CORPORACIÓN MEXICANA DE INVESTIGACIÓN

EN MATERIALES S.A. DE C.V.

"ANÁLISIS DE MULTICOLINEALIDAD EN MODELOS DE REGRESIÓN PARA PROCESOS DE MANUFACTURA"

TESIS

Para obtener el Grado Académico de

Maestro en Ciencia y Tecnología en Ingeniería Industrial y de Manufactura

Presenta:

JORGE SAÚL RODRÍGUEZ VÁZQUEZ

SALTILLO, COAHUILA MARZO 2022

“ANÁLISIS DE MULTICOLINEALIDAD EN MODELOS DE REGRESIÓN PARA PROCESOS DE MANUFACTURA”

Por
Jorge Saúl Rodríguez Vázquez

Tesis

Presentada al Programa Interinstitucional en Ciencia y Tecnología

Sede

Corporación Mexicana de Investigación en Materiales, S.A. de C.V.

Como requisito parcial para obtener el Grado Académico de

Maestría en Ciencia y Tecnología
en Ingeniería Industrial y de Manufactura

Programa Interinstitucional en Ciencia y Tecnología COMIMSA / CONACyT

Saltillo, Coahuila, Marzo del 2022

CORPORACIÓN MEXICANA DE INVESTIGACIÓN EN MATERIALES, S.A. DE C.V.

GERENCIA DE DESARROLLO HUMANO

DIVISIÓN DE ESTUDIOS DE POSTGRADO

Los miembros del Comité Tutorial recomendamos que la Tesis "Análisis de multicolinealidad en modelos de regresión para procesos de manufactura", realizada por el alumno **Jorge Saúl Rodríguez Vázquez** matrícula 1806IM1142 sea aceptada para su defensa como **Maestro en Ciencia y Tecnología en Ingeniería Industrial y de Manufactura.**

El Comité Tutorial

_____

Dr. Rolando Javier Praga Alejo

Director de Tesis

Tutor Académico

_____

Dr. David Salvador González González

Asesor

_____

Dr. Pedro Pérez Villanueva

Coordinación General de Estudios de Posgrado

COMIMSA

CORPORACIÓN MEXICANA DE INVESTIGACIÓN EN MATERIALES, S.A. DE C.V.

GERENCIA DE DESARROLLO HUMANO

DIVISIÓN DE ESTUDIOS DE POSGRADO

Los abajo firmantes, miembros del Jurado de Examen de Grado del C. **Jorge Saúl Rodriguez Vázquez**, una vez leída y revisada la Tesis titulada "Análisis de multicolinealidad en modelos de regresión para procesos de manufactura" aceptamos que la referida tesis revisada y corregida sea presentada por el alumno para aspirar al grado de Maestría en Ciencia y Tecnología en Ingeniería Industrial y de Manufactura durante el Examen de Grado correspondiente.

Y para que así conste firmamos la presente a los 31 días del mes de Marzo del año 2022

_____ _____

Dra. Esmeralda Ramírez Méndez — Presidente

Dr. Marco Antonio Fuentes Huerta — Secretario

_____

Dr. Rolando Javier Praga Alejo

Vocal

DECLARACIÓN DE ORIGINALIDAD

Yo, Jorge Saúl Rodríguez Vázquez, estudiante con matrícula 1806IM1142, del Programa de Posgrado Maestría en Ciencia y Tecnología en Ingeniería Industrial y de Manufactura de la Corporación Mexicana de Investigación en Materiales S.A. de C.V. (COMIMSA), declaro que el presente trabajo terminal con título "Análisis de multicolinealidad en modelos de regresión para procesos de manufactura" es original, de mi autoría y producto de mi contribución intelectual y de investigación.

Así mismo, manifiesto que los datos, imágenes y textos tomados de fuentes publicadas, como artículos y tesis, están debidamente citados y referenciados, dando el crédito a los investigadores y fuentes originales.

_____

Nombre y Firma del estudiante

Agradecimientos

Al concluir esta etapa de mi vida quiero extender un profundo agradecimiento a quienes hicieron posible este sueño, aquellos que durante mi camino en todo momento fueron inspiración, apoyo y fortaleza.

Esta mención en especial es para **DIOS**.

Para mi esposa **Rocío Natali Carranza Santoy** que ha sido el principal apoyo en momentos de flaqueza y debilidad ya que con sus consejos, enseñanzas y tolerancia me ha brindado ese respaldo incondicional para culminar esta etapa de mi vida.

Para mis hijos **Jorge** y **Saúl** que me brindaron su apoyo, me comprendieron, tuvieron tolerancia e infinita paciencia cedieron su tiempo conmigo para permitirme llevar adelante este proyecto, que paso de ser una meta personal a un emprendimiento de familia, a ellos mi infinito cariño y gratitud.

Para mis padres **Jorge Rodriguez Covarrubias** y **Alma Rosa Vázquez Mendoza** porque a pesar de las dificultades que presenta la vida siempre han sabido enseñarme a salir adelante y a no rendirme. Muchas gracias a ustedes por demostrarme que "el verdadero amor no es otra cosa que el deseo inevitable de ayudar al otro para que este se supere".

Para mi hermano **Jaime Raúl** del cuál quiero ser un buen ejemplo y que vea que nunca es tarde para continuar creciendo en el ámbito profesional.

Agradezco también a mi asesor de tesis el **Dr. Rolando Javier Praga Alejo** por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento científico, así como

también haberme tenido toda la paciencia del mundo para guiarme durante todo el desarrollo de la tesis, a quien hago llegar mi más sincero agradecimiento, por permitirme ser partícipe de uno de sus proyectos dentro del centro de investigación y por su entrega incondicional durante el desarrollo de este trabajo de investigación.

De la misma manera al **Dr. David Salvador González González**, por su colaboración directa en el proyecto.

Al **Dr. Pedro Pérez Villanueva** por siempre estar al pendiente de sus estudiantes y dispuesto a brindar apoyo cuando uno lo requería.

A mi gran amigo **Isaac Esaú Cerda Duran** por todos sus consejos, por sus palabras de aliento y motivación que en ocasiones llegue a necesitar.

Además agradezco a **El Consejo Nacional de Ciencia y Tecnología** (CONACyT) y a **La Corporación Mexicana de Investigación en Materiales** (COMIMSA) por haberme aceptado como becario nacional con CVU 939533, ser parte de ella y haberme abierto las puertas de su seno científico para poder estudiar la maestría, así como también a los diferentes docentes que brindaron sus conocimientos y su apoyo para seguir adelante día a día.

También agradezco a todos los que fueron **Mis compañeros** de clase durante todos los cuatrimestres de la maestría, ya que gracias al compañerismo, amistad y apoyo moral, han aportado un alto porcentaje de ganas de seguir adelante en mi vida profesional.

Resumen

Hoy en día el sector industrial se está apoyando en la investigación científica para controlar y mejorar sus procesos, todo esto se debe en gran medida a la expansión industrial en todo el mundo, además la diversidad de procesos industriales hace necesario el apoyo de herramientas, métodos y técnicas capaces de comprender, analizar y controlar los procesos industriales que presenten variabilidad durante su implementación.

La regresión lineal es una herramienta estadística comúnmente utilizada para analizar la posible relación entre las variables, en la mayoría de los casos el análisis que se lleva a cabo mediante el método de Mínimos Cuadrados, el cual es realizado entre una variable de salida y dos o más variables de entrada, también llamadas factores. El problema al utilizar dicho método es cuando las variables de entrada presentan una alta relación entre ellas, (también conocido como multicolinealidad), el modelo que se obtiene mediante mínimos cuadrados no es confiable al momento de analizar el proceso en cuestión.

La regresión Ridge es una alternativa que se utiliza cuando los factores presentan multicolinealidad, sin embargo para llevar a cabo este método es necesario agregar un valor numérico llamado sesgo que permite eliminar la relación entre los factores, no obstante al añadir dicho valor es importante agregar un número lo suficientemente grande para eliminar la multicolinealidad, pero también lo suficientemente pequeño para que no afecte demasiado al modelo.

Es por eso que en este trabajo de investigación se lleva a cabo un análisis comparativo entre diversos métodos para obtener parámetros de sesgo que eliminan la multicolinealidad

entre los factores de entrada, aunado a un porcentaje aceptable del ajuste del modelo. Por lo cual se construyó el Análisis de la Varianza ($ANOVA$ por sus siglas en inglés) para cada método con el fin de realizar un comparativo entre los diversos métodos. Además se contrastan los métodos con métricas estadísticas como lo son el Cuadrado Medio del Error (MSE por sus siglas en inglés) y el ajuste del modelo (R^2) las cuales permitirán elegir el método más adecuado al momento de seleccionar el sesgo que elimine la dependencia lineal entre las variables de entrada. Los resultados obtenidos fueron aplicados a dos procesos industriales, el primero es del maquinado de una pieza y el segundo es un proceso de soldadura.

Índice

Capítulo 1

Introducción

La industria tiene como objetivo aprovechar los recursos naturales para convertirlos en productos útiles para la humanidad, mejorando así, su calidad de vida. En el sector industria existe gran variedad de procesos para la obtención, transformación y transportación de la materia prima, en esta investigación se ahondará en la transformación de los recursos naturales, los cuales son también conocidos como procesos de manufactura.

La manufactura cuenta con una alta gama de procesos industriales y debido a que la empresa desea cumplir con las especificaciones del producto, es necesario que los procesos se encuentren controlados. Al tener los parámetros adecuados de los factores, esto permite obtener las características de calidad que se desea, en caso contrario, se puede mencionar que el proceso presenta variabilidad. Debido a las exigencias industriales, si el producto no cumple con las características deseadas (a pesar de tener los parámetros establecidos para el proceso), una práctica utilizada con mayor frecuencia para ajustar los parámetros de los factores que intervienen durante el proceso es, modificarlos a prueba y error, esperando que el producto llegue a satisfacer las características de calidad. Por tal motivo, es pertinente recurrir a especialistas que analicen el proceso, determinen las razones por las cuales el proceso presenta variabilidad y señalen las condiciones necesarias para obtener las especificaciones deseadas para el producto.

Para analizar el proceso se puede realizar una investigación científica, donde los recursos son más abundantes. Un camino es la estadística, esta ciencia toma relevancia en el área industrial, debido a que cuenta con diversas técnicas que se pueden implementar para dar solución a los problemas en los procesos de manufactura, asimismo, la estadística cuenta con todo un compendio de métricas que avalan, dan certeza y veracidad a los resultados obtenidos de la investigación.

Dentro de la estadística se encuentra el análisis de regresión, empleado con el fin de investigar y modelar la relación entre las variables, para después inferir a partir del modelo detalles importantes acerca del proceso, como lo son: determinar cuáles son los factores más significativos, establecer intervalos de confianza a los factores del proceso para poder cumplir con las características de calidad del producto, etc. Entonces, el análisis de regresión se convierte en una herramienta útil para dar solución a problemáticas industriales.

Un objetivo importante en el análisis de regresión es estimar los coeficientes en el modelo de regresión, también llamado "ajuste del modelo", se calculan los coeficientes aplicando procedimientos matemáticos para resolver un sistema de ecuaciones (generado a partir de una muestra de los datos del proceso) el cual es expresado en notación matricial para mayor comodidad. Seguido de la estimación de los coeficientes, se analiza el modelo de regresión para establecer si es adecuado y definir también la calidad del ajuste; a partir del análisis se determina la utilidad del modelo de regresión, señalando si es razonable o si se debe modificar. El análisis del modelo también es conocido como "comprobación de la adecuación del modelo".

Un problema que influye sobre la utilidad del modelo de regresión es la multicolinealidad entre las variables de entrada. La multicolinealidad implica una dependencia casi lineal entre los factores del proceso, afectando la precisión para estimar los coeficientes de regresión, que son utilizados para aplicar las pruebas de hipótesis sobre los coeficientes individuales del modelo, mediante estas pruebas de hipótesis se determina la contribución individual de dichos coeficiente para el modelo. Por lo tanto, la multicolinealidad afecta directamente las pruebas de hipótesis del modelo, generando problemas como, señalar erróneamente la contribución individual de los coeficientes en el modelo de regresión.

Ante la presencia de multicolinealidad entre las variables de entrada (factores), se debe optar por otros métodos de regresión distintos a Mínimos Cuadrados (MC), como Regresión Ridge (RR), su objetivo es eliminar el efecto de la multicolinealidad, agregando un valor de sesgo a la matriz de diseño, de preferencia entre cero y uno (Montgomery, Peck, & Vining, 2006) aportado por el investigador. Existen diferentes técnicas para determinar el valor de sesgo en RR, entre las que destacan: la traza de Ridge y un método iterativo propuesto por Hoerl et al., (1975), dicho método se ha ido modificando y se han obtenido buenos resultados como se muestra en Golub et al., (1979), Kibria (2003), Alkhamisi et al., (2007), Dorugade et al., (2010), Wong et al., (2015). Al emplear Regresión Ridge es importante obtener el sesgo de tal forma que, la reducción de la multicolinealidad sea mayor que el aumento del Cuadrado Medio del Error (MSE por sus siglas en inglés) se observa que cuando el sesgo aumenta, también aumenta MSE y disminuye el estadístico R^2 conocido como “adecuación general del modelo”, implicando una pérdida de “ajuste” pero se obtienen estimadores más estables.

Capítulo 2

Planteamiento del problema

En este capítulo se presenta la descripción del problema, seguido de las hipótesis, objetivos, tanto el general como los específicos, las ventajas y desventajas de la solución del problema y los resultados esperados, todo esto a través de secciones.

2.1 Descripción del problema

Al aplicar Mínimos Cuadrados (MC) en un proceso industrial con presencia de multicolinealidad entre sus factores, se obtendrán estimadores inadecuados, es necesario recurrir a métodos alternativos para tratar la multicolinealidad y obtener estimadores más estables.

El modelo de regresión múltiple representado en forma matricial es:

$$y = X\beta + \varepsilon \tag{2.1}$$

Donde y es el vector de respuestas, X representa la matriz de diseño, β es el vector de parámetros desconocidos y ε es un vector de errores aleatorios.

Los estimadores mediante Mínimos Cuadrados son:

$$\hat{\beta}_{MC} = (X'X)^{-1}X'y \tag{2.2}$$

Si la matriz $X'X$ presenta dependencia lineal entre sus columnas entonces se obtendrán estimadores inestables debido a la multicolinealidad.

