CAPÍTULO 2

Metodología de superficie de respuesta y Modelo de regresión logística: Aplicación a conjuntos de datos reales sobre la fabricación de circuitos integrados

https://doi.org/10.22533/at.ed.1232517102

René Castro Montoya
Facultad de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa.
https://orcid.org/0000-0002-6746-7559

José Vidal Jiménez Ramírez
Facultad de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa
https://orcid.org/0000-0001-6747-0144

RESUMEN: La Metodología Superficie de Respuesta (MSR) puede definirse como una estrategia que engloba los siguientes puntos: elegir un diseño experimental que permita medir adecuadamente el comportamiento de la respuesta de interés; determinar un modelo que describa el comportamiento de los datos obtenidos mediante el diseño experimental, lo que implica hacer algunas pruebas estadísticas para verificar si el modelo es adecuado. Una vez que se tiene un modelo adecuado se procede a encontrar la combinación de los niveles de los factores de entrada que producen la respuesta óptima. La MSR tuvo su origen en los años 30's en trabajos realizados por Sisar, Yates, y otros, sin embargo, ésta fue desarrollada formalmente por Box y Wilson (1951). El objetivo del presente trabajo es describir tres aplicaciones de MSR a conjuntos de datos reales: el primero es un experimento realizado en una industria electrónica mexicana; cuyo objetivo es disminuir el número de dispositivos electrónicos que se rompen en cierta etapa de su proceso de fabricación, debido a los cambios bruscos de temperatura que allí ocurren, el problema es que algunas obleas no resisten dichos cambios de temperatura y se rompen (Castro Montoya, 1995). Una oblea de silicio es un dispositivo electrónico en el que vienen integrados microcircuitos para ser procesados juntos, es decir, la oblea es el medio que permite procesar al mismo tiempo cientos de dados o chips, lo cuales deben cumplir ciertas propiedades eléctricas. El análisis mediante el modelo de regresión múltiple y el

modelo de regresión logística tuvieron la capacidad para detectar los mismos efectos. Esto se debe tal vez a que un número grande de obleas procedas en cada tratamiento.

PALABRAS CLAVE: Diseño y análisis de experimentos, Modelos, Metodología superficie de respuesta y Pruebas de hipótesis.

Response surface methodology and logistic regression model: Application to real datasets on integrated circuit manufacturing

ABSTRACT: The Response Surface Methodology (MSR) can be defined as a strategy that encompasses the following points: choosing an experimental design that allows to adequately measure the behavior of the response of interest; determine a model that describes the behavior of the data obtained through the experimental design, which implies doing some statistical tests to verify if the model is adequate. Once we have an adequate model, we proceed to find the combination of the levels of the input factors that produce the optimal response. MSR had its origin in the 1930s in work by Sisar, Yates, and others, however, it was formally developed by Box and Wilson (1951). The objective of this work is to describe three applications of MSR to real data sets: the first is an experiment carried out in a Mexican electronics industry; whose objective is to reduce the number of electronic devices that break at a certain stage of their manufacturing process, due to the sudden changes in temperature that occur there, the problem is that some wafers do not resist these changes in temperature and break (Castro Montoya, 1995). A silicon wafer is an electronic device in which microcircuits are integrated to be processed together, that is, the wafer is the medium that allows hundreds of dice or chips to be processed at the same time, which must meet certain electrical properties.

KEYWORDS: Design and analysis of experiments, Models, Response surface methodology and Hypothesis tests.

INTRODUCCIÓN

En este trabajo se presentan los resultados de aplicar Metodología Superficie de Respuesta a tres conjuntos de datos reales. En el primer caso se presenta una aplicación del diseño de experimentos en la industria electrónica mexicana. Algunos aspectos que hacen interesante este experimento son: 1) las consideraciones de ingeniería de proceso que se hicieron previamente, 2) las diferentes alternativas de análisis estadístico, por ser la respuesta una variable binaria y 3) el ahorro económico obtenido. La compañía está interesada en determinar los niveles de los factores que minimizan el número de obleas rotas. Los factores que se controlan son temperatura

de grabado, temperatura de piraña y temperatura de agua. El proceso se realizaba antes del experimento a una temperatura de grabado de -3 °C, una temperatura de piraña de 98 °C y una temperatura de agua de 20 °C y se tenía un rendimiento mecánico del 97% en la solución piranha. Puesto que no se sabe que tan cerca puedan estar los niveles usuales (-3 °C, 98 °C, 20 °C) de los niveles que producen la respuesta óptima, los ingenieros se deciden por correr un diseño factorial 23, con el objetivo de localizar el tratamiento mediante el cual se obtenga un mejor rendimiento mecánico. En cada tratamiento se utilizaron 500 obleas, y se obtienen solo ocho puntos. Debido a que se tienen pocos puntos no es posible detectar si los supuestos de independencia y varianza constante se cumplen, por lo que se decidió considerar sólo 250 obleas por tratamiento, se obteniendo 16 puntos, pues 250 obleas por tratamiento es suficiente para observar al menos una oblea rota por tratamiento.

