^k \times N\right)^{1/2} - 2^k}{2} \right)^{1/2}$$

donde N representa el total de corridas en el diseño. Así, si se tiene un valor específico de α que haga el diseño rotable, se puede determinar el número de puntos centrales que lo harán rotable y ortogonal.

En los diseños en los que se busque rotabilidad y ortogonalidad, al mismo tiempo, se debe cumplir con valores específicos para α y n_0. Los valores que éstos deben tomar son

$$\alpha = \left(2^k\right)^{1/4} \qquad \text{y} \qquad n_0 \approx 4\left(2^k\right)^{1/2} + 4 - 2^k$$

Precisión Uniforme. Se dice que un diseño es de precisión uniforme si la varianza de la respúesta predicha $V[\hat{y}(x)]$ en el centro del diseño (radio $r = 0$), es la misma que en la esfera de radio $r = 1$, esto es, se tiene una varianza constante dentro de la esfera unitaria. Box y Hunter [6] demuestran que para conseguir un diseño rotable con precisión uniforme o casi uniforme se debe seleccionar

$$\alpha = \left(2^k\right)^{1/4} \qquad n_0 \approx \lambda_4\left(2^k + 2\right) + 2^k - 2k$$

donde λ_4 es una constante que depende de k, el número de factores a utilizar. Algunos de sus valores se presentan en la siguiente tabla

k	2	3	4	5	6	7	8
λ_4	0.7844	0.8385	0.8704	0.8918	0.9070	0.9184	0.9274

Tabla 4.1

Entonces, utilizando λ_4, se puede fijar el número de puntos centrales de un diseño.

4.4.2 Diseño Central Compuesto con dos y tres factores

Consideremos un factorial completo 2^2 como la base de un diseño central compuesto con dos factores. Si la distancia desde el centro del diseño a un punto factorial es ± 1 unidad por cada factor, la distancia del centro del diseño a un punto estrella es $\pm\alpha$ con $\alpha > 1$; donde el valor de α depende de algunos aspectos, como se vio anteriormente. En la gráfica 4.3 se ilustra este diseño para el caso de $\alpha = \left(2^k\right)^{1/4} = \left(2^2\right)^{1/4} = 1.4142$.

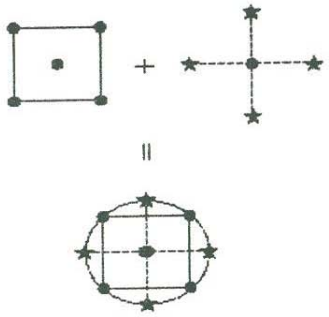


Figura 4.3

Así pues, para un diseño central compuesto, en el que se tienen dos factores, a dos niveles cada uno de ellos, esto es un 2^2, se puede resumir lo anterior en la figura 4.4

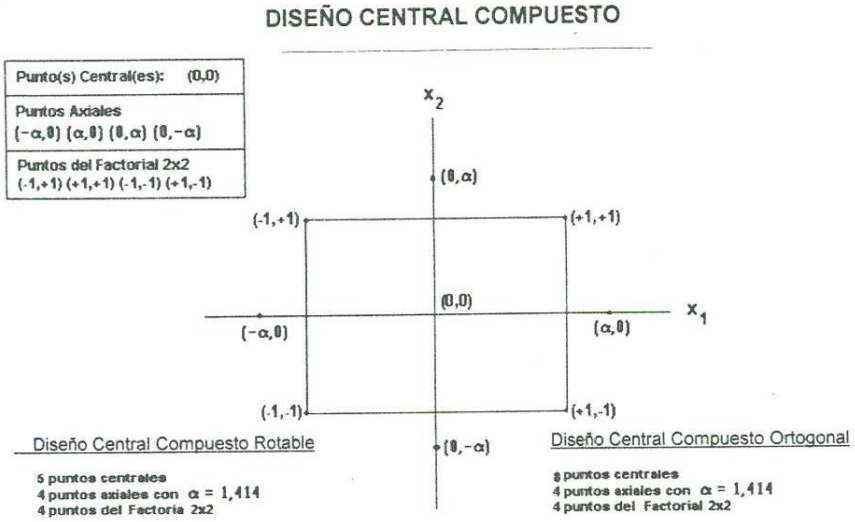


Figura 4.4

El caso de un diseño con tres factores está ilustrado en la figura 4.5

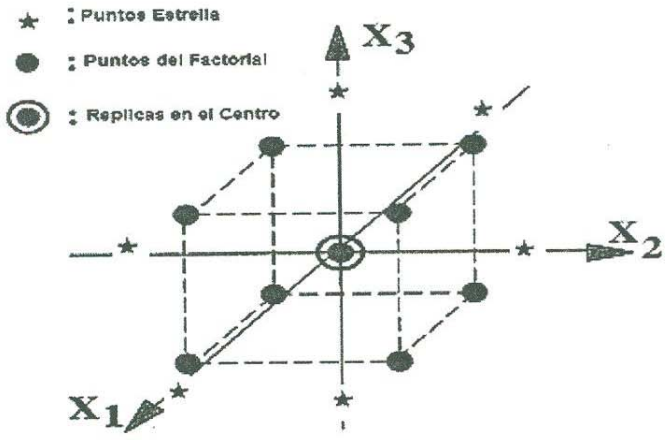


Figura 4.5

	Diseño Central Compuesto Rotable	Diseño Central Compuesto Ortogonal
Puntos Centrales (0,0,0)	6 puntos centrales	9 puntos centrales
Puntos axiales $(0,0,\alpha)$, $(0,0,-\alpha)$, $(\alpha,0,0)$, $(\alpha,0,0)$, $(0,\alpha,0)$, $(0,-\alpha,0)$	6 puntos axiales con $\alpha = 1.682$	6 puntos axiales con $\alpha = 1.682$
Puntos del Factorial 2x2x2 (1,1,1),(1,-1,1),(-1,-1,1), (-1,1,1), (1,-1,-1),(1,1,-1), (-1,1,-1),(-1,-1,-1)	8 puntos del Factorial $2x2x2$	8 puntos del Factorial $2x2x2$

Tabla 4.2

Como ya se mencionó, un diseño central compuesto, cuya base es un factorial completo, consiste de 2^k puntos que forman la base del diseño, $2k$ puntos estrella y varios puntos centrales. En la tabla 4.3 se resumen algunos de estos diseños con diferentes propiedades.

Diseños Centrales Compuestos					
k	2	3	4	5	6
F (Porcion Factorial) $F = 2^k$	4	8	16	32	64
Puntos Axiales	4	6	8	10	12
α	1.414	1.682	2	2.378	2.828
Puntos Centrales (P. U.)=n	5	6	7	10	15
Puntos Centrales (ortogonales)=n	8	9	12	17	24
Pruebas Totales(P. U.)=N	13	20	31	52	91
Pruebas Totales (ortogonales)=N	16	23	36	59	100

Tabla 4.3

Los diseños centrales compuestos son muy eficientes [4], pues proporcionan mucha información sobre los efectos de las variables experimentales y sobre todo del error experimental, todo ello en un número mínimo de corridas posibles.

Aunque en el problema de aplicación que se aborda en esta tesis se utiliza un diseño central compuesto, se explica a continuación un diseño también muy usado para modelar una respuesta de segundo orden, que es conocido como Diseño Box-Behnken.

4.4.3 Diseños Box-Behnken

Este es un tipo de diseño experimental, propuesto por Box y Behnken [3], el cual tiene tres niveles en cada factor, lo que permite la estimación de un modelo cuadrático completo incluyendo interacciones. Este diseño es una alternativa para ajustar modelos cuadráticos que requieren tres niveles de cada factor. Es un

diseño cuadrático independiente que se forma combinando factoriales 2^k con diseños de bloques incompletos balanceados. Estos últimos se conocen como incompletos pues en cada bloque se prueba sólo una parte de los tratamientos, y balanceados cuando cada par de tratamientos se prueba el mismo número de veces.

El Diseño Box-Benkhen se construye con:

- Puntos Centrales que como se explicó anteriormente sirven para examinar la presencia de curvatura, dar información acerca de los efectos cuadráticos y proporcionar una estimación de la magnitud del error experimental. El número de puntos centrales se puede escoger también para establecer rotabilidad.
- Puntos sobre la superficie, de igual distancia del punto central.

Este diseño consiste en localizar los puntos medios de las aristas y el punto central del espacio procesado, como se ilustra en la figura 4.6, para el caso de tres factores. Es un diseño rotable o casi rotable.

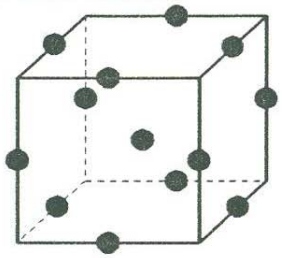


Figura 4.6

Los puntos se generan seleccionando dos factores y perturbando en forma completa sus niveles. Esta estrategia se repite para todos los pares de valores posibles. En la tabla siguiente se muestra la matriz del diseño de Box-Behnken para tres factores. Los puntos del diseño se localizan en la mitad de las aristas del cubo centrado en el origen, lo que permite que los tratamientos tomen valores más pequeños que en un factorial 3^2.

Diseño de Box-Behnken con tres factores

x_1	x_2	x_3
-1	-1	0
-1	1	0
1	-1	0
1	1	0
-1	0	-1
-1	0	1
1	0	-1
1	0	1
0	-1	-1
0	-1	1
0	1	-1
0	1	1
0	0	0
0	0	0
0	0	0

Se puede observar que este tipo de diseño carece de los puntos factoriales que corresponden a los vértices del hipotético cubo, lo cual es una gran ventaja cuando la realización de esos puntos es muy cara o requieren condiciones físicas complejas, pues se encuentran lejos del centro, que es regularmente la zona estándar de trabajo.

4.4.4 Selección de un Diseño de Segundo Orden

Aunque no se presentan en esta tesis, existe una gran variedad de diseños para estimar superficies de segundo orden. De hecho, algunos experimentadores se sienten perdidos cuando, al no conocer sobre el tema, pretenden utilizar algún software estadístico y se encuentran con una gran cantidad de diseños para estimar este tipo de respuesta, todos ellos con características y propiedades diferentes. ¿Cómo seleccionar entonces el diseño más apropiado?. Los criterios a seguir son generalmente:

- Se debe tratar de minimizar el número de tratamientos, pero al mismo tiempo el diseño debe permitir la estimación de todos los parámetros de interés en el modelo de segundo orden.
- Debe ser flexible, es decir, que pueda construirse a partir de diseños de primer orden.
- Propiedades como ortogonalidad, rotabilidad y/o precisión uniforme son convenientes pues están relacionadas con la calidad de la estimación de los parámetros.

4.5 Análisis de una Superficie de Respuesta

Cuando se ha encontrado la región donde se cree se encuentra el óptimo o punto estacionario, se puede caracterizar la superficie para ver si se tiene un máximo, un mínimo o un punto silla. Estas técnicas de optimización dependen del modelo ajustado y existen básicamente tres métodos que son: escalamiento ascendente, análisis canónico y análisis de cordilleras. En esta tesis solamente se utilizan los primeros dos. Enseguida se muestra en qué consiste el análisis canónico y la aplicación de éste en el problema estudiado, se desarrolla en el capítulo cinco.

4.5.1 Análisis canónico

Es una técnica utilizada para analizar los modelos de segundo orden y caracterizar su superficie, estudiando las coordenadas del punto estacionario, el tipo de punto y la orientación de la superficie. Este análisis consiste en reescribir el modelo ajustado de segundo orden en su forma canónica, que consiste en expresarlo en términos de nuevas variables, llamadas variables canónicas, las cuales son transformaciones de las variables codificadas. La ventaja es que la

ecuación canónica proporciona información a simple vista sobre el tipo de superficie que se está observando y sobre su forma.

Para realizar un análisis canónico se realizan los pasos señalados en las secciones anteriores, que son: seleccionar los niveles de los factores para así determinar la región de exploración, correr un diseño de segundo orden para explorar la región experimental determinada antes y ya encontrado el modelo de segundo orden:

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=1}^{k} \hat{\beta}_{ii} x_{ii}^2 + \sum_{i=1}^{k} \sum_{<j=1}^{k} \hat{\beta}_{ij} x_i x_j \ ,$$

determinar el punto estacionario, o sea el punto donde el plano tangente a la superficie tiene pendiente cero, el cual se localiza derivando con respecto a cada variable x_i, igualando a cero y despejando. Todo esto se puede facilitar si el modelo se rescribe en notación matricial como:

$$\hat{Y} = \hat{\beta}_0 + \boldsymbol{x}^T \hat{\boldsymbol{\beta}} + \boldsymbol{x}^T \boldsymbol{B}\boldsymbol{x} \qquad (4.3)$$

donde

$\boldsymbol{x}^T = [x_1, x_2, \ldots x_k]$ es un punto en la región de operabilidad del proceso

$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ es el vector de los coeficientes de la parte lineal del modelo

y $\boldsymbol{B} = \begin{bmatrix} \hat{\beta}_{11} & \hat{\beta}_{12}/2 & \hat{\beta}_{13}/2 & \cdots & \hat{\beta}_{1k}/2 \\ \hat{\beta}_{12}/2 & \hat{\beta}_{22} & \hat{\beta}_{23}/2 & \cdots & \hat{\beta}_{2k}/2 \\ \hat{\beta}_{13}/2 & \hat{\beta}_{23}/2 & \hat{\beta}_{33} & \cdots & \hat{\beta}_{3k}/2 \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{\beta}_{1k}/2 & \hat{\beta}_{2k}/2 & \hat{\beta}_{3k/2} & \cdots & \hat{\beta}_{kk} \end{bmatrix}$

es la matriz de los coeficientes de las interacciones y de los términos cuadráticos puros.

El punto estacionario se encuentra derivando $\hat{Y}$ con respecto al vector $\boldsymbol{x}$ e igualando a cero

$$\partial \hat{Y} / \partial \boldsymbol{x} = \hat{\boldsymbol{\beta}} + 2\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$$

de donde se obtiene $\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\hat{\boldsymbol{\beta}}$, donde $\mathbf{B}^{-1}$ es la inversa de la matriz **B.** Sustituyendo lo anterior en la ecuación (4.3), se obtiene que la respuesta estimada en el punto estacionario es

superficie ajustada (elipsoide) tiene un máximo, como se observa en la figura 4.7.

w_2 w_1

$$\hat{Y}_s = \hat{\beta}_0 + \frac{1}{2}\mathbf{x}_s^T\hat{\beta} \qquad (4.4)$$

La ecuación canónica es una manera efectiva de visualizar la superficie y determinar la sensibilidad relativa de la respuesta a los cambios en cada uno de los factores. El análisis canónico gira los ejes de las variables x_i a un nuevo sistema de coordenadas (ver apéndice C, teorema de ejes principales), y el centro de este nuevo sistema se coloca en el punto de respuesta estacionario.

La ecuación canónica de una superficie puede ser representada de la siguiente manera:

$$\hat{Y} = \hat{Y}_s + \lambda_1 w_1^2 + \lambda_2 w_2^2 + \ldots + \lambda_k w_k^k$$

donde $\hat{Y}_s$ está dada por la ecuación (4.4) o sea es el valor predicho por el modelo sobre el punto estacionario; w_i representa a las variables independientes transformadas, conocidas como variables canónicas, y las λ_i son los valores propios de la matriz ***B***.

Los tamaños y signos de las λ_i determinan el tipo de superficie de respuesta y punto estacionario que se ha encontrado:

- Si los valores propios son todos negativos, el punto estacionario es un máximo.
- Si son todos positivos, el punto estacionario es un mínimo.
- Si los valores propios son de signos mezclados, el punto estacionario es un punto silla.

