UNIVERSIDAD TÉCNICA DE AMBATO

FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA E INDUSTRIAL

CENTRO DE POSGRADO

MAESTRÍA EN MATEMÁTICA APLICADA

Tema: ESTUDIO DE LOS MODELOS DE REGRESIÓN PARAMÉTRICOS POLINOMIALES Y MODELOS DE REGRESIÓN NO PARAMÉTRICOS B-SPLINES. APLICACIONES EN INGENIERÍA.

Trabajo de titulación previo a la obtención del grado académico de Magíster en Matemática Aplicada

Modalidad de Titulación Proyecto de Desarrollo

Autor: Ing. Byron Miguel Toalombo Rojas

Director: Dr. Manuel Antonio Meneses Freire, Ph.D.

Ambato – Ecuador

2021

APROBACIÓN DEL TRABAJO DE TITULACIÓN

A la Unidad Académica de Titulación de la Facultad de Ingeniería en Sistemas, Electrónica e Industrial.

El Tribunal receptor del Trabajo de Investigación presidido por la Ingeniera Elsa Pilar Urrutia Urrutia, Mg., Presidenta del Tribunal, e integrado por los señores: Profesor Saba Rafael Infante Quirpa Dr. y Doctor Freddy Geovanny Benalcázar Palacios Mg., designados por la Unidad Académica de Titulación de la Universidad Técnica de Ambato, para receptar el trabajo de Titulación con el tema: "ESTUDIO DE LOS MODELOS DE REGRESIÓN PARAMÉTRICOS POLINOMIALES Y MODELOS DE REGRESIÓN NO PARAMÉTRICOS B-SPLINES. APLICACIONES EN INGENIERÍA", elaborado y presentado por el Señor Ing, Byron Miguel Toalombo Rojas, para optar por el Grado Académico de Magister en Matemática Aplicada; una vez escuchada la defensa oral del Trabajo de Titulación el Tribunal aprueba y remite el trabajo para uso y custodia en las bibliotecas de la Universidad Técnica de Ambato.

Firmado electrónicamente por:
ELSA PILAR URRUTIA

Ing. Elsa Pilar Urrutia Urrutia, Mg.

Presidenta y Miembro del Tribunal de Defensa

SABA RAFAEL INFANTE QUIRPA

Firmado digitalmente por SABA RAFAEL INFANTE QUIRPA
Fecha: 2021.07.22 13:02:08 -05'00'

Prof. Saba Rafael Infante Quirpa Dr.

Miembro del Tribunal de Defensa

Firmado electrónicamente por:
FREDDY GEOVANNY BENALCAZAR PALACIOS

Dr. Freddy Geovanny Benalcázar Palacios Mg.

Miembro del Tribunal de Defensa

AUTORÍA DEL TRABAJO DE TITULACIÓN

La responsabilidad de las opiniones, comentarios y críticas emitidas en el Trabajo de Titulación presentado con el tema: "ESTUDIO DE LOS MODELOS DE REGRESIÓN PARAMÉTRICOS POLINOMIALES Y MODELOS DE REGRESIÓN NO PARAMÉTRICOS B-SPLINES. APLICACIONES EN INGENIERÍA", le corresponde exclusivamente al: Ing. Byron Miguel Toalombo Rojas, autor bajo la Dirección del Dr. Manuel Antonio Meneses Freire, Ph.D., Director del Trabajo de Investigación; y el patrimonio intelectual a la Universidad Técnica de Ambato.

Firmado electrónicamente por:
BYRON MIGUEL TOALOMBO ROJAS

Ing. Byron Miguel Toalombo Rojas
c.c. 180366934-8
AUTOR

Firmado electrónicamente por:
MANUEL ANTONIO MENESES FREIRE

Dr. Manuel Antonio Meneses Freire, Ph.D.
c.c. 180251584-9
DIRECTOR

DERECHOS DE AUTOR

Autorizo a la Universidad Técnica de Ambato, para que el Trabajo de Titulación, sirva como un documento disponible para su lectura, consulta y procesos de investigación, según las normas de la Institución.

Cedo los Derechos de mi Trabajo de Titulación, con fines de difusión pública, además apruebo la reproducción de este, dentro de las regulaciones de la Universidad Técnica de Ambato.

Firmado electrónicamente por:
BYRON MIGUEL TOALOMBO ROJAS

Ing. Byron Miguel Toalombo Rojas
c.c. 180366934-8

ÍNDICE GENERAL DE CONTENIDOS

CAPÍTULO III

3. MARCO METODOLÓGICO

CAPÍTULO IV

4. RESULTADOS Y DISCUSIÓN

CAPÍTULO V

5. CONCLUSIONES, RECOMENDACIONES, BIBLIOGRAFÍA Y ANEXOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

AGRADECIMIENTO

A Dios, ser trascendente del mundo espiritual.

A mis padres, por su apoyo incondicional.

A mis hermanos, especialmente a Andrea (+).

Al Dr. Antonio Meneses, Ph.D. por su colaboración en la investigación.

A la Universidad Técnica de Ambato, por acogerme en mi formación académica.

Byron Toalombo

DEDICATORIA

A mis seres queridos y amigos cercanos.
To my loved ones and close friends.

Byron Toalombo

"Mathematical science shows what is. It is the language of unseen relations between things. But to use and apply that language, we must be able fully to appreciate, to feel, to seize the unseen, the unconscious".

Ada Lovelace

UNIVERSIDAD TÉCNICA DE AMBATO

FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA E INDUSTRIAL

MAESTRÍA EN MATEMÁTICA APLICADA

TEMA:

Estudio de los modelos de regresión paramétricos polinomiales y modelos de regresión no paramétricos B-Splines. Aplicaciones en Ingeniería.

AUTOR: Ing. Byron Miguel Toalombo Rojas, Ingeniero Mecánico.

DIRECTOR: Dr. Manuel Antonio Meneses Freire, Ph.D.

LÍNEA DE INVESTIGACIÓN: Diseño, Materiales y Producción.

FECHA: 28 de mayo de 2021.

RESUMEN EJECUTIVO

Se realiza el estudio de los modelos de regresión paramétricos polinomiales y modelos de regresión no paramétricos B-Splines, a partir de aplicaciones particulares en ingeniería. Se consideran los casos de una simulación del impacto de un auto contra la carrocería de un autobús como parte del diseño de la estructura y la relación de las variables climatológicas en la estación meteorológica de San Antonio de Pichincha. Se plantea un diseño metodológico no experimental de corte transversal, siendo una investigación de tipo correlacional. Para determinar los modelos de regresión apropiados para cada relación se utiliza el software R y se tienen en cuenta los criterios de: rechazo de la nulidad de los coeficientes de los modelos mediante la prueba de hipótesis t de Student, validez de los modelo mediante la prueba F de Snedecor de la tabla ANOVA, bondad de ajuste, intervalos de confianza al 95%, y cumplimiento de los supuestos de distribución normal, no autocorrelación y homocedasticidad de los residuos para la regresión polinomial (pruebas de Shapiro-Wilk, Kolmogorov-Smirnov corregida por Lilliefors, Durbin-Watson y Breusch-Pagan, respectivamente). Para la selección del modelo de regresión más idóneo se aplicó la prueba no paramétrica de Wilcoxon, a partir de las longitudes de los intervalos de confianza. Con base en los resultados obtenidos, los modelos de regresión paramétricos polinomiales se ajustan bien cuando las curvas tienen forma parabólica o siguen un patrón sin cambios abruptos de curvatura, adaptándose mejor a las relaciones de la simulación de impacto vehicular que tienen como variable explicativa la velocidad del vehículo que impacta. En cambio, los modelos de regresión no paramétricos B-splines brindan un mejor ajuste cuando las curvas tienen forma de campana con cambios de curvatura más abruptos, adaptándose mejor a las condiciones de las variables climatológicas en función de la hora del día.

Descriptores: <Bondad de ajuste>, <Intervalo de confianza>, <Métricas de error>, <Modelos de regresión B-splines>, <Modelos de regresión polinomiales>, <Normalidad>, <Prueba de hipótesis>, <Simulación de impacto vehicular>, <Software R>, <Variables climatológicas>.

UNIVERSIDAD TÉCNICA DE AMBATO

FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA E INDUSTRIAL

MASTER IN APPLIED MATHEMATICS

THEME:

Study of polynomial parametric regression models and non-parametric B-Splines regression models. Engineering Applications.

AUTHOR: Eng. Byron Miguel Toalombo Rojas, Mechanical engineer.

DIRECTED BY: Dr. Manuel Antonio Meneses Freire, Ph.D.

LINE OF RESEARCH: Design, Materials and Production.

DATE: May, 28th 2021.

ABSTRACT

The study of polynomial parametric regression models and non-parametric B-Splines regression models is carried out based on particular applications in engineering. The cases considered are a car against the bus bodywork crash simulation as part of the structural design and the relationship of the climatological variables in the weather station of San Antonio de Pichincha. A non-experimental cross-sectional methodological design is made, being a correlational type of research. The appropriate regression models for each relationship are established with the use of R software and taking into account the criteria: rejection of the nullity of the coefficients of the models by Student's t-hypothesis test, the validity of the models by Snedecor's F test of the ANOVA table, the goodness of fit, 95% confidence intervals, and compliance with the assumptions of normal distribution, no autocorrelation, and homoscedasticity of the residuals for the polynomial regression (Shapiro-Wilk, Kolmogorov-Smirnov corrected by Lilliefors, Durbin-Watson and Breusch-Pagan tests, respectively). The Wilcoxon nonparametric test was used to select the most suitable regression model based on the lengths of the confidence intervals. Based on the results obtained, the polynomial parametric regression models fit well when the curves have a parabolic shape or follow a pattern without abrupt changes in curvature. It means the model can define better to the relationships of the vehicle impact simulation that have the speed of the impacting vehicle as an explanatory variable. In contrast, the nonparametric B-splines regression models provide a better fit when the curves are bell-shaped with more abrupt curvature changes. This model can adapt better to the conditions of the climatological variables as a function of the time of day.

Keywords: <B-Splines regression models>, <Climatic variables>, <Confidence interval>, <Error metrics>, <Goodness-of-fit>, <Hypothesis test>, <Normality>, <Polynomial parametric regression models>, <R Software>, <Vehicle crash simulation>.

CAPÍTULO I

EL PROBLEMA DE INVESTIGACIÓN

1.1. Introducción

El proyecto de titulación se enfoca en la caracterización de los modelos teóricos estadísticos de regresión paramétricos polinomiales y no paramétricos B-splines, que desde un enfoque conceptual son el objeto mismo del estudio. Adicionalmente, se plantea mediante la modelización de dos problemas particulares de la ingeniería, delimitar la aplicabilidad de los referidos modelos y la bondad de ajuste correspondiente. En este sentido, se considerarán dos problemas concretos referentes al campo de la Ingeniería Mecánica y Ambiental, que son de interés y que requieren ser explicados desde un enfoque matemático. Metodológicamente el trabajo que se propone, es un estudio de tipo no experimental probabilístico, en el que se considera el manejo de una muestra de datos de tipo numérico continuo. La función del autor se centra en la modelización matemática, a través de la manipulación de los datos con el uso de herramientas estadísticas y con base en los principios y teoremas que rigen la estadística paramétrica y no paramétrica.

En los diferentes campos de la ingeniería se presentan problemas que se describen con un enfoque cualitativo y/o cuantitativo, mediante modelos matemáticos, que usualmente se establecen a través del uso de un conjunto de datos de variables inherentes a los fenómenos de interés. Particularmente, en el caso de la dinámica del movimiento de vehículos surge el interés por explicar la relación existente entre las magnitudes físicas involucradas en el fenómeno del impacto o colisión. Específicamente se toma como caso particular el análisis del efecto de la velocidad de circulación de un automóvil modelo compacto con la fuerza de impacto, con el tiempo de duración del impacto, con la deformación que produce y con el factor de seguridad (FDS) durante un evento de choque contra la parte posterior de un autobús. Este tipo de relación es de interés en el diseño de la estructura de las carrocerías de autobuses, considerando el estudio del efecto que ocasionan los

impactos o colisiones y estos criterios se tienen en cuenta en la etapa de diseño de la estructura de la carrocería de los autobuses.

Por otra parte, un problema común dentro del campo de la Ingeniería Ambiental corresponde a la necesidad de pronosticar con cierto grado de certidumbre la relación entre los parámetros climatológicos. Por ejemplo, se requiere determinar si a partir de la intensidad de la radiación solar es posible conocer la temperatura ambiental de un determinado espacio geográfico.

En este sentido, el aporte de la Matemática Aplicada permite el establecimiento de relaciones entre las magnitudes inherentes a ambos casos antes indicados. Los modelos matemáticos probabilísticos corresponden a la estadística y pueden ser de dos tipos principales, paramétricos y no paramétricos. Los primeros son tradicionalmente conocidos y relativamente más sencillos de utilizar, sin embargo, conllevan algunas restricciones para ser utilizados para cada caso particular. Adicionalmente es necesario considerar la bondad de ajuste de los modelos, para tomar una decisión respecto al que se adapta de mejor manera a las condiciones de los problemas estudiados.

1.2. Justificación

En la ingeniería están presentes fenómenos que necesitan ser explicados dentro del ámbito científico, para lo cual se puede hacer uso de la matemática, como una herramienta que brinda información probabilística confiable para la definición y el pronóstico de las variables de forma objetiva.

La interrogante de partida es la siguiente: ¿Cuáles son las relaciones existentes entre las variables inherentes al impacto de vehículos y las variables climatológicas mediante el uso modelos de regresión paramétricos polinomiales y modelos de regresión no paramétricos B-splines?

La descripción y simulación del impacto de un auto contra la carrocería de un autobús como parte del diseño de la estructura de la carrocería, demanda de la

obtención de datos de las variables velocidad, fuerza, tiempo de impacto, factor de seguridad y deformación de la estructura. La obtención de datos es factible mediante el empleo de un software de simulación de diseño asistido por computador (CAD) y de elementos de análisis computarizado (CAE). Para el efecto, en primer lugar se realiza el diseño de los vehículos a ser impactados, tomando en cuenta los materiales (estructura del autobús de tubo cuadrado de acero ASTM A500), las dimensiones (modelo interprovincial) y las formas correspondientes. Seguidamente se efectúa una simulación del movimiento vehicular a diferentes velocidades y el software arroja los resultados de las variables respuesta, que son fuerza, deformación, tiempo de impacto y FDS. Estos datos pueden ser exportados a una hoja de cálculo, con lo cual se demuestra la factibilidad de la disponibilidad de los datos para la solución del problema 1, que toma como variable explicativa a la velocidad lineal.

Por su parte, el desarrollo del estudio de la relación de las variables climatológicas también es factible, porque existe la información disponible en la página web de la Secretaría del Ambiente del Distrito Metropolitano Quito, que es de uso público y que contiene información desde el año 2004 hasta el 2020, lo que hace posible contar con una base de miles de datos.

Desde el punto de vista académico, el tema del presente trabajo de titulación se enmarca dentro de las líneas de investigación del programa de Maestría en Matemática Aplicada, por lo que se justifica su importancia.

El impacto que tendrá el desarrollo de la investigación radica en que presenta un estudio académico, como una herramienta que brinda información útil para el desarrollo de tecnologías en la fabricación de carrocerías de autobuses, así como para pronosticar el comportamiento de las variables climatológicas en la ciudad de Quito y en la región Sierra en general.

1.3. Objetivos

1.3.1 General

Estudiar los modelos de regresión paramétricos polinomiales y modelos de regresión no paramétricos B-splines, para comparar su bondad de ajuste en la modelación de problemas de la Ingeniería Mecánica y Ambiental.

1.3.2 Específicos

- Caracterizar los modelos de regresión paramétricos polinomiales y modelos de regresión no paramétricos B-splines.
- Determinar las variables numéricas inherentes a los problemas de la simulación del impacto vehicular entre un vehículo y un autobús, y las variables climatológicas que pueden tener una relación mutua.
- Establecer los modelos de regresión polinomiales y modelos de regresión B-Splines que definen el comportamiento de los problemas mencionados.
- Contrastar la bondad de ajuste de los modelos estudiados.

CAPÍTULO II

ANTECEDENTES INVESTIGATIVOS

2.1 Estado del arte

2.1.1 Modelos de regresión

La regresión tiene por objeto la modelación del efecto que produce un conjunto dado de variables explicativas x_1, x_2, x_3, …, x_k en una variable respuesta o dependiente *y*, que es de interés primario. Las variables explicativas también se denominan covariables, variables independientes o regresores.

Los distintos modelos de regresión se diferencian principalmente por el tipo de variables de respuesta (continuas, binarias, categóricas) y los diferentes tipos de covariables, que también pueden ser continuas, binarias o categóricas. En situaciones más complejas, también es posible incluir escalas de tiempo, variables para describir la distribución espacial o ubicación geográfica, o indicadores de grupo.

Una característica principal de los modelos de regresión es que la relación entre la variable de respuesta *y* y las covariables no es una función determinista $f(x_1, x_2, \ldots, x_k)$ de $x_1, x_2, \ldots, x_k$, sino que muestra errores aleatorios. Esto implica que la respuesta *y* es una variable aleatoria, cuya distribución depende de las variables explicativas [1].

El objetivo principal de la regresión es analizar la influencia de las covariables en el valor medio de la variable de respuesta. En otras palabras, se modela el valor esperado (condicional) $E(y \mid x_1, x_2, \ldots, x_k)$ de *y* dependiendo de las covariables. Por tanto, el valor esperado es una función de las covariables:

$$E(y \mid x_1, x_2, \ldots, x_k) = f(x_1, x_2, \ldots, x_k) \quad (1)$$

Incorporando la desviación aleatoria ϵ es posible descomponer la respuesta en:

$$y = \mathrm{E}(\mathrm{y} \mid x_1, x_2, \dots, x_k) + \epsilon = f(x_1, x_2, \dots, x_k) + \epsilon \qquad (2)$$

El valor esperado *E* usualmente se denota como el componente sistemático del modelo. La desviación aleatoria ε también se denomina componente aleatoria o estocástica, perturbación o término de error.

2.1.2 Modelos de regresión paramétricos

El tipo de modelo de regresión más comúnmente conocido es la regresión lineal simple, la cual contempla la existencia de una sola variable explicativa o regresoras y que está dada por el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ , \quad i = 1, \dots, n. \qquad (3)$$

Donde β_0 y β_1 son parámetros desconocidos o coeficientes de regresión y ϵ_i es un error no observable que se supone que está presente en cada observación. Los errores ϵ_i son independiente e idénticamente distribuidos (i. i .d.), de modo que: $E[\epsilon_i] = 0$ y $V[\epsilon_i] = \sigma^2$ [2].

La suposición de que el valor esperado de los errores es nulo implica que el valor esperado de la respuesta depende sólo de la variable explicativa. La suposición de varianza constante entre los errores se denomina homocedasticidad. En particular, esto implica que los errores son independientes de las covariables. Mientras que la suposición de que los errores no están correlacionados se conoce como no colinealidad.

El modelo anterior contempla la existencia de una sola variable explicativa o regresora *x*, sin embargo cuando el número de variables es mayor que uno, corresponde la formulación de un modelo más general. En este caso corresponde referirse a un modelo de regresión lineal múltiple (MRLM), que también se conoce

como el modelo clásico de regresión lineal. El MRLM en forma matricial se expresa de la siguiente manera:

$$y = \mathrm{X}\beta + \epsilon \tag{4}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathrm{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{5}$$

Donde:

y, vector de *n* valores que toma la variable respuesta, dependiente o a predecir.
X, matriz de las *k* variables regresoras para el conjunto de *n* valores disponibles.
β, vector de *k*+1 parámetros desconocidos o coeficientes de regresión del modelo.
ϵ, vector de errores no observables presentes en las mediciones.

El MRLM también se puede expresar en forma algebraica, una vez que se desarrolla el producto matricial Xβ, a través de la siguiente expresión:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i \ , \quad i = 1, \ldots, n. \tag{6}$$

La función lineal estimada es:

$$\hat{f}(x_1, x_2, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \tag{7}$$

Puede ser utilizado como estimador $\widehat{\mathrm{E}}(\mathrm{y} \mid x_1, x_2, \ldots, x_k)$ para el valor condicional esperado de *y* dado las covariables x_1, x_2, …, x_k. Como tal, se puede utilizar para predecir *y*, denotado como $\hat{y}$ [1].

2.1.3 Modelos de regresión paramétricos polinomiales

La regresión polinomial es un caso especial de regresión múltiple, con solo una variable independiente X. El modelo de regresión polinomial de una variable explicativa o regresora se puede expresar de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \ldots + \beta_k {x_i}^k + \epsilon_i \ , \quad i = 1, \ldots, n. \tag{8}$$

Donde:

n es el número de datos disponibles a ser utilizados para hallar el modelo.
k es el grado del polinomio.

El grado del polinomio también es el orden del modelo. Efectivamente, esto es lo mismo que tener un modelo múltiple con $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$, etc. [3]. El modelo de regresión polinomial de dos variables explicativas o regresoras se puede expresar de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 {x_{1i}}^2 + \beta_4 {x_{2i}}^2 + \beta_5 x_{1i} x_{2i} + \ldots + \beta_l {x_{1i}}^k + \beta_m {x_{2i}}^k + \epsilon_i \ , \ i = 1, \ldots, n. \tag{9}$$

En forma matricial el modelo de regresión polinomial de una variable explicativa o regresora se expresa de la siguiente manera:

$$y = \mathrm{X}\beta + \epsilon \tag{10}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \mathrm{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ 1 & x_2 & x_2^2 & \ldots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \ \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{11}$$

Donde:

y, vector de *n* valores que toma la variable respuesta, dependiente o a predecir.
X, matriz de la variable regresora para los *k* grados del polinomio en el conjunto de *n* valores disponibles.
β, vector de *k*+1 parámetros desconocidos o coeficientes de regresión del modelo.
ϵ, vector de errores no observables presentes en las mediciones [4].

El problema consiste en hallar los componentes del vector β, que se pueden calcular a partir de despejar el vector β de la ecuación (10), lo que implica la resolución de una matriz inversa, como se indica a continuación:

$$\beta = \mathrm{X}^{-1}(y + \epsilon) \tag{12}$$

Una vez establecido un modelo de regresión polinomial para un conjunto de datos, se pueden obtener a partir de él valores predichos de las variables respuesta. En este caso, el vector de valores predichos o ajustados se denota como $\hat{y}$, el cual generalmente difiere del vector *y* de valores observados de las variables respuesta. Es decir que, cualquier modelo de regresión hace una estimación de las variables de respuesta, pero con una desviación y por lo tanto se existen residuos. En este sentido, es de interés el análisis de la denominada matriz "Hat" H, que relaciona ambos vectores, a través de la expresión que se muestra a continuación:

$$\hat{y} = \mathrm{H} \cdot y \tag{13}$$

Entre las aplicaciones más comunes de la matriz H en el análisis de regresión se tiene a la distancia de Hook y al apalancamiento. Al mismo tiempo, el vector de residuos se puede obtener mediante las siguientes expresiones:

$$e = \hat{y} - y = (\mathrm{I} - \mathrm{H}) \cdot y \tag{14}$$

Donde:

e, vector de residuos entre los valores precios y los valores observados.
I, matriz identidad, que es una matriz diagonal cuyos elementos son iguales a 1.
H, matriz Hat.
$\hat{y}$, vector de valores predichos de la variable respuesta o a predecir.
y, vector de valores observados de la variable respuesta o a predecir.

2.1.3.1 Prueba ANOVA para evaluar un modelo de regresión

Para evaluar el grado de fortaleza de la regresión como indicador de la relación entre las variables explicativas y la variable a predecir, se puede aplicar la prueba de análisis de varianza (ANOVA). Para el efecto, se llevan a cabo contrastes de hipótesis para los beta´s β, además se calculan los residuos, el coeficiente de determinación y la Tabla ANOVA [5].

En el caso de la prueba de contraste de hipótesis de los beta´s β, como sirve como evidencia estadística que justifica el modelo, se tienen las siguientes hipótesis:

Nula:

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{15}$$

Alternativa:

$$H_1\colon \beta_j \neq 0 \quad para\ al\ menos\ un\ j \tag{16}$$

La relación entre las variables existe sí y solo sí, existe al menos un β que sea diferente de cero. El contraste de hipótesis se efectúa mediante el estadístico de prueba con distribución F de Snedecor, que se calcula mediante la siguiente expresión:

$$F_0 = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-p}} \tag{17}$$

Donde:

F_0, es el estadístico de prueba F de Snedecor.
MSR, media cuadrática de la regresión.
MSE, media cuadrática del error.
SSR, suma de cuadrados de la regresión.
SSE, suma de cuadrados del error.
n, es el número de datos disponibles a ser utilizados para hallar el modelo.
k, es el grado del polinomio.
$p = k+1$.

La suma de cuadrados totales se calcula a partir de la suma de cuadrados de la regresión y de la suma de cuadrados del error, como sigue:

$$SST = SSR + SSE \tag{18}$$

Donde:

SST, suma de cuadrados totales corregida.

En la expresión (18) la suma de cuadrados total corregida se puede expresar tomando en cuenta los valores predichos y las medias de los datos [6], conforme se indica en la siguiente expresión:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{19}$$

A partir de los parámetros descritos que se utilizan para la evaluación de un modelo de regresión y tomando en cuenta los tres parámetros de la variación: regresión, error y total, se elabora la tabla ANOVA, como se muestra a continuación:

Tabla 2-1. Tabla análisis de varianza, ANOVA.

Fuente de variación	**Grados de libertad**	**Sumas cuadráticas**	**Medias cuadráticas**	**F**	**p-valor**
Regresión	k	$SSR = \hat{\beta}_1 \cdot S_{xy}$	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$	
Error	n - p	$SSE = SST - \hat{\beta}_1 \cdot S_{xy}$	$MSE = \frac{SSE}{n-p}$	-	-
Total	n - 1	SST	-	-	-

Fuente: [6].

Elaborado por: Toalombo, B. (2021).

En la tabla ANOVA, el modelo es perfecto si SSE = 0, error igual a 0. Para un determinado intervalo de confianza (1-α) *x*100%, se rechaza la hipótesis nula H_0 en favor de la alternativa H_1 si el estadístico de prueba F de Snedecor calculado es mayor que el percentil (1-α) *x*100% de F(v_1, v_2) con v_1 = (p -1) grados de libertad en el numerador y v_2 = (n -p) grados de libertad en el denominador. Lo que se representa de la siguiente manera:

$$F = \frac{MSR}{MSE} > F_{\alpha,p-1,n-p} \tag{20}$$

Posterior a que se rechace la hipótesis nula en la expresión (15), se efectúa la prueba individual para conocer los beta´s β que son diferentes de 0 y por consiguiente las variables que realmente aportan al modelo de regresión. En este sentido, el nuevo contraste de hipótesis es de la forma:

Nula:

$$H_0: \beta_i = 0 \tag{21}$$

Alternativa:

$$H_1: \beta_i \neq 0 \quad para\ i = 1, 2, \dots, p-1 \tag{22}$$

Para un determinado intervalo de confianza (1-α) *x*100%, se rechaza la hipótesis nula H_0 en favor de la alternativa H_1 si el estadístico de prueba *t* de Student calculado es mayor que el percentil (1-α) *x*100% de *t* con (*n*-*p*) grados de libertad.

2.1.3.2 Grado apropiado del polinomio

Un problema frecuente en el análisis de regresión lineal es determinar la cantidad de variables independientes que se deben incluir en la función de regresión ajustada. En el presente caso atañe referirse cuando las variables independientes son potencias sucesivas del índice de observación. Generalmente si se incluye en la función de regresión una potencia en particular, también se incluyen todas las potencias inferiores. En este sentido el problema de estudio se direcciona a la determinación de cuantas potencias incluir, lo que significa hallar el grado del polinomio de regresión [7].

Para seleccionar el grado más apropiado del polinomio u orden del modelo de regresión, se debe desarrollar la Tabla ANOVA de diferentes modelos, comparar y

determinar cual tiene el menor error, a simple vista esto implica comparar entre los p-valores. Adicionalmente se debe tener en cuenta la simplicidad del modelo, es decir a p-valores similares, tendrá prioridad el modelo de menor orden o grado del polinomio más sencillo.

2.1.3.3 Métricas de medición del error de los modelos de regresión

El vector de coeficientes $\hat{\beta}$ como estimador de mínimos cuadrados se calcula a partir de la siguiente expresión:

$$\hat{\beta} = (X' \cdot X)^{-1} \cdot X' \cdot y \tag{23}$$

Donde *y* es la matriz de valores que se pretende predecir [8].

El método de mínimos cuadrados tiene por finalidad encontrar los beta´s β que minimicen la suma total de las diferencias al cuadrado entre los valores observados y los predichos:

$$mín\ f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k) = mín\left\{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \right\} \tag{24}$$

$$mín\ f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k) = mín\left\{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 {x_i}^2 - \dots - \hat{\beta}_k {x_i}^k\right)^2\right\}, \qquad i = 1, \dots, n. \tag{25}$$

La media cuadrática del error *MSE* es un estimador insesgado de la varianza σ^2 del término de error aleatorio y se define en la ecuación:

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - (k+1)} \tag{26}$$

Donde y_i son valores observados, $\hat{y}_i$ son los valores ajustados o estimados de la variable dependiente *y* para el i-ésimo caso, *SSE* es la suma de cuadrados de los residuos y df_E son los grados de libertad.

El *MSE* es una medida que expresa que tan bien la regresión se ajusta a los datos. La raíz cuadrada de *MSE* es un estimador de la desviación estándar σ del término de error aleatorio. La raíz del error cuadrático medio *RMSE* = $\sqrt{MSE}$ no es un estimador insesgado de σ, pero sigue siendo un buen estimador.

MSE y *RMSE* son medidas del tamaño de los errores en la regresión y no brindan una indicación sobre el componente explicado del ajuste de regresión. Sin embargo, la medida más útil para comparar la precisión de los pronósticos entre diferentes elementos o productos es el error porcentual absoluto medio *MAPE*, ya que mide el rendimiento relativo. El *MAPE* es una medida de precisión comúnmente utilizada en métodos cuantitativos de pronóstico. Esta medida se define mediante la siguiente ecuación:

$$MAPE = \frac{100\ \%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{27}$$

Si el valor calculado del *MAPE* es inferior al 10%, se interpreta como un pronóstico excelente y preciso; entre un 10 y 20% es un buen pronóstico; entre un 20 y 50% es un pronóstico aceptable y más del 50% un pronóstico inexacto [3].

