ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

MODELOS LINEALES GENERALIZADOS EN EL ANÁLISIS DE DERIVADOS IMPORTADOS Y REFINADOS EN LA EMPRESA PÚBLICA PETROECUADOR EN EL PERIODO 2017-2020

Trabajo de Titulación

Tipo: Proyecto de Investigación

Presentado para optar al grado académico de:

INGENIERA EN ESTADÍSTICA

AUTORA: JOHANNA TANIA BASTIDAS CAIBE

DIRECTORA: Ing. JOHANNA ENITH AGUILAR REYES Mgtr.

Riobamba – Ecuador

2022

Yo, Johanna Tania Bastidas Caibe, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 14 de enero de 2022

Johanna Tania Bastidas Caibe
1724248883

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

El Tribunal del Trabajo de Titulación, certifica que: El trabajo de titulación; tipo proyecto de investigación: "**MODELOS LINEALES GENERALIZADOS EN EL ANÁLISIS DE DERIVADOS IMPORTADOS Y REFINADOS EN LA EMPRESA PÚBLICA PETROECUADOR EN EL PERIODO 2017-2020"**, realizado por la señorita: **JOHANNA TANIA BASTIDAS CAIBE**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación. El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	Firma	**Fecha**
Ing. Natalia Alexandra Perez Londo Ms.C. **PRESIDENTE DEL TRIBUNAL**	Firmado electrónicamente por: NATALIA ALEXANDRA PEREZ LONDO	2022-01-14
Ing. Johanna Enith Aguilar Reyes Mgtr. **DIRECTOR DEL TRABAJO DE TITULACION**	Firmado electrónicamente por: JOHANNA ENITH AGUILAR REYES	2022-01-14
Dra. Jaqueline Elizabeth Balseca Castro Mgs. **MIEMBRO DE TRIBUNAL**	JAQUELINE ELIZABETH BALSECA CASTRO Firmado digitalmente por JAQUELINE ELIZABETH BALSECA CASTRO Fecha: 2022.01.21 10:14:03 -05'00'	2022-01-14

DEDICATORIA

Este trabajo es dedicado a mi familia, por el apoyo moral y económico que me han dado cada uno de los integrantes, por cada consejo y la confianza que me brindaron, pero especialmente a mis padres Juan Bastidas y María Caibe, quienes han inculcado en mi valores y principios, la mentalidad de siempre luchar por mis sueños y de nunca rendirme. A mi hermano Marco Bastidas y hermana Isabel Bastidas, que con su ejemplo y experiencia han sabido guiarme por el buen camino.

Johanna Bastidas

AGRADECIMIENTO

Agradecer a toda la planta docente de la Escuela que semestre a semestre fueron impartiendo conocimientos que me han servido para crecer como persona y profesional. Además, agradecer a la Ing. Johanna Aguilar tutora de mi tesis y Docente de varias materias a lo largo de mi trayectoria en la institución, y a la Dra. Jaqueline Balseca miembro del trabajo de titulación que ha guiado de la mejor manera por varios años mi camino como estudiante y en este trabajo.

Johanna Bastidas

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICAS

ÍNDICE DE ECUACIONES

ÍNDICE DE ANEXOS

RESUMEN

El análisis de los derivados refinados e importados mediante poliductos en la Empresa Pública Petroecuador, tiene como objetivo pronosticar el número de barriles de derivados que se distribuyen a nivel nacional y el poliducto adecuado, considerando cambios en los factores de los que dependen, mediante modelos lineales generalizados (GLM). Para el estudio se consideraron, en principio 224 observaciones registradas en el periodo 2017-2020, con información del volumen transportado en barriles de cada año, en cada poliducto, mediante los diferentes tramos, así como también el producto que se distribuyeron, además, se sabe el caudal, capacidad y el volumen despachado, con esta información se planteó la variable respuesta cuantitativa de interés, que es el volumen transportado y se ajustó un modelo lineal múltiple; se evaluó el rendimiento del modelo; se detectaron datos atípicos multivariados; se mejoró el ajuste mediante selección de variables predictoras significativas; se aplicó un modelo lineal generalizado que es la regresión de Poisson; se evaluó el modelo; y, para seleccionar el mejor modelo de los propuestos, se consideraron dos criterios: la comparación de modelos basada en la suma de cuadrados y el criterio de información de Akaike. Para la variable respuesta cualitativa poliducto, se ajustó un modelo de regresión logística. Tras detectar 87 datos atípicos, con la regresión múltiple, se obtuvo como resultado un coeficiente de determinación del 92%, cumple dos de los supuestos y falla en independencia; mientras que la regresión de Poisson presentó sobredispersión. En tanto que, para la regresión logística dependerá del volumen transportado y el producto. Se concluye que la regresión múltiple es la más adecuada para identificar el volumen transportado, y la regresión logística nominal para identificar el poliducto adecuado. Finalmente, se recomienda una adecuada toma de decisiones por el personal de la empresa frente a los diferentes escenarios que se puede presentar.

Palabras clave: <DERIVADOS DE PETRÓLEO>, <POLIDUCTOS>, <MODELOS LINEALES>, <MODELOS LINEALES GENERALIZADOS (GLM)>, <REGRESIÓN LOGIT>.

ABSTRACT

The analysis of refined and imported derivatives through pipelines in the Petroecuador Public Company, aims to forecast the number of barrels of derivatives that are distributed at the national level and the appropriate pipeline, considering changes in the factors on which they depend, using generalized linear models (GLM). For the study, 224 observations recorded in the 2017-2020 period were considered, with information on the volume transported in barrels of each year, in each pipeline, through the different sections, as well as the product that was distributed, in addition, the flow, capacity and volume dispatched. With this information, the quantitative response variable of interest was proposed, which is the transported volume, and a multiple linear model was adjusted. The performance of the model was evaluated; multivariate outliers were detected; fit was improved by selection of significant predictor variables. A generalized linear model was applied, which is the Poisson regression; the model was evaluated; and, to select the best model from those proposed, two criteria were considered: the comparison of models based on the sum of squares and the Akaike information criterion. For the qualitative multi-pipeline response variable, a logistic regression model was adjusted. After detecting 87 atypical data, with multiple regression, a coefficient of determination of 92% was obtained, it meets two of the assumptions and fails in independence; while the Poisson regression presented over-dispersion. Whereas, for the logistic regression it will depend on the transported volume and the product. It is concluded that multiple egress is the most appropriate to identify the transported volume, and nominal logistic regression to identify the appropriate pipeline. Finally, an adequate decision making by the company's staff is recommended in the face of the different scenarios that may arise.

