TECNOLÓGICO NACIONAL DE MÉXICO
en Celaya

TECNOLÓGICO NACIONAL DE MÉXICO EN CELAYA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

"ANÁLISIS DE VALIDACIÓN CRUZADA BAJO DIFERENTES CONDICIONES DE RUIDO"

TESIS PROFESIONAL
PARA OBTENER EL GRADO DE:

MAESTRA EN INGENIERÍA INDUSTRIAL

PRESENTA:

ING. NATALIA ARCEDALIA RODRÍGUEZ MURILLO

DIRECTOR DE TESIS:

DR. ARMANDO JAVIER RÍOS LIRA

CO-DIRECTOR DE TESIS:

M.C. DARÍO HERNÁNDEZ RIPALDA

CELAYA, GTO., MÉXICO, FEBRERO, 2019

SEP
SECRETARIA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MEXICO
en Celaya

"2019 [illegible] del Sur, Emiliano Zapata"

Asunto: Autorización de impresión de trabajo profesional.

Celaya Gto., **21 de FEBRERO 2019**

M.C. MOISES TAPIA ESQUIVIAS
JEFE DEL DEPARTAMENTO DE INGENIERIA INDUSTRIAL.
P r e s e n t e.

De acuerdo a la convocatoria hecha por esta jefatura a fin de aprobar o no la impresión del trabajo profesional titulado:

"Análisis de validación cruzada bajo diferentes condiciones de ruido"

Presentado por el (a) pasante **C. ING.NATALIA ARCEDALIA RODRIGUEZ MURILLO (M1703012)**
alumno (a) del programa de Maestría en Ingeniería Industrial que ofrece nuestro Instituto.
Hacemos de su conocimiento que éste jurado ha tenido a bien aprobar la impresión de dicho trabajo para los efectos consiguientes.

A T E N T A M E N T E

DR. ARMANDO JAVIER RIOS LIRA
Presidente

DR. SALVADOR HERNANDEZ GONZALEZ
Secretario

DR. JOSE ALFREDO JIMENEZ GARCIA
Vocal

M.C. VICENTE FIGUEROA FERNANDEZ
Vocal Suplente

SECRETARIA DE EDUCACION PUBLICA
TECNOLOGICO NACIONAL DE MEXICO
INSTITUTO TECNOLOGICO DE CELAYA
COORDINACION DE MAESTRIA DE INGENIERIA INDUSTRIAL

Ccp.Escolares
Archivo.
*VFF*MTE*DMVP*

Antonio García [illegible] No. 630 esq. Av. Tecnológico, Col. Alfredo V. Bonfil,
C.P. 38010
Celaya, Gto. Mex. Conmutador (461) 6117575 e-mail: lince@itcelaya.edu.mx
www.itcelaya.edu.mx

DEDICATORIAS

El presente trabajo es dedicado:

A Dios, quien con su amor y cuidado a lo largo de los años con su guía estuvo presente en el caminar de mi vida, bendiciéndome y dándome fuerzas para continuar con mis metas trazadas sin desfallecer. Por la maravillosa familia que me apoya día a día.

A mi familia, porque me han enseñado a no darme por vencida, a superar todas las dificultades que se presentan en la vida. Por siempre apoyarme a cumplir mis sueños. Por amarme incondicionalmente. Por permitirme conocer más allá de este país y querer aspirar a mejores cosas siempre.

A mi madre, quien es el pilar de mi vida, el motor principal para lograr mis metas. Gracias por tanto amor, por tanto cariño, por tanta paciencia. Te amo, eres el mejor ejemplo de vida, una gran guerrera.

A mi hermano Tadeo, contigo he aprendido que no hay caída que nos detenga trazar un camino extraordinario en nuestras vidas. Que el amor es lo más invencible en el mundo y que juntos somos lograremos todo lo que nos propongamos.

A mi novio Luis, tantas veces que he querido desistir has sido la clave de muchos de mis éxitos, tú apoyo incondicional y gran amor me motiva a seguir, a no desistir. Gracias por todo y por lo que falta. Te amo.

AGRADECIMIENTOS

Debo agradecer de manera especial y sincera al Dr. Armando Javier Lira Ríos por aceptarme para realizar esta tesis de maestría bajo su dirección. Su apoyo y confianza en mi trabajo y su capacidad para guiarme, no solamente en el desarrollo de esta tesis, sino también en mi formación como investigador. Le agradezco también el haberme facilitado siempre los medios suficientes para llevar a cabo todas las actividades propuestas durante el desarrollo de esta tesis. Muchas gracias Doctor.

Al departamento de Posgrado de Ingeniería Industrial del Tecnológico Nacional de México en Celaya, a mis profesores durante la maestría quienes con su apoyo y experiencia propiciaron ésta maravillosa experiencia de vida y formación profesional, en quienes sé que puedo contar para seguir creciendo en el ámbito laboral.

A mis compañeros de generación, personas únicas con quien crecí personalmente al conocer sus historias de vida y profesionales, ahora como fruto tenemos una amistad sincera y fructífera.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo brindado en los programas de becas económicos para concluir un posgrado.

ABSTRACT

When submitting information to analyze using a linear regression model, there is doubt about the predictive capacity of the model when predicting new information. For this reason it is necessary to validate the predictive capacity of the model by means of some suitable method. The Cross Validation technique is the most used method, since it divides the data into two parts: a set of training data and a set of test data. The first set is used to estimate the linear regression coefficients, while the second is used to measure the predictive capacity of the model. The existing literature proposes to save 10 percent of the data, but it may not be appropriate in the different situations presented within a research, therefore the optimal number of prediction data to keep is uncertain when performing the cross-validation technique under a noise level and a set information size. The presented research raises the different scenarios to identify the optimal percentage to be saved when using the aforementioned technique.

RESUMEN

Al someter información a analizar mediante un modelo de regresión lineal, existe la duda sobre la capacidad predictiva del modelo al predecir información nueva. Por ello es necesario validar la capacidad predictiva del modelo mediante algún método adecuado. La técnica de Validación cruzada es el método más usado, ya que divide los datos en dos partes: un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El primer conjunto es utilizado para estimar los coeficientes de regresión lineal, mientras que el segundo es utilizado para medir la capacidad predictiva del modelo. La literatura existente propone guardar el 10 por ciento de los datos, pero podría no ser el apropiado en las diferentes situaciones presentadas dentro de una investigación, por lo tanto el número óptimo de datos de predicción a guardar es incierto al realizar la técnica de validación cruzada bajo un nivel de ruido y un tamaño de información del conjunto. La investigación presentada plantea los diferentes escenarios para identificar el porcentaje óptimo a ser guardado al utilizar la técnica anteriormente mencionada.

ÍNDICE

INDÍCE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

INTRODUCCIÓN

La estadística propone encontrar patrones en un río de información confusa. Para ello, surgen diversas cuestiones que deben ser analizadas: ¿Cómo y qué información o datos elegir? ¿Cómo analizar y resumir el análisis efectuado? (Kerner, 1015).

Durante la etapa de construcción de modelos hay factores que pueden afectar significativamente las nuevas observaciones, lo que hace las predicciones menos exactas (García & Lara, 1998). Además, la estructura correlativa entre los regresores puede diferir el rendimiento para el modelo; esto puede resultar en un rendimiento predictivo pobre para el modelo. La validación adecuada de un modelo desarrollado para predecir nuevas observaciones debe involucrar y probar el modelo en el medio a desarrollarse antes de que sea entregado al usuario (Montgomery & Runger, 2014).

Con la finalidad de poder analizar algunos métodos, se realiza una validación del modelo obtenido. Para ello, existen diversas técnicas, entre las cuales destacamos los métodos de validación cruzada. La validación cruzada hace uso de distintos subconjuntos de los datos disponibles para realizar el entrenamiento del modelo y su posterior validación (Pérez-Planells L. , Delegido, Rivera-Caicedo, & Verrelst, 2015).

Muchos casos los nuevos modelos son representados sin una adecuada validación, por ello la validación propia de un modelo de regresión incluye un estudio de coeficientes para determinar si sus signos y magnitudes son razonables, así como el análisis en la estabilidad de los coeficientes de regresión (Montgomery, Peck, & Vining, 2011).

El Análisis de Regresión Lineal Múltiple permite establecer la relación que se produce entre una variable dependiente Y y un conjunto de variables independientes (X_1, X_2, ... X_K). El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción.

En el análisis de regresión lineal múltiple la construcción de su correspondiente ecuación se realiza seleccionando las variables una a una, "paso a paso". La finalidad perseguida es buscar de entre todas las posibles variables explicativas aquellas que más y mejor expliquen a la variable dependiente sin que ninguna de ellas sea combinación lineal de las restantes.

Dentro del análisis existen dos estadísticos a analizar: R^2 ya que mide la capacidad explicativa de la variable X sobre la variable Y. al introducir en el modelo otra variable regresoras el nivel explicativo será mayor entre las dos que solo con la primera o, en todo caso, no disminuirá, pues la primera variable continúa como explicativa, dado esto la interpretación de R^2 no solo debe considerar la muestra, sino también el número de variables explicativas incluidas en el modelo (Martínez Rodríguez, 2005).

Y el estadístico PRESS el cual mide la calidad del modelo de regresión y es definida como la suma de cuadrados de los errores de la predicción para la variable que represente a la desviación al cuadrado entre el valor observado y estimado (Valencia Delfa, Díaz-LLanos, & Calleja, 2003).

CAPÍTULO I. MARCO DE REFERENCIA

En el siguiente capítulo se mostraran las bases necesarias para llevar a cabo la investigación dentro de un enfoque estadístico describiendo el problema a resolver, los objetivos, justificación y alcances de la investigación.

1.1 Planteamiento del problema

Dadas las investigaciones anteriormente realizadas, se ha planteado que la cantidad de datos a guardar en un experimento, para aplicar la Validación cruzada deberá ser del 10%, no obstante dicho planteamiento podría no ser apropiado para cierta cantidad de datos dentro de un conjunto de información pequeño, se desconoce cuál número de datos que debería guardarse.

Esta investigación pretende determinar el número óptimo de datos que deben ser separados para realizar la validación cruzada, en función del nivel de ruido y del tamaño de conjunto de información

1.2 Objetivos

1.2.1 Objetivo General

Analizar la técnica de validación cruzada bajo diferentes condiciones de ruido para determinar la cantidad óptima de datos a guardar dado un nivel de ruido y una cantidad de observaciones en un conjunto determinado.

1.2.2 Objetivos Específicos

a) Determinar la cantidad de datos óptima a guardar para realizar la validación cruzada bajo diferentes condiciones de ruido.

b) Determinar como el número de datos a guardar afecta al estadístico PRESS y a la R^2 de la predicción.
c) Proporcionar lineamientos que faciliten la toma de decisiones al realizar la técnica de validación cruzada.

1.3 Preguntas de Investigación

a) ¿Por qué la literatura sugiere guardar el 10 por ciento de los datos?
b) ¿Cuál es la cantidad óptima de datos a guardar dado un tamaño del conjunto de información y un nivel de ruido?
c) Si se analizan diferentes conjuntos de datos de diversos tamaños bajo diferentes niveles de ruido al aplicar Validación cruzada, ¿Cómo se afectan, el estadístico PRESS y la $R^2_{predicción}$?

1.4 Hipótesis

Es posible determinar la cantidad óptima de datos a guardar dado un nivel de ruido y un número de observaciones para realizar una validación del modelo de regresión lineal utilizando la técnica de validación cruzada.

1.5 Justificación

El proyecto a realizar será de investigación, en el cual se analizarán diferentes tamaños de conjuntos de datos bajo diferentes niveles de ruido mediante simulación Monte Carlo, posteriormente aplicando un análisis de Validación cruzada se determinará la cantidad óptima de datos a guardar dependiendo la cantidad de datos y del nivel de ruido en el conjunto, ya que esto se dificulta cuando la cantidad de información disponible es reducida.

1.6 Alcances y limitaciones

a) Se analizarán conjuntos de información con una distribución uniforme de 30, 50, 100, 500 y 1000 datos.
b) Se trabajara con 3 niveles de ruido bajo, medio y alto.
c) Se utilizará un modelo verdadero de primer orden con interacciones para simular información mediante simulación Monte Carlo.
d) No se analizarán distribuciones con valores extremos.

CAPÍTULO II. MARCO TEÓRICO

El marco teórico que fundamenta esta investigación proporcionara una idea más clara acerca del tema. Se encontrarán los conceptos básicos, complementarios y específicos.

La Estadística analiza la relación o dependencia entre variables, conociendo el efecto que una o varias variables pueden causar sobre otra, de igual manera predice en mayor o menor grado valores en una variable a partir de otra. Una de las técnicas estadísticas más ampliamente utilizadas es la regresión lineal que modela la relación en un sistema a partir de un conjunto de datos (Ríos & Simpson , 2016).

2.1 Antecedentes

Dentro de la literatura consultada para la realización de la investigación que a continuación se muestra, se menciona que existen tres técnicas utilizadas comúnmente las cuales son usadas para validar modelos de regresión:

1. Uso de estadísticos PRESS y $R^2_{\text{predicción}}$.
2. Recopilación de nuevos datos.
3. Validación cruzada.

Hablando de Validación cruzada es recomendable guardar el 10 por ciento de los datos, sin embargo cuando el tamaño del conjunto de información es pequeño, no se conoce el porcentaje a guardar para realizar la Validación cruzada (Cox & Gaudard, 2013).

Con la investigación que se pretende realizar, se analizarán varios escenarios con diferentes niveles de ruido (varianza) y *n* datos para llevar a cabo la Validación cruzada.

2.2 Estadística inferencial

La estadística es una rama de la ciencia matemática encargada del diseño de experimentos, del análisis de datos y los procedimientos para inferir acerca de las características de una población con base en la información obtenida de una muestra. Comprende de un conjunto de técnicas para la estimación estadística de las características (frecuentemente los parámetros) de una población con base en una muestra obtenida de ella y, una vez estimados, tomar decisiones sobre esa población (Peró & Guàrdia Olmos, 2001).

Estas decisiones incluyen un factor de riesgo, dado que las características de la población se infieren aproximadamente, pero no se conocen con certeza. Por ello la estadística inferencial se utilizan conceptos de probabilidad (Rubin & Levin, 2004).

2.3 Regresión lineal

2.3.1 Regresión lineal simple

El modelo de regresión lineal simple consiste en un solo regresor x que tiene una relación con respuesta y, donde la relación es una línea recta. Este modelo de regresión lineal simple es (2.1):

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

Donde la ordenada en el origen β_0 y la pendiente β_1 son constantes desconocidas, y ε es un componente aleatorio de error. Se supone que los errores tienen promedio cero y varianza σ^2 desconocida. Además, se suele suponer que los errores no están correlacionados. Quiere decir que el valor de un error no depende del valor de cualquier otro error (Montgomery, Peck, & Vining, Introducción al análisis de regresión lineal., 2011).

2.3.2 Regresión lineal múltiple

El modelo de regresión lineal múltiple expresa la media de la variable de respuesta Y como función de una o más variables de predicción distintas $x_1, x_2, \dots, x_k$. Asume la forma siguiente (2.2):

$$\mu_{Y|x_1,x_2,\dots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2.2}$$

En este modelo se trata de *k* variables de predicción distintas, cada una de primer grado. La aplicación del método de mínimos cuadrados a la estimación de los parámetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ se efectúa al rescribir el modelo en la siguiente forma (2.3):

$$Y_{|x_1,x_2,\dots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{2.3}$$

Donde $Y_{|x_1,x_2,\dots,x_k}$, denota la variable de repuesta cuando las variables de predicción $X_1, X_2, \dots, X_k$ asumen los valores $x_1, x_2, \dots, x_k$ y ε denota la diferencia aleatoria entre $Y_{|x_1,x_2,\dots,x_k}$ y su valor medio. Una muestra aleatoria de tamaño *n* consiste en un conjunto de *n* (*k*+1)-tuplos y tiene la forma (2.4):

$$\{(x_{1i}, x_{2i}, \dots, x_{ki}, Y|x_{1i}, x_{2i}, \dots, x_{ki}) : i = 1,2,3, \dots, n\} \tag{2.4}$$

Donde cada uno de los primeros *k* miembros de cada *n* (*k*+1)-tuplo denota un número real, y el último, una variable aleatoria. Al omitir la notación condicional, se expresa la muestra como (2.5):

$$\{(x_{1i}, x_{2i}, \dots, x_{ki}, Y_i) : i = 1,2,3, \dots, n\} \tag{2.5}$$

Donde (2.6):

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \tag{2.6}$$

Se parte del supuesto de que los errores aleatorios $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, son independientes, con media 0 y varianza común σ^2.

En lo conceptual, es relativamente sencilla la idea de estimar un modelo de regresión lineal múltiple con el método de mínimos cuadrados. Sólo se amplían los conceptos de regresión lineal simple a un modelo más complejo (Milton & Arnold, 1999). Cualquier

modelo de regresión que es lineal en los parámetros (los valores β) es un modelo de regresión lineal, no importando la forma de la superficie de respuesta que el modelo genere (Anderson, Sweeney, & Williams, 2008), (Mendenhall, Beaver, & Beaver, 2010).

2.3.3 Estimación de los modelos del parámetro

2.3.3.1 Mínimos cuadrados

Es difícil la identificación de fórmulas para los estimadores de mínimos cuadrados en modelos complejos. La técnica busca llegar a la ecuación de regresión minimizando la suma de los cuadrados de las distancias verticales entre los valores y actuales y los valores y anticipados; en la regresión múltiple, en el proceso de inferencia estadística análogo, b_0, b_1, …, b_k denotan los estadísticos que se usan para estimar los parámetros β_0, β_1, …, β_k. (Walpole, Myers, & Ye, 2012). Para evitar estos problemas se hace uso del álgebra matricial.

En resumen, se hace lo siguiente:

a) Expresar el modelo lineal general en forma matricial.
b) Encontrar una expresión matricial de las ecuaciones normales de ese modelo.
c) Encontrar una expresión matricial de las estimaciones de mínimos cuadrados al despejar las ecuaciones normales.
d) Aplicar los resultados obtenidos a los modelos de regresión lineal múltiple.

El modelo de regresión general tiene la forma (2.2):

$$\mu_{Y|x_1,x_2,\dots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (2.2)$$

Y también puede escribirse como (2.7):

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1,2,3,\dots,n \qquad (2.7)$$

La formulación matricial del modelo se aprecia al escribir esas ecuaciones en forma expandida, como se muestra (2.7.1) (Milton & Arnold, 1999):

$$Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2$$

$$Y_3 = \beta_0 + \beta_1 x_{13} + \beta_2 x_{23} + \cdots + \beta_k x_{k3} + \varepsilon_3 \qquad (2.7.1)$$

$$\vdots$$

El método de mínimos cuadrados selecciona las β_s en la ecuación anterior. El desarrollo destallado de éste método se puede verificar en el capítulo 3 (Montgomery, Peck, & Vining, Introducción al análisis de regresión lineal., 2011). Quedando que el estimador de β por mínimos cuadrado (2.8):

$$\widehat{\beta} = (X'X)^{-1}X'y \qquad (2.8)$$

Donde $X'X$ es una matriz simétrica ($p \times p$) y $X'y$ es un vector columna ($p \times 1$). El modelo de regresión ajustado es (2.9):

$$\hat{y} = X\hat{\beta} \qquad (2.9)$$

En forma escalar el modelo ajustado es (2.10):

$$\widehat{y_i} = \hat{\beta}_0 + \sum_{i=1}^{k} \widehat{\beta_J}, x_{ij} \qquad (2.10)$$

Las diferencias entre la observación real y_i y el correspondiente valor ajustado $\hat{y}_i$, es el residuo, digamos $e_i = y_i - \hat{y}_i$. El vector de residuos ($n \times 1$) esta denotado por (2.11):

$$e = y - \hat{y} \qquad (2.11)$$

2.3.3.2 Estimación de σ^2

Para hacer inferencias sobre el modelo (β_0), es necesario encontrar una forma de estimar σ^2 (Gutiérrez & De La Vara , 2008). A partir de la suma de cuadrados de residuales

en la regresión lineal simple es posible desarrollar un estimador de σ^2; la ecuación llamada del error o suma de cuadrados residual, con $n-p$ grados de libertad asociados es (2.12):

$$SS_{Res} = y'y - \hat{\beta}'X'y \tag{2.12}$$

Los cálculos a detalle se pueden observar en el apéndice C.3 de (Montgomery, Peck, & Vining, 2011). El cuadrado medio residual o cuadrado medio de residuales es (2.13):

$$MS_{Res} = \frac{SS_{Res}}{(n-p)} \tag{2.13}$$

Un estimador insesgado de σ^2 está dado por (2.14):

$$\sigma^2 = \frac{SS_{Res}}{(n-p)} \tag{2.14}$$

2.4 Técnicas de validación

La validación de un modelo para usarse como ecuación de predicción se debe concentrar en la determinación de la exactitud del modelo. Dado que el desarrollador consiguientemente no controla el uso del modelo, se recomienda que, siempre que sea posible, se usen todas las técnicas de validación mencionadas (Snee, 1977), (Dette & Munk, 1998) (Arriaga Balderas, 2017).

2.4.1Estadístico PRESS

PRESS (del inglés, Prediction Error Sum of Squares). También se le suele llamar residuales eliminados, se usa para validar el modelo de regresión en términos de validación cruzada de errores o en términos de predicción. Es de notarse que entre más pequeño sea el valor de PRESS mejor será el modelo seleccionado (Kutner, Nachtsheim, & Neter, 2003).

PRESS se definen como la diferencia entre el valor observado y el valor estimado si se elimina la i-ésima observación (Buenaño Cordero, De la Cruz Cedeño, & Zurita, 2015), se ajusta el modelo de regresión a las $n-1$ observaciones restantes, y se calcula el valor predicho de yi correspondiente a la observación omitida, el error de predicción correspondiente es (2.15):

$$e_{(i)} = y_i - \hat{y}_{(i)} \tag{2.15}$$

El cálculo obtenido del error de predicción se repite para cada observación $i=1, 2, \ldots, n$. A esos errores de predicción se les llamar residuales PRESS. La Suma de Cuadrados de Error de Predicción está definida como (2.16) (Montgomery, Peck, & Vining, 2011):

$$e_{(i)} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 \tag{2.16}$$

2.4.2 R^2

Coeficiente de determinación múltiple (R^2) representa el porcentaje de variación de la variable explicado por el juego de variables independientes mostrada en la ecuación (2.17). Dicho criterio R^2 mide la fuerza de la relación lineal entre los componentes del modelo. Entre más alto sea el valor de R^2 las variables dependientes observadas estarán mejor ajustadas por el modelo de regresión aplicado. No obstante, es necesario aclarar que un valor alto de R^2, no necesariamente indicaría un buen modelo de regresión, como tampoco lo indicaría un valor de R^2 pequeño (Kutner, Nachtsheim, & Neter, 2003).