METODOLOGÍA DE SUPERFICIE DE RESPUESTA

La Metodología de Superficie de Respuesta, inventada en 1951 por Box y Wilson, es un conjunto de estrategias de investigación, métodos matemáticos e inferencia estadística que permiten al investigador hacer exploración empírica eficiente en el proceso de su interés. Es un método estadístico que usa información cuantitativa de experimentos apropiados para determinar y resolver ecuaciones multivariantes, utilizado, la mayoría de las veces, para la optimización de procesos.

DISEÑOS Y MODELOS PARA SUPERFICIE DE RESPUESTA

La estrategia experimental y análisis en MSR se basa en el supuesto de que la verdadera respuesta η desconocida es una función $\varphi(x_1, x_2, \ldots, x_k)$ del conjunto de variables de diseño $x_1, x_2, \ldots, x_k$, y que la función puede ser aproximada en alguna región de las x_{i_s} por un polinomio de primero o segundo orden. En la práctica este supuesto es razonable si la respuesta observada es continua y suave, aunque su comportamiento no sea suave esta variable puede describirse con un polinomio de bajo orden si se escoge una región experimental lo suficientemente pequeña.

Por otra parte, un polinomio de grado *n* puede ser aproximado mediante una expansión en serie de Taylor de la verdadera función teórica fundamental $\varphi(x_1, x_2, \ldots, x_k)$ truncada después del término de orden *n*, para lo cual se tiene que:

1. Entre mayor es el grado del modelo polinomial ajustado, mejor es la aproximación que se obtiene mediante las expansiones en serie de Taylor a la verdadera función $\varphi(x_1, x_2, \ldots, x_k)$.
2. En regiones pequeñas se obtienen las mejores aproximaciones de cierto grado dado.

En la práctica, se procede bajo la suposición de que, sobre regiones pequeñas del espacio de factores, un polinomio de primero o segundo orden puede representar adecuadamente la verdadera función. De aquí que los modelos más utilizados en MSR sean los polinomios de primero y segundo orden dados por

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 \beta_{ii} + \sum_{i=1}^{k} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j + \varepsilon$$

$$\square^{i<j}$$

respectivamente, donde es el error aleatorio y los coeficientes (β_i) que aparecen las ecuaciones anteriores deben estimarse mediante regresión lineal, a partir de las observaciones obtenidas de la realización del diseño experimental.

Diseños de Primer Orden

Estos diseños son muy utilizados en la primera etapa de una investigación cuyo objetivo es encontrar las condiciones operación de un proceso, o cuando se tienen muchos factores, se puede utilizar un diseño de primer orden para descartar los factores que no tengan influencia significativa sobre la respuesta. Cuando se utiliza este tipo de diseño se requieren pocos puntos para ajustar un modelo polinomial a la respuesta.

Como su nombre lo indica con los diseños de primer orden se pretende ajustar un modelo de primer orden. Este modelo, como se sabe, para factores tiene parámetros a estimar, así que se requieren al menos observaciones para poderlo ajustar**.**

Cuando se quiere ajustar un modelo de primer orden en k variables $Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon$, es conveniente utilizar diseños que minimizan la varianza de los coeficientes de regresión. Los diseños que satisfacen esta condición se llaman *Diseños Ortogonales*. Un diseño se llama ortogonal cuando los términos en el modelo ajustado son no correlacionados entre sí, lo que hace que los estimadores de los parámetros tampoco estén correlacionados entre sí. Esto hace posible que la varianza de la respuesta estimada en cualquier punto en la region experimental, sea expresada como la suma de la varianza de cada parámetro estimado en el modelo. Los diseños de primer orden pueden fallar debido a que hay curvatura en la superficie de respuesta, o los experimentos se realizaron de forma incorrecta. Cuando esto sucede, el modelo debe ser mejorado mediante la adición de términos de orden mayor o a través de una transformación a las variables. Cuando el modelo de primer orden no describe

adecuadamente el comportamiento de la respuesta, se propone el modelo de segundo orden.

Diseño Factorial 2^k

En este diseño se estudian *k* factores, en dos niveles cada uno. Se le llama diseño factorial completo en *k* factores cuando se seleccionan 2 niveles del primer factor, 2 niveles del segundo factor ,…,2 niveles del k-ésimo factor, y la matriz de diseño se forma por todas las combinaciones de los niveles, que son tantas como 2^k. Por ejemplo, el diseño factorial 2^2 consiste de los cuatro tratamientos que resultan al combinar los dos niveles de cada factor y su matriz de diseño está dada por.

$$X = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{bmatrix} -1 & -1 \\ 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} \end{array}$$

Diseño Factorial para $k=2$

De Segundo Orden

Los diseños de segundo orden son útiles en la etapa final de un estudio de optimización, cuando el punto estacionario está cerca o dentro de la región experimental, y permiten estudiar efectos lineales, de interacción y efectos cuadráticos o de curvatura pura.