Se puede entonces describir la forma de la superficie de respuesta a partir del punto estacionario y de las magnitudes y signos de las λ_i's.

La forma canónica de una ecuación con dos variables es:

$$\hat{Y} - \hat{Y}_s = \lambda_{11} w_1^2 + \lambda_{22} w_2^2$$

Para este caso, si λ_{11} y λ_{22} son ambos negativos, se puede decir que un movimiento en cualquier dirección, alejándose del centro del nuevo sistema de coordenadas, resulta en una pérdida cuadrática de respuesta, es por eso que la superficie ajustada (elipsoide) tiene un máximo, como se observa en la figura 4.7.

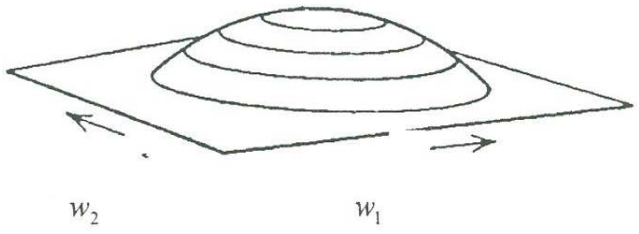


Figura 4.7

En el caso de ser ambos positivos se tendrá un mínimo (también elipsoide). Por otra parte si uno de los términos es negativo y el otro termino es positivo entonces se dice es un punto silla (paraboloide hiperbólico), como en la figura 4.8.

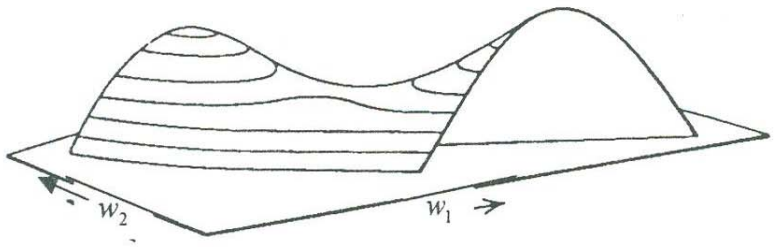


Figura 4.8

Para el caso en que por ejemplo λ_{22} sea aproximadamente cero indica la existencia de algún tipo de teja (cilindro), un ejemplo el representado en la figura 4.9.

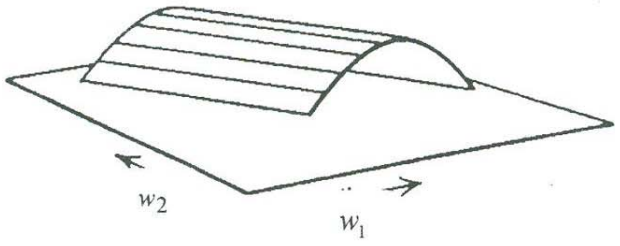


Figura 4.9

Para encontrar el valor de los $\lambda's$ o valores propios de la matriz $\boldsymbol{B}$ hay que recordar que son constantes que cumplen la relación

$$\boldsymbol{B}\boldsymbol{m}_i = \lambda_i \boldsymbol{m}_i$$

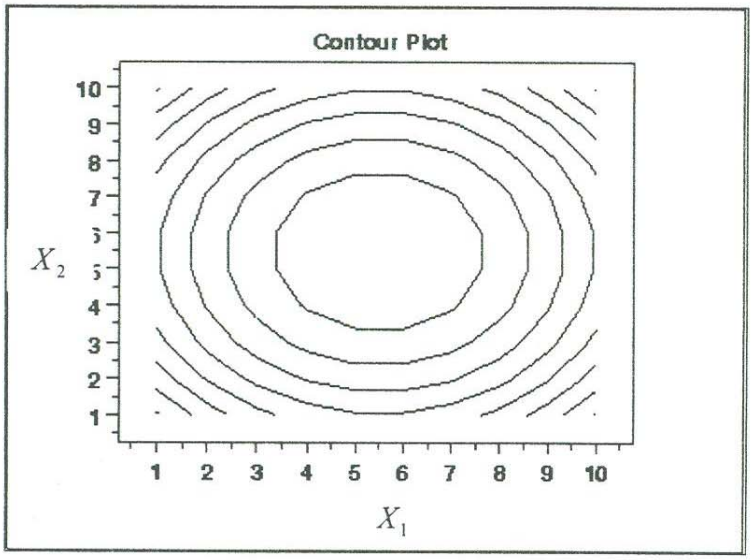
donde $\boldsymbol{m}_i$ es un vector propio, también llamado vector característico, asociado al valor propio λ_i, el cual se obtiene al resolver calculando e igualando a cero, el determinante de la matriz

$$|\boldsymbol{B} - \lambda \boldsymbol{I}| ,$$

donde $\boldsymbol{I}$ es la matriz identidad de dimensión k. Cuando se tienen varios factores en el diseño, resulta difícil resolver el polinomio resultante, por lo cual se recomienda utilizar algún apoyo computacional para encontrar estos valores propios.

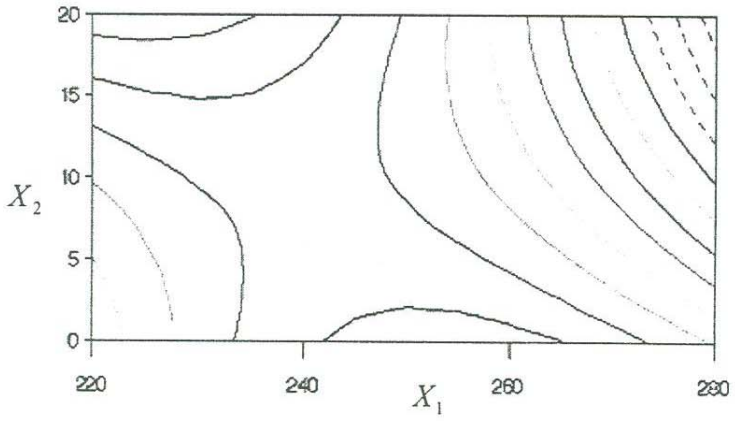
4.5.1 Análisis Gráfico de una Superficie de Respuesta

Una superficie de respuesta se puede analizar también gráficamente por medio de los contornos de ésta, también conocidos como curvas de nivel. En la metodología de superficie de respuesta el método gráfico tiene muchas ventajas, pues permite determinar visualmente cómo se espera que sea el comportamiento de la respuesta en estudio. Por ejemplo un máximo o un mínimo puede estar representado por una como la expuesta a continuación (gráfica 4.1) expuesta a continuación, donde dependiendo de los valores mostrados, se tendrá un máximo si los círculos en el interior tienen valores más grandes, y estos disminuyen conforme se alejan del centro y un mínimo cuando se comporten de una manera contraria a lo anterior.



Gráfica 4.1

En el caso de que la gráfica presente un punto silla, los contornos se verán como en la gráfica 4.2



Gráfica 4.2

El método gráfico también permite visualizar superficies de respuesta simultáneas, esto es, cuando se están manejando los mismos factores o variables independientes y dos o más respuestas al mismo tiempo. De esta manera el experimentador se asegura que el punto óptimo simultáneo encontrado sea en realidad un óptimo global, dado que cuando se utiliza esta metodología por separado, el método detecta la presencia de varias regiones factibles separadas (cuando existan), que pueden dar lugar a óptimos locales, pero no corresponden a un óptimo global. Al optimizar varias respuestas el método gráfico sobrepone sobre la región experimental las superficies de respuesta descritas por los diferentes modelos ajustados y se trata de localizar dentro de ella subregiones en las cuales todos los modelos predicen valores aceptables para las respuestas. Este método es sencillo cuando se tienen dos respuestas pues así las superficies se pueden sobreponer en forma de contornos sobre la región experimental y es fácil visualizar el óptimo aceptable. Cuando intervienen más respuestas, la gráfica de contornos simultánea no es muy clara, por lo que se recomiendan otros procedimientos para hacer el análisis.

CAPÍTULO 5

APLICACIÓN DE LA MSR EN LA INDUSTRIA ACEITERA

La industria aceitera hoy en día es más cuidadosa con el control de cada una de las etapas del procesamiento de los aceites. Los aceites vegetales deben ser sometidos a un proceso de refinación antes de poder ser consumidos por el hombre. Este proceso sirve para lograr un sabor suave y estabilidad oxidativa. Los aceites son sometidos a diferentes etapas unitarias que involucran en conjunto el proceso de refinación (desgomado, neutralización, blanqueo y desodorización).

En la producción de aceites se necesitan implementar procedimientos de optimización que permitan conseguir mejores resultados que los obtenidos por los medios tradicionales. Para la industria mexicana es de suma importancia poder competir con productos de calidad, lo que implica producir aceites con alta calidad oxidativa, bajo color y nutricionalmente aceptables. La elevada demanda de aceites en el mundo ha traído como consecuencia que su producción vaya en aumento y actualmente se le está dando importancia al efecto del procesamiento sobre la calidad nutricional de los aceites, ya que éstos son portadores de vitaminas y ácidos grasos esenciales.

En este trabajo se ha considerado solamente una de las etapas del procesamiento del aceite, en este caso de soya, que consiste en el proceso de blanqueo, que se explicará más delante. En este paso se tiene como objetivo optimizar los tocoferoles (vitamina E), que son antioxidantes naturales y obtener un valor adecuado de peróxidos, lo que proporcionará mejor estabilidad oxidativa en el aceite.

5.1 Procesamiento del Aceite de Soya

El proceso de refinamiento del aceite de soya tiene como finalidad eliminar impurezas liposolubles que se encuentran en el aceite crudo, y así obtener un aceite con altas normas de calidad como son su sabor, apariencia y estabilidad oxidativa. El cumplimiento de estas normas implica generalmente la pérdida de ciertas propiedades nutricionales como son los tocoferoles. En Estados Unidos y otros países desarrollados se cuida cada vez más el no afectar dichas propiedades, poniéndose especial atención durante las etapas de blanqueo y desodorización.

En el diagrama de la figura 5.1 se muestran las etapas a seguir en el procesamiento del aceite de soya, las cuales se explican enseguida.

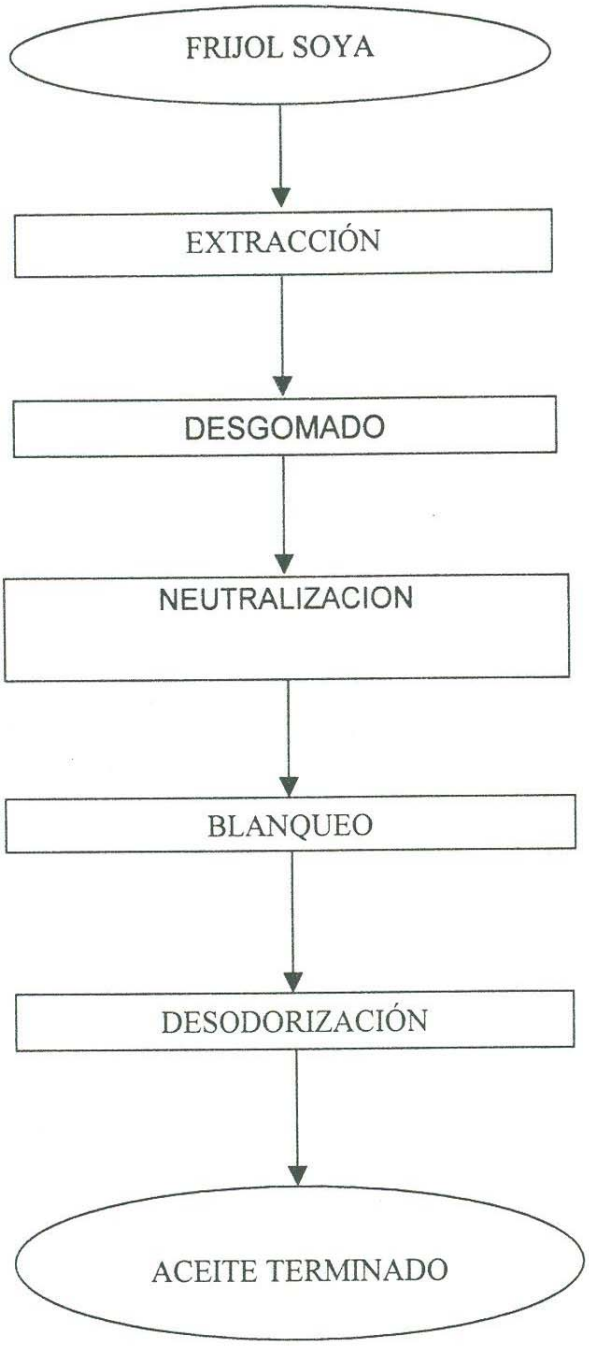

FRIJOL SOYA
EXTRACCIÓN
DESGOMADO
NEUTRALIZACION
BLANQUEO
DESODORIZACIÓN
ACEITE TERMINADO

Figura 5.1

Desgomado. La operación de desgomado es el primer paso de la refinación del aceite crudo y tiene como objetivo separar los fosfolípidos, las partículas metálicas (sales de calcio, magnesio y fierro) y todas las partículas sólidas del aceite, dándole así mayor estabilidad a la oxidación lipídica, haciendo más eficiente su refinación alcalina y blanqueo.

Neutralización. El proceso de refinación puede realizarse por el método químico con sosa cáustica, la cual también es llamada neutralización o por el método físico el cual es una refinación por destilación con vapor. Esta etapa tiene como finalidad reducir al mínimo los fosfolípidos y los ácidos grasos libres.

Blanqueo. El proceso de blanqueo con tierras ácidas activadas, es el paso más importante del procesamiento de aceite de soya. El estudio de esta etapa ha traído como resultado la confirmación de que el blanqueo afecta a los aceites más allá de la remoción de material colorante y que además este proceso es determinante en la estabilidad del aceite blanqueado [34], [36]. La reducción del color del aceite de soya, ocurre durante todos los pasos de su procesamiento: desgomado, refinación con sosa cáustica, blanqueo, hidrogenación y desodorización. Sin embargo, la operación de blanqueo es fundamental en el aceite de soya y actualmente se está teniendo cuidado con las temperaturas que se utilizan, así como en el tipo de tierras y el tiempo de contacto durante dicho proceso, para tener la mayor eliminación de los compuestos coloridos que afectan la calidad de los aceites.

Desodorización. El objetivo de la desodorización es obtener un aceite, sin olor, con apenas un ligero sabor a almendra, un color tenue, un contenido de ácidos grasos libres menor del 0.05% y una buena estabilidad oxidativa [15]. Esta etapa consiste en inyectar vapor de agua al aceite que se encuentra a una presión absoluta baja y a una temperatura lo suficientemente alta para eliminar por arrastre de vapor algunos compuestos que si no son eliminados proporcionan al aceite olores y sabores desagradables.

5.1.1 Factores que Afectan el proceso de Blanqueo

En la etapa de blanqueo del proceso de refinación de los aceites existen diferentes factores que son de gran importancia, y que deben ser manejados con un estricto control y son: tiempo de contacto, temperatura de blanqueo, cantidad de tierras, y presión.

Ti*empo.* El tiempo necesario para la remoción máxima del color depende de la calidad de tierras de blanqueo y de la temperatura de blanqueo. El tiempo de contacto óptimo para muchos aceites se encuentra en el rango de 20 a 30 minutos y una temperatura de 90°C a 100°C. Se recomiendan tiempos cortos cuando se utilizan temperaturas altas.