Definiendo la multicolinealidad de la matriz de diseño X en términos de la dependencia lineal entre las columnas, donde X_j es la $j-ésima$ columna de la matriz X de

modo que $X = [X_1, X_2, \dots, X_p]$ van a ser linealmente dependientes si existe un conjunto de constantes $t_1, t_2, \dots, t_p$ no todas cero, tales que:

$$\sum_{j=1}^{p} t_j X_j = 0 \qquad (2.3)$$

Si la ecuación (2.3) es exactamente válida para un subconjunto de las columnas X, entonces la matriz X es Linealmente Dependiente (LD), por lo cual las columnas de X no son ortogonales, eso implica que las columnas de la matriz generada $X'X$ sean no ortogonales, entonces el rango de la matriz $X'X$ es menor que su número de columnas, por lo que la matriz $X'X$ es singular y su determinante es cero $|X'X| = 0$, ocasionando la inexistencia de su inversa $(X'X)^{-1}$ y no poder calcular los estimadores mediante Mínimos Cuadrados.

Ahora si la ecuación (2.3) es aproximadamente válida para un subconjunto de las columnas X , entonces va a existir dependencia casi lineal en la matriz X, esto implica que la matriz generada $X'X$ también presenta dependencia casi lineal, ocasionando que el rango de la matriz $X'X$ sea menor o igual a su número de columnas, entonces la matriz $X'X$ es casi singular y su determinante es muy cercano a cero $|X'X| \approx 0$ afectando el cálculo de los estimadores del modelo mediante MC.

Si la dependencia lineal es muy alta en la matriz, podrán ocurrir los siguientes síntomas al modelar mediante Mínimos Cuadrados.

- Pequeños cambios en los datos provocan grandes cambios en las estimaciones de los coeficientes.
- Las estimaciones de los coeficientes pueden presentar signos distintos a los esperados y magnitudes poco razonables.
- Se incrementan las varianzas de los coeficientes estimados por MC.

- Se obtienen valores altos del coeficiente de determinación R^2 aun cuando los valores de los estadísticos t de forma individual son bajos.

Un serio problema de modelar un proceso mediante regresión, es la presencia de multicolinealidad en la matriz de diseño, debido a que el "ajuste del modelo" mediante Mínimos Cuadrados es inadecuado, generando estimadores muy sensibles. Es necesario detectar y eliminar la presencia de multicolinealidad para obtener estimadores más estables del modelo de regresión.

Para detectar la multicolinealidad se utilizan los métodos:

a) *Examen de la matriz de correlación*: inspecciona los elementos r_{ij} no diagonales de la matriz de diseño, si se observan valores próximos a 1 entonces los elementos x_i y x_j son casi linealmente dependientes.

$$r_{x_i x_j} = \frac{\sigma_{x_i x_j}}{\sigma_{x_i}\sigma_{x_j}} \tag{2.4}$$

b) *Factores de Inflación de la Varianza* (VIF por sus siglas en inglés): miden el efecto combinado que tienen las dependencias entre los regresores sobre la varianza de cada término, si los VIF obtenidos son mayores que 10 existe multicolinealidad.

$$VIF_j = \frac{1}{(1 - R_j^2)} \tag{2.5}$$

c) *Análisis del Eigensistema*: si uno o más eigenvalores son muy pequeños, implica que existe dependencia casi lineal entre las columnas de la matriz. Se examina el número de condición que se define como:

$$\eta = \frac{\lambda_{max}}{\lambda_{min}} \tag{2.6}$$

Si $k < 100$ no existen problemas graves de multicolinealidad, si k esta entre 100 y 1000 implica multicolinealidad de moderada a fuerte y si $k > 1000$ es indicio de fuerte multicolinealidad.

A continuación se muestra un primer caso de estudio de un proceso de maquinado para evidenciar la multicolinealidad y señalar el impacto negativo que tiene sobre los estimadores obtenidos mediante regresión.

El proceso de rectificado utiliza una maquina con programación CNC para dar un acabado superficial a la pieza de la Figura 2.1. La herramienta de corte que se emplea acumula residuos en ella, es necesario remover el exceso mediante un proceso llamado dresado, para que así, el proceso de rectificado se realice de manera correcta. Las variables del proceso son:

x_1= Avance de herramienta (mm/rev) y= Diámetro 30.187 mm

x_2= Profundidad de corte (mm)

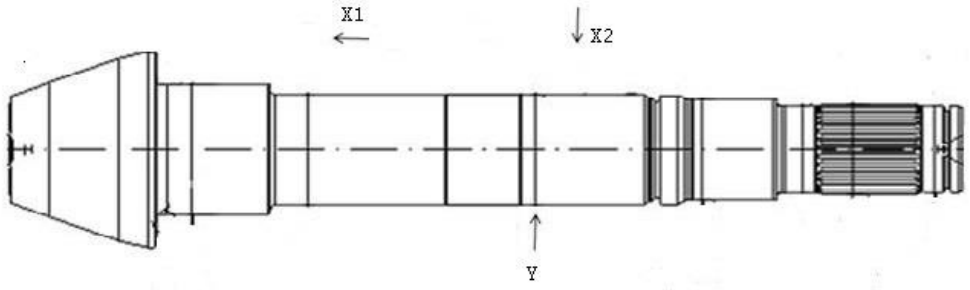

Figura 2.1 Pieza proceso de maquinado

La siguiente tabla muestra 27 observaciones del proceso de rectificado para obtener la pieza de la figura (2.1). Se genera la matriz de diseño para detectar la multicolinealidad entre las variables de entrada.

Tabla 2.1. Matriz de diseño del proceso de maquinado

x_1	x_2	x_1x_2	${x_1}^2$	${x_2}^2$	y
Avance de corte	Profundidad de corte	AC*PC	AC^2	PC^2	D 30.187
180	30	5400	32400	900	30.184
180	40	7200	32400	1600	30.187
180	50	9000	32400	2500	30.188
200	30	6000	40000	900	30.190
200	40	8000	40000	1600	30.191
200	50	10000	40000	2500	30.190
220	30	6600	48400	900	30.196
220	40	8800	48400	1600	30.195
220	50	11000	48400	2500	30.194
180	30	5400	32400	900	30.190
180	40	7200	32400	1600	30.187
180	50	9000	32400	2500	30.189
200	30	6000	40000	900	30.193
200	40	8000	40000	1600	30.194
200	50	10000	40000	2500	30.190
220	30	6600	48400	900	30.195
220	40	8800	48400	1600	30.199
220	50	11000	48400	2500	30.198
180	30	5400	32400	900	30.196
180	40	7200	32400	1600	30.194
180	50	9000	32400	2500	30.194
200	30	6000	40000	900	30.197
200	40	8000	40000	1600	30.197
200	50	10000	40000	2500	30.199
220	30	6600	48400	900	30.199
220	40	8800	48400	1600	30.200
220	50	11000	48400	2500	30.201

La tabla (2.2) muestra los Factores de Inflación de la Varianza (VIF), resultados obtenidos a partir de la matriz de diseño.

Tabla 2.2. Factores de Inflación de la Varianza. Proceso de maquinado

Variable del Polinomio	VIF
x_1	1225
x_2	343
x_1x_2	175
${x_1}^2$	1201
${x_2}^2$	193

Dado los resultados en la tabla (2.2) se observa que los VIF de los coeficientes de regresión son mayores a 10 en todas las variables de entrada del proceso (indicio de multicolinealidad), además el número de condición obtenido mediante la ecuación (2.6) es $\eta = 7230$ por mucho mayor a 1000. Estos resultados nos indican que se obtendrán estimadores inestables mediante Mínimos Cuadrados y por consecuencia se generan problemas en las pruebas de hipótesis, como puede ser, cometer el error Tipo 1 (rechazar H_0 cuando no se debería) o el error Tipo 2 (aceptar H_0 cuando no se debería).

Otro caso de estudio que evidencia la presencia de multicolinealidad y con ello los problemas antes mencionados, es un proceso de soldadura PTA (Plasma Transferred Arc) donde intervienen 4 variables de entrada y se estudia la relación contra una variable de salida. Se realizó un diseño de experimentos obteniendo 33 observaciones del proceso, las cuales muestran a continuación:

Tabla 2.3. Observaciones del proceso de soldadura PTA.

x_1	x_2	x_3	x_4	y
Tasa de alimentación de polvo [%]	Tasa de alimentación de polvo [g/min]	Velocidad del proceso [cm/min]	Corriente de soldadura [Amp]	Zona afectada por el calor Área [mm2]
70	26	80	60	4.908
60	21	100	60	5.184
50	18	80	80	4.859
60	21	100	80	3.693
60	21	100	70	3.623
70	26	100	70	3.504
70	26	120	80	3.711
60	21	100	70	3.737
70	26	120	60	3.371
50	18	100	70	3.345
60	21	100	70	3.541
60	21	100	70	4.092
50	18	80	60	4.674
60	21	100	70	4.023
60	21	100	70	3.47
60	21	100	70	3.478
50	18	120	80	3.202
50	18	120	60	3.081
60	21	100	70	3.447
60	21	100	70	4.016
60	21	120	70	2.719
60	21	100	70	3.526
70	26	80	80	4.584
60	21	80	70	4.662
60	21	100	50	3.498
60	21	80	50	4.449
60	21	120	50	3.197
50	18	120	50	2.985
60	21	120	50	3.028
70	26	120	50	2.099
70	26	160	50	1.813
70	26	140	50	2.406
70	26	180	50	1.945

La tabla (2.4) muestra los resultados de los VIF obtenidos a partir de la matriz de diseño, observando como en el caso anterior valores mayores a 10 en dos variables de entrada del proceso. El número de condición para este caso de estudio fue de $\eta = 159$

Tabla 2.4. Factores de Inflación de la Varianza. Soldadura PTA

Variable del Polinomio	VIF
x_1	33.77
x_2	34.86
x_3	1.48
x_4	1.27

A partir de los dos casos de estudio analizados en este trabajo, se concluye que la multicolinealidad es un problema grave que afecta la estimación de los coeficientes del modelo y la inferencia mediante el método Mínimos Cuadrados. Por lo cual, es necesario emplear un tratamiento para la matriz de diseño que obtenga estimadores más estables, que permitan generar un modelo representativo y determinar los factores más importantes del proceso de maquinado.

La regresión Ridge es un método que trata el problema de la multicolinealidad, propuesto por Hoerl et al., (1970a), la idea central es eliminar el requisito de que el estimador β sea insesgado. La propiedad de Gauss-Markov asegura que el estimador mediante Mínimos Cuadrados tiene varianza mínima de entre todos los estimadores lineales insesgados, pero no se garantiza que esa varianza sea pequeña, entonces se puede definir un estimador sesgado de β llamado $\hat{\beta}_R$, que tenga menor varianza que $\hat{\beta}_{MC}$ estimador insesgado de Mínimos Cuadrados obtenido con la Ecuación (2.2). Al definir el Cuadrado Medio del Error del estimador sesgado resulta:

$$
\begin{aligned}
MSE(\hat{\beta}_R) &= E(\hat{\beta}_R - \beta)^2 \\
&= Var(\hat{\beta}_R) + [E(\hat{\beta}_R) - \beta]^2 \\
&= Var(\hat{\beta}_R) + [sesgo\ \hat{\beta}_R]^2 \qquad (2.7)
\end{aligned}
$$

Al aceptar una pequeña cantidad de sesgo en $\hat{\beta}_R$, su varianza se puede hacer pequeña, de tal modo que el Cuadrado Medio del Error de $\hat{\beta}_R$ va a ser menor que el del estimador insesgado de Mínimos Cuadrados $\hat{\beta}_{MC}$ (Piña & Diaz, 2005). En consecuencia, los intervalos de confianza para β serán mucho más angostos mediante el estimador sesgado. Además, la pequeña varianza del estimador $\hat{\beta}_R$ implica que es más estable en comparación con el estimador $\hat{\beta}_{MC}$. El estimador de Ridge se calcula resolviendo una versión modificada de Mínimos Cuadrados (Hoerl & Kennard, 1970a).

$$(X'X + kI)\hat{\beta}_R = X'y \qquad (2.8)$$

El estimador para Ridge del modelo de regresión se obtiene mediante:

$$\hat{\beta}_R = (X'X + kI)^{-1}X'y \qquad (2.9)$$

Donde $k \geq 0$ es una constante que determina el analista, se observa que si $k = 0$ el estimador Ridge es igual al estimador $\hat{\beta}_{MC}$. Ahora si $k > 0$ el sesgo en $\hat{\beta}_R$ incrementa al aumentar k, sin embargo la varianza disminuye al aumentar k. Por lo tanto, es importante escoger un valor de k, tal que la reducción de la varianza sea mayor que el aumento del sesgo. Además el aumento de k trae como consecuencia la disminución del estadístico R^2, debido a esto, el "ajuste del modelo" mediante Regresión Ridge no llegará a ser el mejor, pero lo importante es que se obtendrán estimadores más estables.

El parámetro k de sesgo se puede obtener mediante diversas técnicas, por mencionar algunas, existe la traza de Ridge y un proceso iterativo propuesto por Hoerl et al., (1976) a continuación se ahonda en dichas técnicas:

Para la Traza de Ridge se inicia resolviendo la ecuación (2.9) para diversos valores k de preferencia entre 0 y 1, después se realiza una gráfica, donde se visualiza el comportamiento individual de los coeficiente $\hat{\beta}_R$, la traza ilustrará la inestabilidad de la solución por Mínimos Cuadrados cuando se observen grandes cambios en los coeficientes con pequeños valores de k, se requiere juicio para interpretar el gráfico y seleccionar un valor adecuado de la constante (Montgomery, Peck, & Vining, 2006), debido a que se debe elegir un valor de sesgo k lo suficientemente grande para producir coeficientes estables, pero que no sea innecesariamente grande, porque así se introduce más sesgo y aumenta el Cuadrado Medio del Error, este detalle convierte a la traza de Ridge en un método subjetivo para la elección de k.

El Método Analítico realiza un proceso iterativo propuesto por Hoerl et al., (1976) mediante la fórmula:

$$\hat{\beta}_R(k_i) = k_{i+1} = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_i)\hat{\beta}_R(k_i)} \tag{2.10}$$

Considerando para la primera iteración la solución obtenida mediante Mínimos Cuadrados, el contador iniciaría en $i = 0$, el proceso se detiene cuando se cumpla:

$$\frac{k_{i+1} - k_i}{k_i} < 20T^{-1.3} \tag{2.11}$$

La variable T es la traza de la matriz inversa del polinomio dividida entre el número de variables de entrada, se escoge este criterio de paro debido a que el valor de T aumenta

con la dispersión de los eigenvalores de la matriz $X'X$, permitiendo mayor contracción al aumentar el grado del deterioro en los datos.