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_{Res}}{SS_T} \tag{2.17}$$

Por lo tanto, es igual a 1 menos la proporción de la suma de cuadrados totales (SS_T), que no es explicada por la regresión (SS_{Res}). De acuerdo con la definición de, debe cumplirse que:

$$0 \leq R^2 \leq 1$$

2.4.3 R^2 predicción basado en PRESS

Refleja el porcentaje de variabilidad que explica el modelo con los nuevos datos provenientes del estadístico PRESS calculando un estadístico parecido al R^2 (Arriaga Balderas, 2017)(2.18).

$$R^2_{predicción} = 1 - \frac{PRESS}{SS_T} \tag{2.18}$$

Provee cierta indicación de la capacidad predictiva del modelo de regresión y proporciona el porcentaje de la variabilidad cuando se predigan nuevas observaciones, en comparación con el porcentaje de variabilidad en los datos originales, explicado por el ajuste de mínimos cuadrados. (Montgomery, Peck, & Vining, 2007) (Valencia Delfa J. L., 2009).

2.5 Validación cruzada

La estrategia de Validación Cruzada consiste en la división de un conjunto de muestras que se pueden analizar en dos conjuntos disjuntos de datos. Uno de estos conjuntos entrenará las muestras que contiene, y los resultados obtenidos se aplicarán al otro conjunto que será utilizado para la clasificación de muestras. Hay que tener en cuenta que la división de las muestras en dos conjuntos fijos es una simplificación de la implementación de una división en k subconjuntos. El resultado se obtiene tras una optimización en cada iteración, acotando la probabilidad de error estimado como promedio de los errores en cada iteración (Hurtado, 2007).

La Validación Cruzada tiene una clara y grave desventaja, puesto que la división aleatoria de un pequeño conjunto de datos para el análisis implica la casi segura pérdida de información que no podrá ser recuperada. La pregunta clave sería determinar el número de conjuntos en los que se debe dividir para obtener el rendimiento óptimo. Aunque en cada

iteración se hace un promedio del error producido, existe el problema de que no hay representatividad de las muestras. Este grave problema lo solucionamos utilizando la técnica de la estratificación (Gong, 1982).

La forma más común de aplicar la técnica de Validación Cruzada es dejar el 10% de las muestras para realizar la evaluación y entrenar el 90% restante según la figura 2.1 (Cox & Gaudard, 2013):

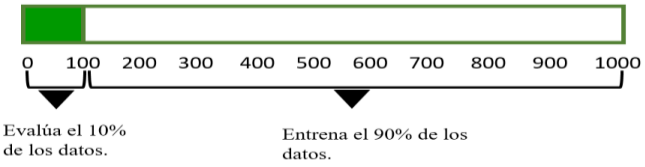

Figura 2.1 Técnica de Validación Cruzada sobre 1000 muestras (Fuente: Elaboración propia).

El costo principal de la Validación Cruzada es la intensidad computacional, éste tema se está convirtiendo cada vez más en secundario debido al rápido aumento de la potencia de cálculo y aunque la Validación Cruzada toma más de tiempo aplicarla inicialmente, de manera gradual proporciona más confianza y seguridad en las conclusiones resultantes para medir la exactitud de la predicción en el uso del modelo y simular la reproducción total o parcial de un estudio (Stone, 1974), (Fortmann-Roe, 2012).

2.5.1 Tipos de Validación Cruzada

2.5.1.1 Validación Cruzada dejando uno fuera (LOOCV)

Este es el caso límite de Validación Cruzada donde el conjunto de muestras X no se subdivide en N subconjuntos, sino que se realiza, $X = |Z|$, lo que proporciona una mejor estimación al utilizar todos los datos menos uno en el entrenamiento (Hoffman & van der

Merwe, 2002). Las muestras no se dividen en subconjuntos, simplemente omiten en cada iteración un dato que será aplicado para la clasificación, y entrena con los datos restantes.

El resultado con el dato que se clasifica se almacena, se pasa el siguiente dato a la fase de clasificación y se entrena con todos los demás incluyendo el dato que se clasificó la primera vez. Una vez clasificado el último dato de la muestra, calcula con todos los datos el porcentaje de acierto (Liu , Sun , Wei , & Liu , 2008) (Pérez-Planells L. , Delegido, Rivera-Caicedo, & Verrelst, 2015). Las ventajas de este algoritmo son que utiliza un mayor número de datos para entrenar, y que tiene un resultado determinista, pues el experimento siempre ofrecerá los mismos resultados (Kohavi, 1995).

Un inconveniente es que necesita ser ejecutado sobre conjuntos de muestras relativamente pequeños, ya que, si la muestra a evaluar es muy grande, el coste computacional de recursos puede ser excesivo gastando demasiado tiempo en la ejecución del proceso. Otra característica es que el uso del LOOCV no puede ser combinado con la estratificación porque como hemos dicho, esta forma de evaluación extrae las muestras de una en una, lo que significa que, aunque el orden sea distinto no habrá variación de los resultados (Ke-Lin & Swamy, 2014).

El error puede ser calculado como (2.19):

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{2.19}$$

Donde (2.20):

$$MSE_i = (y_i - \hat{y}_i)^2 \tag{2.20}$$

Suponiendo 10 muestras, el proceso LOOCV que aparece en la Figura 2.2, representa en cada iteración una muestra del conjunto total para la clasificación, entrenando con todas las demás. Al finalizar el entrenamiento hace una media de los resultados extraídos en cada iteración

obteniendo el porcentaje de acierto del algoritmo para ese conjunto de muestras (Refaeilzadeh, Tang, & Liu, 2008).

Figura 2.2 Leave-One-Out sobre diez muestras (Fuente: Elaboración propia).

Una estimación de la precisión obtenida usando LOOCV es conocida por ser algo imparcial y tener un alto valor de varianza, dejando estimaciones poco fiables (Efron & Tibshirani, 1993).

2.5.1.2 Validación cruzada de k- interacciones

Los datos de muestra se dividen en k subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto $(k-1)$ de los subconjuntos como datos de entrenamiento. El proceso es repetido durante k-interacciones, con cada uno de los posibles subconjuntos de datos de prueba. Es así como se obtiene el estadístico PRESS de los resultados de cada una de las interacciones para obtener un único resultado del modelo.

El error se calcula como la media aritmética de los errores de cada iteración para obtener un único resultado (2.21).

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \quad (2.21)$$

Donde MSE_i denota el error en la iteración i-ésima (2.22):

$$MSE_i = (y_i - \hat{y}_i)^2 \quad (2.22)$$

El método es muy preciso evalúa a partir de las k combinaciones de datos de entrenamiento y de prueba, teniendo aun así una desventaja ya que es lento desde el punto de vista computacional. La elección del número de interacciones depende de la medida del conjunto de datos. Lo más común es utilizar la validación cruzada de 10-interacciones (Joanneum, 2005) (Werner, Granzow, & Berrar, 2007).

2.5.1.3 Validación Cruzada de 10 interacciones

Cuando el conjunto de muestras se separa en bloques *X* y el número de estos es 10, se determina el *10-fold Cross-Validation*. Validación Cruzada no es una técnica desconocida ya que diversos estudios la han considerado una medida fiable que ofrece buena complejidad en tiempo real (Bouckaert , 2008). El modelo repite diez veces el proceso de entrenamiento y el porcentaje de acierto resultante es utilizado como una medida de bondad del algoritmo evaluado (Lakshmanan, Fritz , Smith, Hondl, & Stumpf, 2007).

2.5.1.4 Validación cruzada Monte Carlo

Conocida como submuestreo aleatorio repetido, repite el proceso de división de datos dentro de un conjunto de entrenamiento y un conjunto de prueba, en base al modelo desarrollado sobre el conjunto de entrenamiento la tasa de error evaluada sobre el conjunto de prueba. El conjunto de prueba estimado es promediado en el conjunto de prueba dividido aleatoriamente. Para cada partición de entrenamiento-prueba, cada caso aparece en cualquiera de los dos, pero no en ambos, debido a que la división entrenamiento-prueba es aleatoria, el caso aparece en el conjunto de prueba un número variable de veces (Arriaga Balderas, 2017).

El método consiste en dividir aleatoriamente el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Para cada división la función de aproximación se ajusta a partir de los datos de entrenamiento y calcula los valores de salida para el conjunto de datos de prueba. El resultado final corresponde a la media aritmética de los valores obtenidos para las diferentes divisiones. Con este método hay algunas muestras que quedan sin evaluar y otras que se evalúan más de una vez, es decir, los subconjuntos de prueba y entrenamiento se pueden solapar (Zhang, 2011) (Goos, Hartmanis, & Leeuwen, 2003) (Arriaga Balderas, 2017).

2.5.2 Medidas de ajuste

2.5.2.1 Error cuadrático medio

Se le llama error cuadrático, y es el que nos define el error que tenemos con el valor verdadero al tomar como valor de este último el más probable, el cual ya dijimos era la media aritmética (Ansley & Kohn, 1986). Si llamamos ε_m a éste, su valor será (2.23):

$$\varepsilon_m = \pm\frac{s}{\sqrt{n}} = \pm\sqrt{\frac{\sum \delta^2}{n(n-1)}} \tag{2.23}$$

Y por lo tanto podemos decir que (2.24):

$$x = \bar{x} \pm \varepsilon_m \tag{2.24}$$

2.5.2.2 Media cuadrática

La media cuadrática es igual a la raíz cuadrada de la suma de los cuadrados de los valores dividida entre el número de datos (2.25):

$$\bar{x} = \sqrt[2]{\frac{\sum_{i=1}^{n} a_i^2}{n}} = \sqrt[2]{\frac{a_1^2+a_2^2+\cdots+a_n^2}{n}} \tag{2.25}$$

A veces la variable toma valores positivos y negativos, como ocurre, por ejemplo, en los errores de medida. En tal caso se puede estar interesado en obtener un promedio que no recoja los efectos del signo (Vargas Sabadías, 1995).

Este problema se resuelve, mediante la denominada media cuadrática. Consiste en elevar al cuadrado todas las observaciones (así los signos negativos desaparecen), en obtener después su media aritmética y en extraer, finalmente, la raíz cuadrada de dicha media para volver a la unidad de medida original (Sominski, 1975).

2.6 Nivel de ruido

Se considera como la variación de la salida que es impredecible respecto de las entradas. Este ruido, está claramente relacionado con la idea física de ruido, definido como la parte de una señal que es debida a un proceso estocástico (aleatorio). El ruido en los datos limita la capacidad de predicción de los modelos a que dan lugar (Arriaga Balderas, 2017).

El ruido en las variables de entrada es un serio problema para la construcción de modelos, aunque sus efectos son más difíciles de analizar. En las zonas del espacio de entrada donde la función a aproximar tiene poca pendiente, el ruido en el regresor tiene poca importancia en las zonas donde la pendiente es alta, este ruido puede tener efectos catastróficos (Arahal, Berenguel Soria, & Rodríguez Díaz, 2006).

La mejor forma de evitar los malos efectos del ruido consiste en tener una gran cantidad de datos, de manera que las contribuciones del ruido en varias medidas similares se cancelen unas a otras; la presencia de ruido en la variable de salida provoca que el error de generalización (error cuadrático medio en los datos futuros) nunca pueda ser menor que la varianza del mismo, independientemente de la cantidad de datos disponibles para entrenamiento o del tipo de modelo usado. De hecho, si se intenta reducir el error de predicción en los datos usados para entrenamiento a niveles por debajo de dicha varianza hay grandes probabilidades de que el modelo resultante esté sobre parametrizado (Arriaga Balderas, 2017).

2.7 Simulación Monte Carlo

Los métodos Monte Carlo son cálculos numéricos que utilizan una secuencia de números aleatorios para llevar a cabo una simulación estadística, con el fin de conocer algunas propiedades estadísticas del sistema. Estos métodos de simulación están en contraste con los métodos numéricos de discretización aplicados para resolver ecuaciones diferenciales parciales que describen el comportamiento de algún sistema físico o matemático (Piñeiro Redondo, 2007).

La característica esencial de Monte Carlo es el uso de técnicas de toma de muestras aleatorias para llegar a una solución del problema físico, mientras una solución numérica convencional inicia con un modelo matemático del sistema físico, discretizando las ecuaciones diferenciales para luego resolver un grupo de ecuaciones algebraicas (Taha, 2004).

Se requiere una forma rápida y efectiva de generar números aleatorios con cierta distribución, dando inicio a la simulación generando muestreos aleatorios de la misma. Posteriormente de múltiples simulaciones, el resultado deseado será tomado con el valor promedio de los resultados obtenidos en cada simulación (Figura 2.3).

En muchas aplicaciones prácticas se puede predecir un error estadístico (varianza) para este promedio y por tanto una estimación del número de simulaciones necesarias para conseguir un error dado. De entre todos los métodos numéricos basados en evaluación de n organismos en espacios de dimensión r, los métodos de Monte Carlo tienen asociado un error absoluto de estimación que decrece como $\sqrt{n}$ mientras que para el resto de métodos tal error decrece en el mejor de los casos como $\sqrt[r]{n}$ (Drakos, 1995).

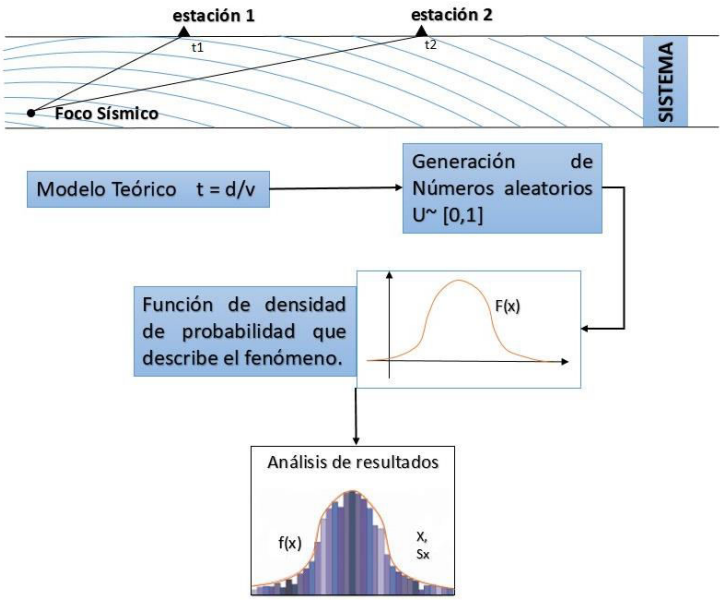

Figura 2.3 Esquematización de una simulación mediante el método Monte Carlo.

(Fuente: *(Arriaga Balderas, 2017)*).

2.8 Estado del arte

Tabla 2.1 Estado del arte.

Título de la investigación	*Autores*	*Revista/Año*	*Resumen*
Selecting the number of components in principal component analysis using cross-validation approximations	Julie Josse ,François Husson	Computational Statistics & Data Analysis, Volume 56, Issue 6, June 2012, Pages 1869-1879	El autor prueba la eficacia del método de Validación cruzada para seleccionar el número de componentes en el Análisis de Componentes Principales (PCA), sin embargo, muestra que el principal inconveniente es su coste computacional. Señala que en una regresión (o en una regresión no paramétrica) se establecen criterios tales como la Validación Cruzada General (GCV) que proporcionan aproximaciones convenientes con la licencia-uno-fuera de la Validación Cruzada. Se basa en la relación entre el error de predicción y la suma residual de cuadrados ponderados. Tal relación se establece entonces en el PCA con una presentación original de PCA con una matriz de proyección única. Permitiendo la definición de dos criterios de Validación Cruzada de aproximación: la aproximación de suavizar el criterio de Validación Cruzada (SACV) y el criterio GCV.

A survey of cross-validation procedures for model selection	Sylvain Arlot, Alain Celisse	Statist. Surv. Volume 4 (2010), 40-79	El artículo menciona la utilidad de la Validación Cruzada para estimar el riesgo de un estimador o para llevar a cabo la selección del modelo siendo una estrategia generalizada debido a su sencillez y su universalidad (aparente). Señala que existen muchos resultados sobre selección de modelo de procedimientos de validación cruzada. Aporta una encuesta que tiene la intención de relacionar estos resultados con los más recientes avances de la teoría de selección del modelo, con un especial énfasis en distinguir afirmaciones empíricas de resultados teóricos rigurosos. Proporcionan pautas para elegir el mejor procedimiento de validación cruzada según las características particulares del problema en mano.

Monte Carlo Cross Validation	Qing-Song Xu , Yi-Zeng Liang	Chemometrics and Intelligent Laboratory Systems 56 (2001) 1–11	Para escoger correctamente la dimensión del modelo de calibración en química, un nuevo método simple y efectivo llamado Validación Cruzada con Monte Carlo (MCCV). A diferencia de la licencia uno fuera el procedimiento comúnmente utilizado en Quimiometría es Validación Cruzada (CV), la Validación Cruzada con Monte Carlo es un método asintóticamente consistente en determinar el número de componentes en el modelo de calibración. Se puede evitar un innecesario modelo grande y por lo tanto disminuye el riesgo de sobre ajuste para el modelo de calibración. Los resultados obtenidos del estudio de simulación demostraron que MCCV tiene una probabilidad más grande que la licencia uno fuera CV en elegir el número correcto de los componentes que debe contener el modelo. Los resultados de conjuntos de datos reales demostraron que MCCV con éxito podría elegir el modelo adecuado, la licencia uno fuera CV no.

Applicability of Monte Carlo Cross Validation technique for model development and validation using generalised least squares regression	Khaled Haddad, Ataur Rahman , Mohammad A Zaman, Surendra Shrestha	Journal of Hydrology Volume 482, 4 March 2013, Pages 119–128	En el análisis de regresión hidrológico regional, la validación y selección del modelo se consideran pasos importantes. Aquí, la selección de modelo generalmente se basa en unas medidas de bondad de ajuste entre la predicción del modelo y los datos observados. El Análisis Regional de la Frecuencia de Inundación (RFFA), con una Validación de licencia uno fuera (LOO) o dejar un porcentaje fijo fuera de la validación (por ejemplo, 10%) es adoptado comúnmente para evaluar la capacidad predictiva de las ecuaciones de predicción basadas en regresión. Este trabajo desarrolla una técnica de Validación Cruzada con Monte Carlo (MCCV) (que ha sido adoptada ampliamente en Quimiometría y Econometría) en RFFA utilizando regresión de mínimos cuadrados generalizados (GLSR) y compara con la mayoría que comúnmente adoptó el enfoque de validación LOO.

CAPÍTULO III. MARCO METODOLÓGICO

3.1 Marco metodológico

El siguiente apartado muestra las técnicas y procesos empleados para llevar a cabo dicha investigación, usando herramientas como la simulación Monte Carlo y el software MATLAB, es una herramienta muy útil y de alto nivel para desarrollar aplicaciones técnicas, fácil de utilizar y que puede ayudar a aumentar significativamente la productividad de los programadores respecto a otros entornos para el desarrollo de cálculos numérico, visualizaciones avanzadas y resultados aptos para el trabajo científico y la ingeniería (Benítez López & Hueso Pagoaga, 1999).

La siguiente figura (3.1) describe el procedimiento del marco metodológico que se seguirá en esta investigación.

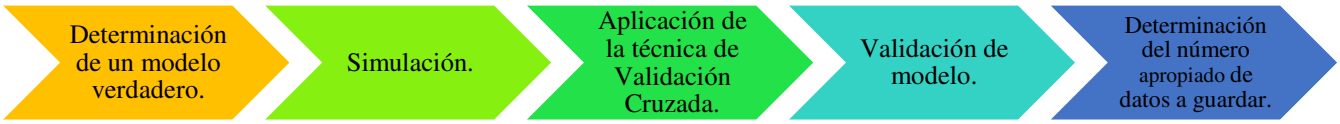

Figura 3.1 Proceso del marco metodológico (Fuente: Elaboración propia).

En el cual muestra como primera parte la determinación de un modelo verdadero que contenga variables de primer orden e interacciones de dos factores únicamente; posteriormente se realizará una simulación de información mediante Simulación Monte Carlo; dicho conjunto de información se separa aleatoriamente en dos bloques, uno de predicción y otro de estimación por el método de Validación Cruzada, se utilizarán los datos obtenidos para ajustar un nuevo modelo a partir de una regresión paso a paso, la cual,

permitirá conocer las variables e interacciones significativas para el modelo de regresión lineal.

El modelo de regresión obtenido se probará en los datos de predicción para obtener el estadístico PRESS (Suma de Cuadrados de las Errores de Predicción) y de igual manera se calculará la R^2 aproximada para predicción; con ambos datos se analizará cual conjunto de datos presenta el menor porcentaje de variabilidad, también conocido como desempeño predictivo del modelo.

Se observará la capacidad predictiva del modelo ajustado comparando los porcentajes de separación de datos realizado con una muestra de n tamaño y nivel de ruido deseado (alto, medio y bajo); se llevaran a cabo simulaciones de conjuntos de datos mediante la programación de un algoritmo en el software de MATLAB y determinar el número óptimo de datos que deben ser guardados sin perder la capacidad predictiva del modelo ni sacrificar un número excesivo de conjunto de información comparando los resultados de cada conjunto.

Cabe mencionar que esta metodología propuesta se sustenta por la metodología empleada en la investigación previa "*Determinación número óptimo de datos para realizar una validación cruzada"* de Andrea Balderas (2017) donde es analizado un solo nivel de ruido con *n* tamaño de muestras.

Dicho lo anterior, se muestra en los siguientes puntos de manera más detallada las etapas de la metodología.

3.1.1 Determinación de un modelo verdadero

Identificar la información que cumpla con los requisitos para la determinación de un modelo verdadero. Se seleccionará un modelo lineal múltiple, con tres variables regresoras (x_1, x_2, x_3), tres interacciones de dos factores (x_1x_2, x_1x_3, x_2x_3) y un error aleatorio U (0,1). Los coeficientes de regresión son al menos tres veces mayores que la varianza del error, para ser detectados como significativos.

Dicha propuesta tiene como objetivo crear conjuntos de información que tengan variables con un determinado nivel de ruido. Se define el modelo verdadero con la fórmula general (3.1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon \qquad (3.1)$$

El modelo verdadero como ejemplo utilizado que cumple las características anteriores se estructura de la siguiente manera, con un nivel bajo de ruido (4.1):

$$y = 10 + 6x_1 - 5x_2 + 7x_3 - 6.8x_1 x_2 + 6.3x_1 x_3 + 0.6x_2 x_3 + \varepsilon(0,2.1) \qquad (4.1)$$

Una vez determinado el modelo, se realizan los diferentes escenarios propuestos para el desarrollo de la experimentación.

3.1.2 Simulación

Se simularán conjuntos de *n* (*n*=30, 50, 100, 500 y 1000) cantidad de datos y un nivel de ruido (bajo ($3\sigma^2$), medio ($2\sigma^2$) y alto (σ^2)) según los requisitos del experimento, se usará la simulación Monte Carlo para generarlos ya que utiliza una secuencia de números aleatorios para llevar a cabo una simulación estadística. Recordando que una característica esencial de Monte Carlo es el uso de técnicas de toma de muestras aleatorias para llegar a una solución del problema físico (Taha, 2004).

Los datos uniformes que serán simulados tendrán como característica abarcar un rango de (U (-1,1)), los cuales serán generados en el programa MATLAB y se utilizarán para generar una matriz que contenga los efectos principales para el modelo de regresión lineal múltiple, con los cuales se calcularán los valores para la variable de respuesta *y*.

El comando que se usará es el siguiente:

```
random('Unif',-1,1)
```

Los datos generados (tabla 3.1) por el comando proporcionan los valores de los efectos principales del modelo de regresión, los cuales son usados para determinar la variable de respuesta *y* y las interacciones tomadas en cuenta para aplicar la regresión paso a paso.