Otro indicador que se emplea para evaluar la eficacia de un modelo de regresión paramétrico polinomial es el cuadrado R^2 o coeficiente de determinación, que se define como:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{28}$$

Donde *SST* es la suma total de cuadrados e $\bar{y}$ es la media aritmética de la variable *y*. El R^2 mide el porcentaje de variación en la variable de respuesta *y* explicado por la variable explicativa *x*. Por lo tanto, el coeficiente de determinación es una medida importante de qué tan bien se ajusta el modelo de regresión a los datos. El valor de R^2 siempre está entre 0 y 1 ($0 < R^2 < 1$). Un valor de R^2 de 0.9 o superior es muy bueno, un valor superior a 0.8 es bueno y un valor de 0.6 o superior puede ser

satisfactorio en algunas aplicaciones, aunque se debe considerar que, en tales casos, pueden producirse errores de predicción que podrían ser relativamente altos. Cuando el valor de R^2 es 0.5 o menos, la regresión explica sólo el 50% o menos de la variación de los datos; por lo tanto, la predicción puede ser pobre. El coeficiente de determinación ajustado R^2_{ajustado} se calcula mediante la siguiente expresión:

$$R^2{}_{ajustado} = R^2 - \frac{(1 - R^2) \cdot k}{n - (k + 1)} \tag{29}$$

La fórmula (29) muestra explícitamente el proceso de "ajuste" y también demuestra que el R^2_{ajustado} es siempre menor que R^2. R^2_{ajustado} se ajusta por el número de variables incluidas en la ecuación de regresión. Si el valor de R^{*2} es mucho más bajo que el valor de R^2, es una indicación de que la ecuación de regresión puede estar sobre ajustada a la muestra y de generalización limitada. Siempre se prefiere R^2_{ajustado} a R^2 cuando se examinan datos, debido a la necesidad de protección contra relaciones espurias.

2.1.3.4 Intervalo de confianza de los parámetros beta´s β

Un intervalo de confianza es un rango entre el que fluctúan los datos o valores que se obtienen de un modelo estadístico. El intervalo es construido a partir de los parámetros estimados del modelo y considera un determinado nivel de confianza (generalmente 95 o 99%). Por consiguiente, un intervalo de confianza se establece entre dos límites (inferior y superior) en torno a la curva que define el modelo, cuanto más alto sea el nivel de confianza los límites estarán más cercanos a la curva del modelo [9]. La importancia de establecer un intervalo de confianza radica en que permite hacer pronósticos de los datos con un buen nivel de exactitud.

2.1.4 Modelos de regresión no paramétricos

Los modelos de regresión paramétricos se establecen a partir de determinados parámetros del modelo y por consiguiente su interpretación suele ser relativamente sencilla, pero al mismo tiempo presentan el inconveniente de que la

parametrización puede ser demasiado rígida imposibilitando que el ajuste a los datos sea óptimo.

En aplicaciones complejas con regresores más continuos, la búsqueda de transformaciones adecuadas se vuelve muy difícil o intratable incluso para investigadores muy experimentados. Para conseguir una modelación matemática que proporcione una mejor capacidad de ajuste de los datos y por ende que sea más eficaz desde un punto de vista predictivo, se requiere la aplicación de modelos más flexibles y generales, lo que es posible recurriendo a la estadística funcional no paramétrica.

Los modelos de regresión no paramétricos permiten una estimación flexible de efectos no lineales. No requieren ningún supuesto restrictivo con respecto a una determinada forma funcional paramétrica. En el caso de una sola covariable continua *x*, el modelo estándar para la regresión no paramétrica se define de la siguiente manera:

$$y_i = f(x_i) + \epsilon_i \ , \quad i = 1, \dots, n. \tag{30}$$

En el modelo anterior, para la variable de error ϵ_i, se hacen los mismos supuestos que fueron tomados en el modelo de regresión lineal simple:

Se supone que la función f presenta ciertas características de suavidad, como continuidad o diferenciabilidad, pero no se especifica una forma paramétrica específica. Se estima en base a datos a través de enfoques no paramétricos.

En los modelos de regresión no paramétricos, la función de regresión f que define la relación (30) no presentan una forma paramétrica fija, más bien dependen de una función desconocida que demanda de cierto grado de suavidad. En este punto toma fuerza el concepto de métodos de suavización, tales como: Suavizadores Polinómicos locales, Suavizadores Kernel, Cubic smoothing splines y Splines de regresión [10].

2.1.5 Modelos de regresión no paramétricos B-splines

El modelado de regresión no paramétrico ha recibido una atención considerable y se han propuesto muchos métodos de suavizado para extraer información de datos con una estructura no lineal compleja. Los B-splines constituyen un método atractivo para la estimación no paramétrica de una variedad de objetos estadísticos de interés, como la estimación de una media condicional, es decir, la "función de regresión" [11].

El uso de modelos de regresión no paramétricos B-splines estimados por el método de máxima verosimilitud penalizada son una alternativa viable para la modelización del comportamiento de una variable dependiente a partir de una independiente. Los puntos cruciales en el suavizado B-splines son las elecciones de un parámetro de suavizado y el número de funciones básicas, para las cuales se han realizado varios intentos utilizando la validación cruzada y el criterio de información de Akaike conocido como AIC [12].

Una spline es una función que se construye por partes a partir de funciones polinomiales. El término proviene del “listón elástico”, una herramienta utilizada por los constructores y delineantes navales para construir formas suaves que tienen las propiedades deseadas. Los dibujantes han hecho uso durante mucho tiempo de una tira flexible fijada en posición en varios puntos que se relaja para formar una curva suave que pasa por esos puntos [11]. Entonces los splines “son curvas polinómicas por trozos continuamente diferenciables hasta un orden prescrito” [13]. Entre los ejemplos más característicos se tiene el C^0, que es una spline lineal por trozos y el C^1 que es un spline cúbico. Previo a profundizar en la teoría que sustenta a las curvas splines conviene hacer una revisión de las curvas de Bézier.

2.1.5.1 Curvas de Bézier

Una curva Bézier de grado n (orden m) está compuesta de $m = n + 1$ términos y está dado por:

$$B(x) = \sum_{i=0}^{n} \beta_i \binom{n}{i} (1-x)^{n-1} x^i = \sum_{i=0}^{n} \beta_i \, B_{i,n}(x) \qquad 31)$$

Donde:

$$\binom{n}{i} = \frac{n!}{(n-i)!\, i!} \qquad (32)$$

La fórmula (31) se puede expresar recursivamente como:

$$B(x) = (1-x)\left(\sum_{i=0}^{n-1} \beta_i \, B_{i,n-1}(x)\right) + x\left(\sum_{i=1}^{n} \beta_i \, B_{i-1,n-1}(x)\right) \qquad (33)$$

De manera que una curva Bézier de grado *n* es una interpolación lineal entre dos curvas Bézier de grado *n* – 1 [11].

2.1.5.2 Nodos B-spline

Las curvas B-spline se componen de muchas piezas polinomiales y, por lo tanto, son más versátiles que las curvas Bézier. Considere $N + 2$ valores reales t_i, llamados nodos ($N \geq 0$ se llaman nudos interiores y siempre hay dos extremos, t_0 y t_{N+1}), con:

$$t_0 \leq t_1 \leq t_{N+1}$$

Cuando los nudos son equidistantes se dice que son uniformes, de lo contrario se dice que son no uniformes. Las curvas de Bézier poseen dos nodos de punto final, 98t_0 y t_1, y no hay nudos interiores, por consiguiente son un caso límite, es decir, un B-spline para el que $N = 0$.

2.1.5.3 Función base B-spline

Una función B-spline es la función de base interpolativa de máxima diferenciación. La B-spline es una generalización de la curva de Bézier. Los B-splines se definen

por su orden *m* y el número de nodos interiores *N* (hay dos puntos finales que son en sí mismos nodos, por lo que el número total de nudos será *N* + 2). El grado del polinomio B-spline será el orden de spline *m*-1 [11].

Sea $t = \{t_i \mid i \in \mathbb{Z}\}$ una secuencia de números reales no decrecientes ($t_i \leq t_{i+1}$) de manera que:

$$t_0 \leq t_1 \leq \ldots \leq t_{N+1}$$

Definir el nudo aumentado:

$$t_{-(m-1)} = t_0 \leq t_1 \leq \ldots 5t \leq t_N \leq t_{N+1} = \ldots = t_{N+m}$$

Donde se han agregado los nudos de límite superior e inferior t_0 y t_1 $n = m - 1$ veces (esto es necesario debido a la naturaleza recursiva del B-spline). Para cada uno de los nodos aumentados t_i, $i = 0, \ldots, N + 2m–1$, se define recursivamente un conjunto de funciones de valor real $B_{i,j}$ (para $j = 0, 1, \ldots, n$, siendo n los grados de la B-spline) como sigue:

$$B_{i,0}(x) = \begin{cases} 1 & Si\ t_i \leq x \leq t_{i+1} \\ 0 & En\ otro\ caso \end{cases} \qquad (34)$$

Para los cálculos 0/0 se define como 0.

Definiciones:

Empleando la notación de arriba:

- La secuencia *t* se conoce como secuencia de nodos y el término individual de la secuencia es un nodo.
- La función $B_{i,j}$ son llamados *i*-ésimas funciones base B-spline de orden *j* y la relación de recurrencia se llama relación de recurrencia de Boor.
- Dado cualquier entero *j* no negativo, el espacio vectorial $V_j(t)$ sobre $\mathbb{R}$, generado por el conjunto de todas las funciones básicas B-spline de orden *j*

se denomina B-spline de orden *j*. En otras palabras, el B-spline $V_j(t)$ = lapso $\{B_{i,j}(x) \mid i = 0, 1, \ldots\}$ sobre $\mathbb{R}$. Cualquier elemento de $V_j(t)$ es una función B-spline de orden *j*. El primer término $B_{0,n}$ es a menudo referido como intercepto.

Una B-spline de grado *n* (de orden spline $m = n + 1$) es una curva paramétrica compuesta de una combinación lineal de B-splines de base $B_{i,n}(x)$ de grado *n* dada por:

$$B(x) = \sum_{i=0}^{N+n} \beta_i \cdot B_{i,n}(x), \quad x \in [t_0, t_{N+1}] \tag{35}$$

La función *f* tiene la siguiente estructura:

$$f(x) = a_1 \cdot B_1(x) + \ldots + a_k \cdot B_k(x) \tag{36}$$

Donde:

k, número de bases.
$a_1, \ldots, a_k$, parámetros desconocidos, y
$B_1, \ldots, B_k$, funciones conocidas que dependen solamente de los nodos.

En este sentido, en la regresión Spline se puede establecer que un problema de regresión no paramétrico se reduce a un problema paramétrico, siendo requerida la estimación de los coeficientes $a_1, \ldots, a_k$ ajustando un modelo de regresión lineal. De acuerdo al tipo de base se tendrán diferentes tipos de regresión, como los B-splines [10].

Para el efecto se requieren seleccionar *M* nodos interiores $C_1, \ldots, C_M$ de manera que:

$$X_{min} < C_1 < \ldots < C_k \leq X_{max}$$

A continuación, en el Gráfico 2-1 se muestran las curvas B-spline de base para un determinado grado del polinomio, número de vértices del polígono de control y nudos internos.

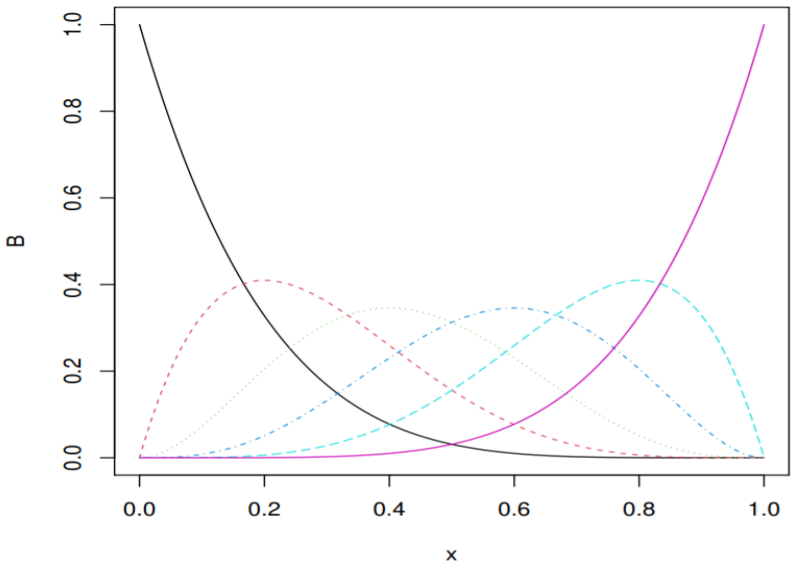

Gráfico 2-1. B-spline base para un modelo de regresión.

Fuente: [14].

La regresión basada en B-spline de orden *p* se considera de la estructura:

$$f(x) = a_0 + a_1 x + \ldots + a_p x^p + B_1 (x - C_1)^p + \ldots + B_k (x - C_k)^p \quad (37)$$

Donde:

a_0, …, a_p y B_1, …, B_k, coeficientes a determinar.

$$(x - t)^p = \begin{cases} (x - t)^p & ; si\ t > x \\ 0 & ; si\ t \leq x \end{cases} \quad (38)$$

La expresión (38) se denomina función potencia truncada de orden *p*. Por consiguiente, el spline *f* puede ser expresado como una combinación lineal de la forma:

$$f(x) = a_0 \cdot B_0(x) + \ldots + a_p \cdot B_p(x) + a_{p+1} \cdot B_{p+1}(x) + \ldots + a_{p+M} \cdot B_{p+M}(x) \quad (39)$$

En donde las funciones mostradas a continuación, constituyen una base de funciones polinómicas del spline:

$$B_0(x) = 1, B_1(x) = x, B_p(x) = x^p \quad (40)$$

$$B_{p+1}(x) = (x - C_1), \dots, B_{p+M} = (x - C_M) \quad (41)$$

De esa manera ya se tiene un modelo de regresión paramétrico, cuyos coeficientes $a_0, \dots, a_{p+M+1}$ se obtienen mediante mínimos cuadrados.

2.1.6 Bondad de ajuste

La bondad de ajuste de un modelo de regresión es el criterio que se emplea para valorar su capacidad de predicción de la variable respuesta. Para el efecto se mide la variabilidad explicada por el modelo sobre la variable a predecir y la forma de hacerlo es a través del uso de diferentes métricas, tales como el error estándar de los residuos (RSE), el test F de bondad de ajuste que es calculado en la tabla ANOVA y el coeficiente de determinación ajustado R^2 ajustado [15]. Acerca de los dos últimos, la descripción correspondiente consta en los numerales 2.1.3.1 y 2.1.3.3, respectivamente. De igual manera, las fórmulas de cálculo son la (17) y (28).

En el caso del estadístico F, las hipótesis del modelo son las siguientes:

Nula:

$$H_0: \textit{El modelo de regresión no explica bien la variable respuesta} \quad (42)$$

Alternativa:

$$H_1: \textit{El modelo de regresión explica bien la variable respuesta} \quad (43)$$

Una vez calculado el valor del estadístico F, se lo compara con el valor de F crítico obtenido de tablas para un determinado nivel de significancia α (generalmente igual

a 0.05). En caso de que el valor de F calculado sea mayor al de tablas, se rechaza la hipótesis nula y se acepta el modelo, esto implica que el p-valor < α.

Respecto al coeficiente de determinación R^2 ajustado, su valor oscila entre 0 y 1 ($0 \leq R^2 \leq 1$), mientras el coeficiente más se acerque a 1, representa que el modelo tiene buena bondad de ajuste. Se suele considerar que son aceptables valores en el intervalo entre 0.6 y 1. Sin embargo, el R^2 ajustado por sí solo no puede ser la métrica que determine la idoneidad de un modelo de regresión, sino que sirve como un indicador que se complementa con las otras métricas de bondad de ajuste y con el diagnóstico de los residuos del modelo [15].

En cuanto al error estándar de los residuos (RSE), es una métrica relativa a la escala de medida utilizada. En virtud de que el valor del RSE depende de las magnitudes de las variables involucradas, no se puede hacer una generalización en torno a un rango aceptable. No obstante, se espera que sea lo menor posible en proporción a los valores de los datos de la variable respuesta.

2.1.7 Análisis de residuos de la regresión y transformaciones

Una vez establecido un modelo de regresión y verificada la bondad de ajuste, corresponde hacer un análisis de los residuos, que consiste en verificar el cumplimiento de los denominados supuestos del modelo.

2.1.7.1 Análisis de los residuos

El análisis se centra en la aplicación de pruebas estadísticas destinadas a determinar el cumplimiento de determinadas hipótesis de partida respecto a los residuos de un modelo [15]. A continuación, se enlistan los supuestos que deben ser comprobados:

- El número de observaciones debe ser mucho mayor que el de parámetros β.
- Los residuos del modelo deben seguir una distribución normal $\epsilon \sim N(\mu, \sigma^2)$.

- Debe existir homocedasticidad de los residuos, es decir la esperanza matemática del error tiene que ser igual a cero $E(\epsilon) = 0$, con la varianza constante.
- Los residuos del modelo no deben estar correlacionados entre sí o lo que es lo mismo no debe haber autocorrelación.

Existen varias pruebas de hipótesis que son aplicables para cada supuesto. De la revisión de la literatura se desprende que las principales pruebas estadísticas para verificar la normalidad de los datos son entre otras: Shapiro-Wilk, Kolmogorov-Smirnov, Kolmogorov-Smirnov corregida por Lilliefors, Anderson-Darling, D'Agostino-Pearson, Jarque-Bera y Shapiro-Francia [16] [17] [18].

Para verificar la hipótesis de no autocorrelación de los residuos se suele emplear la prueba de Durbin-Watson [19]. Mientras que las pruebas aplicables para comprobar la homocedasticidad de los residuos son las siguientes: Breusch-Pagan, Levene, Bartlett, O'Brien, Brown y Forsythe, Fligner-Killeen, y White [20].

Adicionalmente, en la evaluación de la idoneidad de los modelos se suelen emplear gráficos, particularmente son de interés los que muestran información de los residuos. Entre los gráficos más usualmente utilizados constan: Valores ajustados (predichos) vs residuos, cuantiles teóricos vs residuos estandarizados (gráfico Q-Q Plot), valores ajustados vs raíz cuadrada de los residuos estandarizados y el grado de influencia (Leverage) vs residuos estandarizados [21].

2.1.7.2 Problemas respecto a los supuestos del modelo

En caso de que no se cumplan los supuestos acerca de los residuos de un modelo, éste será válido, pero no se podrán establecer intervalos de confianza al 95% para el modelo de regresión [15]. Para que se cumplan los supuestos de un modelo de regresión se puede considerar la eliminación de los outliers (valores atípicos) de las variables regresoras y/o la transformación de una de las variables (generalmente la variable a predecir o respuesta).

2.1.7.3 Transformaciones de Box-Cox

En las situaciones en las que los supuestos de los residuos de un modelo de regresión se incumplen gravemente, se pueden tomar algunas decisiones, que van desde aceptar el modelo a pesar del incumplimiento de todos los supuestos, hasta diseñar un nuevo modelo que tenga aspectos importantes del modelo original y satisfaga todos los supuestos. En el último caso, es útil optar por una transformación de los datos. Las más aplicadas son las transformaciones paramétricas de la potencia propuestas por Box-Cox [22], que presentan algunas versiones, como las siguientes:

$$y_i{}^{\lambda} = \begin{cases} y_i{}^{\lambda}; & \lambda \neq 0 \\ \ln y_i; & \lambda = 0 \end{cases} \qquad (44)$$

Donde:

λ, es un exponente que minimiza la desviación estándar de una variable transformada estandarizada, para el caso de la función (44) se requiere que sea un valor conocido. En caso de que se desconozca su valor, se puede emplear una forma más compleja de la función. El logaritmo natural puede ser sustituido por un logaritmo de cualquier base.

Para tener en cuenta la discontinuidad de $\lambda = 0$, la función (44) se puede expresar como sigue:

$$y_i{}^{\lambda} = \begin{cases} \dfrac{y_i{}^{\lambda} - 1}{\lambda}; & \lambda \neq 0 \\ \ln y_i; & \lambda = 0 \end{cases} \qquad (45)$$

Manly [22] sugirió otra alternativa que puede utilizarse con observaciones negativas y que se afirma que es eficaz para convertir las distribuciones unimodales sesgadas en distribuciones casi simétricas de tipo normal y es de la forma:

$$y_i{}^{\lambda} = \begin{cases} \dfrac{e^{\lambda y_i} - 1}{\lambda}; & \lambda \neq 0 \\ y_i; & \lambda = 0 \end{cases} \qquad (45)$$

La aplicación de las transformaciones de Box-Cox en los modelos de regresión polinomiales, determinan que la forma general de los modelos aplicando la función (44) sea de la siguiente manera:

$$y_i{}^{\lambda} = \beta_0 + \beta_1 x_i + \beta_2 x_i{}^2 + \ldots + \beta_k x_i{}^k + \epsilon_i \ , \quad i = 1, \ldots, n. \qquad (46)$$

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i{}^2 + \ldots + \beta_k x_i{}^k + \epsilon_i \ , \quad i = 1, \ldots, n. \qquad (47)$$

2.1.8 Programas estadísticos

La generación de modelos de regresión se suele realizar mediante el empleo de un software estadístico con la capacidad de efectuar los cálculos respectivos para hallar los coeficientes de los modelos, las métricas de error, los intervalos de confianza, las pruebas de hipótesis y los gráficos correspondientes. Existen varias opciones de softwares y hojas de cálculo con capacidad de establecer modelos de regresión paramétricos polinomiales. No obstante, el número de los softwares que pueden ser utilizados para trabajar con modelo de regresión no paramétricos es menor.

Entre los principales aspectos que se consideran a la hora de escoger uno en particular se destacan: el tipo de licencia de uso, la funcionalidad y la reputación que tienen entre los usuarios estadísticos. En este sentido, se consideran las opciones de softwares y lenguajes de programación de la Tabla 2-2 como las más viables.

De entre las opciones presentadas en la Tabla 2-2, se identifica que los lenguajes de programación R y Pyhton destacan de las demás, por ser de uso libre, por tener mayor difusión entre los estadísticos y científicos de datos y dado que poseen la posibilidad de realizar los dos tipos de modelos de regresión de interés. Se puede utilizar cualquiera de las dos herramientas indistintamente, pero el usuario debe

decidirse en función de su familiaridad con la programación. Python suele ser más fácil de aprender para los usuarios informáticos, mientras que R para quienes son estadísticos o ingenieros.

Tabla 2-2. Softwares y lenguajes estadísticos.

Software IDE*	Lenguaje	Licencia	Tipo de propósito	Orden por número de usuarios**
MATLAB	M	Propietario	Cálculo numérico	13
RStudio	R	Licencia Pública General de GNU	Estadístico	5
Spyder Jupyter Notebook	Pyhton	Licencia libre permisiva PSF 2.5	General (librerías: Pandas, Numpy, Scipy, Matplotlib, Seaborn)	2
SAS Studio	SAS	Propietario	Estadístico	7
SPSS	Java	Propietario	Estadístico	14

* IDE = Entorno de desarrollo integrado.
** Ranking de lenguajes de programación más demandados en Ciencia de Datos en función de plazas laborales y número de usuarios. R Bloggers (2017). Data Science Job Report 2017 [23].

Fuente: [23] [24].

Elaborado por: Toalombo, B. (2021).

2.1.9 Aplicaciones de los modelos de regresión en ingeniería

Las aplicaciones de los modelos de regresión son numerosas, básicamente se dirigen a la predicción del comportamiento de los datos a partir de determinadas variables regresoras. O bien se utilizan para la clasificación de datos, esto último se utiliza en Machine Learning (aprendizaje automático) y suele ser considerado como algo ajeno a la regresión, pero se debe tener en cuenta que el modelo establece una curva, que por sí misma es la que permite hacer la clasificación.

En el caso de su uso para la explicación de fenómenos en la ingeniería, se pueden destacar algunos casos, tales como: estimación del desplazamiento sísmico del terreno en ingeniería geológica [25], descripción del comportamiento de variables

climatológicas [10], estudio de la aerodinámica no lineal [26], investigaciones de niveles de mortalidad de una población y mapeo de enfermedades [27], control de puentes y estructuras en ingeniería civil [28], recalibración de dispositivos de seguimiento ocular [29], modelado y animación de las expresiones faciales en aprendizaje automático [30], exploración de las relaciones entre la velocidad del impacto del vehículo y la gravedad de las lesiones ocasionadas sobre las personas [31], entre otras.

En el presente estudio la atención se centra en el uso de modelos de regresión paramétrico polinomiales y no paramétricos B-spline, para explicar el comportamiento de las variables inherentes al impacto de vehículos y de las variables climatológicas. Por esta razón a continuación se abordan la revisión del estado del arte respectiva.

2.1.9.1 Simulación de impacto de vehículos

Los accidentes de tránsito representan un problema que atenta contra la salud de las personas llegando al punto de ocasionar altos niveles de mortalidad. Los principales factores que inciden en la ocurrencia de este tipo de acontecimientos son entre otros la impericia de los conductores y las condiciones de las vías. Ante tal situación los fabricantes de automotores y de autobuses no pueden intervenir directamente, no obstante, están en capacidad de minimizar la afectación de los accidentes en las personas involucradas, a través de la fabricación de componentes que brinden buena resistencia al impacto y a la deformación. Esto constituye un aspecto inherente a la ingeniería mecánica automotriz y particularmente dentro del campo de la dinámica, ya que en el diseño de los vehículos se consideran aspectos como la forma y los materiales de los componentes automotrices.

En el análisis estático y dinámico de una colisión intervienen algunas magnitudes físicas tales como: velocidad lineal, fuerza, esfuerzo, tiempo de impacto, factor de seguridad de diseño (FDS), entre otras. El FDS es el cociente calculado entre el esfuerzo de cedencia de un material y el esfuerzo dado por las condiciones a las que está sometido el diseño [32]. En el presente caso, los valores del FDS están

representados de manera inversa, es decir indican la proporción de sobredimensionamiento del material requerida en función de la velocidad a la que recibe el impacto la carrocería del autobús. Por ejemplo, un valor de FDS de 2 indica que, en el diseño y selección de los materiales para la estructura de la carrocería, se debe considerar un sobredimensionamiento de 2 para asegurarse que el diseño brinde resistencia al impacto recibido.

Bajo el contexto indicado, los fabricantes de vehículos y de autobuses suelen implementar herramientas tecnológicas para el análisis de las colisiones de vehículos. La simulación de impacto suele ser una de las herramientas indispensables, que consiste en la utilización de un software de Elementos de Análisis Computarizado (CAE) para representar un choque real. De esta manera se puede conocer el efecto que provocan en las estructuras de los automotores las cargas de impacto, concretamente esto se traduce en estados de tensiones y deformaciones [33].

Desde una visión más amplia, las acciones para evitar la afectación a las personas como consecuencia de un accidente de tránsito no solamente dependen de los fabricantes de los automotores, sino también de los conductores y de las autoridades. En este punto, se destaca que la velocidad de circulación de los vehículos es un aspecto muy importante, pues no se experimenta el mismo resultado en un siniestro de tránsito a 35 km/h que en uno a 90 km/h [34]. Una de las principales medidas que pueden adoptar los conductores es conducir a velocidades moderadas. En este sentido, el presente estudio considera el análisis del efecto que produce la velocidad en las otras magnitudes físicas relacionadas a la dinámica del movimiento de vehículos.

2.1.9.2 Parámetros climatológicos

Los parámetros climatológicos suelen presentar un comportamiento estocástico a lo largo del tiempo, por lo que resulta complicado establecer modelos explicativos de su comportamiento. Sin embargo, a partir de la disponibilidad de una base de datos robusta obtenida de un registro de un historial de información es posible

modelar el comportamiento con un nivel de certeza aceptable. Los parámetros climatológicos son diversos y las relaciones que se pueden conseguir mediante un modelo de regresión pueden ser algunas, aunque existe cierto grado de dificultad en la determinación de cuales son las variables explicativas o independientes, por lo que en primer lugar se puede determinar la existencia de una correlación y de ahí hallar un modelo de regresión. De entre las variables climatológicas se pueden destacar la temperatura ambiente, humedad relativa, presión atmosférica, radiación solar, velocidad de viento y precipitación. De entre las variables indicadas, la temperatura suele ser una de las que más interés tiene en la población puesto que interfiere directamente en la sensación de confort térmico de las personas, en el caso ecuatoriano ésta suele ser distinta según la región territorial, siendo que en la región Sierra suele oscilar de media entre 13.7 y 18.7ºC [35].

Otro aspecto de interés es el estudio de los contaminantes atmosféricos, relacionados con la presencia de enfermedades y la merma de la calidad de vida de las personas. De entre los contaminantes más reconocidos constan el monóxido de carbono (CO), el dióxido de azufre (SO_2), el dióxido de nitrógeno (NO_2) y el material particulado PM_{10}. En la práctica no se suelen hallar relaciones directas entre las variables, pero haciendo uso de bases de datos grandes es posible modelar alguna relación existente [36].

CAPÍTULO III

MARCO METODOLÓGICO

3.1 Metodología

La realización del presente proyecto de desarrollo demanda la aplicación de una metodología de investigación, orientada a establecer la manera en que se lleva a cabo la consecución de los objetivos planteados, así como el proceso de recopilación y procesamiento de los datos, en función de las variables de estudio. En este sentido, se optó por la modelación matemática mediante la utilización de modelos de regresión paramétricos y no paramétricos, con la finalidad de describir el comportamiento de las variables inherentes al impacto de vehículos en movimiento y de los parámetros climatológicos.

3.2 Equipos y materiales

Para el desarrollo de la investigación se emplearon los siguientes equipos, materiales y recursos digitales:

Tabla 3-1. Equipos y materiales.

Equipo / material	Detalle
Computadora	Equipo informático utilizado para el procesamiento de los datos y la redacción del informe escrito.
Software CAD-CAE	Programa destinado al diseño y simulación de un evento de impacto vehicular.
Software estadístico R	Programa de acceso libre especializado en el manejo y procesamiento de datos, mediante el uso de códigos.
Libros y artículos digitales	Recursos disponibles en internet que sirven de referente y fuente de consulta para el desarrollo de la investigación.
Materiales de oficina	Implementos empleados para hacer apuntes y llevar el registro de la información utilizada.

Elaborado por: Toalombo, B. (2021).

3.3 Tipo de investigación

Desde el punto de vista del alcance, la investigación es de tipo correlacional y explicativa, en vista de que se establecen relaciones de correlación bivariada y de causalidad entre las variables de interés.

De acuerdo al enfoque, la investigación es de tipo cuantitativa, porque la información corresponde a datos numéricos discretos y continuos, que sirven para medir las variables de la investigación y que permiten establecer las relaciones existentes entre las mismas.

Según la modalidad, la investigación es bibliográfica y de campo, dado que se recurre a la revisión de publicaciones académicas en revistas y textos inherentes al campo de la Matemática Aplicada. En cuanto a la obtención de los datos, para el caso de la simulación de impacto vehicular se lo realizó mediante el uso de un software CAD-CAE y los datos de las variables climatológicas fueron recopilados en la estación meteorológica de San Antonio de Pichincha, perteneciente al Distrito Metropolitano de Quito. Los datos se muestran en los Anexos A y B del presente documento, respectivamente.

El diseño de la investigación es no experimental de corte transversal o transeccional, debido a que los datos fueron obtenidos en las condiciones regulares de ocurrencia de los eventos y fueron analizados en una sola oportunidad para obtener modelos de regresión explicativos de los fenómenos de ingeniería considerados.