Keywords: <PETROLEUM DERIVATIVES>, <POLIDUCTS>, <LINEAR MODELS>, <GENERALIZED LINEAR MODELS (GLM)>, <LOGIT REGRESSION>

Firmado electrónicamente por:
NANCI MARGARITA INCA CHUNATA

INTRODUCCIÓN

El mundo desde hace varias décadas, depende de grandes fuentes de energía como el petróleo y sus derivados, pues la necesidad de estos recursos se ha incrementado por el crecimiento poblacional, e incluso a pesar de que se han implementado nuevas energías como la nuclear, hidroeléctrica y otras energías renovables; siguen siendo el petróleo, el gas natural y el carbón, agentes principales para el consumo energético de la población mundial (Cubillos & Estenssoro, 2011, p.103). En América y a nivel mundial, Estados Unidos es uno de los mayores consumidores de energía, especialmente para cubrir la demanda del sector transporte. En América Latina, muchos países son principales exportadores de petróleo del mundo, ya que de acuerdo a su situación geográfica se puede explotar este recurso, así como Ecuador es el país más pequeño que tiene petróleo en América Latina.

En Ecuador la explotación del petróleo ha ido disminuyendo, y esto se debe en parte a que los campos petroleros se encuentran en declinación, pero también depende de la disminución de explotación de crudo que se está realizando, debido a la falta de pago del gobierno a petroleras privadas (Supurrier, 1996, p.155). En EP Petroecuador, la distribución de derivados mediante poliductos a centros de distribución para su expendio, no abastece la demanda interna del país, por lo que se opta por importar derivados, pues la cantidad de barriles que se distribuyen dependen de varios factores como la demanda por sector, producto, tramo, viscosidad, entre otros. Sin embargo, ¿Todos los poliductos trabajan por igual?, es una de las preguntas que se harían sin tomar en cuenta de los factores que depende. Además, si la situación sigue así ¿Cuántos barriles de derivados se distribuirán si varían estos factores?, pues la Estadística es una ciencia que ayuda a responder este tipo de preguntas.

Gracias a un análisis estadístico se pueden identificar los agentes de los que depende el número de barriles de derivados e incluso la intensidad con la que lo hacen, Así como también el poliducto más adecuado para transportar un producto. Sin embargo, estos factores, no quedan solamente descritos, pues son la base para poder pronosticar mediante modelo lineal generalizado, la cantidad de barriles de derivados que se deben distribuir en periodos futuros en circunstancias diferentes, lo cual sería conveniente para poder visualizar una aproximación real de los posibles panoramas, ya que varios entendidos en el tema petrolero, coinciden en que ciertas circunstancias afectaría a la economía, no sólo de la empresa sino del país, pero muchas de las veces al ser subjetivo puede alejarse de la realidad, y al tener datos de un aproximación a la vida real, sería un aporte para la toma de decisiones por parte del personal encargado para evitar algunas consecuencias o a su vez mejorar el rendimiento de la empresa.

Los modelos lineales generalizados son procesos estadísticos, en donde intervienen variables numéricas, categóricas o la combinación de ellas, de tal forma que existen modelos que trabajan dependiendo el tipo de variables, entre estos se puede mencionar los métodos de regresión simple,

múltiple, logit, modelos lineales generalizados, entre otros. Todos son de suma utilidad, para obtener buenas predicciones, pues se basa en la información obtenida y las variables predictoras que en este intervienen, tomando en cuenta la significancia de cada variable predictora en el modelo a ajustar. Las proyecciones que dan como resultado, son escenarios de los posibles valores que toman las variables independientes, lo cual ayudaría a tomar acciones en cada una de las situaciones.

Antecedentes

Las refinerías en la cuenca Atlántica han tenido un crecimiento significativo especialmente en Estados Unidos ya que tuvo más posibilidades de exportar este recurso incluso evadiendo las adversidades de la época. También, las exportaciones de Diésel se daban mejor en el continente europeo (Tapia, 2017, p.2).

El petróleo se ha convertido en la principal fuente de energía, pues es base de economía de cualquier país, a pesar de ser considerado como una fuente de contaminación. En los últimos tiempos, ha decrecido en proporción por la aparición de nuevos sustitutos, este recurso es destinado como combustibles de motores, de transporte y la obtención de productos químicos, medicamentos, entre otros productos de uso diario, representando así un 80% y el porcentaje restante proviene de energías renovables (EP Petroecuador, 2013, p.27).

En la década de los 70, Ecuador tuvo un cambio importante en su economía, refiriéndose al auge de exportaciones de este recurso el petróleo, pues se incrementaron notablemente su precio a nivel internacional, lo cual beneficio a la economía del Ecuador pues en ese tiempo se manejaba el modelo agroexportador (EP Petroecuador, 2013, p.8).

La producción nacional de derivados en 1971 se incorpora de manera definitiva en el país, pues en este año se evidencia por primera vez, exportaciones que bordeaban los ciento sesenta y dos millones de dólares, siendo este monto significativo a periodos anteriores pues hasta 1981 las exportaciones crecieron hasta dos mil quinientos millones de dólares, volviendo al país apto para inversiones a nivel internacional (Acosta, 2006, p.121).

En el trabajo realizado por Juan Pablo Murillo Fajardo de "MODELACIÓN PARA LA PROGRAMACIÓN DEL TRANSPORTE DE PRODUCTOS REFINADOS EN LA RED NACIONAL DE POLIDUCTOS DE ECOPETROL S.A", en donde menciona el funcionamiento de la red de transportes de derivados mediante poliductos a nivel nacional, de tal forma que se puede apreciar los factores de los que depende, ya que influye aspectos económicos, como físicos y geográficos (Macías y Martínez, 2012, p.3).