Cuando se está cerca del punto estacionario, a veces la verdadera respuesta tiene curvatura y no puede describirse adecuadamente con un modelo de primer orden. Si la curvatura existe, se utiliza un modelo de segundo orden. Como el siguiente:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j + \varepsilon$$

$\square^{i<j}$

A continuación, se presentan los diseños de segundo orden más utilizados en la práctica.

Diseño Factorial 3^k

En este diseño es necesario que la respuesta sea observada en todas las combinaciones de las *k* variables del diseño, las cuales tienen 3 niveles cada una.

Diseño de Composición Central

Box & Wilson proponen diseños más económicos, que tienen la ventaja de que se puede estudiar los efectos lineales no confundidos y efectos de interacción de segundo orden. En estos diseños cada variable anexa dos puntos, más replicas en el centro, en lugar de aumentar el número de niveles en los factores. Así, un diseño de composición central consiste de tres tipos de puntos, a saber:

1. Un diseño factorial 2^k completo (o fraccionado).
2. De n_0 puntos en el centro.
3. Dos puntos axiales en cada variable diseño a una distancia α del centro del diseño.

El número total de puntos en el diseño es $N=2^k+2k+n_0$. El diseño central compuesto puede hacerse rotable tomando $\alpha=(F)^{\frac{1}{4}}$ donde $F=2^K\left(o\,F=2^{k-m}\right)$ y además el diseño de composición central puede hacerse un diseño ortogonal, caso en el cual los efectos individuales de las k variables pueden ser evaluadas independientemente. Ahora si se quiere que el diseño de composición central sea ortogonal y rotable se toman $\alpha=(F)^{\frac{1}{4}}$ y $n_0=4\sqrt{F}+4-2k$.Esto es el número de réplicas en el centro puede escogerse de tal manera que el diseño sea rotable.

Técnicas de Optimización

Una vez que se tiene el modelo debidamente ajustado y validado se puede proceder a encontrar la combinación de los niveles de los factores que producen la respuesta óptima. Para localizar esta combinación de niveles, a partir del modelo ajustado, existen básicamente tres métodos, a saber:

1. Escalamiento Ascendente (o descendente)
2. Análisis Canónico
3. Análisis de Cordillera.

El uso de estos métodos depende del orden del modelo ajustado y la situación particular que se presenta con el punto óptimo que se quiere encontrar. A continuación, se describen cada uno de estos tres métodos.

Escalamiento ascendente y descendente

Este método es utilizado con el modelo de primer orden. Su objetivo es encontrar la dirección de máximo incremento de la variable de respuesta sobre el plano. En el caso de que se busque el máximo decremento, estaremos hablando del método descendente.

Consiste en la realización secuencial de experimentos a lo largo de la trayectoria de escalamiento ascendente, es decir, en la dirección del máximo incremento de la respuesta, a partir del centro del diseño. Cuando ya se tiene la dirección en la cual la respuesta crece, se realizan los experimentos secuenciales sobre puntos espaciados adecuadamente, hasta que el valor de la respuesta cambia de tendencia. En este momento se corre otro diseño de primer orden con centro en el punto anterior al cambio de tendencia. Se procede de la misma manera hasta encontrar otro cambio en la tendencia, es decir, se localiza la dirección en la cual la respuesta crece, de igual forma, se realizan experimentos en puntos espaciados hasta encontrar un nuevo punto donde hay cambio de tendencia. Se corre un tercer diseño con centro en el punto anterior al cambio de tendencia y se encuentra la nueva dirección de crecimiento. Mediante este proceso se llega a una vecindad del punto óptimo, lo cual se detecta mediante la falta de ajuste del modelo de primer orden.

Ya que se cuenta con el modelo ajustado $\hat{Y}(x)=\hat{\beta}_0+\sum_{i=1}^{k}\hat{\beta}_i x_i$ que describe adecuadamente el comportamiento de la variable respuesta, el objetivo es trasladarse a una distancia de r unidades a partir el centro del diseño en la dirección de máximo incremento de la respuesta. Por consiguiente, el problema se traduce a un problema de máximos (o mínimos, en el caso de escalamiento descendente) con ciertas restricciones. La maximización de la función respuesta se lleva a cabo mediante el uso de multiplicadores de Lagrange. El problema se fórmula de la siguiente manera:

Maximizar $\hat{Y}(x)=\hat{\beta}_0+\sum_{i=1}^{k}\hat{\beta}_i x_i$

sujeto a $r=\sqrt{\sum_{i=1}^{k}x_i^2}$

Ahora supóngase que $x=(x_1, x_2,\ldots,x_k)^t$ es tal que $r=\sqrt{\sum_{i=1}^{k}x_i^2}$ y Y(x) es máximo, entonces se cumple que

$$x_i = \beta_i /(2\mu);\ i = 1, 2,\ldots, k.$$

donde μ es el multiplicador de Lagrange. Nótese que el cambio de las variables x_i es directamente proporcional a los coeficientes estimados β_i, y por consiguiente los incrementos a lo largo de la trayectoria de escalamiento ascendente son proporcionales a los coeficientes $\{\beta_i\}$. Cabe señalar que el tamaño de paso se elige con base en el conocimiento del proceso.