Se observa el aumento del Cuadrado Medio del Error (MSE) en ambas técnicas, debido al sesgo agregado, que repercute directamente en el cálculo del error estándar (se), el cual se obtiene de la siguiente manera:

$$se = \sqrt{MSE * C_{jj}} \tag{2.12}$$

Donde C_{jj} son los elementos de la diagonal de la matriz de covarianza de β. Al eliminar la multicolinealidad, la obtención del error estándar estará afectado por el sesgo, el cálculo de este estadístico es muy relevante debido a que es utilizado para calcular t_0, importante para probar la significancia individual de cada coeficiente en el modelo, su ecuación es:

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{2.13}$$

La significancia individual de cada coeficiente son llamadas también "pruebas de hipótesis del modelo", que determinan la importancia de cada coeficiente del modelo de regresión, como cada coeficiente está asociado a un factor del proceso, estas pruebas de hipótesis establece sí los factores son relevantes para el proceso. Por lo tanto, el sesgo afecta gravemente las pruebas de hipótesis del modelo y repercute directamente para determinar la importancia de cada factor en el proceso.

Como se dijo anteriormente, el sesgo también causa un nivel bajo del estadístico R^2, el cual es un valor porcentual que determina la "adecuación general del modelo", se obtiene mediante la ecuación:

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.14}$$

Donde SSE es la Suma de Cuadrados del Error y SST es la Suma de Cuadrados del Total, debido a que SST es fija, al aumentar k aumenta SSE y disminuye R^2, el cual es también llamado coeficiente de determinación del modelo.

Por lo cual, la elección del parámetro de sesgo k toma gran relevancia, es importante determinar la técnica adecuada en la obtención de k o si es pertinente, elaborar otro método para encontrar el parámetro de sesgo que mejore la estabilidad del estimador, de la mano con un menor incremento sobre MSE. Además, en la literatura se menciona la obtención de un valor k para eliminar la multicolinealidad (Hui et al., 2017, Xiang et al., 2017, Li et al., 2010, Qi et al., 2008) y se ha planteado poco la posibilidad de agregar más de un valor distinto de k a la matriz de diseño (El-Dereny et al., 2011). Esta técnica es llamada Regresión Ridge Generalizada (Hoerl et al., 1970) en donde se obtienen parámetros separados de sesgo para cada regresor.

Se transforma el modelo lineal al espacio de los regresores ortogonales para obtener el estimador mediante Regresión Ridge Generalizada (GRR por sus siglas en inglés) de la forma

$$(\Lambda + K)\hat{\alpha}_{GR} = Z'y \tag{2.15}$$

Donde $\Lambda = T'X'TX$ con T como la matriz ortogonal de eigenvectores de X, K es una matriz diagonal con los elementos $(k_1, k_2, \dots, k_p)$ de sesgo, por último $Z = XT$. Para determinar los valores de la matriz K (Hoerl et al., 1970) sugieren un método iterativo iniciando a partir de la solución de Mínimos Cuadrados considerando que la primera iteración es construida a partir de $\hat{\alpha} = \Lambda^{-1}Z'y$, generando así, la matriz con los elementos en la diagonal de la forma:

$$k_j = \frac{\hat{\sigma}^2}{\hat{\alpha}_j} \tag{2.16}$$

Para después comenzar las iteraciones considerando:

$$\hat{\alpha}^i{}_{GR} = \left(\Lambda + K^i\right) Z' y \tag{2.17}$$

Donde $K^i = diag\left(k_1^i,\ k_2^i, \dots, k_p^i\right)$ estos valores se utilizan para corregir los valores de α, el proceso iterativo continua hasta obtener estimadores estables del parámetro. Hasta el momento, no se ha determinado una matriz óptima de K para eliminar la multicolinealidad debido a que depende de parámetros desconocidos $\hat{\sigma}^2$ y $\hat{\alpha}_j$.

Se ha destacado como importante los VIF y η (número de condición) que son medidas que dimensionan el problema de multicolinealidad presente entre las variables de entrada, lo cual afecta la estimación de los coeficientes del modelo $\hat{\beta}$ obtenidos mediante regresión. El agregar valores de sesgo elimina la multicolinealidad, generando así coeficientes más estables, pero afecta y hace que disminuya el estadístico R^2 (coeficiente de determinación) él cual se emplea como criterio que establece el porcentaje de explicación del modelo sobre el proceso.

Es por lo anterior que, se tiene como enfoque dos funciones como objetivo para el modelo de regresión: eliminar la multicolinealidad minimizando los VIF y η además de mantener o maximizar el coeficiente de determinación R^2, lo cual se pretende llevar a cabo partiendo de la idea de la Regresión Ridge Generalizada en donde su idea radica en agregar más de un valor distinto de sesgo en la matriz de diseño. Al contar con más de un objetivo en específico para el modelo de Regresión, es preciso utilizar una optimización multiobjetivo para determinar los valores de sesgo en la matriz K que elimine la multicolinealidad y estabilice el coeficiente de determinación R^2.

Esta herramienta ha sido utilizada por diferentes autores Munner et al., (2015) implementaron una optimización multiobjetivo mediante un algoritmo evolutivo llamado MOGA por sus siglas en inglés (Multi-objective Optimization Genetic Algorithm) debido a que deseaban minimizar dos funciones en específico.

Otra aportación realizada por Jun et al., (2015) aplicando una optimización multiobjetivo mediante otro algoritmo evolutivo de nombre NSGA-II por sus siglas en inglés (Non-dominated Sorting Genetic Algorithm) donde el interés es maximizar la vida útil de la herramienta y al mismo tiempo minimizar la rugosidad de la superficie las cuales son demandas industriales comunes.

Existe además otra ramificación de la optimización multiobjetivo aplicada por ejemplo en Elham et al., (2017) donde utiliza un método de filtro de región de confianza SQP (Sequential Cuadratic Programming) para optimizar la estructura del ala de un avión, teniendo como función objetivo minimizar el consumo de combustible.

Después de haber revisado las aportaciones de diversos autores con respecto a la optimización, además de la eficacia de la regresión Ridge Generalizada para afrontar el problema de multicolinealidad, surgen las siguientes preguntas de investigación.

2.2 Preguntas de investigación

1. ¿Qué impacto tendrá agregar más de un valor k distinto de sesgo en la matriz de diseño?
2. El agregar más de un valor de sesgo k ¿Ayudara a generar estimadores más estables?
3. ¿Qué método de optimización se debe utilizar para encontrar los valores de k que eliminen la multicolinealidad?
4. ¿Se deberá plantear una optimización multiobjetivo o global para obtener los valores k que produzcan menos sesgo a los estimadores?
5. ¿Cuál es el efecto de k sobre las variables del proceso de manufactura?

2.3 Hipótesis

Se establecen las siguientes hipótesis, las cuales se demostraran con la metodología propuesta en esta investigación.

1. Se podrá agregar diferentes valores k en la matriz de diseño para obtener estimadores más estables mediante regresión Ridge.
2. Es posible que la optimización de los valores k elimine la multicolinealidad, disminuya el Cuadrado Medio del Error, además de estabilizar el coeficiente de determinación del modelo R^2.

2.4 Objetivos

Se plantean los siguientes objetivos con la finalidad de obtener mejores modelos y optimizaciones de los procesos que se están investigando.

2.4.1 Objetivo General

Determinar los valores k de sesgo que generen estimadores más estables para el modelo de regresión Ridge, que permitan explicar mejor las variables del proceso de manufactura.

2.4.2 Objetivos Específicos

- Establecer que técnicas se pueden implementar para calcular los valores k de sesgo que eliminen la multicolinealidad.
- Realizar un método de optimización que determine los valores k de sesgo que eliminen la multicolinealidad.
- Aplicar Regresión Ridge a los datos del proceso de manufactura para eliminar la multicolinealidad en la matriz de diseño utilizando los valores de sesgo obtenidos.
- Comparar los estimadores $\hat{\beta}$ de cada modelo a partir de los estadísticos Factores de Inflación de la Varianza (VIF), Cuadrado Medio del Error (MSE) y el Coeficiente de determinación del modelo (R^2).

2.5 Justificación

Debido a la alta competitividad en el mercado industrial, las empresas desean que sus productos cumplan, en gran medida, con las especificaciones establecidas, si no se cumplen dichas especificaciones, se considera que el proceso presenta alta variabilidad; una manera muy recurrente en la industria para tratar de controlar el proceso, es modificando los parámetros de los factores de forma arbitraria, es decir, "a prueba y error", con lo cual, no se garantiza un proceso controlado, debido a la poca fiabilidad de dicha técnica.

Por lo cual, es pertinente realizar un análisis estadístico al proceso industrial, para determinar un modelo que sea representativo, que permita inferir y controlar el proceso, generando así, mayor estabilidad en el proceso industrial.

En la industria, la mayoría de las ocasiones, los factores del proceso están altamente correlacionados, entonces para poder inferir adecuadamente, a partir del modelo de regresión, es importante eliminar la multicolinealidad entre las variables de entrada y obtener estimadores estables, para que las conclusiones del proceso, basadas en el análisis del modelo, sean acertadas.

Capítulo 3

Estado del Arte

Se presenta el estado del arte referente a la problemática de la multicolinealidad, la aplicación de métodos alternativos de regresión a mínimos cuadrados, además de los avances sobre las técnicas de optimización.

3.1 Revisión de Literatura

En (El-Dereny & Rashwan, 2011) se realizó un comparativo de métodos de regresión para eliminar la multicolinealidad, como la Regresión Ridge (RR), Regresión Ridge Generalizada (RRG), Regresión Ridge Dirigida (RRD) y Mínimos Cuadrados (MC), de tal forma que se generan diferentes técnicas para obtener los parámetros de sesgo, todo esto se lleva a cabo con datos simulados, concluyendo que los métodos RR, RRG y RRD son mejores en presencia de multicolinealidad y los modelos RRG y RRD son mejores utilizando como medida el estadístico R^2 también llamado "adecuación general del modelo" con un 91% en comparación con el 77% obtenido mediante MC.

En (Wong & Chiu, 2015) se propuso una nueva técnica para obtener el parámetro k de sesgo, el resultado es comparado con otros 26 métodos existentes y tomando como medida de desempeño de cada método el Cuadrado Medio del Error. Utilizan datos simulados con diferentes escenarios, como considerar diferente número de regresores, considerar los valores de los regresores de forma aleatoria, además variando la colinealidad entre los datos. Concluyendo que en la mayoría de los escenarios la técnica propuesta era mejor en comparación con los demás métodos además de Mínimos Cuadrados.

En (Shengzheng, Baoxian, Jiansen, Wei, & Tie, 2017) se propuso un nuevo modelo de predicción de consumo de combustible para buques empleando el algoritmo de regresión LASSO, debido a que algunas de las variables características están altamente correlacionadas, por lo tanto surge un problema de colinealidad múltiple. Se tuvieron en cuenta tanto el conjunto de datos operativos de buques realistas como los datos meteorológicos, y los resultados de la predicción se compararon cuantitativamente con los datos reales de consumo de combustible. Además, el método propuesto fue validado tanto en precisión de predicción y rendimiento computacional. El método supera a otros métodos tradicionales como ANN (Red Neuronal Artificial), SVR (Regresión Soporte Vector), GP (Programación Genética), y también tiene varias buenas características tales como la capacidad de interpretación, la generalización y la estabilidad numérica.

En (Liu, Miao, Yuan, & Dong, 2017) se estudió el error térmico de máquinas herramienta CNC tipo Leaderway V-450 durante diferentes temporadas, utiliza Regresión Ridge para establecer un modelo de error térmico para inhibir la mala influencia de la colinealidad sobre la robustez pronosticada por error térmico. Proponen el "método de modelado por error térmico de la herramienta de Regresión Ridge de Robustez", "método RRR" abreviado. Además, en dicho método el coeficiente de correlación se usa para medir la correlación entre los puntos sensibles a la temperatura y el error térmico, lo comparan contra otros dos métodos y los resultados muestran que el "método RRR" puede mejorar significativamente la precisión predicha a largo plazo y la solidez del error térmico. Finalmente, el efecto de la aplicación de la compensación práctica muestra que el "método RRR" es utilizable y eficaz.

En (Wang, Liang, Wang, & Zhang, 2017) se crearon modelos empíricos para predecir las fuerzas de la herramienta que actúan sobre una única selección de arrastre, porque son

parámetros básicos en el diseño de las unidades de trabajo de las máquinas de excavación y la evaluación de su rendimiento. Los coeficientes relevantes de la resistencia de la roca se obtuvieron usando el análisis de regresión del componente principal y el análisis de Regresión Ridge para el desarrollo del modelo de fuerza cortante de picos cónicos y radiales, respectivamente. Verifican el rendimiento de predicción de los nuevos modelos mediante la prueba de hipótesis y el análisis de regresión entre los valores de fuerza de corte medidos y predichos. En conclusión se evita la multicolinealidad grave en modelos de regresión lineal múltiple, y los coeficientes de regresión inducidos y las ecuaciones son más razonables físicamente.

En (Demirhan, 2014) propuso un nuevo modelo no lineal para la estimación de la radiación solar horizontal promedio diaria haciendo uso de la técnica de programación genética para superar la multicolinealidad, estimar y predecir con éxito la cantidad de radiación solar diaria promedio. Revisan algunos de los modelos desarrollados para Turquía, donde observan que estos modelos se han identificado como precisos bajo ciertas estructuras de multicolinealidad, y cuando se elimina la multicolinealidad, la precisión de estos modelos es controversial. De acuerdo al modelo propuesto, la variable con mayor impacto relativo en la radiación solar diaria promedio es la altitud. Las variables asociadas a este modelo son: efectos de temperatura, precipitación, altitud, longitud y promedio mensual de radiación solar horizontal, la humedad y la temperatura del suelo no están incluidas en el modelo debido a su alta correlación con precipitación y temperatura, respectivamente. No existe un problema de multicolinealidad en el nuevo modelo, y su exactitud de estimación es mejor que los modelos revisados en términos de numerosas medidas de rendimiento estadístico.

En (Ying-Ze, y otros, 2013) se investigaron los efectos de la colinealidad de la fuente y la presencia de fuentes desconocidas en el modelo NCPCRCMB (balance de masa química

de la regresión de componentes principales no negativos restringidos) que se comparó con USEPA CMB8.2, el modelo NCPCRCMB puede obtener resultados más estables, incluso si distribuye los conjuntos de datos con un fuerte problema de colinealidad, fue propuesto y validado por conjuntos de datos sintéticos, así como un conjunto de datos ambientales de Kaifeng, China. El modelo NCPCRCMB puede tolerar niveles más altos de colinealidad de la fuente como proporciones de fuentes desconocidas. Ambos modelos se realizaron para distribuir las contribuciones de seis fuentes conocidas (Fuente A, polvo de suelo, ceniza volante de combustión de carbón, polvo de cemento, sal marina y acería) a los receptores sintéticos para los 1000 conjuntos de datos sintéticos. En conclusión NCPCRCMB resolver el problema de colinealidad entre los perfiles de origen.