3.4 Prueba de hipótesis

La hipótesis de la investigación se centra en la existencia de diferencias significativas entre el ajuste obtenido mediante los modelos de regresión paramétricos polinomiales y los de los modelos de regresión no paramétricos B-splines, considerando algunos casos de aplicación en ingeniería. El planteamiento de las hipótesis es el siguiente:

3.4.1 Hipótesis nula

No existe diferencia significativa entre las distancias de un modelo de regresión polinómico y un modelo B-spline en las variables climatológicas y en las variables inherentes al impacto de vehículos.

$$H_0: \tilde{x}_{rp} = \tilde{x}_{rbs} \tag{48}$$

3.4.2 Hipótesis alterna

Existe diferencia significativa entre las distancias de un modelo de regresión polinómico y un modelo B-spline en las variables climatológicas y en las variables inherentes al impacto de vehículos.

$$H_1: \tilde{x}_{rp} \neq \tilde{x}_{rbs} \tag{49}$$

En el presente caso, la prueba estadística que se utiliza para la verificación de las hipótesis planteadas es la prueba no paramétrica de Wilcoxon, en virtud de que las diferencias entre las distancias no son normales.

Adicionalmente, la determinación de la validez de los modelos de regresión y el establecimiento de intervalos de confianza llevan implícitos las siguientes pruebas de hipótesis:

- Prueba de rechazo de la nulidad de los parámetros β´s del modelo polinomial y de los coeficientes del modelo B-spline, mediante el estadístico t de Student $\beta' s \neq 0$.
- Prueba de diferencias significativas para la validez de los modelos, mediante la prueba F de Snedecor presentada en la tabla ANOVA.
- Pruebas de distribución normal de los residuos de los modelos de regresión $\epsilon \sim N(\mu, \sigma^2)$, para medir la bondad de ajuste del modelo. Las pruebas aplicadas son: Shapiro-Wilk, conjuntamente con la de Kolmogorov-Smirnov corregida por Lilliefors.

- Pruebas de Durbin-Watson para la independencia de las variables (no autocorrelación) y prueba de homocedasticidad de Breusch-Pagan.

3.5 Población y muestra

La población corresponde a los datos disponibles de las variables involucradas en determinadas aplicaciones en ingeniería. Para la presente investigación se consideró el caso particular del estudio de los resultados de una simulación de impacto vehicular y los parámetros climatológicos. En este sentido, se estableció una muestra, que corresponde a una porción del universo de datos existentes. En la Tabla 3-2 se detalla la muestra considerada en la investigación:

Tabla 3-2. Muestra utilizada en el estudio.

Aplicación	Variable	Tamaño de muestra	Archivo de datos
Simulación de impacto vehicular	Velocidad (km/h)	51	Impacto.csv (Anexo A)
	Fuerza (N)	51	
	Factor de seguridad FDS	51	
	Tiempo de impacto (s)	51	
	Deformación (mm)	48	
Parámetros climatológicos	Radiación solar (W/m^2)	605	Rad_Temp.csv
	Hora del día	24	Climatologicos.csv (Anexo B)
	Temperatura (ºC)	605 24	Rad_Temp.csv Climatologicos.csv (Anexo B)
	Humedad relativa (%)	24	Climatologicos.csv (Anexo B)
	Presión atmosférica (mbar)	24	Climatologicos.csv (Anexo B)

Elaborado por: Toalombo, B. (2021).

3.6 Recolección de información

Para la medición de las variables climatológicas, la Secretaría del Ambiente del Distrito Metropolitano Quito dispone de equipos e instrumentos de medición. Los datos son presentados de manera independiente para cada una de las variables en formato .xlsx, con información de la fecha, hora y estación meteorológica a la que corresponden los datos. Se consideró la información generada en la estación meteorológica de San Antonio de Pichincha, localizada en las coordenadas Latitud: 0.00617º Sur y Longitud: 78.4355º Oeste. Los datos son de libre acceso y están disponibles en la página de la Secretaría del Ambiente del Distrito Metropolitano Quito, cuyo link de acceso es:

http://www.quitoambiente.gob.ec/index.php/descarga-datos-historicos

Por otra parte, los datos de las variables que intervienen en la simulación de un impacto o choque de un vehículo con un autobús fueron obtenidos directamente por parte del autor de la presente investigación. Los datos fueron generados en un software CAD-CAE y exportados a un archivo .xlsx.

La información presentada y los datos utilizados en el desarrollo de la presente investigación son originales y auténticos. Esto garantiza que los resultados son reales y que éstos representan las condiciones reales de los problemas estudiados.

3.7 Procesamiento de la información y análisis estadístico

El procesamiento de la información demandó de la aplicación de una serie de fases o pasos a seguir, conforme se describe a continuación:

3.7.1 Tipos de datos

En primer lugar, se identificaron los tipos de datos. Los datos disponibles son de tipo numérico discreto (velocidad y hora del día) o numérico continuo (para todos los otros casos). Se aclara que la velocidad de desplazamiento del auto, se considera

como una variable numérica discreta y no continua, en vista que se planteó una situación en la que se simuló la situación de impacto a diferentes velocidades de choque, con incrementos de una unidad en un rango que fluctúa entre 45 y 95 km/h.

Los datos tienen unidades de medición específicas por tratarse de variables físicas, de acuerdo a lo indicado en la Tabla 3-2.

3.7.2 Depuración y limpieza de los datos

Previo a la realización de los análisis correspondientes para la obtención de los modelos de regresión, se efectuó una clasificación, depuración y limpieza de los datos. Para el efecto se utilizó una hoja de cálculo y el software R. En el presente caso se efectuaron las siguientes actividades para la depuración de datos:

- Elaboración de archivos .csv que contienen los datos de las variables utilizadas en los modelos de regresión. Los datos de cada variable se agrupan por columna.
- Identificación de los posibles modelos que se pueden establecer entre las variables existentes, para lo cual se utilizó el software R (se consideraron potencialmente relacionadas las variables que tenían una correlación superior a 0.6).
- Verificación de ausencia de datos en blanco y/o erróneos. La idea fue evitar que existan datos faltantes para todas las parejas de las dos variables a ser consideradas en los modelos de regresión.
- Eliminación de datos aberrantes en las bases de datos. Se suprimieron aquellos que estaban muy fuera de rango respecto a los demás, para lo cual se emplearon diagramas de cajas con la finalidad de identificar los valores que podrían distorsionar los resultados. Para los datos de la radiación solar y temperatura ambiente también se eliminaron los datos atípicos.

3.7.3 Preparación del código

Se realizó la programación de un script con una codificación destinada a establecer los modelos de regresión polinomial y B-splines, tomando en cuenta el lenguaje de código correspondiente y la teoría estadística inherente a los modelos referidos. En este caso se utilizó el lenguaje de programación estadística R, por ser de uso libre y porque dispone de paquetes y funciones especializadas para trabajar con modelos de regresión. Los códigos correspondientes a la obtención de los modelos de regresión polinomiales, B-splines y las gráficas conjuntas de los dos modelos se presentan en los Anexos C, D y E, respectivamente.

3.7.4 Supuestos del modelo de regresión paramétrico polinomial

Los supuestos a ser verificados para el modelo de regresión paramétrico polinomial fueron los siguientes:

- El número de observaciones *n* es mucho mayor que el de parámetros β.
- Homocedasticidad, que implica que la esperanza matemática del error sea 0. $E(\epsilon) = 0$ con varianza constante, a través de aplicar la prueba de Breusch-Pagan estudiantilizada.
- Los residuos del modelo no están correlacionados entre sí, mediante la prueba de Durbin-Watson.

3.7.5 Prueba de normalidad de los residuos

Con la finalidad de corroborar los supuestos de partida de los modelos paramétricos (regresión polinomial), se aplicaron pruebas de normalidad de los residuos con un nivel de significancia de 0.05. En caso de que el p-valor calculado a partir de los datos sea mayor a 0.05 implica que la distribución de los residuos es normal:

$$p - valor > 0.05 \ \therefore \ \epsilon \sim N(\mu, \sigma^2)$$

En caso contrario ($p - valor \leq 0.05$), los residuos no se distribuyen normalmente y por consiguiente no se cumplen los supuestos iniciales del modelo de regresión polinomial, por lo que un modelo con dichas características se consideró como válido, pero sin intervalos de confianza al 95%. Las pruebas se aplicaron de manera independiente para cada modelo de regresión.

La teoría indica que en función del tipo de distribución de los datos y del tamaño de la muestra se deben seleccionar la o las pruebas correspondientes. Sin embargo, la prueba de normalidad que ofrece el mayor poder de detección es la de Shapiro-Wilk [37] [16]. Por esta razón es la que se utilizó en el estudio realizado y para corroborar los resultados se empleó la prueba de Kolmogorov-Smirnov corregida por Lilliefors [38].

Complementariamente se obtuvieron gráficos cuantil-cuantil (Q-Q Plot) para corroborar los resultados de las pruebas de normalidad. En caso de que todos los puntos estén muy próximos a la línea recta representa que los residuos se distribuyen normalmente, caso contrario significa que no tienen distribución normal y no se establecen intervalos de confianza para el modelo de regresión paramétrico polinomial. El mismo análisis se realiza para el modelo de regresión B-spline, aunque en este caso el cumplimiento de los supuestos no es una condición de obligatoriedad para el establecimiento de los intervalos de confianza.

3.7.6 Prueba de diferencias para validar el modelo

Se aplicó la prueba no paramétrica de Wilcoxon para diagnosticar las diferencias entre las distancias de los modelos de regresión polinómicos y los modelos B-splines. De esta manera se determinó si los dos modelos tienen la misma capacidad de definir la relación existente entre las dos variables (p-valor >0.05), o si por el contrario uno de los dos modelos define la relación de mejor manera (p-valor ≤ 0.05).

Adicionalmente se hizo uso de los diagramas de cajas para representar la distribución de los datos de los residuos de los modelos de regresión obtenidos. De

esta forma se puede visualizar las diferencias de ambos modelos, así como también se representan gráficos cuantil-cuantil (Q-Q Plot).

3.7.7 Generación del modelo

Una vez preparados los datos y teniendo en cuenta las condiciones requeridas para la validación de los modelos de regresión, se procedió a determinar los modelos de regresión polinomial y B-splines que se ajustaron a los datos disponibles en cada caso, haciendo uso del software estadístico. En el presente caso se efectuaron las siguientes actividades en el script previamente elaborado:

- Importación y lectura de los datos de los archivos .csv.
- Inspección de los posibles modelos de regresión que se podrían establecer entre las variables de la base de datos, mediante la generación de gráficos de dispersión de puntos y la información de los coeficientes de correlación.
- Preparación de los datos, mediante filtración de las columnas en donde constan las variables independientes (predictoras) y dependientes (a predecir) que son de interés.
- Asignación de nombres para las variables, la predictora se denominó con “x” y la variable a predecir con “y”.

A continuación, para la obtención del modelo de regresión polinomial:

- Carga de las librerías adicionales para las funciones especiales de R.
- División del set en datos de entrenamiento (80%) y de prueba (20%).
- Ajuste mediante modelo de regresión polinomial usando el comando "poly".
- Prueba de diferencias significativas ANOVA para obtener el grado más apropiado de la regresión polinomial.
- Asignación del grado del polinomio.
- Obtención de los parámetros del modelo (intercepto, coeficientes, desviación estándar, t de Student para comprensión de la no nulidad de los coeficientes, significancia, error estándar residual, grados de libertad, F, coeficiente de determinación, coeficiente de determinación ajustado, p-valor).

- Obtención de la tabla ANOVA para validar el modelo.
- Estimación de los valores predichos por el modelo, intervalo de confianza, error estándar y residuos.
- Determinación de las métricas de error del modelo (coeficiente de determinación, coeficiente de correlación, media cuadrática del error, suma de cuadrados del error, raíz de la media cuadrática del error y suma de cuadrados totales).
- Verificación del cumplimiento de los supuestos de normalidad, no autocorrelación de los residuos del modelo y homocedasticidad, mediante la aplicación de las pruebas estadísticas de Shapiro-Wilk, Kolmogorov-Smirnov corregida por Lilliefors, Durbin-Watson y Breusch-Pagan. Además se generan los gráficos de valores ajustados (predichos) vs residuos, cuantiles teóricos vs residuos estandarizados (gráfico Q-Q Plot), valores ajustados vs raíz cuadrada de los residuos estandarizados y el grado de influencia (Leverage) vs residuos estandarizados.
- Generación del gráfico de dispersión de puntos y de la curva del modelo de regresión conjuntamente con las curvas del intervalo de confianza al 95% (esto último en caso de que se cumplan los supuestos del modelo).

Para el caso del modelo de regresión no paramétrico B-spline:

- Carga de las librerías adicionales para las funciones b-spline de R.
- Asignación del grado de la curva B-spline.
- Asignación del número de puntos de control o vértices del polígono de control.
- Generación del B-spline base para el modelo de regresión.
- Ajuste mediante modelo de regresión B-spline usando el comando "bs".
- Obtención de los parámetros del modelo (intercepto, coeficientes, desviación estándar, t de Student para comprensión de la no nulidad de los coeficientes, significancia, error estándar residual, grados de libertad, F, coeficiente de determinación, coeficiente de determinación ajustado, p-valor).
- Obtención de la tabla ANOVA para validar el modelo.

- Estimación de los valores predichos por el modelo, intervalo de confianza, error estándar y residuos.
- Determinación de las métricas de error del modelo (coeficiente de determinación, coeficiente de correlación, media cuadrática del error, suma de cuadrados del error, raíz de la media cuadrática del error y suma de cuadrados totales).
- Verificación del cumplimiento de los supuestos de normalidad, no autocorrelación de los residuos del modelo y homocedasticidad, mediante la aplicación de las pruebas estadísticas de Shapiro-Wilk, Kolmogorov-Smirnov corregida por Lilliefors, Durbin-Watson y Breusch-Pagan. También se generan los gráficos de valores ajustados (predichos) vs residuos, cuantiles teóricos vs residuos estandarizados (gráfico Q-Q Plot), valores ajustados vs raíz cuadrada de los residuos estandarizados y el grado de influencia (Leverage) vs residuos estandarizados.
- Generación del gráfico de dispersión de puntos y de la curva del modelo de regresión conjuntamente con las curvas del intervalo de confianza al 95% (esto último en caso de que se cumplan los supuestos del modelo).

Con la finalidad de comparar la bondad de ajuste de los dos modelos obtenidos para cada relación se efectuaron las siguientes actividades:

- Aplicación de la prueba de Wilcoxon para identificar si existen diferencias significativas entre ambos modelos (polinomial y B-splines).
- Generación de un gráfico Q-Q Plot para identificar las diferencias entre los intervalos de confianza de ambos modelos, con la finalidad de corroborar los resultados de la prueba de Wilcoxon.
- Elaboración de los diagramas de cajas de la variabilidad de la longitud del intervalo de confianza de los dos modelos.
- Construcción de los dos gráficos (modelo polinomial y B-spline) de la dispersión de puntos, curvas e intervalos de confianza en una sola gráfica general.
- Decisión de la aceptación o rechazo de los modelos de regresión obtenidos.

- Declaración del modelo más idóneo o que exprese de mejor manera la relación entre las variables.

3.8 Variables respuesta o resultados alcanzados

Las variables consideradas como parte de los problemas de ingeniería que fueron explicados mediante los modelos de regresión polinomial y B-splines se exponen en la Tabla 3-3, mostrada como sigue:

Tabla 3-3. Variables y características.

Aplicación	Variable	Tipo	Rol en el modelo
Simulación de impacto vehicular	Velocidad (km/h)	Numérica discreta	Regresora
	Fuerza (N)	Numérica continua	Regresora y Respuesta
	Factor de seguridad (adimensional)	Numérica continua	Respuesta
	Tiempo de impacto (s)	Numérica continua	Respuesta
	Deformación (mm)	Numérica continua	Respuesta
Parámetros climatológicos	Radiación solar (W/m^2)	Numérica continua	Regresora
	Hora del día	Numérica discreta	Regresora
	Temperatura (ºC)	Numérica continua	Respuesta
	Humedad relativa (%)	Numérica continua	Respuesta
	Presión atmosférica (mbar)	Numérica continua	Respuesta

Elaborado por: Toalombo, B. (2021).

En el caso de la simulación de impacto vehicular, la variable regresora velocidad es no estocástica, en tanto que la fuerza es estocástica. Por su parte, en el caso de los fenómenos climatológicos, la variable regresora hora del día es no estocástica y la variable radiación solar sí lo es.

Todas las variables dependientes o de respuesta son estocásticas. Por este motivo, en lo que respecta a las variables climatológicas, los datos considerados en el

análisis corresponden a valores promedios obtenidos a partir de un volumen de datos superior a 32000. De esta manera se minimizó el sesgo producto de la aleatoriedad de los datos.

Las relaciones que establecen los modelos de regresión polinomial y B-splines en el presente estudio son:

- Velocidad vs fuerza.
- Velocidad vs FDS.
- Velocidad vs tiempo de impacto.
- Fuerza vs deformación.
- Radiación solar vs temperatura ambiente.
- Hora del día vs temperatura ambiente.
- Hora del día vs humedad relativa.
- Hora del día vs presión atmosférica.

Las cuatro primeras relaciones corresponden al fenómeno de la simulación de impacto vehicular y las cuatro restantes a las variables climatológicas. También se trabajó sobre la relación existente entre los contaminantes CO, SO_2, PM_{10} y NO_2, pero no se hallaron relaciones significativas, motivo por el cual no consta información al respecto en el análisis de los resultados.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Análisis de resultados

A partir de los datos disponibles para las dos aplicaciones de ingeniería consideradas en el presente trabajo, se establecieron los modelos de regresión polinomial y B-splines, siguiendo los lineamientos establecidos en el marco metodológico. Se hace una comparación directa entre ambos modelos de regresión con el fin de determinar en que condiciones una u otra alternativa resulta de utilidad, así como las limitaciones que presentan en función de la bondad de ajuste (pruebas de hipótesis y métricas de error), residuos e intervalos de confianza.

4.1.1 Simulación de impacto vehicular

Las variables inherentes a la simulación de impacto de un automóvil con un autobús son: velocidad, fuerza, factor de seguridad (FDS), tiempo de impacto y deformación. La simulación consistió en representar el impacto de la parte frontal de un auto contra la parte posterior de la carrocería de un autobús, con el objeto de determinar la afectación que experimenta la estructura de la carrocería. Se planteó una situación de desplazamiento del auto a diferentes velocidades de choque, en un rango que fluctúa entre 45 y 95 km/h, con incrementos de una unidad, de manera que se dispone de 51 observaciones por cada variable.

Para llevar a cabo la simulación se utilizó un software CAD-CAE, siendo necesario el diseño del vehículo y de la carrocería del bus de acuerdo a las dimensiones reales y mediante asignación de los materiales correspondientes a ambas estructuras. De esta manera se representó el fenómeno en las condiciones reales. A continuación, se presenta una captura de pantalla del proceso de simulación del problema planteado en el software:

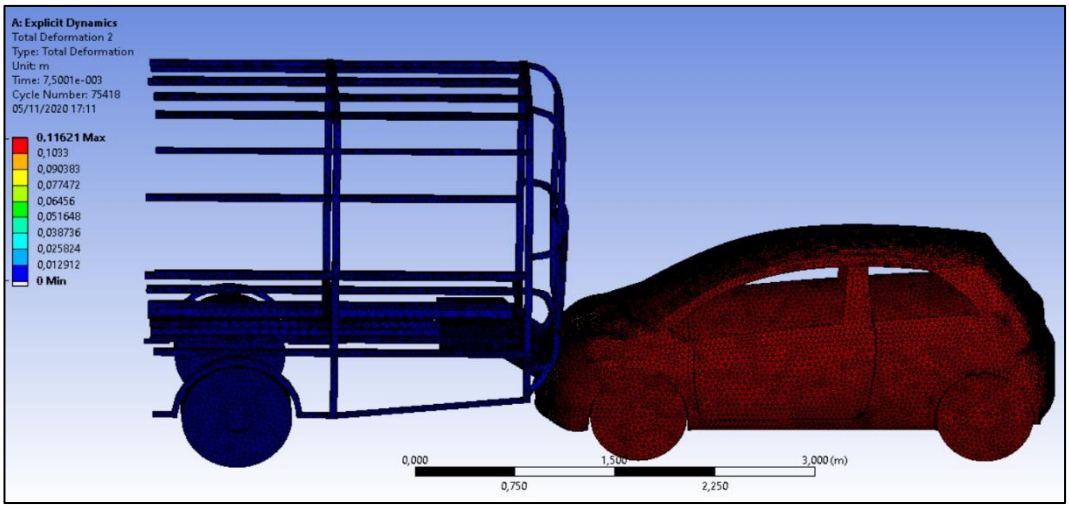

Figura 4-1. Simulación del impacto del vehículo contra la carrocería del autobús.

Elaborado por: Toalombo, B. (2021).

Posterior a la simulación del fenómeno se procedió a recopilar los datos y a depurarlos para efectuar el análisis correspondiente. Se establecieron un total de ocho modelos de regresión para cuatro relaciones entre las variables (en todos los casos se consideró una variable regresora y una de respuesta). La base de datos utilizada se denomina "Impacto.csv", la cual se muestra en el Anexo A del presente documento.

Para la presentación de los resultados se emplean tablas que detallan la información concerniente a: las variables explicativas y regresoras, número de datos procesados, tipo de modelo de regresión, grado del polinomio, número de vértice del polígono de control (modelo B-spline), intercepto y coeficientes del modelo, prueba de hipótesis F, p-valor, RSE, R^2 ajustado, MSE, SSE, RSME, SST, pruebas de normalidad de los residuos (estadístico y p-valor), prueba de residuos no correlacionados, prueba de homocedasticidad de residuos, prueba de diferencias entre los dos modelos y los códigos de significancia. Todas las tablas se subdividen de forma vertical en dos partes, una para el modelo de regresión polinomial y otra para el B-spline, de modo que visualmente se puede comparar directamente las características de ambos modelos y sobre todo para identificar la bondad de ajuste de cada uno. A continuación, se presenta el desarrollo de cada caso:

4.1.1.1 Velocidad vs Fuerza

Se utilizaron 51 datos de la velocidad de desplazamiento del auto pequeño y la correspondiente fuerza generada por el impacto del mismo contra la estructura de la carrocería del autobús. En la Tabla 4-1 se presentan los modelos de regresión polinomial y B-spline de la variable velocidad versus la variable fuerza:

Tabla 4-1. Modelos de regresión de la velocidad y fuerza.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Velocidad (km/h)	
Respuesta:		Fuerza (N)	
Número de datos *n*:		51	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	8	Grado:	5
		Vértices del polígono de control:	7
Intercepto β_0	5.789831 *x* 10^{15}	Intercepto a_0	2.425 *x* 10^{12}
β_1	-7.329609 *x* 10^{14}	a_1	-2.435 *x* 10^{12}
β_2	4.027116 *x* 10^{13}	a_2	-2.379 *x* 10^{12}
β_3	-1.254354 *x* 10^{12}	a_3	-2.547 *x* 10^{12}
β_4	24228335129	a_4	-2.152 *x* 10^{12}
β_5	-297211927	a_5	-2.802 *x* 10^{12}
β_6	2261648	a_6	-2.092 *x* 10^{12}
β_7	-9762.751	a_7	-2.679 *x* 10^{12}
β_8	18.30744	a_8	0
F	21730	F	4941
p-valor	0.000***	p-valor	0.000***
RSE	7.676 *x* 10^9	RSE	1.72 *x* 10^{10}
R^2 ajustado	0.9997124	R^2 ajustado	0.998556
MSE	5.892849 *x* 10^{19}	MSE	2.958214 *x* 10^{20}
SSE	2.474997 *x* 10^{21}	SSE	1.272032 *x* 10^{22}
RSME	7676489551	RSME	17199460010

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Velocidad (km/h)	
Respuesta:		Fuerza (N)	
Número de datos *n*:		51	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
SST	$1.024441\ x\ 10^{25}$	SST	$1.024441\ x\ 10^{25}$
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.95393	W:	0.96121
p-valor:	0.04607* (Sin normalidad)	p-valor:	0.09389
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.074996	D:	0.10141
p-valor:	0.6733	p-valor:	0.2121
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			
D-W:	0.819884	D-W:	0.6755046
p-valor:	0.05238	p-valor:	0.041540*
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	14.723	BP:	28.966
p-valor:	0.06475	p-valor:	0.0001468***
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	78	p-valor	$2.874\ x\ 10^{-16}$ *** (Diferencias significativas entre ambos modelos)

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial de grado 8 se muestran a continuación:

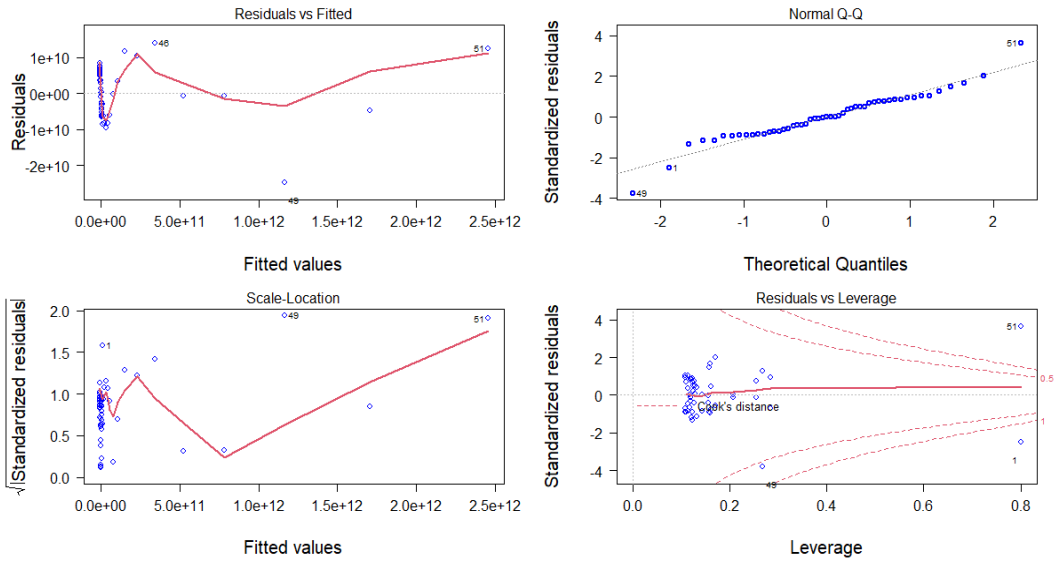

Gráfico 4-1. Gráficos para evaluar la idoneidad del modelo de regresión polinomial de grado 8.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-1 se observa que los valores predichos y los residuos no están vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente en torno a la línea roja y la mayoría está en torno a cero. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que los puntos de posición coinciden aproximadamente con la recta diagonal. En el gráfico de localización de escala se identifica que la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, entonces la variabilidad de las magnitudes es pequeña en función de los valores ajustados. En la última gráfica (grado de influencia Leverage vs residuos estandarizados) se aprecia que existen tres valores atípicos en los residuos que producen apalancamiento y/o están fuera de la distancia de Cook, sin embargo, este efecto es casi imposible de eliminar debido a que la escala de valores de la variable fuerza es logarítmica. En términos generales los gráficos sugieren que se verifican las hipótesis de los supuestos de homocedasticidad y de errores normalmente distribuidos. Se debe tener en cuenta que la prueba de Shapiro-Wilk arroja un p-valor de 0.04607, que implica ausencia de distribución normal de los residuos, no obstante su valor es muy cercano a 0.05, además la prueba de Kolmogorov-Smirnov corregida por Lilliefors presenta un p-valor de 0.6733 que implica normalidad. Por este motivo se acepta el modelo.

Para el caso del modelo de regresión B-spline de grado 5, la base utilizada es la que se muestra como sigue:

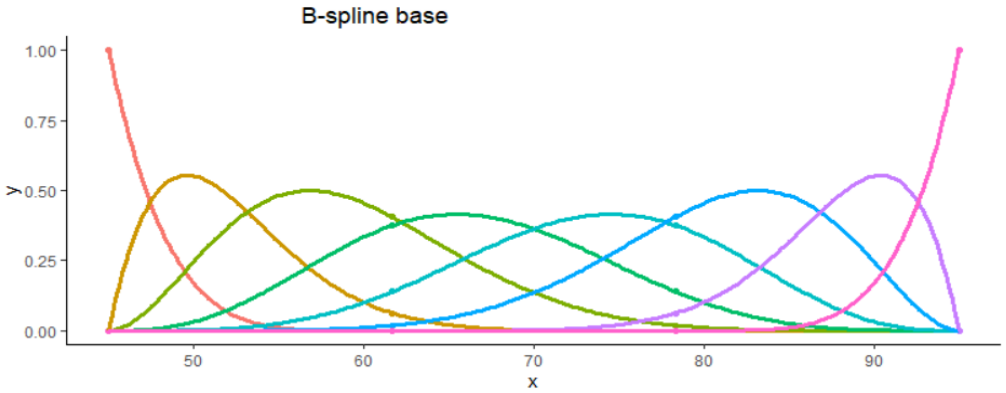

Gráfico 4-2. B-spline base para el modelo de regresión velocidad vs fuerza.
Elaborado por: Toalombo, B. (2021).

Similar al caso del modelo de regresión polinomial de grado 8, para el modelo B-spline de grado 5 se obtuvieron los gráficos de evaluación del modelo:

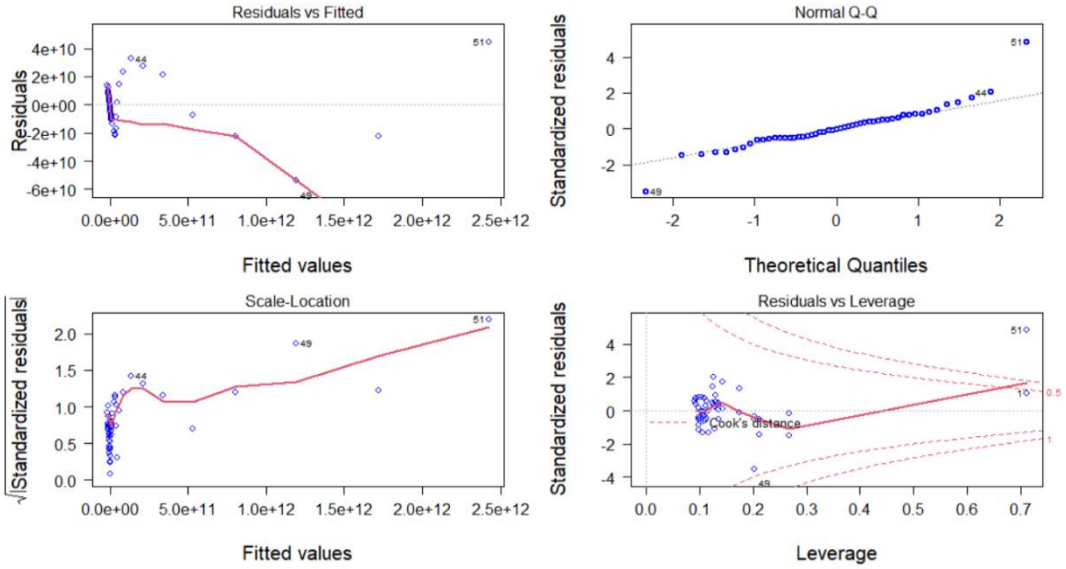

Gráfico 4-3. Gráficos para evaluar el modelo de regresión B-spline de grado 5.
Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-3 se identifica que los residuos están correlacionados y que tampoco existe homocedasticidad. Sin embargo, el modelo B-spline de grado 5 es aceptado en vista de que es no paramétrico.