Tabla 3.5 Matriz de efectos principales del modelo de regresión lineal, nivel bajo de ruido.

	x_1	x_2	x_3
1	0.0643	-0.5046	-0.1254
2	0.3382	0.0954	0.2181
3	0.7263	-0.2386	0.4979
4	-0.6866	-0.8838	-0.3206
5	0.6344	-0.2449	0.9452
⋮	⋮	⋮	⋮
30	0.46495	-0.6741	0.84215

Una vez obtenido los valores de los efectos principales, determinamos el valor del error aleatorio (tabla 3.2) para obtener el valor de la variable de respuesta *y* (tabla 3.3), dentro del software utilizamos el siguiente comando:

```
vectorerror(ll,1)=icdf('Normal',(random('Unif',0,1)),media, (varianza^1/2));
```

Tabla 3.6 Valor del error aleatorio para determinar *y*.

	Ɛ (0,2.1)
1	-0.8029
2	-1.4503
3	-1.5212
4	0.7742
5	0.9519
⋮	⋮
30	0.2828

Manteniendo la distribución inicial en todo momento, en este caso una distribución normal tomando en cuenta su media (μ) y desviación estándar (σ).

Tabla 3.3 Matriz de efectos principales con sus interacciones y su variable de respuesta.

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	y
1	0.0643	-0.5046	-0.1254	-0.0324	-0.0081	0.0633	11.6361
2	0.3382	0.0954	0.2181	0.0323	0.0738	0.0208	11.5374
3	0.7263	-0.2386	0.4979	-0.1733	0.3616	-0.1188	20.1035
4	-0.6866	-0.8838	-0.3206	0.6068	0.2201	0.2833	6.7728
5	0.6344	-0.2449	0.9452	-0.1554	0.5996	-0.2315	25.7816
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	0.465	-0.6741	0.8422	-0.3134	0.3916	-0.5677	25.2481

3.1.3 Aplicación de la técnica de Validación cruzada

Teniendo los conjuntos de datos de acuerdo a las especificaciones del experimento, es decir *n* cantidad de datos y nivel de ruido mencionados anteriormente se llevar a cabo la técnica de Validación Cruzada que consiste en la división de un conjunto de muestras que se analizará en dos conjuntos de datos. Uno de los conjuntos entrenará las muestras que contiene y los resultados obtenidos se aplicarán al otro conjunto. El resultado se obtiene tras una optimización en cada iteración, acotando la probabilidad de error estimado como promedio de los errores en cada iteración (Hurtado, 2007).

Se realizará la técnica mediante la selección aleatoria sin reemplazo en MATLAB para la separación del conjunto de información en datos de estimación y predicción; donde se establece que la selección se hace por bloques con una cantidad determinada de datos fijada por el porcentaje (10%, 20%, 30%, 40%) de datos a analizar (Figura 3.2).

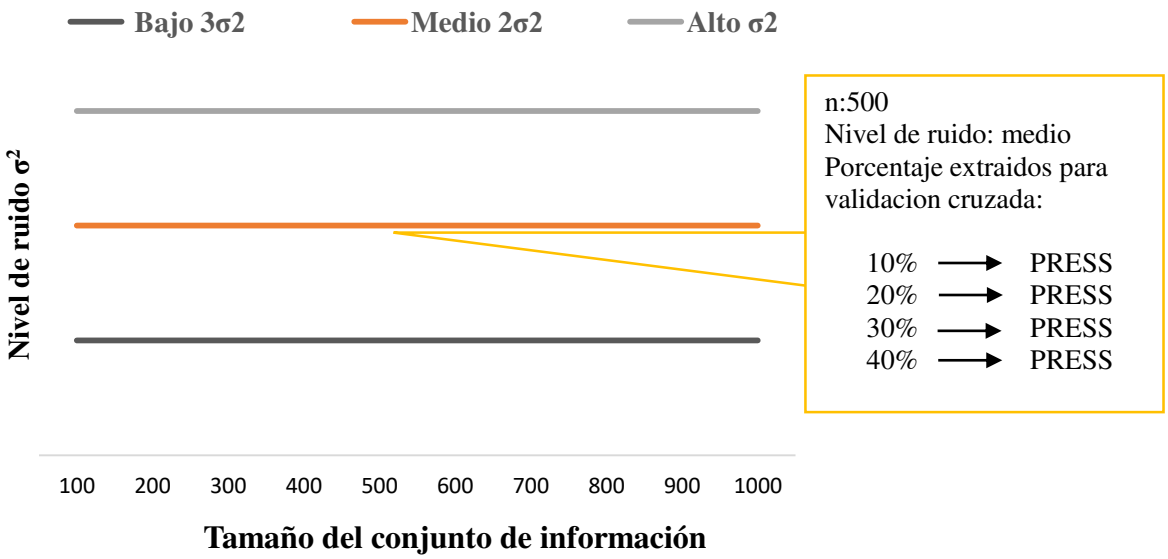

Figura 3.2 Planteamiento para realizar validación cruzada en los diferentes escenarios posibles (Fuente: Elaboración propia).

3.1.4 Determinación del número apropiado de datos a guardar

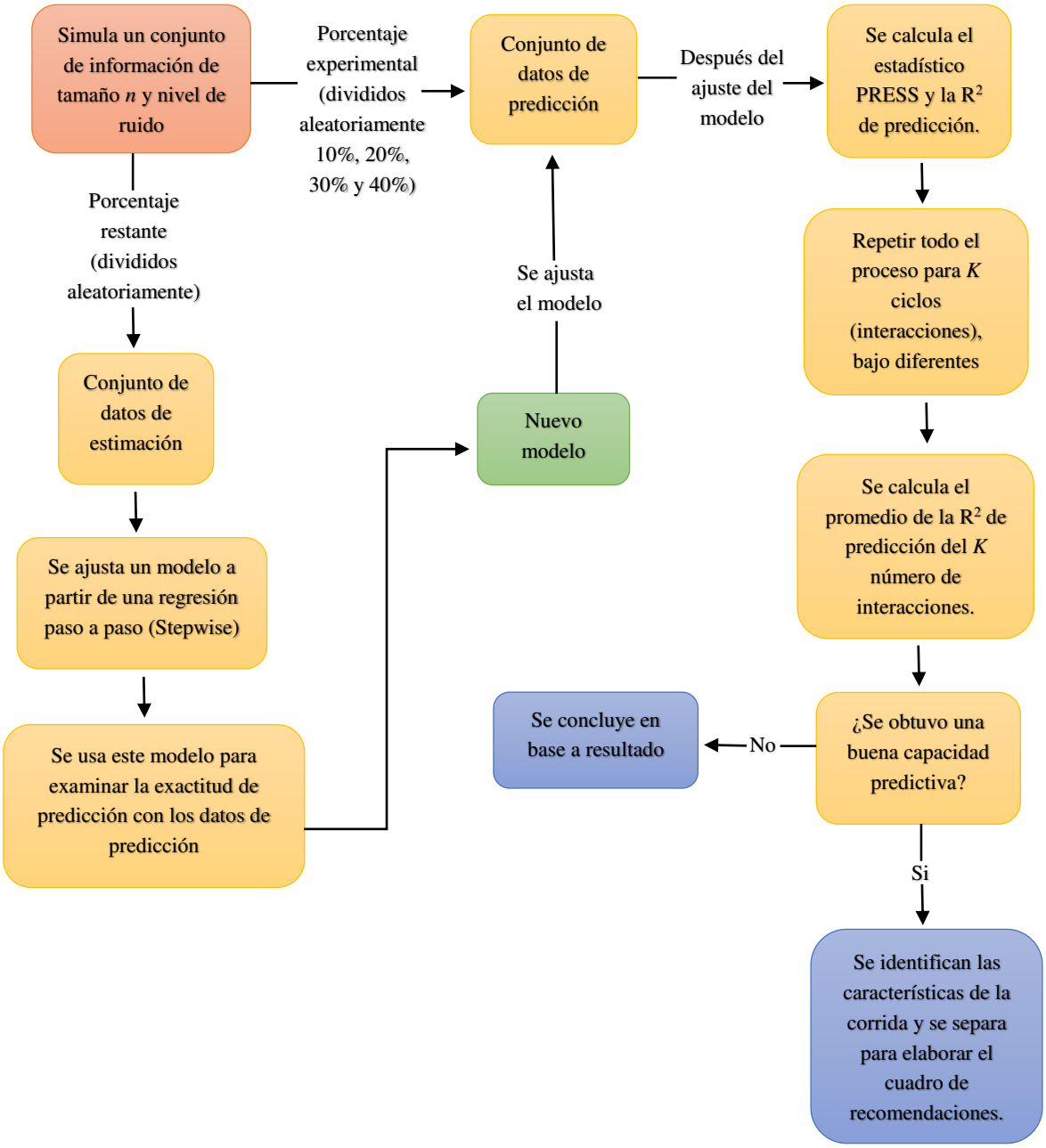

Figura 3.3 Planteamiento de la secuencia de pasos a realizar en el software MATLAB (Fuente: Elaboración propia).

Como se menciona anteriormente en la figura 3.3, los datos de predicción serán divididos aleatoriamente en 10%, 20%, 30% y 40% respectivamente; como ejemplo

tomaremos el 10% de los datos que anteriormente se mostraron. Teniendo datos de estimación de ***27x7***, es decir 27 renglones por siete columnas, las cuales incluyen la variable *y;* y los datos de predicción de ***3x7***, tres renglones y siete columnas.

Después de obtener las matrices divididas aleatoriamente se comprueba la no repetitividad de la numeración de una matriz en otra ("número aleatorio"), entonces el algoritmo se programa para que tomar la información correspondiente a lo largo de cada renglón y formar de esta manera dos nuevos conjuntos de datos siendo unos las matriz de predicción (tabla 3.6) y los otros matriz de estimación (tabla 3.5).

Tabla 3.4 Número aleatorio otorgado ya sea de predicción o estimación.

Estimación	Predicción
13	2
20	12
22	18
7	
3	
⋮	⋮

Tabla 7.5 Matriz de estimación.

Estimación	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	y
13	0.2973	0.5961	-0.5591	0.1772	-0.1662	-0.3333	0.1733
20	-0.9418	-0.7276	0.3891	0.6853	-0.3665	-0.2831	3.3979
22	0.5874	0.0037	-0.4467	0.0022	-0.2624	-0.0017	10.2797
7	0.7968	0.7014	-0.4864	0.5589	-0.3876	-0.3412	2.9359
3	0.7263	-0.2386	0.4979	-0.1733	0.3616	-0.1188	20.1035

Tabla 3.6 Matriz de Predicción.

Predicción	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	y
2	0.3382	0.0954	0.2181	0.0323	0.0738	0.0208	11.5374
12	-0.7343	0.5170	0.1305	-0.3796	-0.0958	0.0675	7.3984
18	-0.1376	-0.0487	0.5704	0.0067	-0.0785	-0.0278	11.3900

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	y
13	0.2973	0.5961	-0.5591	0.1772	-0.1662	-0.3333	0.1733
20	-0.9418	-0.7276	0.3891	0.6853	-0.3665	-0.2831	3.3979
22	0.5874	0.0037	-0.4467	0.0022	-0.2624	-0.0017	10.2797
7	0.7968	0.7014	-0.4864	0.5589	-0.3876	-0.3412	2.9359
3	0.7263	-0.2386	0.4979	-0.1733	0.3616	-0.1188	20.1035
21	0.0314	0.0852	0.6170	0.0027	0.0193	0.0525	13.7327
28	0.6432	-0.3902	-0.3613	-0.2510	-0.2323	0.1410	14.9154
16	-0.6595	0.3280	0.0717	-0.2163	-0.0473	0.0235	5.5629
30	0.4650	-0.6741	0.8422	-0.3134	0.3916	-0.5677	25.2481
17	0.6582	-0.4653	-0.6477	-0.3063	-0.4263	0.3014	12.5076

Estimación	Predicción
13	2
20	12
22	18
7	
3	
21	
28	
16	
30	
17	
29	
27	

	x1	x2	x3	x1x2	x1x3	x2x3	y
1	0.0643	-0.5046	-0.1254	-0.0324	-0.0081	0.0633	11.6361
2	0.3382	0.0954	0.2181	0.0323	0.0738	0.0208	11.5374
3	0.7263	-0.2386	0.4979	-0.1733	0.3616	-0.1188	20.1035
4	-0.6866	-0.8838	-0.3206	0.6068	0.2201	0.2833	6.7728
5	0.6344	-0.2449	0.9452	-0.1554	0.5996	-0.2315	25.7816
6	0.2106	-0.3235	0.8560	-0.0681	0.1803	-0.2769	19.6137
7	0.7968	0.7014	-0.4864	0.5589	-0.3876	-0.3412	2.9359
8	-0.4290	0.5599	0.4028	-0.2402	-0.1728	0.2255	9.3345
9	-0.0150	0.9353	-0.0477	-0.0140	0.0007	-0.0446	4.9605
10	0.9898	-0.0188	0.0069	-0.0187	0.0068	-0.0001	17.0760
11	0.5375	-0.2238	-0.0935	-0.1203	-0.0503	0.0209	14.1487
12	-0.7343	0.5170	0.1305	-0.3796	-0.0958	0.0675	7.3984
13	0.2973	0.5961	-0.5591	0.1772	-0.1662	-0.3333	0.1733
14	0.7158	0.8095	-0.4160	0.5794	-0.2977	-0.3367	4.1683
15	0.4517	-0.3211	-0.4546	-0.1451	-0.2054	0.1460	11.8268
16	-0.6595	0.3280	0.0717	-0.2163	-0.0473	0.0235	5.5629
17	0.6582	-0.4653	-0.6477	-0.3063	-0.4263	0.3014	12.5076
18	-0.1376	-0.0487	0.5704	0.0067	-0.0785	-0.0278	11.3900
19	-0.7387	-0.8973	0.2550	0.6628	-0.1884	-0.2288	4.5519

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	y
2	0.3382	0.0954	0.2181	0.0323	0.0738	0.0208	11.5374
12	-0.7343	0.5170	0.1305	-0.3796	-0.0958	0.0675	7.3984
18	-0.1376	-0.0487	0.5704	0.0067	-0.0785	-0.0278	11.3900

Figura 3.4 Esquematización sobre el análisis de la no repetitividad en la matriz principal en base a las matrices de estimación y predicción (Elaboración propia).

Una vez llevado a cabo cada escenario propuesto dentro del experimento se realizará una comparación entre las respuestas obtenidas en cada caso (PRESS y R^2) seleccionando el escenario más favorable para la respuesta a la hipótesis planteada en Capitulo I.

3.1.5 Validación del modelo

Se validará el modelo mediante la selección del número de factores óptimos que aseguren la capacidad predictiva sin que se produzca sobreajuste, esto se debe a que los factores están jerarquizados, de tal manera que cada nuevo factor que se añade al modelo, describe menos variación sistemática y más variación aleatoria o ruido.

El procedimiento construye de manera iterativa una secuencia de modelos de regresión con los datos de "estimación" mediante la adición o eliminación de variables en cada paso. El criterio para añadir o eliminar una variable en cualquier paso se expresa de manera usual en términos de una prueba parcial F. La regresión por pasos forma un modelo con una variable que utiliza la variable de regresión que tiene la correlación más grande con la variable de respuesta y; ésta también es la variable de regresión que produce el estadístico F más grande (Arriaga Balderas, 2017).

A continuación, se calcula el estadístico F de cada variable y de cada variable de regresión del modelo, y se elimina la que tenga el valor observado de F más pequeño y el procedimiento continúa hasta que ya no pueda añadirse o eliminarse ninguna otra variable de regresión.

Para efectuar la regresión paso a paso, en el software MATLAB se utiliza el siguiente comando:

```
[b,se,pval,inmodel,stats,nextstep,history] = stepwisefit(X,Y)
```

El comando obtiene un vector $p\times1$ de los coeficientes estimados para todos los términos en x. Mientras que la función $stepwisefit$ calcula los valores estimados de los coeficientes en b de dos maneras:

a) Si un término no está dentro del modelo final, entonces el coeficiente de estimación en b correspondiente agrega sólo ese término a los predictores en el modelo final.

b) Si un término está en el modelo final el coeficiente de estimación en b para ese término es el resultado del modelo final, ya que $stepwise$ no toma en cuenta los términos excluidos del modelo al calcular dichos valores.

De la información que arroja la función "$stepwisefit$" se tomará en cuenta los siguientes datos para la programación que se llevara a cabo posteriormente:

- $inmodel$: es un vector lógico, con longitud igual al número de columnas en X, especificando cuáles términos se encuentran en el modelo final.
- se: es un vector con los errores estándar para b.
- $pval$: es un vector $p-value$ para probar si los elementos de b son 0.
- $stats$: representa una estructura de estadísticas adicionales en donde todas las estadísticas corresponden al modelo final excepto el observado (figura 3.5).

1x1 struct with 17 fields

Field	Value	Min	Max
source	'stepwisefit'		
dfe	20	20	20
df0	6	6	6
SStotal	1.2971e+03	1.2971...	1.2971...
SSresid	24.7297	24.7297	24.7297
fstat	171.4982	171.49...	171.49...
pval	4.0453e-16	4.0453...	4.0453...
rmse	1.1120	1.1120	1.1120
xr	[ ]		
yr	27x1 double	-2.4921	1.8332
B	[7.0774;-4.8258;6.5752;-4.9353;6.1651;2.6078]	-4.9353	7.0774
SE	[0.5655;0.4503;0.6016;0.9868;0.7162;1.1522]	0.4503	1.1522
TSTAT	[12.5155;-10.7170;10.9303;-5.0013;8.6075;2.2634]	-10.71...	12.5155
PVAL	[6.4429e-11;9.7528e-10;6.9475e-10;6.8528e-05;3.6949e-08;0.0349]	6.4429...	0.0349
covb	6x6 double	-0.0953	1.3276
intercept	9.8829	9.8829	9.8829
wasnan	27x1 logical		

Figura 3.5 Ejemplo de *stats* obtenido de un escenario de *n*=30 con un nivel bajo de ruido en MATLAB.

Como se ha mencionado anteriormente se encuentra la matriz de predicción y la matriz de estimación, aplicando la programación obtenemos la regresión $stepwisefit$ donde la matriz de estimación es la que se somete a este análisis, de dicha manera obtenemos los factores significativos para el modelo mediante el $inmodel$ ya que es un vector lógico, con longitud igual al número de columnas en X y especifica cuáles términos se encuentran en el modelo final.

El tamaño del vector $inmodel$ corresponde a la cantidad de efectos principales e interacciones del modelo. Donde además uno, significa que se encuentra dentro del modelo y cero se considera fuera del mismo. Por lo que se planteará un nuevo modelo de regresión lineal en base a los valores obtenidos en "$inmodel$".

$$tamaño\ del\ vector\ inmodel = [x_1,_2,x_3,x_1x_2,x_1x_3,x_2x_3\ ]$$

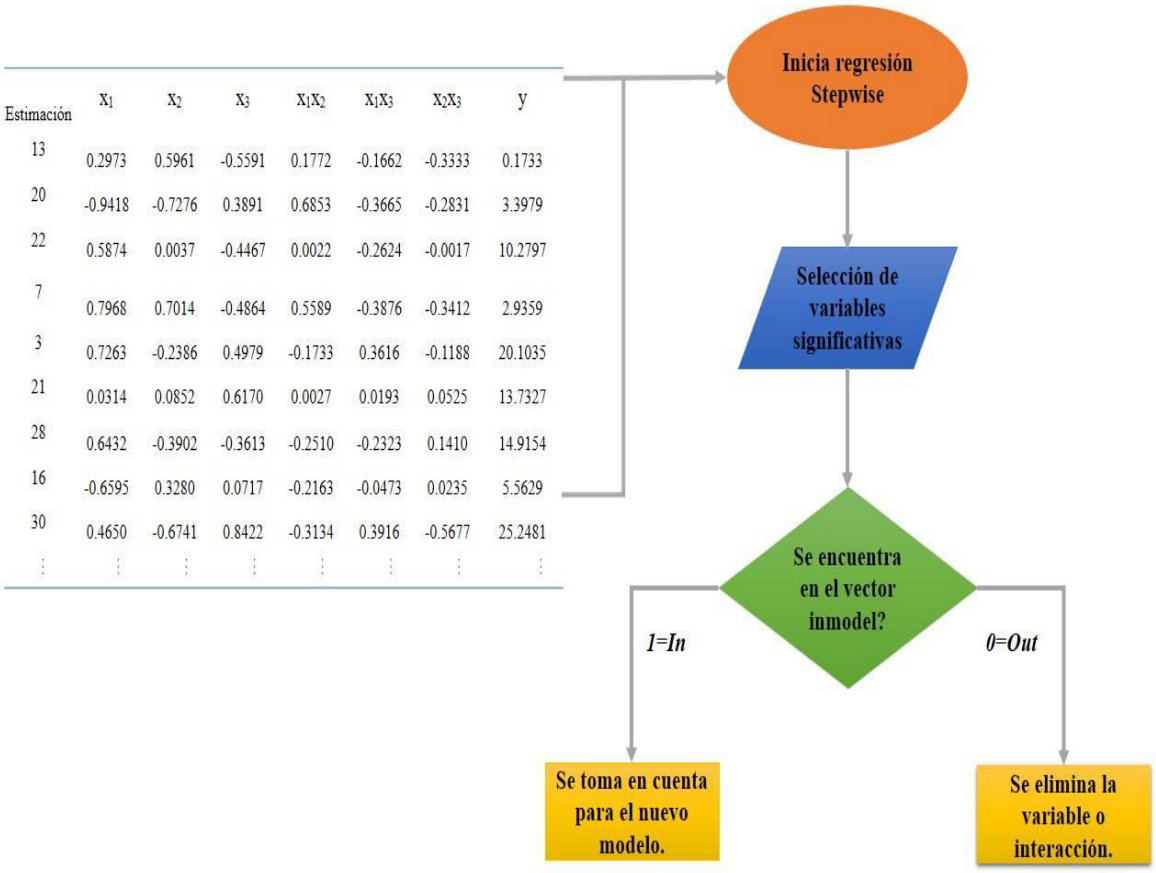

Figura 3.6 Proceso de *Stepwise* con *inmodel* (Fuente: Elaboración propia)

Al terminar la regresión Stepwise se obtienen los valores para el nuevo modelo, siguiendo la ecuación 3.1.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon \qquad (3.1)$$

Una vez obtenido el nuevo modelo con los factores significativos que lo componen, así como el error, el cual permitirá determinar el estadístico PRESS que ayude a explicar la variabilidad de las nuevas observaciones (R^2). El valor en el PRESS deberá ser el mínimo encontrado dentro de los cuatro porcentajes de separación analizados (10%, 20%, 30%, 40%) y máximo para el valor de la R^2.