En (Yan-Fu, Min, & Thong-Ngee, 2010) mejoraron el rendimiento de la regresión para la estimación de costos de software en conjuntos de datos multicolineales. Para lograr este objetivo, propusieron un enfoque holístico (ARR) que combina transformación de datos, diagnóstico de multicolinealidad, regresión Ridge y optimización multiobjetivo. Con la técnica de optimización multiobjetivo, ARR es capaz de maximizar la precisión de la estimación y minimizar la colinealidad múltiple con cierto equilibrio predeterminado entre estos dos criterios. Las comparaciones contra Red Neuronal Artificial (ANN), Arboles de Clasificación y Regresión (CART) y Razonamiento Basado en Casos (CBR) revelan que la regresión Ridge puede lograr predicciones igualmente buenas o incluso mejores que los métodos de aprendizaje automático y la regresión Ridge tienen mejores interpretaciones que los métodos de aprendizaje automático. Los resultados muestran que la regresión Ridge podría ser una mejora prometedora de las regresiones en el contexto de la estimación de costos.

En (Li & Shao, 2008) se investigo la estimación de la composición de destilación en línea en base a un nuevo método de análisis de matriz sensible y regresión Ridge Kernel para implementar la estimación en línea de composiciones de destilación, analiza la matriz de sensibilidad para seleccionar las variables secundarias más adecuadas para usarlas como entradas del estimador. Mediante la elección óptima de las segundas variables y la construcción del modelo de composición KRR, el resultado de la simulación muestra que el método es eficiente. Con el desarrollo de los estimadores de composición, la investigación permite implementar un control avanzado de las variables de calidad del proceso de destilación en línea. Por lo anterior, la instalación exitosa de los estimadores de composición en una refinería existente puede garantizar un mejor control de calidad del producto con mayor productividad.

En (Lipovetsky & Conklin, 2000) desarrollaron una nueva técnica multivariada para producir regresiones con coeficientes interpretables que se aproximan y tienen los mismos signos que los coeficientes de regresión pairwise. Utilizando un enfoque multiobjetivo para incorporar regresiones múltiples y por parejas en un objetivo, reduciendo la técnica a un problema propio que representa un híbrido entre la regresión y el análisis de componentes principales. Muestra que el enfoque corresponde a un esquema específico de regresión Ridge con una matriz total añadida a la matriz de correlaciones. Esta solución produce un valor menor de determinación múltiple R^2 pero un mejor conjunto de coeficientes en comparación con regresión regular.

En (Serkan & Rasit, 2019) utilizaron un algoritmo de colonia de abejas artificiales (ABC) para resolver la Cinemática Inversa (CI) de un brazo robótico de 7 grados de libertad, mencionan difícil la CI dado que convertir la posición y orientación del efector final del manipulador del robot del espacio cartesiano al espacio articular es una ecuación más

compleja, no lineal e imposible de resolver por métodos convencionales. El algoritmo ABC lo han utilizado para la solución de la CI y sus resultados se analizaron en términos de error de posición y tiempo de cálculo, sus resultados los compararon con la optimización de enjambre de partículas. El problema de optimización fue encontrar el valor del ángulo óptimo para cada articulación con la coordenada cartesiana inicial dada y la coordenada objetivo, las simulaciones mostraron que el algoritmo ABC busco con éxito los ángulos de articulación óptimos del manipulador robótico, además para la precisión del algoritmo aplicaron un segundo escenario en 100 pruebas diferentes.

En (Shakya, Mishra, Maity, & Santarsiero, 2019) anexaron el algoritmo Colonia de Hormigas (ACO por sus siglas en inglés) a los patrones de búsqueda Hooke-Jeeves, los cuales son utilizados en ingeniería civil para el diagnóstico oportuno de daños a infraestructuras. Consideraron cuatro casos de estructuras: viga en voladizo, armadura plana, edificio de dos pisos y un marco 2D, los cuales tienen diversa complejidad para escenarios de daño de bajo y alto nivel; los daños se consideran desde la fatiga, deterioro del material hasta daños por accidentes o terremotos. Realizaron un comparativo con otros métodos antes utilizados por otros investigadores como GA y PSO, concluyen que su modelo propuesto obtuvo los mejores resultados debido a que es más preciso, basándose en el porcentaje de error como medida de comparación.

En (Guanghui & Xiaohong, 2017) incorporaron la técnica de Optimización de Enjambre de Partículas (PSO) para identificar los parámetros físicos de un sistema ETC (Electronic Throttle Control) real de un vehículo, dicho mecanismo es utilizado para que el motor funcione de manera eficiente, mejorando la capacidad de conducción, el ahorro de combustible y el rendimiento de emisiones del vehículo. El problema es que el rendimiento del control del sistema ETC es afectado por los inciertos parámetros físicos del sistema

relacionados con la fricción, el resorte de retorno, la reacción del engranaje, el proceso de su producción y el envejecimiento de los componentes. Su estrategia mediante PSO garantizo una velocidad de seguimiento satisfactorio de la apertura del acelerador ETC incluso al imponer valores límites reales de voltaje, además verificaron su estrategia de control comparando con otras existentes en el entorno MATLAB/Simulink como también por sus resultados experimentales llevados a cabo en una plataforma de prueba de hardware realizando varios casos operativos reales.

Se concluye del estado del arte:

La multicolinealidad es un serio problema que afecta la utilidad de los modelos de regresión disminuyendo su poder para describir las variables, estimar los parámetros del modelo, predecir nuevos datos y controlar el proceso.

Además la regresión Ridge es una vía para eliminar la multicolinealidad, obteniendo estimadores más estables a partir de un parámetro de sesgo k, no se ha profundizado en la posibilidad de agregar más de una constante k para que el sesgo no sea tan severo en el modelo de regresión.

Finalmente la optimización utilizando algoritmos basados en la naturaleza es de gran ayuda debido a su practicidad de uso, además dado el planteamiento del problema (agregar más de un valor de sesgo) estos algoritmos permiten obtener soluciones óptimas en un espacio de búsqueda determinado.

Capítulo 4

Marco Teórico

A continuación se realiza una introducción al marco de referencias y conceptual de los temas que se abordarán y se utilizarán en la investigación.

4.1 Regresión Lineal Múltiple

Para modelar mediante regresión lineal múltiple se define la ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{4.1}$$

De donde y es la variable respuesta, x_i son las variables regresoras, β_j son los coeficientes de regresión, ε es el error de estimación.

Si se disponen de n observaciones $n > k$, y_i es la i-ésima respuesta observada, x_{ij} es la i-ésima observación o regresor de x_i. Suponer que el error ε del modelo tiene $E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$ y que los errores no están correlacionados. Se considera:

Tabla 4.1. Estructura de un sistema de ecuaciones generado.

Observaciones	Respuesta	Regresores			
i	y	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Entonces la ecuación (4.1) se puede escribir de la siguiente forma

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \ , \qquad i = 1,2, \dots, n \tag{4.2}$$

La función de mínimos cuadrados es

$$\begin{aligned} S(\beta_0, \beta_1, \dots, \beta_k) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \end{aligned} \tag{4.3}$$

Se debe minimizar la función S respecto a los estimadores $\beta_0, \beta_1, \dots, \beta_k$ de tal forma que por mínimos cuadrados deben satisfacer

$$\left. \frac{\partial S}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right) = 0$$

$$\left. \frac{\partial S}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right) x_{ij} = 0 \ , \qquad j = 1,2, \dots, k \tag{4.4}$$

Al simplificar, se obtienen las ecuaciones normales de mínimos cuadrados.

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ \hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i \end{aligned} \tag{4.5}$$

Es más comodo representar el modelo en notación matricial, al quedar más compacto los datos y los resultados, quedando planteado:

$$y = \boldsymbol{X}\beta + \varepsilon \tag{4.6}$$

En donde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

y es un vector de $nx1$ respuestas, X es una matriz de nxk variables regresoras, β es un vector de $kx1$ coeficientes de regresión, ε es un vector de $nx1$ errores aleatorios.

Los coeficientes de regresión β mediante mínimos cuadrados minimizando son:

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\,\varepsilon = (y - X\beta)'\,(y - X\beta) \tag{4.7}$$

Desarrollando $S(\beta)$

$$S(\beta) = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \tag{4.8}$$

Donde $\beta'X'y$ resulta ser una matriz de 1x1 (un escalar) y su transpuesta $(\beta'X'y)' = y'X\beta$ siendo el mismo escalar, por lo que se simplifica

$$S(\beta) = \; y'y - 2\,\beta'X'y + \beta'X'X\beta \tag{4.9}$$

Los estimadores por mínimos cuadrados deben satisfacer

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} \; = \; -2X'y + 2X'X\hat{\beta} \; = 0 \tag{4.10}$$

Simplificando

$$X'X\hat{\beta} = X'y \tag{4.11}$$

Siendo estas las ecuaciones normales de mínimos cuadrados, para resolverlas se multiplica ambos lados por la inversa $X'X$, quedando el estimador β por MC.

$$\hat{\beta} = (X'X)^{-1}X'y \tag{4.12}$$

Para solucionar la ecuación 4.12 es necesario que la matriz X sea linealmente independiente, de no ser así la matriz podría presentar dependencia lineal perfecta o dependencia casi lineal, a continuación se abordara cada uno de los conceptos en términos matemáticos.

4.2 Independencia Lineal

Una matriz es linealmente independiente si y solo si la única solución al sistema.

$$\sum_{j=1}^{p} t_j X_j = 0 \tag{4.13}$$

Es la solución trivial $t_1 = t_2 = \cdots = t_p = 0$.

El rango de una matriz es igual a su número de columnas.

$$rango(X) = p$$

Si una matriz es $L.I.$ entonces su rango es igual a su número de columnas

$$Si\ X\ es\ L.I. \Rightarrow\ rango(X) = p$$

Una matriz $X'X$ es invertible si y solo si su rango es igual al número de columnas.

$$X'X\ es\ invertible\ \Leftrightarrow\ rango(X'X) = p$$

Una matriz $X'X$ es invertible si y solo si existe una matriz $(X'X)^{-1}$ tal que su multiplicación es igual a la matriz identidad.

$$X'X\ es\ invertible \iff (X'X)(X'X)^{-1} = (X'X)^{-1}(X'X) = I$$

Si una matriz $X'X$ es invertible entonces su determinante es distinto de cero.

$$X'X\ es\ invertible \implies |X'X| \neq 0$$

Si el determinante de $X'X$ es distinto de cero entonces se puede calcular su inversa.

$$|X'X| \neq 0 \implies (X'X)^{-1} = \frac{1}{|X'X|} Adj(X'X)$$

4.3 Dependencia Lineal

Una matriz X con vectores columna $[X_1, X_2, \dots, X_p]$ son linealmente dependiente si y solo si la solución al sistema $\sum_{j=1}^{p} t_j X_j = 0$ es exacta para algunos $t_j \neq 0$.

$$X\ es\ L.D. \Leftrightarrow \sum_{j=1}^{p} t_j X_j = 0\ , con\ algun\ t_j \neq 0$$

Si X es una matriz linealmente dependiente entonces $X'X$ también es linealmente dependiente.

$$Si\ X\ es\ L.D. \Rightarrow X'X\ tambien\ es\ L.D.$$

Si $X'X$ una matriz de dimensión nxp es linealmente dependiente, entonces su rango es menor que su número de columnas.

$$Si\ X'X\ es\ linealmente\ dependiente \Rightarrow rango(X) < p$$

La matriz $X'X$ es no invertible si y solo si su rango es menor al número de columnas.

$$X'X\ es\ no\ invertible\ \Leftrightarrow\ rango(X'X) < p$$

Si una matriz $X'X$ es no invertible entonces su determinante es igual a cero.

$$X'X\ es\ no\ invertible\ \Longrightarrow |X'X| = 0$$

Si una matriz $X'X$ es no invertible entonces no existe una matriz $(X'X)^{-1}$ tal que su multiplicación es igual a la matriz identidad.

$$X'X\ es\ no\ invertible\ \Longrightarrow \nexists (X'X)^{-1}\ talq\ que\ (X'X)^{-1}(X'X) = I$$

4.4 Dependencia Casi Lineal

Una matriz X con vectores columna $[X_1,\ X_2, \dots, X_p]$ son casi linealmente dependiente si y solo si la solución al sistema $\sum_{j=1}^{p} t_j X_j = 0$ es aproximada para algunos $t_j \neq 0$.

$$X\ es\ casi\ L.D. \Leftrightarrow \sum_{j=1}^{p} t_j X_j \approx 0\ , con\ algun\ t_j \neq 0$$

Existe multicolinealidad aproximada cuando las columnas de la matriz X son casi linealmente dependientes.

Si X es una matriz casi linealmente dependiente entonces $X'X$ también es casi linealmente dependiente.

$$Si\ X\ es\ casi\ L.D. \Rightarrow X'X\ tambien\ es\ casi\ L.D.$$

Por lo tanto, si $X'X$ es casi linealmente dependiente entonces su rango es igual a su número de columnas

$$Si\ X'X\ es\ casi\ L.D. \Rightarrow\ rango(X'X) = p$$

Una matriz $X'X$ es casi singular si y solo si su rango es igual al número de sus columnas.

$$X'X\ es\ casi\ singular\ \Leftrightarrow\ rango(X'X) = p$$

Si una matriz $X'X$ es casi singular entonces su determinante es muy próximo a cero.

$$X'X\ es\ casi\ singular \Longrightarrow |X'X| \approx 0$$

Si el determinante de $X'X$ es próximo a cero entonces se puede calcular su inversa mediante el método de la matriz adjunta.

$$|X'X| \approx 0\ \Longrightarrow\ (X'X)^{-1} = \frac{1}{|X'X|} Adj(X'X)$$

4.5 Técnicas para detectar la Multicolinealidad

4.5.1 Matriz de Correlación

Para detectar la multicolinealidad se utiliza la técnica de la matriz de correlación que inspecciona los elementos r_{ij} no diagonales de la matriz $X'X$, de tal forma que si los regresores x_i y x_j son casi linealmente dependientes los elementos $|r_{ij}|$ serán próximos a la unidad. Esto se lleva a cabo mediante la formula

$$r_{x_i x_j} = \frac{\sigma_{x_i x_j}}{\sigma_{x_i}\sigma_{x_j}} \qquad (4.14)$$

De donde

$r_{x_i x_j}$ es la correlación que existe entre los elementos x_i y x_j, $\sigma_{x_i x_j}$ es la covarianza de x_i y x_j, σ_{x_i} es la desviación estándar de x_i, σ_{x_j} es la desviación estándar de x_j

4.5.2 Factores de Inflación de la Varianza

Los factores de inflación de la varianza (VIF) miden el efecto combinado que tienen las dependencias entre los regresores sobre la varianza de ese término, si los VIF obtenidos son mayores que 10 entonces existe multicolinealidad, se calculan:

$$VIF_j = \frac{1}{(1 - R_j^2)} \tag{4.15}$$

De donde:

R_j^2 es el coeficiente obtenido cuando se hace la regresión de x_j

4.5.3 Análisis del Eigensistema

Al trabajar con el eigensistema de la matriz se detecta la multicolinealidad, al detectar uno o más valores propios pequeños, implica dependencia entre las columnas. Se lleva a cabo examinando el número de condición de $X'X$ mediante la fórmula:

$$k = \frac{\lambda_{max}}{\lambda_{min}} \tag{4.16}$$

Donde

λ_{max} es el máximo valor propio de la matriz $X'X$, λ_{min} es el mínimo valor propio de la matriz $X'X$. Si k es menor que 100 no existen problemas graves de multicolinealidad. Si k esta entre 100 y 1000 implica multicolinealidad de moderada a fuerte. Si k es mayor que 1000 hay fuerte multicolinealidad.