Cantidad de Tierras. La remoción de las impurezas se incrementa con la cantidad de adsorbente (arcillas). Sin embargo, es difícil de calcular la mínima cantidad de tierra de blanqueo necesarias para un proceso eficiente. Esto se debe a que los distintos tipos de aceites contienen diferentes fuentes y cantidades de impurezas, que pueden interaccionar de manera distinta con las tierras de blanqueo. El aceite después del proceso de blanqueo debe ser de un color claro, esto se logra por efecto de la presión utilizada en el filtro cuando la concentración de tierras es la adecuada, para no dejar pasar ninguna impureza en éste. Se ha visto que altas concentraciones de tierra de blanqueo fortalecen la purificación de las grasas y aceites.

Temperatura. Las temperaturas de blanqueo usualmente se encuentran dentro del intervalo de noventa a ciento diez grados centígrados. El uso de temperaturas altas disminuye la viscosidad del aceite favoreciéndose las reacciones que influyen en la calidad y la vida media del mismo (reacciones de oxidación y reacciones de fijación de color). La temperatura adecuada depende del tipo de aceite, de los productos y sus impurezas [1].

Presión. En el pasado el blanqueo era realizado en condiciones atmosféricas. Sin embargo, varios experimentos han probado el efecto determinante del oxígeno en la calidad o estabilidad oxidativa, y se ha reportado un incremento en el valor de peróxidos cuando el blanqueo se lleva a cabo a presión atmosférica (presencia de oxigeno), mientras que, este efecto no se presenta con el uso de vacío. Entonces el proceso de blanqueo ofrece una mejor protección del aceite a la oxidación cuando es procesado al vacío, es decir todas las grasas y aceites deben de ser procesados protegiéndolos del oxígeno, para obtener un aceite sin sabor, con estabilidad oxidativa y con mayor vida de anaquel [30].

5.2 Aplicación de la MSR en la producción de aceite de soya

Por todo lo anterior, para la optimización de tocoferoles e índice de peróxidos, en la operación de blanqueo se considera solamente el experimento realizado en ausencia de oxígeno, aunque en la práctica también se llevó a cabo en presencia de éste.

La existencia de investigaciones similares realizadas en el extranjero y estudios previos efectuados por investigadores del Departamento de Investigaciones Científicas y Tecnológicas de la Universidad de Sonora, donde se llevó a cabo este experimento, permitieron retomar las primeras etapas de una metodología de superficie de respuesta, como son el diseño factorial completo o incompleto y el método de escalamiento ascendente, y proponer un diseño para modelar una superficie de segundo grado. Todo esto por la razón de que se conocía el rango de operabilidad de las variables en estudio y solamente se deseaba optimizar las respuestas en estos rangos.

El diseño seleccionado fue un diseño central compuesto para tres factores, con las propiedades de rotabilidad y precisión uniforme. Las corridas se efectuaron de manera aleatoria y los resultados se muestran en la siguiente tabla:

Orden de las corridas	*Temperatura (°C)*	*Cantidad de Tierras (gms)*	*Tiempo (min.)*	*Índice de Peróxido*	*(%)Ret. De Tocoferoles*
19	90	1	20	0.113	86.95
8	130	1	20	0.174	82.45
20	90	3	20	0.102	93.01
12	130	3	20	0.113	87.94
3	90	1	60	0.163	85.05
6	130	1	60	0.250	75.6
1	90	3	60	0.099	89.41
10	130	3	60	0.119	66.65
9	77	2	40	0.225	83.4
2	143	2	40	0.418	66.49
4	110	0.318	40	0.285	86.84
5	110	3.68	40	0.121	85.83
16	110	2	6.4	0.158	83.25
14	110	2	74	0.323	83.52
15	110	2	40	0.123	90.31
7	110	2	40	0.123	87.87
17	110	2	40	0.125	87.41
11	110	2	40	0.125	88.33
13	110	2	40	0.121	88.16
18	110	2	40	0.121	88.31

Tabla 5.1

Los valores asignados a los puntos axiales, mostrados en la tabla anterior, se encuentran redondeados. Para obtener los valores correctos para las corridas se debe crear el diseño estableciendo que:

	Nivel Bajo (-1)	*Valor Central (0)*	*Nivel Alto (+1)*
Temperatura	90	110	130
C. de Tierra	1	2	3
Tiempo	20	40	60

Tabla 5.2

Los datos fueron analizados utilizando el software estadístico JMP IN, y sobre los resultados obtenidos con el mismo, se derivan las conclusiones. Se plantean, asimismo, las hipótesis para cada uno de estos resultados.

Utilizando la tabla de análisis de varianza para la retención de tocoferoles, que se muestra a continuación, se puede probar la hipótesis siguiente:

$$H_0: El\ modelo\ no\ es\ adecuado$$
$$H_1: El\ modelo\ es\ adecuado$$

Fuente de Variación	*Grados de libertad*	*Suma de cuadrados*	*Cuadrado Medio*	*Razón F*	*Prob > F*
Modelo	9	840.83621	93.4262	8.7440	0.0011
Error	10	106.84637	10.6846		
C. Total	19	947.68258			

Tabla 5.3

La probabilidad obtenida para el modelo es de 0.0011, lo que implica un rechazo de la hipótesis nula, esto es concluimos que el modelo es adecuado con un nivel de significancía de 0.05.

Por otra parte el coeficiente de determinación calculado es:

$$R^2 = \frac{840.83621}{947.68258} = 0.8872$$

Este coeficiente indica que el modelo explica el 88.72% de la variabilidad en la respuesta, por lo que se podría pensar que el modelo explica de forma adecuada esta variabilidad.

Sin embargo la prueba de falta de ajuste, planteada como:

$$H_0: No\ hay\ falta\ de\ ajuste$$
$$H_1: Hay\ falta\ de\ ajuste$$

cuyo resultado se puede obtener de la tabla 5.4, sale significativa.

Fuente de Variación	*Grados de libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	*Razón F*	*Prob > F*
Falta de ajuste	5	101.86669	20.3733	20.4565	0.0024
Error puro	5	4.97968	0.9959		
Error Total	10	106.84637			

Tabla 5.4

Esto es, se rechaza H_0, pues Prob > F es 0.0024, por lo tanto hay falta de ajuste, lo que tal vez implica considerar reformas en el modelo.

La estimación de los parámetros para la retención de tocoferoles, obtenidas en el JMP fueron las siguientes:

Parámetro	*Estimación*	*Error estándar*	*Estadístico T*	*Valor p*
Intersección	-73.37992	32.71097	-2.24	0.0487*
Temperatura	2.5551901	0.502935	5.08	0.0005*
CantTierr	14.987765	7.641236	1.96	0.0783
Tiempo	1.1183478	0.382062	2.93	0.0151*
(Temp..)(Temp.)	-0.010708	0.002153	-4.97	0.0006*
(CantTierr)(Temp)	-0.08675	0.057784	-1.50	0.1642
(CantTierr)(CantTierr)	-0.256252	0.86105	-0.30	0.7721
(Tiempo)(Temp)	-0.007075	0.002889	-2.45	0.0343*
(Tiempo)(CantTierr)	-0.100875	0.057784	-1.75	0.1114
(Tiempo)(Tiempo)	-0.003248	0.002153	-1.51	0.1623

Tabla 5.5

*Significativo para $\alpha = 0.05$

Recordemos que las pruebas sobre los coeficientes individuales de la regresión, como ya se explicó en el capítulo tres, se plantean

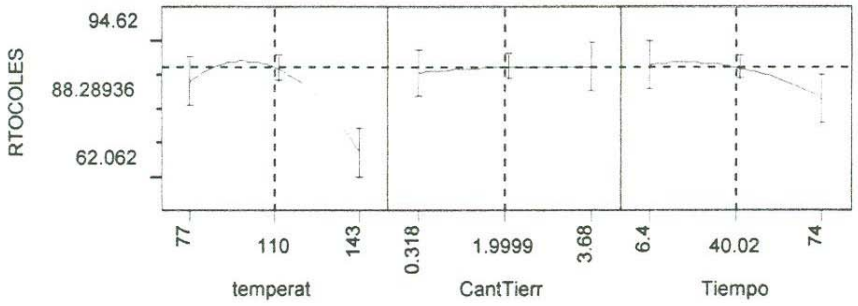
$$H_o : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

y se rechazará la hipótesis nula, en el caso de que la probabilidad calculada sea menor que el nivel de significancia especificado, en este caso se utilizó $\alpha = 0.05$, por lo que probabilidades menores que este valor, se señalaron con un asterisco.

En la tabla 5.5 se puede apreciar que la retención de tocoferoles (RTOCOLES) fue influenciada significativamente por la temperatura en su término cuadrático y lineal, por el tiempo de contacto y por la interacción entre tiempo y temperatura.

La temperatura es la variable que presenta un cambio mas drástico en la retención de tocoferoles, siendo a partir de su valor central (110 °C) cuando ésta empieza a disminuir.

Algo similar parece suceder con el tiempo de contacto, mas el cambio no es tan radical, como se observa en la gráfica 5.1.



Gráfica 5.1

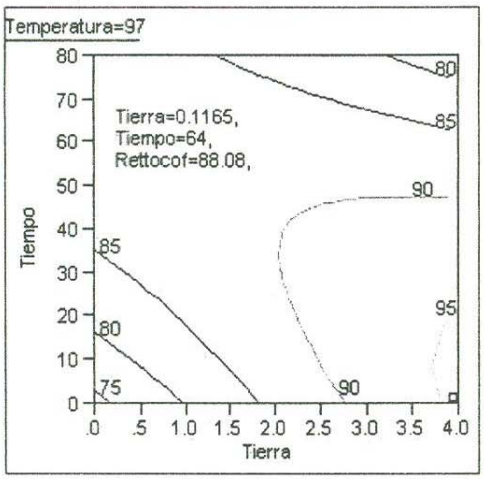
La solución estimada por mínimos cuadrados y obtenida en JMP IN es:

Solution	
Variable	Critical Value
Temperatura	97.72058
Tierra	0.1135926
Tiempo	63.963202
Solution is a	SaddlePoint
Critical values outside data range	
Predicted Value at Solution	88.085213

Los valores de temperatura y tiempo propuestos como solución se encuentran dentro del rango de experimentación, más no así la cantidad de tierras. La retención de tocoferol predicha por el modelo es de 88.08%. La solución obtenida es un punto silla como se puede confirmar de los resultados de los valores propios obtenidos, pues estos son de diferente signo.

EigenStructure			
EigenValues and EigenVectors			
Variable	0.0083	-0.0054	-0.2731
Temperatura	0.36075	0.91835	0.16275
Tierra	-0.23144	-0.08090	0.96948
Tiempo	0.90349	-0.38740	0.18336

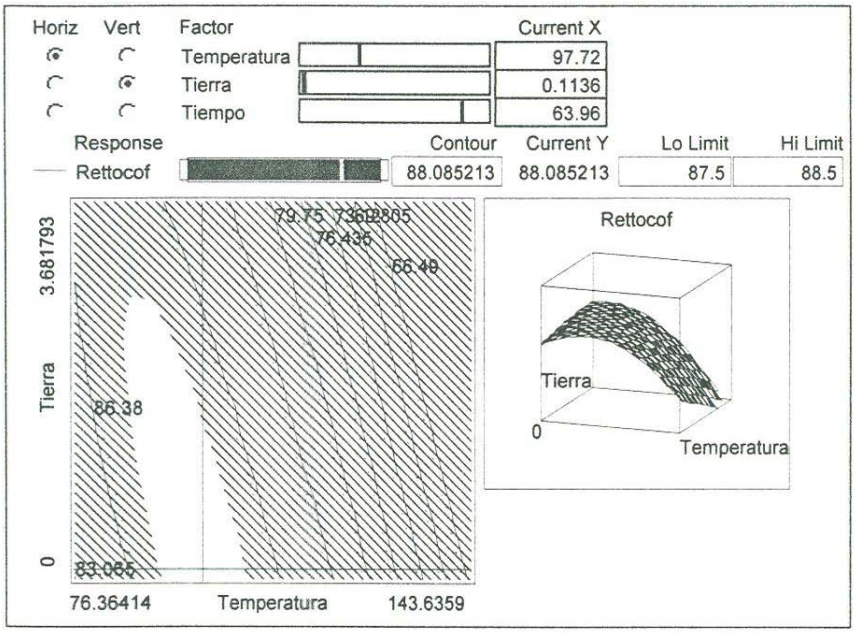
La gráfica 5.2 muestra los contornos para la interacción entre tiempo y cantidad de tierras, dejando una temperatura fija de 97°C



Gráfica 5.2

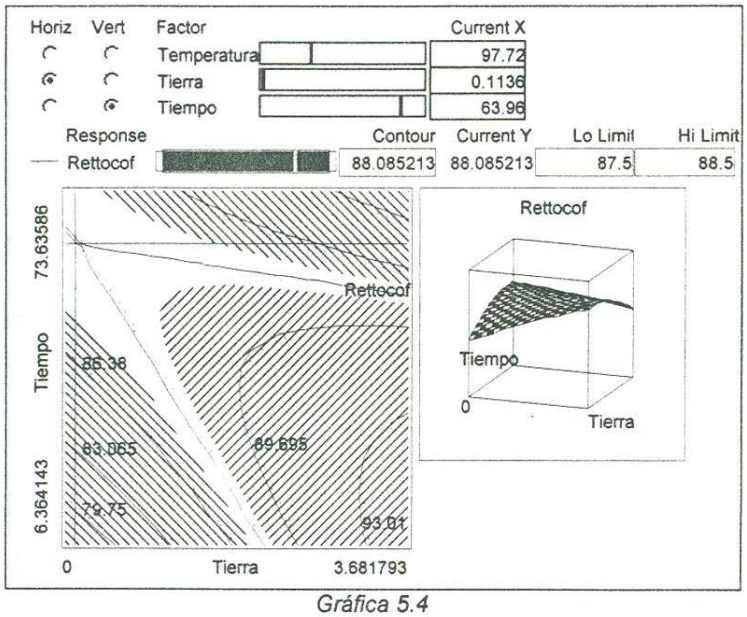
Al explorar manualmente, en el JMP IN, la superficie de respuesta utilizando la gráfica de contornos, se puede ver aproximadamente la localización de la solución propuesta (gráfica 5.2).

En la gráfica 5.3 se ilustran los contornos para la retención de tocoferol con los factores de temperatura, tiempo y cantidad de tierras, propuestos como solución. Las region no sombreada (en blanco) es donde se encuentra una retención de tocoferol de entre 87.5% a 88.5%. Dentro de ésta se muestran dos rectas que se cortan perpendicularmente mostrando la solución predicha.



Gráfica 5.3

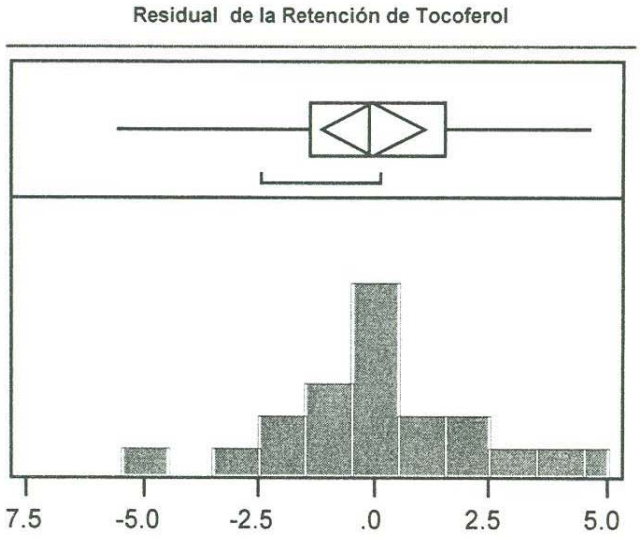
Enseguida la gráfica 5.4 muestra los contornos con los factores de cantidad de tierra y tiempo, y con las mismas especificaciones anteriores.