La gráfica de dispersión de puntos de la velocidad versus la fuerza de impacto se muestra en el Gráfico 4-4, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

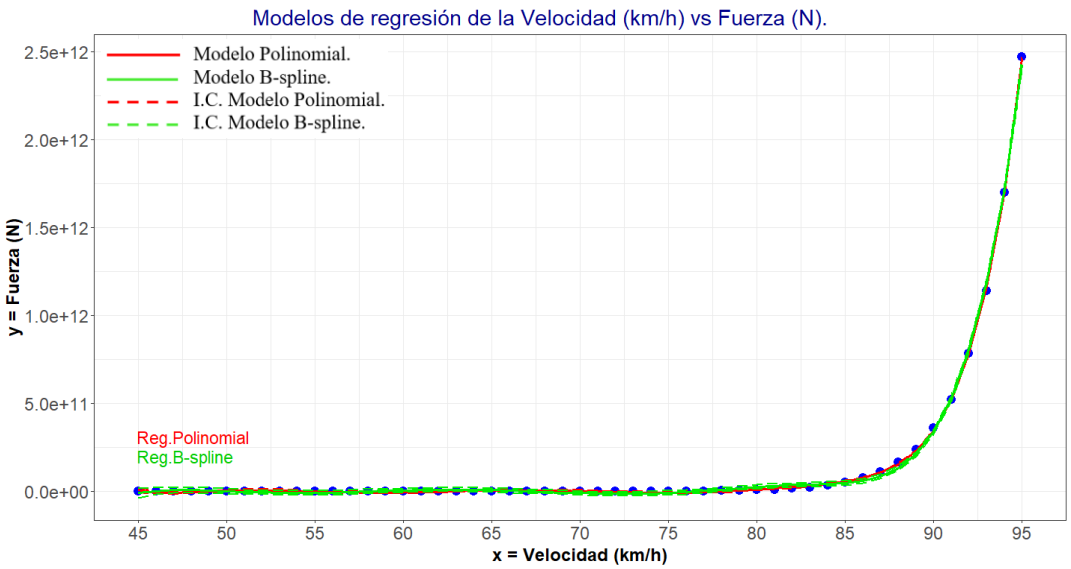

Gráfico 4-4. Modelos de regresión polinomial y B-spline de la velocidad vs fuerza.
Elaborado por: Toalombo, B. (2021).

De acuerdo a la información que proporciona el Gráfico 4-4, la distribución de los puntos sigue una curva en forma de parábola que crece de forma pronunciada, debido a que a medida que se incrementa la velocidad de movimiento del auto, la fuerza del impacto se eleva ostensiblemente, a tal punto que dicha variable respuesta está expresada en una escala logarítmica. Además, se observa que ambos modelos ofrecen una buena bondad de ajuste con intervalos de confianza muy estrechos, aunque son complejos en el sentido de que fue necesario considerar altos grados del polinomio y del B-spline de base (8 y 5, respectivamente).

Con el fin de identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 78 (p-valor = $2.874\ x\ 10^{-16}$***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta

el Gráfico 4-5 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

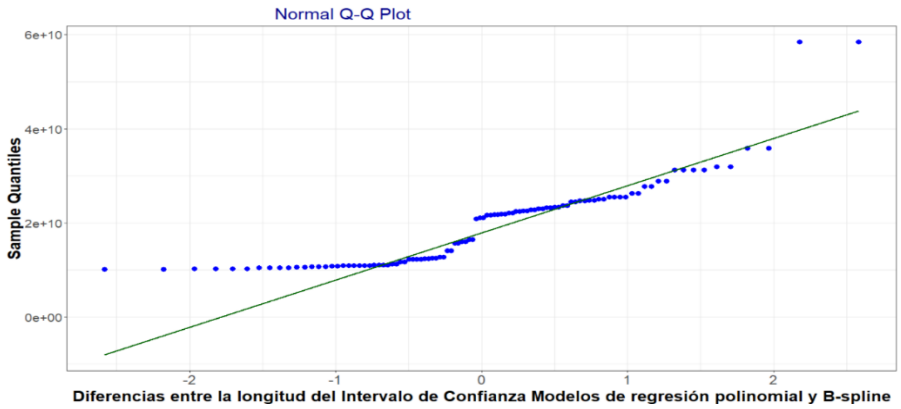

Gráfico 4-5. Ajuste normal Q-Q Plot de la variable fuerza.

Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-5 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.9997124 y un RSE de 7.676 x 10^9, mientras que para el modelo B-spline un R^2 ajustado de 0.998556 y un RSE de 1.72 x 10^{10}. Por lo tanto, la mejor bondad de ajuste corresponde al modelo polinomial.

Los dos modelos son válidos, sin embargo, el modelo de regresión polinomial es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo B-spline. Es decir que la relación de la variable explicativa velocidad y la variable respuesta fuerza con las condiciones dadas del problema se expresa de mejor manera con el uso de la regresión paramétrica polinomial de grado 8.

4.1.1.2 Velocidad vs FDS

Se emplearon 51 datos de la velocidad de desplazamiento del auto pequeño y el factor de seguridad (FDS) requerido para el diseño de la estructura del autobús en función de la velocidad. En la Tabla 4-2 se presentan los modelos de regresión polinomial y B-spline de la variable velocidad en km/h versus la variable FDS:

Tabla 4-2. Modelos de regresión de la velocidad y FDS.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Velocidad (km/h)	
Respuesta:		Factor de seguridad FDS (adimensional)	
Número de datos *n*:		51	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	5	Grado:	3 (Cúbico)
		Vértices del polígono de control:	5
Intercepto β_0	-258.6344	Intercepto a_0	33.57894
β_1	21.7104	a_1	-32.38416
β_2	-0.7155	a_2	-32.64817
β_3	0.01167	a_3	-32.84946
β_4	-0.0000948	a_4	-28.93164
β_5	$3.1105\ x\ 10^{-7}$	a_5	-17.16638
F	1537000	F	362900
p-valor	0.000***	p-valor	0.000***
RSE	0.02288	RSE	0.04572
R^2 ajustado	0.9999935	R^2 ajustado	0.9999724
MSE	0.00049367	MSE	0.00209067
SSE	0.0222152	SSE	0.09407996
RSME	0.0222187	RSME	0.0457238
SST	3793.162	SST	3793.162
MAPE	1.00%	-	-
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.95485	W:	0.97033
p-valor:	0.05038	p-valor:	0.2282
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.074616	D:	0.095047
p-valor:	0.6809	p-valor:	0.2983
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			
D-W:	0.8168112	D-W:	0.3692189
p-valor:	0.06107	p-valor:	0.02033*
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	11.256	BP:	5.0543

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Velocidad (km/h)	
Respuesta:		Factor de seguridad FDS (adimensional)	
Número de datos *n*:		51	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
p-valor:	0.04654*	p-valor:	0.4093
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	82	p-valor	$3.584\ x\ 10^{-16}$*** (Diferencias significativas entre ambos modelos)

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial de grado 5 se muestran a continuación:

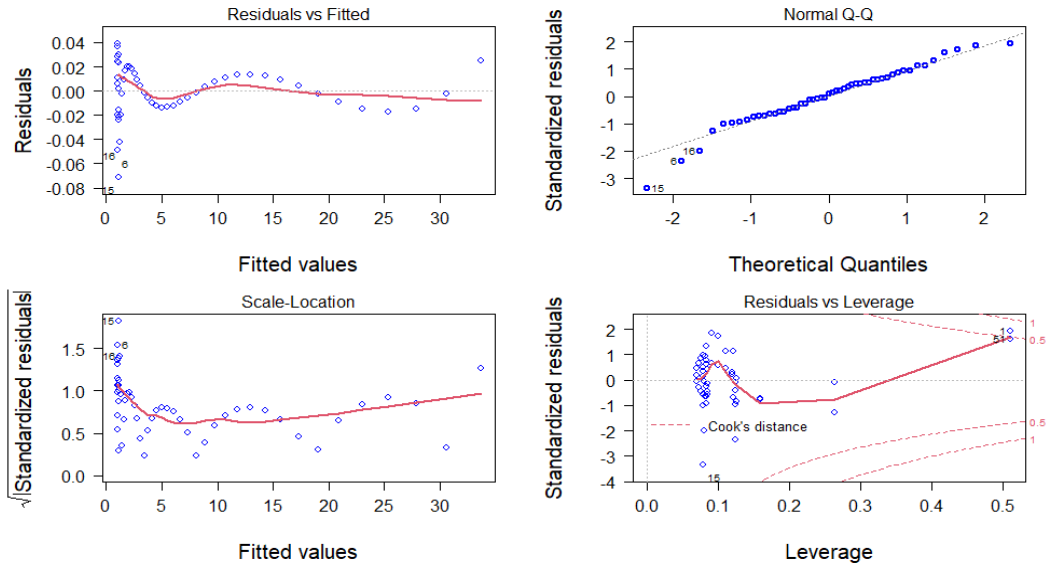

Gráfico 4-6. Gráficos para evaluar el modelo de regresión B-spline de grado 5.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-6 se aprecia que los valores predichos y los residuos no están vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente

en torno a la línea roja y la mayoría está en torno a cero. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que los puntos de posición coinciden aproximadamente con la recta diagonal. En el gráfico de localización de escala se observa que la dispersión alrededor de la línea roja no varía con los valores ajustados, aunque existe una tendencia en la distribución de los puntos en torno a un patrón de curva ondulada, entonces la variabilidad de las magnitudes varía ligeramente en función de los valores ajustados. En el gráfico del grado de influencia (Leverage) vs residuos estandarizados se aprecia que existen dos valores atípicos en los residuos que producen apalancamiento y/o están fuera de la distancia de Cook. Los gráficos sugieren que se cumplen las hipótesis de no autocorrelación y de errores normalmente distribuidos, pero no existe homocedasticidad. Se debe tener en cuenta que la prueba de Breusch-Pagan estudiantilizada arroja un p-valor de 0.04654, que implica ausencia de distribución normal de los residuos, no obstante su valor es muy cercano a 0.05, por lo que se entiende que la heterocedasticidad no es marcada. Por este motivo se acepta el modelo.

Para el caso del modelo de regresión B-spline cúbico, la base utilizada es la que se muestra como sigue:

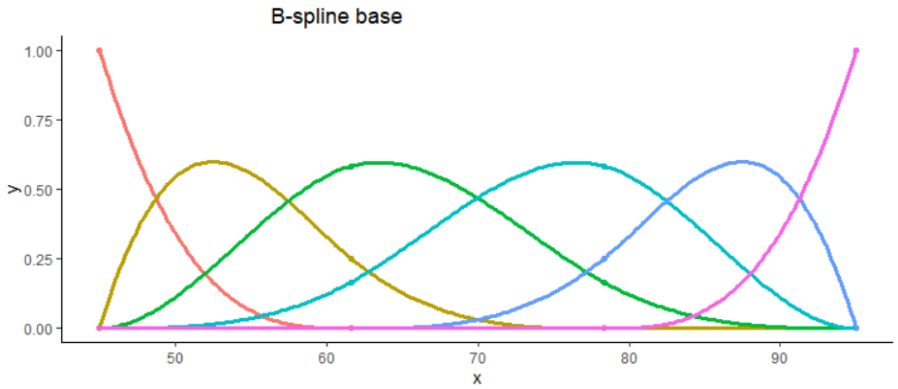

Gráfico 4-7. B-spline base para el modelo de regresión velocidad vs FDS.
Elaborado por: Toalombo, B. (2021).

Igual al caso del modelo de regresión polinomial de grado 5, para el modelo B-spline cúbico se obtuvieron los gráficos de evaluación del modelo:

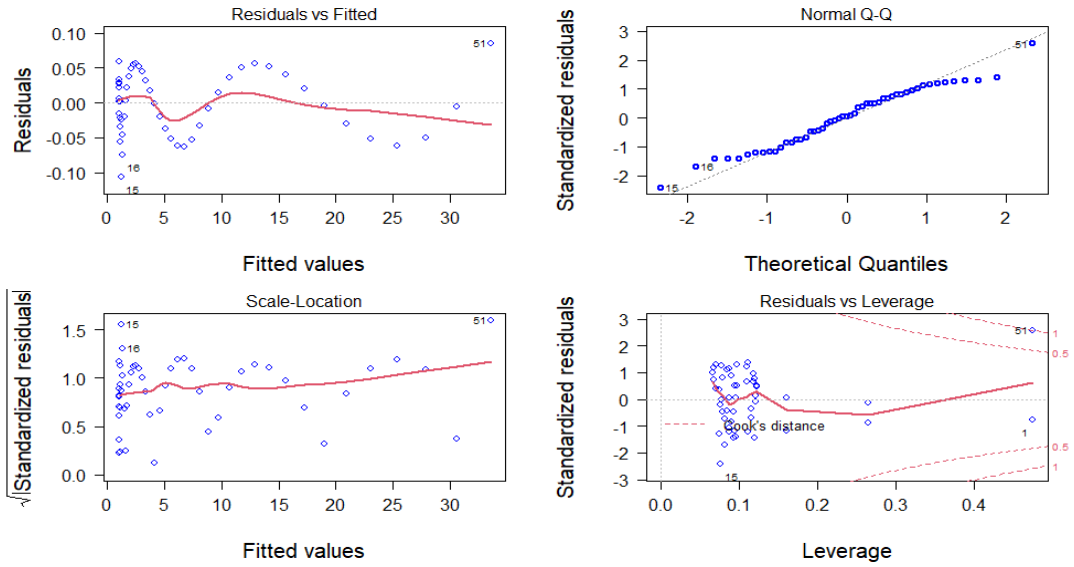

Gráfico 4-8. Gráficos para evaluar el modelo de regresión B-spline cúbico.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-8 se identifica que los residuos están correlacionados. Sin embargo, el modelo B-spline cúbico es aceptado en vista de que es no paramétrico.

La gráfica de dispersión de puntos de la velocidad versus el FDS se muestra en el Gráfico 4-9, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

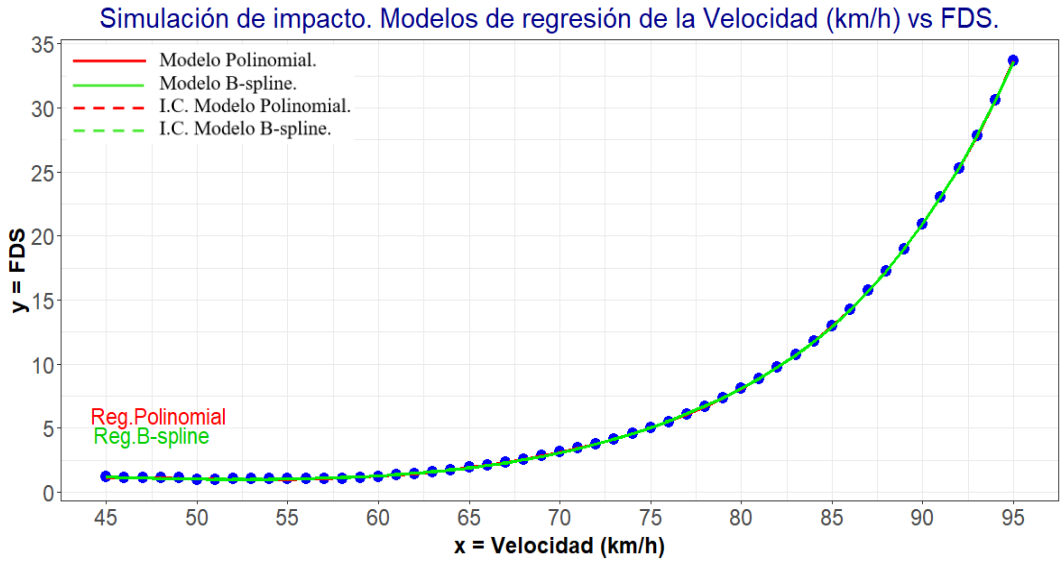

Gráfico 4-9. Modelos de regresión polinomial y B-spline de la velocidad vs FDS.

Elaborado por: Toalombo, B. (2021).

De acuerdo a la información que proporciona el Gráfico 4-9, la distribución de los puntos sigue una curva en forma de parábola creciente, debido a que conforme se incrementa la velocidad de movimiento del auto, el FDS de diseño requerido también crece. Comparativamente cuando la velocidad de desplazamiento del vehículo que impacta a la carrocería del autobús no excede de 60 km/h, el FDS de diseño aproximadamente no supera el valor de 2, pero a medida que se eleva la velocidad el FDS requerido para que la estructura no se deforme está en torno a 35. En cuanto a los dos modelos, ambos ofrecen una buena bondad de ajuste con intervalos de confianza muy estrechos, al mismo tiempo que tampoco son demasiado complejos al estar constituidos por polinomios de grado 3 y 5, respectivamente para el modelo polinomial y B-spline.

Para identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 82 (p-valor = $3.584 \; x \; 10^{-16}$***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se muestra el Gráfico 4-10 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

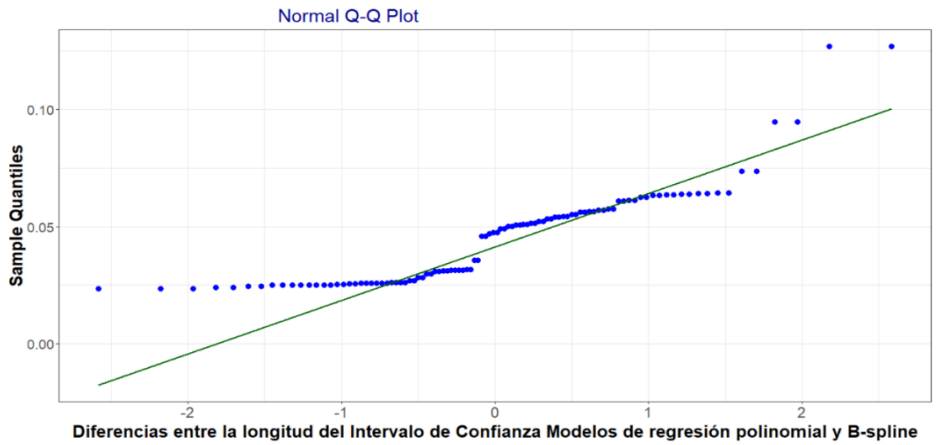

Gráfico 4-10. Ajuste normal Q-Q Plot de la variable FDS.

Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-10 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo

suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.9999935 y un RSE de 0.02288, mientras que para el modelo B-spline un R^2 ajustado de 0.9999724 y un RSE de 0.04572. Por lo tanto, la mejor bondad de ajuste corresponde al modelo polinomial.

Los dos modelos son válidos, no obstante, el modelo de regresión polinomial es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo B-spline. Es decir que la relación de la variable explicativa velocidad y la variable respuesta FDS con las condiciones dadas del problema se expresa de mejor manera con el uso de la regresión paramétrica polinomial de grado 5.

4.1.1.3 Velocidad vs Tiempo de impacto

Se consideraron 51 datos de la velocidad de desplazamiento del auto pequeño y el tiempo de duración del impacto contra el autobús. En la Tabla 4-3 se presentan los modelos de regresión polinomial y B-spline de la variable velocidad en km/h versus el tiempo de impacto en segundos:

Tabla 4-3. Modelos de regresión de la velocidad y tiempo de impacto.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Velocidad (km/h)	
Respuesta:		Tiempo de impacto (s)	
Número de datos *n*:		51	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	3 (Cúbico)	Grado:	3 (Cúbico)
		Vértices del polígono de control:	6
Intercepto β_0	-2.694863	Intercepto a_0	0.63982
β_1	0.09930049	a_1	-0.64031
β_2	-0.001079244	a_2	-0.57289
β_3	4.236292×10^{-6}	a_3	-0.37788
β_4	0	a_4	-0.18385
β_5	0	a_5	-0.13317

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Velocidad (km/h)	
Respuesta:		Tiempo de impacto (s)	
Número de datos *n*:		51	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
F	593.8	F	305.3
p-valor	0.000***	p-valor	0.000***
RSE	0.0309	RSE	0.03051
R^2 ajustado	0.9726526	R^2 ajustado	0.9733435
MSE	0.0009547156	MSE	0.00093059
SSE	0.04487163	SSE	0.040946
RSME	0.0308985	RSME	0.0305056
SST	1.74553	SST	1.74553
Test de normalidad de los residuos:			
Prueba: Shapiro-Wilk			
W:	0.97539	W:	0.98146
p-valor:	0.365	p-valor:	0.6024
Prueba: Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.090466	D:	0.082547
p-valor:	0.3729	p-valor:	0.5212
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			
D-W:	1.235732	D-W:	1.34797
p-valor:	0.062	p-valor:	0.020*
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	4.4028	BP:	10.165
p-valor:	0.2211	p-valor:	0.1179
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	356	p-valor	$2.647\ x\ 10^{-10}$***

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial cúbico se muestran a continuación:

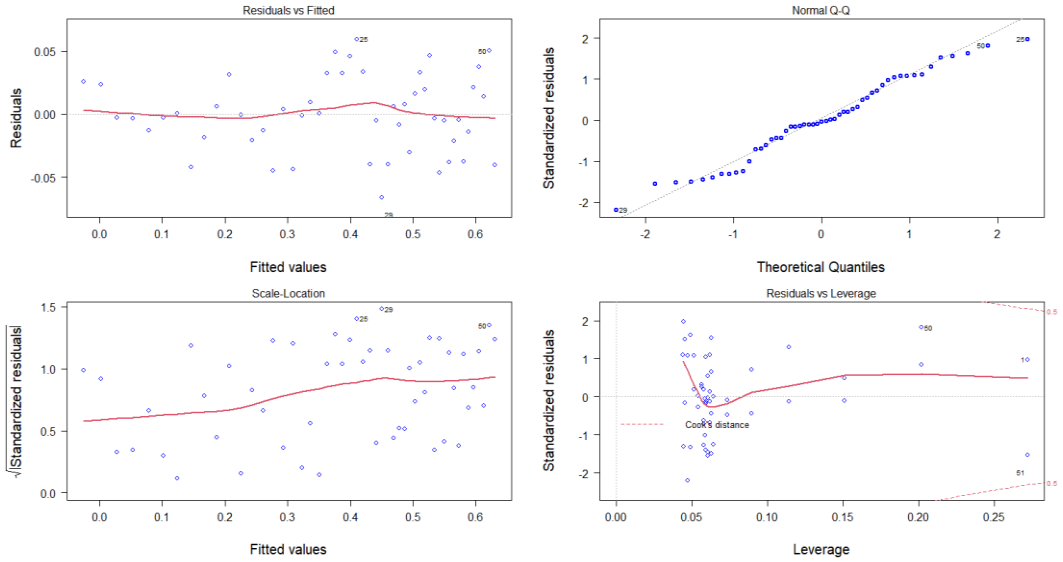

Gráfico 4-11. Gráficos para evaluar el modelo de regresión polinomial cúbico.
Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-11 se observa que los valores predichos y los residuos no están vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente en torno a la línea roja que es casi horizontal. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados evidencia que los puntos de posición coinciden aproximadamente con la recta diagonal. En el gráfico de localización de escala se identifica que la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, es decir que la variabilidad de las magnitudes no cambia en función de los valores ajustados. En la gráfica del grado de influencia (Leverage) vs residuos estandarizados se aprecia que los valores atípicos en los residuos no tienen efecto sobre la línea roja. Los gráficos sugieren que se cumplen las hipótesis de los supuestos de no autocorrelación, homocedasticidad y distribución normal de los residuos. Los gráficos concuerdan con los resultados de las pruebas estadísticas de Shapiro-Wilk, Kolmogorov-Smirnov corregida por Lilliefors, Durbin-Watson y Breusch-Pagan estudiantilizada, en favor de las hipótesis que comprueban los supuestos del modelo. Bajo estas circunstancias se acepta el modelo de regresión polinomial.

Por otra parte, para el caso del modelo de regresión B-spline cúbico, la base utilizada es la que se muestra como sigue:

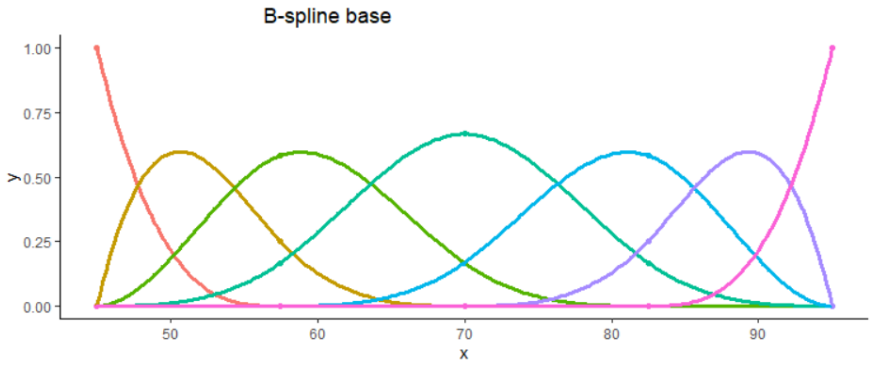

Gráfico 4-12. B-spline base para el modelo de regresión velocidad vs tiempo de impacto.

Elaborado por: Toalombo, B. (2021).

Del mismo modo que en el caso del modelo de regresión polinomial de grado 3, para el modelo B-spline cúbico, se obtuvieron los gráficos de evaluación del modelo:

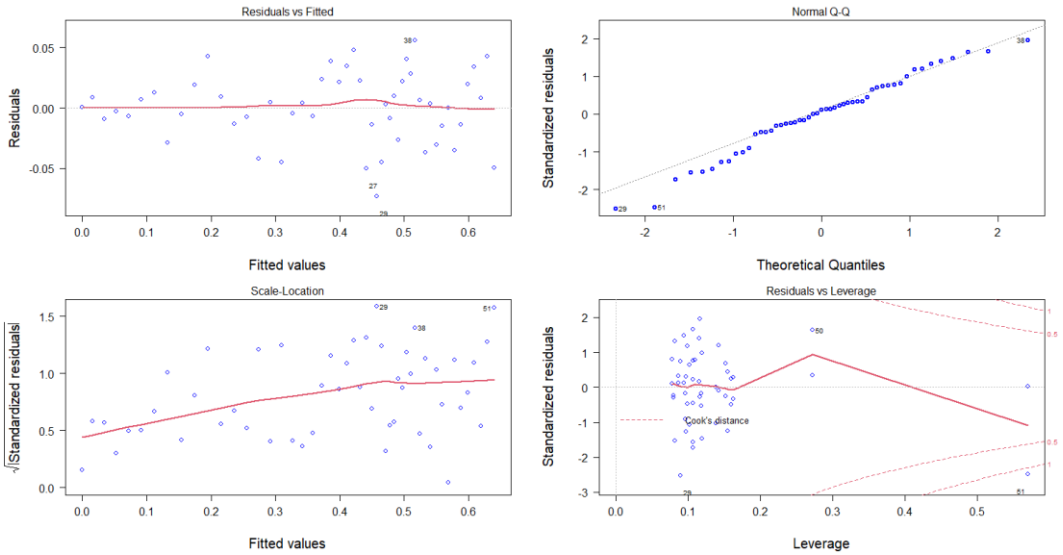

Gráfico 4-13. Gráficos para evaluar el modelo de regresión B-spline cúbico.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-13 se identifica que los residuos están correlacionados. Sin embargo, el modelo B-spline cúbico es aceptado en vista de que es no paramétrico.

La gráfica de dispersión de puntos de la velocidad versus el tiempo de impacto se muestra en el Gráfico 4-14, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

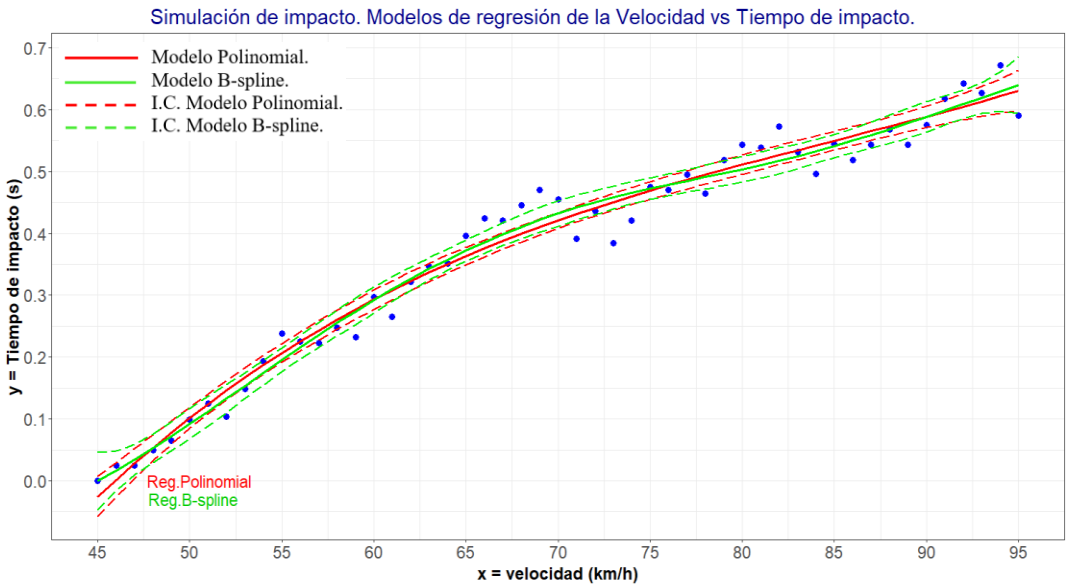

Gráfico 4-14. Modelos de regresión polinomial y B-spline de la velocidad vs tiempo de impacto.

Elaborado por: Toalombo, B. (2021).

Según la información que proporciona el Gráfico 4-14, la distribución de los puntos sigue una curva con forma parecida a una función logarítmica. Conforme se incrementa la velocidad de movimiento del auto también aumenta el tiempo de duración del impacto, aunque el ritmo de crecimiento va disminuyendo a medida que la velocidad se acerca a la máxima del ensayo de 95 km/h. Además, se observa que ambos modelos ofrecen una aceptable bondad de ajuste, aunque con intervalos de confianza que dejan por fuera a algunos puntos. Los modelos obtenidos son relativamente simples, ya que se consideran grados cúbicos tanto para el modelo de regresión polinómico cuanto para la base del B-spline.

Con el fin de identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 365 (p-valor = 2.647×10^{-10}***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de

ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta el Gráfico 4-15 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

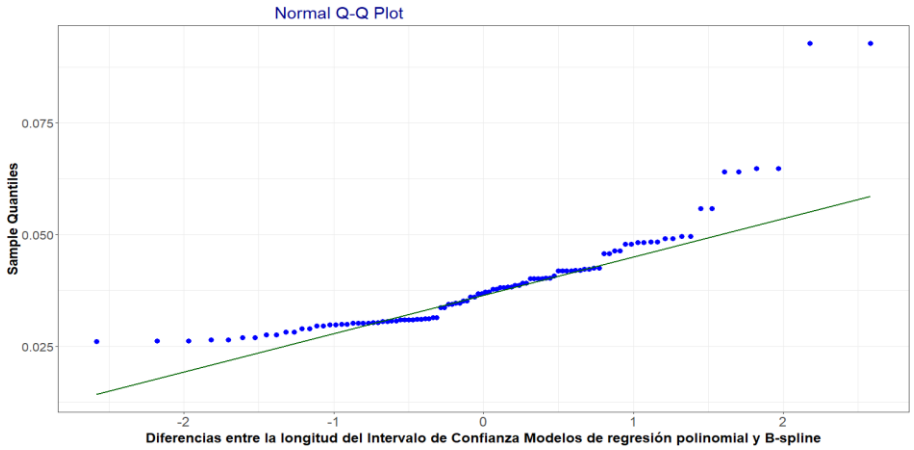

Gráfico 4-15. Ajuste normal Q-Q Plot de la variable Tiempo de impacto.
Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-15 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.9726526 y un RSE de 0.0309, mientras que para el modelo B-spline un R^2 ajustado de 0.9733435 y un RSE de 0.03051. Por lo tanto, la mejor bondad de ajuste corresponde al modelo B-spline.