Análisis Canónico

Los principales objetivos del análisis canónico son encontrar las coordenadas del punto estacionario, expresar el modelo en su forma canónica y encontrar la relación entre las variables canónicas y las variables codificadas. Este método es de gran utilidad, ya que es a través de él como se puede expresar e interpretar de manera sencilla el modelo de segundo orden utilizado.

El punto estacionario es aquél sobre el cual, dentro de una superficie de respuesta, el plano tangente a la superficie tiene pendiente cero. Es importante localizarlo porque en dicho punto, la variable respuesta es un máximo, un mínimo, o un punto silla, lo que significa que podría ser el punto óptimo que se busca. Puede suceder que se tenga una superficie estacionaria, en lugar de punto estacionario. La situación ideal es cuando dicho punto resulta ser del tipo que buscamos, máximo o mínimo, y que se encuentre dentro de la región experimental, pero en la práctica, lo más común es que el punto estacionario no sea el que buscamos, y se procede a encontrar el mejor punto dentro de la región, utilizando el método de análisis de cordillera.

Las coordenadas del punto estacionario, $x_0=(x_{10},x_{20},...,x_{k0})^t$ se obtienen derivando la respuesta ajustada con respecto a cada x_i, igualando a cero esas derivadas y resolviendo las k ecuaciones simultáneamente. Esto es, consideremos el modelo de segundo orden, dado por:

$$\hat{Y}(x)=\hat{\beta}_0+\sum_{i=1}^{k}\hat{\beta}_i x_i+\sum_{i=1}^{k}\hat{\beta}_{ii}x_i^2+\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\hat{\beta}_{ij}x_i x_j$$

¿

que en forma matricial se puede escribir como

$$\hat{Y}(x)=\hat{\beta}_0+X^t\beta+X^tBX$$

donde $x^t=(x_1,x_2,\ldots,x_k),\beta^t=(\beta_1,\beta_2,\ldots,\beta_k)$ y

$$B=\begin{bmatrix} \beta_{11} & \frac{\beta_{12}}{2} & \cdots & \frac{\beta_{1k}}{2} \\ \frac{\beta_{12}}{2} & \beta_{22} & \cdots & \frac{\beta_{2k}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_{1k}}{2} & \frac{\beta_{2k}}{2} & \cdots & \beta_{kk} \end{bmatrix}.$$

Observemos que $\hat{Y}(x) : R^k \to R$ es diferenciable. Así pues, el punto estacionario de la superficie de respuesta pertenece al conjunto de puntos $x \in R^k$ que satisfacen

$$grad\ \hat{Y}(x)=0$$

Por lo tanto, el punto estacionario esta dado $x_0=\frac{-B^{-1}\beta}{2}=(x_{10},x_{20},\ldots,x_{k0})^t$. Nótese que el punto estacionario se puede obtener fácilmente de los coeficientes del modelo ajustado.

La forma bilineal simétrica

$$Hf(x)=\begin{bmatrix} \frac{\partial^2 Y(x)}{\partial x_1^2} & \frac{\partial^2 Y(x)}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 Y(x)}{\partial x_k \partial x_1} \\ \frac{\partial^2 Y(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 Y(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 Y(x)}{\partial x_k \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 Y(x)}{\partial x_1 \partial x_k} & \frac{\partial^2 Y(x)}{\partial x_2 \partial x_k} & \cdots & \frac{\partial^2 Y(x)}{\partial x_1^2} \end{bmatrix}$$

sirve para caracterizar la superficie de respuesta de la siguiente manera:

1. Si det $Hf(x_0)\neq 0$
2. Si det $Hf(x_0)>0$ entonces mínimo local si todos eigenvalores son positivos.
3. Si det $Hf(x_0)>0$ entonces máximo local si todos los eigenvalores son negativos.

4. Si det $Hf(x_0)<0$ entonces es un punto silla.
5. Si det $Hf(x_0)=0$

Si el eigenvalor distinto de cero es positivo entonces se tiene una variedad mínima local. Si el eigenvalor diferente de cero es negativo entonces se tiene una variedad máxima local.