Para los índices de condición de la matriz $X'X$ se utiliza la fórmula:

$$k = \frac{\lambda_{max}}{\lambda_j} \quad para\ j = 1,2,\dots,p \tag{4.17}$$

De donde:

λ_j es el j-esimo valor propio de la matriz $X'X$

Con índices de condición mayores que 1000 implica dependencia casi lineal.

4.6 Técnicas para tratar la Multicolinealidad

4.6.1 Regresión Ridge

El problema de Mínimos Cuadrados es que $\hat{\beta}$ sea un estimador insesgado de β. La propiedad de Gauss-Markov asegura que el estimador de Mínimos Cuadrados tiene varianza mínima, pero no hay garantía que sea pequeña. Para aliviar el problema se elimina el requisito de que β sea insesgado. Suponer que se puede determinar un estimador sesgado de β por decir un $\hat{\beta}^*$ que tenga menor varianza que el estimador insesgado $\hat{\beta}$, el error cuadrático medio del estimador $\hat{\beta}^*$ se define:

$$MSE(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2 = Var(\hat{\beta}^*) + [E(\hat{\beta}^*) - \beta]^2 \tag{4.18}$$

Es decir:

$$MSE(\hat{\beta}^*) = Var(\hat{\beta}^*) + (sesgo\ en\ \hat{\beta}^*)^2 \tag{4.19}$$

La pequeña varianza del estimador sesgado implica también que $\hat{\beta}^*$ es un estimador más estable de β que el estimador insesgado $\hat{\beta}$. La regresión Ridge es un procedimiento para obtener estimadores sesgados de coeficientes de regresión, se determina resolviendo una versión modificada de las ecuaciones normales, el estimador Ridge $\hat{\beta}_R$ se define como la solución de:

$$(X'X + kI)\hat{\beta}_R = X'y \tag{4.20}$$

Que es:

$$\hat{\beta}_R = (X'X + kI)^{-1}X'y \tag{4.21}$$

Donde $k \geq 0$ es el parámetro de sesgo (constante que selecciona el analista)

Para determinar el valor de k se puede recurrir a las siguientes técnicas:

- **Traza de Ridge:** es una gráfica de los elementos $\hat{\beta}_R$ en función de k lo más recomendable es tomar valores entre 0 y 1 (Montgomery, Peck, & Vining, 2006).
- **Fórmula analítica:** propuesta por Hoerl et al.(1975)

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{4.22}$$

Mediante un proceso iterativo por Hoerl et al., (1976)

$$\hat{\beta} = k_0 = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$

$$\hat{\beta}_R(k_0) = k_1 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_0)\hat{\beta}_R(k_0)}$$

$$\hat{\beta}_R(k_1) = k_2 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_1)\hat{\beta}_R(k_1)}$$

$$\vdots$$

El proceso se detiene cuando

$$\frac{k_{j+1} - k_j}{k_j} < 20T^{-1.3} \tag{4.23}$$

Siendo

$$T = Tr(X'X)^{-1}/p \tag{4.24}$$

La matriz de covarianza de $\hat{\beta}_R$ es

$$Var(\hat{\beta}_R) = \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1} \tag{4.25}$$

El Cuadrado Medio del Error es

$$MSE(\hat{\beta}_R) = Var(\hat{\beta}_R) + (sesgo\ en\ \hat{\beta}_R\ )^2$$

De lo que resulta

$$MSE(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\beta'(X'X + kI)^{-2}\beta \tag{4.26}$$

4.6.2 Regresión Ridge Generalizada

Se transforma el modelo lineal $y = \boldsymbol{X}\beta + \varepsilon$ al espacio de los regresores ortogonales para obtener el estimador mediante Regresión Ridge Generalizada (GRR por sus siglas en inglés) considerando que Λ es una matriz diagonal de pxp cuyos elementos de la diagonal principal son los eigenvalores $\lambda_1, \lambda_1, \dots, \lambda_1$ de $X'X$ y T es la matriz ortogonal correspondiente de eigenvectores, es decir:

$$T'X'XT = \Lambda \tag{4.27}$$

Si se considera

$$\mathrm{Z} = \mathrm{XT} \tag{4.28}$$

$$\alpha = \mathrm{T}'\beta \tag{4.29}$$

El modelo lineal se transforma

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= (ZT)(T\alpha) + \varepsilon \\ &= Z\alpha + \varepsilon \end{aligned} \tag{4.30}$$

El estimador de α por Mínimos Cuadrados queda de la forma

$$\begin{aligned} (Z'Z)\hat{\alpha} &= Z'y \\ \Lambda\hat{\alpha} &= Z'y \\ \hat{\alpha} &= \Lambda^{-1}Z'y \end{aligned} \tag{4.31}$$

Por lo tanto el estimador Ridge Generalizado es la solución de

$$(\Lambda + K)\hat{\alpha}_{RG} = Z'y \tag{4.32}$$

Donde K es una matriz diagonal con los elementos $(k_1, k_2, \dots, k_p)$. Para la elección de los parámetros de sesgo en K se considera el Cuadrado Medio del Error en la Regresión Ridge Generalizada

$$\begin{aligned} MSE(\hat{\beta}_{RG}) &= E\left[(\hat{\beta}_{RG} - \beta)'(\hat{\beta}_{RG} - \beta)\right] \\ &= E[(\hat{\alpha}_{RG} - \alpha)'(\hat{\alpha}_{RG} - \alpha)] \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k_j)^2} + \sum_{j=1}^{p} \frac{\alpha_j^2 k_j^2}{(\lambda_j + k_j)^2} \end{aligned} \tag{4.33}$$

El primer término del lado derecho de la ecuación (4.33) es la suma de las varianzas de los estimadores de parámetro y el segundo término es el sesgo elevado al cuadrado. El cuadrado medio del error de la ecuación (4.33) se minimiza al escoger adecuadamente los valores de sesgo k_j de la ecuación

$$k_j = \frac{\hat{\sigma}^2}{\hat{\alpha}_j}, \qquad j = 1,2,\dots,p \tag{4.34}$$

Se sugiere un método iterativo (Hoerl & Kennard, 1970a) a partir de la solución de Mínimos Cuadrados se obtiene un estimado inicial de las k_j es decir:

$$k_j^0 = \frac{\hat{\sigma}^2}{\hat{\alpha}_j}, \qquad j = 1,2,\dots,p$$

Para después comenzar las iteraciones del método considerando:

$$\hat{\alpha}^i{}_{RG} = \left(\Lambda + K^i\right)Z'y \tag{4.35}$$

Donde $K^i = diag\left(k_1^i,\ k_2^i, \dots, k_p^i\right)$ estos valores se utilizan para corregir los valores de α, el proceso iterativo continua hasta obtener estimadores estables del parámetro.

4.7 Optimización

La optimización es parte importante de la investigación de operaciones, su finalidad consiste en encontrar el valor de las variables para hacer óptima la función objetivo satisfaciendo un conjunto de restricciones, se puede expresar matemática como

$$\min_x f(x)$$

$$s.a \quad g_i(x) \leq 0$$

Para el proceso de búsqueda de la solución óptima existen tres mecanismos los cuales son: las técnicas analíticas, de enumeración y de búsqueda heurística. La búsqueda analítica se basa en cálculos. Los algoritmos de búsqueda pueden guiarse por el gradiente o la arpillera de la función, lo que lleva a una solución mínima local. La búsqueda y enumeración aleatorias son métodos de búsqueda no guiados que simplemente enumeran el espacio de búsqueda y buscan exhaustivamente la solución óptima. La búsqueda heurística es una búsqueda guiada que en la mayoría de los casos produce soluciones de

alta calidad. (Du & Swamy, 2016). De acuerdo al número de funciones, se puede clasificar a la optimización como: optimización global y optimización multiobjetivo.

4.7.1 Optimización Global

Consiste en encontrar los mejores conjuntos de parámetros que optimizan una función objetivo dada, sin embargo, solo se pueden dar condiciones de optimalidad global bajo la restricción de tener una función objetivo y una región factible, convexa ambas, esto hace muy difícil resolver exactamente los problemas de optimización global. La formulación matemática del problema de optimización global es:

$$\begin{aligned} &\min f(x) \\ &s.a \quad g_i(x) = 0, \\ &\qquad h_j(x) \leq 0, \end{aligned}$$

4.7.2 Optimización Multiobjetivo

La mayor parte de los problemas de optimización del mundo real son naturalmente multiobjetivo. Esto es, suelen tener dos o más funciones objetivo que deben satisfacerse simultáneamente y que posiblemente están en conflicto entre sí. Sin embargo, a fin de simplificar su solución, muchos de estos problemas tienden a modelarse como mono-objetivo usando solo una de las funciones originales y manejando las adicionales como restricciones. El problema de optimización multiobjetivo se puede formular como:

Encontrar el vector

$$\vec{x}^* = [x_1^*, x_2^*, \dots, x_n^*]^T$$

Que satisfaga las m restricciones de desigualdad:

$$g_i(\vec{x}) \geq 0 \qquad i = 1,2, \dots, m$$

Con p restricciones de igualdad

$$h_i(\vec{x}) = 0 \qquad i = 1,2, \dots, p$$

Y que optimice

$$\vec{f}(\vec{x}) = [f_1(\vec{x}), f_2(\vec{x}), \dots, f_k(\vec{x})]^T$$

4.7.3 Algoritmos Inspirados en la naturaleza

Existen problemas para los que no se puede garantizar encontrar una solución óptima en un tiempo razonable y estos se clasifican según la teoría de la complejidad computacional como "difíciles". Cuando se aborda un problema difícil, evaluar parte de las soluciones y no todo el conjunto de soluciones, es una estrategia de búsqueda de una solución aplicada por los algoritmos inspirados en la naturaleza, también son llamados algoritmos bio-inspirados. Estos sacrifican la garantía de encontrar la mejor solución y son capaces de encontrar soluciones "buenas" con un tiempo y consumo de recursos computacionales aceptables. La forma de operación de un algoritmo bio-inspirado es una búsqueda continua por mantener el equilibrio entre diversificación e intensificación. El primero se refiere a la exploración de nuevas regiones del espacio de búsqueda, mientras que el segundo a la explotación de alguna región concreta. La existencia de este balance, identifica rápidamente las regiones prometedoras del espacio de búsqueda y evita el consumo de tiempo en las regiones que ya han sido exploradas o que no contienen soluciones de alta calidad.

A continuación se da una breve explicación de algunos algoritmos inspirados en la naturaleza.

a) Ant Colony Optimization (ACO)

Es una metaheurística de optimización basada en colonia de hormigas y fue propuesta por Marco Dorigo en 1992 en su tesis doctoral, como un método para resolver problemas de optimización combinatorios duros (POCs).

El investigador S. Goss en 1989 desarrollo un modelo para explicar el comportamiento observado en un experimento del puente de dos brazos, el modelo determina que después de t unidades de tiempo, desde el comienzo del experimento, m_1 hormigas han

usado el primer brazo del puente y m_2 el segundo, la probabilidad p_1 para la $(m+1)-$ $ésima$ hormiga de elegir el primer brazo del puente está dada por:

$$p_{1(m+1)} = \frac{(m_1 + k)^h}{(m_1 + k)^h + (m_2 + k)^h} \tag{4.36}$$

Donde los parámetros k y h son necesarios para ajustar el modelo a los datos experimentales. La probabilidad para que la misma hormiga elija el segundo brazo es $p_{2(m+1)} = 1 - p_{1(m+1)}$, utilizó simulación Monte Carlo para probar si el modelo corresponde a los datos reales obteniendo muy buen ajuste para valores de $k \approx 20$ y $h \approx 2$.

b) Artificial Bee Colony (ABC)

El algoritmo fue propuesto por Dervis Karaboga en 2005 y está basado en el comportamiento de las abejas, diseñado originalmente para problemas de optimización numérica sin restricciones, aunque puede ser utilizado para resolver problemas de combinatoria. El proceso de búsqueda de néctar por parte de las abejas es un proceso de optimización y su comportamiento se modela como una heurística de optimización basada en el modelo bilógico y consta de los siguientes elementos:

1. Fuentes de alimento: es un valor numérico que indica su potencial.
2. Abejas recolectoras empleadas: explotan una fuente de alimento, son encargadas de comunicar su ubicación y rentabilidad a las abejas observadoras.
3. Abejas recolectoras desempleadas: se encuentran buscando fuentes de alimento para explotar, se dividen en dos tipos:
    a) Las exploradoras: buscan nuevas fuentes de alimento
    b) Las observadoras: esperan en la colmena para elegir alguna de las fuentes de alimento que se encuentran en el proceso de exploración por las abejas empleadas.

El algoritmo básico se describe como las fuentes de alimento que representan a cada solución como un vector $\mathbb{D} - dimensinal$ y a las abejas como los operadores de variación, ya que cuando ellas visitan las fuentes de alimento, calcularán una nueva solución.

$$\vec{v}_{i,g} = \vec{x}_{i,g} + \phi\left(\vec{x}_{i,g} - \vec{x}_{k,g}\right) \tag{4.37}$$

Donde $\vec{x}_{i,g}$ representa la fuente de alimento donde se encuentra la abeja en ese momento, $\vec{x}_{k,g}$ es una fuente de alimento seleccionada aleatoriamente y diferente de $\vec{x}_{i,g}$, g es el ciclo actual y ϕ es un número real aleatorio entre $[-1,1]$, $\vec{v}_{i,g}$ denota la ubicación de la nueva solución con respecto a la posición actual $\vec{x}_{i,g}$.