Gráfica 5.4

En las gráficas 5.3 y 5.4 se puede apreciar que la solución propuesta para el factor cantidad de tierras está fuera del rango de operabilidad propuesto.

El análisis para verificar la adecuación del modelo se muestra a continuación, iniciando con un histograma y un diagrama de caja para los residuales obtenidos del ajuste de éste, los cuales parecen indicar normalidad.



Gráfica 5.5

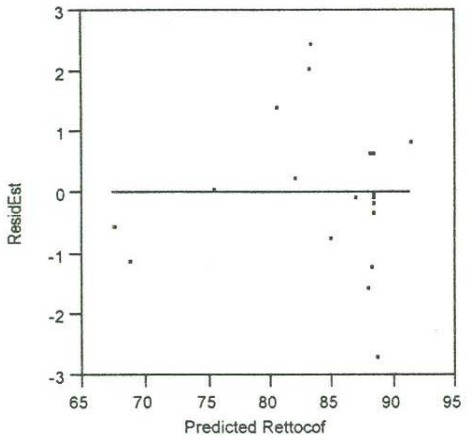
Para comprobarlo se efectuó una prueba de hipótesis para probar la bondad del ajuste a una distribución normal y las hipótesis planteadas fueron:

H_0: *Los residuales siguen una distribución normal*

H_1: *Los residuales no siguen una distribución normal*

Test for Normality	
Shapiro-Wilk W Test	
W	Prob<W
0.978808	0.9015

Como se puede ver en los resultados obtenidos, no se rechaza la hipótesis nula, por lo que el supuesto de normalidad se satisface. Se muestra a continuación la gráfica 5.6 de valores predichos contra residuales estandarizados.



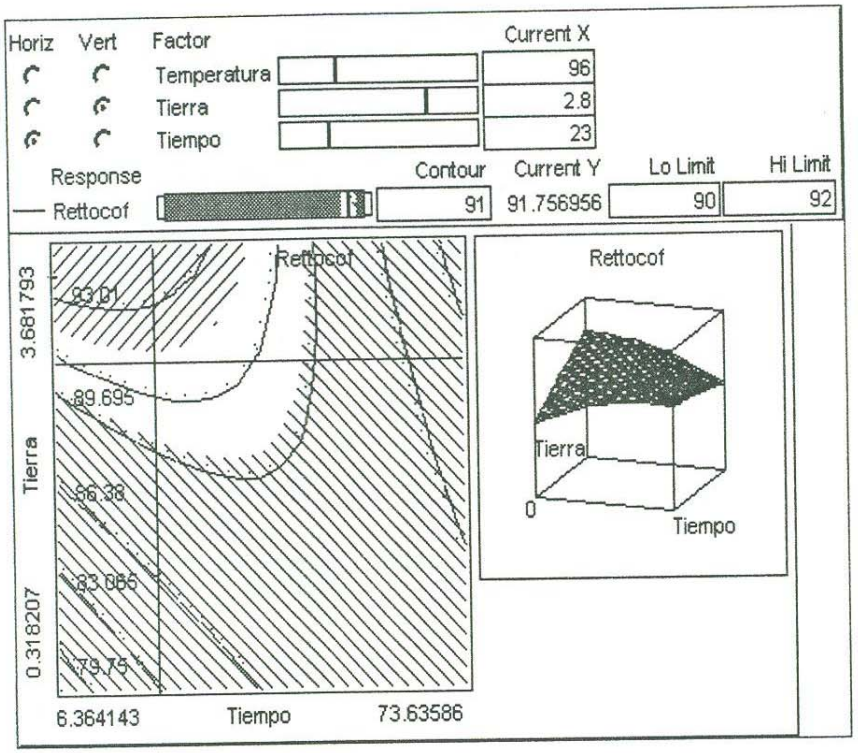
Gráfica 5.6

Se calcularon también las siguientes medidas descriptivas, y un intervalo al 95% de confianza para la media de los residuales.

Mean	-0.00000
Std Dev	2.37139
Std Error Mean	0.53026
Upper 95% Mean	1.10984
Lower 95% Mean	-1.10984
N	20.00000

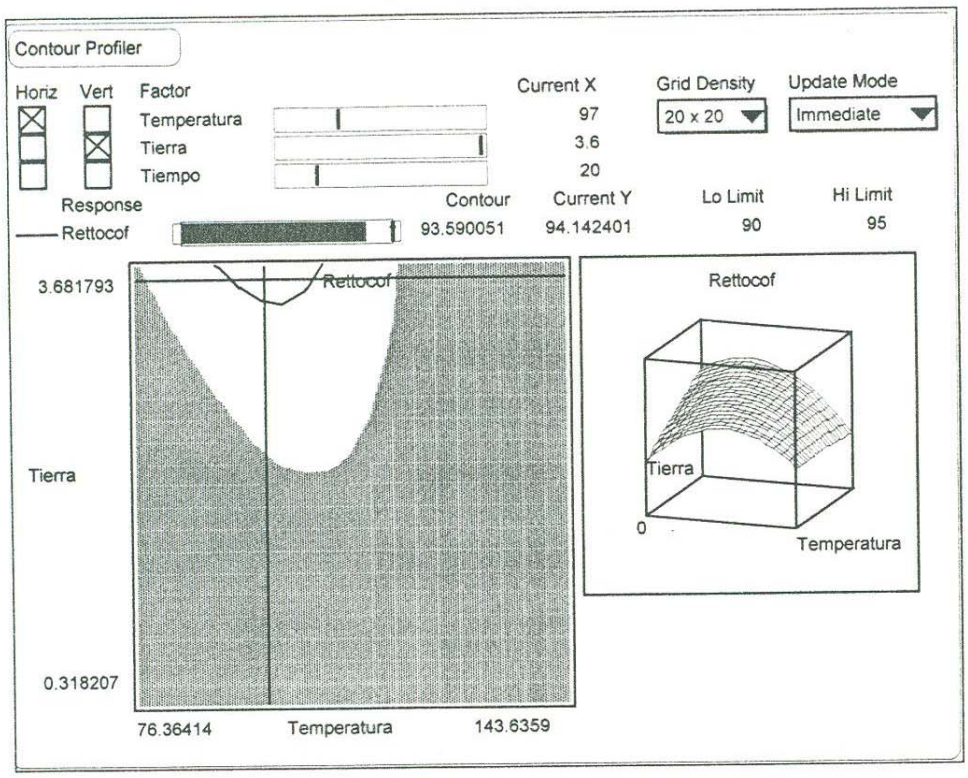
Se muestra también la gráfica 5.7 de residuales obtenida en JMP IN, considerando el orden de las corridas.

Hay que recordar que la solución encontrada no satisface los rangos de operabilidad de las variables, por ser ésta un punto silla. El software utilizado permite moverse en la dirección en la cual se incremente la retención de tocoferol y hay algunas soluciones que brindan una retención aceptable, en puntos dentro del rango de operabilidad de las variables. Por ejemplo, en la siguiente gráfica se ilustra una retención de tocoferoles de 91.75%, utilizando una temperatura de 96 °C, con cantidad de tierras de 1.4% (2.8 gramos por cada 200 gramos) y con un tiempo de contacto de 23 minutos. Estos valores sugeridos se validaron en el laboratorio, con excelentes resultados.



Gráfica 5.8

Por otra parte, al explorar un poco más la superficie, y tomando como base los resultados anteriores, se obtuvo una retención de tocoferoles de 94.14%, con una temperatura de 97 °C, una cantidad de tierras de 1.8%, y con un tiempo de contacto de 20 minutos. Sin embargo, hay que aclarar que estas condiciones no se validaron. Enseguida se muestra la gráfica correspondiente.



Gráfica 5.9

En el modelo anterior el factor cantidad de tierras no es significativo, por lo cual se efectuó el análisis de los datos excluyendo éste y con los siguientes resultados.

H_0 : *El modelo no es adecuado*

H_1 : *El modelo es adecuado*

En la tabla de análisis de varianza que se muestra a continuación, podemos ver que Prob > F es menor de 0.0001, por lo que hay evidencia suficiente para rechazar la hipótesis nula, y concluir que el modelo es adecuado.

ANÁLISIS DE VARIANZA

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	*Razón F*	*Prob > F*
Modelo	5	781.21865	156.244	13.1405	<.0001
Error	14	166.46393	11.890		
C. Total	19	947.68258			

Tabla 5.6

Con respecto al coeficiente de determinación, éste explica ahora el 82.13% de la variabilidad en la retención de tocoferoles, un poco menos que el modelo que incluía el factor cantidad de tierras, pero este resultado era de esperarse.

$$R^2 = \frac{778.36563}{947.68258} = 0.821336$$

Hay que notar que ahora la prueba de falta de ajuste no es significativa al nivel de 0.05, ya que Prob >F es 0.0899, o sea no se rechaza la hipótesis nula, por lo tanto no hay falta de ajuste y entonces este modelo, sin el factor cantidad de tierras, parece ajustarse mejor a los datos.

H_0 : No hay falta de ajuste

H_1 : Hay falta de ajuste

Falta de Ajuste

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	*Razón F*	*Prob > F*
Falta de ajuste	3	74.45330	24.8178	2.8778	0.0844
Error Puro	11	94.86365	8.6240		
Error Total	14	169.31695			

Tabla 5.7

Las pruebas para cada uno de los parámetros de la ecuación de regresión se muestran a continuación.

$H_0 : \beta_i = 0$

H_1 : al menos un $\beta_i \neq 0$

Estimación de Parámetros

Variable	Estimado	Error Std	t Radio	Prob>|t|
Intersección	-43.76766	30.78789	-1.42	0.1770
Temperatura	2.3676957	0.513966	4.61	0.0004*
Tiempo	0.9115089	0.383737	2.38	0.0324*
(Temp.)(temp)	-0.010644	0.00226	-4.71	0.0003*
(tiempo)(temp)	-0.007075	0.003048	-2.32	0.0359*
(tiempo)(tiempo)	-0.003184	0.00226	-1.41	0.1806

Tabla 5.8

Podemos ver que el efecto cuadrático del tiempo y la intersección no fueron significativos y todos los demás si. Lo que nos conduce a un modelo de la siguiente forma:

$$Y = 2.3676957X_1 + 0.9115089X_3 - 0.010644X_1^2 - 0.007075X_1X_3$$

donde

$$X_1 = Temperatura, X_3 = Tiempo$$

La solución propuesta es la siguiente

Variable	Critical Value
temperat	101.37
Tiempo	30.218084

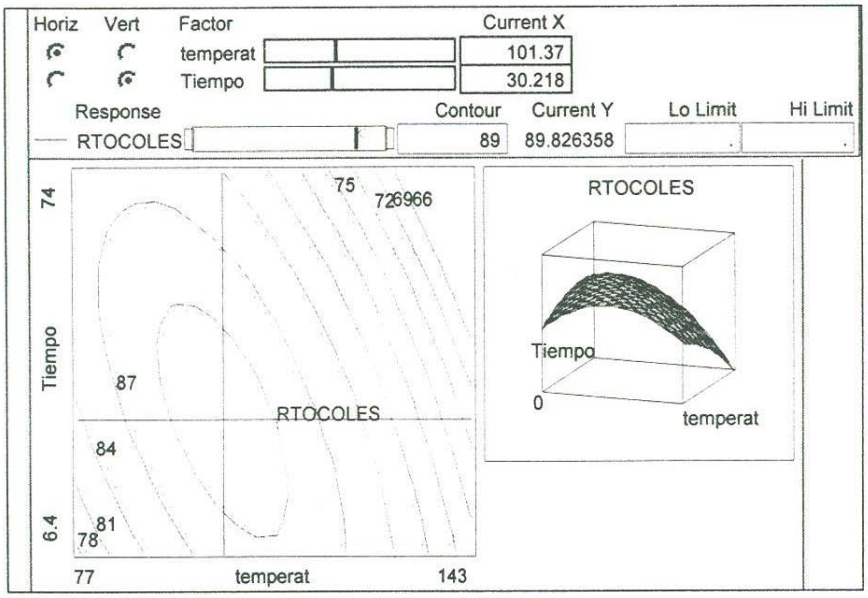
donde se obtiene una respuesta aproximada de 89.82% y la tabla para calcular su curvatura canónica, se muestra a continuación.

Curvatura Canónica

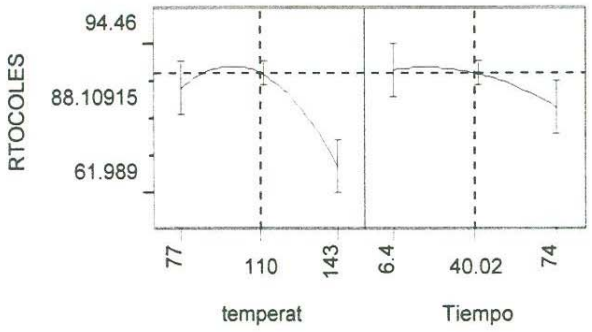
Tabla de Eigenvalores y Eigenvectores

Variable	-0.0017	-0.0123
Temperatura	-0.35673	0.93421
Tiempo	0.93421	0.35673

Tabla 5.9



Gráfica 5.10



Gráfica 5.11

La solución es un máximo, como se puede deducir del análisis canónico y observar en la gráfica de contornos, con valores de tiempo y temperatura dentro del rango de operabilidad y su valor predicho para la retención de tocoferoles es de 89.83, que supera a la obtenida anteriormente cuando se consideraba el factor cantidad de tierras en el diseño.

También podemos ver en las gráficas de perfiles de predicción (gráfica 5.11) que el factor temperatura influye significativamente en la retención de tocoferol y que el máximo común se alcanza para una temperatura de 101.37 grados centígrados y tiempo de 30.218 minutos. Hay que aclarar que este último resultado no se validó en el laboratorio y simplemente fue una sugerencia al efectuar el análisis.

Por otra parte también se analizó otra respuesta, que corresponde al índice o valor de peróxido. Para este caso el modelo no es significativo pues tenemos que Prob>F es de 0.2167, como se puede ver en la tabla 5.10 y no hay evidencia suficiente para rechazar la hipótesis nula mostrada a continuación.

H_0:*El modelo no es adecuado*

H_1:*El modelo es adecuado*

Análisis de Varianza para Índice de Peróxido

Fuente de Variación	*Grados de libertad*	*Suma de cuadrados*	*Cuadrado Medio*	*Razón F*	*Prob > F*
Modelo	9	0.08666191	0.009629	1.6742	0.2167
Error	10	0.05554504	0.005555		
C. Total	19	0.14220695			

Tabla 5.10

$$R^2 = \frac{0.08666191}{0.14220695} = 0.60914$$

y con este coeficiente de determinación, solo podemos explicar el 60.914% de la variabilidad en la respuesta, que es un porcentaje relativamente bajo y se confirma esto, pues ninguno de los efectos fue significativo para un $\alpha = 0.05$.