En el trabajo "ANÁLISIS DEL IMPACTO ECONÓMICO QUE TENDRÁ LA REFINERÍA DEL PACÍFICO "ELOY ALFARO" EN LA ECONOMÍA ECUATORIANA" realizado por Guillín Medina Carlos César, hace referencia a escenarios futuros en la economía del país,

refiriéndose a la producción de las refinerías y los factores que influyen en la economía del país. Conocer la importancia del sector petrolero en la economía del Ecuador, obtener información de producción y tecnología de las actuales plantas refinadoras del país e interpretar la incidencia en la Balanza Comercial (Muirragui y Guillín, 2013).

En el "Análisis de las Exportaciones e Importaciones de Derivados de Petróleo y su Incidencia en la Balanza Comercial del Ecuador." Realizado por Srta. Vanessa Piedad Mosquera Medina Sr. Jonathan Marcelo Simbaña Cajo, en donde menciona las importaciones de derivados que se realizan en el país los agentes principales que en ellos intervienen, y dando un panorama general en la situación del país, en donde se propone un modelo econométrico para realizar el análisis de la incidencia de las variables y como podrían afectar a la balanza comercial del país, para una toma adecuada de decisiones (Mosquera y Simbaña, 2019).

Planteamiento del Problema

Enunciado del problema

En EP Petroecuador una de sus funciones es distribuir a nivel nacional derivados mediante poliductos a centros de acopio, derivados que son necesarios e imprescindibles para el consumo diario de la población ecuatoriana, ya sea directamente como la gasolina, o como materia prima para la elaboración de otros productos; estos son distribuidos desde las refinerías de Esmeraldas, La libertad y Shushufindi, y a pesar de que en el Ecuador, el petróleo es uno de los principales productos de exportación, estas no abastecen la demanda interna del país, por lo que un porcentaje de estos es importado. Hay que considerar que la cantidad de barriles de derivados que se distribuyen en cada centro de acopio depende de varios factores, por lo que a la empresa le interesa saber las proyecciones de la cantidad de barriles que se necesitan para los próximos años en diferentes situaciones; así como también el Poliducto adecuado que se requiere para cada producto. Este proyecto ayudará a aproximar estos requerimientos y a su vez con las proyecciones ayudará a la toma de decisiones del personal encargado de esta área en la empresa.

Formulación (Incógnita)

¿Cuántos barriles de derivados refinados e importados se necesitan transportar a cada centro de distribución en el Ecuador o cuál sería el poliducto adecuado, si los factores de los que depende varían?

Objetivo General

Aplicar modelos lineales generalizados en el análisis de derivados importados y refinados en la Empresa Pública Petroecuador en el periodo 2017-2020.

Objetivos Específicos

- Análisis estadístico descriptivo de cada una de las variables en estudio.

- Identificar los factores de los que depende el Volumen de derivados refinados e importados en los modelos lineales.

- Ajustar un modelo lineal generalizado para obtener los pronósticos de la cantidad de derivados refinados e importados y un modelo lineal generalizado para identificar el poliducto adecuado.

- Validar los modelos ajustados para evitar pronósticos erróneos.

- Comparar las proyecciones en los diferentes escenarios que se podrían dar en la distribución de derivados a los centros de acopio.

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL

1.1. Bases conceptuales

1.1.1. Análisis exploratorio de datos

Estadística

La estadística recolecta, organiza, presenta y analiza datos, así como también se obtiene conclusiones y contribuye a la toma de decisiones razonables y válidas (Murray & Larry, 2009, p.1).

Medidas de Tendencia Central

Cuando se dispone de un conjunto de observaciones, es de interés encontrar el valor en torno al cual se agrupan la mayoría de ellos o el centro de estas, las medidas descriptivas que permiten identificar estos valores se denominan medidas de localización o medidas de tendencia central (Pulido y Salazar, 2008, p.360).

Media Aritmética

Es el promedio simple de n datos de un conjunto dado. La media poblacional (μ) y la media muestral con $\bar{x}$. Las fórmulas de cada una de estas son:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad\qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1-1}$$

Mediana

Si hay n datos de un conjunto, para el cálculo de la mediana se debe ordenar de forma creciente, de tal manera que la mediana ($\tilde{x}$), es:

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & si\ n\ es\ impar \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & si\ n\ es\ par \end{cases} \tag{2-1}$$

Medidas de Dispersión

Son parámetros estadísticos que indican como se alejan los datos respecto de la media aritmética. Sirven como indicador de la variabilidad de los datos. Las medidas de dispersión más utilizadas la desviación estándar y la varianza.

Desviación Estándar

La desviación estándar de n datos de un conjunto de observaciones, corresponde a la raíz cuadrada de la varianza (positiva).

$$S = \sqrt{\frac{\sum_{i=1}^{n} d_i}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{3-1}$$

Varianza

La varianza de n datos de un conjunto de observaciones (Congacha, 2016, p.70).

$$S^2 = \frac{\sum_{i=1}^{n} d_i}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{4-1}$$

Coeficiente de Asimetría

El coeficiente de Asimetría o también denominado tercer momento centrado, indica el sesgo de una distribución de probabilidad, este puede ser positivo o negativo o nulo es decir en el caso de que la distribución sea simétrica será igual a 0 (Monroy, 2008, p.70).

$$A = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3} \tag{5-1}$$

Si A>0 asimetría positiva: Para este caso la media está a la derecha de la mediana y la mediana está a la derecha de la moda.

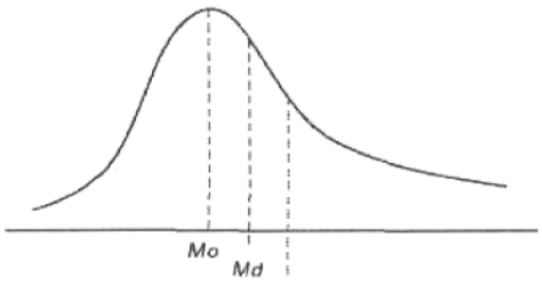


Figura 1-1: Asimetría positiva

Fuente: Monroy, Saldivar, 2008.