Mediante el software obtenemos el valor de PRESS, proveniente de los e^2 como se muestra en la siguiente tabla 3.7:

Tabla 3.7 Matriz de los Datos de predicción para determinarlos errores (PRESS)

Predicción	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
2	0.3382	0.0954	0.2181	0.0323	0.0738	0.0208	11.5374	13.5998696	-2.0624	4.25366087
12	-0.7343	0.5170	0.1305	-0.3796	-0.0958	0.0675	7.3984	4.50780264	2.8906	8.35557077
18	-0.1376	-0.0487	0.5704	0.0067	-0.0785	-0.0278	11.3900	12.3051087	-0.9151	0.83748548
										13.4467171

Con los valores de PRESS se calculó un estadístico importante para analizar cuánto cabe esperar que el modelo explique la variabilidad de las nuevas observaciones ($R^2_{predicción}$), con la ecuación vista en el capítulo 2:

$$R^2_{predicción} = 1 - \frac{PRESS}{SS_T} \tag{2.18}$$

El valor de SS_T se obtiene mediante la programación efectuada en MATLAB (Figura 3.7).

Name	Value	Min	Max
se	[0.5655;0.4503;0.6016;0.9868;0.7162;1.1522]	0.4503	1.1522
ss	1	1	1
SSTotalcorregidaprediccion	11.0286	11.0286	11.0286
stats	*1x1 struct*		
sumatoriayobservadaalcuadrado	919.6541	919.65...	919.65...
tamanodeinformacion	30	30	30
y	6	6	6

Figura 3.7 Valor de SS_T obtenido en MATLAB.

Para visualizar de una manera más completa el comportamiento del PRESS (figura 3.9) y $R^2_{predicción}$ (figura 3.10), se utilizará un histograma para observar la acumulación o tendencia de los datos, la variabilidad y la distribución que siguen, como se muestra en la figura 3.8:

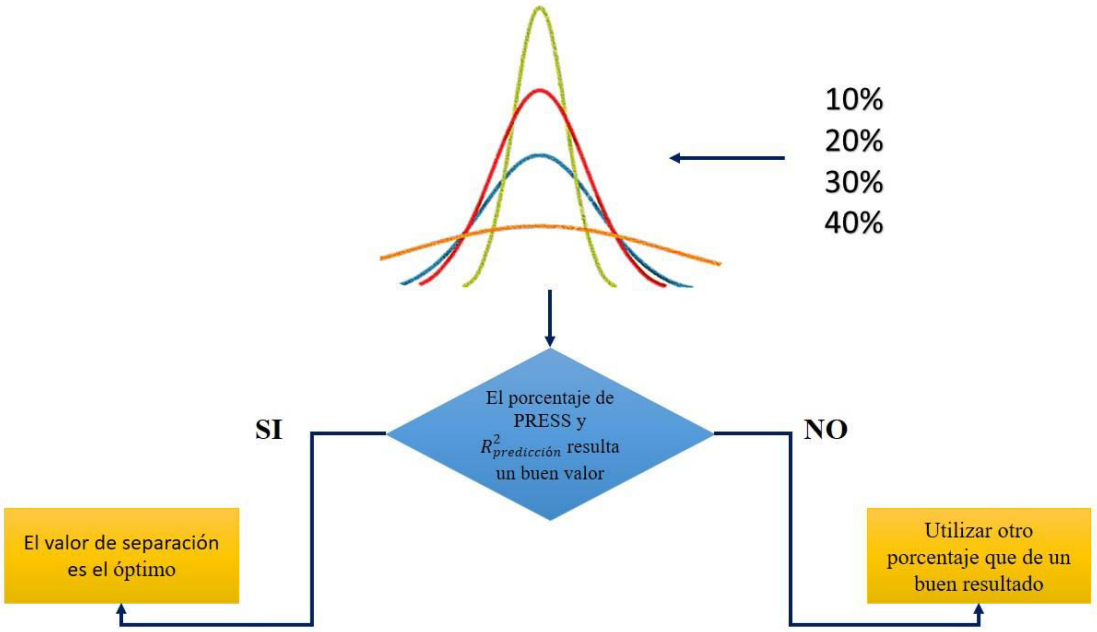

Figura 3.8 Esquematización del histograma a observar para PRESS y $R^2_{predicción}$ (Fuente: Elaboración propia)

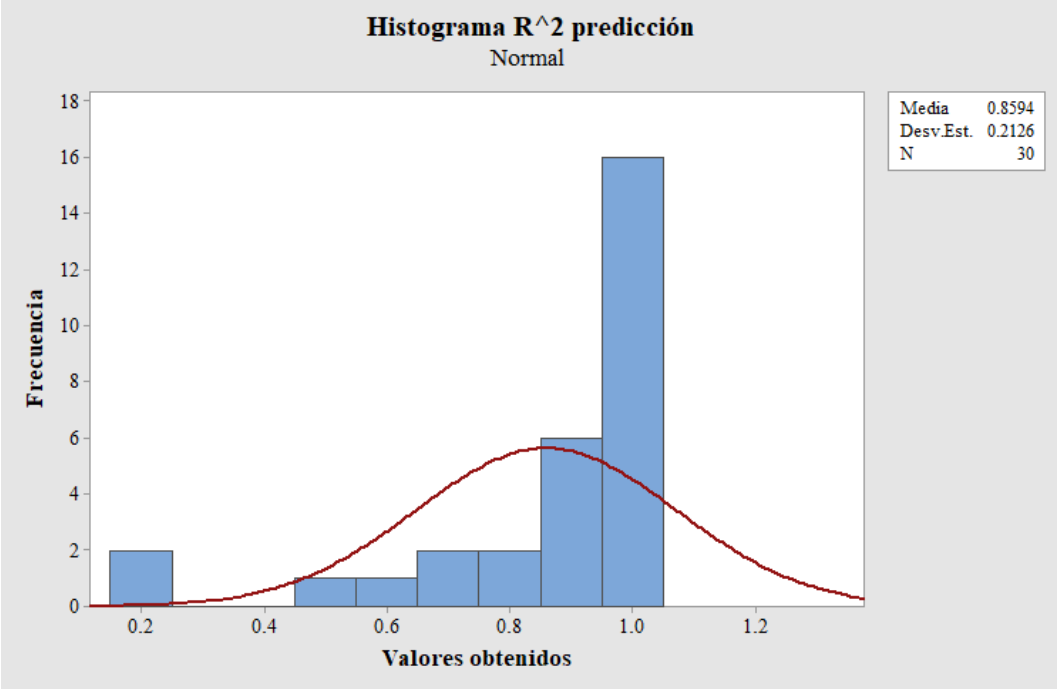

Figura 3.9 Gráfica de $R^2_{predicción}$ en 30 datos a nivel de ruido bajo con 10% de separación.

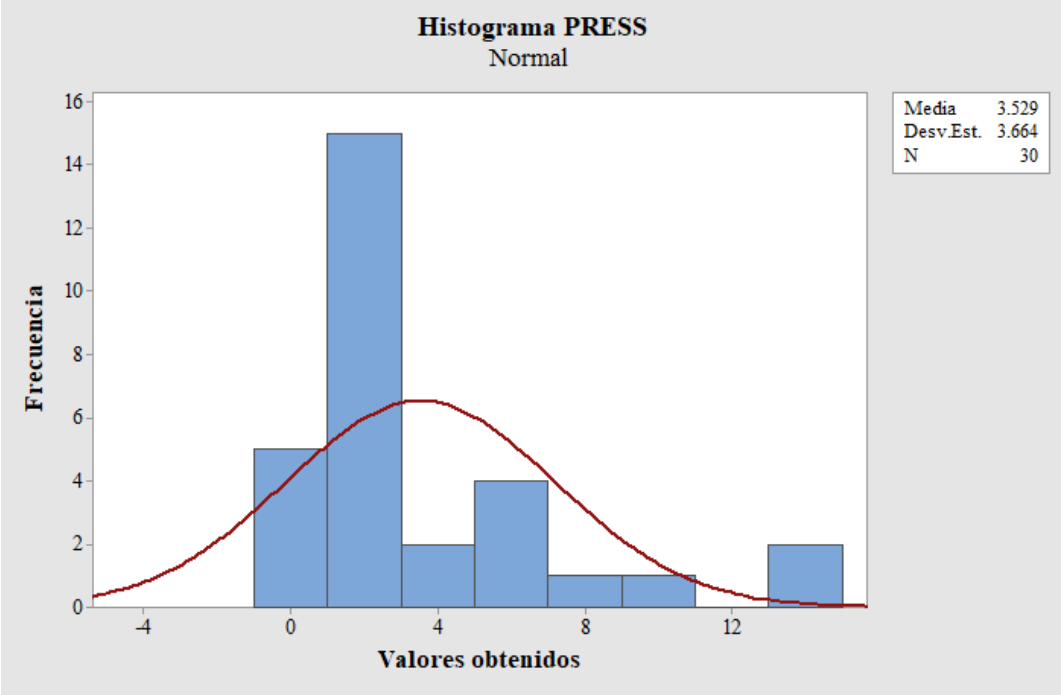

Figura 3.10 Gráfica de PRESS en 30 datos a nivel de ruido bajo con 10% de separación.

CAPÍTULO IV. RESULTADOS

En el presente capítulo se dan a conocer los resultados de la experimentación realizada con el objetivo de encontrar el número óptimo de datos a separar para llevar a cabo el método de Validación Cruzada aplicado en un modelo de regresión lineal múltiple y de igual manera verificar la capacidad predictiva del modelo bajo ciertos argumentos estadísticos como un valor bajo de PRESS (*Predicción Error Sum of Squares)* y el valor más alto de $R^2_{\text{predicción}}$, el cual muestra cuánto cabe esperar que el modelo explique la variabilidad de las nuevas observaciones.

En el capítulo anterior se especifican los pasos a seguirse a lo largo de la investigación, así como los escenarios planteados, como lo muestra la figura 4.1:

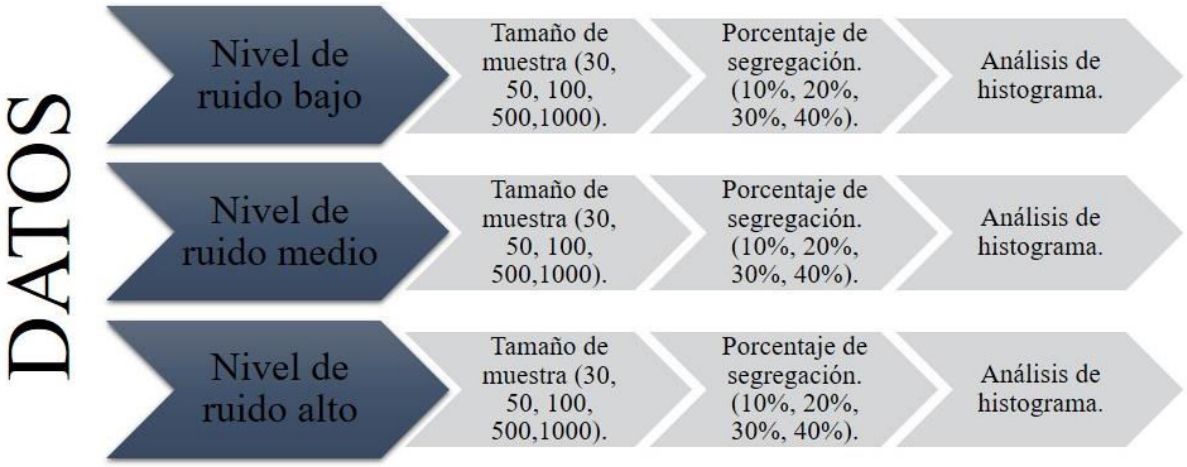

Figura 4.2 Escenarios planteados para la investigación (Elaboración propia).

En la parte superior derecha de los histogramas se muestran los colores asignados (líneas dentro de la gráfica) a los diferentes porcentajes de división de datos de predicción; para una separación de los datos de predicción del 10% , 20%,30% y 40%. De igual manera debajo de la información anterior, se observa los resultados promediados de la media y desviación estándar para las réplicas según la *n* para cada porcentaje de división.

4.1 Nivel de ruido bajo

Modelo verdadero (4.1):

$$y = 10 + 6x_1 - 5x_2 + 7x_3 - 6.8x_1x_2 + 6.3x_1x_3 + 0.6x_2x_3 + \varepsilon(0,2.1) \qquad (4.1)$$

4.1.1 Tamaño de muestra (30)

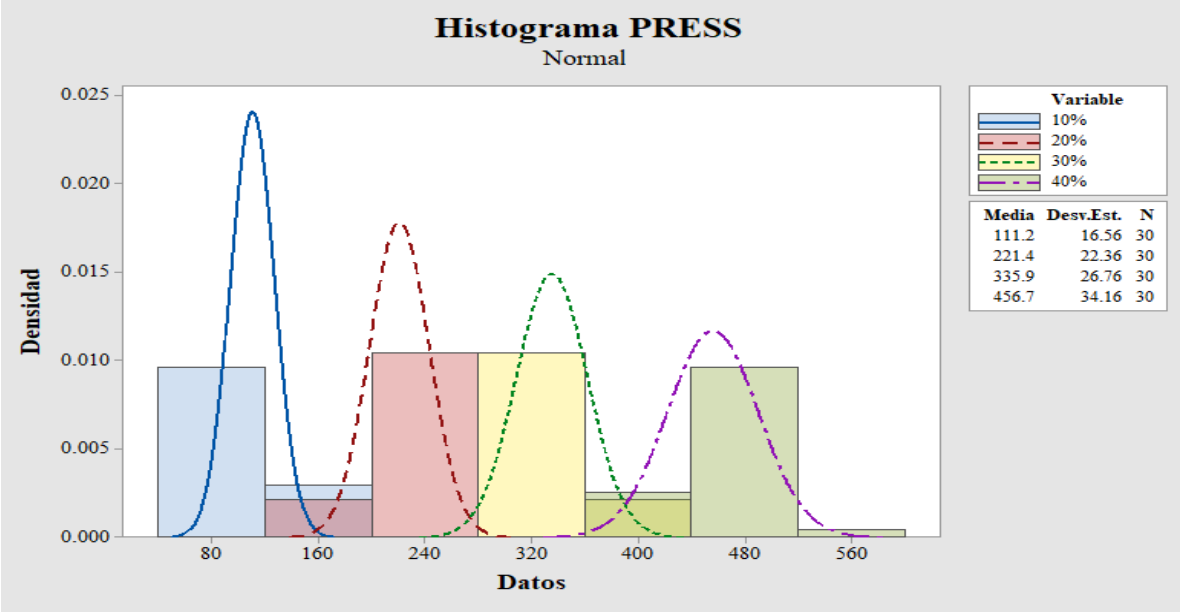

Figura 4.2 Histograma PRESS n=30, nivel de ruido bajo (Minitab).

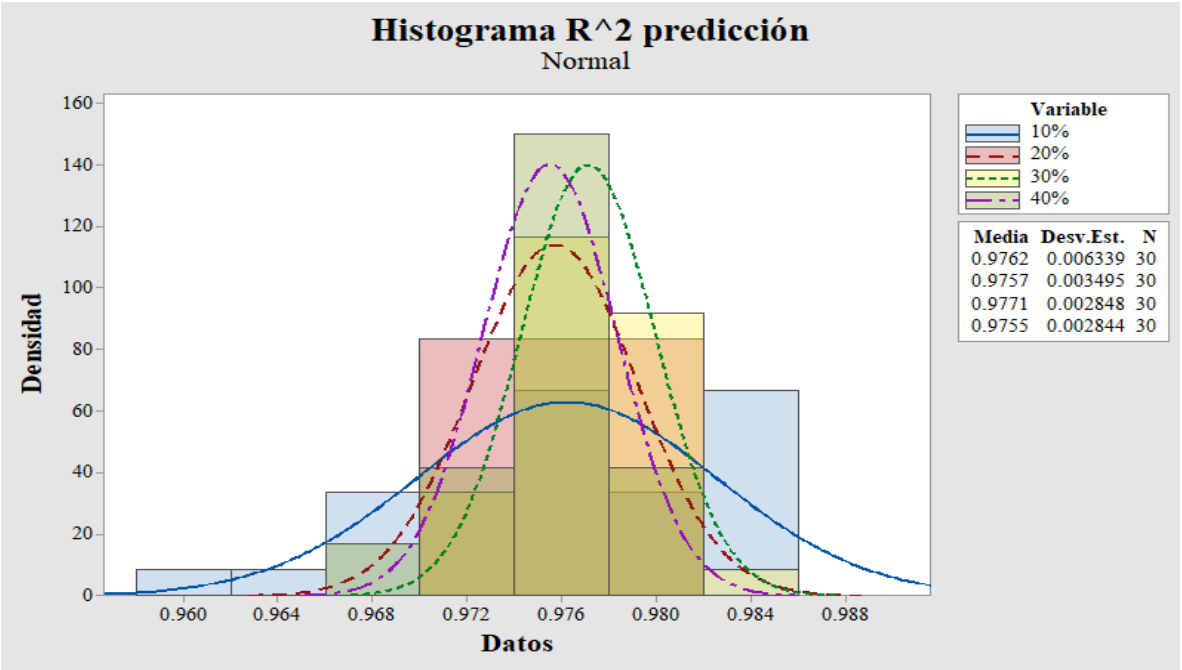

Figura 4.3 Histograma $R^2_{predicción}$ n=30, nivel de ruido bajo (Minitab).

4.1.2 Tamaño de muestra (50)

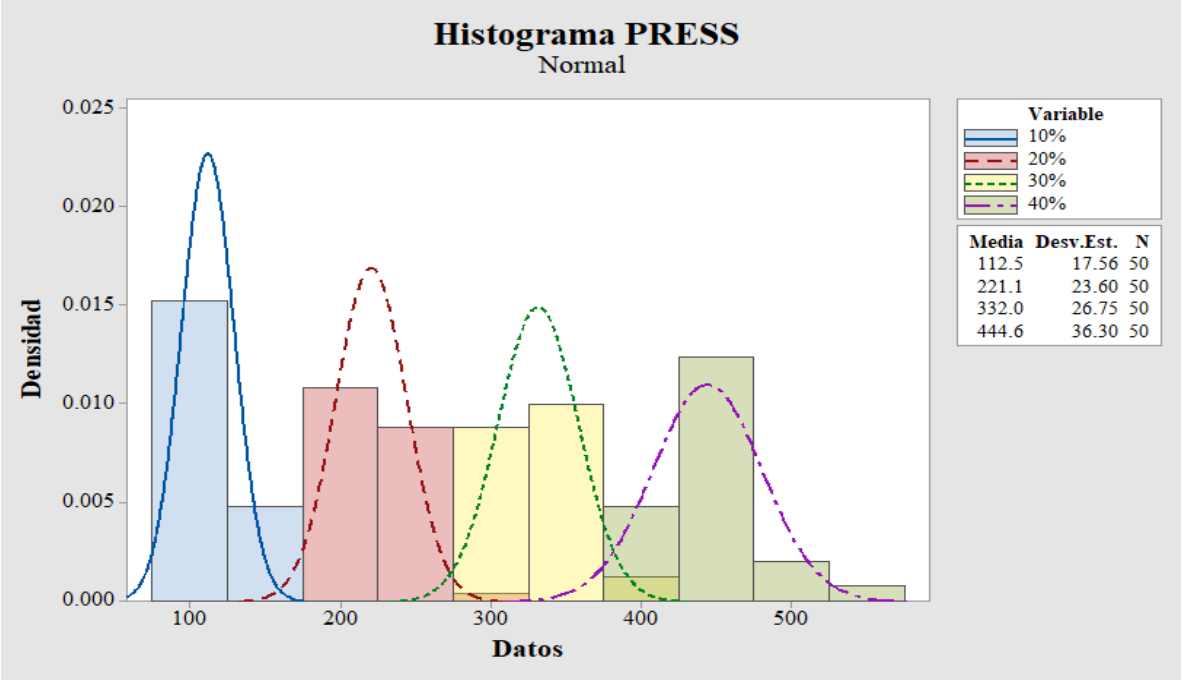

Figura 4.4 Histograma PRESS n=50, nivel de ruido bajo (Minitab).

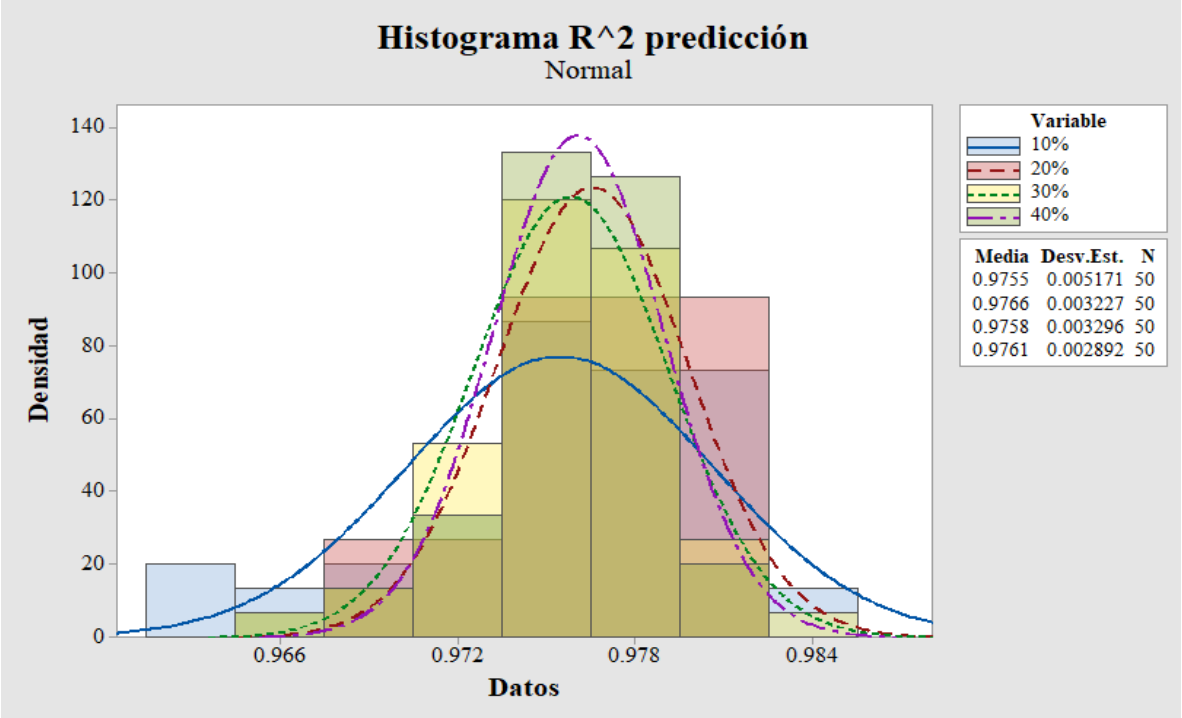

Figura 4.5 Histograma $R^2_{predicción}$ n=50, nivel de ruido bajo (Minitab).