Esto se puede corroborar más aún, efectuando la prueba de falta de ajuste mostrada a continuación

H_0 : *No hay falta de ajuste*

H_1 : *Hay falta de ajuste*

Falta de ajuste para Índice de Peróxido

Fuente de Variación	*Grados de libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	*Razón F*	*Prob > F*
Falta de ajuste	5	0.05552904	0.011106	3470.565	<.0001
Error puro	5	0.00001600	0.000003		
Error Total	10	0.05554504			

Tabla 5.11

la cual indica que hay falta de ajuste, pues Prob>F es menor de 0.0001, lo que lleva al rechazo de la hipótesis nula. Aún así, se muestra la estimación de los parámetros del modelo, para la respuesta índice de peróxido, obtenidos por mínimos cuadrados:

Parámetro	*Estimación*	*Error estándar*	*Estadístico*	*Valor p*
Intercepción	1.1923304	0.745824	1.60	0.1410
Temperatura	-0.021642	0.011467	-1.89	0.0884
C.Tierra	0.060762	0.174223	0.35	0.7345
Tiempo	-0.001363	0.008711	-0.16	0.8788
(temp)(temp)	0.0001114	0.000049	2.27	0.0466*
(C.Tierra)(temp)	-0.000731	0.001317	-0.56	0.5911
(C.Tierra)(C.Tierra)	0.0026697	0.019632	0.14	0.8945
(Tiempo)(Temp.)	0.0000109	0.000066	0.17	0.8714
(Tiempo)(C.Tierra)	-0.000769	0.001317	-0.58	0.5725
(Tiempo)(Tiempo)	0.0000398	0.000049	0.81	0.4361

Tabla 5.12

Se puede ver que solamente el cuadrado de la temperatura es significativo al nivel de $\alpha = 0.05$. La ecuación de regresión se formaría entonces de la siguiente manera, la cual tal vez no tiene mucho sentido.

$$Y = 0.0001114X_1^2$$

Para este diseño con tres factores e índice de peróxido como respuesta, no se muestran las gráficas de contorno respectivas de temperatura y tiempo, temperatura y cantidad de tierras y tiempo con cantidad de tierras, ni los perfiles de predicción, pues la solución obtenida corresponde a valores fuera del rango en algunos de los factores, como se puede ver enseguida:

Solution

Variable	Critical Value	
Temperatura	87.359333	
Tierra	-3.388873	
Tiempo	-27.5957	
Solution is a	SaddlePoint	
Critical values outside data range		
Predicted Value at Solution		0.162843

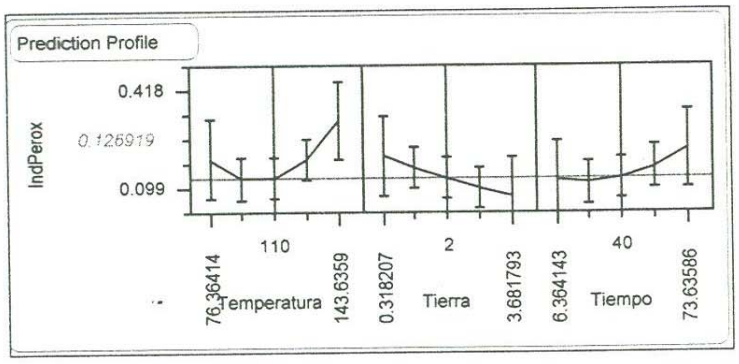
EigenStructure

EigenValues and EigenVectors

Variable	0.0028	0.0001	-0.0000
Temperatura	-0.13501	0.89798	0.41882
Tierra	0.98116	0.06221	0.18291
Tiempo	-0.13819	-0.43562	0.88946

Estos valores fueron obtenidos con el paquete estadístico JMPIN y se concluye que la solución es un punto silla, pues tiene valores propios positivos y negativos, y la respuesta predicha es de 0.162843. Solamente para el caso de temperatura se obtienen valores dentro del rango de experimentación. Pero hay que recordar que en el análisis de varianza, el modelo no fue significativo, el coeficiente de determinación es muy bajo y existe falta de ajuste.

En la gráfica 5.12 se puede observar, por separado, el efecto que tiene cada uno de los factores, sobre el índice de peróxido.

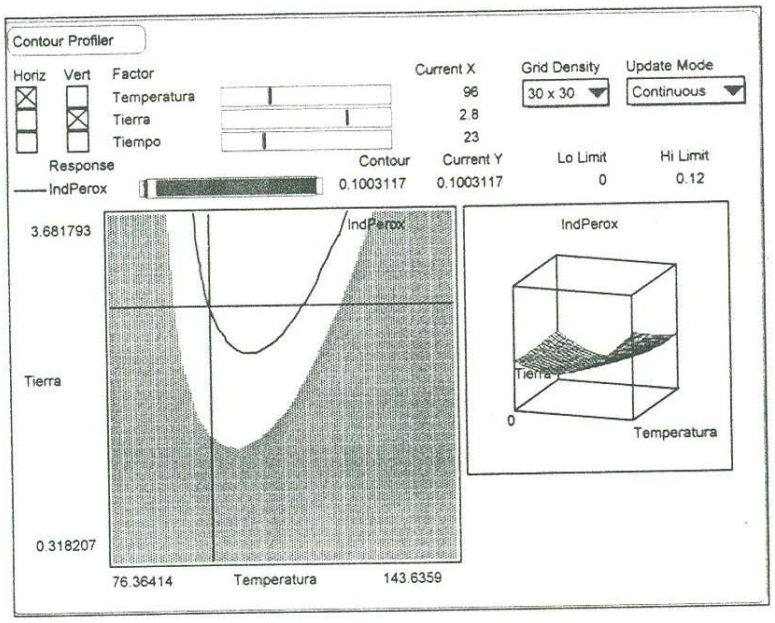


Gráfica 5.12

Recordemos que los rangos de operabilidad para temperatura son de 77°C a 143°C, en cantidad de tierras son de 0.318 g. a 3.68 g. por cada 200 gramos y para tiempo de 6.4 a 74 minutos. Así, los valores sugeridos como solución, son imposibles de alcanzar pues, la cantidad de tiempo y la cantidad de tierras son negativos.

Posteriormente se corrieron diseños utilizando dos factores y la variable respuesta, esto es, se analizó la influencia de temperatura y tiempo en el índice de peróxido, luego temperatura y cantidad de tierras y por último tiempo y cantidad de tierras. En los dos últimos casos los resultados obtenidos corresponden a valores fuera de rango en los factores o imposibles de alcanzar.

Ahora, experimentando con los mismos valores, validados en el laboratorio, para la retención de tocoferol, esto es, una temperatura de 96 grados centígrados, cantidad de tierra de 2.8% y un tiempo de contacto de 23 minutos, se tiene un índice de peróxido de aproximadamente 0.10. Esto se muestra en la siguiente gráfica, donde la parte no sombreada es la región donde el índice de peróxido se encuentra entre 0 y 0.12, y el contorno dibujado es para una respuesta de 0.10, aproximadamente.

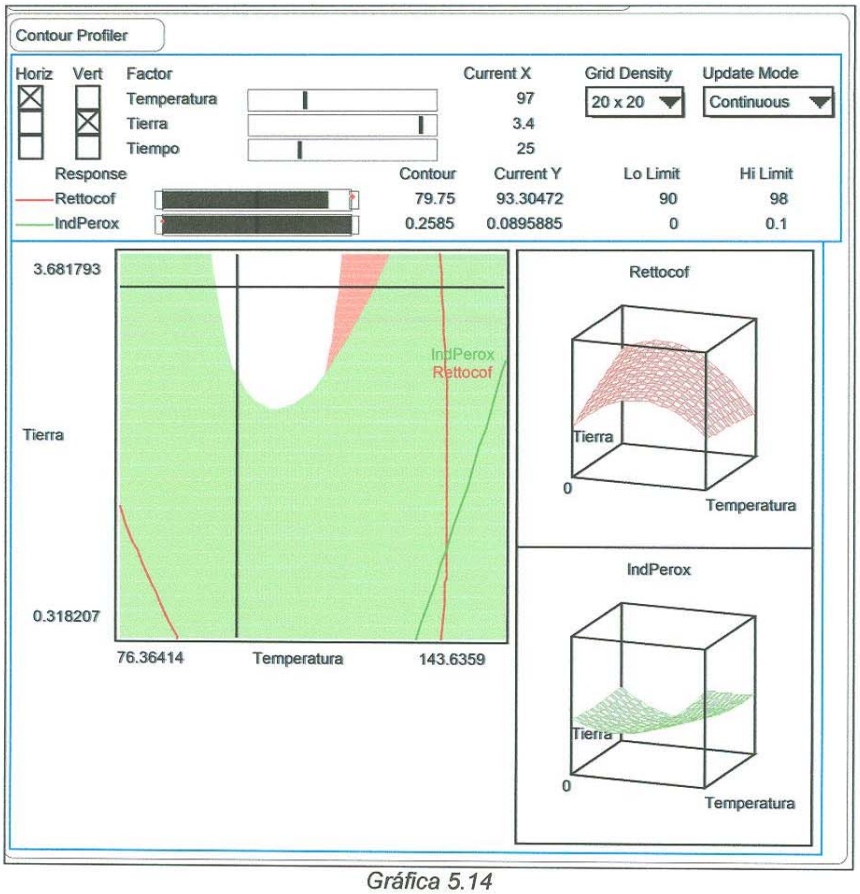


Gráfica 5.13

5.3 Método Gráfico.

Todo el análisis anterior se hizo por separado, hay que efectuarlo ahora en forma conjunta para ver si se puede encontrar una mejor solución considerando ambas respuestas, esto es, obtener menor índice de peróxido y mayor retención de tocoferol.

Enseguida se muestra el método gráfico para analizar varias superficies de respuesta. En este caso se consideraron solamente las dos respuestas ya mencionadas anteriormente. Fijando un tiempo de 25 minutos, pues el tiempo de contacto óptimo en muchos aceites se encuentra entre veinte y treinta minutos, se pidió delimitar la retención de tocoferol entre 90% y 98%, mientras que el índice de peróxido se desea que esté entre 0.0 y 0.1 (mEq/Kg de aceite). El sombreado verde delimita lo correspondiente al índice de peróxido, mientras que el rojo corresponde a la retención de tocoferol; dejando un área en blanco que corresponde al área que satisface las condiciones antes indicadas para ambas respuestas. La intersección mostrada en la región en blanco, corresponde a una temperatura de 97 grados centígrados, cantidad de tierras de 3.4 gramos (por cada 200 gramos), y conduce a una retención de tocoferol del 93.3%, con un índice de peróxido de 0.089 mEq/Kg.



Gráfica 5.14

Considerando que para el tiempo fijado, la temperatura debiera estar entre 90 a 100 grados centígrados, pues de lo contrario grandes tiempos de exposición del aceite con las tierras de blanqueo y las altas temperaturas, disminuyen la estabilidad oxidativa, las condiciones y respuestas obtenidas se consideran aceptables.

5.4 Conclusiones

Los factores que, por experiencia y revisión de la literatura especializada, se considera influyen en la retención de tocoferoles en la etapa de blanqueo, son: temperatura, cantidad de adsorbente (tierras) y tiempo de contacto.
También se piensa que estos factores influyen en el valor de los peróxidos, cuya cantidad se busca reducir. Es decir, en el blanqueo se desea eliminar peróxidos, también conocidos como productos de oxidación, pero con las condiciones óptimas para la retención de tocoferoles, pues lo que se busca hoy en día es maximizar este importante antioxidante natural.

Como se pudo ver, realizar el análisis estadístico por separado para cada respuesta, no es conveniente y nos conduce a diferentes soluciones, ya que cada una de ellas trata de optimizar solamente esa respuesta. Además, se

puede inferir que cuando se obtiene una solución que representa un punto silla, en una superficie de respuesta, muchas veces éste no corresponde al óptimo que se está buscando. Así pues, es recomendable explorar esta superficie, utilizando alguna herramienta computacional, para localizar la región de respuesta óptima. En el experimento desarrollado se presentó este problema en ambas respuestas y se siguió un procedimiento gráfico para explorar cada una de las superficies. Al analizarlas por separado, se puede llegar, como ya se mencionó, a resultados lejos del óptimo común. Se sugiere, entonces, analizar conjuntamente las respuestas, pues se tendrá una región de operabilidad que las optimice simultáneamente. Fue por lo tanto conveniente utilizar una metodología de multirrespuesta para encontrar ese óptimo común. El método que se utilizó aquí fue el gráfico, el cual es sencillo cuando solamente se tienen dos respuestas.

El investigador puede escoger, dentro de esta región, valores específicos para cada uno de los factores, y validarlos en el laboratorio. En la tabla siguiente se muestran los valores que el investigador escogió, dentro de esta región común de operabilidad, para mejorar sus respuestas, así como lo predicho por el modelo para éstas y lo obtenido en su validación:

Niveles de variables			*V. de Peróxidos*	*RTOCOLES*
X_1	X_2	X_3	*(mEq/kg.)*	*(%)*
96	2.8	23		
Predichas por el modelo			0.10	91.75
Validadas en el laboratorio			0.094 ± 0.04	91.79 ± 1.33

$$X_1 = Temperatura, X_2 = \%\ \ de\ \ tierras, X_3 = Tiempo\ \ de\ \ contacto$$

Para poder analizar gráficamente las dos respuestas y sus tres factores involucrados, se tuvo que fijar uno de ellos a cierto nivel para estudiar el comportamiento de los otros. El valor al que se fija este factor, depende de las condiciones de operabilidad, de la experiencia del investigador y también puede ser sugerido por análisis gráficos o experimentos previos. Por ejemplo, para este experimento, se podría sugerir validar lo siguiente en un experimento posterior.

Niveles de variables			*V. de Peróxidos*	*RTOCOLES*
X_1	X_2	X_3	*(mEq/kg.)*	*(%)*
97	3.4	25		
Valores predichos			0.089	93.30

Ya que al explorar en estas superficies descubrimos puntos que nos dan mejor retención de tocoferol dentro de los rangos del experimento, como una temperatura de 97 grados centígrados, una cantidad de tierras de 3.4 gramos (por cada 200 gramos), es decir 1.7% de cantidad de tierras, y un tiempo de 25

Se puede concluir, finalmente, que los resultados obtenidos fueron satisfactorios, pues la región experimental propuesta como viable para maximizar tocoferol y minimizar índice de peróxido, predice una pérdida de tocoferol menor al diez por ciento y un índice de peróxido menor que 1 mEq/kg, lo cual efectivamente se verificó en la validación realizada.