Si A=0 la distribución es simétrica: Cuando coinciden la media, la moda y la mediana, como muestra en la imagen.

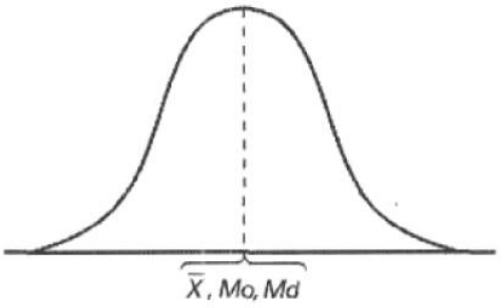


Figura 2-1 :Distribución simétrica

Fuente: Monroy, Saldivar, 2008.

Si A<0 asimetría negativa: Para este caso la moda y la media están a la izquierda de la mediana

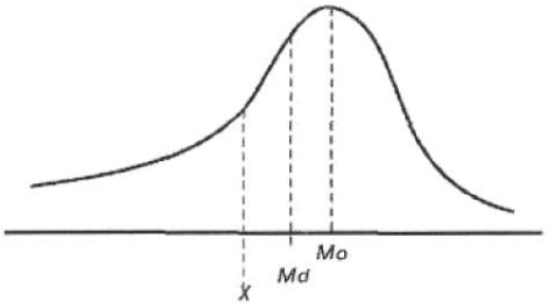


Figura 3-1: Asimetría negativa

Fuente: Monroy, Saldivar, 2008.

Coeficiente de Kurtosis

El coeficiente de Kurtosis indica la cantidad de agrupamiento con respecto a una medida que sea de tendencia central, a dicho coeficiente también se lo denomina el cuarto momento centrado de una distribución de probabilidad (Murray & Larry, 2009, p.126).

El coeficiente de Kurtosis viene dado por:

$$K = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4} \qquad (6\text{-}1)$$

Donde se tiene que:

- Si K>0 la distribución es leptocúrtica: Indica que la curva presenta un apuntamiento positivo, es decir los individuos están más agrupados (Viedma, 2018, p.59).

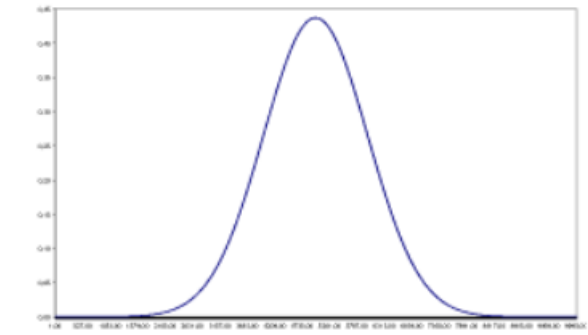


Figura 4-1: Distribución es leptocúrtica

Fuente: Viedma, Carlos, 2018.

- SI K=0 la distribución es mesocúrtica: la curva no presenta apuntamiento.

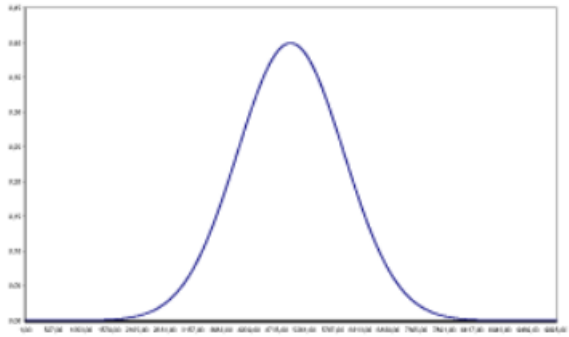


Figura 5-1: Distribución mesocúrtica

Fuente: Viedma, Carlos, 2018.

- Si K<0 la distribución de platicurtica: La curva presenta un apuntamiento negativo, es los individuos en estudio están menos agrupados, más dispersos.

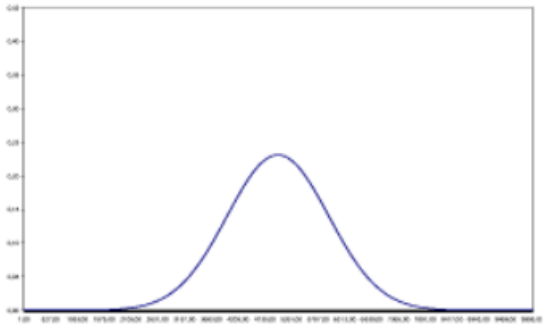


Figura 6-1: Distribución platicúrtica

Fuente: Viedma, Carlos, 2018

1.1.2. Identificación de Grupos atípicos

Sea $\underline{X} = (X_1, X_2, X_3, X_4, .., X_p)^t$ vector estadístico p-variante con componentes cuantitativas.

Se tiene una distribución estadística unitaria p-variante; $x = \{x_i\}_{i=1}^{n}$, con $\boldsymbol{n}$ la numerosidad del colectivo.

Para la **identificación de grupos de atípicos** se tiene:

- Estandarizar los datos de forma multivariante:

Donde la transformación está dada por $\underline{Y} = S_x^{-\frac{1}{2}}(\underline{X} - \bar{x})$, donde $S_x^{-\frac{1}{2}}$ inversa de la matriz de la raíz cuadrada simétrica de S_x

Se tiene $\underline{Y} = \underline{Z}^{(j)}$. Por lo tanto:

Tabla 1-1: Matrices de proyecciones de máxima y mínima kurtosis

Matrices $Z^{(j)}$ de las proyecciones ortogonales de máxima kurtosis		
$\underline{Z}^{(1)}$	$\underline{Z}^{(3)} = A_2\underline{Z}^{(2)}$	$\underline{Z}^{(5)} = A_4\underline{Z}^{(4)}$
$\underline{Z}^{(2)} = A_1\underline{Z}^{(1)}$	$\underline{Z}^{(4)} = A_3\underline{Z}^{(3)}$	…
Matrices $Z^{(j)}$ de las proyecciones ortogonales de mínima kurtosis		
$\underline{Z}^{(6)} = \underline{Z}^{(1)}$ coincide con el inicial	$\underline{Z}^{(8)} = A_7\underline{Z}^{(7)}$	…
$\underline{Z}^{(7)} = A_6\underline{Z}^{(6)}$	$\underline{Z}^{(9)} = A_8\underline{Z}^{(8)}$	$\underline{Z}^{(2p)}$

Elaborado por: Bastidas, Johanna, 2022.