Método de Análisis de Cordillera

Durante el análisis de una superficie de respuesta puede suceder que el punto estacionario este afuera de la región experimental, pero todavía se desea encontrar el mejor punto dentro de esta región. El análisis de cordillera es parecido a un escalamiento ascendente, pero sobre una superficie de segundo orden. El método de análisis de cordillera sirve para encontrar el máximo (mínimo) de $\hat{Y}(x)$ sobre esferas de radio variable r_j $(i=1,\ 2,\ldots)$ centradas en el origen $(x_1,x_2,\ldots,x_k)=(0,0,\ldots,0)$ y contenidas en la region experimental. El objetivo es encontrar el máximo valor de $\hat{Y}(x)$ en la superficie de cada esfera. Como el modelo ajustado describe el comportamiento de la respuesta se espera que el mejor punto de operación sea el máximo (o mínimo) sobre alguna esfera.

El modelo ajustado de segundo orden sobre la región de las k variables codificadas $x^t=(x_1,x_2,\ldots,x_k)$ y $\beta=(\hat{\beta}_1,\ \hat{\beta}_2,\ \ldots,\ \hat{\beta}_k)$ se puede expresar en forma matricial como

$$\hat{Y}(x)=\ \hat{\beta}_0+X^t\beta+X^tBX$$

y se desea maximizar $\hat{Y}(x)=\ \hat{\beta}_0+X^t\beta+X^tBX$ sujeto a la restricción $x^tx-r^2=0$ supongamos que $x^t=(x_1,x_2,\ldots,x_k)$ es tal que $r=\sqrt{\sum_{i=1}^{k}x_i^2}$ y ademas $\hat{Y}(x)$ es un máximo (o mínimo) entonces se cumple que

$$grad\ \left(\hat{Y}(x)\right)=\mu\ \ grad\ \left(\sum_{i=1}^{k}x_i^2-r^2\right)$$

donde μ es el multiplicador de Lagrange. De aquí se deduce que

$$2\begin{bmatrix} \beta_{11} & \frac{\beta_{12}}{2} & \cdots & \frac{\beta_{1k}}{2} \\ \frac{\beta_{12}}{2} & \beta_{22} & \cdots & \frac{\beta_{2k}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_{1k}}{2} & \frac{\beta_{2k}}{2} & \cdots & \beta_{kk} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = 2\mu \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} - \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Así, el mejor punto sobre cada esfera debe cumplir la restricción

$$(B - \mu I)x = \frac{-\beta}{2},$$

donde el valor de μ se elige de acuerdo a lo que se busca; si se quiere un máximo μ debe ser mayor que el más grande valor propio de la matriz B, y si se busca un mínimo debe ser menor que el más pequeño valor propio de la matriz B Cada valor de μ corresponde a una esfera de cierto radio, y se debe localizar aquella sobre la cual el modelo predice el mejor valor de la respuesta en la región experimental.

Ejemplo 1: Una Aplicación de MSR en la Industria Electrónica.

La compañía está interesada en determinar los niveles de los factores que minimizan el número de obleas rotas. Los factores que se controlan son *temperatura de grabado, temperatura de piranha y temperatura de agua*. El proceso se realizaba antes del experimento a una temperatura de grabado de - , una temperatura de piranha de y una temperatura de agua de y se tena un rendimiento mecánico del en la solución piranha. Se utilizo un diseño factorial , en cada tratamiento se utilizaron obleas, y se obtienen solo ocho puntos, se decidió considerar solo obleas por tratamiento es suficiente para observar al menos una oblea rota por tratamiento.

Se supone que la variable porcentaje de obleas rotas depende de los factores temperatura de grabado, temperatura de piraña y temperatura de agua, en grados centígrados respectivamente. El modelo regresión lineal múltiple a considerar está dado por

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_{12} X_{i12} + \beta_{13} X_{i13} + \beta_{23} X_{i23} + \varepsilon_i;\ i = 1,2,\ldots,16,$$

donde

Y_i = Porcentaje de obleas rotas

X_{i1} = Temperatura de grabado

X_{i2} = Temperatura de piranha

X_{i3} = Temperatura de agua

β = $(\beta_0,\ \beta_1,\ \beta_2,\ \beta_3, \beta_{13}, \beta_{12}, \beta_{23})$, es el vector de parámetros, y

ε_i = $N(0,\ \sigma^2)$, donde es la varianza.

En base a la muestra y aplicando regresión lineal múltiple (mediante paquete Statistica) se obtiene el modelo ajustado siguiente:

$$\hat{E}(Y_i) = 0.015 - 0.009\ X_{i1} - 0.003\ X_{i2} - 0.004\ X_{i3} + 0.004\ X_{i13}; i = 1,2,\ldots,\ n$$

De la expresión anterior se ve que el factor que más afecta a la variable número de obleas rotas (en porcentajes) es la temperatura de grabado. También se observa que mantener la temperatura de grabado en su nivel alto causa una disminución en la variable respuesta, mientras que, manteniendo la temperatura de piraña en su nivel alto y la temperatura de agua en su nivel bajo, implicando esto que se eliminen los términos correspondientes a la temperatura de piraña y el factor de interacción. Es decir, la combinación $(1,\ 1,\ -1)$ es el mejor tratamiento, debido a que esto causa que los términos correspondientes a temperatura de grabado y temperatura de piraña contribuyen a una disminución en la variable porcentaje de obleas rotas mientras que los otros factores se eliminan.