Los dos modelos son válidos, no obstante, el modelo de regresión B-spline es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo polinomial. Es decir que la relación de la variable explicativa velocidad y la variable respuesta tiempo de duración del impacto con las condiciones dadas del problema se expresa de mejor manera con el uso de la regresión no paramétrica B-spline de grado 3.

4.1.1.4 Fuerza vs Deformación

Se utilizaron 48 datos de la fuerza de impacto del auto pequeño y la deformación generada por el impacto del mismo contra la estructura de la carrocería del autobús.

En la Tabla 4-4 se presentan los modelos de regresión polinomial y B-spline de la variable fuerza en Newton versus la variable deformación en milímetros:

Tabla 4-4. Modelos de regresión de la fuerza y deformación.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Fuerza (N)	
Respuesta:		Deformación (mm)	
Número de datos *n*:		48	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	8	Grado:	4
		Vértices del polígono de control:	6
Intercepto β_0	3.986846	Intercepto a_0	12.3073
β_1	8.576639 *x* 10^{-10}	a_1	-11.5149
β_2	-3.331115 *x* 10^{-20}	a_2	-7.5879
β_3	5.50025 *x* 10^{-31}	a_3	-3.6727
β_4	-4.520568 *x* 10^{-42}	a_4	4.4501
β_5	1.980222 *x* 10^{-53}	a_5	-7.4979
β_6	-4.64129 *x* 10^{-65}	a_6	0
β_7	5.421413 *x* 10^{-77}	a_7	0
β_8	-2.448582 *x* 10^{-89}	a_8	0
F	11.87	F	449.7
p-valor	2.228 *x* 10^{-8}***	p-valor	0.000***
RSE	2.185	RSE	0.4832
R^2 ajustado	0.649063	R^2 ajustado	0.9828409
MSE	4.774958	MSE	0.2334722
SSE	186.2233	SSE	9.572361
RSME	2.185168	RSME	0.4831896
SST	639.4965	SST	639.4965
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.96138	W:	0.95802
p-valor:	0.1148	p-valor:	0.08395
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.066969	D:	0.10452
p-valor:	0.8524	p-valor:	0.2109
Test de residuos no correlacionados:			

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Fuerza (N)	
Respuesta:		Deformación (mm)	
Número de datos *n*:		48	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Prueba de Durbin-Watson			
D-W:	0.09927003	D-W:	0.2574849
p-valor:	0.072152	p-valor:	0.049731*
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	9.7254	BP:	9.5885
p-valor:	0.2848	p-valor:	0.1431
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	2233	p-valor	0.000 *** (Diferencias significativas entre ambos modelos)

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial de grado 8 se muestran a continuación:

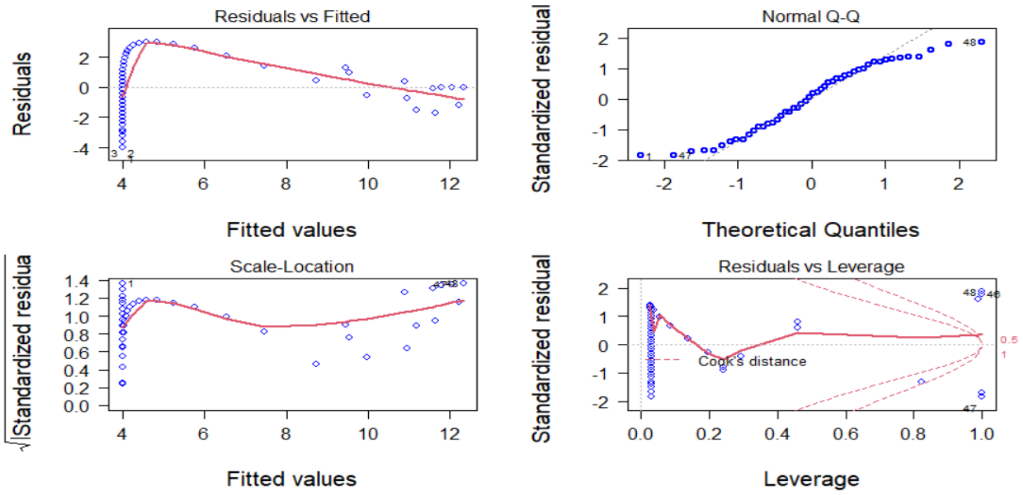

Gráfico 4-16. Gráficos para evaluar el modelo de regresión polinomial de grado 8.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-16 se observa que los valores predichos y los residuos parecen no estar vinculados o correlacionados, ya que al menos la mitad de los puntos están distribuidos aleatoriamente en torno a la línea roja. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que los puntos de posición coinciden aproximadamente con la recta diagonal. En el gráfico de localización de escala se identifica que la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, entonces la variabilidad de las magnitudes no varía en función de los valores ajustados. En la gráfica del grado de influencia (Leverage) vs residuos estandarizados se observa que existen tres valores atípicos en los residuos que producen apalancamiento y/o están fuera de la distancia de Cook. Los gráficos sugieren que se cumplen las hipótesis de los supuestos de no autocorrelación, homocedasticidad y de errores normalmente distribuidos. Los gráficos concuerdan con los resultados de las pruebas estadísticas de Shapiro-Wilk, Kolmogorov-Smirnov corregida por Lilliefors, Durbin-Watson y Breusch-Pagan estudiantilizada, en favor de las hipótesis que comprueban los supuestos del modelo. En tal virtud se acepta el modelo de regresión polinomial de grado 8.

Para el caso del modelo de regresión B-spline de grado 4, la base utilizada es la que se presenta a continuación:

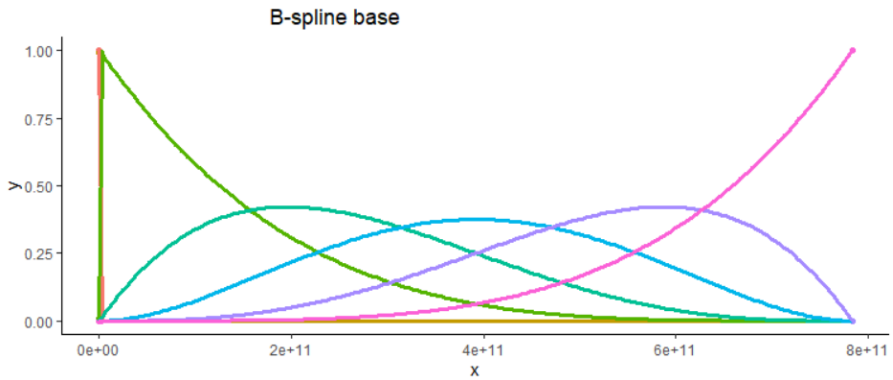

Gráfico 4-17. B-spline base para el modelo de regresión fuerza vs deformación. Elaborado por: Toalombo, B. (2021).

Del mismo modo que en el caso del modelo de regresión polinomial de grado 8, para el modelo B-spline de grado 4 se obtuvieron los gráficos de evaluación del modelo:

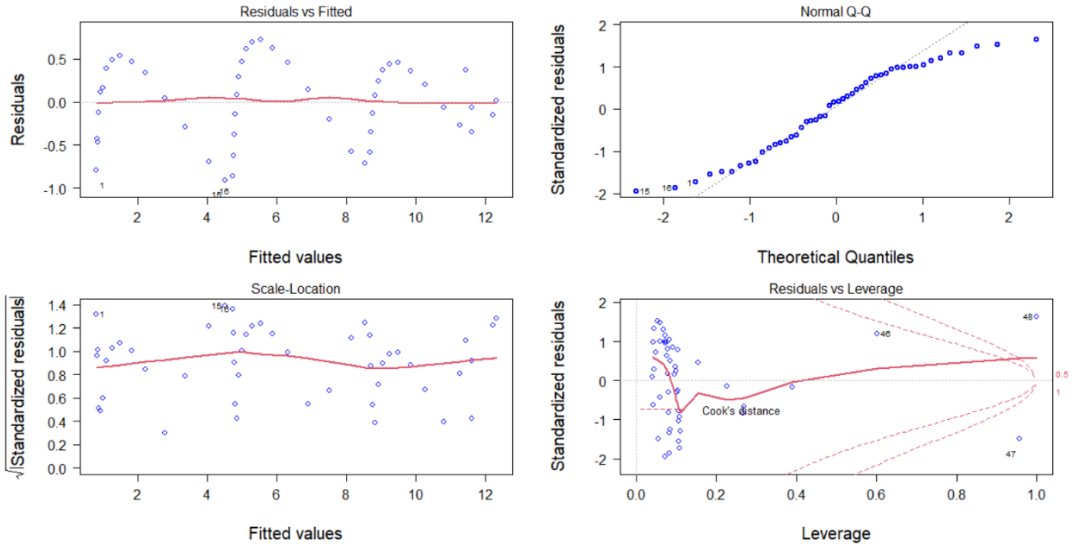

Gráfico 4-18. Gráficos para evaluar el modelo de regresión B-spline de grado 4.
Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-18 se identifica que los residuos están correlacionados. Sin embargo, el modelo B-spline de grado 4 es aceptado en vista de que es no paramétrico.

La gráfica de dispersión de puntos de la fuerza versus la deformación se indica en el Gráfico 4-19, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

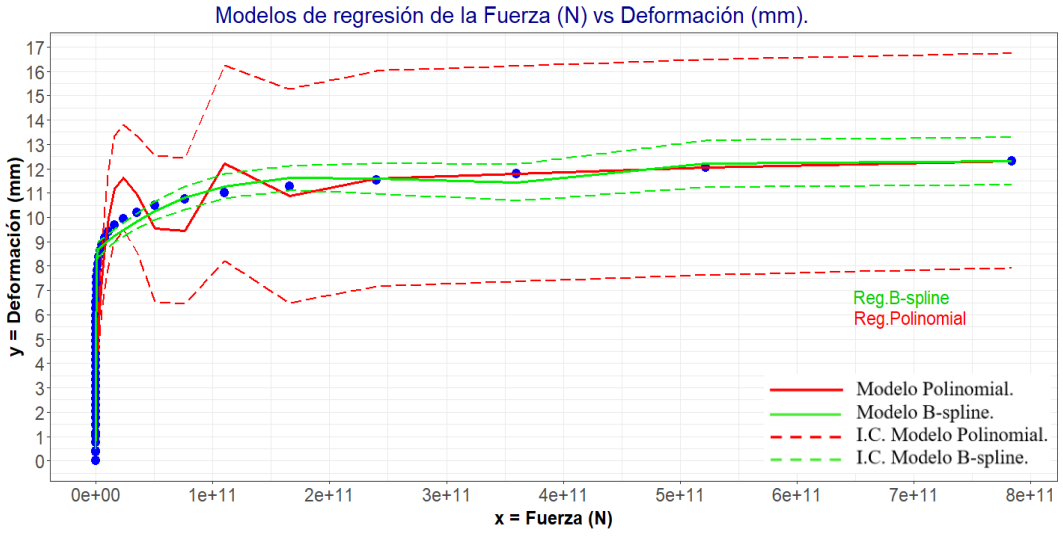

Gráfico 4-19. Modelos de regresión polinomial y B-spline de fuerza vs deformación.
Elaborado por: Toalombo, B. (2021).

De acuerdo a la información que proporciona el Gráfico 4-19, la distribución de los puntos sigue una curva con forma parecida a una función logarítmica. Conforme se incrementa la fuerza de impacto del auto sobre la carrocería del autobús también aumenta la deformación de la estructura, aunque el ritmo de crecimiento va disminuyendo a medida que la fuerza se hace más grande. La curva del modelo polinomial no ofrece un ajuste aceptable, dado que el intervalo de confianza al 95% es demasiado ancho al mismo tiempo que la curva tiene aristas en forma de dientes de sierra. Por el contrario, la curva del modelo de regresión B-spline brinda un ajuste aceptable a pesar de que existen algunos puntos que quedan ligeramente fuera de intervalo de confianza. Adicionalmente se destaca que el modelo de regresión polinomial es complejo en vista de que es de grado 8, en tanto que el B-spline no lo es al tener un grado 4. Se visualiza como una mejor opción el modelo de regresión no paramétrico.

Para identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 2233 (p-valor = 0.000***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta el Gráfico 4-20 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

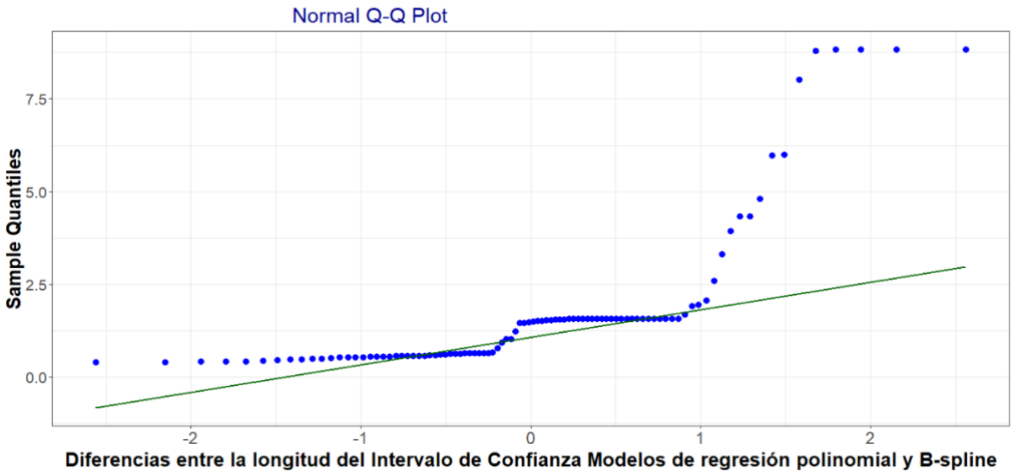

Gráfico 4-20. Ajuste normal Q-Q Plot de la variable deformación.

Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-20 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.649063 y un RSE de 2.185, mientras que para el modelo B-spline un R^2 ajustado de 0.9828409 y un RSE de 0.4832. Por lo tanto, la mejor bondad de ajuste corresponde al modelo B-spline.

Al comparar la distribución de los diagramas de cajas de ambos modelos, es evidente que el de regresión B-spline tiene una menor longitud del intervalo de confianza, lo que significa que tiene una mejor capacidad de minimizar los residuos. Por consiguiente, la relación de la variable explicativa fuerza de impacto y la variable respuesta deformación de la estructura tubular de acero de la carrocería del autobús, se expresa de mejor manera con el uso de la regresión no paramétrica B-spline de grado 4. En este caso no se podría considerar el modelo polinomial, por no ofrecer un ajuste aceptable.

4.1.2 Variables climatológicas

Las variables climatológicas de San Antonio de Pichincha que fueron consideradas en el presente estudio son: radiación solar, temperatura ambiente, humedad relativa y presión atmosférica. El volumen de datos disponibles de las variables es robusto, debido a que se dispone de más de 32000 datos, correspondientes a la situación climatológica de los años 2017 a 2020. Al efectuar una inspección de las correlaciones existentes entre las mencionadas variables no se hallaron correlaciones significativas, salvo el caso de la relación entre la radiación solar y la temperatura, pero solamente para el año 2020. Por este motivo se procedió a establecer correlaciones entre la hora del día y cada una de las variables, para lo cual se calcularon los valores promedio de cada magnitud según la hora del día. De esta manera se obtuvieron 24 pares de datos, que provienen de una base original de miles de datos.

Las bases de datos utilizadas se denominan “Climatologicos.csv” y “Rad_Temp.csv”, la primera está conformada por 24 pares de datos y se presenta

en el Anexo B del presente documento; mientras que la segunda contiene 605 pares de datos. Igual al caso de la simulación de impacto vehicular, para la presentación de los resultados se emplean tablas que detallan la información concerniente al análisis y validación de los modelos de regresión polinomial y B-spline. A continuación, se presenta el desarrollo de cada caso:

4.1.2.1 Radiación solar vs Temperatura

El conjunto de datos utilizados para analizar la relación entre la variable regresora radiación solar y temperatura ambiente es de 605 para cada variable. Los datos corresponden al periodo comprendido entre el 7 de noviembre del año 2020 y el 28 de diciembre de 2020, en el horario comprendido entre las 7 am y las 6 pm, excluyéndose los horarios en los que no existe radiación solar. Al efectuarse el análisis con las condiciones reales y una vez eliminados los datos atípicos, no se encontraron modelos de regresión válidos (para los supuestos). Por este motivo se procedió a realizar una transformación logarítmica, siendo conveniente utilizar la de la forma expresada en la ecuación (44) con logaritmo natural de la variable respuesta. En la Tabla 4-5 se presentan los modelos de regresión polinomial y B-spline de la variable radiación solar versus la variable temperatura ambiente:

Tabla 4-5. Modelos de regresión de la radiación solar y temperatura.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Radiación solar (W/m^2)	
Respuesta:		Logaritmo natural de la Temperatura (ºC)	
Número de datos *n*:		605	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	3	Grado:	3 (Cúbico)
		Vértices del polígono de control:	5
Intercepto β_0	2.577	Intercepto a_0	3.0819
β_1	0.0009332	a_1	-0.48371
β_2	$-9.330\ x\ 10^{-7}$	a_2	-0.45612
β_3	$4.379\ x\ 10^{-10}$	a_3	-0.25008

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Radiación solar (W/m^2)	
Respuesta:		Logaritmo natural de la Temperatura (ºC)	
Número de datos *n*:		605	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
β_4	0	a_4	-0.19141
β_5	0	a_5	0
F	462.8	F	278.6
p-valor	0.000***	p-valor	0.000***
RSE	0.09131	RSE	0.09125
R^2 ajustado	0.6963733	R^2 ajustado	0.696761
MSE	0.0083375	MSE	0.00832686
SSE	5.01084	SSE	4.98779
RSME	0.0913099	RSME	0.091252
SST	16.58567	SST	16.58567
MAPE	0.48%	-	-
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.99499	W:	0.99532
p-valor:	0.04604* (Sin normalidad)	p-valor:	0.06454
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.035188	D:	0.034606
p-valor:	0.07176	p-valor:	0.08159
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			
D-W:	0.8482043	D-W:	0.8506689
p-valor:	0.000***	p-valor:	0.000***
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	118.5	BP:	133.06
p-valor:	0.000***	p-valor:	0.000***
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	59527	p-valor	0.000*** (Diferencias significativas

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Radiación solar (W/m^2)	
Respuesta:		Logaritmo natural de la Temperatura (ºC)	
Número de datos *n*:		605	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
			entre ambos modelos)

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial cúbico se muestran a continuación:

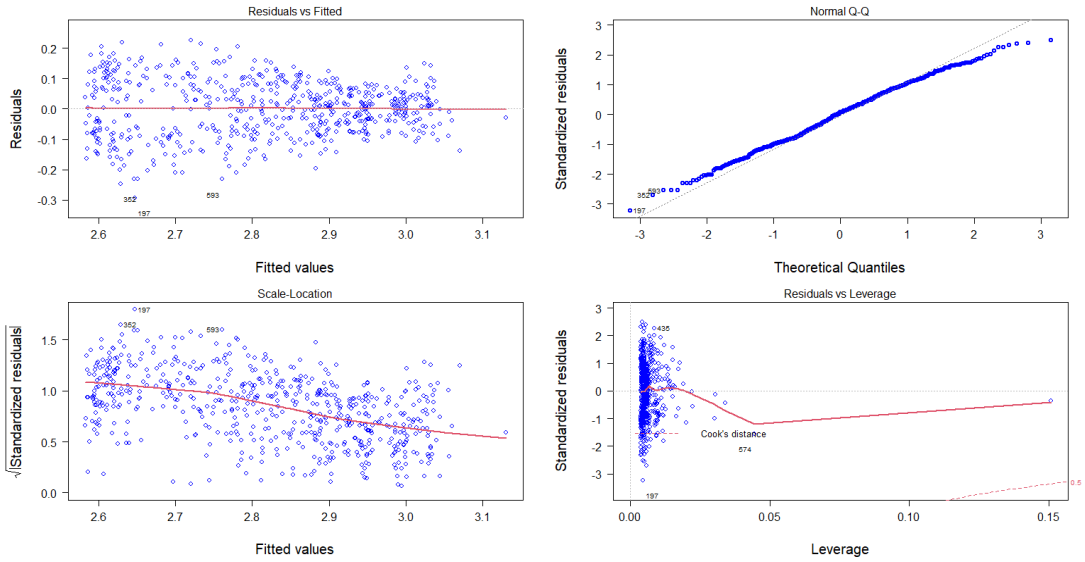

Gráfico 4-21. Gráficos para evaluar el modelo de regresión polinomial cúbico.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-21 se observa que los valores predichos y los residuos parecen no estar vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente en torno a la línea roja, pero la mayoría no están en torno a cero. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que los puntos de posición coinciden con la recta diagonal, siendo evidente la existencia de normalidad en los residuos. En el gráfico de localización de escala se identifica que

la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, entonces la variabilidad de las magnitudes es pequeña en función de los valores ajustados. En el gráfico del grado de influencia (Leverage) vs residuos estandarizados se aprecia que existen varios valores atípicos en los residuos que producen apalancamiento y están fuera de la distancia de Cook. Los gráficos sugieren que se cumplen las hipótesis de errores normalmente distribuidos, pero los supuestos de homocedasticidad y no autocorrelación parecen no verificarse. Se debe tener en cuenta que la prueba de Shapiro-Wilk arroja un p-valor de 0.04604, que implica ausencia de distribución normal de los residuos, no obstante su valor es muy cercano a 0.05, además la prueba de Kolmogorov-Smirnov corregida por Lilliefors presenta un p-valor de 0.07176 que implica normalidad. No obstante, las pruebas de Durbin-Watson y de Breusch-Pagan estudiantilizada arrojan p-valores de 0, lo que implica que los residuos están correlacionados y sin homocedasticidad. Bajo estas circunstancias no se puede establecer un intervalo de confianza al 95%.

La base utilizada para el modelo de regresión B-spline cúbico es la siguiente:

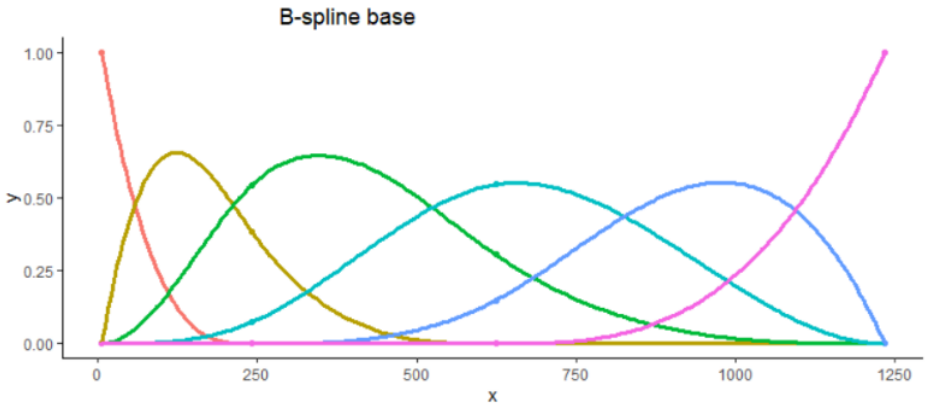

Gráfico 4-22. B-spline base para el modelo de regresión radiación solar vs temperatura.

Elaborado por: Toalombo, B. (2021).

Similar al caso del modelo de regresión polinomial de grado 3, para el modelo B-spline cúbico se obtuvieron los gráficos de evaluación del modelo:

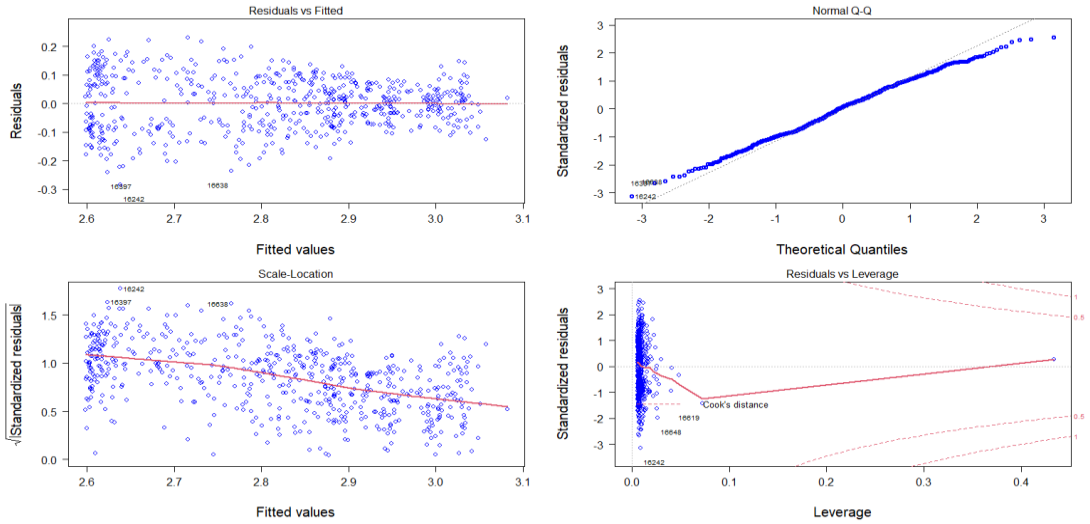

Gráfico 4-23. Gráficos para evaluar el modelo de regresión B-spline cúbico.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-23 se identifica que los residuos están correlacionados y que no existe homocedasticidad. Por este motivo, no se pueden establecer intervalos de confianza al 95%.

La gráfica de dispersión de puntos de la radiación solar versus la temperatura ambiente se muestra en el Gráfico 4-24, conjuntamente con las curvas de los dos modelos de regresión:

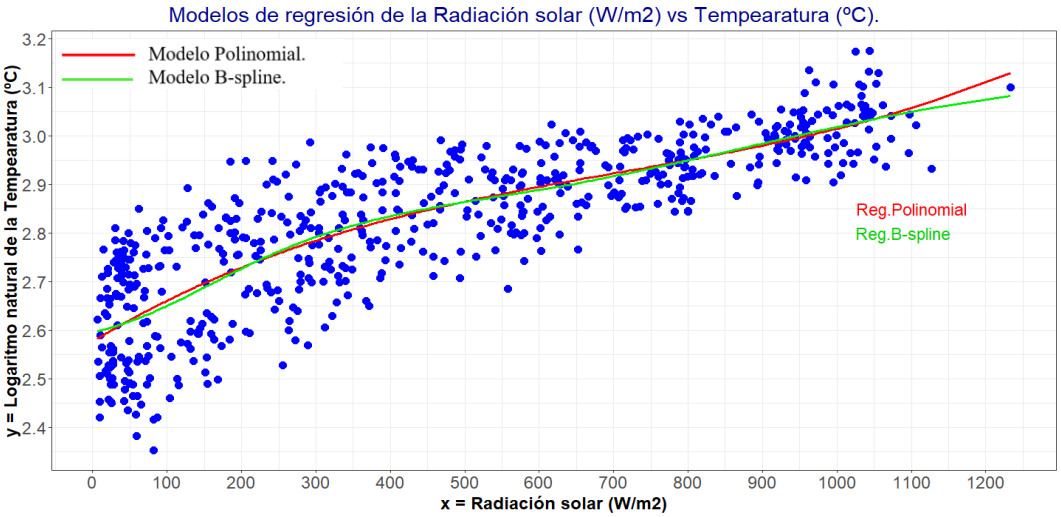

Gráfico 4-24. Modelos de regresión polinomial y B-spline de la radiación solar vs. temperatura.

Elaborado por: Toalombo, B. (2021).

Según la información que proporciona el Gráfico 4-24, la distribución de los puntos es dispersa y sigue una curva casi lineal inclinada hacia la derecha, mientras mayor es la radiación solar también es más elevada la temperatura ambiente. La ventaja que se puede apreciar en ambos modelos, es que tienen expresiones sencillas, ya que son de grado 3. Los dos modelos de regresión presentan una bondad de ajuste aceptable, aunque no es posible establecer intervalos de confianza al 95%, tomando en cuenta los resultados de las pruebas de Shapiro-Wilk, Durbin-Watson y Breusch-Pagan estudiantilizada.

Con la finalidad de identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 59527 (p-valor = 0.000***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta el Gráfico 4-25 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

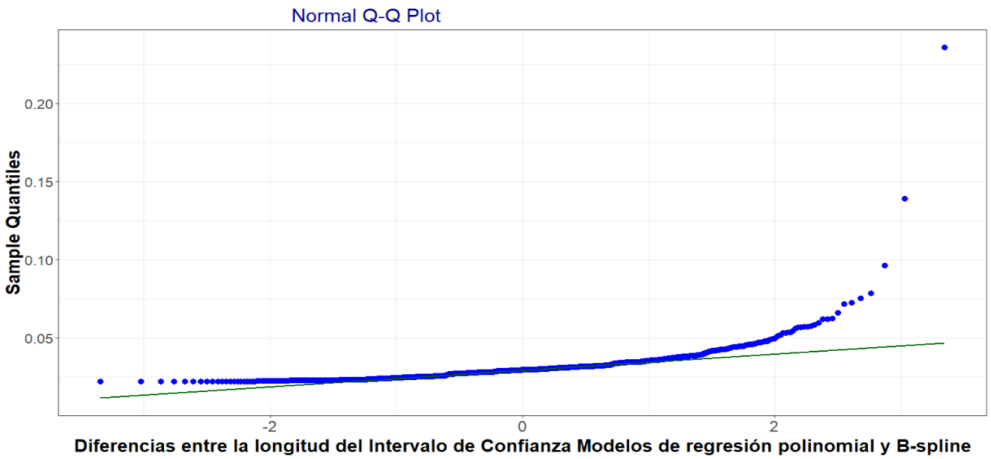

Gráfico 4-25. Ajuste normal Q-Q Plot del logartimo natural de la temperatura.
Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-25 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.6963733 y un RSE de 0.09131,

mientras que para el modelo B-spline un R^2 ajustado de 0.696761 y un RSE de 0.09125. Por lo tanto, la mejor bondad de ajuste corresponde al modelo B-spline.

Los dos modelos son válidos, no obstante, el modelo de regresión B-spline es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo polinomial. Es decir que, la relación de la radiación solar y el logaritmo natural de la temperatura en la estación meteorológica de San Antonio de Pichincha se expresa de mejor manera con el uso de la regresión no paramétrico B-spline de grado 3.