Las abejas observadoras seleccionan las fuentes de alimento de acuerdo a una probabilidad p_i asociada a la fuente de alimento, la cual es calculada de la siguiente manera:

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \tag{4.38}$$

Donde fit_i es el valor de aptitud de la solución y este valor es proporcional a la cantidad de néctar que tiene la solución i. Los parámetros del algoritmo son los siguientes:

1. SN: es el número de fuentes de alimento.
2. MCN: es el número total de iteraciones que ejecutara el algoritmo.
3. Límite: es el número de ciclos que será conservada una solución sin mejora antes de ser reemplazada por una nueva solución generada por una abeja exploradora.

c) Particles Swarm Optimization. (PSO)

El algoritmo de optimización por enjambre de partículas (PSO), fue originalmente desarrollado por James Kennedy y Russell Eberhart en 1995, es un algoritmo del área de la inteligencia artificial de la rama de inteligencia de enjambres. Está inspirada en el comportamiento social de los seres vivos. El algoritmo de optimización por enjambre de

partículas se inspira en la evolución en el comportamiento colectivo, principalmente trata de imitar el comportamiento social de varios grupos de animales como lo son los cardúmenes, parvadas, manadas, etc. Esto método no garantiza dar el mejor resultado, pero sí un resultado aceptable. Otra característica de este algoritmo, es que es no determinista (estocástico), esto quiere decir que los resultados obtenidos no siempre serán los mismos aunque se trate de una misma función.

Este algoritmo pretende representar el proceso natural de comunicación grupal para compartir conocimiento individual cuando grupos de animales se desplazan, migran o cazan. Si un miembro detecta un camino deseable para desplazarse, el resto de la colonia lo sigue inmediatamente. En PSO, este comportamiento animal es imitado por partículas con ciertas posiciones y velocidades en un espacio de búsqueda, donde la población es llamada swarm, y cada miembro del swarm es llamado partícula. La población inicial se determina aleatoriamente y cada partícula se desplaza a través del espacio de búsqueda y recuerda la mejor posición que ha encontrado. Cada partícula comunica las buenas posiciones a las demás y dinámicamente ajustan su propia posición y su velocidad con base en las buenas posiciones. La velocidad se ajusta con el comportamiento histórico de las partículas. De esta forma, las partículas tienden a dirigirse hacia un mejor espacio de búsqueda en el proceso de minimización de la función objetivo. Este procedimiento de búsqueda se describe:

$$v_i^{k+1} = w \cdot v_i^k + c_1 \cdot rand_1 \cdot \left(pbest_i - x_i^k\right) + c_2 \cdot rand_2 \cdot \left(gbest - x_i^k\right) \qquad (4.39)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \qquad (4.40)$$

Donde c_1 y c_2 son constantes positivas, definidas como coeficientes de aceleración; w es el factor inercial; $rand_1$ y $rand_1$ son dos números aleatorios (con distribución de probabilidad uniforme) en el rango [0,1]; x_i^k representa la $i - esima$ partícula y $pbest$ la

mejor posición previa de x_i^k; $gbest$ es la posición de la mejor partícula de toda la población; y v_i^k es la razón de cambio de la posición (velocidad) de la partícula x_i^k. Los cambios de velocidad se componen de tres partes: momentum, cognitiva y social. De esta forma se obtiene una velocidad que tiende a acercar la partícula a $pbest$ y $gbest$.

El algoritmo de optimización de enjambre de partículas original ha tenido varios cambios y han surgido variaciones del mismo según el problema que se quiera resolver. Sin embargo, en 2006, se trató de establecer un estándar (SPSO), y que posteriormente se le ha contribuido con algunas sugerencias y cambio en el 2007 y 2011. Comparado con el algoritmo clásico, el estándar, agregó el factor social, cognitivo, de inercia y constricción. También se cambió el modo en que las partículas se comunicaban por medio de topologías definidas.

A continuación se describirán los factores utilizados en el algoritmo estándar.

a) Cognitiva y social

   La cognitiva contribuye para que la partícula tenga una especie de memoria y pueda saber si anteriormente había adoptado una mejor posición.

   El factor social, es el responsable de que la partícula sea influenciada por otras partículas en una mejor posición, provocando ser atraída al óptimo encontrado.

b) Topología gbest y lbest

   El concepto del modelo gbest es que todas las partículas se comunican entre sí y el mejor punto encontrado, es comunicado a las demás partículas. Tiene la ventaja de ser muy rápido para encontrar un valor óptimo, pero esto genera un problema, ya que puede provocar una convergencia prematura en un óptimo local, que queremos evitar en la mayoría de los casos.

En contraste, el modelo lbest es más lento pero hay más posibilidad de evitar una convergencia prematura, en la práctica, el modelo mantiene varios puntos de atracción que hace escapar de óptimos locales.

c) Inercia y constricción

La inercia es un factor que se añadió para dar mayor estabilidad a la fórmula, permite que la partícula siga una trayectoria constante ignorando la influencia de otros.

El factor de constricción, a diferencia de la inercia, es variable, permite un equilibrio entre búsquedas locales y globales. Cuando este factor es menor a 4, el enjambre se mueve lentamente y tiene una convergencia lenta, mientras que si es mayor a 4, logra una convergencia rápida.

El factor de inercia modulada, brinda un comportamiento similar al de la constricción. Consiste en comenzar con una inercia alta, e ir disminuyendo con el paso del tiempo. Está comprobado que este factor da mejores resultados que una inercia constante.

Capítulo 5

Metodología

El siguiente esquema describe la secuencia de actividades necesarias para llevar a cabo la elaboración de la investigación.

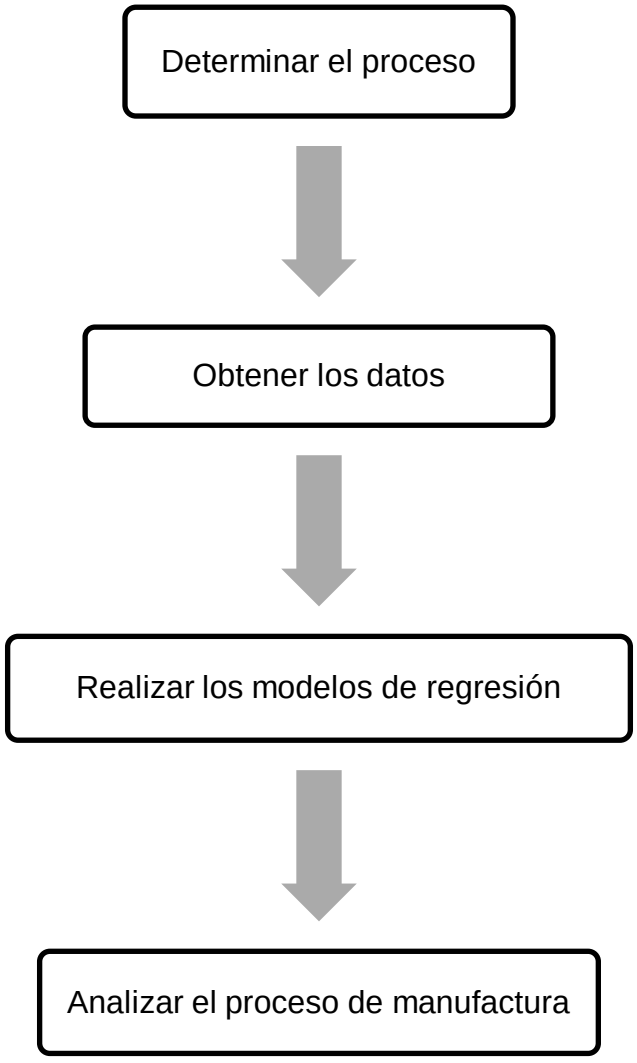

Figura 5.1 Metodología general

5.1 Determinar el proceso

Seleccionar un proceso de manufactura en específico para la investigación, debido a la gran cantidad de procesos que existen y porque cada proceso tiene condiciones particulares. Además, establecer las variables involucradas durante el proceso, llevando esta tarea en base a una consulta de especialistas además de investigación científica.

5.2 Obtener los datos

Los datos del proceso se pueden obtener mediante un diseño de experimentos o por fuentes históricas, para esta investigación los datos obtenidos son proporcionados por la empresa debido a que ya se habían recolectado anteriormente, dichos datos son utilizados para realizar el modelo de regresión.

5.3 Realizar los modelos de regresión

Para llevar a cabo este paso se realiza lo siguiente:

Se genera una matriz de diseño para posteriormente aplicar la técnica de los Factores de Inflación de la Varianza (VIF), que indica la presencia de multicolinealidad entre los factores del proceso, si los valores obtenidos son mayores a 10.

Se considera la Regresión Ridge para eliminar el problema de la multicolinealidad, se cuenta con diferentes técnicas para elegir un valor k de sesgo, cada una de las técnicas genera distintas soluciones.

Se realiza un método de optimización para calcular los valores k de sesgo, se establece como funciones objetivo minimizar los Factores de Inflación de la Varianza (VIF) y el Cuadrado Medio del Error (MSE), además de maximizar el coeficiente de determinación general del modelo (R^2).

Se realiza una comparación entre los modelos de regresión, generados a partir de las distintas técnicas para calcular k, después el comparativo se plantea mediante un análisis de la varianza, se obtienen los estadísticos VIF, MSE, R^2 con sus respectivas pruebas de hipótesis, en base a los resultados de las métricas, se determina el modelo más adecuado y representativo del proceso de maquinado.

5.4 Analizar el proceso de manufactura

Mediante el modelo de Regresión Ridge se podrá inferir cuales variables del proceso son más significativas y así obtener las especificaciones deseadas para la pieza de maquinado, generando un proceso controlado y con menos variabilidad.

Capítulo 6

Desarrollo experimental y resultados

En este capítulo se muestran los resultados obtenidos a partir del desarrollo experimenta llevado a cabo en dos procesos de manufactura, utilizando los contenidos planteados en el Marco Teórico.

6.1 Proceso de maquinado

El primer caso de estudio en analizar es un proceso de maquinado, para eliminar la multicolinealidad se obtienen los valores de k mediante la Traza de Ridge (ec.2.9), el Método Iterativo (ec.2.10 y 2.11), Regresión Ridge Generalizada (ec.2.16 y 2.17), además de los algoritmos inspirados en la naturaleza PSO, ACO y ABC.

Tabla 6.1. Valores k de sesgo del proceso de maquinado.

Traza Ridge	Método Iterativo	Ridge Generalizada	PSO	ACO	ABC
$k = 0.5$	$k = 0.899$	$k = \begin{bmatrix} 0.0433 \\ 0.4629 \\ 0.1391 \\ 0.1219 \\ 0.8908 \end{bmatrix}$	$k = \begin{bmatrix} 0.4324 \\ 0.4427 \\ 0.04 \\ 0.01 \\ 0.5022 \end{bmatrix}$	$k = \begin{bmatrix} 0.0870 \\ 0.8546 \\ 0.3855 \\ 0.0080 \\ 0.0274 \end{bmatrix}$	$k = \begin{bmatrix} 0.2650 \\ 0.5510 \\ 0.01 \\ 0.01 \\ 0.1651 \end{bmatrix}$

La Tabla 6.1 no muestra una columna de sesgo del método de Mínimos Cuadrado, pero cabe mencionar que cuando $k = 0$ el estimador Ridge es igual al estimador por Mínimos Cuadrados. Además todos los valores de sesgo en la Tabla 6.1 son entre $0 \leq k \leq 1$ dado que el interés es generar estimadores más estables, pero no agregar demasiado sesgo (Marquardt & Snee, 1975).

Después se agrega el valor k a la matriz de diseño para obtener los estimadores de regresión para cada método, posteriormente se calculan los Factores de Inflación de la Varianza ($VIF's$) donde se verifica la eliminación de la multicolinealidad.

Tabla 6.2. Estimadores de regresión del proceso de maquinado.

$\hat{\beta}_{MC}$	$\hat{\beta}_{R_{traza}}$	$\hat{\beta}_{R_{iter}}$	$\hat{\beta}_{RG}$	$\hat{\beta}_{PSO}$	$\hat{\beta}_{ACO}$	$\hat{\beta}_{ABC}$
30.1796	30.1708	30.1737	30.1619	30.1741	30.1728	30.1738
-6.67E-05	7.21E-05	6.25E-05	1.31E-04	4.04E-06	1.54E-05	6.38E-06
7.22E-05	-9.54E-06	-6.87E-06	-3.27E-06	-2.65E-06	1.58E-06	-3.19E-06
8.33E-07	2.13E-07	1.89E-07	1.22E-07	1.25E-07	3.87E-08	1.93E-07
5.56E-07	1.81E-07	1.57E-07	1.19E-07	4.45E-07	4.27E-07	4.33E-07
-2.78E-06	-1.46E-07	-1.01E-07	-3.71E-08	-5.71E-08	7.48E-08	-2.13E-07

Tabla 6.3. $VIF's$ del proceso de maquinado.

VIF_{MC}	$VIF_{R\ traz}$	$VIF_{R\ iter}$	VIF_{RG}	VIF_{PSO}	VIF_{ACO}	VIF_{ABC}
1225	0.1503	0.1120	0.5590	0.0047	0.0971	0.0122
343	0.0974	0.0747	0.0572	0.0304	0.0066	0.0174
175	0.0936	0.0695	0.4344	0.8489	0.0599	1.431
1201	0.1505	0.1120	0.0972	1.0403	0.9229	1.0856
193	0.1031	0.0764	0.0202	0.0398	0.8705	0.3244

En la Tabla 6.3 la primera columna muestra los $VIF's$ mediante Mínimos Cuadrados donde se visualiza valores muy por encima de 10, lo cual indica la multicolinealidad entre las variables de entrada. Para las demás columnas el caso es totalmente opuesto, porque se observan valores muy por debajo de 10, esto indica que en todos los métodos al agregar cierta cantidad de sesgo se elimina la multicolinealidad. Aunado a lo anterior, la Tabla 6.4 proporciona los resultados del Número de Condición (otra métrica para detectar la multicolinealidad) donde se reafirma la eliminación de la dependencia lineal entre las variables de entrada, ya que si el valor obtenido es mayor a 100 es indicio de

multicolinealidad y el único método que muestra esa condición es el de Mínimos Cuadrados, él cual no agrega sesgo a sus datos.

Tabla 6.4. Número de Condición proceso de maquinado.