Se realizaron las gráficas de diagnóstico no se observa alguna violación seria a la suposición de normalidad, no se viola el supuesto de independencia, ni el supuesto de varianza constante.

La significancia de la regresión se prueba mediante la hipótesis

H_0: $\beta_1 = \beta_2 = \cdots = \beta_{3\text{vs}}$ H_1: $\beta_i \neq 0$ para alguna

En la tabla siguiente aparece el análisis de varianza y se concluye que al menos una variable contribuye significativamente a la regresión porque $F_o = 24.84 > F_{0.05,4,11} = 12.002$.

Efecto	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Estadístico F	P_{valor}
Regresión	.002340	4	.000585	24.84556	.000018
Residuos	.000259	11	.000024		
Total	.002599				

Técnica de Optimización

Con el método de mínimos cuadrados se ajustó el modelo de segundo orden a los datos y se obtuvo el modelo dado por

$$\hat{Y} = 0.015 - 0.009\ X_1 - 0.003\ X_2 - 0.004\ X_3 + 0.004\ X_{i13};$$

Considerando que el objetivo es encontrar los niveles de los factores que producen la respuesta mínima, se utiliza el análisis canónico y se obtienen los siguientes resultados el valor mínimo de los valores positivos de se encuentran sobre la esfera de radio 0.8, sobre el punto con coordenadas $(1.001, 0.334, 0.572)$.

Análisis mediante un Modelo de Regresión Logística

La función distribución que está siendo propuesta para el uso de análisis de la variable respuesta binaria, es regresión logística.

En esta situación se puede expresar el valor de la variable respuesta, dado **x**, como

$$Y = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_{12} x_{12}+\beta_{23} x_{23}+\beta_{13} x_{13}+\beta_{13}\ x_{13}}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_{12} x_{12}+\beta_{23} x_{23}+\beta_{13} x_{13}+\beta_{13}\ x_{13}}} + \varepsilon = p_i x + \varepsilon$$

Aquí ε asume únicamente dos posibles valores.

$$\varepsilon = \begin{cases} 1 - p_i\ x, & \& con\ \ probabilidad\ \ p_i x\ \ , \ si\ \ y = 1 \\ p_i, & con\ \ probabilidad\ \ 1 - p_i x\ \ , \ \ si\ \ y = 0 \end{cases}$$

Así, ε tiene una distribución binomial con media cero y varianza $(\pi,\ 1-\pi)$. Esto es, la distribución binomial describe la distribución de los errores. La distribución condicional de la variable de respuesta sigue una distribución binomial, con probabilidad dada por la media condicional $p_i x$

En regresión logística se puede estimar directamente la probabilidad de ocurrencia de un evento. Para el caso de más de una variable, el modelo de regresión logística puede ser escrito como

$$P_r(Y=1|x)=\frac{e^{g(x)}}{1+e^{g(x)}}=\frac{1}{1+e^{-g(x)}}=p_i x$$

Donde $g(x)= e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_{12} x_{12}+\beta_{23} x_{23}+\beta_{13} x_{13}+\beta_{13}\ x_{13}}$

La probabilidad que un evento no ocurra se estima por

$$P_r(Y=0|x)=\frac{P_r(Y=1|x)}{1+e^{g(x)}}=1-p_i x$$

En regresión logística, los parámetros del modelo son estimados mediante Máxima Verosimilitud. Todas las ecuaciones de verosimilitud se obtienen derivando la ecuación de log-verosimilitud respecto a cada parámetro. Así, para la pareja $(x_i|y_i)$, $y_i=1$ con , la contribución a la función verosimilitud es $p_i x$, y para las parejas donde $1-\ p_{ix}$ la contribución a la función de verosimilitud es . La cantidad $p_i(x)$ denota el valor de $p_i(x)$ calculado en x_i. Puesto que las observaciones son independientes la función de verosimilitud tiene la forma:

$$L(\beta)=\prod_{i=1}^{16} p_i(x)^{y_i}\left(1-p_i(x)\right)^{1-y_i}$$

El principio de máxima verosimilitud establece que el estimador de será el que maximice la ecuación 4.4 . Sin embargo, es más fácil trabajar con la ecuación de log-verosimilitud definida como

$$\ln\ L(\beta)=\sum_{i=1}^{16} y_i \ln\left(p_i(x)\right)+\left(1-y_i\right)\ln\left(1-p_i(x)\right)$$

El valor de que maximiza se obtiene derivando respecto a $\beta_0, \beta_1,\ \beta_{2,}\ \beta_{13}, \beta_{23},$ e igualando a cero las ecuaciones resultantes:

$$\sum_{i=1}^{16}\ x_{ij}\quad \left(y_i-\left(p_i(x)\right)\right)=0\,,\ para\ \ 1=1,2,3$$

La significancia de cada coeficiente se muestra en la tabla (4.3). En ella se observa que los efectos X_1, X_2, X_{13} son significativos con . Nótese que es más significativo X_1, seguido por X_{13} y después X_2.