Donde A_j una matriz simétrica

Tabla 2-1: Matriz Aj una matriz simétrica

$A_1 = I - d_1d_1^t$	$A_2 = I - d_2d_2^t$	$A_3 = I - d_3d_3^t$	$A_4 = I - d_4d_4^t$	$A_5 = I - d_5d_5^t \ldots$
$A_6 = I - d_6d_6^t$	$A_7 = I - d_7d_7^t$	$A_8 = I - d_8d_8^t$	$A_9 = I - d_9d_9^t$	$A_{10} = I - d_{10}d_{10}^t \ldots$

Elaborado por: Bastidas, Johanna, 2022

- Sea $\{d_1, d_2, d_3, d_4, d_5\}$ direcciones ortogonales de máxima kurtosis, y $\{d_6, d_7, d_8, d_9, d_{10}\}$ direcciones ortogonales de mínima kurtosis.

- Cálculo de los $y_i^{(j)} = d_j^t z_i^{(j)}$, siendo esta también una transformación lineal se tiene las siguientes relaciones

$$Y^{(j)} = Z^{(j)}d_j \qquad (7\text{-}1)$$

Tabla 3-1: Proyecciones Ortogonales de máxima y mínima Kurtosis

$Y^{(j)}$ **de las proyecciones ortogonales de máxima kurtosis**		
$Y^{(1)} = Z^{(1)}d_1$	$Y^{(3)} = Z^{(3)}d_3$	$Y^{(5)} = Z^{(5)}d_5$
$Y^{(2)} = Z^{(2)}d_2$	$Y^{(4)} = Z^{(4)}d_4$	…
$Y^{(j)}$ **de las proyecciones ortogonales de mínima kurtosis**		
$Y^{(6)} = Z^{(6)}d_6$	$Y^{(8)} = Z^{(8)}d_8$	$Y^{(10)} = Z^{(10)}d_{10}$
$Y^{(7)} = Z^{(7)}d_7$	$Y^{(9)} = Z^{(9)}d_9$	…

Elaborado por: Bastidas, Johanna, 2022

- Considerar como sospechosos aquellos puntos que en alguna de estas 2p direcciones están claramente alejados del resto, es decir, verifican

$$\frac{|y_i^{(i)} - med(y^{(j)})|}{Meda(y^{(j)})} > 5 \qquad (8\text{-}1)$$

- A continuación, se eliminan todos los valores sospechosos detectados y se vuelve al primer paso para analizar los datos restantes. hasta que no se detecten más datos sospechosos o se haya eliminado una proporción de datos prefijada, por ejemplo, un máximo del 40% de los datos.
- Con la matriz de datos libre de valores sospechosos, se calcula $\bar{x}_R$ y S_R y calcular la distancia de mahalanobis con los datos sospechosos:

$$d_R^2(x_i, \bar{x}_R) = (x_i, -\bar{x}_R)S_R^{-1}(x_i, -\bar{x}_R)^t \qquad (9\text{-}1)$$

Y se tiene:

Si $d_R^2(x_i, \bar{x}_R) > p + 3\sqrt{2p}$ se considera atípico

Si $d_R^2(x_i, \bar{x}_R) \leq p + 3\sqrt{2p}$ se considera sospechoso no atípico (Peña, 2002, pp. 122-124).

1.1.3. Análisis de regresión

El análisis de regresión tiene como objetivo modelar en forma matemática el comportamiento de una variable de respuesta en función de una o más variables independientes (factores). Por ejemplo, suponga que el rendimiento de un proceso química está relacionado con la temperatura de operación. Si mediante un modelo matemático es posible describir tal relación, entonces este modelo puede ser usado para propósitos de predicción, optimización o control. Para estimar los parámetros de un modelo de regresión son necesarios los datos, los cuales pueden obtenerse de experimentos planeados, de observaciones de fenómenos no controlados o de registros históricos (Pulido y Salazar, 2008, pp.360-472).

Método de mínimos cuadrados

Procedimiento para estimar los parámetros de un modelo de regresión que minimiza los errores de ajuste del modelo.

$$S = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n])^2 \quad (10\text{-}1)$$

De esta forma, se quieren encontrar valores de β_0 y β_1 que minimizan la suma de los errores cuadrados. Es decir, se busca ajustar la recta de manera que la suma de las distancias en forma vertical de los puntos a la recta se minimice.

$$\frac{\partial S}{\partial \beta_0} = -\sum_{i=1}^{n} 2(y_i - [\beta_0 + \beta_1 x_i])$$
$$\frac{\partial S}{\partial \beta_1} = -\sum_{i=1}^{n} 2x_i(y_i - [\beta_0 + \beta_1 x_i]) \quad (11\text{-}1)$$

1.1.4. Modelos lineales

Regresión lineal Simple

Sean dos variables X y Y, se quiere explicar él comportamiento de Y con respecto a los valores que toma X. Para esto, se mide el valor de Y sobre el conjunto de todos los valores de X, con lo que se obtienen n parejas de puntos (x1, y1), (x2, y2), …, (xn, yn). A Y se le denomina la variable dependiente o la variable de respuesta o a predecir y a X se le conoce como variable independiente o variable regresora o predictora. La variable X no necesariamente es aleatoria, Y sí es una variable aleatoria. Una manera de estudiar el comportamiento de Y con respecto a X es mediante un modelo de regresión que consiste en ajustar un modelo matemático:

$$Y = f(X) \quad (12\text{-}1)$$

a las n parejas de puntos. Con esto, se puede ver si dado un valor de la variable explicativa X es posible predecir el valor promedio de Y. Suponga que las variables X y Y están relacionadas linealmente y que, para cada valor de X, la variable dependiente, Y, es una variable aleatoria. Es decir, que cada observación de Y puede ser descrita por el modelo (Pulido y Salazar, 2008, pp.360-472).