APÉNDICE A
TABLAS ESTADÍSTICAS

Tabla A.1 **Distribución Normal Estándar** $F(z) = P(Z \leq z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7157	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7969	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8513	.8554	.8577	.8529	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9215	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9492	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Tabla A.2 **Distribución *t* de Student** $F(t) = P(T \leq t)$

g. de l	.60	.70	.75	.80	.90	.95	.975	.99	.995	.9995
1	.325	.727	1.000	1.367	3.078	6.314	12.706	31.821	63.657	636.619
2	.289	.617	.816	1.061	1.886	2.920	4.303	6.965	9.925	31.598
3	.277	.584	.765	.978	1.638	2.353	3.182	4.541	5.841	12.924
4	.271	.569	.741	.941	1.533	2.132	2.776	3.747	4.604	8.61
5	.267	.559	.727	.920	1.476	2.015	2.571	3.365	4.032	6.869
6	.265	.553	.718	.906	1.440	1.943	2.447	3.143	3.707	5.959
7	.263	.549	.711	.896	1.415	1.895	2.365	2.998	3.499	5.408
8	.262	.546	.706	.889	1.397	1.860	2.306	2.896	3.355	5.041
9	.261	.543	.703	.883	1.383	1.833	2.262	2.821	3.250	4.781
10	.260	.542	.700	.879	1.372	1.812	2.228	2.764	3.169	4.587
11	.260	.540	.697	.876	1.363	1.796	2.201	2.718	3.106	4.437
12	.259	.539	.695	.873	1.356	1.782	2.179	2.681	3.055	4.318
13	.259	.538	.694	.870	1.350	1.771	2.160	2.650	3.012	4.221
14	.258	.537	.692	.868	1.345	1.761	2.145	2.624	2.977	4.140
15	.258	.536	.691	.866	1.341	1.753	2.131	2.602	2.947	4.073
16	.258	.535	.690	.865	1.337	1.746	2.120	2.583	2.921	4.015
17	.257	.534	.689	.863	1.333	1.740	2.110	2.567	2.898	3.965
18	.257	.534	.688	.862	1.330	1.734	2.101	2.552	2.878	3.922
19	.257	.533	.688	.861	1.328	1.729	2.093	2.539	2.861	3.883
20	.257	.533	.687	.860	1.325	1.725	2.086	2.528	2.845	3.850
21	.257	.532	.686	.859	1.323	1.721	2.080	2.518	2.831	3.819
22	.256	.532	.686	.858	1.321	1.717	2.074	2.508	2.819	3.792
23	.256	.532	.685	.858	1.319	1.714	2.069	2.500	2.807	3.767
24	.256	.531	.685	.857	1.316	1.708	2.060	2.485	2.787	3.745
25	.256	.531	.684	.856	1.316	1.708	2.060	2.485	2.787	3.725
26	.256	.531	.684	.856	1.315	1.706	2.056	2.479	2.779	3.707
27	.256	.531	.684	.855	1.314	1.703	2.052	2.473	2.771	3.690
28	.256	.530	.683	.855	1.313	1.701	2.048	2.467	2.763	3.674
29	.256	.530	.683	.854	1.310	1.697	2.042	2.457	2.750	3.659
30	.256	.530	.683	.854	1.310	1.697	2.042	2.457	2.750	3.646
40	.255	.529	.681	.851	1.303	1.684	2.021	2.423	2.704	3.551
60	.254	.527	.679	.848	1.296	1.671	2.000	2.390	2.660	3.460
120	.254	.526	.677	.845	1.289	1.658	1.980	2.358	2.617	3.373
∞	.253	.524	.674	.842	1.282	1.645	1.960	2.326	2.576	3.291

Tabla A.3 Porcentajes superiores de la distribución *F*

	áreas	1	2	3	4	5	6	7	8	9	10	11	12	15
1	.05	161	200	216	225	230	234	237	239	241	242	243	244	246
	.01	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106	6157
2	.05	18.51	19.00	19.16	19.25	19.39	19.33	19.36	19.37	19.38	19.39	19.4	19.4	19.4
	.01	98.49	99.01	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.4	99.4	99.4
3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.70
	.01	34.12	30.81	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.87
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.86
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.95	14.37	14.2
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.62
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89	9.72
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94
	.01	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.51
	.01	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.31
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.22
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.52
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.85
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.56
20	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.31	2.28	2.20
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.30	3.23	3.09
30	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.12	2.09	2.01
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.90	2.84	2.70
40	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52
50	.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.98	1.95	1.87
	.01	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.62	2.56	2.42
60	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.35
70	.05	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.82
	.01	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.32
80	.05	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.80
	.01	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41	2.28
100	.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.88	1.85	1.77
	.01	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.36	2.22
200	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.83	1.80	1.72
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.28	2.13
400	.05	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.70
	.01	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.23	2.08
1000	.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.80	1.76	1.68
	.01	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20	2.06
Inf	.05	3.85	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.66
	.01	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.03

A.4 Valores críticos del estadístico de Durbin-Watson

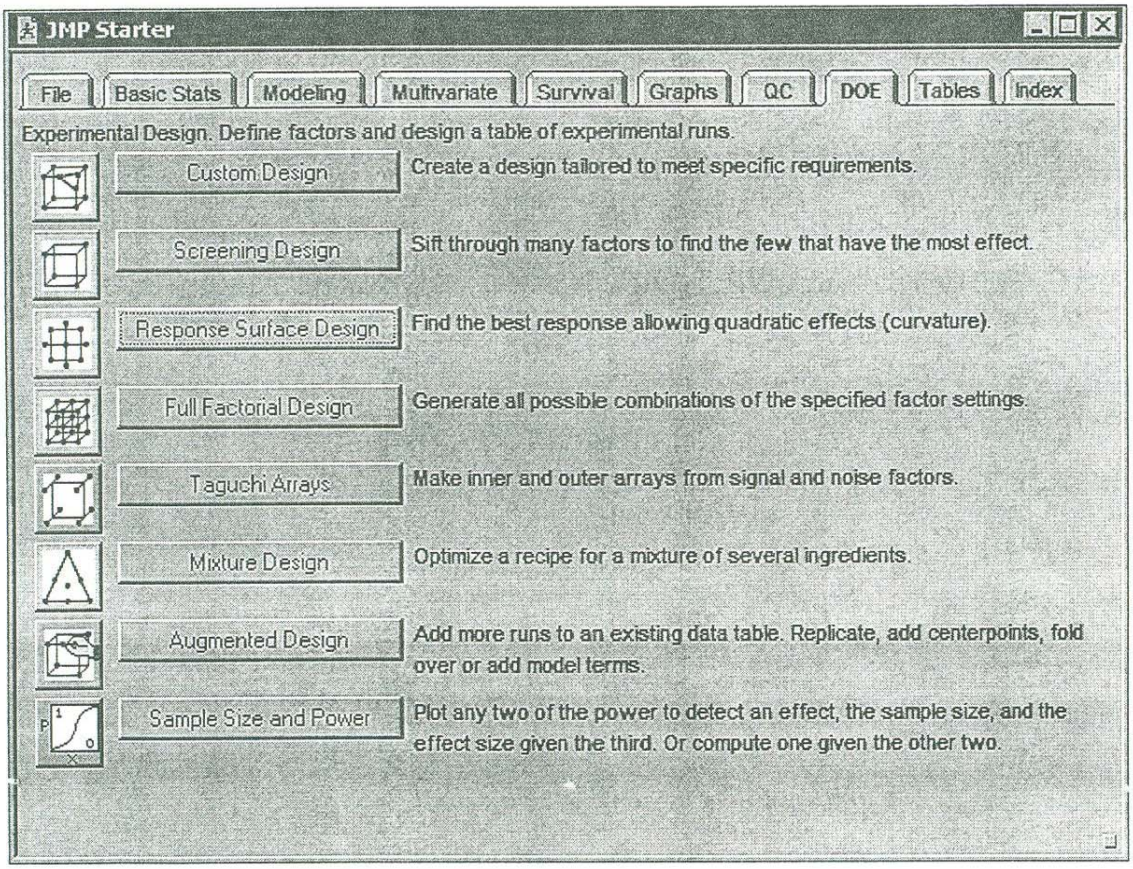
Tamaño de muestra	Probabilidad en la cola inferior (nivel de significancia = α)	k = Cantidad de regresores (excepto ordenada al origen)									
		1		2		3		4		5	
		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	.01	.81	1.07	.70	1.25	.59	1.46	.49	1.70	.39	1.96
	.025	.95	1.23	.83	1.40	.71	1.61	.59	1.84	.48	2.09
	.05	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
20	.01	.95	1.15	.86	1.27	.77	1.41	.63	1.57	.60	1.74
	.025	1.08	1.28	.99	1.41	.89	1.55	.79	1.70	.70	1.87
	.05	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
25	.01	1.05	1.21	.98	1.30	.90	1.41	.83	1.52	.75	1.65
	.025	1.13	1.34	1.10	1.43	1.02	1.54	.94	1.65	.86	1.77
	.05	1.20	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89
30	.01	1.13	1.26	1.07	1.34	1.01	1.42	.94	1.51	.88	1.61
	.025	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	.98	1.73
	.05	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	.01	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
	.025	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
	.05	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	.01	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
	.025	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
	.05	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	.01	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
	.025	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
	.05	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
80	.01	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
	.025	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
	.05	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
100	.01	1.52	1.56	1.50	1.58	1.48	1.60	1.45	1.63	1.44	1.65
	.025	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72
	.05	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

APÉNDICE B

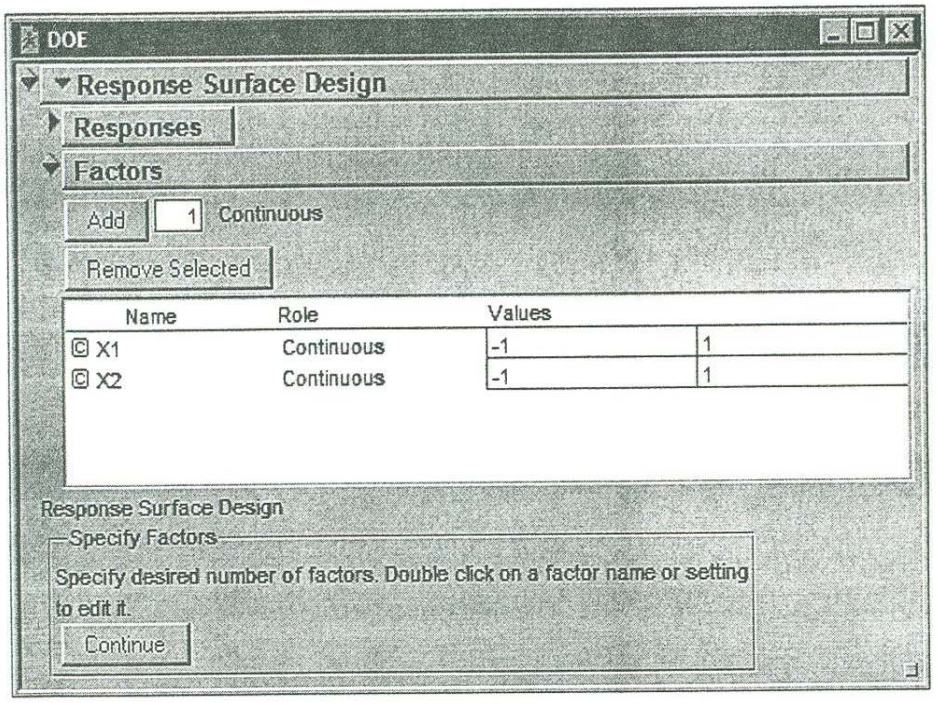
USO DE JMP IN EN SUPERFICIES DE RESPUESTA

Para analizar un diseño de superficie de respuesta en el JMP IN podemos seguir los pasos que a continuación se detallan:

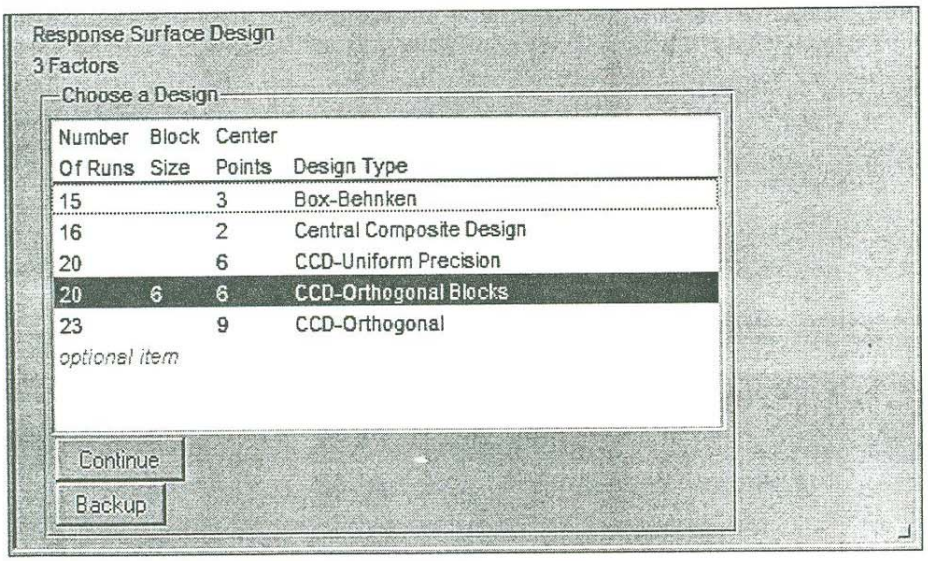
1. En el menú **JMP Starter** seleccionamos la opción que aparecerá en una pestaña con el nombre **DOE** (Design of Experiments), en donde escogemos **Response Surface Designs**, pues así indicamos que se desea un modelo de superficie de respuesta en JMP IN.



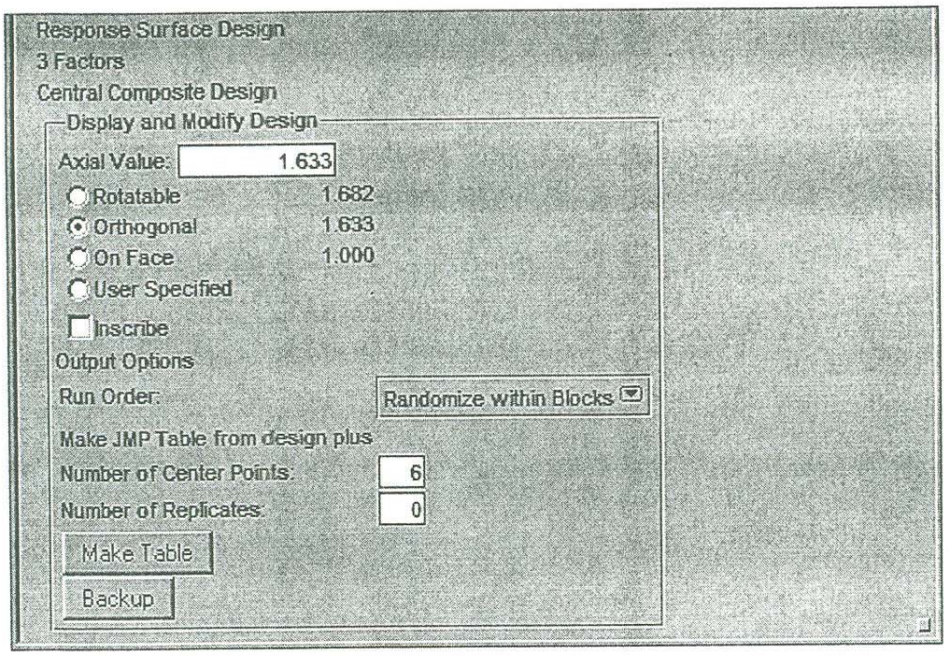
Al escoger esta opción se introduce el número de factores que vamos a analizar y sus nombres o etiquetas.



Al oprimir **Continue** aparecen los posibles diseños con diferentes características donde se escoge el que se piensa utilizar



y oprimiendo **Continue** se establece según los intereses del experimentador el valor de los puntos axiales.



Con **Make Table** se generará una hoja para la captura de los valores que toma la variable Respuesta (Y), a partir de los factores. Se puede llenar esta tabla, o bien establecer un orden para la captura de datos en este paso.