En la tabla (4.4) se presentan los coeficientes estimados y los estadísticos relacionados del modelo de regresión logística.

Que predice el porcentaje de obleas rotas en términos del intercepto y las variables temperatura de grabado, temperatura de pirahna y temperatura de agua.

Variable	Coeficientes	Error estándar	Estadístico Wald	Grados de Libertad	Nivel de significancia	Estadístico R	Factor
X_1	-.8926	.2082	18.3753	1	0.0000	-0.1610	0.4096
X_2	-.4384	.2082	4.4330	1	0.0353	-.0621	0.6451
X_1X_2	-.3144	.2082	2.2804	1	0.1310	-.0211	0.7302
X_1X_3	.3292	.1361	5.8522	1	0.0156	.0781	1.3899
Constante	-4.614	.2124	471.7991	1	0.0000		

Tabla (4.4) Parámetros estimados para el modelo de regresión logística.

Dados los coeficientes, la ecuación de regresión logística para la probabilidad de oblea rota puede expresarse en la forma

$$P_r(Y=1|x)=\frac{e^{g(x)}}{1+e^{g(x)}}$$

Donde $g(x)= e^{-4.6141-0.8926_{\square}x_1-0.4384\,x_2-0.3144\,x_{12}+0.3292\,x_{13}}$

Este modelo se obtuvo con SPSS (Paquete Estadístico para Ciencias Sociales), utilizando el método Backward. La significancia de cada coeficiente se muestra en la tabla (4.3). En ella se observa que los efectos X_1, X_2 y X_{13} son significativos con $\alpha=0.05$. Nótese que es más significativo X_1, seguido por X_{13} y después X_2.

$$H_0: \hat{\beta}_i=0 \qquad \text{vs} \qquad H_1: \hat{\beta}_i\neq 0$$

La prueba que el coeficiente es cero se basa en el estadístico Wald, el cual tiene una distribución ji-cuadrada. Puesto que la variable tiene un simple grado de libertad, la estadística de prueba es justamente,

$$W = \left(\frac{\widehat{\beta}_i - \beta_i}{\sqrt{Var\left(\widehat{\beta}_i\right)}} \right)^2$$

Donde la estadística W tiene una distribución Ji-cuadrada con n-3 grados de libertad bajo la hipótesis H_0. Para un nivel de significancia $\alpha = 0.05$, la región crítica es de la forma.

$$P\left(|T| > W\right) = 0.05$$

La hipótesis H_0 se rechaza si $|T| > W$. En la tabla (4.4) se presentan los valores de la estadística W para los parámetros $\beta_0, \beta_1, \ \beta_2, \ \beta_{12}, \beta_{13}$.

La contribución de una variable individual en regresión logística es difícil de determinar. La contribución de cada variable depende de las otras variables en el modelo. En nuestro caso las variables no están correlacionadas.

Mediante el estadístico R se estima la correlación parcial entre la variable dependiente y cada variable independiente. De la tabla (4.4) se ve que el estadístico está entre -1 y 1.

La ecuación (4.6) para el estadístico R es

$$R = \sqrt{\frac{W - 2k}{-2 \ll (0)}}$$

donde k son los grados de libertad para la variable y W es el estadístico de Wald. El denominador es menos dos veces la log-verosimilitud de un modelo base que contiene únicamente al intercepto. El valor de 2k en la ecuación (4.6) es un ajuste por número de parámetros estimados.

Para interpretar los coeficientes en un modelo de regresión logística, modelo logístico puede ser reescrito en términos de los eventos que están ocurriendo. Primeramente, expresamos el modelo logístico en términos de razón de logaritmos, el cual es llamado un modelo logit:

$$\log\left(\frac{P_r(Y=1|x)}{P_r(Y=0|x)} \right) = -4.6141 - 0.8926\, x_1 - 0.4384\, x_2 - 0.3292\, x_{13}$$

De la ecuación anterior, se ve que los coeficientes pueden ser interpretados como el cambio en la razón de logaritmos asociado con una unidad de cambio en la variable independiente. Puesto que es más fácil considerar los razón, mejor que la log-razón, la ecuación logística puede ser expresada de la siguiente manera:

$$\left(\frac{P_r(Y=1|x)}{P_r(Y=0|x)}\right)=e^{-4.6141-0.8926_{\square}x_1-0.4384\,x_2-0.3292\,x_{13}}$$

$$\left(\frac{P_r(Y=1|x)}{P_r(Y=0|x)}\right)=e^{-4.6141}e^{-0.8926_{\square}x_1}e^{-0.4384\,x_2}e^{-0.3292\,x_{13}}$$

Entonces e es elevado a la potencia β_i es el factor por el cual el odds cambio cuando la i-ésima variable independiente crece por una unidad. Puesto que $\beta_0, \beta_1, \beta_2,$, son negativos, sus respectivos factores son menor que uno, lo cual significa que se tiene un descenso; como β_{13} es positivo, este factor será mayor que uno, lo cual significa que se tiene un incremento.