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (13\text{-}1)$$

Regresión lineal múltiple

En muchas situaciones prácticas existen varias variables independientes o explicativas, que se cree que influyen o están relacionadas con una variable de respuesta Y, y por lo tanto será necesario tomar en cuenta si se quiere predecir o entender mejor el comportamiento de Y relación a todas las posibles variables predictoras. Por ejemplo, para explicar o predecir el consumo de electricidad en una casa, tal vez sea necesario considerar el tipo de residencia, el número de personas que la habitan, etcétera (Pulido y Salazar, 2008, pp. 360-472).

Sea $X_1, X_2, \ldots, X_k$ variables independientes o regresoras, y sea Y una variable de respuesta o a predecir, entonces el modelo de regresión lineal múltiple con k variables independientes es el polinomio de primer orden:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \tag{14-1}$$

donde los β_j son los parámetros del modelo que se conocen como coeficientes de regresión y ε es el error aleatorio, con media cero, $E(\varepsilon) = 0\ y\ V(\varepsilon) = \sigma^2$.

Coeficiente de determinación R^2

Para evaluar la calidad del ajuste es observar la forma en que el modelo se ajustó a los datos. En el caso de la regresión lineal simple esto se distingue al observar si los puntos tienden a ajustarse razonablemente bien a la línea recta x=y. Pero otro criterio más cuantitativo es el que proporciona el coeficiente de determinación o cuadrado del coeficiente de correlación múltiple:

$$R^2 = \frac{Variabilidad\ explicada\ por\ el\ modelo}{variabilidad\ total} = \frac{SC_R}{S_{yy}} \tag{15-1}$$

Coeficiente de determinación ajustado, R^2_{aj}

Este coeficiente se calcula de la siguiente manera:

$$R^2_{aj} = \frac{CM_{Total} - CM_E}{CM_{Total}} \tag{16-1}$$

Coeficiente de correlación

Mide la intensidad de la relación lineal entre dos variables X y Y.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{17-1}$$

Residuos

Es la diferencia entre lo observado y lo estimado o predicho. Ayudan a analizar el error de ajuste de un modelo (Pulido & Salazar, 2008, pp.360-472).

$$e_i = y_i - \hat{y}_i \tag{18-1}$$

Supuestos para modelos lineales

Normalidad

Para verificar la eficacia del modelo es necesario realizar una prueba de normalidad a los residuos es común utilizar la prueba de Jarque Bera la cual se basa en los coeficientes de asimetría y kurtosis, la expresión viene dada por:

$$JB = n\left[\frac{A^2}{6} + \frac{(K-3)^2}{24}\right] \tag{19-1}$$

Donde:

A =coeficiente de asimetría.

K =coeficinente de kurtosis.

Homocedasticidad

- **La prueba de Bartlett** es quizá la técnica que es utilizada para probar homogeneidad de varianza. En esta prueba los r_i en cada tratamiento no necesitan ser iguales; sin embargo, se recomienda que los r_i no sean menores que 3 y muchos de los r_i deben ser mayores de 5, La estadística de prueba es (Pulido y Salazar, 2008, p.350).

$$q = (N-k)logS_p^2 - \sum_{i=1}^{k}(n_i - 1)logS_i^2$$

$$= 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(n_i-1)^{-1} - (N-k)^{-1}\right) \tag{20-1}$$

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

$$X_0^2 = X_c^2 = 2.3026\frac{q}{c}$$

Cuando la hipótesis nula es cierta, la estadística tiene distribución aproximadamente χ_α^2 con t-1 grados de libertad.; cuando el muestreo se realiza en poblaciones normales.

- **Prueba de Goldfeld-Quandt:** Esta prueba asume que existe un punto que puede definir y se utiliza para diferencias la varianza del término de error. Lo individuos en estudio se dividen en dos grupos, se basa en evidencias de heteroscedasticidad en una comparación de la suma cuadrados de residuales, basado en le F de Fisher.

$$G - Q = F = \frac{SCE_1/n - p - 1}{SCE_2/n - p - 1} \tag{21-1}$$

La hipótesis nula que presenta la prueba G-Q es que los residuos cumplen homocedasticidad. Mientras mayor sea el F (G-Q), hay más evidencia estadística en contra de la hipótesis de homocedasticidad y es más probable que tenga heterocedasticidad (Aparicio, 2018, p.145).

Independencia

Con la prueba de Durbin Watson se diagnostica la presencia de correlación entre los residuos consecutivos, que es una posible manifestación de la falta de independencia ya que lo óptimo para este supuesto es que estén incorrelados (Pulido y Salazar, 2008, pp.360-472)

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{22-1}$$

La decisión sobre la hipótesis dada:

$$\begin{gathered} Si\ d < d_L\ Se\ rechaza\ H_0 \\ Si\ d > d_{ij}\ \ No\ se\ rechaza\ H_0 \\ Si\ d_L \leq d \leq d_{ij}\ Sin\ decisión \end{gathered} \tag{23-1}$$

1.1.5. Modelos lineales generalizados

Los GLM o modelos lineales generalizados como su nombre lo dice es algo más general de los modelos lineales conocidos como los que se ha mencionado el simple o múltiple, se sabe que los modelos considerados al inicio en este trabajo deben cumplir con los supuestos que son principalmente independencia en los residuos, es decir deben estar incorrelados, pero también

igualdad de varianzas, así como también normalidad en los residuos, esto ayuda a identificar que tan bueno es el ajuste propuesto. Si falla algún supuesto el ajuste no es tan bueno, por lo que se puede recurrir a otros modelos que no se requiera cumplan estos supuestos, pero si cumplan otras condiciones, como lo son los modelos lineales generalizados. Para saber cuándo usar un modelo se puede observar la variable a predecir.