Central Composite Design

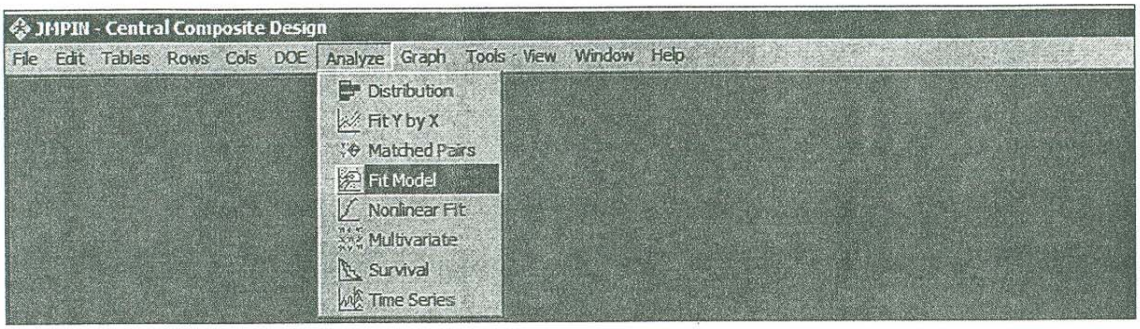
Central Composite Desi
Design Central Composite D
Model

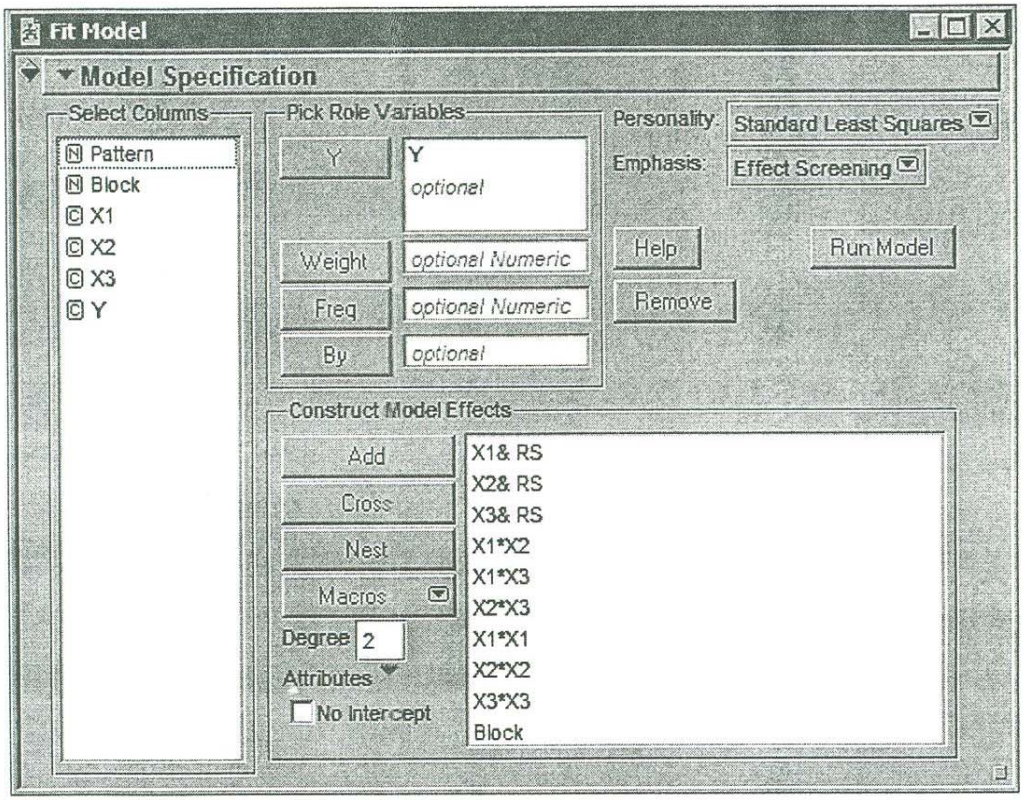
Columns (6/0)
Pattern
Block
X1
X2
X3
Y

Rows
All Rows 20
Selected 0
Excluded 0
Hidden 0
Labelled 0

	Pattern	Block	X1	X2	X3	Y
1	++-	1	1	1	-1	.
2	---	1	-1	-1	-1	.
3	000	1	0	0	0	.
4	000	1	0	0	0	.
5	+-+	1	1	-1	1	.
6	-++	1	-1	1	1	.
7	--+	2	-1	-1	1	.
8	000	2	0	0	0	.
9	+++	2	1	1	1	.
10	-+-	2	-1	1	-1	.
11	+--	2	1	-1	-1	.
12	000	2	0	0	0	.
13	000	3	0	0	0	.
14	0a0	3	0	-1.6329932	0	.
15	00a	3	0	0	-1.6329932	.
16	0A0	3	0	1.63299316	0	.
17	a00	3	-1.6329932	0	0	.
18	000	3	0	0	0	.
19	00A	3	0	0	1.63299316	.
20	A00	3	1.63299316	0	0	.

2. Una vez que los datos se han capturado, se recomienda salvar el diseño. Para analizar la superficie de respuesta, en el menú principal de JMP IN se escogerá **Analyze**, luego **Fit Model**





y aparecerá una pantalla donde en la parte izquierda se enlistan las columnas capturadas. Se envía a **Y** la columna correspondiente a la variable respuesta, luego se seleccionan los factores o variables independientes y del menú que aparece en **Macros** se escoge **Response Surface.** Aparte de este último menú aparecen en la misma pantalla un menú llamado **Personality** y otro de **Emphasis**, en ellos debe escogerse **Standard Least Squares** y **Effect Screening** respectivamente. Por último se oprime el icono **Run Model** y aparecerán los resultados del análisis.

APÉNDICE C

MATERIAL SUPLEMENTARIO

C.1 Prueba de Bartlett

La prueba de Bartlett es una de las técnicas más usadas para probar homogeneidad de varianza. En las tareas de la investigación científica es poco probable que al medir las observaciones en dos o más muestras poblacionales, la variación sea idéntica. Por lo general, esas varianzas tienen números diferentes y el investigador cae en la incertidumbre de decidir si las fuentes de error fueron las mismas o si intervinieron uno o más agentes de variación. La prueba de ji cuadrada de Bartlett permite saber, en función de la probabilidad, si la discrepancia entre varianzas fue dada por el azar o por otros factores de error no deseados por el experimentador. La varianza corresponde a la suma de las diferencias de los valores individuales en relación con el promedio, elevadas al cuadrado y divididas entre los grados de libertad, es decir, son variaciones alrededor de la medida de tendencia central, representativa de la muestra con la cual se estudia un fenómeno, sin embargo, no se puede saber si esas variaciones se deben a errores dados por el fenómeno en sí o a errores del observador o del método para efectuar las mediciones.

La hipótesis a plantear para la prueba de Bartlett son:

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2$

H_1 : al menos una σ_i^2 es diferente de las demás

El estadístico χ^2 utilizado en esta prueba se calcula de la siguiente manera:

$$\chi^2_{Bartlett} = 2.3026\frac{q}{c},$$

donde

$$q = (N-k)\log_{10} S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log_{10} S_i^2$$

y

$$c = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(n_i - 1)^{-1} - (N-k)^{-1}\right)$$

con

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{N-k}$$

Donde

$\chi^2_{Bartlett}$ = valor estadístico de esta prueba.

n = tamaño de la muestra del grupo,
k = número de grupos participantes
N = tamaño total (sumatoria de las muestras).

Bajo el supuesto de que la hipótesis nula sea cierta, el estadístico tiene una distribución aproximadamente χ^2 con $k-1$ grados de libertad, cuando el muestreo se realiza en poblaciones normales. Para todo valor de probabilidad menor que el nivel de significancía establecido en la prueba α, se rechazará H_0.

C.2 Prueba de Shapiro-Wilks

Es un contraste adecuado para muestras aleatorias pequeñas *(n<30)*. Se basa en el estudio del ajuste de los datos observados en la muestra a una recta dibujada en papel probabilístico normal. Si la hipótesis nula, de normalidad en la población, es cierta, entonces se puede afirmar que los valores muestrales provienen de una distribución $N(\mu,\sigma^2)$ y sus correspondientes valores tipificados y ordenados, serían una muestra ordenada procedente de una N(0,1).

Para la prueba de Shapiro-Wilks, se plantean las hipótesis

H_0:*los datos proceden de una distribución normal*

H_1:*los datos no proceden de una distribución normal*

y se necesita:

- Ordenar los datos de menor a mayor los cuales se denotan los datos ordenados por $x_1, x_2, ..., x_n$,
- Obtener los coeficientes $a_1, a_2, ..., a_k$ siendo $a_i = \dfrac{c_i}{\sum_{i=1}^{k} c_i^2}$, donde $k = n/2$, si n es par, y $k = (n-1)/2$ si n es impar,
- Se calcula después el estadístico W definido por:

$$W = \frac{1}{(n-1)S^2}\left[\sum_{i=1}^{k} a_i (x_{n-i+1} - x_i)\right]^2$$

donde S^2 es la varianza muestral. Los valores de los a_i y del estadístico W están tabulados.

- Finalmente la hipótesis de normalidad será rechazada cuando el valor W observado en la muestra, sea menor que el valor crítico dado en la tabla, puesto que no hemos de olvidar que se está midiendo la bondad del ajuste a la recta y no la discrepancia con la hipótesis nula de normalidad. En escala probabilística normal se representa en el eje horizontal, para cada valor observado en nuestros datos, la función de distribución o probabilidad acumulada observada, y en el eje vertical la prevista por el modelo de distribución normal. Si el ajuste es bueno, los puntos se deben distribuir aproximadamente según una recta a cuarenta y cinco grados.

C.3 Estimación por Máxima Verosimilitud de los Parámetros del Modelo de Regresión Lineal Simple

Si se conoce la forma de distribución de los errores, un método alternativo para estimar parámetros es este método, también llamado máxima verosimilitud.
Se tienen los datos $(y_i, x_i), i = 1,2,...,n$. Si se supone que los errores en el modelo de regresión son $NID(0,\sigma^2)$, las observaciones y_i en esa muestra son variables aleatorias normal e independientemente distribuidas (NID) con promedio $\beta_0 + \beta_1 x_i$, y varianza σ^2. La función de verosimilitud o de posibilidad se determina con la distribución conjunta de las observaciones. Si se considera esta distribución conjunta con las observaciones dadas y los parámetros β_0, β_1 y σ^2 son constantes desconocidas, se tiene la función de posibilidad. Para el modelo de regresión lineal simple con errores normales, la función de posibilidad es:

$$L(y_i, x_i, \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

Los estimadores de máxima posibilidad son los valores de los parámetros, por ejemplo $\tilde{\beta}_0, \tilde{\beta}_1$ y $\tilde{\sigma}^2$, que maximizan a L, o lo que es lo mismo a $\ln L$. Así,

$$\ln L(y_i, x_i, \beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right)\ln 2\pi - \left(\frac{n}{2}\right)\ln\sigma^2 - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

y los estimadores de posibilidad máxima, $\tilde{\beta}_0, \tilde{\beta}_1$ y $\tilde{\sigma}^2$ deben satisfacer

$$\left.\frac{\partial \ln L}{\partial \beta_0}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial \ln L}{\partial \beta_1}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)x_i = 0$$

y

$$\left.\frac{\partial \ln L}{\partial \sigma^2}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 = 0$$

La solución de las ecuaciones anteriores determina los estimadores de máxima verosimilitud:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n}$$

Obsérvese que los estimadores de máxima verosimilitud de la ordenada al origen y la pendiente, $\tilde{\beta}_0, \tilde{\beta}_1$, son idénticos a los obtenidos con los mínimos cuadrados y también, $\tilde{\sigma}^2$ es un estimador sesgado de σ^2.

C.4 Propiedades de los estimadores de mínimos cuadrados del modelo de regresión lineal simple.

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ por mínimos cuadrados son **estimadores insesgados** de los parámetros β_0 y β_1 del modelo. Para demostrarlo con $\hat{\beta}_1$, considérese

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i)$$

$$= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

ya que, se supone que, $E(\varepsilon_i) = 0$. Ahora se puede demostrar en forma directa que $\sum_{i=1}^{n} c_i = 0$ y que $\sum_{i=1}^{n} c_i x_i = 1$, entonces

$$E(\hat{\beta}_1) = \beta_1$$

Esto es, si se supone que el modelo es correcto $[que\ E(y_i) = \beta_0 + \beta_1 x_1]$, entonces $\hat{\beta}_1$ es un estimador insesgado de β_1. De igual manera se puede demostrar que $\hat{\beta}_0$ es un estimador insesgado de β_0, es decir,

$$E(\hat{\beta}_0) = \beta_0$$

La varianza de $\hat{\beta}_1$ se calcula como sigue:

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i y_i\right)$$

$$= \sum_{i=1}^{n} c_i^2 Var(y_i)$$

ya que las observaciones y_i son no correlacionadas, por lo que la varianza de la suma es igual a la suma de las varianzas. La varianza de cada término en la suma es $c_i^2 Var(y_i)$ y hemos supuesto que $Var(y_i) = \sigma^2$; en consecuencia,

$$Var(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2}$$
$$= \frac{\sigma^2}{S_{xx}}$$

La varianza de $\hat{\beta}_0$ es

$$Var(\hat{\beta}_1) = Var(\bar{y} - \hat{\beta}_1 \bar{x})$$
$$= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1)$$

Ahora bien, la varianza de $\bar{y}$ no es más que $Var(\bar{y}) = \sigma^2/n$, y se puede demostrar que la covarianza entre $\bar{y}$ y $\hat{\beta}_1$ es cero. Así,

$$Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1)$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

C.5 Propiedades de las Sumas de Cuadrados en el Modelo de Regresión Lineal.