4.1.2.2 Hora del día vs Temperatura

El conjunto de datos de las variables hora del día y temperatura ambiente está conformado por 24 pares, uno por cada hora. Para el efecto se obtuvieron los valores promedios de la temperatura por hora a partir de un total de 32759 datos correspondientes al periodo comprendido entre el 23 de marzo de 2017 y 31 de diciembre de 2020. En la Tabla 4-6 se presentan los modelos de regresión polinomial y B-spline de la variable hora del día versus la temperatura ambiente:

Tabla 4-6. Modelos de regresión de la hora del día y temperatura.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Hora del día	
Respuesta:		Temperatura (ºC)	
Número de datos *n*:		24	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	6	Grado:	3 (Cúbico)
		Vértices del polígono de control:	6
Intercepto β_0	9.712	Intercepto a_0	13.4014
β_1	5.067	a_1	0
β_2	-2.245	a_2	0
β_3	0.3815	a_3	-5.7623
β_4	-0.02841	a_4	11.6989
β_5	0.0009585	a_5	2.4513

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Hora del día	
Respuesta:		Temperatura (ºC)	
Número de datos *n*:		24	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
β_6	-0.00001206	a_6	0
F	262	F	685.3
p-valor	$8.697\ x\ 10^{-16}$***	p-valor	0.000***
RSE	0.3764	RSE	0.2335
R^2 ajustado	0.985523	R^2 ajustado	0.9944296
MSE	0.141694	MSE	0.0545216
SSE	2.40879	SSE	0.9268676
RSME	0.376422	RSME	0.2334987
SST	225.1161	SST	225.1161
MAPE	1.8%	-	-
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.96802	W:	0.97476
p-valor:	0.6183	p-valor:	0.7834
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.13468	D:	0.063491
p-valor:	0.3139	p-valor:	0.9988
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			
D-W:	1.163421	D-W:	1.295452
p-valor:	0.05321	p-valor:	0.052
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	9.7938	BP:	10.808
p-valor:	0.1336	p-valor:	0.09449
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	540	p-valor	$2.133\ x\ 10^{-7}$*** (Diferencias significativas entre ambos modelos)

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial de grado 6 se muestran a continuación:

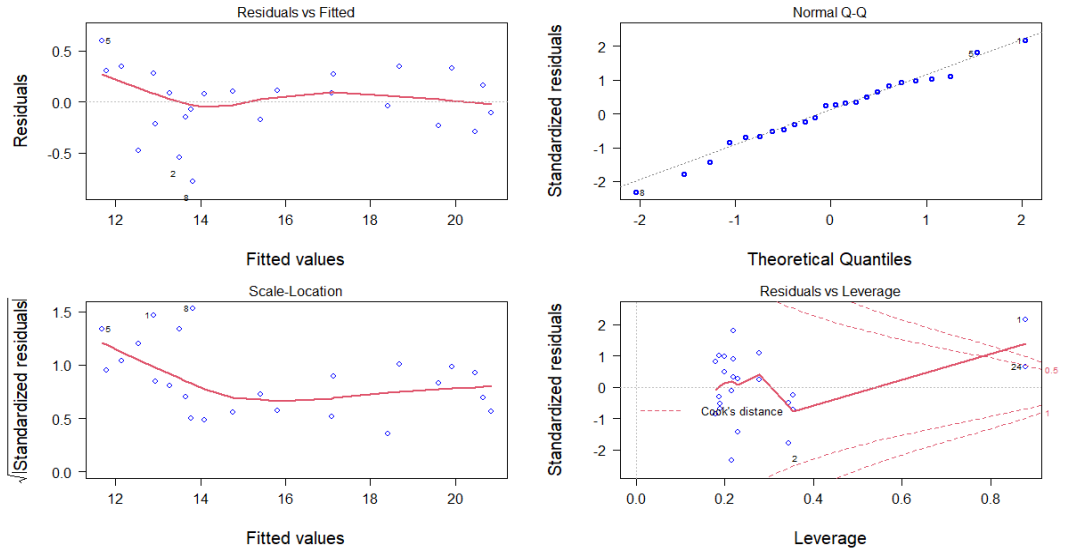

Gráfico 4-26. Gráficos para evaluar el modelo de regresión polinomial de grado 6.
Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-26 se observa que los valores predichos y los residuos no están vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente en torno a la línea roja y la mayoría está en torno a cero. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que los puntos de posición coinciden aproximadamente con la recta diagonal. El gráfico de localización de escala muestra que la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, por ende la variabilidad de las magnitudes en función de los valores ajustados es pequeña. En el gráfico del grado de influencia (Leverage) vs residuos estandarizados se observa que existe un solo valor atípico en los residuos y que está fuera de la distancia de Cook. Los gráficos sugieren que se verifican las hipótesis de los supuestos de no autocorrelación, homocedasticidad y de errores normalmente distribuidos. Por este motivo se corroboran los resultados de las pruebas estadísticas correspondientes y en consecuencia se acepta el modelo de regresión polinomial de grado 6.

La base utilizada para el modelo de regresión B-spline cúbico es la siguiente:

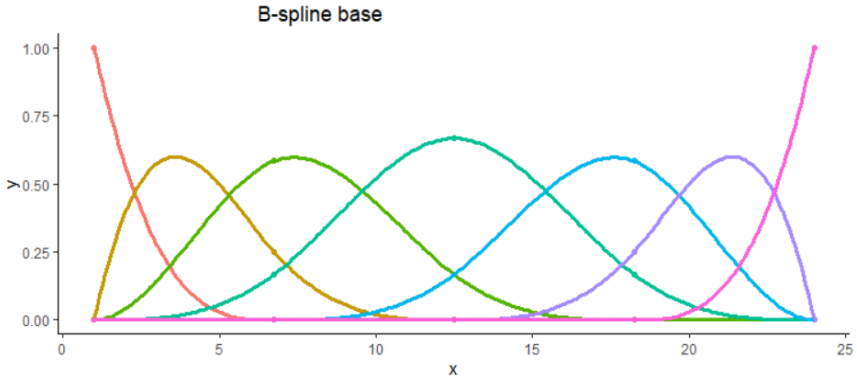

Gráfico 4-27. B-spline base para el modelo de regresión hora vs temperatura.

Elaborado por: Toalombo, B. (2021).

Igual al caso del modelo de regresión polinomial de grado 6, para el modelo B-spline cúbico, se obtuvieron los gráficos de evaluación del modelo:

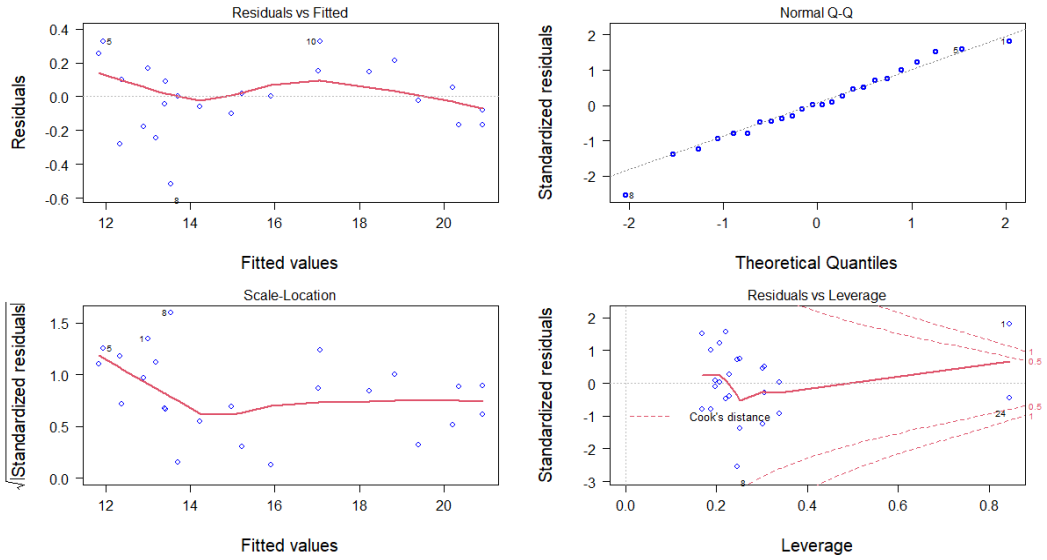

Gráfico 4-28. Q-Q Plot para el modelo de regresión B-spline cúbico.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-28 se identifica que no existe autocorrelación de los residuos, tienen homocedasticidad y se distribuyen normalmente. Por lo tanto, el modelo B-spline cúbico es idóneo.

La gráfica de dispersión de puntos de la hora del día versus la temperatura ambiente se muestra en el Gráfico 4-29, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

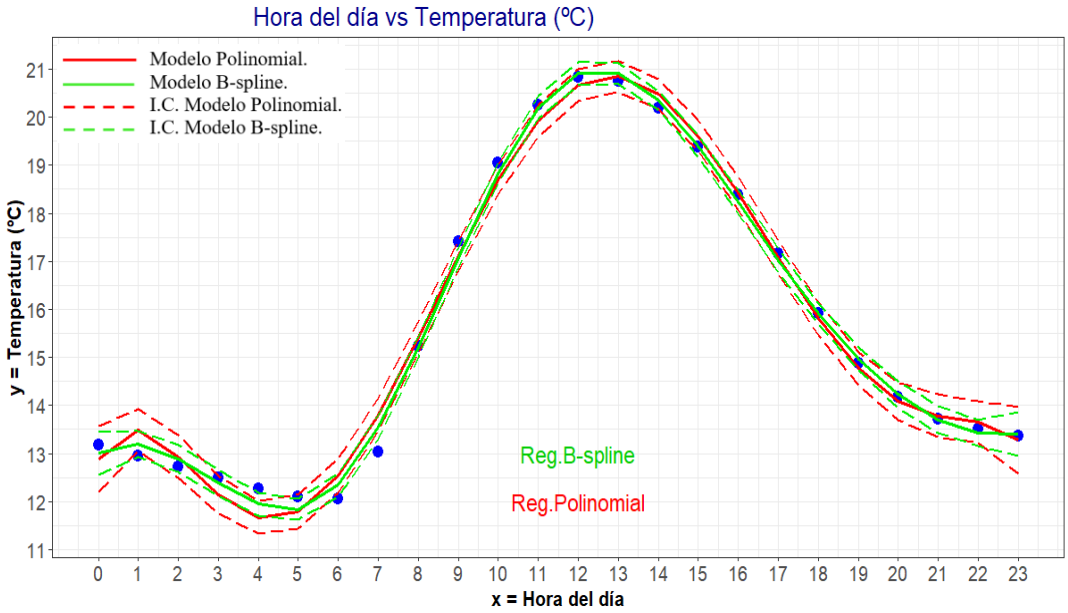

Gráfico 4-29. Modelos de regresión polinomial y B-spline de la hora del día vs. temperatura.

Elaborado por: Toalombo, B. (2021).

De acuerdo a la información que proporciona el Gráfico 4-29, la distribución de los puntos sigue una curva con forma de campana. En el intervalo comprendido entre las 6 am y las 12 del medio día la temperatura ambiente tiende a ir en aumento hasta aproximarse a los 21ºC, mientras que desde las 2 pm en adelante la tendencia es hacia la disminución de la temperatura ambiente. Las dos curvas ofrecen una buena bondad de ajuste, siendo que los intervalos de confianza al 95% cubren casi todos los puntos y la longitud de los intervalos es pequeña. El modelo de regresión polinomial es relativamente complejo dado que es de grado 6, en tanto que el B-spline no lo es al tener un grado 3. De acuerdo al análisis realizado, parece ser una mejor opción el modelo de regresión no paramétrico.

Para identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 540 (p-valor = 2.133×10^{-7}***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta

el Gráfico 4-30 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

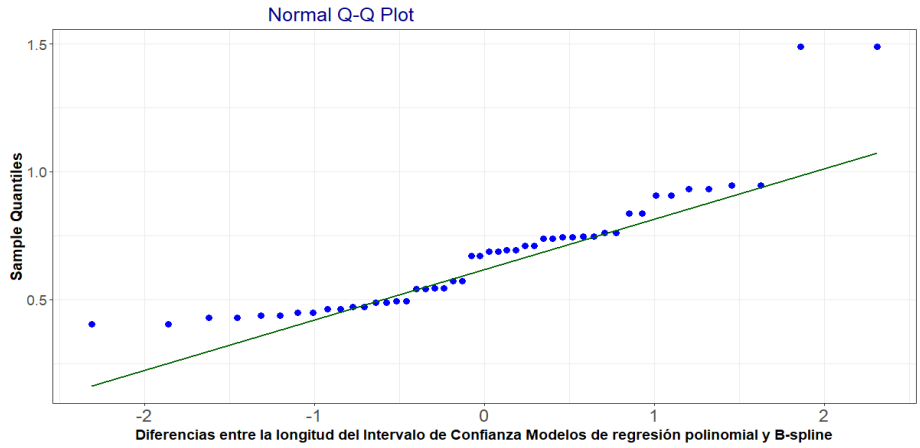

Gráfico 4-30. Ajuste normal Q-Q Plot de la variable fuerza.
Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-30 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.985523 y un RSE de 0.3764, mientras que para el modelo B-spline un R^2 ajustado de 0.9944296 y un RSE de 0.2335. Por lo tanto, la mejor bondad de ajuste corresponde al modelo B-spline.

Los dos modelos son válidos, no obstante, el modelo de regresión B-spline es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo polinomial. Es decir que la relación de la variable hora del día y la variable respuesta temperatura ambiente se expresa de mejor manera con el uso de la regresión no paramétrico B-spline de grado 3.

4.1.2.3 Hora del día vs Humedad relativa

El conjunto de datos de las variables hora del día y humedad relativa está conformado por 24 pares, uno por cada hora. Para el efecto se obtuvieron los valores promedios de la humedad relativa por hora a partir de un total de 32768 datos correspondientes al periodo comprendido entre el 23 de marzo de 2017 y 31 de

diciembre de 2020. En la Tabla 4-7 se presentan los modelos de regresión polinomial y B-spline de la variable hora del día versus la variable humedad relativa:

Tabla 4-7. Modelos de regresión de la hora del día y humedad relativa.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Hora del día	
Respuesta:		Humedad relativa (%)	
Número de datos *n*:		24	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	6	Grado:	3 (Cúbico)
		Vértices del polígono de control:	6
Intercepto β_0	103.7	Intercepto a_0	89.85988
β_1	-20.98	a_1	0
β_2	8.978	a_2	-5.59761
β_3	-1.546	a_3	15.14513
β_4	0.1168	a_4	-57.74115
β_5	-0.003977	a_5	-10.75390
β_6	5.032×10^{-5}	a_6	0
F	211	F	463.4
p-valor	5.327×10^{-15}***	p-valor	0.000***
RSE	1.924	RSE	1.303
R^2 ajustado	0.982073	R^2 ajustado	0.9917776
MSE	3.70188	MSE	1.697902
SSE	62.93196	SSE	28.86433
RSME	1.924027	RSME	1.303036
SST	4749.439	SST	4749.439
MAPE	1.94%	-	-
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.9524	W:	0.97822
p-valor:	0.3051	p-valor:	0.861
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.16473	D:	0.10814
p-valor:	0.09115	p-valor:	0.6631
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Hora del día	
Respuesta:		Humedad relativa (%)	
Número de datos *n*:		24	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
D-W:	0.9655242	D-W:	0.961553
p-valor:	0.09235	p-valor:	0.10314
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	7.1755	BP:	8.616
p-valor:	0.3049	p-valor:	0.1964
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	532	p-valor	$5.103\ x\ 10^{-7}$*** (Diferencias significativas entre ambos modelos)

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Elaborado por: Toalombo, B. (2021).

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial de grado 6 se muestran a continuación:

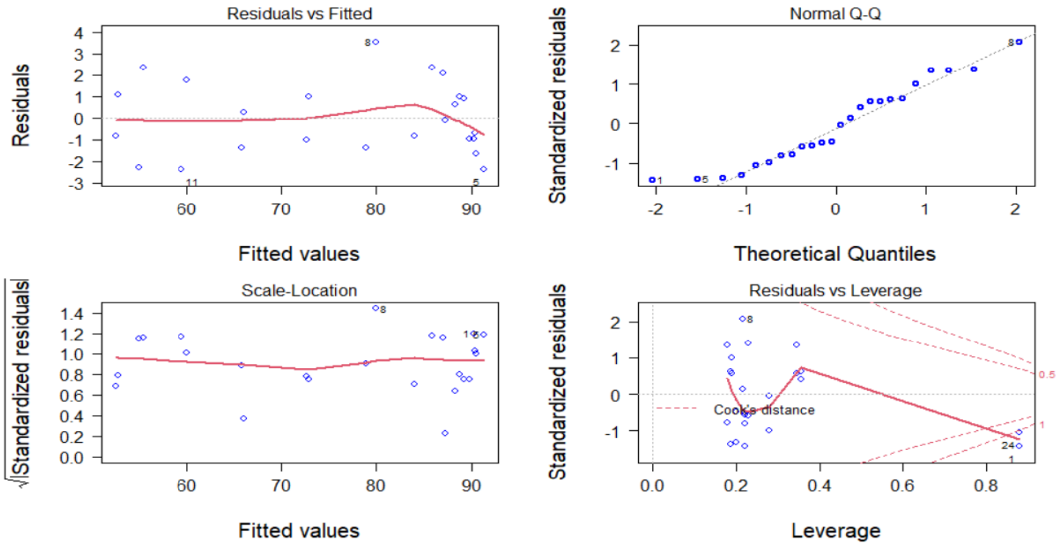

Gráfico 4-31. Gráficos para evaluar el modelo de regresión polinomial cuadrático.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-31 se observa que los valores predichos y los residuos no están vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente en torno a la línea roja. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que los puntos de posición coinciden aproximadamente con la recta diagonal. El gráfico de localización de escala muestra que la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, por tanto la variabilidad de las magnitudes es pequeña en función de los valores ajustados. En el gráfico del grado de influencia (Leverage) vs residuos estandarizados se observa que existen dos valores atípicos en los residuos y que están fuera de la distancia de Cook, lo que sugiere que podría haber valores atípicos que influyan en el modelo. Pero al revisar los resultados de las pruebas estadísticas de comprobación de las hipótesis de los supuestos de no autocorrelación, homocedasticidad y de errores normalmente distribuidos, se concluye que se cumplen los supuestos. Por esta razón se acepta el modelo de regresión polinomial de grado 6.

La base utilizada para el modelo de regresión B-spline cúbico es la siguiente:

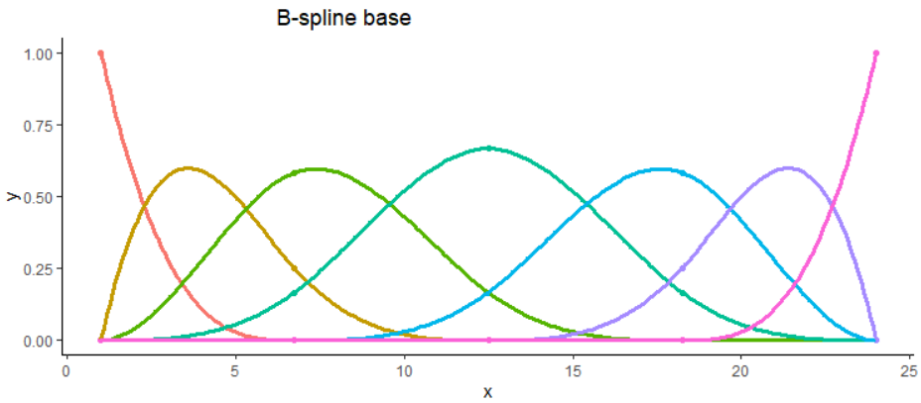

Gráfico 4-32. B-spline base para el modelo de regresión hora vs humedad.
Elaborado por: Toalombo, B. (2021).

Igual al caso del modelo de regresión polinomial de grado 6, para el modelo B-spline cúbico, se obtuvieron los gráficos de evaluación del modelo:

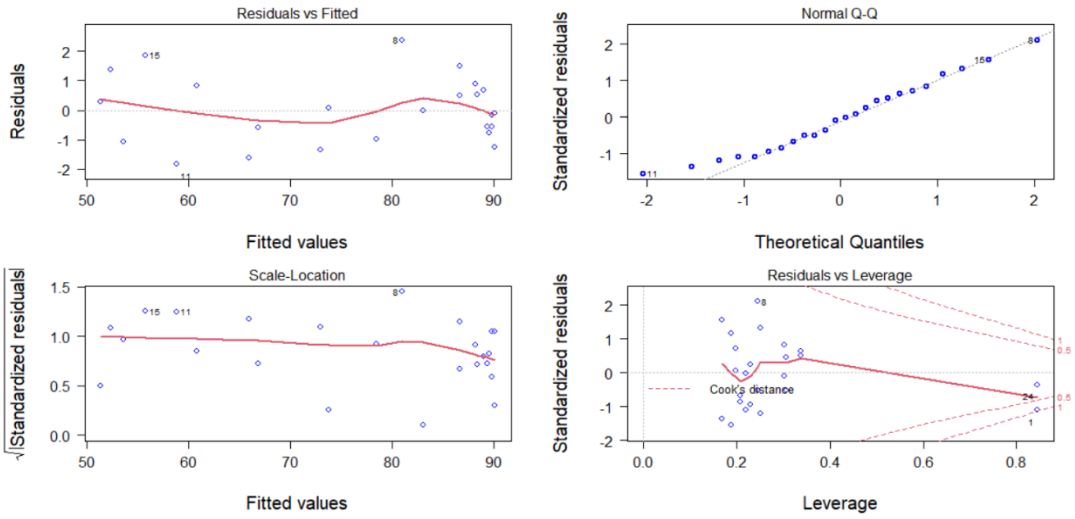

Gráfico 4-33. Gráficos para evaluar el modelo de regresión B-spline cúbico.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-33 se identifica que no existe autocorrelación de los residuos, tienen homocedasticidad y se distribuyen normalmente. Por lo tanto, el modelo B-spline cúbico es idóneo.

La gráfica de dispersión de puntos de la hora del día versus la humedad relativa se muestra en el Gráfico 4-34, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

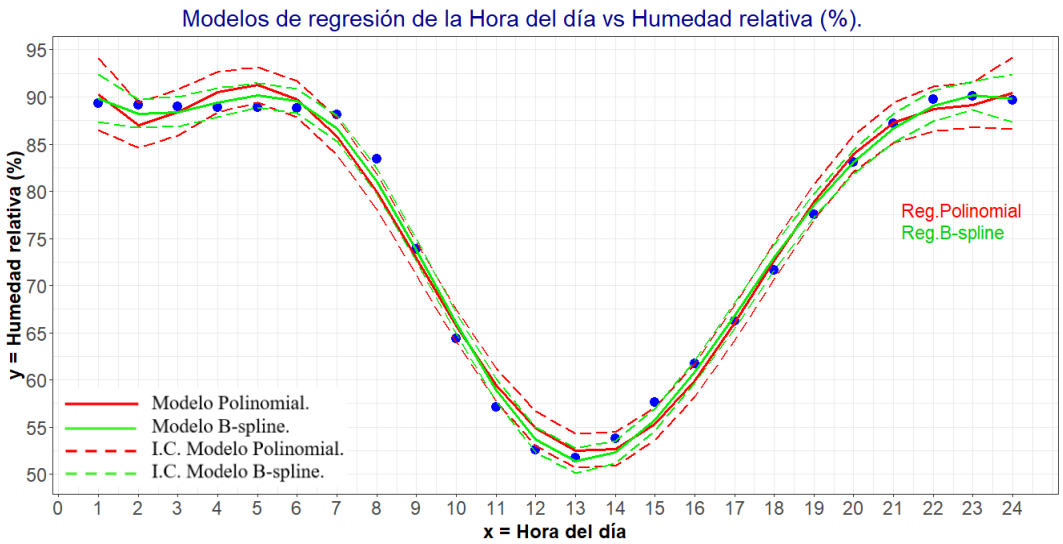

Gráfico 4-34. Modelos de regresión polinomial y B-spline de la hora de día vs humedad relativa.

Elaborado por: Toalombo, B. (2021).

De acuerdo a la información que proporciona el Gráfico 4-34, la distribución de los puntos sigue una curva con forma de campana invertida. En el intervalo comprendido entre las 6 am y las 12 del medio día la humedad relativa tiende a ir en decreciendo hasta ubicarse cerca del 50%, mientras que desde la 1 pm en adelante la tendencia es hacia la elevación de la humedad relativa cerca del valor de 90%. Las dos curvas ofrecen una buena bondad de ajuste, siendo que los intervalos de confianza al 95% cubren casi todos los puntos y la longitud de los intervalos es pequeña. El modelo de regresión polinomial es relativamente complejo dado que es de grado 6, en tanto que el B-spline no lo es al tener un grado 3. De acuerdo al análisis realizado, parece ser una mejor opción el modelo de regresión no paramétrico.

Para identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 532 (p-valor = $5.103 \; x \; 10^{-7}$***), que implica que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta el Gráfico 4-35 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

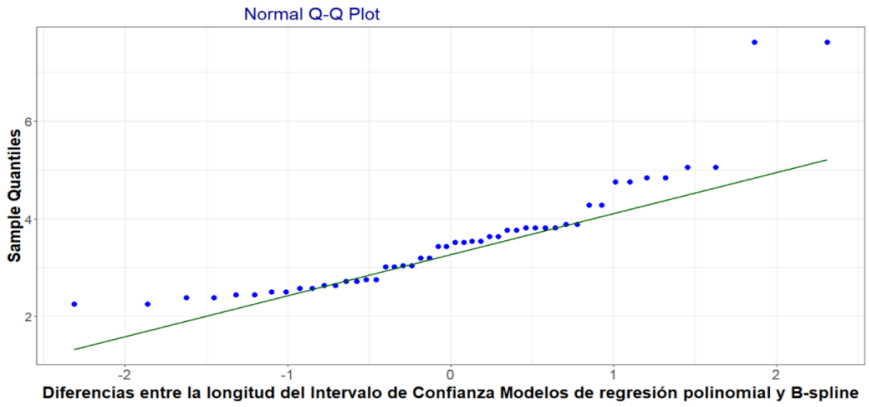

Gráfico 4-35. Ajuste normal Q-Q Plot de la variable humedad relativa.

Elaborado por: Toalombo, B. (2021).

De acuerdo al Gráfico 4-35 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se

tiene para el modelo polinomial un R^2 ajustado de 0.982073 y un RSE de 1.924, mientras que para el modelo B-spline un R^2 ajustado de 0.9917776 y un RSE de 1.303. Por lo tanto, la mejor bondad de ajuste corresponde al modelo B-spline.

Los dos modelos son válidos, no obstante, el modelo de regresión B-spline es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo polinomial. Es decir que la relación de la variable hora del día y la variable respuesta humedad relativa se expresa de mejor manera con el uso de la regresión no paramétrico B-spline de grado 3.

4.1.2.4 Hora del día vs Presión atmosférica

El conjunto de datos de las variables hora del día y presión atmosférica está conformado por 24 pares, uno por cada hora. Para el efecto se obtuvieron los valores promedios de la presión por hora a partir de un total de 32779 datos correspondientes al periodo comprendido entre el 23 de marzo de 2017 y el 31 de diciembre de 2020. En la Tabla 4-8 se presentan los modelos de regresión polinomial y B-spline de la variable hora del día versus la variable presión atmosférica:

Tabla 4-8. Modelos de regresión de la hora del día y presión.

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Hora del día	
Respuesta:		Presión (mbar)	
Número de datos *n*:		24	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
Grado:	7	Grado:	4
		Vértices del polígono de control:	7
Intercepto β_0	764.1	Intercepto a_0	765.29012
β_1	2.105	a_1	0
β_2	-1.391	a_2	-0.33896
β_3	0.3270	a_3	-3.80244
β_4	-0.03560	a_4	3.71204

Modelo de regresión			
Variables de la regresión			
Tipo de variable		**Denominación (Unidad de medida)**	
Explicativa o regresor:		Hora del día	
Respuesta:		Presión (mbar)	
Número de datos *n*:		24	
Modelo de regresión *Polinomial*		**Modelo de regresión *B-spline***	
β_5	0.001949	a_5	-5.38236
β_6	$-5.219\ x\ 10^{-5}$	a_6	-2.84074
β_7	$5.442\ x\ 10^{-7}$	a_7	0.47724
F	1075	F	890.6
p-valor	0.000***	p-valor	0.000***
RSE	0.05855	RSE	0.06432
R^2 ajustado	0.9969504	R^2 ajustado	0.99632
MSE	0.00342799	MSE	0.0041365
SSE	0.0548479	SSE	0.06618475
RSME	0.058549	RSME	0.064316
SST	25.854	SST	25.854
MAPE	0%	-	-
Test de normalidad de los residuos:			
Prueba de Shapiro-Wilk			
W:	0.9642	W:	0.92687
p-valor:	0.5284	p-valor:	0.08299
Prueba de Kolmogorov-Smirnov corregida por Lilliefors			
D:	0.11341	D:	0.17424
p-valor:	0.5886	p-valor:	0.05764
Test de residuos no correlacionados:			
Prueba de Durbin-Watson			
D-W:	1.445399	D-W:	1.174276
p-valor:	0.802	p-valor:	0.0574
Test de Homocedasticidad de residuos:			
Prueba de Breusch-Pagan estudiantilizada			
BP:	10.345	BP:	3.4937
p-valor:	0.1698	p-valor:	0.8359
Prueba no paramétrica de Wilcoxon para comprobar diferencias entre las distancias de los IC de los modelos:			
W	156	p-valor	0.006673** (Diferencias significativas entre ambos modelos)

Códigos de significancia: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1.

Los gráficos para corroborar las pruebas de evaluación de los supuestos del modelo de regresión polinomial de grado 7 se muestran a continuación:

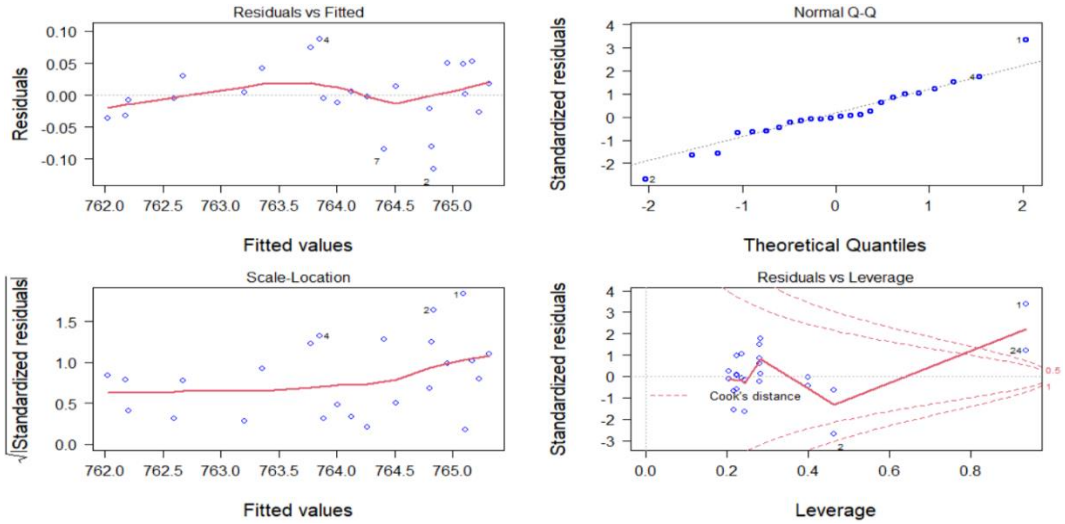

Gráfico 4-36. Gráficos para evaluar el modelo de regresión polinomial de grado 7.
Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-36 se observa que los valores predichos y los residuos no están vinculados o correlacionados, ya que los puntos están distribuidos aleatoriamente en torno a la línea roja. La gráfica Q-Q Plot de los cuantiles teóricos y residuos estandarizados muestra que casi todos los puntos de posición coinciden aproximadamente con la recta diagonal. El gráfico de localización de escala muestra que la dispersión alrededor de la línea roja no varía con los valores ajustados porque no existe una tendencia, por tanto la variabilidad de las magnitudes es pequeña en función de los valores ajustados. En el gráfico del grado de influencia (Leverage) vs residuos estandarizados se observa que existen dos valores atípicos en los residuos y que están fuera de la distancia de Cook, lo que sugiere que podría haber valores atípicos que influyan en el modelo. Pero al revisar los resultados de las pruebas estadísticas de comprobación de las hipótesis de los supuestos de no autocorrelación, homocedasticidad y de errores normalmente distribuidos, se concluye que se cumplen los supuestos. Por esta razón se acepta el modelo de regresión polinomial de grado 7.