4.1.3 Tamaño de muestra (100)

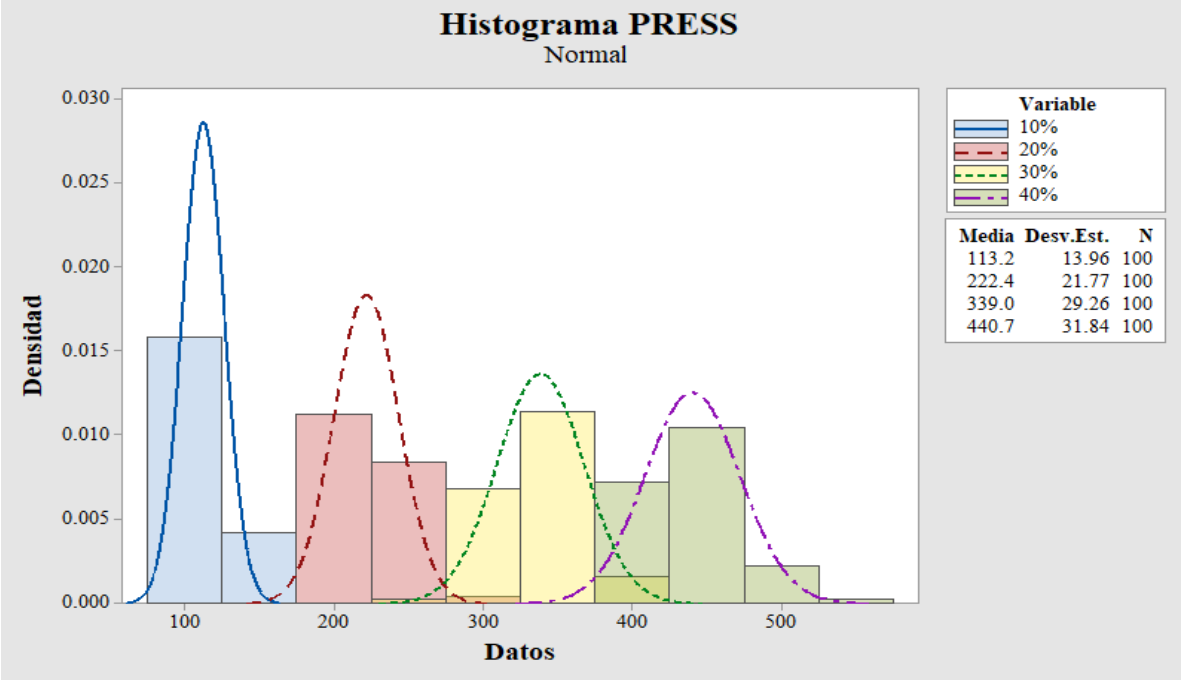

Figura 4.6 Histograma PRESS n=100, nivel de ruido bajo (Minitab).

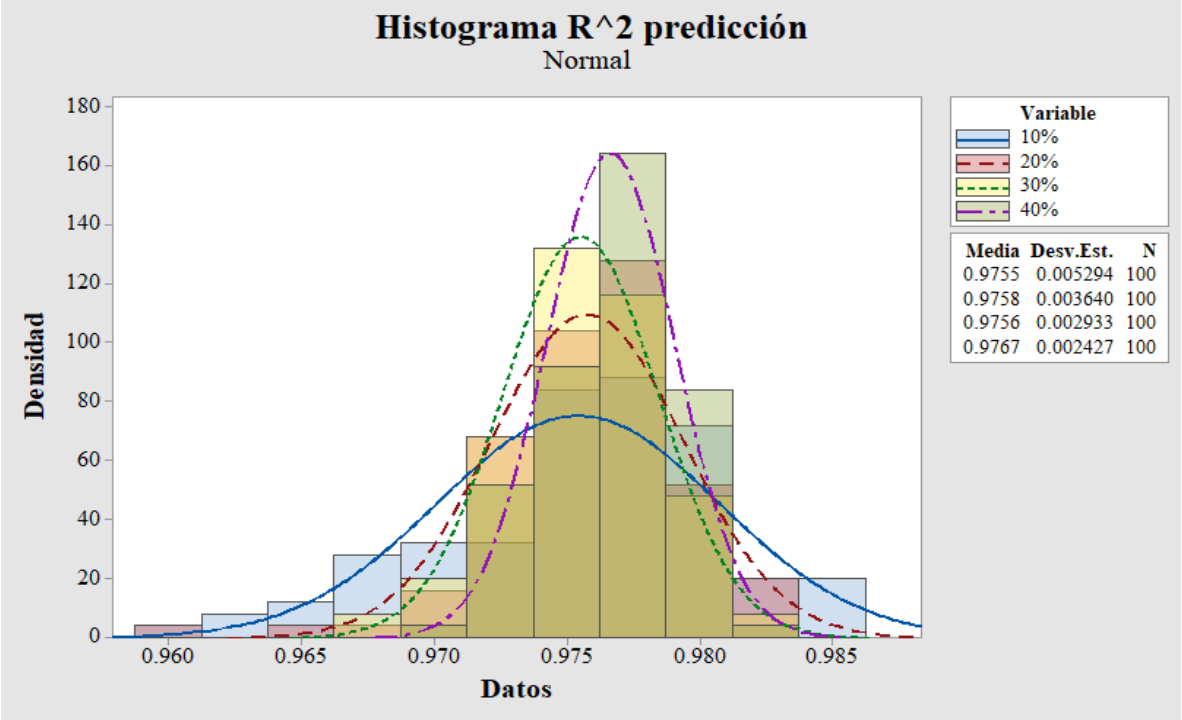

Figura 4.7 Histograma $R^2_{predicción}$ n=100, nivel de ruido bajo (Minitab).

4.1.4 Tamaño de muestra (500)

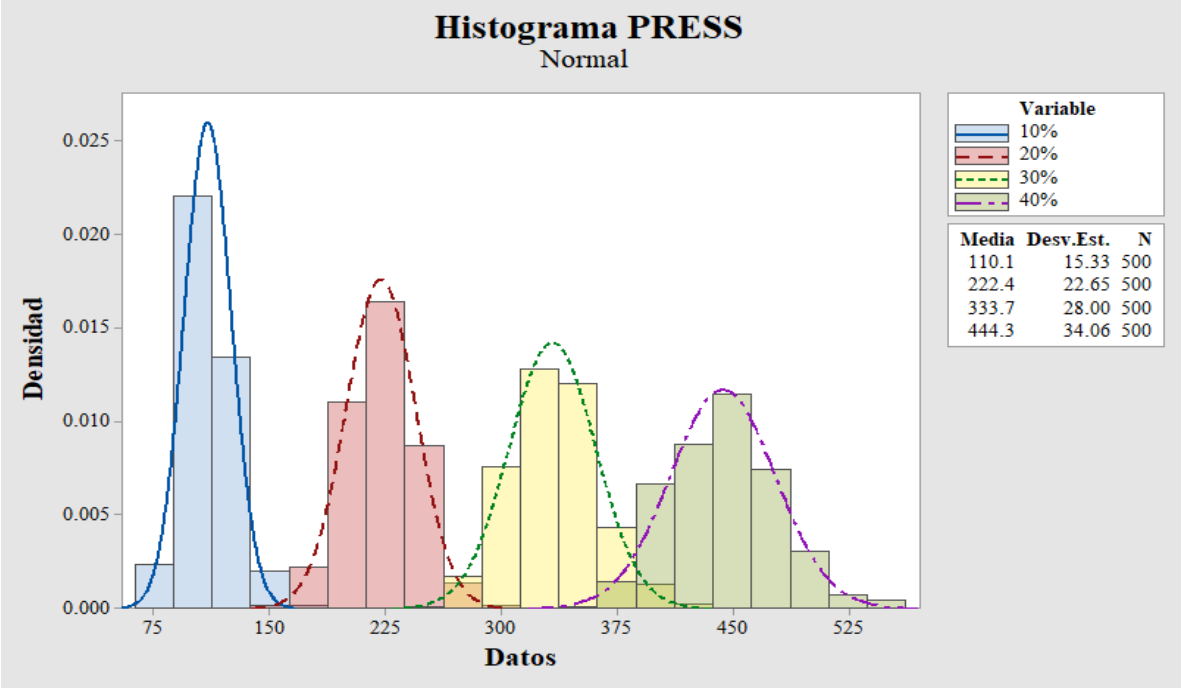

Figura 4.8 Histograma PRESS n=500, nivel de ruido bajo (Minitab).

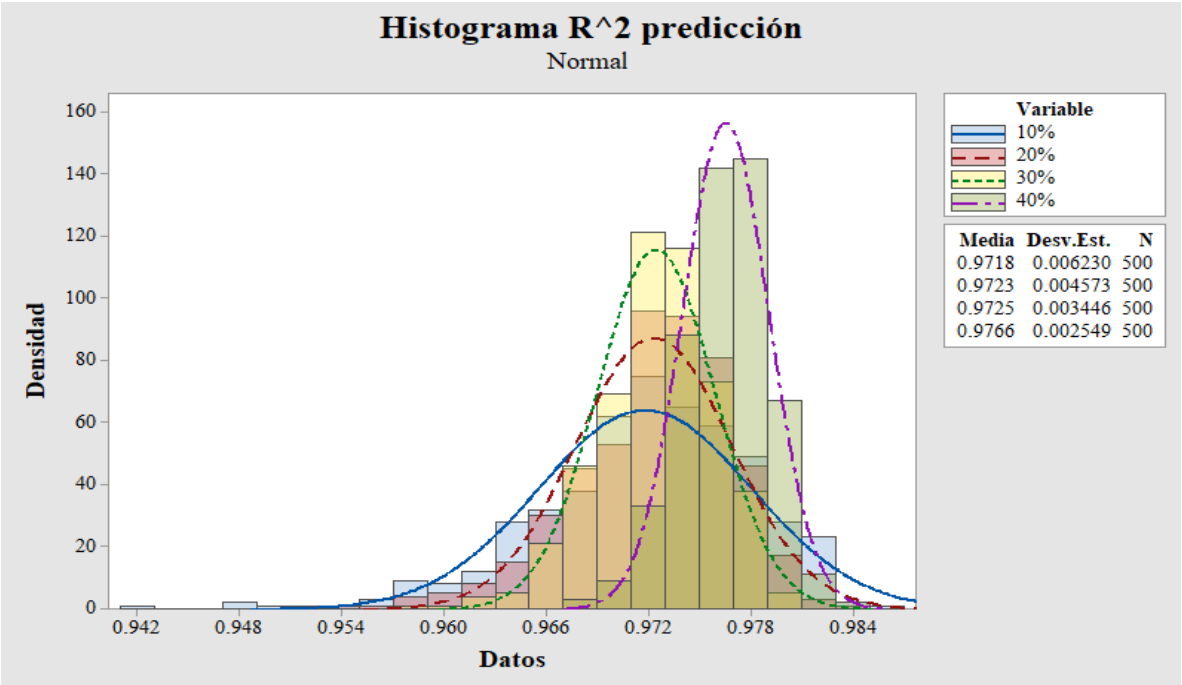

Figura 4.9 Histograma $R^2_{predicción}$ n=500, nivel de ruido bajo (Minitab).

4.1.5 Tamaño de muestra (1000)

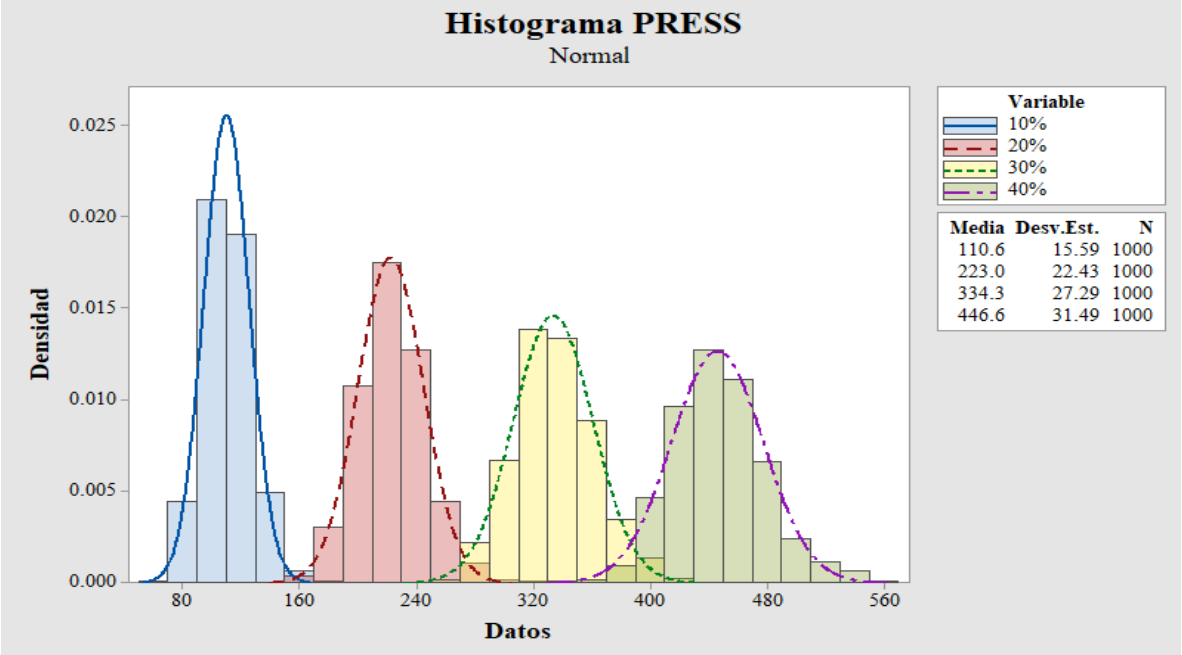

Figura 4.10 Histograma PRESS n=1000, nivel de ruido bajo (Minitab).

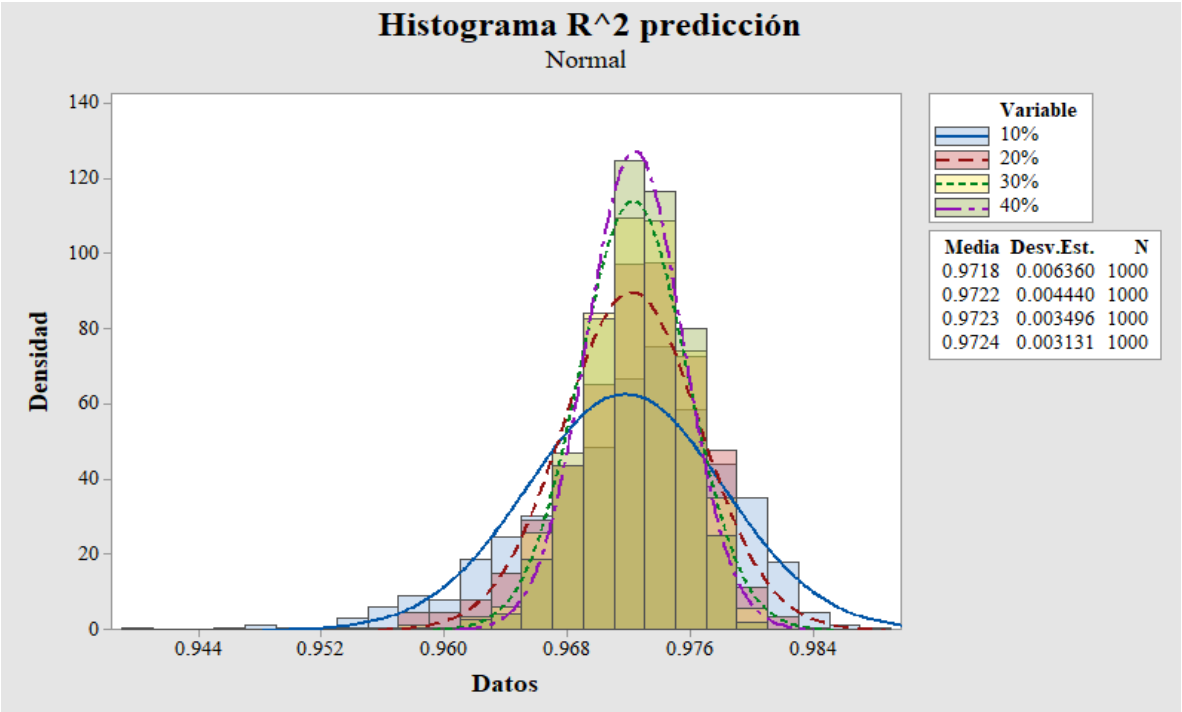

Figura 4.11 Histograma $R^2_{predicción}$ n=1000, nivel de ruido bajo (Minitab).

4.2 Nivel de ruido medio

Modelo verdadero (4.2):

$$y = 6 + 4x_1 - 2x_2 + 5x_3 - 4.8x_1x_2 + 4.3x_1x_3 + 0.9x_2x_3 + \varepsilon(0,2.1) \qquad (4.2)$$

4.2.1 Tamaño de muestra (30)

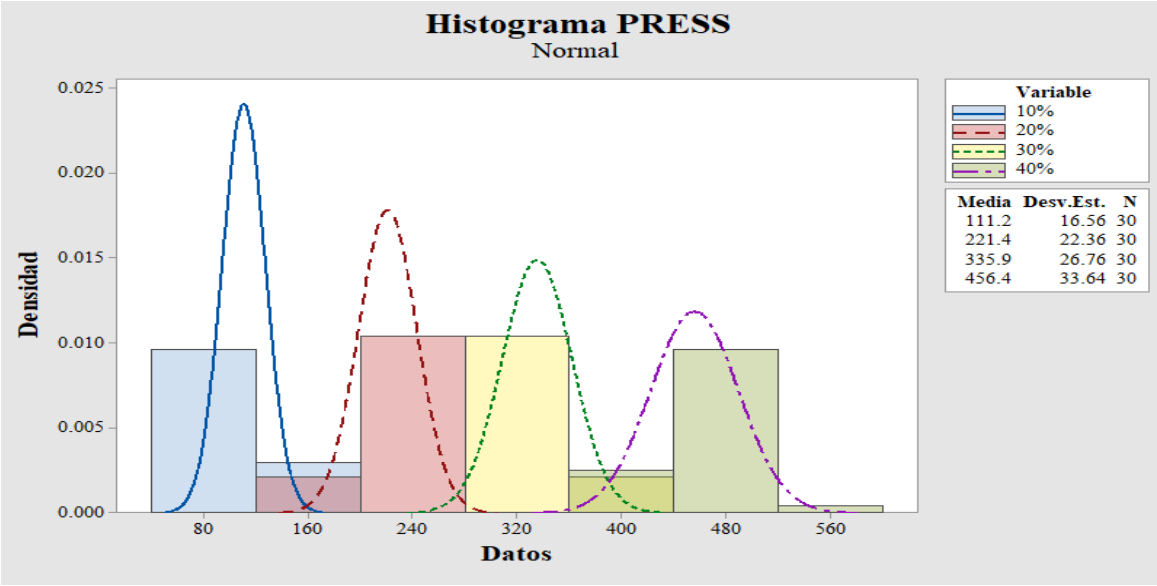

Figura 4.12 Histograma PRESS n=30, nivel de ruido medio (Minitab).

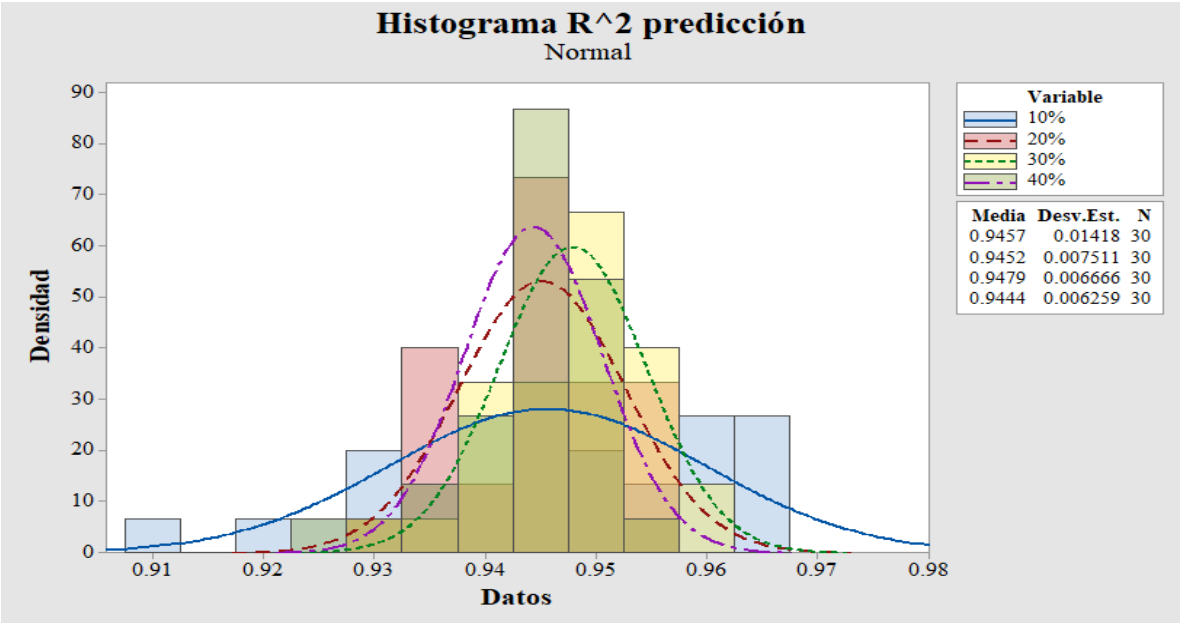

Figura 4.13 Histograma $R^2_{predicción}$ n=30, nivel de ruido medio (Minitab).

4.2.2 Tamaño de muestra (50)

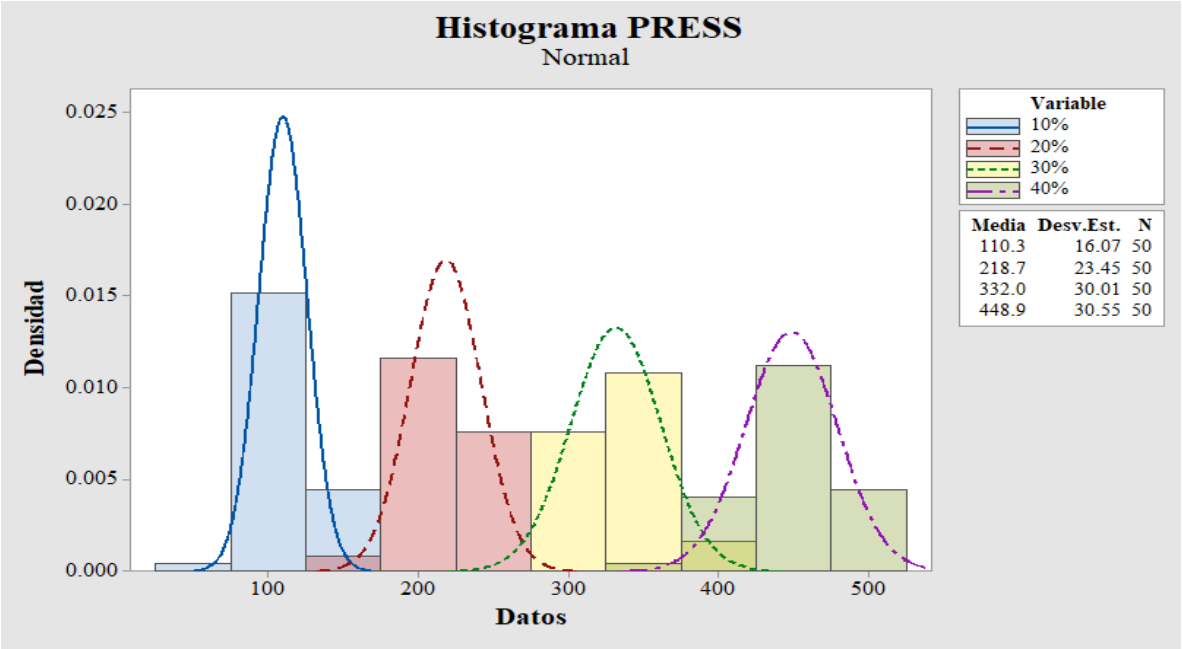

Figura 4.14 Histograma PRESS n=50, nivel de ruido medio (Minitab).