η_{MC}	$\eta_{R\ traz}$	$\eta_{R\ iter}$	η_{RG}	η_{PSO}	η_{ACO}	η_{ABC}
7230	6.98	4.33	42.61	24.80	73.09	46.59

También se realiza un Análisis de la Varianza (ANOVA por sus siglas en inglés) para cada método de regresión de donde se obtienen las métricas estadísticas R^2 (ajuste del modelo) y MSE (Cuadrado Medio del Error), los cuales se utilizarán como punto de comparación entre los métodos de regresión propuestos.

Tabla 6.5. ANOVA de $\hat{\beta}_{MC}$ con $k = 0$. Proceso de maquinado

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	5	0.48539	0.09707	3.9616
Error	21	0.51461	0.02450	
Total	26	1		

Tabla 6.6. ANOVA de $\hat{\beta}_{R\ traz}$ con $k = 0.5$. Proceso de maquinado

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	5	0.39197	0.07839	2.7076
Error	21	0.60803	0.02895	
Total	26	1		

Tabla 6.7. ANOVA de $\hat{\beta}_{R\ iter}$ con $k = 0.899$. Proceso de maquinado

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	5	0.34048	0.068097	2.1683
Error	21	0.65952	0.031405	
Total	26	1		

Tabla 6.8. ANOVA de $\hat{\beta}_{RG}$. Proceso de maquinado

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	5	0.46919	0.093838	3.7125
Error	21	0.53081	0.025276	
Total	26	1		

Tabla 6.9. ANOVA de $\hat{\beta}_{PSO}$. Proceso de maquinado

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	5	0.47951	0.095901	3.8693
Error	21	0.52049	0.024785	
Total	26	1		

Tabla 6.10. ANOVA de $\hat{\beta}_{ACO}$. Proceso de maquinado

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	5	0.48028	0.096056	3.8813
Error	21	0.51972	0.024748	
Total	26	1		

Tabla 6.11. ANOVA de $\hat{\beta}_{ABC}$. Proceso de maquinado

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	F_0
Modelo	5	0.47977	0.095953	3.8733
Error	21	0.52023	0.024773	
Total	26	1		

Se obtiene de los Análisis de la Varianza de la Tabla 6.5 a la Tabla 6.11 las métricas estadísticas que ayudaran como medida comparativa entre los métodos aplicados, para poder determinar y elegir el método de regresión que cumpla con el objetivo, eliminar la multicolinealidad sin afectar demasiado el ajuste del modelo, en consecuencia los estimadores del modelo serán más estables para la obtención de nuevas predicciones.

Tabla 6.12. Comparación de resultados métricas estadísticas. Proceso de maquinado

Estadístico	$\widehat{\boldsymbol{\beta}}_{MC}$	$\widehat{\boldsymbol{\beta}}_{R_{traza}}$	$\widehat{\boldsymbol{\beta}}_{R_{iterativo}}$	$\widehat{\boldsymbol{\beta}}_{RG}$	$\widehat{\boldsymbol{\beta}}_{PSO}$	$\widehat{\boldsymbol{\beta}}_{ACO}$	$\widehat{\boldsymbol{\beta}}_{ABC}$
MSE	0.02450	0.02895	0.03140	0.02527	0.02478	0.02474	0.02477
F_0	3.9616	2.7076	2.1683	3.7125	3.8693	3.8813	3.8733
R^2	48.53	39.19	34.04	46.91	47.95	48.02	47.97

La Figura 6.1 muestra el grafico de las predicciones generadas mediante los métodos Ridge Generalizado y ACO, se consideró graficar estos dos últimos debido a su mejor comportamiento en base a las métricas estadísticas con las cuales se compararon. Cabe mencionar que en estos dos métodos ya está eliminada la multicolinealidad, por lo cual sus estimadores son más estables en comparación con mínimos cuadrados (donde no se ha tratado la multicolinealidad). Se grafica además los datos deseados del diámetro del maquinado.

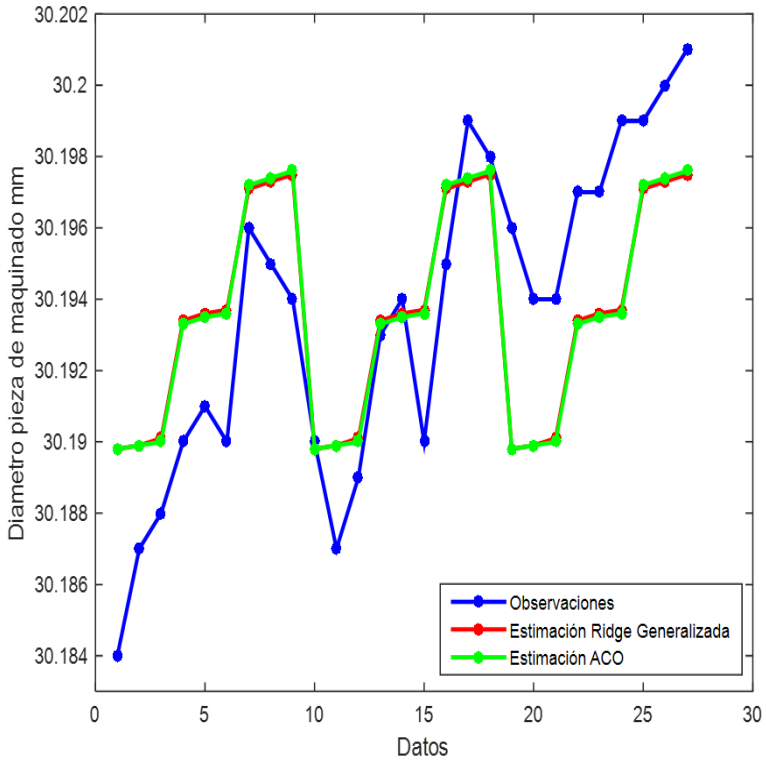

Figura 6.1. Gráfica observaciones y estimaciones de RG y ACO. Proceso Maquinado

6.2 Proceso de Soldadura PTA

Se presenta además un segundo caso de estudio de un proceso de soldadura PTA, donde también los datos presentan problemas de multicolinealidad entre las variables de entrada y se corroboran resultados similares a favor de los algoritmos inspirados en la naturaleza ACO, ABC y PSO en comparación con las otras técnicas que también calculan y agregar sesgo en la matriz de datos para eliminar la dependencia lineal y generar estimadores de regresión más estables.

Tabla 6.13. Valores k de sesgo del proceso soldadura PTA.

Traza Ridge	Método Iterativo	Ridge Generalizada	PSO	ACO	ABC
$k = 0.3$	$k = 0.998$	$k = \begin{bmatrix} 0.3328 \\ 0.1553 \\ 0.1992 \\ 0.2713 \end{bmatrix}$	$k = \begin{bmatrix} 0.5459 \\ 0.0708 \\ 0.01 \\ 0.1522 \end{bmatrix}$	$k = \begin{bmatrix} 0.3763 \\ 0.2040 \\ 0.0083 \\ 0.9296 \end{bmatrix}$	$k = \begin{bmatrix} 0.9048 \\ 0.3426 \\ 0.01 \\ 0.0772 \end{bmatrix}$

Tabla 6.14. Estimadores de regresión del proceso soldadura PTA.

$\hat{\beta}_{MC}$	$\hat{\beta}_{R_{traza}}$	$\hat{\beta}_{R_{iterativo}}$	$\hat{\beta}_{RG}$	$\hat{\beta}_{PSO}$	$\hat{\beta}_{ACO}$	$\hat{\beta}_{ABC}$
6.6899	5.6577	5.0234	5.8966	6.5135	6.6698	6.5134
-1.94E-02	-8.23E-04	-3.57E-03	-6.88E-05	-3.23E-04	4.66E-04	6.00E-04
5.97E-02	-5.49E-03	-1.23E-02	-2.69E-03	1.19E-02	8.84E-03	7.65E-03
-3.23E-02	-2.31E-02	-1.46E-02	-2.55E-02	-3.17E-02	-3.21E-02	-3.15E-02
3.59E-03	8.68E-03	9.38E-03	7.27E-03	3.42E-03	1.84E-03	3.71E-03

Los Factores de Inflación de la Varianza ($VIF's$) y el número de condición (η), son las métricas estadísticas más utilizadas para determinar la multicolinealidad entre las variables de entrada. Como se observa en las Tablas 6.15 y 6.16 mediante los métodos propuestos en esta investigación, se elimina la multicolinealidad agregando los valores de sesgo obtenidos en la Tabla 6.13 para el proceso de soldadura PTA.

Tabla 6.15. $VIF's$ del proceso de soldadura PTA.

VIF_{MC}	$VIF_{R\ traz}$	$VIF_{R\ iter}$	VIF_{RG}	VIF_{PSO}	VIF_{ACO}	VIF_{ABC}
33.77	0.2804	0.1197	0.2029	0.0841	0.1874	0.0712
34.86	0.2695	0.1157	0.5158	0.8569	0.4418	0.3827
1.48	0.6379	0.2286	0.7878	1.2786	1.1397	1.2707
1.27	0.6222	0.2354	0.6659	0.8886	0.2664	1.0491

Tabla 6.16. Número de Condición proceso soldadura PTA.

η_{MC}	$\eta_{R\ traz}$	$\eta_{R\ iter}$	η_{RG}	η_{PSO}	η_{ACO}	η_{ABC}
159.09	8.37	3.29	10.20	9.19	9.03	5.40

Se realiza el Análisis de la Varianza a partir de los estimadores de la Tabla 6.14 para determinar la afectación de la regresión debido al sesgo agregado, el cual se utiliza para eliminar la multicolinealidad entre las variables de entrada del proceso de soldadura PTA.

Tabla 6.17. ANOVA de $\hat{\beta}_{MC}$ con $k = 0$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	4	0.74513	0.18628	20.465
Error	28	0.25487	0.00910	
Total	32	1		

Tabla 6.18. ANOVA de $\hat{\beta}_{R\ traz}$ con $k = 0.3$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	4	0.58502	0.14625	9.8682
Error	28	0.41498	0.01482	
Total	32	1		

Tabla 6.19. ANOVA de $\hat{\beta}_{R\ iter}$ con $k = 0.9981$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	4	0.40706	0.10176	4.8056
Error	28	0.59294	0.02117	
Total	32	1		

Tabla 6.20. ANOVA de $\hat{\beta}_{RG}$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	4	0.62751	0.15688	11.792
Error	28	0.37249	0.01330	
Total	32	1		

Tabla 6.21. ANOVA de $\hat{\beta}_{PSO}$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	4	0.73672	0.18418	19.587
Error	28	0.26328	0.00940	
Total	32	1		

Tabla 6.22. ANOVA de $\hat{\beta}_{ACO}$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$\boldsymbol{F_0}$
Modelo	4	0.73699	0.18425	19.615
Error	28	0.26301	0.00939	
Total	32	1		

Tabla 6.23. ANOVA de $\hat{\beta}_{ABC}$. Proceso de soldadura PTA

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	F_0
Modelo	4	0.73659	0.18415	19.575
Error	28	0.26341	0.00940	
Total	32	1		

En la Tabla 6.24 se observa el comparativo de las métricas estadísticas donde se destaca que el ajuste de modelo R^2 para el método Mínimos Cuadrados es de 74.51, lo cual sería algo aceptable, sin embargo las variables de entrada presentan multicolinealidad, es por esto que los resultados obtenidos por los métodos inspirados en la naturaleza PSO, ACO y ABC, son más favorables para el proceso de soldadura PTA.

Tabla 6.24. Comparación de resultados métricas estadísticas. Soldadura PTA

Estadístico	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MC}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R_{traza}}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R_{iterativo}}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{RG}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{PSO}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{ACO}}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{ABC}}$
MSE	0.00910	0.01482	0.02117	0.01330	0.00940	0.00939	0.00940
F_0	20.465	9.8682	4.8056	11.792	19.587	19.615	19.575
R^2	74.51	58.50	40.70	62.75	73.67	73.69	73.65

En la Figura 6.2 se graficó las observaciones del área afectada por el calor del proceso de soldadura PTA, además de la estimación por media de la regresión Ridge Generalizada la cual obtuvo mejores resultados entre los métodos exactos, también se graficó las estimaciones obtenidas mediante ACO la cual obtuvo el mejor resultado entre las metaheuristicas, tomando como métrica estadística el ajuste del modelo R^2.

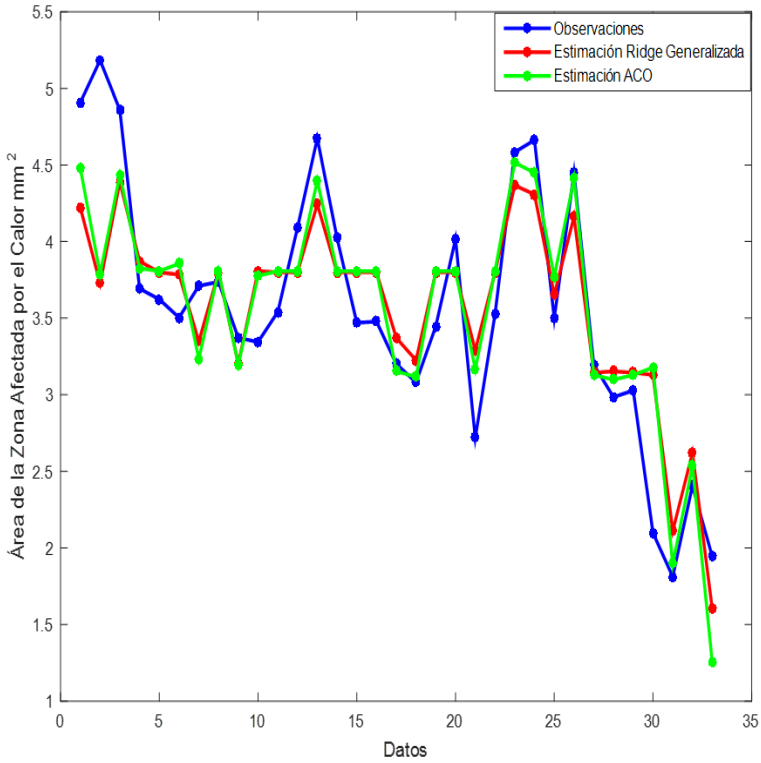

Figura 6.2. Grafica observaciones y estimaciones de RG y ACO. Proceso Soldadura PTA

En términos generales, los resultados de la Tabla 6.12 del proceso de maquinado y la Tabla 6.24 del proceso de soldadura PTA, muestran que introducir sesgo en la matriz de las variables de entrada es una buen método para eliminar la multicolinealidad, además el comparar entre diversos métodos para obtener los valores *k* de sesgo, amplia la posibilidad de obtener mejores resultados. En cuanto a los métodos exactos: traza de Ridge, el método iterativo y Ridge Generalizado, este último fue quien se comportó de mejor manera ante el objetivo de eliminar la multicolinealidad sin afectar en gran medida el ajuste del modelo. Por otro lado los algoritmos inspirados en la naturaleza PSO, ACO y ABC se comportaron mejor que los métodos exactos y también cumplieron con el objetivo antes mencionado.