La base utilizada para el modelo de regresión B-spline de grado 4 es la siguiente:

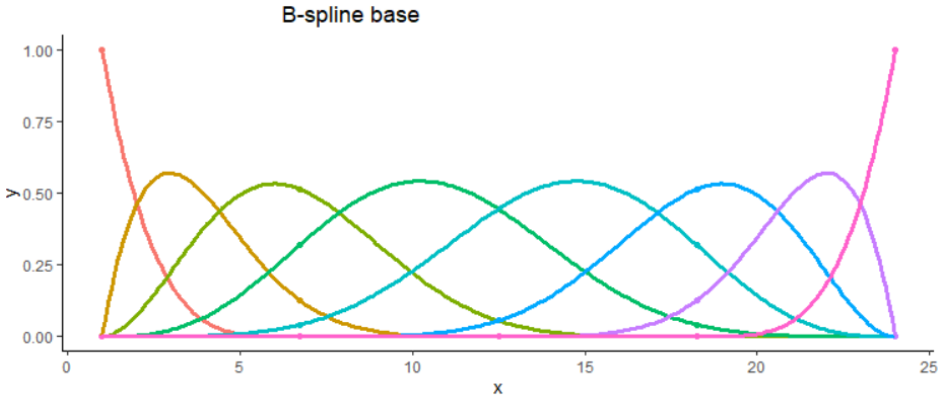

Gráfico 4-37. B-spline base para el modelo de regresión hora vs presión atmosférica.

Elaborado por: Toalombo, B. (2021).

Similar al caso del modelo de regresión polinomial de grado 7, para el modelo B-spline de grado 4, se obtuvieron los gráficos de evaluación del modelo:

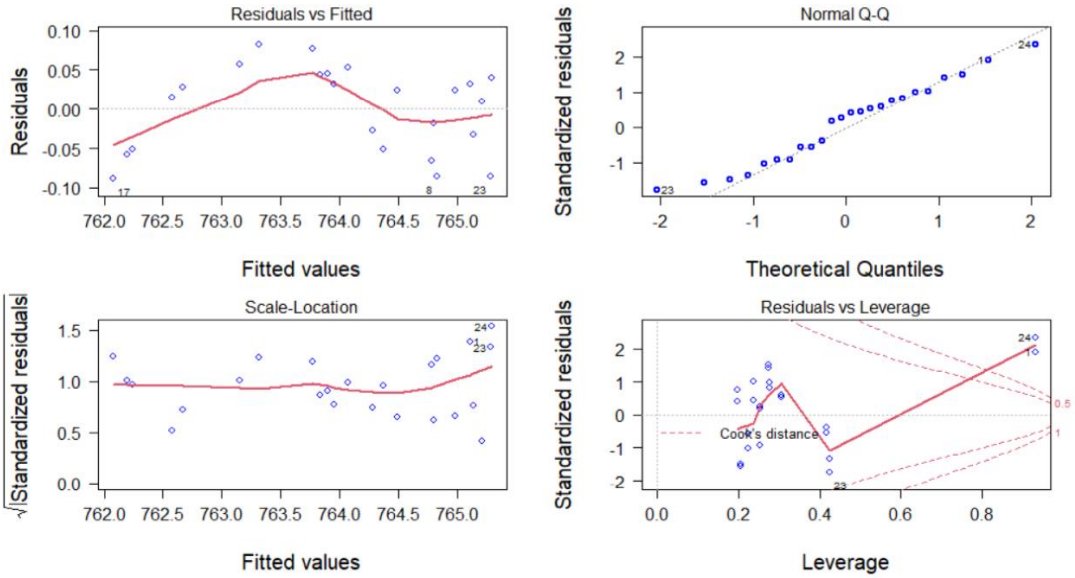

Gráfico 4-38. Gráficos para evaluar el modelo de regresión B-spline de grado 4.

Elaborado por: Toalombo, B. (2021).

En el Gráfico 4-38 se identifica que no existe autocorrelación de los residuos, tienen homocedasticidad y se distribuyen normalmente. Por lo tanto, el modelo B-spline cúbico es idóneo.

La gráfica de dispersión de puntos de la hora del día versus la presión atmosférica se muestra en el Gráfico 4-39, conjuntamente con las curvas de los dos modelos de regresión y las líneas de los intervalos de confianza al 95%:

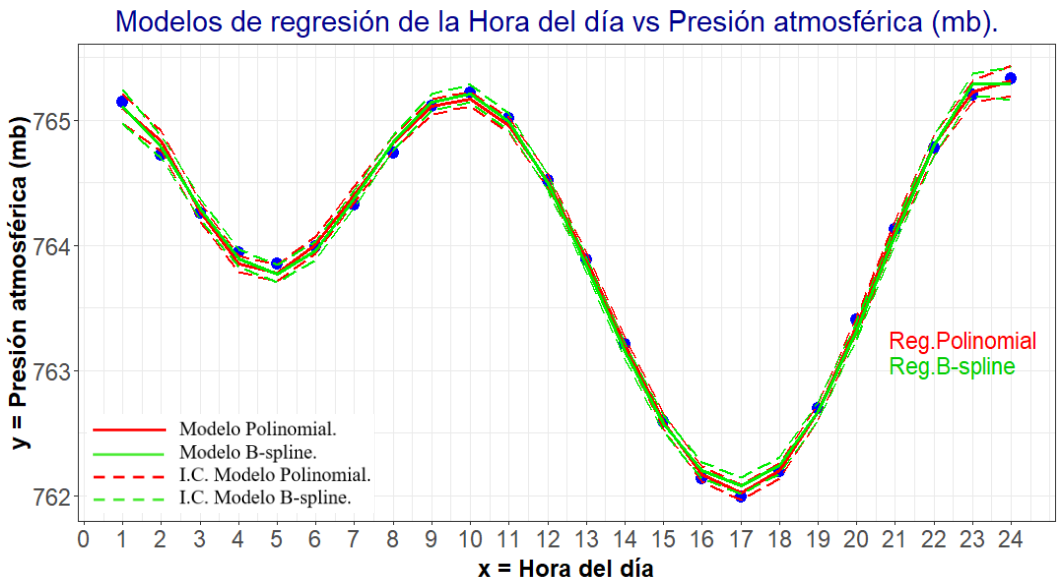

Gráfico 4-39. Modelos de regresión polinomial y B-spline de la presión atmosférica vs. temperatura.

Elaborado por: Toalombo, B. (2021).

Según la información que proporciona el Gráfico 4-39, la distribución de los puntos sigue una curva ondulada con intervalos crecientes y decrecientes. En el intervalo comprendido entre la 1 am y las 5 am la presión atmosférica tiende a decrecer entre 765.2 y 763.8 mbar; luego hay un incremento hasta 765.3 mbar a las 10 am; se produce una nueva disminución hasta 762 mbar a las 5 pm, y finalmente una subida de la presión atmosférica hasta el valor de 765.4 mbar. Las dos curvas ofrecen una buena bondad de ajuste, siendo que los intervalos de confianza al 95% cubren casi todos los puntos y la longitud de los intervalos es muy pequeña. Los dos modelos de regresión son complejos, dado que el polinomial es de grado 7 y el B-spline tiene una base de grado 4, pero ambos son idóneos.

Para identificar si existen diferencias significativas entre los dos modelos de regresión se aplicó la prueba no paramétrica de Wilcoxon, con una significancia $\alpha = 0.05$, la misma que arrojó un valor W = 156 (p-valor = 0.006673***), que indica

que los modelos de regresión polinomial y B-spline ofrecen una bondad de ajuste diferente. Para corroborar el resultado de la prueba de hipótesis, se presenta el Gráfico 4-40 del Q-Q Plot de las diferencias entre la longitud del intervalo de confianza versus los cuantiles:

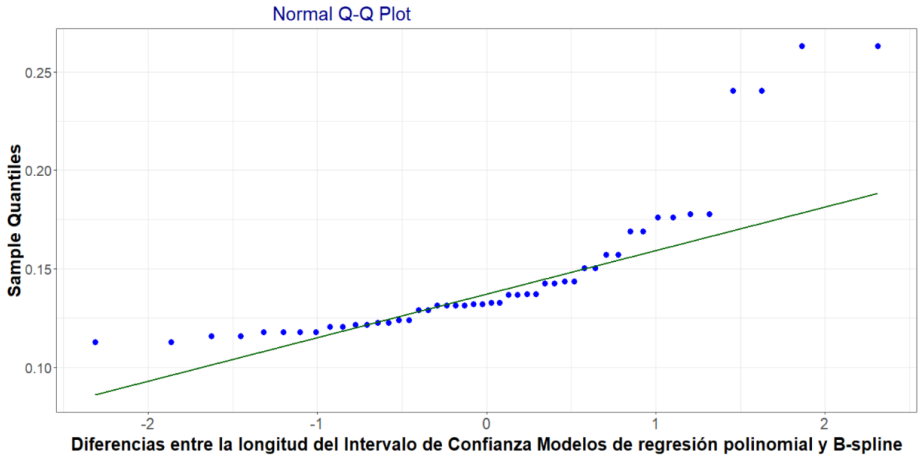

Gráfico 4-40. Ajuste normal Q-Q Plot de la variable presión atmosférica.
Elaborado por: Toalombo, B. (2021).

Según el Gráfico 4-40 se corrobora la hipótesis alterna de que existen diferencias significativas entre los dos modelos, dado que los residuos son lo suficientemente diferentes. Al comparar la bondad de ajuste de los dos modelos, se tiene para el modelo polinomial un R^2 ajustado de 0.9969504 y un RSE de 0.05855, mientras que para el modelo B-spline un R^2 ajustado de 0.99632 y un RSE de 0.06432. Por lo tanto, la mejor bondad de ajuste corresponde al modelo polinomial.

Los dos modelos son válidos, no obstante, el modelo de regresión polinomial es el que ofrece una mejor bondad de ajuste, ya que los residuos son menores que los del modelo B-spline. Es decir que la relación de la variable hora del día y la variable respuesta presión atmosférica se expresa de mejor manera con el uso de la regresión paramétrica polinomial de grado 7.

4.2 Discusión

Se han considerado dos aplicaciones en ingeniería de los modelos de regresión paramétricos polinomiales y no paramétricos B-splines. Los resultados obtenidos en la presente investigación son particulares para las situaciones específicas estudiadas y si bien puede tomarse como un referente para estudios posteriores, sin embargo, no se pueden hacer generalizaciones al respecto.

A continuación, se resumen los resultados obtenidos para el problema de la simulación de impacto de un vehículo en la parte posterior de la carrocería de un autobús y para el estudio de la relación entre las variables climatológicas en la localidad de San Antonio de Pichincha:

Tabla 4-9. Resumen general para comparación de los dos modelos de regresión.

			Modelo de regresión	
Problema	**Variables**	**Aspecto**	**Modelo de regresión *Polinomial***	**Modelo de regresión *B-spline***
Simulación de impacto vehicular entre un auto y un bus.	Velocidad vs Fuerza	Validez del modelo	✓	✓
		Grado	8	5, 7 v. p. c.
	Velocidad vs FDS	Validez del modelo	✓	✓
		Grado	5	3, 5 v. p. c.
	Velocidad vs Tiempo de impacto	Validez del modelo	✓	✓
		Grado	3	3, 6 v. p. c.
	Fuerza vs Deformación	Validez del modelo	✗	✓
		Grado	8	4, 6 v. p. c.
Situación climatológica de San Antonio de Pichincha, en Quito.	Radiación solar vs Temperatura ambiente	Validez del modelo	✓	✓
		Grado	3	3, 5 v. p. c.
	Hora del día vs Temperatura ambiente	Validez del modelo	✓	✓
		Grado	6	3, 6 v. p. c.

Modelo de regresión				
Problema	**Variables**	**Aspecto**	**Modelo de regresión *Polinomial***	**Modelo de regresión *B-spline***
	Hora del día vs Humedad relativa	Validez del modelo	✓	✓
		Grado	6	3, 6 v. p. c.
	Hora del día vs presión atmosférica	Validez del modelo	✓	✓
		Grado	7	4, 7 v. p. c.

✓ Modelo válido.

✘ Modelo no válido.

v. p. c. vértices del polígono de control.

Sombreado en color azul el modelo más apropiado para cada caso.

Elaborado por: Toalombo, B. (2021).

Los resultados indican que ambos modelos de regresión brindan una buena bondad de ajuste para un volumen de datos relativamente pequeño (n < 100), ya que es posible identificar y eliminar con relativa facilidad los datos atípicos que pueden ocasionar una distorsión en el modelo (en caso de existir). En cambio, cuando el volumen de datos es grande, resulta difícil hallar los datos atípicos y determinar si la eliminación de los mismos puede mejorar el modelo, pero sin alterar la relación de las variables. El principal problema que se presenta es el incumplimiento de los supuestos de la distribución normal de los residuos, no autocorrelación y homocedasticidad de los mismos.

En el presente estudio, para el caso de la relación entre la radiación solar y la temperatura ambiente, a pesar de realizarse una transformación logarítmica de la variable respuesta (temperatura) y de suprimirse algunos valores atípicos, no fue posible conseguir un modelo que satisfaga los supuestos requeridos para el modelo paramétrico. Por esta razón, se consideraron únicamente los datos disponibles del período comprendido entre noviembre y diciembre del año 2020 y no fue posible establecer un intervalo de confianza al 95%, ya que de hacerlo una importante proporción de los datos quedarían fuera del intervalo.

En consideración de las dificultades expuestas en el párrafo anterior, se establecieron ocho modelos de regresión simples polinomiales e igual número de modelos de regresión B-splines, aunque en la práctica se analizaron una mayor cantidad de relaciones, que no fueron factibles bien porque no presentaron una correlación significativa y/o una buena bondad de ajuste.

En cuanto a la bondad de ajuste, el coeficiente de determinación R^2 ajustado en todos los casos analizados presentó un valor cercano a 1 y el test F de bondad de ajuste de la tabla ANOVA arrojó que los modelos son idóneos sin mayor complicación.

En términos generales se determinó que los modelos de regresión paramétricos polinomiales se ajustan bien cuando las curvas tienen forma parabólica o siguen un patrón sin cambios abruptos de curvatura. Por su parte, los modelos de regresión no paramétricos B-spline brindan un mejor ajuste cuando las curvas tienen forma de campana con cambios de curvatura más abruptos. De acuerdo a la simplicidad de los modelos, se consideraron modelos polinomiales de hasta grado 8 y B-splines de hasta grado 5 con 7 vértices del polígono de control.

En la práctica no se suelen establecer modelos de regresión con polinomios de grado elevado, debido a la complejidad que representa la obtención de los valores de la variable respuesta a partir de los valores de la variable regresora. Aunque este problema es subestimado al disponer de programas estadísticos y de propósito de cálculo general que tienen un poder suficiente como para procesar los cálculos de forma veloz. Otra desventaja de obtener modelos complejos está en el hecho de que no resulta factible determinar una relación de crecimiento directa o inversamente proporcional entre las variables. Por este motivo los modelos de regresión estudiados pueden resultar útiles para la predicción de los valores de las variables dependientes, pero no para resolver problemas de clasificación que son de interés en el ámbito de aprendizaje automático, que suelen establecerse con regresiones lineales [8].

Desde el punto de vista de la estructura de los dos modelos, la regresión polinomial es más sencilla de comprender dado que básicamente consiste en una expresión algebraica con la variable elevada a distintos exponentes en forma progresiva, ajustada con el uso de coeficientes. De su parte la regresión B-spline está estructurada por una base formada por curvas splines que también tienen polinomios y en los que es necesario establecer un determinado número de vértices del polígono de control.

Finalmente se debe destacar la utilidad práctica de la aplicación de los modelos de regresión en los problemas estudiados. En la simulación de impacto como parte del diseño de la estructura de la carrocería de autobuses, se detecta que la fuerza del impacto crece ostensiblemente a medida que se va incrementando la velocidad del vehículo que está en movimiento, el FDS de diseño también aumenta con la velocidad al igual que el tiempo de impacto, pero de forma más lenta. El FDS cuanto mayor es implica que se necesitaría sobredimensionar el diseño, lo que se traduce en utilizar un espesor mayor para el diseño de la estructura del autobús [32]. La relación de la fuerza de impacto con la deformación de la estructura del autobús fue complicada de modelar y se observó que a partir de la magnitud de 10^{11} N de fuerza la deformación se mantuvo casi constante.

Respecto a la relación de las variables climatológicas, tomando en cuenta los datos disponibles en el período comprendido entre los años 2017 y 2020 y el comportamiento de las magnitudes en promedio según la hora del día. Estas relaciones se expresaron de mejor forma con el modelo de regresión B-spline para el caso de la temperatura y humedad relativo, no así para la presión atmosférica que se define con el modelo de regresión polinomial. Los resultados obtenidos para modelar la relación de las variables climatológicas concuerdan con los alcanzados en la investigación desarrollada por Chariguamán en el año 2015 [10], acerca de las condiciones en la estación meteorológica ubicada en la Escuela Superior Politécnica de Chimborazo de la ciudad de Riobamba, en la que se concluyó que los modelos más idóneos fueron los de regresión B-spline de grado 3 por sobre los de regresión polinomial de grado 7.

CAPÍTULO V

CONCLUSIONES, RECOMENDACIONES, BIBLIOGRAFÍA Y ANEXOS

5.1 Conclusiones

- Se determinó que los modelos estadísticos de regresión paramétrica polinomial y no paramétrica B-splines tienen sus propias particularidades. En el primer caso, demanda el cumplimiento de supuestos para los residuos: distribución normal, no autocorrelación y homocedasticidad. La estructura de un modelo de regresión polinomial consiste en una expresión algebraica con la variable elevada a exponentes consecutivos de entre los cuales el mayor determina el grado del polinomio y para el ajuste se emplean diferentes coeficientes para cada término. La regresión B-spline está estructurada por una base formada por curvas splines y en los que es necesario establecer un determinado número de vértices del polígono de control.

- Respecto a la simulación de impacto entre un vehículo y un autobús, se estableció como variable regresora a la velocidad del auto pequeño en el momento de la colisión, cuyos valores oscilaron entre 45 y 95 km/h. Según la distribución de los puntos de la velocidad versus la fuerza, son directamente proporcionales, la elevación de la velocidad provocó que la fuerza tienda a incrementarse ostensiblemente, con un rango de valores fluctuantes entre 11935 y $2.47 \, x \, 10^{12}$ N. La distribución de los puntos de la velocidad versus el FDS y de la velocidad vs el tiempo de impacto también expresaron relaciones directamente proporcionales, siendo que los rangos de datos fluctuaron entre 1 y 33.66 para el FDS, y entre 0 y 0.67 segundos para el tiempo. La distribución de puntos de la fuerza de impacto versus la deformación en la estructura es directamente proporcional, aunque la deformación tiende a presentar incrementos menores conforme se eleva la fuerza de impacto. El rango de datos de la deformación estuvo comprendido entre 0 y 13.11 mm.

- En cuanto a la relación de las variables climatológicas, la distribución de puntos de la radiación solar versus la temperatura ambiente mostró que son directamente proporcionales, siendo que los rangos de valores fluctuaron entre 0 y 1237.24 W/m^2, y entre 8.28 y 25.6ºC, respectivamente. La relación de la hora del día y la temperatura ambiente promedio muestra que los puntos se distribuyen en forma de campana con valores que oscilan entre 12 y 21ºC, los valores más altos corresponden al medio día. De igual manera la relación de la hora del día y la humedad relativa promedio muestra que los puntos se distribuyen en forma de campana invertida con valores que oscilan entre 52.5 y 90%, los valores más bajos corresponden al medio día. En tanto que la relación de la hora del día y la presión atmosférica promedio muestra que los datos fluctúan formado una curva ondulada con intervalos crecientes y decrecientes, con un rango de datos de entre 765.2 y 763.8 mbar.

- En el problema de la simulación de impacto entre un vehículo y un autobús, la relación de la velocidad y la fuerza se explica de manera óptima mediante un modelo de regresión polinomial de grado 8. La relación de la velocidad y el FDS mediante un modelo de regresión polinomial de grado 5. La relación de la velocidad y el tiempo de impacto se explica con un modelo de regresión polinomial cúbico. La fuerza y deformación se relacionan con un modelo de regresión B-spline de grado 4, con 6 vértices del polígono de control.

- La relación de la radiación solar y la temperatura ambiente en las horas de sol durante el período comprendido entre noviembre y diciembre de 2020 en San Antonio de Pichincha se explica a través de un modelo de regresión B-splines cúbicos con 5 vértices del polígono de control. Los comportamientos de la temperatura ambiente y de la humedad relativa durante el transcurso del día se explican mediante modelos de regresión B-splines cúbicos con 6 vértices del polígono de control. Mientras que el comportamiento de la presión atmosférica en el transcurso del día se explica mediante un modelo de regresión polinomial de grado 7.

- Los modelos de regresión paramétricos polinomiales se ajustan bien cuando las curvas tienen forma parabólica o siguen un patrón sin cambios abruptos de curvatura, adaptándose mejor a las relaciones de la simulación de impacto vehicular. Por su parte, los modelos de regresión no paramétricos B-splines brindan un mejor ajuste cuando las curvas tienen forma de campana con cambios de curvatura más abruptos, adaptándose mejor a las relaciones entre las variables climatológicas.

5.2 Recomendaciones

- En la simulación de impacto para el diseño de las carrocerías de autobuses se recomienda hacer numerosas pruebas, considerando el desarrollo de experimentos bajo diferentes escenarios, tanto para los vehículos que participen en la colisión, como para varios materiales constructivos.

- Hacer una revisión del comportamiento climático principalmente en las variables radiación solar y temperatura, a fin de establecer criterios para la eliminación de los datos atípicos y aberrantes. De esta manera se podrá contar con datos que permitan establecer modelos de regresión polinomiales o B-spline, ya sean simples o múltiples.

- Realizar estudios de las variables climatológicas con datos de otras estaciones meteorológicas de localidades con características diferentes a las de la ciudad de Quito, para identificar si los modelos aquí propuestos pueden ser generalizables o si por el contrario solamente se adaptan a las condiciones particulares estudiadas.

- Desarrollar investigaciones destinadas a comparar los dos modelos aquí estudiados, pero destinados a otras aplicaciones, con la finalidad de corroborar los resultados obtenidos en el presente trabajo.

5.3 Bibliografía

[1] L. Fahrmeir, T. Kneib, S. Lang, y B. Marx, *Regression. Models, Methods and Applications*, 1.ª ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-34333-9.

[2] E. Ponsot-Balaguer, «Tema 3: Modelos Lineales. Segunda parte: Regresión simple», presentado en Campamento de Datos, 2020.

[3] E. Ostertagová, «Modelling using Polynomial Regression», *Procedia Engineering*, vol. 48, pp. 500-506, 2012, doi: 10.1016/j.proeng.2012.09.545.

[4] N. E. Helwig, «Regression with Polynomials and Interactions», 2017. [En línea]. Disponible en: http://users.stat.umn.edu/~helwig/notes/polyint-Notes.pdf

[5] S. P. Verma, «Evaluation of polynomial regression models for the Student t and Fisher F critical values, the best interpolation equations from double and triple natural logarithm transformation of degrees of freedom up to 1000, and their applications to quality control in science and engineering», *Revista Mexicana de Ciencias Geológicas*, vol. 26, n.º 1, pp. 79-92, 2009.

[6] D. C. Montgomery y G. C. Runger, *Applied statistics and probability for engineers*, 3.ª ed. New York, USA: John Wiley & Sons, Inc., 2003.

[7] T. W. Anderson, «The choice of the degree of a polynomial regression as a multiple decision problem», *The Annals of Mathematical Statistics*, pp. 255-265, 2006.

[8] A. Peˇckov, «A machine learning approch to polynomial regression», Tesis Doctoral, Joˇzef Stefan International Postgraduate School, Ljubljana, Slovenia, 2012. [En línea]. Disponible en: http://kt.ijs.si/theses/phd_aleksandar_peckov.pdf

[9] P. Bruce, A. Bruce, y P. Gedeck, *Practical Statistics for Data Scientists*, 2.ª ed. Sebastopol, CA, USA: O'Reilly Media, Inc., 2020.

[10] N. E. Chariguamán-Maurisaca, «Análisis y ajustes, con modelos de regresión B-splines utilizando el software R, en las variables climatológicas: Temperatura, humedad, radiación y velocidad del viento, de la Estación Meteorológica Solar del Centro de Energías Alternativas, Facultad de Ciencias, Escuela de Física y Matemática», Tesis de Maestría, Escuela Superior Politécnica del Chimborazo, Riobamba, Ecuador, 2015.

[11] J. S. Racine, «A primer on Regression Splines». 2019.

[12] S. Imoto y S. Konishi, «B-spline nonparametric regression models and information criteria», en *Proceedings of 2nd Int. Symp. on Frontiers of Time Series Model*, Fukuoka, Japan, 2000, pp. 240-241.

[13] M. Paluszny, H. Prautzsch, y W. Böhm, *Métodos de Bézier y B-splines*. Karlsruhe: Univ.-Verl, 2005.

[14] J. S. Racine, «The crs Package». 2021. Accedido: abr. 03, 2021. [En línea]. Disponible en: https://cran.r-project.org/web/packages/crs/vignettes/crs.pdf

[15] J. Aparicio, M. Martínez, y J. Morales, *Modelos lineales aplicados en R*, 1.ª ed. Elche, España, 2006. Accedido: abr. 29, 2021. [En línea]. Disponible en: https://umh3067.edu.umh.es/wp-content/uploads/sites/240/2013/02/Modelos-Lineales-Aplicados-en-R.pdf

[16] I. Pedrosa, J. Juarros-Basterretxea, A. Robles-Fernández, J. Basteiro, y E. García-Cueto, «Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar?», *Univ Psychol*, vol. 14, n.º 1, pp. 245-254, 2014, doi: 10.11144/Javeriana.upsy14-1.pbad.

[17] J. C. Alonso y S. Montenegro, «Estudio de Monte Carlo para comparar 8 pruebas de normalidad sobre residuos de mínimos cuadrados ordinarios en presencia de procesos autorregresivos de primer orden», *Estudios Gerenciales*, vol. 31, n.º 136, pp. 253-265, 2015, doi: 10.1016/j.estger.2014.12.003.

[18] V. Díaz, «Errores estadísticos frecuentes al comparar dos poblaciones independientes», *Rev. chil. nutr.*, vol. 36, n.º 4, pp. 1136-1138, 2009, doi: 10.4067/S0717-75182009000400011.

[19] J. Durbin y G. S. Watson, «Testing for Serial Correlation in Least Squares Regression. III», *Biometrika*, vol. 58, n.º 1, pp. 1-19, 1971, doi: 10.2307/2334313.

[20] K. Beyene y S. Bekele, «Assessing Univariate and Multivariate Homogeneity of Variance: A Guide For Practitioners», *Journal of Mathematical Theory and Modeling*, vol. 6, n.º 5, pp. 13-17, 2016.

[21] R. D. Cook y S. Weisberg, «Graphics for Assessing the Adequacy of Regression Models», *Journal of the American Statistical Association*, vol. 92, n.º 438, pp. 490-499, 1997, doi: 10.1080/01621459.1997.10474002.

[22] R. M. Sakia, «The Box-Cox Transformation Technique: A Review», *The Statistician*, vol. 41, n.º 2, pp. 169-178, 1992, doi: 10.2307/2348250.

[23] T. Colliau, G. Rogers, Z. Hughes, y C. Ozgur, «MatLab vs. Python vs. R», *Journal of Data Science*, vol. 15, n.º 3, pp. 355-372, 2017.

[24] M. Yli-Heikkilä, «Data Science Languages», p. 2, 2019.

[25] R. Jibson, «Regression models for estimating coseismic landslide displacement», *Engineering Geology*, vol. 91, n.º 2-4, pp. 209-218, 2007, doi: 10.1016/j.enggeo.2007.01.013.

[26] P. D. Bruce y M. G. Kellett, «Modelling and Identification of Nonlinear Aerodynamic Functions using B-splines», pp. 907-912, 1998.

[27] M. Durbán, «Splines con Penalizaciones: Teoría y aplicaciones». 2007.

[28] D. García-Sánchez, «Control estadístico y modelos de regresión lineal: una forma práctica de control de puentes», Tesis Doctoral, Universidad de Cantabria, Santander, España, 2016.

[29] F. Gonzalez-Herrera, C. Lemus-Olalde, A. Ochoa-Zezzatti, y C. Lara-Alvarez, «Los mejores modelos polinomiales para recalibración de dispositivos de seguimiento ocular», 2019. Accedido: jul. 26, 2020. [En línea]. Disponible en: https://www.researchgate.net/profile/Fernando_Gonzalez_Herrera/publication/332766507_Los_mejores_modelos_polinomiales_para_recalibracion_de_dispositivos_de_seguimiento_ocular/links/5cc89e794585156cd7bd9d53/Los-mejores-modelos-polinomiales-para-recalibracion-de-dispositivos-de-seguimiento-ocular.pdf

[30] M. Hoch, G. Fleischmann, y B. Girod, «Modeling and animation of facial expressions based on B-Splines», *The Visual Computer*, vol. 11, n.º 2, pp. 87-95, feb. 1994, doi: 10.1007/BF01889979.

[31] C. Jurewicz, A. Sobhani, J. Woolley, J. Dutschke, y B. Corben, «Exploration of Vehicle Impact Speed – Injury Severity Relationships for Application in Safer Road Design», *Transportation Research Procedia*, vol. 14, pp. 4247-4256, 2016, doi: 10.1016/j.trpro.2016.05.396.

[32] D. Cárdenas, J. Escudero, S. Quizhpi, y M. Amaya, «Propuesta de diseño estructural para buses de carrocería interprovincial», *Ingenius. Revista de Ciencia y Tecnología*, vol. 11, n.º 1, pp. 42-52, 2014.

[33] M. Portillo, R. Chacón, M. Moreno, y F. Bongiorno, «Simulación y análisis de una prueba de choque de un automóvil tipo deportivo, utilizando un software basado en el método de los elementos finitos», *Revista Ciencia e Ingeniería*, vol. 32, n.º 1, pp. 69-77, 2011.

[34] A. Pikûnas, V. Pumputis, y V. Sadauskas, «The influnce of vehicles speed on accident rates and their consequences», *Transport*, vol. 19, n.º 1, pp. 15-25, 2004.

[35] F. Portilla, *Agroclimatología del Ecuador*, 1.ª ed. Quito, Ecuador: Editorial Universitaria Abya-Yala, 2018.