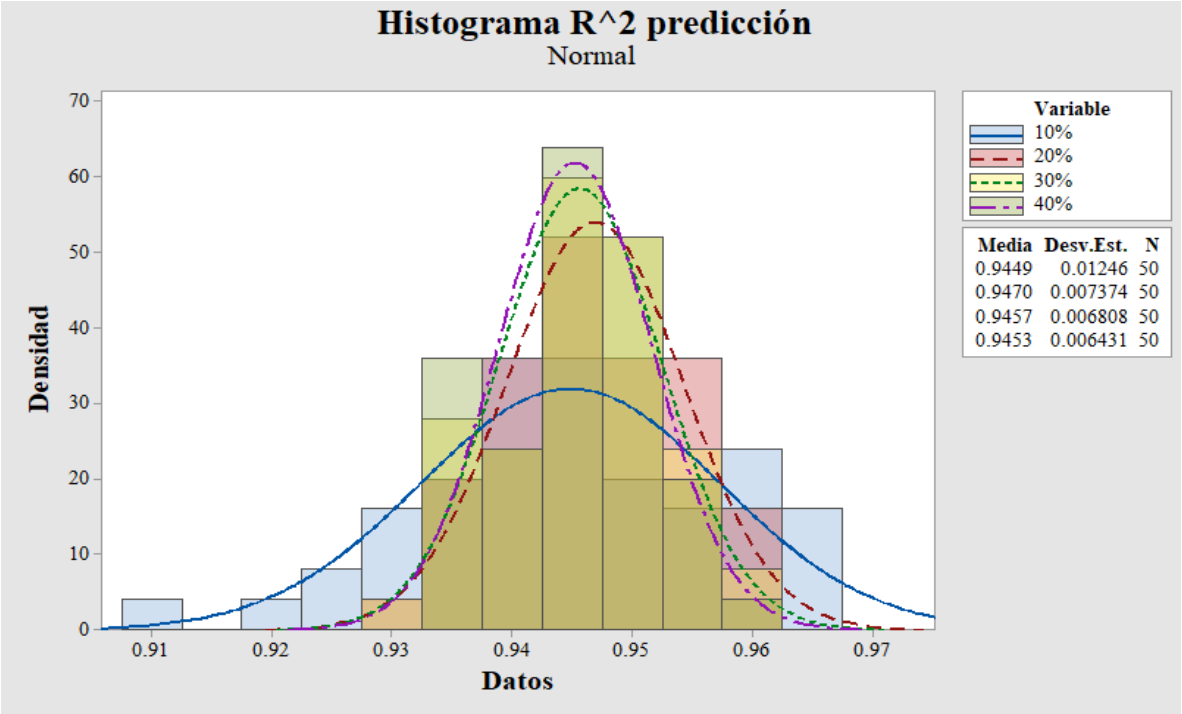

Figura 4.15 Histograma $R^2_{predicción}$ n=50, nivel de ruido medio (Minitab).

4.2.3 Tamaño de muestra (100)

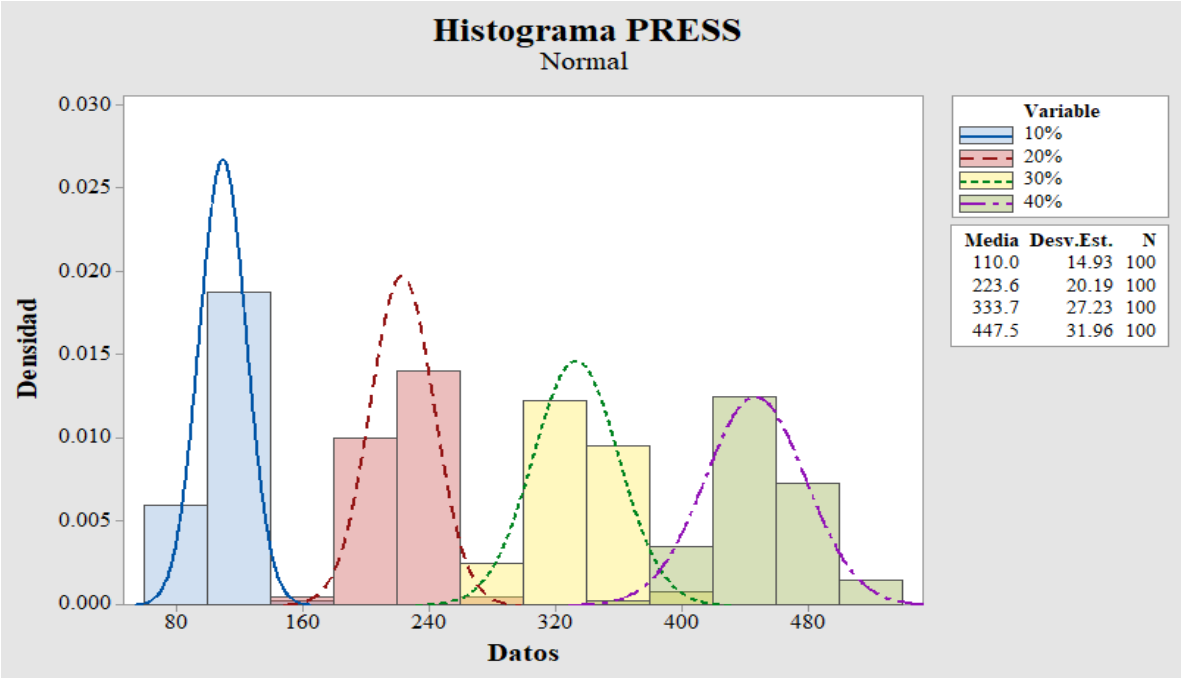

Figura 4.16 Histograma PRESS n=100, nivel de ruido medio (Minitab).

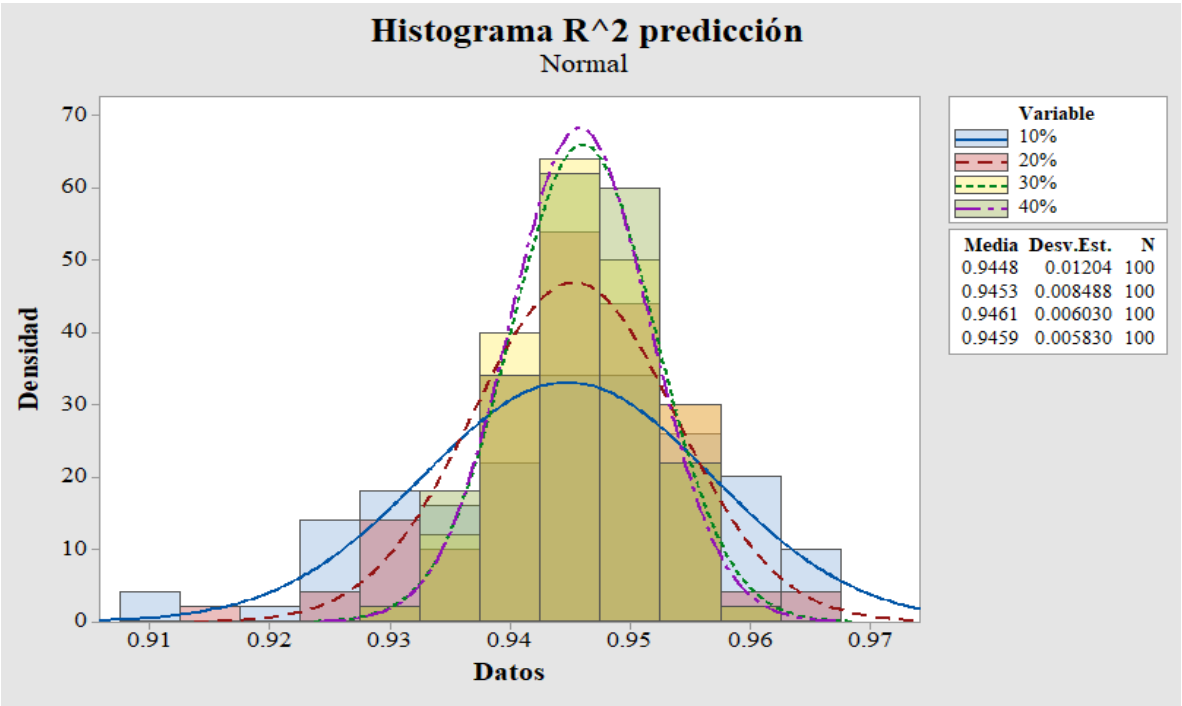

Figura 4.17 Histograma $R^2_{predicción}$ n=100, nivel de ruido medio (Minitab).

4.2.4 Tamaño de muestra (500)

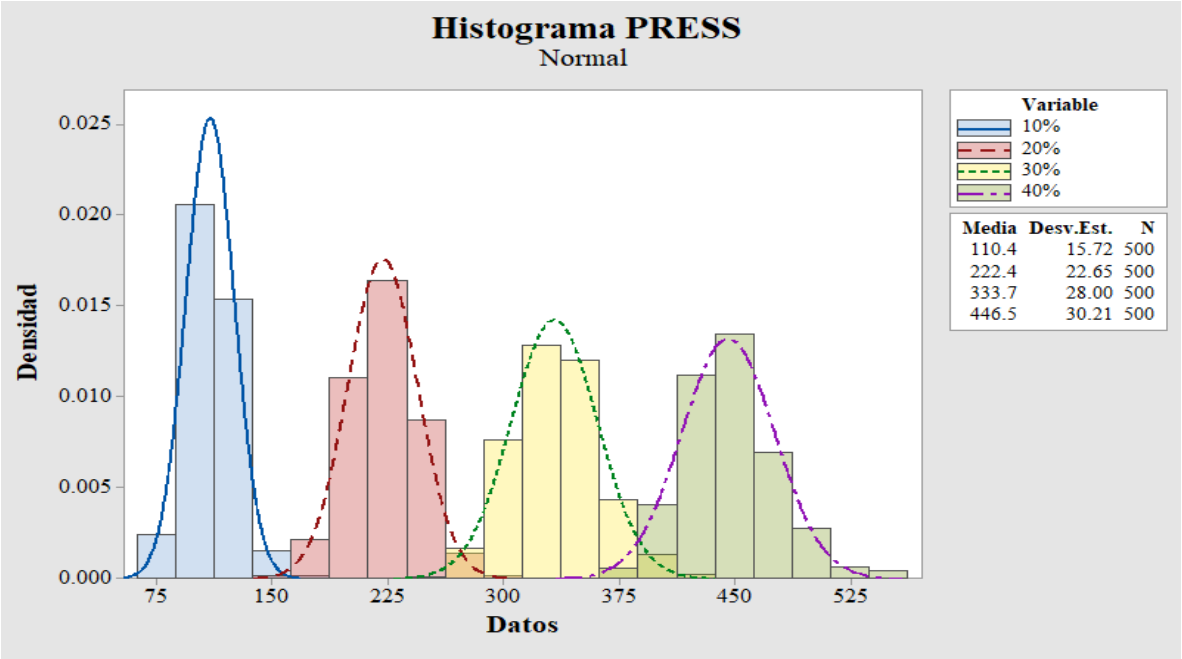

Figura 4.18 Histograma PRESS n=500, nivel de ruido medio (Minitab).

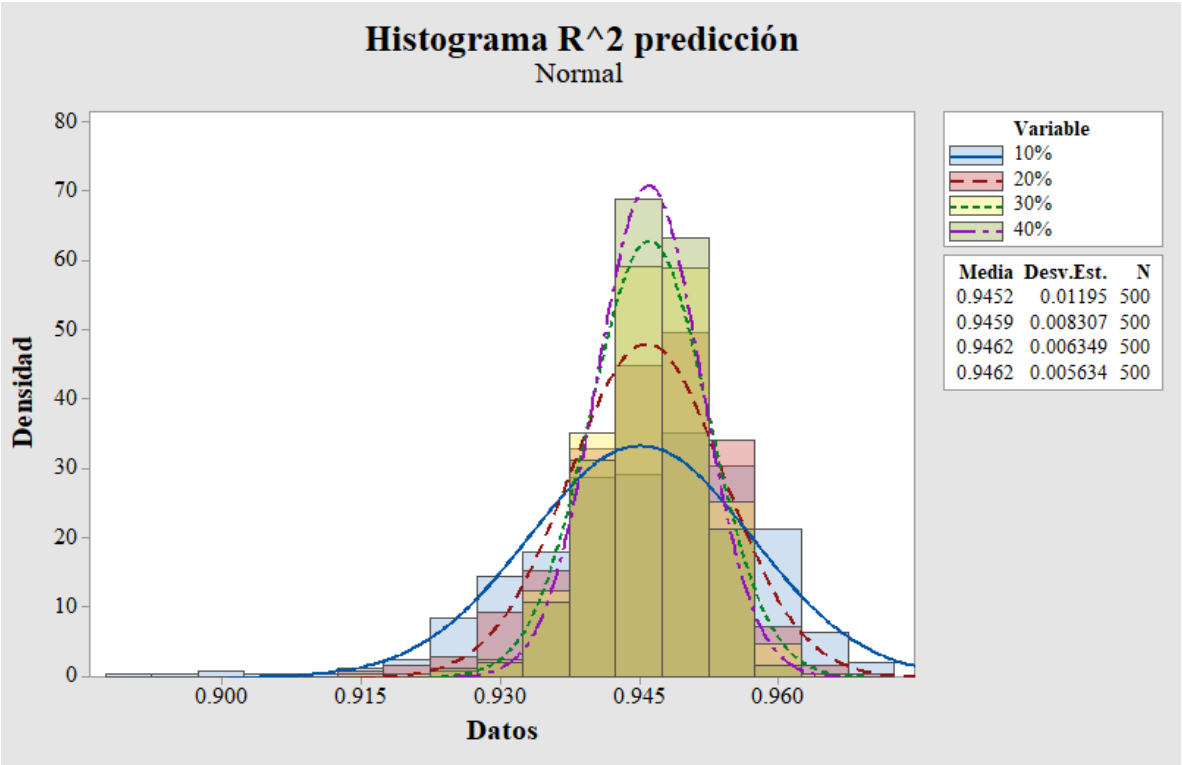

Figura 4.19 Histograma $R^2_{predicción}$ n=500, nivel de ruido medio (Minitab).

4.2.5 Tamaño de muestra (1000)

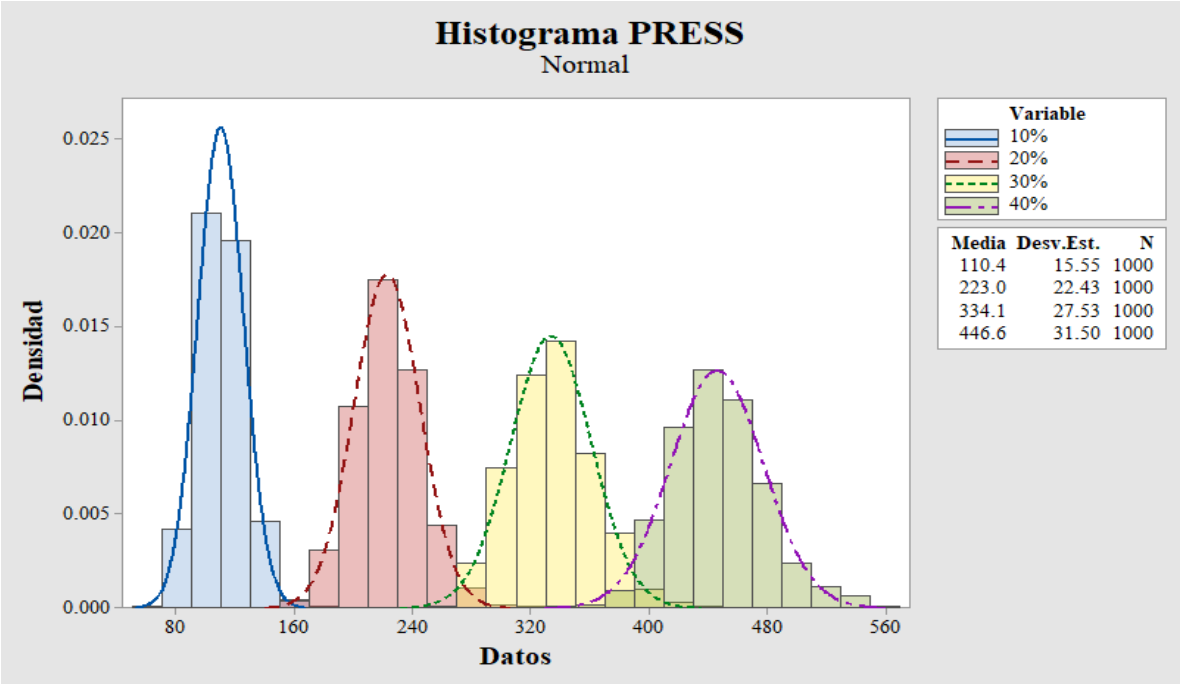

Figura 4.20 Histograma PRESS n=1000, nivel de ruido medio (Minitab).

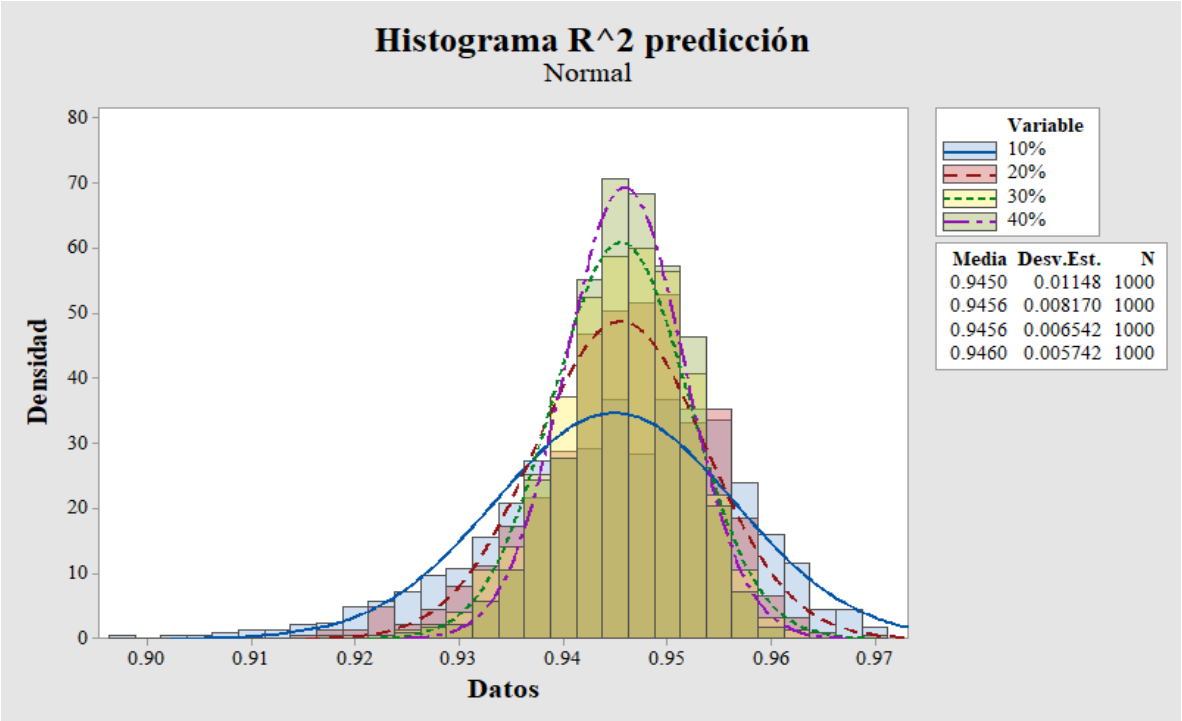

Figura 4.21 Histograma $R^2_{predicción}$ n=1000, nivel de ruido medio (Minitab).

4.3 Nivel de ruido alto

Modelo verdadero:

$$y = 4 + 2x_1 - 1 + 3x_3 - 2.8x_1x_2 + 3.3x_1x_3 + 2.9x_2x_3 + \varepsilon(0,2.1) \qquad (4.3)$$

4.3.1 Tamaño de muestra (30)

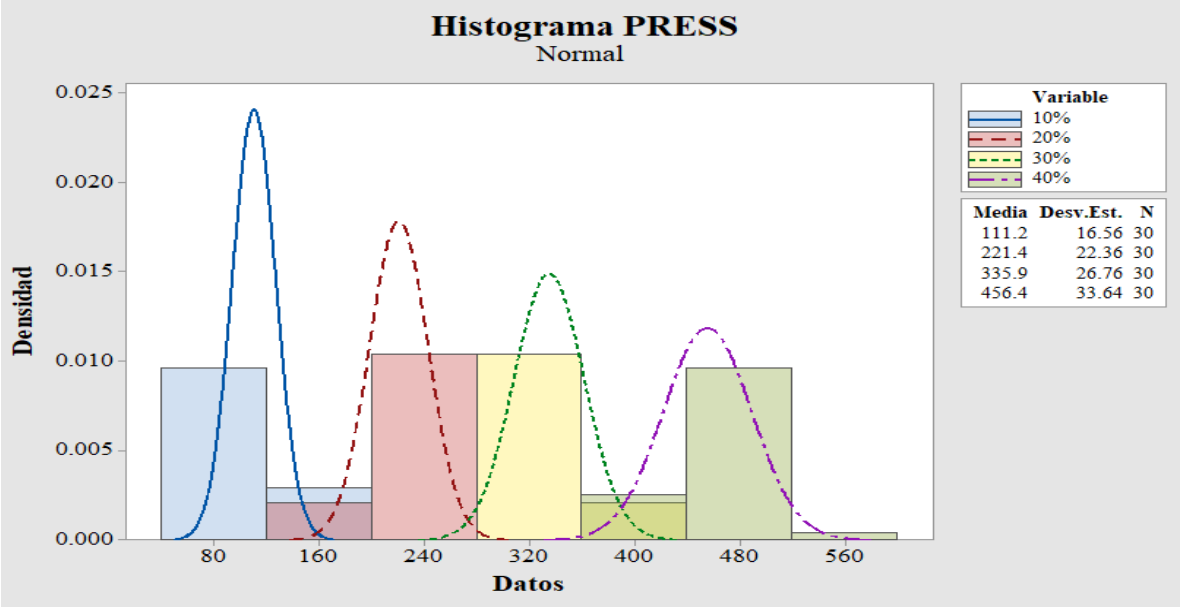

Figura 4.22 Histograma PRESS n=30, nivel de ruido alto (Minitab).