Capítulo 7

Conclusiones

En esta tesis se presenta el trabajo de investigación sobre el análisis de la multicolinealidad en modelos de regresión para dos procesos de manufactura. Los resultados obtenidos permitieron llegar a las siguientes conclusiones:

Se realizó un análisis a la matriz de diseño para determinar el grado de multicolinealidad entre las variables de entrada del proceso de manufactura, basados en la literatura (Montgomery, Peck, & Vining, 2006) se estableció que los Factores de Inflación de Varianza ($VIF's$) y el número de condición (η), son mejores indicadores para determinar la dependencia lineal entre los factores, en comparación con la matriz de varianzas y covarianzas.

Al realizar una regresión lineal entre las variables de entrada y salida, se obtuvieron estimadores mediante mínimos cuadrados, se hizo notar que dichos estimadores del modelo presentan problemas de multicolinealidad y no son adecuados debido a que presentan un grado fuerte de multicolinealidad entre las variables de entrada, generando intervalos de confianza grandes para los estimadores de $\hat{\beta}_{MC}$, esto causa problemas al momento de realizar el análisis de la varianza, porque al momento de determinar mediante las pruebas de hipótesis cuales son las variables significativas para el proceso se puede cometiendo el error tipo 1 o tipo 2, es decir, aceptar o rechazar la importancia de una variable de entrada del proceso.

Esto motivo la búsqueda de otras alternativas para calcular los estimadores de regresión, se aplicó regresión Ridge teniendo como resultado la eliminación de la multicolinealidad, pero al ser la dependencia lineal muy significativa, el modelo de regresión pierde ajuste, lo

cual se evidenció mediante el estadístico R^2, estableciendo que la alternativa de regresión Ridge elimina la multicolinealidad generando estimadores $\hat{\beta}_R$ más estables, por lo tanto, ya no se cometerá error tipo 1 o error tipo 2 al momento de realizar el análisis de la varianza, pero el ajuste del modelo se ve muy afectado.

Lo anterior motivo la investigación hacia la búsqueda de métodos capaces de eliminar la multicolinealidad sin afectar el ajuste del modelo de regresión. Las alternativas a desarrollar fueron la Regresión Ridge Generalizada y algunos algoritmos inspirados en la naturaleza PSO, ACO y ABC; ambas vertientes con la idea de añadir más de un valor distinto k de sesgo en la matriz de diseño, que permita eliminar la multicolinealidad, maximizando el ajuste del modelo, medido mediante el estadístico R^2, esto impacta de manera positiva y beneficia para explicar adecuadamente las variables de los procesos de manufactura. Se observó que los cuatro métodos eliminaron la multicolinealidad, resaltando que los algoritmos PSO, ACO y ABC obtuvieron mejores resultados con respecto al ajuste del modelo en ambos casos de estudio: proceso de maquinado ($R^2_{PSO} = 47.95, R^2_{ACO} = 48.02\ y\ R^2_{ABC} = 47.97$) y soldadura PTA ($R^2_{PSO} = 73.67, R^2_{ACO} = 73.69\ y\ R^2_{ABC} = 73.65$) manteniendo valores muy similar al de Mínimos Cuadrados ($R^2_{MC} = 48.53$) y ($R^2_{MC} = 74.51$) respectivamente, con la ventaja que los estimadores obtenidos mediante $\hat{\beta}_{PSO}, \hat{\beta}_{ACO}\ y\ \hat{\beta}_{ABC}$ son más estables. Lo anterior contesta las preguntas de investigación 1. ¿Qué impacto tendrá agregar más de un valor k distinto de sesgo en la matriz de diseño? y 2. El agregar más de un valor de sesgo k ¿Ayudara a generar estimadores más estables?

Utilizando los algoritmos inspirados en la naturaleza PSO, ACO y ABC, se obtuvieron distintos valores k de sesgo, los resultados se aplicaron en la matriz de sesgo; además se consideró como función objetivo el maximizar el estadístico R^2, también verificando la

eliminación de la multicolinealidad entre las variables de entrada, se consideró como medidas estadísticas los Factores de Inflación de la Varianza, donde los resultados fueron menores a diez en ambos casos de estudio (ver tablas 6.3 y 6.15) y el número de condición, donde los resultados fueron menores a cien (ver tablas 6.4 y 6.16) lo cual, es lo que determina eliminar la multicolinealidad. Esto responde la pregunta de investigación 3. ¿Qué método de optimización se debe utilizar para encontrar los valores de k que eliminen la multicolinealidad?

Se planteó una optimización global debido que la función objetivo a considerada fue maximizar el estadístico R^2, asimismo se contempló como métodos para calcular el sesgo k que elimine la multicolinealidad entre las variables de entrada, a los algoritmos inspirados en la naturaleza, debido al gran auge que tienen hoy en día, además de presentar buenos resultados en otras investigaciones (Farmani, Jaamialahmadi, & Babaie, 2011), (Xu & Yang, 2015), (Zhang, y otros, 2016). Los algoritmos implementados fueron Particles Swarm Optimization (PSO), Ant Colony Optimization (ACO) y Artificial Bee Colony (ABC). Esto contesta la pregunta de investigación 4. ¿Se deberá plantear una optimización multiobjetivo o global para obtener los valores k que produzcan menos sesgo a los estimadores?

Además, la búsqueda optima del parámetro de sesgo k mediante los algoritmos ACO, ABC y PSO ayudarán a mencionar de forma acertada, cuales variables son más significativas para los procesos de manufactura que se consideraron en este trabajo, el efecto es positivo debido a que no se cometerá error tipo 1 o error tipo 2 al momento de realizar el análisis de la varianza. Con esto se responde la pregunta de investigación 5. ¿Cuál es el efecto de k sobre las variables del proceso de manufactura?

La idea básica de la regresión Ridge Generalizada es obtener una matriz de sesgo con valores k distintos para tratar la multicolinealidad entre las variables de entrada, entonces a partir de este hecho podemos mencionar como afirmativa la **Hipótesis 1** de esta investigación, además y en base a esta idea de la regresión Ridge Generalizada, al aplicar los algoritmos ACO, ABC y PSO para la búsqueda de valores k de sesgo donde la función objetivo fue maximizar el ajuste del modelo eliminando la multicolinealidad, se obtuvieron resultados positivos concluyendo también como afirmativa la **Hipótesis 2**.

En base a todo lo anterior fue posible determinar el grao de multicolinealidad entre las variables de entrada, eliminar la dependencia lineal utilizando los métodos de regresión Ridge Generalizada y los algoritmos ACO, ABC y PSO obteniendo valores de sesgo k que permitieron obtener estimadores de regresión más estables permitiendo con esto poder mencionar cuales son las variables más significativas en cada proceso de manufactura y así cumplir con **el objetivo general y el objetivo específico uno**, además para obtener los resultados del estadístico R^2, se planteó como idea, mejorar la búsqueda del parámetro k de sesgo en base a una optimización, cumpliendo así **los objetivos específicos dos y tres.**

Para cumplir **el objetivo específico cuatro**, se compararon los resultados de la optimización mediante los algoritmos antes mencionados y se concluyó que fueron mejores los algoritmos PSO, ACO y ABC con respecto a los resultados que se obtuvieron mediante Regresión Ridge Generalizado, estableciendo como parámetro comparativo el ajuste del modelo R^2 y los $VIF´s$ para la eliminación de la multicolinealidad.

Los resultados que se obtuvieron en esta investigación nos dejan una clara idea de lo importante que es realizar un análisis estadístico correcto, detectar y eliminar la multicolinealidad entre las variables de entrada, para entender y predecir la variable de salida,

debido a que todo el análisis se realizó considerando solo una respuesta (variable de salida) en cada proceso, se deja una línea de investigación abierta para trabajo futuro, el estudio de procesos de manufactura que presenten multicolinealidad y en los cuales se consideren controlar más de una variable de salida.

Por último, en este trabajo de investigación se realizó un análisis de la multicolinealidad utilizando la regresión Ridge Generalizada como pieza importante para dar solución a la problemática de la dependencia lineal, donde se obtuvieron buenos resultados, pero sobre todo incorporar los algoritmos inspirados en la naturaleza ACO, ABC y PSO fue de gran ayuda para obtener los resultados deseados, eliminar la multicolinealidad para obtener estimadores estables sin afectar el ajuste del modelo.

Bibliografía

Alkhamisi, M., & Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. *Communication in Statistic. Simulation and Computation*, 535-547.

Demirhan, H. (2014). The problem of multicollinearity in horizontal solar radiation estimation models and a new model for Turkey. *Energy Conversion and Management*, 334-345.

Dorugade, A., & D.N., K. (2010). Alternative method for chossing ridge parameter for regression. *Applied Mathematical Sciences*, 447-456.

Du, K.-L., & Swamy, M. (2016). *SEARCH AND OPTIMIZATION BY METAHEURISTICS.* Suiza: Birkhäuser.

El-Dereny, M., & Rashwan, N. (2011). Solving multicollinearity problem using ridge regression models. *Int. J. Contemp. Math. Sciences*, 585-600.

Elham, A., & van Tooren, M. J. (2017). Multi-fidelity wing aerostructural optimization using a trust region filter-SQP algorithm. *Struct Multidisc Optim*, 1773-1786.

Farmani, M., Jaamialahmadi, A., & Babaie, M. (2011). Multiobjective optimization for force and moment balance of a four-bar linkage using evolutionary algorithms. *Journal of Mechanical Science and Technology*, 2971-2977.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 215-223.

Guanghui, L., & Xiaohong, J. (2017). Synthesis and validation of finite time servo control with PSO identification for automotive electronic throttle. *Nonlinear Dynamics*, 1165-1177.

Hoerl, A., & Kennard, R. (1970a). Ridge Regression: Biades Estimation for Nonorthogonal Problems. *Technometrics*, 55-67.

Hoerl, A., & Kennard, R. (1970b). Ridge regression:Applications to nonorthogonal problems. *Technometrics*, 69-82.

Hoerl, A., & Kennard, R. (1976). Ridge regression: Iterative estimation of the biasing parameter. *Commun Stat*, 77-88.

Hoerl, A., Kennard, R., & Baldwin, K. (1975). Ridge Regression: some simulations. *Commun State*, 105-123.

Jun, L., Xiaoyong, Y., Chengzu, R., Guang, C., & Yan, W. (2015). Multiobjective optimization of cutting parameters in Ti-6Al-4V milling process using nondominated sorting genetic algorithm-II. *International Journal Advanced Manufacturing Technology*, 941-953.

Kibria, G. B. (2003). Performance of some new ridge regression estimators. *Communication in Statistics. Simulation and Computation*, 419-435.

Li, Q., & Shao, C. (2008). Estimation of Distillation Compositions Using Sensitivity Matrix Analysis and Kernel Ridge Regression. *IFAC Proceedings Volumes*, 11943-11948.

Lipovetsky, S., & Conklin, M. (2000). Multiobjective regression modifications for collinearity. *Computers & Operations Research*, 1333-1345.

Liu, H., Miao, E. M., Yuan, W. X., & Dong, Z. X. (2017). Robust modeling method for thermal error of CNC machine tools based on ridge regression algorithm. *International Journal of Machine Tools & Manufacture*, 35-48.

Marquardt, D., & Snee, R. (1975). Ridge regression in practice. *The american statistician*, 3-20.

Montgomery, D. C. (2002). *Diseño y Analisis de Experimentos.* México D.F.: Limusa Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introducción al análisis de Regresión Lineal.* México: Continental.

Munner, K. M., Abdulrahman, A.-A., & Usama, U. (2015). Multiobjective optimization of Nd:YAG direct laser writing of microchannels for microfluidic applications. *International Journal Advanced Manufacture Technology*, 1363-1377.

Piña, M. R., & Diaz, J. (2005). Superioridad de la Regresión General Ridge sobre Mínimos Cuadrados. *CULCYT*, 50-62.

Piña, M., Manuel, R., & Aguirre, J. (2007). Regresión Ridge y la distribución central t. *Ciencia Ergo Sum*, 191-196.

Serkan, D., & Rasit, K. (2019). Simulation based calculation of the inverse kinematics solution of 7-DOF robot manipulator using artificial bee colony algorithm. *SN Applied Sciences*, 27.

Shakya, A., Mishra, M., Maity, D., & Santarsiero, G. (2019). Structural health monitoring based on the hybrid ant colony algorithm by using Hooke–Jeeves pattern search. *SN Applied Sciences*, 799.

Shengzheng, W., Baoxian, J., Jiansen, Z., Wei, L., & Tie, X. (2017). Predicting ship fuel consumption based on LASSO regression. *Transportation Research Part D*, 1-8.

Siarry, P. (2016). *Metaheuristics.* Suiza: Springer.

Stanley, G. (1996). *Algebra Lineal.* Edo. de México: Mc Graw Hill.

Wang, X., Liang, Y., Wang, Q., & Zhang, Z. (2017). Empirical models for tool forces prediction of drag-typed picks based on principal component regression and ridge regression methods. *Tunnelling and Underground Space Technology*, 75-95.

Wong, K. Y., & Chiu, S. N. (2015). An iterative approach to minimize the mean squared error in ridge regression. *Computational Statistics*, 625-639.

Xu, G., & Yang, Z. (2015). Multiobjective optimization of process parameters for plastic injection molding via soft computing and grey correlation analysis. *International Journal of Advanced Manufacturing Technology*, 525-536.

Yan-Fu, L., Min, X., & Thong-Ngee, G. (2010). Adaptive ridge regression system for software cost estimating on multi-collinear datasets. *The Journal of Systems and Software*, 2332-2343.

Ying-Ze, T., Gui-Rong, L., Cai-Yan, Z., Jian-Yu, W., Fang, Z., Guo-Liang, S., & Yin-Chang, F. (2013). Effects of collinearity, unknown source and removed factors on the NCPCRCMB receptor model solution. *Atmospheric Environment*, 76-83.

Zhang, J., Wang, J., Lin, J., Guo, Q., Chen, K., & Ma, L. (2016). Multiobjective optimization of injection molding process parameters based on Opt LHD, EBFNN and MOPSO. *Inernational Journal of Advanced Manufacturing Technology*, 2857-2872.

Índice de Tablas

Índice de Figuras