[36] S. Ramos-Herrera, R. Bautista-Margulis, y A. Valdez-Manzanilla, «Estudio estadístico de la correlación entre contaminantes atmosféricos y variables meterorológicas en la zona norte de Chiapas, México», *Universidad y Ciencia*, vol. 26, n.º 1, pp. 65-80, 2010.

[37] B. W. Yap y C. H. Sim, «Comparisons of various types of normality tests», *Journal of Statistical Computation and Simulation*, vol. 81, n.º 12, pp. 2141-2155, 2011, doi: 10.1080/00949655.2010.520163.

[38] M. Saculinggan y E. A. Balase, «Empirical Power Comparison of Goodness of Fit Tests for Normality in the Presence of Outliers», *J. Phys.: Conf. Ser.*, vol. 435, pp. 1-11, 2013, doi: 10.1088/1742-6596/435/1/012041.

[39] P. Dalgaard, *Introductory statistics with R*, 2nd ed. New York, USA: Springer, 2008.

[40] A. J. Dobson y A. G. Barnett, *An Introduction to Generalized Linear Models*, 4th ed. Boca Raton: CRC Press, Taylor & Francis Group, 2018.

[41] A. C. Rencher y G. B. Schaalje, *Linear models in Statistics*, 2nd ed. Hoboken, N.J: Wiley-Interscience, 2008.

[42] N. Martinez-Luna, «Curvas y Superficies B-splines», Tesis de Pregrado, Universidad Tecnológica de la Mixteca, Huajuapán de León, México, 2015.

[43] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, 2.ª ed. Marcel Dekker, Inc., 1999. Accedido: jul. 26, 2020. [En línea]. Disponible en: http://read.pudn.com/downloads153/doc/672818/Eubank(1999)NonparametricRegressionandSplineSmoothing2ed.pdf

[44] S. Konishi, «Generalised information criteria in model selection», *Biometrika*, vol. 83, n.º 4, pp. 875-890, dic. 1996, doi: 10.1093/biomet/83.4.875.

[45] P. H. C. Eilers y B. D. Marx, «Flexible smoothing with B-splines and penalties», *Statist. Sci.*, vol. 11, n.º 2, pp. 89-121, may 1996, doi: 10.1214/ss/1038425655.

[46] G. Rodríguez, «Smoothing and Non-Parametric Regression», p. 12, 2001.

[47] S. M. El-sayed, M. R. Abonazel, y M. M. Seliem, «B-spline Speckman Estimator of Partially Linear Model», *International Journal of Systems Science and Applied Mathematics*, vol. 4, n.º 4, pp. 53-59, 2019, doi: 10.11648/j.ijssam.20190404.12.

[48] B. A. Craig, «Nonparametric Regression», 2019. [En línea]. Disponible en: https://www.stat.purdue.edu/~bacraig/notes526/topic16a.pdf

[49] P. H. C. Eilers y B. D. Marx, «Splines, knots, and penalties», *WIREs Comp Stat*, vol. 2, n.º 6, pp. 637-653, nov. 2010, doi: 10.1002/wics.125.

[50] G. L. Griepentrog, J. M. Ryan, y L. D. Smith, «Linear Transformations of Polynomial Regression Models», *The American Statistician*, vol. 36, n.º 3a, pp. 171-174, ago. 1982, doi: 10.1080/00031305.1982.10482822.

[51] H. Dette, «Discrimination Designs for Polynomial Regression on Compact Intervals», *The Annals of Statistics*, vol. 22, n.º 2, pp. 890-903, 1994.

[52] C. W. Guerra, A. Cabrera, y L. Fernández, «Criterios para la selección de modelos estadísticos en la investigación científica», *Revista Cubana de Ciencia Agrícola*, vol. 37, n.º 1, pp. 3-10, 2003.

[53] G. Celant y M. Broniatowski, *Interpolation and Extrapolation Optimal Designs. 1, Polynomial Regression and Approximation Theory*, 1.ª ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2016. [En línea]. Disponible en: http://93.174.95.29/main/6423579D926B173CE5D0FDA3B4FD3B6A

[54] F. Schönbrodt, «Testing fit patterns with polynomial regression models», p. 20, 2016.

[55] D. K. Dalal y M. J. Zickar, «Some Common Myths about Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression», *Organizational Research Methods*, vol. 15, n.º 3, pp. 339-362, jul. 2012, doi: 10.1177/1094428111430540.

[56] H. Vega-Mejía, «Función de Regresión Polinomial para la Estimación del Volumen de la Laguna Palcacocha en Sus Diferentes Niveles de Cota», *Revista de Glaciares y Ecosistemas de Montaña*, n.º 3, pp. 59-66, 2017.

[57] J. R. Edwards y M. E. Parry, «On the use of polynomial regression as an alternative to difference scores in organizational research», *Academy of Management Journal*, vol. 36, n.º 6, pp. 1577-1613, 1993.

[58] K. Höllig, *Finite Element Methods with B-Splines*, 1.ª ed. SIAM, 2003.

[59] F. Trebuňa, E. Ostertagová, P. Frankovský, y O. Ostertag, «Application of Polynomial Regression Models in Prediction of Residual Stresses of a Transversal Beam», *American Journal of Mechanical Engineering*, vol. 4, n.º 7, pp. 247-251, 2016, doi: 10.12691/ajme-4-7-3.

[60] BIOST 515, «Polynomial regression», 2004. Accedido: jul. 27, 2020. [En línea]. Disponible en: https://courses.washington.edu/b515/l10.pdf

[61] J. A. Stimson, E. G. Carmines, y R. A. Zeller, «Interpreting Polynomial Regression», *Sociological Methods & Research*, vol. 6, n.º 4, pp. 515-524, may 1978, doi: 10.1177/004912417800600405.

[62] R. Lockhart, «Polynomial regression», 2008. Accedido: jul. 27, 2020. [En línea]. Disponible en: http://people.stat.sfu.ca/~lockhart/richard/350/08_2/lectures/Polynomial/web.pdf

[63] A. M. Anile, G. Gallo, M. Spagnuolo, y S. Spinello, «Modeling uncertain data with fuzzy B-splines», *Fuzzy Sets and Systems*, pp. 397-410, 2000.

[64] P. Kempthorne, «Regression Analysis», 2013. [En línea]. Disponible en: https://ocw.mit.edu/courses/mathematics/18-s096-topics-in-mathematics-with-applications-in-finance-fall-2013/lecture-notes/MIT18_S096F13_lecnote6.pdf

[65] P. E. DeWitt, S. MaWhinney, y N. E. Carlson, «cpr: An R Package for Finding Parsimonious B-Spline Regression Models via Control Polygon Reduction and Control Net Reduction», *arXiv:1705.04756 [stat]*, may 2017, Accedido: oct. 15, 2020. [En línea]. Disponible en: http://arxiv.org/abs/1705.04756

[66] D. C. Montgomery, E. A. Peck, y G. G. Vining, *Introduction to Linear Regression Analysis*, 5.ª ed. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2012.

[67] *Norma Ecuatoriana de Calidad del Aire*. 2015. Accedido: feb. 18, 2021. [En línea]. Disponible en: http://www.quitoambiente.gob.ec/ambiente/images/Secretaria_Ambiente/red_monitoreo/informacion/norma_ecuato_calidad.pdf

[68] L. Fernández-Jambrina, «Curvas de Bézier», Madrid, España, 2014.

[69] L. Fernández-Jambrina, «Curvas spline», Madrid, España, 2010. [En línea]. Disponible en: https://dcain.etsin.upm.es/~leonardo/pres4.pdf

[70] P. Bruce y A. Bruce, *Practical Statistics for Data Scientists*, 1.ª ed. Sebastopol, CA, USA: O'Reilly, 2017. [En línea]. Disponible en: https://math2510.coltongrainger.com/books/2017-bruce-and-bruce-pratical-statistics-for-data-scientists.pdf

[71] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, y M. Schmid, «A review of spline function procedures in R», *BMC Med Res Methodol*, vol. 19, n.º 1, p. 46, dic. 2019, doi: 10.1186/s12874-019-0666-3.

[72] M. Kaňka, «Segmented Regression Based on B-Splines with Solved Examples», p. 20, 2015.

[73] C. Toquica y W. Pineda, «Aproximación Bayesiana de un modelo semiparamétrico», pp. 1-31, 2017.

Anexos

Anexo A. Datos de la simulación de impacto entre un vehículo y un autobús.

Velocidad (m/s)	Fuerza (N)	FDS	Tiempo de impacto (s)	Deformación (mm)
45	11935.1635	1.17	0	0
46	16709.2288	1.115	0.025001	0.38638
47	23392.9204	1.11	0.025	0.36645
48	32750.0885	1.105	0.050001	0.75283
49	45850.1239	1.10	0.065001	1.033
50	64190.1735	1.00	0.098983	1.1539
51	93075.7516	1.004	0.1246949	1.48718
52	139613.627	1.01	0.1039835	1.75143
53	202439.76	1.015	0.1486784	2.01568
54	303659.64	1.024	0.1933733	2.27993
55	440306.477	1.034	0.2380682	2.54418
56	660459.716	1.043	0.2246949	2.80844
57	957666.588	1.059	0.2227631	3.07269
58	1436499.88	1.075	0.247458	3.33694
59	2082924.83	1.089	0.2321529	3.60119
60	3124387.24	1.1979	0.2968478	3.86544
61	4530361.5	1.31769	0.2646949	4.1297
62	6795542.26	1.44946	0.3215427	4.39395
63	9853536.27	1.5944	0.3462376	4.6582
64	14780304.4	1.75385	0.3509325	4.92245
65	21431441.4	1.92923	0.3956274	5.1867
66	32147162.1	2.12215	0.4246949	5.45096
67	46613385	2.33437	0.4203223	5.71521
68	69920077.5	2.56781	0.4450172	5.97946
69	101384112	2.82459	0.4697121	6.24371
70	152076169	3.10704	0.454407	6.50796
71	220510445	3.4177	0.3915427	6.7722
72	330765667	3.7595	0.4362376	7.0365
73	479610217	4.1355	0.3846949	7.3007
74	719415325	4.549	0.4203223	7.565
75	1043152221	5.0039	0.4750172	7.8292

Velocidad (m/s)	Fuerza (N)	FDS	Tiempo de impacto (s)	Deformación (mm)
76	1564728332	5.5043	0.4697121	8.0935
77	2268856081	6.0547	0.494407	8.3577
78	3403284122	6.6602	0.4646949	8.622
79	4934761977	7.3262	0.5191019	8.8862
80	7402142966	8.0589	0.5437968	9.1505
81	1.0733E+10	8.8647	0.5384917	9.4147
82	1.61E+10	9.7512	0.5731866	9.679
83	2.3345E+10	10.7263	0.5309325	9.9432
84	3.5017E+10	11.7990	0.4956274	10.2075
85	5.0774E+10	12.9789	0.5446949	10.4717
86	7.6161E+10	14.2768	0.5191019	10.736
87	1.10E+11	15.7044	0.5437968	11.0002
88	1.66E+11	17.2749	0.5684917	11.2645
89	2.40E+11	19.0024	0.5431866	11.5288
90	3.60E+11	20.9026	0.5746949	11.793
91	5.22E+11	22.9929	0.6178815	12.0573
92	7.84E+11	25.2922	0.6425764	12.3215
93	1.14E+12	27.8214	0.6272713	12.5858
94	1.70E+12	30.6035	0.6719662	12.85
95	2.47E+12	33.6639	0.5903223	13.1143

Elaborado por: Toalombo, B. (2021).

Anexo B. Datos promedio de Temperatura, humedad relativa y presión atmosférica según la hora del día en la estación San Antonio, Quito.

n	Hora (HH:MM)	Temperatura (ºC)	Humedad relativa (%)	Presión atmosférica (mbar)
1	0:00	13.17	89.31	765.14
2	1:00	12.95	89.12	764.72
3	2:00	12.72	88.96	764.26
4	3:00	12.49	88.85	763.94
5	4:00	12.27	88.89	763.85
6	5:00	12.09	88.80	763.99
7	6:00	12.06	88.15	764.32
8	7:00	13.03	83.43	764.74
9	8:00	15.24	73.85	765.11
10	9:00	17.40	64.37	765.22
11	10:00	19.04	57.07	765.01
12	11:00	20.25	52.58	764.52
13	12:00	20.83	51.69	763.88
14	13:00	20.74	53.74	763.21
15	14:00	20.18	57.63	762.59
16	15:00	19.38	61.71	762.14
17	16:00	18.38	66.28	761.99
18	17:00	17.17	71.65	762.19
19	18:00	15.91	77.53	762.7
20	19:00	14.87	83.10	763.4
21	20:00	14.17	87.15	764.13
22	21:00	13.71	89.71	764.78
23	22:00	13.51	90.05	765.2
24	23:00	13.36	89.68	765.33

Fuente: Secretaría del Ambiente del Distrito Metropolitano Quito (2020). http://www.quitoambiente.gob.ec/index.php/descarga-datos-historicos

Elaborado por: Toalombo, B. (2021).

Anexo C. Codificación de los modelos de regresión polinomial en R.

 # MODELO DE REGRESIÓN PARAMÉTRICO POLINOMIAL.

 # Importación del set de datos:

```
dataset <- read.csv('C:/Users/Usuario/Documents/BYRON/Impacto.csv',
                    header = TRUE, sep = ";")
dataset <- read.csv('C:/Users/Usuario/Documents/BYRON/Climatologicos.csv',
                    header = TRUE, sep = ";")
```

 # Carga de las librerías:

```
library(nortest) # Carga de la librería 'nortest' para aplicar la prueba de normalidad de Kolmogorov-Smirnov corregida por Lilliefors.
library(car) # Carga de librería para las pruebas de independencias de las variables y de homocedasticidad.
library(MASS) # Carga de la librería para la transformación de Box-Cox.
library(caTools) # Carga de la librería 'caTools'.
library(rio) # Carga de la librería 'rio'.
library(openxlsx) # Librería para la función export (para exportar los datos a Excel).
library(ggplot2) # Librería para graficar.
```

 # Pre-inspección de los posibles modelos de regresión que se podrían establecer entre todas las variables de la base de datos:

```
library(GGally) # Librería para activar la función 'ggpairs'.
ggpairs(dataset) # Gráficos de dispersión de puntos y coeficientes de correlación entre todas las variables de las bases de datos.
```

 # Preparación de los datos de las variables a ser utilizadas para los modelos:

```
dataset <- dataset[1:2] # Depuración de los datos a ser utilizados.
```

 # Renombre de las variables:

```
names(dataset) <- c("x", "y")
```

 # División del set de datos en datos de entrenamiento y datos de prueba:

```
# Fijación de la semilla:
set.seed(123)
# Proporción de la división del set de datos:
split = sample.split(dataset$y, SplitRatio = 0.8)
# Datos de entrenamiento:
```

```
training_set = subset(dataset, split == TRUE)
# Datos de prueba:
test_set = subset(dataset, split == FALSE)
# Escala de características:
training_set = scale(training_set)
 test_set = scale(test_set)
# Verificación de las dimensiones de los sets de datos de entrenamiento y de prueba:
dim(training_set)
dim(test_set)
```

 # Ajuste mediante modelo de regresión polinomial usando el comando "poly":

```
lin_reg <- lm(dataset$y ~ poly (dataset$x, degree = 1, raw = T))
poly_reg2 <- lm(dataset$y ~ poly(dataset$x, degree = 2, raw =T))
poly_reg3 <- lm(dataset$y ~ poly(dataset$x, degree = 3, raw =T))
poly_reg4 <- lm(dataset$y ~ poly(dataset$x, degree = 4, raw =T))
poly_reg5 <- lm(dataset$y ~ poly(dataset$x, degree = 5, raw =T))
poly_reg6 <- lm(dataset$y ~ poly(dataset$x, degree = 6, raw =T))
poly_reg7 <- lm(dataset$y ~ poly(dataset$x, degree = 7, raw =T))
poly_reg8 <- lm(dataset$y ~ poly(dataset$x, degree = 8, raw =T))
```

 # Prueba de diferencias significativas ANOVA para obtener el grado de la regresión polinomial:

```
anova(lin_reg,poly_reg2)
anova(poly_reg2,poly_reg3)
anova(poly_reg3,poly_reg4)
anova(poly_reg4,poly_reg5)
anova(poly_reg5,poly_reg6)
anova(poly_reg6,poly_reg7)
anova(poly_reg7,poly_reg8)
```

```
if(anova(lin_reg, poly_reg2)[2,6]>0.05){"El modelo más idóneo es lineal"}else{
if(anova(poly_reg2, poly_reg3)[2,6]>0.05){"El modelo más idóneo es polinómico cuadrático"}else{
if(anova(poly_reg3, poly_reg4)[2,6]>0.05){"El modelo más idóneo es polinómico cúbico"}else{
if(anova(poly_reg4, poly_reg5)[2,6]>0.05){"El modelo más idóneo es polinómico de grado 4"}else{
```

```
if(anova(poly_reg5, poly_reg6)[2,6]>0.05){"El modelo más idóneo es polinómico de grado 5"}else{
if(anova(poly_reg6, poly_reg7)[2,6]>0.05){"El modelo más idóneo es polinómico de grado 6"}else{
if(anova(poly_reg7, poly_reg8)[2,6]>0.05){"El modelo más idóneo es polinómico de grado 7"}else{"El modelo más idóneo es polinómico de grado 8"}}}}}}
```

 # Modelo polinómico:

```
modelo <- poly_reg8 #Colocar el número del grado más idóneo del polinomio:
modelo
summary(modelo)
anova(modelo)
if(anova(modelo)[1,5]<0.05){"Modelo válido"}else{"Modelo inválido"}
#Residuos del modelo:
residuals(modelo)
residuos <- as.data.frame(summary(modelo)[3])
residuos
# Valores predichos, intervalo de confianza y error estándar:
y_pred <- predict(modelo, dataset, se.fit = TRUE, interval="confidence", level = 0.95)
```

 # Métricas de medición del error del modelo:

```
# Coeficiente de determinación R^2.
 R2 <- summary(modelo)[8]
 R2 <- as.numeric(R2)
 # Coeficiente de determinación R^2 ajustado.
R2aj <- summary(modelo)[9]
R2aj <- as.numeric(R2aj)
# Media cuadrática del error MSE.
# MSE <- anova(poly_reg)[5,3]
MSE <- anova(modelo)[2,3]
# Suma de cuadrados del error SSE.
SSE <- anova(modelo)[2,2]
# Raíz de la media cuadrática del error RSME:
```

```
RSME <- sqrt(MSE)
# Suma de cuadrados totales SST:
SST <- SSE/(1-R2)
# Error porcentual absoluto medio MAPE.
MAPE <- (100/nrow(dataset))*sum(abs(residuos/dataset$y))
cat("MAPE:",round(MAPE,2),"%")
```

 # Intervalo de confianza:

```
intervalo_conf <- data.frame(inferior = y_pred$fit[,2],superior = y_pred$fit[,3])
```

 # Comprobación de los supuestos del modelo.

```
# Prueba de normalidad de los residuos del modelo:
# Shapiro-Wilk:
shapiro.test(dataset_modelo$residuals)
# Kolmogorov-Smirnov corregida por Lilliefors:
lillie.test(dataset_modelo$residuals)

# Prueba de no correlación de los residuos Durbin-Watson:
durbinWatsonTest(modelo)

# Prueba Breusch-Pagan estudiantilizada para la homocedasticidad:
bptest(modelo, studentize = TRUE)

# Gráficos para comprobar los supuestos del modelo:
par(mar=c(6,6,1.5,2), mfrow=c(2,2))
plot(modelo, lwd=2.5, col='blue', cex.axis=1.2, cex.lab=1.5, las=1,
     mgp=c(4,1,0), mar=c(3.1,4.1,4.1,2.1))
```

 # Transformación de la variable a predecir mediante una función de Box-Cox (solamente en caso de que no se cumplan los supuestos):

```
bc <- boxcox(modelo, lambda =seq(-3,3,0.01))
lambda <- bc$x[which.max(bc$y)]
dataset$y_bc <- dataset$y^lambda
dataset$y_bc <- ((dataset$y^lambda)-1)/lambda
dataset$y_ln <- log(dataset$y)
dataset$y_ex <- (exp(lambda*dataset$y)-1)/lambda
```

 # Creación de un nuevo dataset con los datos x, y, y predichos y residuos:

```
dataset_modelo<- data.frame(x=dataset$x,y=dataset$y,
                            y_pred_inf=intervalo_conf$inferior,
                            y_pred=y_pred$fit[,1], y_pred_sup=intervalo_conf$superior,
                            y_pred_dist=intervalo_conf$superior-intervalo_conf$inferior,
                            y_residuos=residuos)
```

 # Exportación de los datos del nuevo dataset:

```
# Librería para la función export (para exportar los datos a Excel):

library(openxlsx)

export(dataset_modelo,"C:/Users/Usuario/Documents/BYRON/Modelo_regresion
_polinomica.xlsx",
header_Style=createStyle(halign="center",textDecoration='Bold',fontSize=12))
```

 # Gráfico del modelo de regresión polinomial y del set de datos:

```
ggplot(dataset, aes(x, y) )+geom_point(colour = "red",size=2.5) +
   stat_smooth(method = lm, formula = y ~ poly(x, degree = 8, raw=TRUE))+
     geom_line(aes(x = dataset_modelo$x, y = dataset_modelo$y_pred_inf),
               colour = 'blue3', size=0.7, linetype = "longdash") +
     geom_line(aes(x = dataset_modelo$x, y = dataset_modelo$y_pred_sup),
               colour = 'blue3', size=0.7, linetype = "longdash") +
   labs(title='Regresión Polinomial') +
   xlab('Variable regresora x') + ylab('Variable respuesta y')
```

Anexo D. Codificación de los modelos de regresión B-spline en R.

 # MODELO DE REGRESIÓN NO PARAMÉTRICO B-SPLINE.

 # Importación del set de datos:

```
dataset <- read.csv('C:/Users/Usuario/Documents/BYRON/Impacto.csv',
                    header = TRUE, sep = ";")
dataset <- read.csv('C:/Users/Usuario/Documents/BYRON/Climatologicos.csv',
                    header = TRUE, sep = ";")
```

 # Carga de librerías:

```
# Carga de las librerías 'lattice', 'ggfortify ', 'caret', 'splines', 'splines2'
library(lattice)
library(ggfortify) # Para la función autoplot.
library(caret)
library(splines) # Para el uso de la función bs.
library(splines2) # Para el uso de la función bSpline.
```

 # Renombre de las variables:

```
names(dataset) <- c("x","y")
```

 # Ajuste del modelo B-spline

```
#Grado del polinomio de la curva B-spline:
grado <- 3
#Número de puntos de control o vértices del polígono de control n:
gl <- 6
#Número de nudos internos:
nudos <- attr(bs(dataset$x, df = gl, degree = grado), "knots")
modelo_splines <- lm(y ~ bs(x=dataset$x, df = gl, degree = grado, knots = nudos,
Boundary.knots=range(dataset$x),intercept=TRUE),data=dataset)
```

 # B-spline base del modelo:

```
bspline_basis <- bSpline(dataset$x, df=gl,knots=nudos,degree = grado,
                         Boundary.knots = range(dataset$x), intercept = TRUE)
# Graficación del modelo:
autoplot(bspline_basis) + theme_classic()+geom_line(size=1.1) +
  labs(title="B-spline base")+theme(plot.title = element_text(hjust = 0.3))
```

 # Modelo B-spline y resumen:

```
modelo_splines
summary(modelo_splines)
```

#Prueba ANOVA de validez del modelo:

```
anova(modelo_splines)
if(anova(modelo_splines)[1,5]<0.05){"Modelo válido"}else{"Modelo inválido"}

# Residuos del modelo:
residuos2 <- data.frame(residuals(modelo_splines))

# Valores predichos, intervalo de confianza y error estándar:
y_pred2 <- predict(modelo_splines, newdata = dataset,
                   interval="confidence",se.fit = TRUE, level = 0.95)
```

 # Métricas de medición del error del modelo:

Coeficiente de determinación R^2:

```
 R2 <- summary(modelo_splines)[8]
 R2 <- as.numeric(R2)
```

Coeficiente de determinación R^2 ajustado:

```
 R2aj <- summary(modelo_splines)[9]
 R2aj <- as.numeric(R2aj)
```

Media cuadrática del error MSE:

```
 MSE <- anova(modelo_splines)[2,3]
```

Suma de cuadrados del error SSE:

```
 SSE <- anova(modelo_splines)[2,2]
```

Raíz de la media cuadrática del error RSME:

```
 RSME <- sqrt(MSE)
```

Suma de cuadrados totales SST:

```
 SST <- SSE/(1-R2)
```

 # Intervalo de confianza:

```
 intervalo_conf2 <- data.frame(inferior=y_pred2$fit[,2],superior=y_pred2$fit[,3])
```

 # Comprobación de los supuestos del modelo.

Prueba de normalidad de los residuos del modelo:

```
 # Shapiro-Wilk:
 shapiro.test(modelo_splines$residuals)
 # Kolmogorov-Smirnov corregida por Lilliefors:
 lillie.test(modelo_splines$residuals)

# Prueba de no correlación de los residuos Durbin-Watson:
 durbinWatsonTest(modelo_splines)
 # Prueba Breusch-Pagan estudiantilizado para la homocedasticidad:
 bptest(modelo_splines, studentize = TRUE)

 # Gráficos para comprobar los supuestos del modelo:
par(mar=c(6,6,1.5,2), mfrow=c(2,2))
plot(modelo_splines, lwd=2.5, col='blue', cex.axis=1.2,cex.lab=1.5,las=1,
     mgp=c(4,1,0), mar=c(3.1,4.1,4.1,2.1))
```

 # Creación de un nuevo dataset con los datos x, y, y predichos y residuos:

```
 dataset_modelo2<- data.frame(x=dataset$x,y=dataset$y,
                      y_pred2_inf=intervalo_conf2$inferior,
                      y_pred2=y_pred2$fit[,1],
                      y_pred2_sup=intervalo_conf2$superior,
                      y_pred2_dist=intervalo_conf2$superior-intervalo_conf2$inferior,
                       y_residuos=residuos2)
```

 # Exportación de los datos del nuevo dataset:

```
export(dataset_modelo,"C:/Users/Usuario/Documents/BYRON/Modelo_regresion
_bspline.xlsx", header_Style = createStyle(halign="center",textDecoration='Bold',
                           fontSize=12))
```

 # Gráfico del modelo de regresión B-spline y del set de datos:

```
ggplot() +
   geom_point(aes(x = dataset$x, y = dataset$y),
              colour = "red",size=2.5) +
   geom_line(aes(x = dataset$x, y = predict(modelo_splines, newdata = dataset)),
             colour = 'green3', size=1.) +
   geom_line(aes(x = dataset_modelo2$x, y = dataset_modelo2$y_pred2_inf),
             colour = 'green3', size=0.7, linetype = "longdash") +
   geom_line(aes(x = dataset_modelo2$x, y = dataset_modelo2$y_pred2_sup),
             colour = 'green3', size=0.7, linetype = "longdash") +
   labs(title='Regresión B-spline') +
   xlab('Variable regresora x') + ylab('Variable respuesta y')+
  theme(plot.title = element_text(hjust = 0.3))
```

Anexo E. Codificación para la comparación de los modelos de regresión en R.

 # COMPARACIÓN DE LOS 2 MODELOS Y GRÁFICOS.

 # Test de Wilcoxon para identificar diferencias entre los residuos de ambos grupos:

```
wilcox.test(dataset_modelo$y_pred_dist,dataset_modelo2$y_pred2_dist,
alternative = "two.sided", paired = FALSE,conf.level = 0.95)
```

 # Diagrama de cajas de las variabilidades de los 2 modelos:

```
boxpl<- data.frame(u=c(dataset_modelo$y_pred_dist,
dataset_modelo2$y_pred2_dist),v=c(rep("Polinómico",
NROW(dataset_modelo$y_pred_dist)),rep("B-
spline",NROW(dataset_modelo2$y_pred2_dist))))

ggplot(boxpl,aes(x=v,y=u))+theme_bw()+
  geom_boxplot(fill=c('green','orange'), color=c('darkblue','black'), size=0.75) +
  labs(title = "Variabilidad de los modelos de regresión",
       x="Tipo de modelo de regresión", y='Longitud IC') +
  theme(title = element_text(size=16, colour = 'darkblue'),
        axis.title.x = element_text(size=18,face = 'bold',vjust = 0.5, colour='black'),
        axis.title.y = element_text(size=16,face = 'bold',vjust=1, colour='black'),
        axis.text.x = element_text(size=18),
        axis.text.y=element_text(size=14))+
 theme(plot.title=element_text(hjust=0.3))
```

 # Gráfica Q-Q Plot:

```
 ggplot(boxpl) + geom_qq(aes(sample = u),size=3, col="blue")+
 geom_qq_line(aes(sample = u), size=1.0, col="darkgreen")+
 theme_bw()+labs(title = "Normal Q-Q Plot",
 x="Diferencias entre la longitud del Intervalo de Confianza Modelos de regresión
polinomial y B-spline", y='Sample Quantiles')+
 theme(title = element_text(size=16, colour = 'darkblue'),
  axis.title.x = element_text(size=14,face = 'bold',vjust = 0.5, colour='black'),
  axis.title.y = element_text(size=16,face = 'bold',vjust=1, colour='black'),
  axis.text.x = element_text(size=18),
  axis.text.y = element_text(size=14))+theme(plot.title = element_text(hjust = 0.3))
```

 # Gráfico de los dos modelos:

```
ggplot(dataset, aes(x, y)) +
   labs(title='Modelos de regresión del x vs y', x = 'x ', y = 'y ') +
   geom_point(colour = "blue",size=1.5) +
   geom_line(aes(x = x, y = predict(modelo, newdata = dataset)),
             colour = 'red',size=1.2) +
   geom_line(aes(x = x, y = dataset_modelo$y_pred_inf),
             colour = 'red',size=0.8, linetype = "longdash") +
   geom_line(aes(x = x, y = dataset_modelo$y_pred_sup),
             colour = 'red',size=0.8, linetype = "longdash") +
   geom_line(aes(x = x, y = predict(modelo_splines, newdata = dataset)),
             colour = 'green2', size=1.2) +
   geom_line(aes(x = dataset_modelo2$x, y = dataset_modelo2$y_pred2_inf),
             colour = 'green2', size=0.8, linetype = "longdash") +
   geom_line(aes(x = dataset_modelo2$x, y = dataset_modelo2$y_pred2_sup),
             colour = 'green2', size=0.8, linetype = "longdash") + theme_bw() +
  theme(title = element_text(size=16, colour = 'darkblue'),
        axis.title.x = element_text(size=15,face = 'bold',vjust = 0.5, colour='black'),
        axis.title.y = element_text(size=15,face = 'bold',vjust=1, colour='black'),
        axis.text = element_text(size=15)) +
  annotate(geom = "text", x = 7.8, y = 40, label = "Reg.Polinomial",
           colour='red', size=6) +
  annotate(geom = "text", x = 7.7, y =30, label = "Reg.B-spline",
           colour='green3', size=6) +
  theme(plot.title=element_text(hjust=0.4))+
scale_x_continuous(breaks=seq(0,10,1))+
scale_y_continuous(breaks=seq(0,200,10))
```