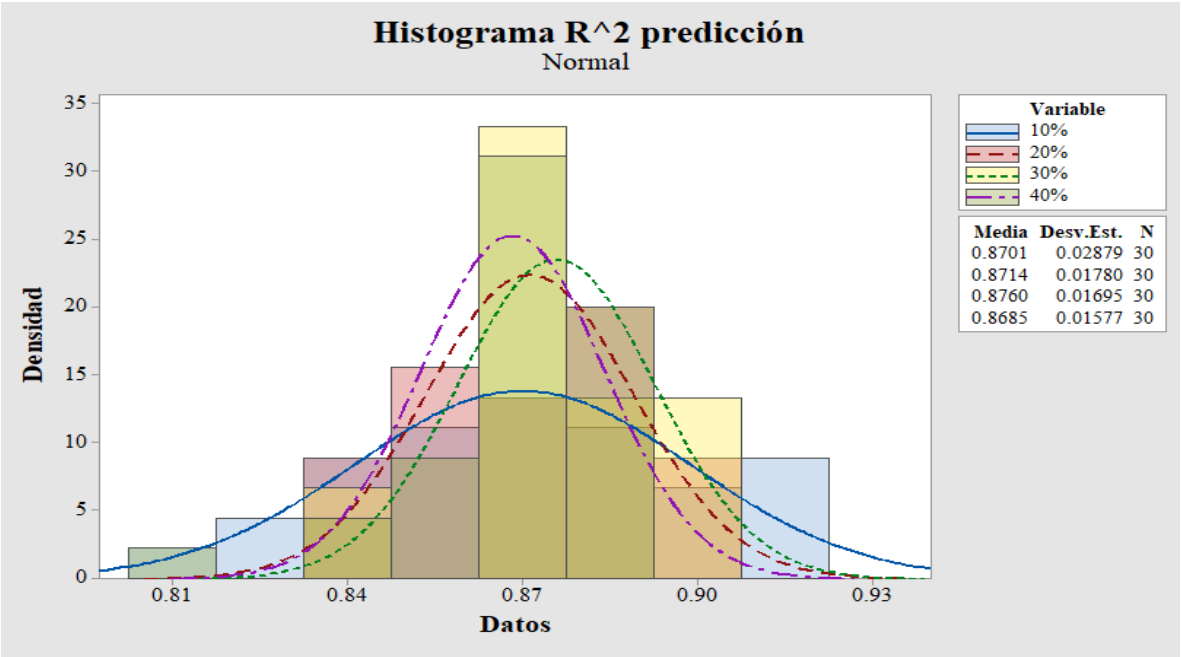

Figura 4.23 Histograma $R^2_{predicción}$ n=30, nivel de ruido alto (Minitab).

4.3.2 Tamaño de muestra (50)

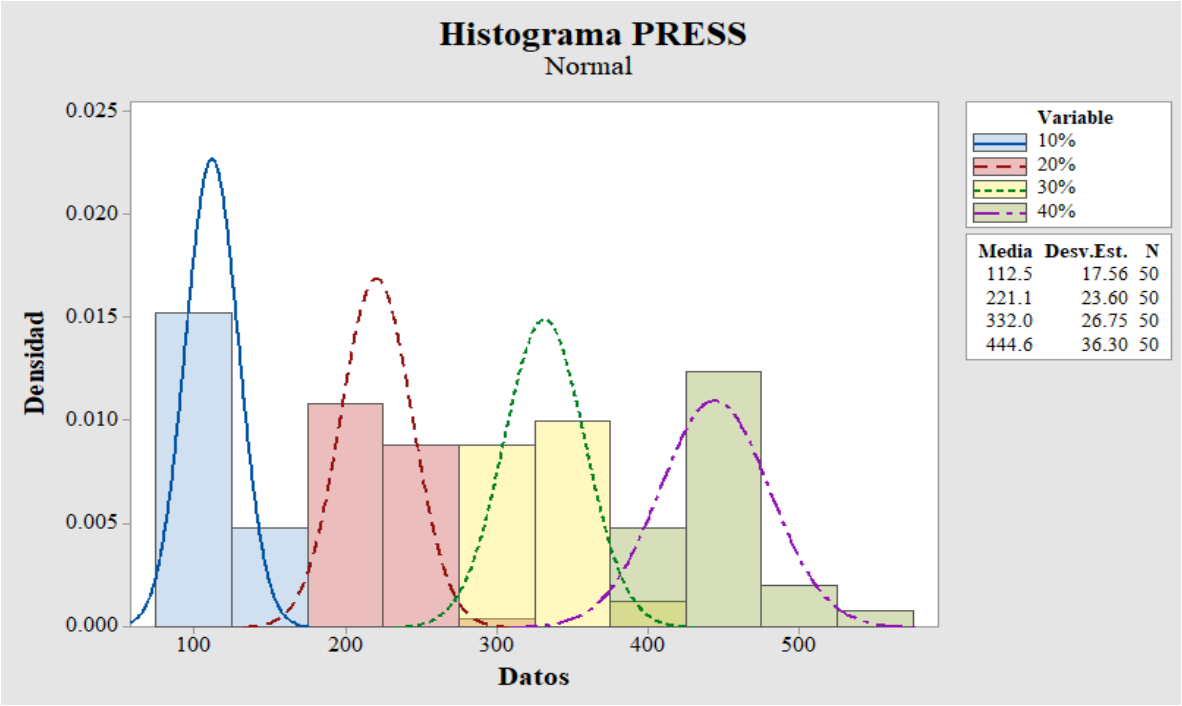

Figura 4.24 Histograma PRESS n=50, nivel de ruido alto (Minitab).

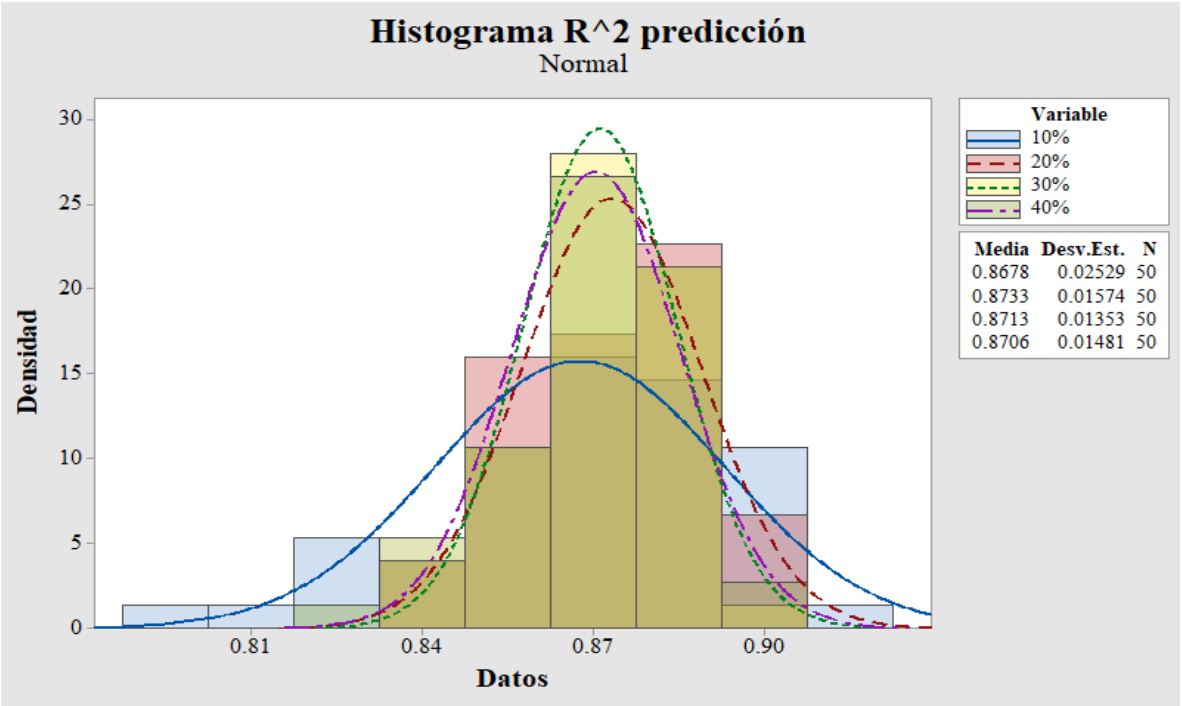

Figura 4.25 Histograma $R^2_{predicción}$ n=50, nivel de ruido alto (Minitab).

4.3.3 Tamaño de muestra (100)

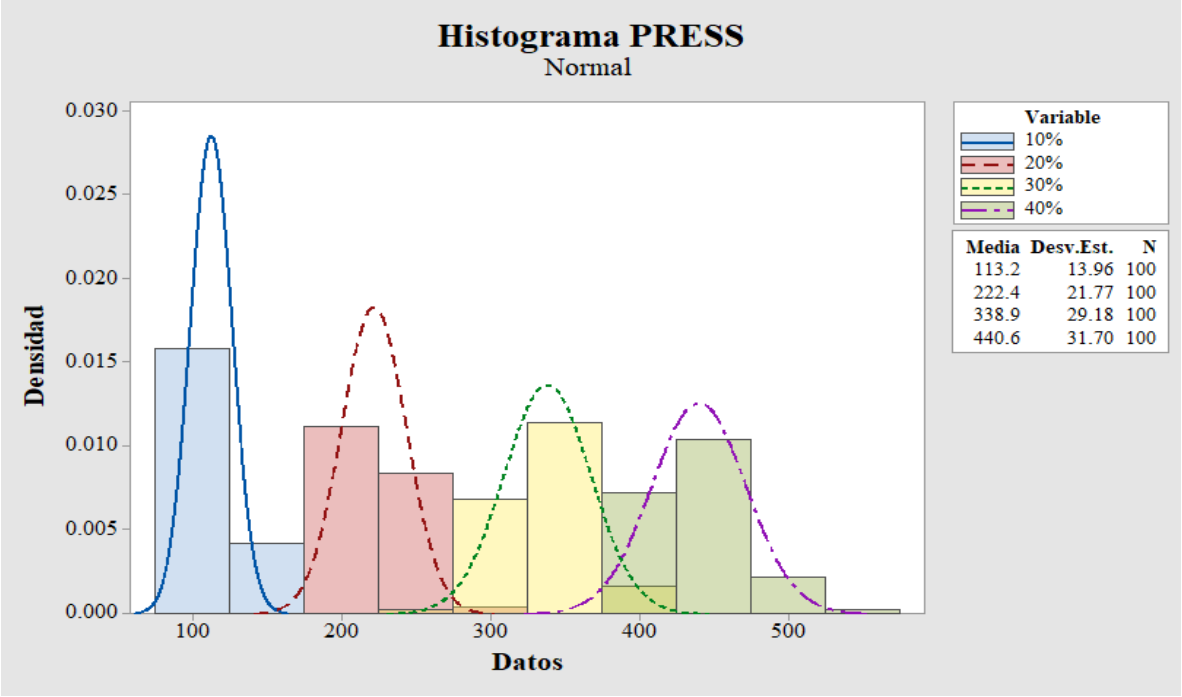

Figura 4.26 Histograma PRESS n=100, nivel de ruido alto (Minitab).

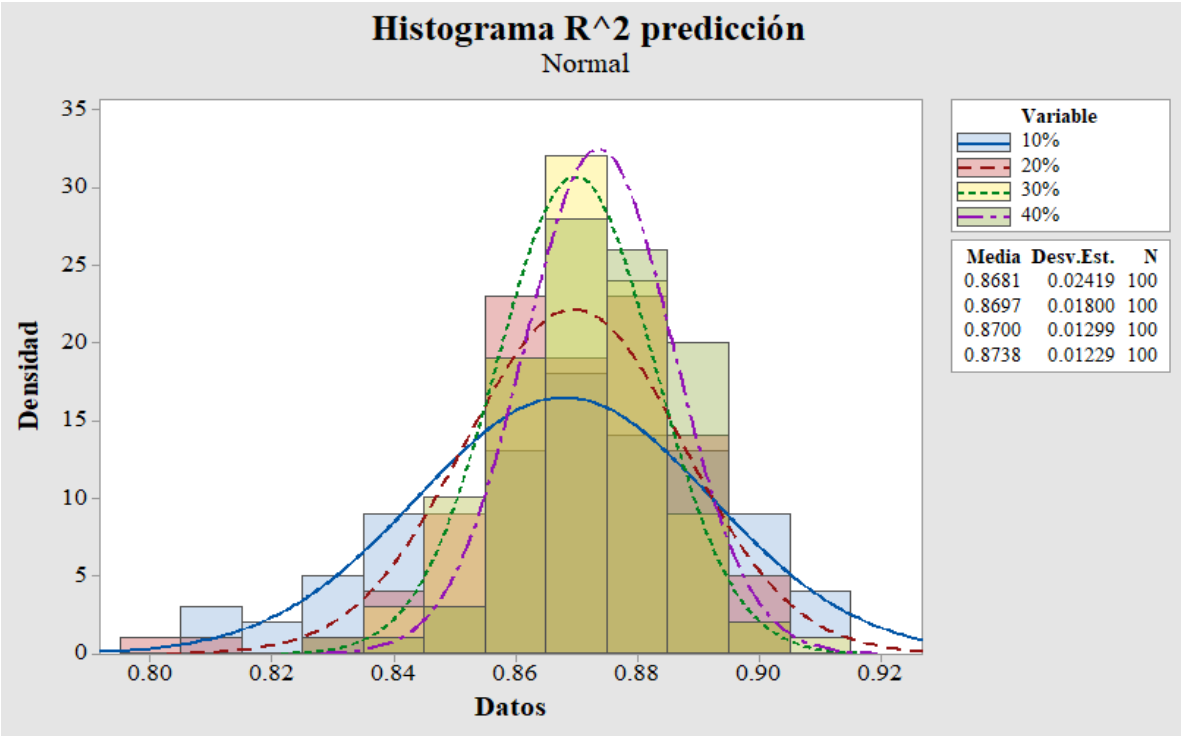

Figura 4.27 Histograma $R^2_{predicción}$ n=100, nivel de ruido alto (Minitab).

4.3.4 Tamaño de muestra (500)

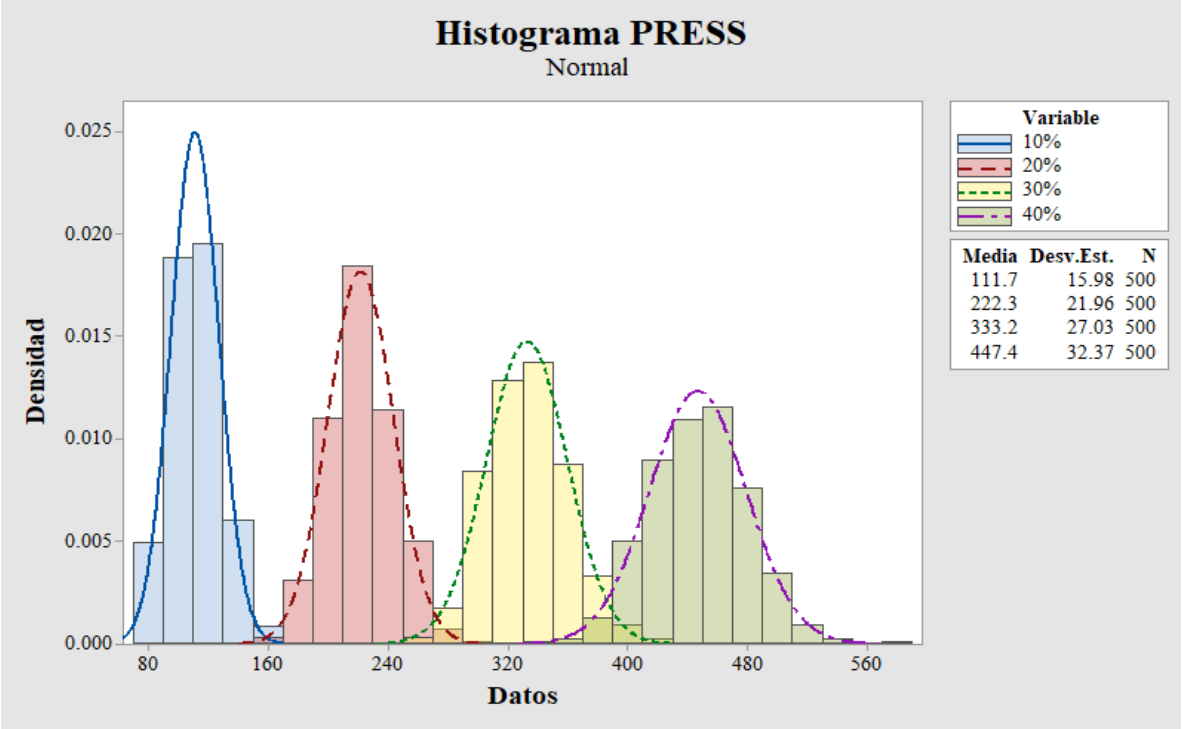

Figura 4.28 Histograma PRESS n=500, nivel de ruido alto (Minitab).

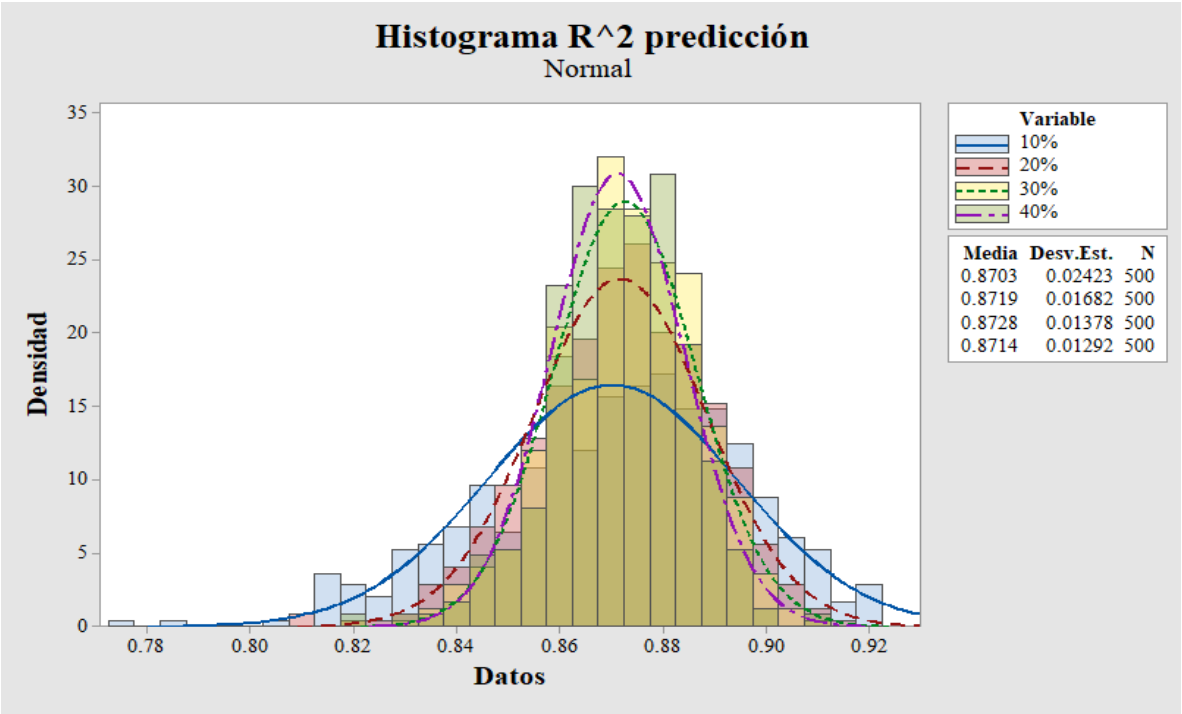

Figura 4.29 Histograma $R^2_{predicción}$ n=500, nivel de ruido alto (Minitab).

4.3.5 Tamaño de muestra (1000)

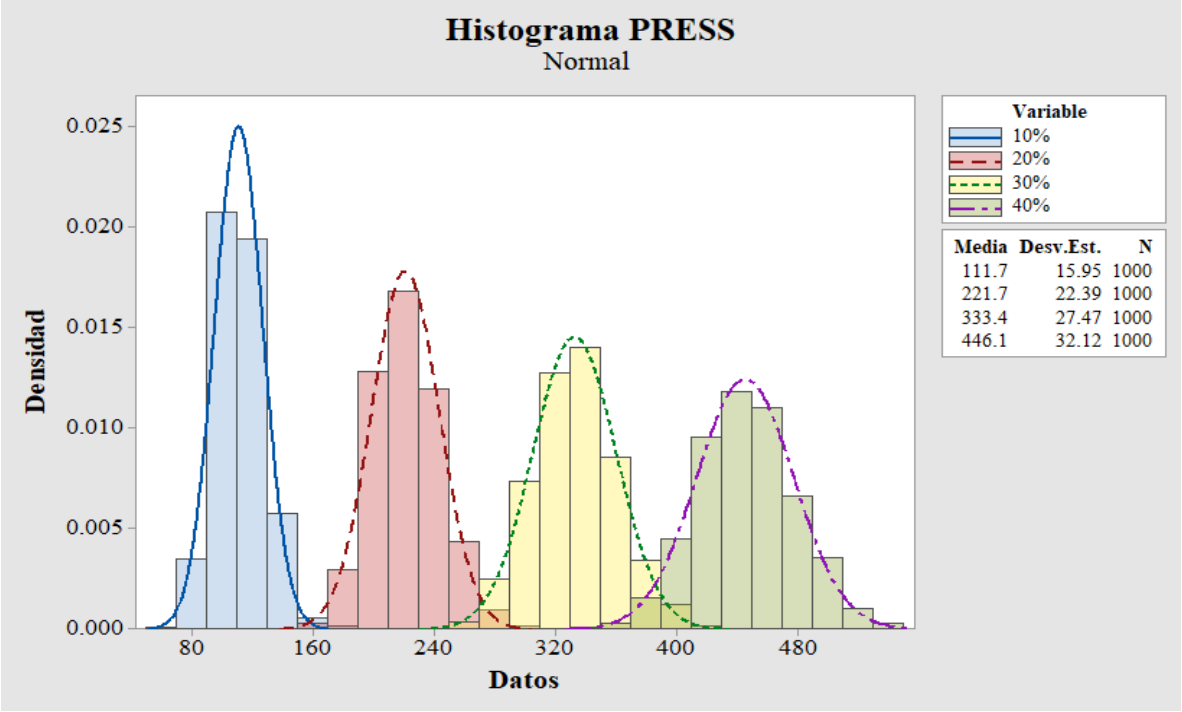

Figura 4.30 Histograma PRESS n=1000, nivel de ruido alto (Minitab).