- La variable dependiente trata de un conteo de casos
- La variable respuesta esta expresada en proporciones.
- una respuesta dicotómica, entre otros (Fox, 2016, pp.418-425).

Un modelo lineal generalizado (o GLM) consta de tres componentes:

1. Un componente aleatorio, que especifica la distribución condicional de la variable de respuesta, Y_i (para el i-ésimo de n observaciones muestreadas independientemente), dados los valores de las variables explicativas o predictoras del modelo a ajustar. En el modelo de GLM, la distribución de Y_i asume sigue una distribución que pertenece a una familia exponencial, como la Gaussiana, binomial, Poisson, gamma o gaussianas inversas.

2. Debe tener un predictor lineal, es decir, una función lineal de regresores:

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \qquad (24\text{-}1)$$

3. Una función enlace que linealiza es suave e invertible $g(\cdot)$, que transforma la expectativa de la variable de a predecir, $\mu_i = E(Y_i)$, en el predictor lineal (Fox, 2016, p.421).

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} \qquad (25\text{-}1)$$

Las funciones enlaces específicos que se pueden utilizar van variando de una familia a otra e incluso en una implementación de software de GLM a otra. Por ejemplo, no sería conveniente utilizar la identidad, enlaces inversos, cuadrados inversos o de raíz cuadrada con datos binomiales, así como también no sería prometedor usar el enlace logit, probit, log-log o log-log complementario con datos que no binomiales.

Las funciones de enlace más comúnmente utilizadas, $g(\mu_i) = \eta_i$, se observa en la Tabla 1-4. El vínculo canónico es la función que transforma la media en un parámetro canónico, de la familia de dispersión exponencial (Henrik & Thyregod, 2010, p.102).

Tabla 4-1: Enlace canónico, rango de respuesta y Función

Familia	Enlace canónico	Rango de Y_i	$V(Y_i\|\eta_i)$
Gaussiana	Identidad	$(-\infty, +\infty)$	ϕ
Binomial	logit	$\frac{0, 1, \ldots, n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	log	0, 1, 2, …	μ_i
Gamma	Inversa	$(0, \infty)$	$\phi{\mu_i}^2$
Gaussiana Inversa	Inversa Cuadrada	$(0, \infty)$	$\phi{\mu_i}^3$

Fuente: Fox, John, 2016.

Con ϕ es el parámetro de dispersión, η_i es el predictor lineal y μ_i es la expectativa de Y_i.

Modelo de regresión de Poisson

Sean $Y_1, Y_2, \ldots, Y_N$ variables aleatorias independientes, donde cada una representa el número de eventos observados de exposición n_i de acuerdo con el i-ésimo patrón de covariables, por lo que el valor esperado de variable respuesta se puede escribir como:

$$E(Y_i) = \mu_i = n_i\theta_i$$

Donde θ_i depende de las variables predictoras:

$$\theta_i = e^{x_i^T\beta}$$

Por lo que el GLM puede expresarse

$$E(Y_i) = \mu_i = n_i e^{x_i^T\beta}; Y_i \sim Pois(\mu) \quad (26\text{-}1)$$

Aplicando la función de enlace que se recomienda para la distribución de Poisson que es la función logarítmica, el modelo queda expresado como:

$$\log(\mu_i) = \log(n_i) + x_i^T\beta \quad (27\text{-}1)$$

Donde, $\log(n_i)$, se lo conoce como el Offset o la compensación (Dobson y Barnett, 2018, p.198).
La función de probabilidad de $y_1, y_2, \ldots, y_n$, se la puede calcular con:

$$L(y, \mu) = \prod_{i=1}^{n} p_i y_i = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} = \frac{\left(\prod_{i=1}^{n} \mu_i^{y_i}\right) e^{-\sum_{i=1}^{n}\mu_i}}{\prod_{i=1}^{n} y_i!} \quad (28\text{-}1)$$

Aplicando el logaritmo natural, se tiene:

$$ln\, L(y,\mu) = \sum_{i=1}^{n} y_i\, ln(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} ln\,(y_i!)$$

El parámetro μ_i se relaciona con los $\beta_0, \dots, \beta_n$, mediante la función de enlace:

$$\lambda_i = g^{-1}(x_i^T \beta)$$

Después de elegir la función de enlace que más se adecúe, la función de log-verosimilitud se maximiza, usando algunas técnicas de optimización para un determinado conjunto de individuos (Henrik & Thyregod, 2010, p.123).

Evaluación de la bondad de ajuste del modelo lineal generalizado

La función varianza del modelo Poisson es $V(\mu) = \mu$. El parámetro ϕ en el caso de que se trate de una distribución de Poisson se prefija en 1. Sin embargo, si requerimos estimar este parámetro igual que en el caso de los modelos lineales generalizados GLM, se define:

$$\hat{\sigma}^2 = \frac{X^2}{n-p} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \Big/ (n-p) \qquad (29\text{-}1)$$

Estimar el parámetro de dispersión lleva preliminarmente a una detección de no existencia de equidispersión asumida en un modelo de regresión Poisson (Figueroa, 2005, p.49).

Evaluación de la adecuación del GLM

Sobredispersión

El modelo lineal generalizado presenta mejoras para representar de datos de conteos, pero puede resultar inapropiado debido al incumplimiento de ciertos supuestos, la más común es la equidispersión. En varios casos puede presentar una subdispersión o sobredispersión, por lo que se propone pruebas para evaluar este supuesto.

La sobredispesión ocurre cuando $V(Y) > E(Y)$, es decir $\sigma^2 > 1$, Cuando existe exceso de variación en los datos, las estimaciones de los residuos pueden estar sesgadas, por lo que se puede presentar errores en las inferencias a partir de los parámetros del modelo de regresión (Figueroa, 2005, p.51).

Causas de sobredispersión

- Presencia de alta variabilidad en los datos.
- Los datos de la variable respuesta, no provienen de una distribución Poisson.
- Los eventos a considerar, no ocurren independientemente a través de un tiempo dado.
- Errores de especificación de la media μ como omitir variables explicativas.
- Errores al elegir la función de enlace.