A continuación se presentan algunos resultados importantes acerca de SC_R y SC_E

- En el caso de SC_R

Por definición,

$$SC_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Se observa que $\hat{y} = X(X^T X)^{-1} X^T y$, y que

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} 1^T y$$

donde 1 es un vector de $nx1$, con todos sus elementos iguales a unos. Además, $n = 1^T 1$, y en consecuencia $\bar{y} = (1^T 1)^{-1} 1^T y$. Por consiguiente, se puede escribir SC_R en la forma

$$SC_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
$$= [\hat{y} - 1\bar{y}]^T [\hat{y} - 1\bar{y}]$$

$$= \left[X\left(X^T X\right)^{-1} X^T y - 1\left(1^T 1\right)^{-1} 1^T y\right]^T \left[X\left(X^T X\right)^{-1} X^T y - 1\left(1^T 1\right)^{-1} 1^T y\right]$$
$$= y^T \left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right]^T \left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right] y$$

Nótese que $X = [1X_R]$, siendo X_R la matriz formada por los valores reales de los regresores, por lo que SC_R es un caso especial de una matriz dividida. Por lo anterior, se puede usar la identidad especial para matrices divididas, para demostrar que

$$X\left(X^T X\right)^{-1} X^T 1 = 1, \qquad \text{y} \qquad 1^T X\left(X^T X\right)^{-1} X^T = 1^T$$

Por consiguiente, se puede demostrar que $\left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right]$ es idempotente. De acuerdo con la premisa de que $Var(\varepsilon) = \sigma^2 I$,

$$\frac{SC_R}{\sigma^2} = \frac{1}{\sigma^2} y^T \left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right] y$$

sigue una distribución χ^2 no central, con parámetro λ de no centralidad, y con grados de libertad iguales al rango de $\left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right]$ Como esta matriz es idempotente, su rango es igual a su traza. Se observa que

$$\begin{aligned} Traza\left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right] &= traza\left[X\left(X^T X\right)^{-1} X^T\right] - traza\left[1\left(1^T 1\right)^{-1} 1^T\right] \\ &= traza\left[X^T X\left(X^T X\right)^{-1}\right] - traza\left[1^T\left(1^T 1\right)^{-1} 1\right] \\ &= traza\left(I_p\right) - traza(1) \\ &= p - 1 = k \end{aligned}$$

Suponiendo que el modelo es correcto,

$$E(y) = X\beta = [1\beta_R]\begin{bmatrix}\beta_0 \\ \beta_R\end{bmatrix} = \beta_0 1 + X_R \beta_R$$

Así, el parámetro de no centralidad es

$$\lambda = \frac{1}{\sigma^2} E(y)^T \left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right] E(y)$$

$$= \frac{1}{\sigma^2} \beta^T X^T \left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right] X\beta$$

$$= \frac{1}{\sigma^2} \left[\beta_0 \quad \beta_R^T \begin{bmatrix} 1^T \\ X_R^T \end{bmatrix} \left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^{-1} 1^T\right] \begin{bmatrix} 1 & X_R \end{bmatrix} \begin{bmatrix}\beta_0 \\ \beta_R\end{bmatrix}\right.$$

$$= \frac{1}{\sigma^2} \left[\beta_0 \quad \beta_R^T \begin{bmatrix} 1^T X\left(X^T X\right)^{-1} X^T - 1^T 1\left(1^T 1\right)^{-1} 1^T \\ X_R^T X\left(X^T X\right)^{-1} X^T - X_R^T 1\left(1^T 1\right)^{-1} 1^T \end{bmatrix} \begin{bmatrix} 1 & X_R \end{bmatrix} \begin{bmatrix}\beta_0 \\ \beta_R\end{bmatrix}\right.$$

$$= \frac{1}{\sigma^2} \left[\beta_0 \quad \beta_R^T \begin{bmatrix} 0^T \\ X_R^T - X_R^T 1\left(1^T 1\right)^{-1} 1^T \end{bmatrix} \begin{bmatrix} 1 & X_R \end{bmatrix} \begin{bmatrix}\beta_0 \\ \beta_R\end{bmatrix}\right.$$

$$= \frac{1}{\sigma^2}\left[\beta_0 \quad \beta_R^T\right]\begin{bmatrix} 0 & 0^T \\ 0 & X_R^T X_R - X_R^T 1\left(1^T 1\right)^{-1} 1^T X_R \end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_R\end{bmatrix}$$

$$= \frac{1}{\sigma^2}\beta_R^T\left[X_R^T X_R - X_R^T 1\left(1^T 1\right)^{-1} 1^T X_R\right]\beta_R$$

Si se define como sigue la matriz de los valores de los regresores **centrados** X_C:

$$X_C = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_1 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

siendo $\bar{x}_1$ el valor promedio del primer regresor, $\bar{x}_2$ el valor promedio del segundo regresor, y asi sucesivamente, se puede demostrar con facilidad que el parámetro de no centralidad se expresa como sigue:

$$\lambda = \frac{1}{\sigma^2}\beta_R^T\left[X_C^T X_C\right]\beta_R$$

El valor esperado de SC_R es

$$\begin{aligned} E(SC_R) &= E\left(y^T\left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^T 1\right]y\right) \\ &= traza\left(\left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^T 1\right]\sigma^2 I\right) \\ &\quad + E(y)^T\left[X\left(X^T X\right)^{-1} X^T - 1\left(1^T 1\right)^T 1\right]E(y) \\ &= k\sigma^2 + \beta_R^T X_C^T X_C \beta_R \end{aligned}$$

Por consiguiente,

$$E(CM_R) = E\left(\frac{SC_R}{k}\right) = \sigma^2 + \frac{\beta_R^T X_C^T X_C \beta_R}{k}$$

- Ahora para la SC_E

Por definición,

$$SC_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Se ve que SC_E se puede escribir en la forma siguiente:

$$\begin{aligned} SC_E &= (y - \hat{y})^T (y - \hat{y}) \\ &= \left[y - X\left(X^T X\right)^{-1} X^T y\right]^T\left[y - X\left(X^T X\right)^{-1} X^T y\right] \\ &= y^T\left[I - X\left(X^T X\right)^{-1} X^T\right]y \end{aligned}$$

La demostración que $\left[I - X\left(X^T X\right)^{-1} X^T\right]$ es simétrica e idempotente es trivial; en consecuencia

$$\frac{SC_E}{\sigma^2} = \frac{1}{\sigma^2} y^T \left[I - X\left(X^T X\right)^{-1} X^T \right] y$$

sigue una distribución χ^2. Los grados de libertad se deben al rango de $\left[I - X\left(X^T X\right)^{-1} X^T \right]$ que a su vez es su traza. La demostración de que la traza es $n-p$ es directa. Bajo la premisa que el modelo es correcto,

$$E(y) = X\beta .$$

Así, el parámetro de no centralidad es

$$\begin{aligned} &\frac{1}{\sigma^2} E(y)^T \left[I - X\left(X^T X\right)^{-1} X^T \right] E(y) \\ &= \frac{1}{\sigma^2} \beta^T X^T \left[I - X\left(X^T X\right)^{-1} X^T \right] X\beta \\ &= \frac{1}{\sigma^2} \beta^T \left[X^T X - X^T X\left(X^T X\right)^{-1} X^T X \right] \beta = 0 \end{aligned}$$

Y como resultado,

$$\frac{SC_E}{\sigma^2} \sim \chi^2_{n-p}$$

El valor esperado de SC_E es

$$\begin{aligned} E(SC_E) &= E\left(y^T \left[I - X\left(X^T X\right)^{-1} X^T \right] y \right) \\ &= traza\left(\left[I - X\left(X^T X\right)^{-1} X^T \right] \sigma^2 I \right) + E(y)^T \left[I - X\left(X^T X\right)^{-1} X^T \right] E(y) \\ &= (n-p)\sigma^2 \end{aligned}$$

Y como resultado,

$$E(CM_E) = E\left(\frac{SC_E}{n-p} \right) = \sigma^2$$

C.6 La prueba F global o general

Un estadístico F es la razón de dos variables independientes χ^2, cada una dividida entre sus respectivos grados de libertad. Aquí se ha demostrado que tanto SC_R / σ^2 como SC_E / σ^2 siguen distribuciones χ^2; lo fundamental ahora es demostrar que son independientes. De acuerdo con la teoría básica de los modelos lineales, SC_R y SC_E son independientes, suponiendo que $Var(\varepsilon) = \sigma^2 I$, si

$$\left[X\left(X^T X\right)^{-1} X^T - 1\left(I^T I\right)^T 1 \right] \sigma^2 I \left[I - X\left(X^T X\right)^T X \right] = 0$$

Se observa que

$$
\begin{aligned}
&\left[X(X^TX)^{-1}X^T - 1(1^T1)^T1\right]\sigma^2 I\left[I - X(X^TX)^TX\right] \\
&= \sigma^2\left[X(X^TX)^{-1}X^T - 1(1^T1)^T1\right]\left[I - X(X^TX)^TX\right] \\
&= \sigma^2\left[X(X^TX)^{-1}X^T - 1(1^T1)^T1 - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T + 1(1^T1)^T1^TX(X^TX)^{-1}X^T\right] \\
&= X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^T - 1(1^T1)^T1^T + 1(1^T1)^T1^T = 0
\end{aligned}
$$

Por consiguiente, SC_R y SC_E son independientes. A continuación se observa que

$$\frac{SC_R}{k\sigma^2} = \frac{CM_R}{\sigma^2} \quad \text{y} \quad \frac{SC_E}{(n-p)\sigma^2} = \frac{CM_E}{\sigma^2}$$

son variables aleatorias χ^2, cada una dividida entre su grado de libertad respectivo. Entonces

$$\frac{CM_R}{CM_E} \sim F'_{k,n-p,\lambda}$$

siendo

$$\lambda = \frac{1}{\sigma^2}\beta'_R X'_C X_C \beta_R$$

En el caso especial de las regresiones lineales simples sólo se tiene un regresor; entonces $\beta_R = \beta_1$ y

$$X'_C X_C = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Como resultado, para la regresión lineal simple,

$$\frac{CM_R}{CM_E} \sim F_{1,n-2,\lambda}$$

siendo $\lambda = \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$.

C.7 Teoremas de Álgebra y Análisis Lineal

Teorema 4.1

Si $f: R^3 \to R$ es diferenciable, entonces todas las derivadas direccionales existen. La derivada direccional en x en la dirección de v está dada por

$$Df(x)v = gradf(x)\cdot v = \nabla f(x)\cdot v = \left[\frac{\partial f}{\partial x}(x)\right]v_1 + \left[\frac{\partial f}{\partial y}(x)\right]v_2 + \left[\frac{\partial f}{\partial z}(x)\right]v_3$$

donde $v = (v_1, v_2, v_3)$.

Demostración: Sea $c(t) = x + tv$, de manera que $f(x + tv) = f(c(t))$. Por un caso particular de la regla de la cadena, $(d/dt) f(c(t)) = \nabla f(c(t)) \cdot c'(t)$. Sin embargo, $c(0) = x$ y $c'(0) = v$ y entonces

$$\frac{d}{dt} f(x+tv)_{t=0} = \nabla f(x) \cdot v$$

Teorema 4.2

Supongamos que $\nabla f(x) \neq 0$. Entonces $\nabla f(x)$ apunta en la dirección a lo largo de la cual f crece más rápidamente.

Demostración: Si n es un vector unitario, la razón de cambio de f en la dirección n está dada por $\nabla f(x) \cdot n = |\nabla f(x)| \cos\theta$, donde θ es el ángulo entre n y $\nabla f(x)$. Éste es máximo cuando $\theta = 0$; esto es, cuando n y ∇f son paralelos. Si $\nabla f(x) = 0$, esta razón de cambio es 0 para cualquier n.

Teorema de Ejes Principales en R^2.

Sea $ax^2 + bxy + cy^2 = d$, una ecuación cuadrática en las variables x y y. Entonces existe un único número θ en $[0, 2\pi)$ tal que la ecuación mencionada puede ser escrita en la forma

$$a'x'^2 + c'y'^2 = d$$

donde x', y' son los ejes obtenidos al rotar los ejes x y y un ángulo de θ en el sentido de las manecillas del reloj. Además, los números a' y c' son los valores propios de la matriz $A = \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$. Los ejes x' y y' son llamados los *ejes principales* de la gráfica de la ecuación cuadrática.

Bibliografía

[1] Bockisch, M. (1998). *Fats and oils handbook*. Champaign, IL: AOCS Press, 29-35.

[2] Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: John Wiley & Sons.

[3] Box, G. E. P. & Behnken, D. W. (1960). *Some new three level designs for the study of quantitative variables. Technometrics* 2, 455-476.

[4] Box, G. E. P. & Hunter, J. S. *Multifactor experimental designs for exploring response surfaces. Annals of Mathematical Statistics.* (vol. 28),195-242.

[5] Box, G. E. P., Hunter, J. S. y Hunter, W. G. (,1998). *Estadística para investigadores*. Barcelona, España: Reverté.

[6] Box, G. E. P. , Hunter, J. S. & Hunter, W. G. (1978). *Statistics for experimenters*. New York: John Wiley & Sons.

[7] Box, G. E. P. y Wilson, K. G. (1951). *On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, B* 13, 1-45.

[8] Buyske, S. & Trout, R. (2001). *Advanced design of experiments*. Rutgers University. Retrieved, from, http://www.stat.rutgers.edu/~buyske/591/lect07.pdf

[9] Cornell, J. A. (1984). *How to apply response surface methodology*. . Milwaukee, WI: American Society for Quality Control.

[10] De La Vara , R. & Domínguez, J. (1998). *Metodología de Superficie de Multirrespuesta*. Articulo publicado en Guanajuato, Gto: CIMAT.

[11] Draper, N. R. & Smith H. (1981). *Applied regression analysis*. (2nd ed.). New York: John Wiley & Sons.

[12] Dudewicz, E. J. & Mishra, S. N. (1988). *Modern mathematical statistics*. New York: John Wiley & Sons.

[13] Durbin, J. & Watson, G. S. (1951). *"Testing for serial correlation in least square regression II", Biometrika,* **38,** 159-178.

[14] Engineering statistics handbook. Retrieved January 17, 2003, from, http://www.itl.nist.gov/div898/handbook/pri/section3/pri3361.htm

[15] Erickson, D. R. (1995). *Practical handbook of soybean processing and utilization*. Champaign, IL: AOCS Press, 186, 187, 219.

[16] Figueroa, G. (Mayo 2002). *Optimización de una superficie de respuesta utilizando JMP IN*. Arenario. (Vol. 2), Número 2. Hermosillo, Sonora: Universidad de Sonora.

[17] Freund, J. E., Miller, I., & Miller, M. (2000). *Estadística matemática con aplicaciones*. (6ª ed.). Pearson Education.

[18] Grossman, Stanley I. (1980). *Elementary linear algebra*. (2ª ed.). Wadsworth.

[19] Gutierrez Pulido, H & Vara Salazar, R. (2003). *Análisis y diseño de experimentos*. México: McGraw-Hill/Internamericana.

[20] Hogg, V. R. & Craig A. T. (1978). *Introduction to mathematical statistics*. (4th ed.). New York : Macmillan Publishing.

[21] Khuri, A. I. & Cornell, J. A. (1997). *Response surfaces: Designs and Analyses*. (2nd ed.). New York: Marcel Dekker.

[22] Kuehl, R. O. (2001). *Diseño de Experimentos*. (2ª ed.). Thomson Learning.

[23] Marsden, J. E. & Tromba, A. J.(1998). *Cálculo vectorial*. (4ª ed). Pearson Education.

[24] Medina-Juárez L. A, et al.(1998). *Effects of processing on the oxidative stability of soybean oil produced in Mexico*. J. Am. Oil Chem, 75(12):1729-1733.

[25] Medina-Juárez L. A, et al. (2000). *Trans fatty acid composition and tocopherol content in vegetable oils produced in Mexico*. J. Am. Oil Chem, 77(7): 721-724.

[26] Montgomery, D. C.(2002). Diseño y Análisis de Experimentos.(2ª ed). Editorial Limusa Wiley.

[27] Montgomery, D. C.(1991). *Diseño y Análisis de Experimentos*. Grupo Editorial Iberoamérica.

[28] Montgomery, D. C., Peck, E.A. & Vining, G. (2002). *Introducción al Análisis de Regresión Lineal*. (3ª Edición). Cd. de México: CECSA, México.

[29] Myers, R. H., & Montgomery, D. C.(1995). *Response surface methodology: Process and product optimization using designed experiments*. New York: John Wiley & Sons.

[30] O´Brien, R. F., Farr, W. E., & Wan, P. J. (2000). *Introductions to fats and Oils technology*. (2nd ed.). Champaign, IL: AOCS, Press, 158-178.

[31] Ortrega G., J. (2002). *"Determinación de las Condiciones Óptimas de Blanqueo y su Influencia Sobre la Calidad y Contenido de Tocoferoles en Aceite de Soya"*.

[32] Response surface methodology. Retrieved April 23, 2002, from, http://www.mne.psu.edu./me82/Learning/RSM/rsm.html

[33] Ruiz-Maya, L. y Martín Pliego, F. J. (2002). *Estadística II: Inferencia*. (2ª ed.). International Thomson Editores Spain Paraninfo.

[34] Villa , V. M., (1996). *Notas del curso internacional sobre refinación de aceites vegetales*. Hermosillo, Sonora: Universidad de Sonora, 30-33.

[35] Villa , V. M., (2001). *Notas del curso sobre refinación de aceites de soya de alta calidad para uso en productos alimenticios*. Hermosillo, Sonora: Universidad de Sonora, 1-12.

[36] Ziller, S., (1996). *Grasas y aceites alimentarios*. (7ª ed.). Zaragoza, España: Acribia, 4,37.