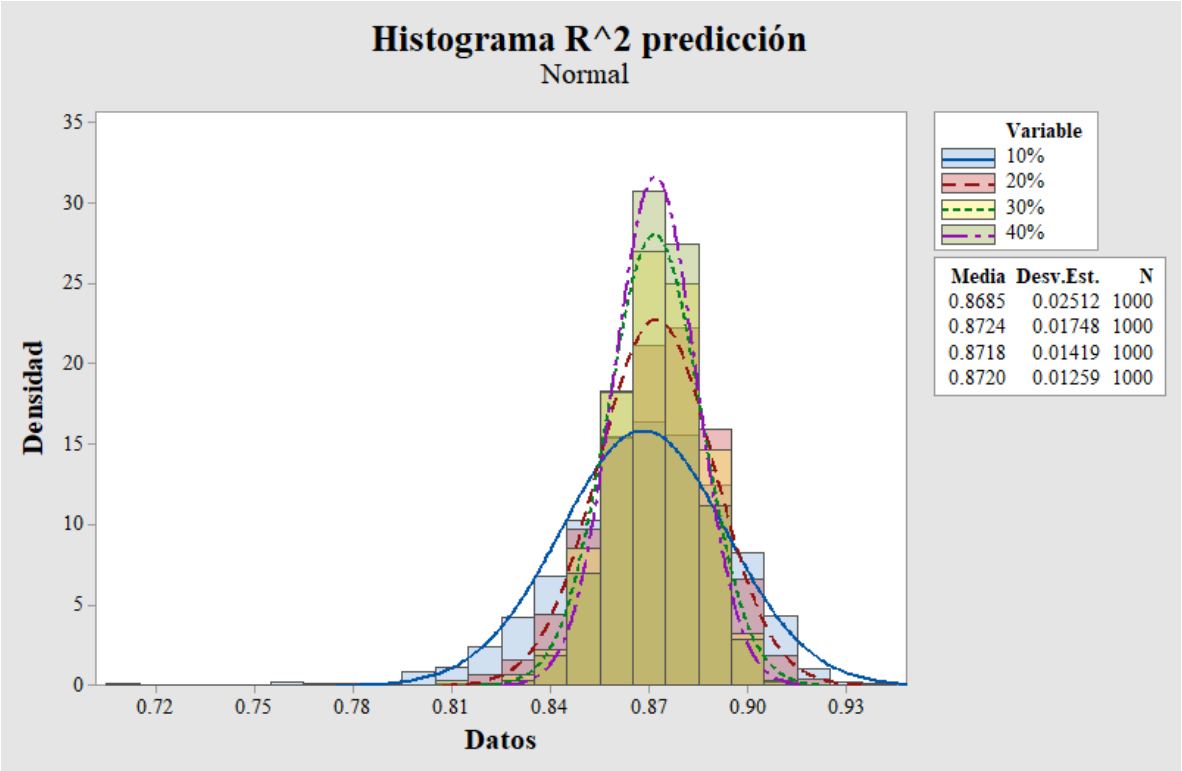

Figura 4.31 Histograma $R^2_{predicción}$ n=1000, nivel de ruido alto (Minitab).

4.4 Resumen

4.4.1 $R^2_{predicción}$

Tabla 8.1 Valores para la $R^2_{predicción}$ en nivel de ruido bajo para los cuatro porcentajes de separación y *n* diferentes.

$R^2_{predicción}$		Tamaño del conjunto de información (nivel de ruido bajo)				
		30	50	100	500	1000
Porcentaje de datos (conjunto de predicción)	10	98.04%	97.89%	96.24%	97.01%	97.34%
	20	97.07%	96.95%	97.38%	97.34%	97.54%
	30	97.93%	97.67%	97.56%	96.61%	97.04%
	40	97.81%	97.81%	97.68%	96.88%	96.84%

Tabla 4.2 Valores para la $R^2_{predicción}$ en nivel de ruido medio para los cuatro porcentajes de separación y *n* diferentes.

$R^2_{predicción}$		Tamaño del conjunto de información (nivel de ruido medio)				
		30	50	100	500	1000
Porcentaje de datos (conjunto de predicción)	10	95.58%	94.64%	94.89%	93.72%	89.76%
	20	93.35%	94.87%	95.21%	94.87%	94.77%
	30	95.41%	93.52%	94.73%	93.52%	94.53%
	40	94.89%	94.92%	94.50%	94.21%	94.02%

Tabla 4.3 Valores para la $R^2_{predicción}$ en nivel de ruido medio para los cuatro porcentajes de separación y *n* diferentes.

$R^2_{predicción}$		Tamaño del conjunto de información (nivel de ruido alto)				
		30	50	100	500	1000
Porcentaje de datos (conjunto de predicción)	10	88.84%	88.67%	81.32%	84.40%	86.79%
	20	84.32%	84.18%	87.88%	86.78%	84.12%
	30	89.38%	87.64%	87.02%	86.44%	89.01%
	40	87.49%	87.47%	87.13%	88.08%	89.48%

4.4.2 *PRESS*

Tabla 4.4 Valores para el PRESS en nivel de ruido bajo para los cuatro porcentajes de separación y *n* diferentes.

PRESS		Tamaño del conjunto de información (nivel de ruido bajo)				
		30	50	100	500	1000
Porcentaje de datos (conjunto de predicción)	10	109.455137	137.336065	154.631401	116.813634	127.716982
	20	263.791744	287.020585	220.807645	216.394257	220.988674
	30	302.238539	308.390769	381.186692	357.719843	347.590536
	40	458.835454	411.642399	453.658005	509.22016	417.544415

Tabla 4.5 Valores para el PRESS en nivel de ruido medio para los cuatro porcentajes de separación y *n* diferentes.

PRESS		Tamaño del conjunto de información (nivel de ruido medio)				
		30	50	100	500	1000
Porcentaje de datos (conjunto de predicción)	10	109.455137	96.1998951	105.708743	100.798171	172.384255
	20	263.791744	188.070845	211.636562	216.394257	220.988674
	30	302.238539	377.073672	331.601402	357.719843	352.411825
	40	458.835454	410.860445	438.883688	500.846209	417.544415

Tabla 4.6 Valores para el PRESS en nivel de ruido alto para los cuatro porcentajes de separación y *n* diferentes.

PRESS		Tamaño del conjunto de información (nivel de ruido alto)				
		30	50	100	500	1000
Porcentaje de datos (conjunto de predicción)	10	109.455137	137.336065	154.631401	117.031057	131.332295
	20	263.791744	287.020585	220.807645	230.673495	241.579281
	30	302.238539	308.390769	381.186692	349.080184	311.169041
	40	458.835454	411.642399	453.658005	443.440227	364.469413

CÁPITULO V. CONCLUSIONES

El modelo de regresión lineal múltiple permite manipular un grupo de datos que darán como resultado por medio de la aplicación de la técnica de Validación cruzada bajo diferentes condiciones de ruido y conjuntos de *n* tamaño obtener la cantidad de datos a guardad óptima.

Dando solución a las preguntas planteadas al inicio de la investigación tomado como referencia los resultado obtenidos por el método propuesto el Capítulo III.

Como se muestra en las tablas 4.1 a la 4.6, los valores una vez utilizada la programación en MATLAB, se determina que el estadístico PRESS en los tres niveles de ruido, el porcentaje óptimo es el 10 por ciento en todos los casos. Por ello es necesario validar con el segundo estadístico $R^2_{predicción}$ donde los resultado obtenidos muestran valores que propician un panorama diferente a lo esperado, ya que para un nivel bajo el tamaño del conjunto de los analizados es el de n=30 con un porcentaje del 10 por ciento de separación; en un nivel de ruido medio obtenemos el mismo resultado donde el mejor es un tamaño del conjunto n=30 con un porcentaje de separación del 10%; mientras tanto para un nivel de ruido alto el tamaño del conjunto óptimo es de n=1000 con un porcentaje del 40 por ciento de separación.

Con los resultados obtenidos se determina que el tamaño de muestra afecta el estadístico $R^2_{predicción}$ de manera significativa para seleccionar el tamaño del conjunto óptimo para la aplicación de la técnica de Validación cruzada.

Por lo tanto se concluye que mientras el nivel de ruido sea bajo o medio el porcentaje a segregar óptimo será del 10 por ciento sin que afecte la capacidad predictiva; mientras que en un nivel de ruido alto el porcentaje será del 40 por ciento.

ANEXOS

I Código del Programa en MATLAB

```
clc, clear
matrizRcuadradaprediccion=zeros(1000,1);%igualar (n,1)
matrizpress=zeros(1000,1);%igualar (n,1)
for e=1:1000;%igualar(n,1)
for z=1:1;
tamanodeinformacion=1000;
efectosprincipales=3;
interacciones=3;
porcentajeseparacion=30;%porcentaje base 100
b0=6;
b1=4;
b2=-2;
b3=5;
b4=-4.8;
b5=4.3;
b6=0.9;
media=0;
varianza=2.1;
%Matriz de efectos principales
matrizefectosprincipales =
zeros(tamanodeinformacion,efectosprincipales);
for i=1:tamanodeinformacion;
for ii=1:efectosprincipales;
matrizefectosprincipales(i,ii)=random('Unif',-1,1);
end
end
%Matriz de interacciones
matrizinteracciones =
zeros(tamanodeinformacion,interacciones);
interaccion1=zeros(tamanodeinformacion,1);
interaccion2=zeros(tamanodeinformacion,1);
interaccion3=zeros(tamanodeinformacion,1);
for j=1:tamanodeinformacion;
for jj=1:interacciones;
interaccion1(j,1)=(matrizefectosprincipales(j,1))*(matrizef
ectosprincipales(j,2));
interaccion2(j,1)=(matrizefectosprincipales(j,1))*(matrizef
ectosprincipales(j,3));
interaccion3(j,1)=(matrizefectosprincipales(j,2))*(matrizef
ectosprincipales(j,3));
matrizinteracciones(:,1)=interaccion1;
matrizinteracciones(:,2)=interaccion2;
```

```
matrizinteracciones(:,3)=interaccion3;
end
end
%Matriz de efectosprincipales e interacciones
matrizdeefectosprincipaleseinteracciones=[matrizefectosprin
cipales,matrizinteracciones];
%Variable de respuesta
matrizrespuesta=zeros(tamanodeinformacion,1);
vectorerror=zeros(tamanodeinformacion,1);
%Error para la VR
for ll=1:tamanodeinformacion;
vectorerror(ll,1)=icdf('Normal',(random('Unif',0,1)),media,
(varianza^1/2));
l=1:tamanodeinformacion;
%Calculo de la VR
matrizrespuesta(l,:)=b0+(b1*matrizdeefectosprincipaleseinte
racciones(l,1))+(b2*matrizdeefectosprincipaleseinteraccione
s(l,2))+(b3*matrizdeefectosprincipaleseinteracciones(l,3))+
(b4*matrizdeefectosprincipaleseinteracciones(l,4))+(b5*matr
izdeefectosprincipaleseinteracciones(l,5))+(b6*matrizdeefec
tosprincipaleseinteracciones(l,6))+(vectorerror(l,1));
end
%Matriz de diseño
matrizdediseno=[matrizdeefectosprincipaleseinteracciones,ma
trizrespuesta];
%matriz de 450 y 50
porcentajeprediccion=(tamanodeinformacion*porcentajeseparac
ion)/100;
porcentajeenestimacion=tamanodeinformacion-
porcentajeprediccion;
corridasaleatorias=(randperm(tamanodeinformacion,porcentaje
enestimacion))';
matrizestimacion=zeros(porcentajeenestimacion,(efectosprinc
ipales+interacciones+1));
for k=1:porcentajeenestimacion;
corrida=corridasaleatorias(k,1);
matrizestimacion(k,:)= matrizdediseno(corrida,:);
end
corridasaleatoriasdeprediccion=zeros(porcentajeprediccion,1
);
c=1;%para separar la matriz de predicción de la de
estimación
for n=1:(porcentajeprediccion+porcentajeenestimacion);
bb=0;
for nn=1:porcentajeenestimacion;
```

```
if (n==corridasaleatorias(nn,1));%compara cada renglón, si
no es igual se extraerá
bb=1;
nnn=porcentajeenestimacion;
end
end
if (bb==0);
corridasaleatoriasdeprediccion(c,1)=n;
c=c+1;
end
end
corridasaleatoriasdeprediccion;
matrizdeprediccion=zeros(porcentajeprediccion,(efectosprinc
ipales+interacciones+1));
for o=1:porcentajeprediccion;
corridab=corridasaleatoriasdeprediccion(o,1);
matrizdeprediccion(o,:)= matrizdediseno(corridab,:);
end
%Separación de la matriz de respuesta para los datos de
precicción
matrizrespuestayobservadadeprediccion=zeros(porcentajepredi
ccion,1);
for p=1:porcentajeprediccion;%Calculo de la nueva VR
matrizrespuestayobservadadeprediccion(p,:)=matrizdepredicci
on(p,end);
end
X=matrizestimacion(:,1:(efectosprincipales+interacciones));
Y=matrizestimacion(:,end);
[b,se,pval,inmodel,stats,nextstep,history] =
stepwisefit(X,Y);
[columna]=find(inmodel==1);
[columna2]=find(inmodel==0);
columnain=(columna);
columnaout=(columna2);
numerodecolumnasinmodelin=(numel(inmodel(inmodel==1)));
numerodecolumnasinmodelout=(numel(inmodel(inmodel==0)));
matrizinmodelinestimacion=zeros(porcentajeenestimacion,nume
rodecolumnasinmodelin);
for v=1:numerodecolumnasinmodelin;
vv=columnain(1,v);
matrizinmodelinestimacion(:,v)= matrizestimacion(:,vv);
end
%se calculan los valores de las simulación que se incluirán
%si son significativos sus coeficientes
matrizinmodelindeprediccion=zeros(porcentajeprediccion,nume
rodecolumnasinmodelin);
```

```
for w=1:numerodecolumnasinmodelin;
ww=columnain(1,w);
matrizinmodelindeprediccion(:,w)=matrizdeprediccion(:,ww);
end
%se estructura la nueva ecuación con los coeficientes que
determinó
%inmodel para introducirle a ésta, los datos de predicción
con los
%que se probará PRESS
coeficientesregressstpwsfitin=(stats.B)';
for qqq=1:numerodecolumnasinmodelin;
qqqq=columnain(1,qqq);
coeficientesin(:,qqq)=coeficientesregressstpwsfitin(:,qqqq)
;
end
matrizvariableMULTcoefpred=zeros(porcentajeprediccion,numer
odecolumnasinmodelin);
for a=1:porcentajeprediccion;
for aa=1:numerodecolumnasinmodelin;
matrizvariableMULTcoefpred(a,aa)=((matrizinmodelindepredicc
ion(a,aa)).*(coeficientesin(1,aa)));
end
end
%calculamos la variable de respuesta con los nuevos
%coeficientes para los datos de predicción
matrizrespuestayajustadadeprediccion=zeros(porcentajepredic
cion,1);
for d=1:porcentajeprediccion;
matrizrespuestayajustadadeprediccion(d,:)=(stats.intercept(
1,1))+(sum(matrizvariableMULTcoefpred(d,:)));
end
%Se calcula la diferencia entre las y´s
matrizdiferenciacalculadaestimada=zeros(porcentajeprediccio
n,1);
for s=1:porcentajeprediccion;
for ss=1,1;
matrizdiferenciacalculadaestimada(s,1)=(matrizrespuestayobs
ervadadeprediccion(s,1))-
(matrizrespuestayajustadadeprediccion(s,1));
end
end
matrizerror=(matrizdiferenciacalculadaestimada);
matrizerrorcuadrada=(matrizerror).^2;
PRESS=sum(sum(matrizerrorcuadrada));
%se obtiene la SSTotal (suma de cuadrados corregida de
cuadrados)
```

```
%de las respuestas en el conjunto de datos de predicción
yprimay=sum(sum((matrizrespuestayobservadadeprediccion).^2)
);
sumatoriayobservadaalcuadrado=((sum(sum(matrizrespuestayobs
ervadadeprediccion)))^2);
zz=porcentajeprediccion;
SSTotalcorregidaprediccion=(yprimay-
(sumatoriayobservadaalcuadrado/zz));
Rcuadradaprediccion=(1-(PRESS/SSTotalcorregidaprediccion));
matrizpress(e,z)=PRESS(1,1);
matrizRcuadradaprediccion(e,z)=Rcuadradaprediccion(1,1);
end
end
%SE APARTARÁN LOS VALORES DE R2 DE LA PREDICCIÓN QUE SEAN
NEGATIVOS PARA
%SÓLO QUEDARSE CON LOS POSITIVOS.
[columna3]=find(matrizRcuadradaprediccion>0);
columnadepositivosR2=columna3;
numeroderenglonespositivosR2=(numel(columnadepositivosR2(co
lumnadepositivosR2>0)));
matrizR2PREDICCIONPOSITIVOS=zeros(numeroderenglonespositivo
sR2,1);
for f=1:numeroderenglonespositivosR2;
r2positivos=columnadepositivosR2(f,1);
matrizR2PREDICCIONPOSITIVOS(f,:)=matrizRcuadradaprediccion(
r2positivos,:);
end
```

BIBLIOGRAFÍA

Allen, D. M. (1971). *The prediction sum of squares as a criterion for selecting predictor variables.* Kentuchy: Department of statistics. University of Kentuchy.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2008). *Estadística para administración y economía.* México, D. F.: Cengage Learning Editores, S. A. de C. V.

Ansley, C. F., & Kohn, R. (1986). Prediction mean squared error for state space. *Biometrika*, 467-473.

Arahal, M. R., Berenguel Soria, M., & Rodríguez Díaz, F. (2006). *Técnicas de predicción con aplicaciones en Ingeniería.* Sevilla, España: Secretariado de Publicaciones de la Universidad de Sevilla.

Arriaga Balderas, A. V. (2017). *Determinación número óptimo de datos para realizar una validación cruzada.* Celaya: Instituto Tecnológico de Celaya.

Baumann, K. (2003). "Cross-validation as the objetive function for variable-selection techiniques". *Trends in Analytical Chemistry*, 395-406.

Benítez López, J., & Hueso Pagoaga, J. L. (1999). Obtenido de http://personales.upv.es/jbenitez/data/matlab.pdf

Bouckaert , R. R. (2008). Practical Bias Variance Decomposition. *Australasian Joint Conference on Artificial Intelligence*, 247–257.

Bryan, S. L. (2012). *Herramientas para el ingeniero industrial. Retrieved from El modelo de regresión lineal.* Obtenido de http://www.udc.es/dep/mate/estadistica2/sec6_3.html

Buenaño Cordero, J. C., De la Cruz Cedeño, C., & Zurita, H. G. (2015). *Verificación de la calidad de modelos en regresión lineal; software estadístico de regresión ERLA.* Obtenido de https://www.dspace.espol.edu.ec/bitstream/123456789/17152/1/RESUMEN_CICYT_Tesina_Buena%C3%B1o_DeLaCruz%20_ICM.pdf

Carmona, P. F. (2005). *Modelos Lineales.* Barcelona, España.: Publicacions edicions de la universitat de barcelona.

Cox, L., & Gaudard, M. (2013). *Discovering Partial Least Squares with JMP.* North Carolina, USA: SAS Institute, Inc.

Dette, H., & Munk, A. (1998). Validation of linear regression models. *The annals of statistics; Institute of mathematical statistics*, 778-800.

Drakos. (1995). *Introduction to Monte Carlo methods.* Comp. Sci. Edu. Proj., , 1 edition.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* United States of America: Chapman & Hall/Crc.

Fortmann-Roe, S. (Mayo de 2012). *Accurately Measuring Model Prediction Error.* Obtenido de http://scott.fortmann-roe.com/docs/MeasuringError.html

García, L., & Lara, P. (1998). *Diseño estadístico de experimentos. Análsisi de la Varianza.* España: Grupo Editorial Universitario.

Gong, G. (1982). *Cross-validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression.* United States of America.: Stanford University.

Goos, G., Hartmanis, J., & Leeuwen, J. V. (2003). *Artificial Neural Networks and Neural Information Processing-ICANN/ICONIP.* Berlín, Alemania.: Springer-Verlag.

Gutiérrez, P. H., & De La Vara , S. R. (2008). *Análisis y diseño de experimentos.* México, D. F.: McGraw-Hill Interamericana.

Hoffman , A. J., & van der Merwe, N. T. (2002). The application of neural networks to vibrational diagnostics for multiple fault conditions. *Elsevier Science Publishing Company, Inc.*, 139-149.

Hurtado, C. (2007). *Evaluación de modelos de clasificación.* Departamento de ciencias de computación, Universidad de Chile.

Joanneum, F. H. (2005). *Cross-Validation Explained.* Obtenido de Institute for Genomics and Bioinformatics - Graz University of Technology: http://genome.tugraz.at/proclassify/help/pages/XV.html

Ke-Lin, D., & Swamy, M. N. (2014). *Neural Networks and Statistical Learning.* London: Springer-Verlag.

Kerner, G. (1015). *Análisis Estadístico con el Método Bootstrap: Aplicaciones en Problemas de Regresión.* Argentina: Universidad de Buenos Aires.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Obtenido de http://citeseer.ist.psu.edu/kohavi95study.html.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2003). *Applied Linear Regression Models.* McGraw-Hill Higher Education.

Lakshmanan, V., Fritz , A., Smith, T., Hondl, K., & Stumpf, G. J. (2007). An automated technique to quality control radar reflectivity data. *Applied Meteorology*, 46, 288-305.

Liu , D., Sun , J., Wei , G., & Liu , X. (2008). RBF Neural Networks and Cross Validation-based Signal Reconstruction for Nonlinear Multi-functional sensor. *International Conference on Signal Processing*, 1512 - 1515 .

Liu, H., Weiss, R. E., Jenrich, R. I., & Wenger, N. (1999). PRESS model selection in repeated measures data. Computational statistics & daya analysis. 16.

Martínez Rodríguez, E. (2005). Errores frecuentes en la interpretación del coeficiente de determinacion lineal. *Anuario Jurídico y Económico Escurialense*, 315-332.

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2010). *Introducción a la probabilidad y estadística.* . México, D. F.: Cengage Learning.

Milton, S. J., & Arnold, J. C. (1999). *Probabilidad y Estadística con Aplicaciones para Ingeniería y ciencias computacionales.* México: Mc Graw-Hill.

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probalility for Engineers.* United Satates of America.: John Wiley & Sons, Inc.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2011). *Introducción al análisis de regresión lineal.* México. D.F.: Grupo Editorial Patria.

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de teledetección*, 44, 55-65.

Peró, M., & Guàrdia Olmos, J. (2001). *Esquemas de estadística: aplicaciones en intervención ambiental.* Universitat de Barcelona: Edicions Universitat Barcelona.

Piñeiro Redondo, Y. (2007). *Simulación de Monte Carlo de sistemas complejos en red.* España: Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico.

Refaeilzadeh, P., Tang, L., & Liu, H. (2008). *Cross-Validation.* United States of America: Arizona State University.

Ríos, A. J., & Simpson , J. R. (2016). A sequential augmentation method to eliminate multicollinearity. *Journal Quality Engineering* , 588-604.

Rubin, D. S., & Levin, R. I. (2004). *Estadistica para administración y economía.* Mexico: Pearson Educación.

Snee, R. D. (1977). Validation of regression models: methods and examples. *Technometrics C.*, 415-428.

Sominski, I. S. (1975). *Método de la inducción Matemática.* Editorial MIR.

Stone, M. (1974). Cross-Validatory and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B.*, 111-147.

Taha, H. A. (2004). *Investigación de operaciones.* México: Pearson Educación.

Valencia Delfa, J. L. (2009). *Regresión PLS en las ciencias Esperimentales.* Madrid, España: Editorial Complutense.

Valencia Delfa, J., Díaz-LLanos, F., & Calleja, S. (2003). *Regresión PLS en las ciencias Esperimentales.* Madrid, España: Editorial Complutense.

Vargas Sabadías, A. (1995). *Estadística descriptiva e inferencial.* España: Universidad de Castilla La Mancha.

Walpole, R. E., Myers, R. H., & Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias.* . México.: Pearson Ecucación.

Werner, D., Granzow, M., & Berrar, D. (2007). *Fundamentals of Data Mining in Genomics and proteomics.* New York, NY, USA.: Springer Science Business Media, LLC.

Zhang, F. (2011). *Cross-validation and Regression Analysis in High-dimensional Sparse Linear Models.* United Stated of America: Stanford University.