Una manera de evaluar el ajuste del modelo es comparar los valores ajustados con los valores observados. En la tabla de clasificación para obleas rotas (4.5) se muestra la clasificación

	Ajustados		Porcentaje correcto
Observado	0	1	
0	3939	0	
1	61	0	
			98.48 %

Tabla (4.5) de clasificación para obleas rotas

De la tabla anterior se ve que 3939 fueron correctamente ajustadas por el modelo, los cuales no están rotas. De las 4000 obleas, el 98.48 % de ellas fueron correctamente especificado.

La probabilidad de los resultados observados, dado los parámetros estimados, es conocido como la verosimilitud es menor que 1, se utiliza -2 veces la log-verosimilitud (-2LL) como una medida, de cómo el modelo estimado ajusta a los datos. Para el modelo de regresión logística que contiene únicamente la constante, -2LL es 592.298.

Otra medida de como el modelo se ajusta es el estadístico de bondad de ajuste, el cual compara las probabilidades observadas de las ajustadas del modelo. El estadístico de bondad de ajuste es definido como

$$Z^2 = \sum \frac{r_i^2}{P_i(1 - P_i)}$$

donde el residual es la diferencia entre el valor observado, y el valor ajustado, P_i.

En la tabla (4.6) se presenta el estadístico de bondad de ajuste para el modelo con todas las variables independientes. Para el modelo actual, valor de -2LL es 592.298, el cual es menor (mayor) que -2LL para el modelo conteniendo únicamente la constante.

	Chi-cuadrada	Grados de libertad	Significancia
Modelo Chi-cuadrada	39.114	4	0.0000
Mejoramiento	-.736	1	.3909

En la tabla (4.6), el modelo chi-cuadrada es la diferencia entre -2LL conteniendo únicamente la constante y -2LL para el modelo actual. Esto es mediante el modelo chi-cuadrada se prueba la hipótesis de que todos los términos, excepto la constante, son cero.

CONCLUSIONES

De los resultados obtenidos del análisis realizado en el presente trabajo se puede resaltar lo siguiente:

1. Se presenta una aplicación del diseño de experimentos en la industria electrónica mexicana. Algunos aspectos que hacen interesante este experimento son: 1) las consideraciones de ingeniera de proceso que se hicieron previamente, 2) las diferentes alternativas de análisis estadístico, por ser la respuesta una variable binaria y 3) el ahorro económico obtenido.

2. Los diferentes análisis que se realizaron detectaron como significativos los efectos X_1, X_2 y X_{13}. El análisis mediante el modelo de regresión múltiple y el modelo de regresión logística tuvieron la capacidad para detectar los mismos efectos. Esto se debe tal vez a que un número grande de obleas procedas en cada tratamiento.

Antes del experimento se utilizaba la combinación de temperaturas (X_1, X_2, X_3) = (-1, 1, -1) y después de analizar los datos mediante las técnicas, se encontró que un mejor punto es (1, 1, -1).

3. Desde el punto de vista económico, se importante mencionar que antes del experimento se rompan obleas por cada mil procesadas, lográndose este número a 15 por cada mil. Esta mejora representa un ahorro aproximado de $ 8000.00 dólares mensuales. Esta aplicación del diseño de experimentos muestra que para tener mejoras importantes no se requieren diseños complicados, ni análisis estadísticos sofisticados, sino experimentos bien conducidos. Aunque hubiese sido mejor haber corrido un diseño factorial 2^3 con puntos al centro.

REFERENCIAS

1)Box, G. E. P. (1952). Multi-factorial designs of first orders. Biométrica.

2)Box, Ep, Box., and Draper, R. Norma (1963). Empirical Model-Building and Response Surfaces, John Wiley \& Sons, Inc.

3)Castro Montoya, R. 1995. Metodología de Superficie de Respuesta: Una aplicación en la fabricación de circuitos integrados. Tesis de licenciatura. Escuela de Ciencias Físico Matemáticas, UAS.

4)Gutiérrez Pulido, H. y De la Vara Salazar, R. (2003). Análisis y Diseño de Experimentos. Ed. McGraw Hill. México, DF.

5)Khuri A.I. and Cornell J.A. (1987). Response Surfaces. New York: Marcel Dekker.

6)Montgomery D.C. (1991). Design and Analysis of Experiments. Third edition. New York: Wiley.

7)Myers, R. (1971). Response Surface Methodology. Boston: Allyn and Bacon.