La aplicación de pruebas para detectar sobredispersión, implican evaluar la relación entre χ^2o la correspondiente discrepancia y sus grados de libertad: $\frac{\chi^2}{gl}$ y $\frac{D}{gl}$.

Pruebas de modelos anidados

El modelo de regresión de Poisson son modelos anidados que se comparan en presecia de sobredispersión, para este modelo el modero de regresión de Poisson está anidado dentro del modelo de regresión binomial negativa, si se cumple $H_0: \alpha = 0$, por lo que se tiene:

$$Var(y_i|x_i) = \mu_i$$

Prueba de Wald

El test de Wald se implementa como una prueba t, que tiene una masa de 0.5, en 0 y una distribución normal para valores estrictamente positivos. Si es el caso se aplica el valor crítico de contraste de hipótesis unilateral $z_{1-\alpha}$.
La Hipotesis nula se basa en que no presenta sobredispersión (Brosa, 2002, p.68).

Estadística chi-cuadrado de Pearson

La estadística Chi-cuadrado de Pearson en el caso de la regresión de recuento o de Poisson es la estadística Pearson χ^2 está definida como:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{V(\widehat{\mu_i})} \qquad (30\text{-}1)$$

La medida mencionada se puede usar como medida de bondad de ajuste, ya que se calcula a partir de los datos y del modelo que se ajusta (Figueroa, 2005, p.48).

1.1.6. Modelo de regresión logística nominal

Distribución multinomial

Esta distribución es una generalización de la distribución binomial, es decir si un conjunto de n individuos puede clasificarse en k clases distintas las cuales deben ser excluyentes y exhaustivas, se tiene (Díaz Monroy y Morales Rivera, 2009, p.10).

1. Una muestra de n objetos de estudio, y a cada una de ellas se asigna una de las k clases.
2. $n_1 + n_2 + \cdots + n_k = n$.
3. Se puede definir una variable aleatoria $X_i (i = 1,2, \dots, k)$, que contiene el conteo del número de objetos que están ubicados en cada clase.
4. La probabilidad de que haya n_1 casos en la clase 1, y así para cada uno, se calcula:

$$P(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} \qquad (31\text{-}1)$$

Regresión logística

Una regresión logística puede expresar la probabilidad de ocurrencia de un evento como función de varias variables explicativas. En el caso general se expresa como un modelo donde hay p-variables predictoras $X_1, \dots, X_p$

$$P(Y = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} \qquad (32\text{-}1)$$

Donde: $\beta_0, \dots, \beta_p$ son los parámetros del modelo.

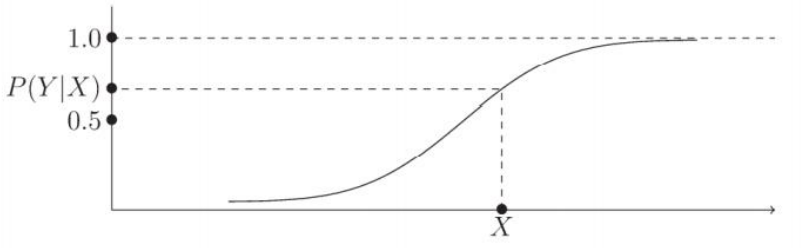


Figura 7-1: Función logística univariada

Fuente: Díaz, M; Morales, R. 2009.

Las variables predictoras o explicativas pueden ser de tipo cualitativo ordinal, cualitativo nominal o cuantitativa.

Regresión logística con respuesta politómica

Ocurre cuando la variable respuesta presenta más de dos modalidades, estas pueden ser de tipo ordinal o nominal.

Regresión logística nominal

La variable Y toma valores $1, 2, \dots, k$ $(k \geq 2)$, es decir dos o más categorías. Se denota como $A_1, \dots, A_k$ con k modalidades de la variable a predecir. Entonces se dice que la variable Y toma el valor de 1 si el individuo incurre en el evento A_1, en general toma el valor de k si el individuo incurre en el evento A_k y, además, se tiene p variables predictoras, se puede expresar la probabilidad de que Y tome cada uno de los k valores en función de las variables explicativas.
Por lo que se propone una función de k-1 funciones, de manera que se deban estimar p+1 coeficientes tal que para calcular la probabilidad sería (Díaz Monroy y Morales Rivera, 2009, p.173).

$$P(Y = j) = \frac{exp\ (\beta_j^T X)}{1 + \sum_{i=1}^{k-1} exp\ (\beta_j^T X)};\ \ para\ j = 1, \dots, k-1 \qquad (33\text{-}1)$$

Donde: $exp\left(\beta_j^T X\right) = exp\ (\alpha_i + \beta_{i1} X_1 + \cdots + \beta_{ip} X_p)$

Regresión logística ordinal

Se da cuando las categorías de la variable respuesta presentan algún tipo de ordenamiento, por lo que una regresión logística nominal, no sería adecuada ya que no tomaría en cuenta la relación de la variable respuesta con las variables predictoras, por lo que se propone algunos modelos como logit acumulativo, el modelo de categoría adyacente, modelo logit de continuación de razón, y también se puede considerar el modelo de odds proporcionales, el cual es el más común en softwares estadísticos, para la implementación en este tipo de modelos.
El modelo odds proporcionales se rige en el supuesto que considera el efecto de covariables de $X_1, \dots, X_p$ es igual para todas las categorías en la escala logarítmica, por lo que el modelo se expresa:

$$log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_k}\right) = \beta_{0j} + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (34\text{-}1)$$

El productor lineal tiene un intercepto el cual depende de la j-ésima categoría, sin embargo, las variables predictoras, no dependen de la categoría j (Díaz Monroy y Morales Rivera, 2009, p.174).

Interpretación de los coeficientes de regresión

Una interpretación de un modelo de regresión logística para los coeficientes puede ser mediante:

Riesgo relativo

$$RR = \frac{O(X^k)}{O(X^l)} = exp\left[\sum_{i=1}^{p} \beta_i (X_{ki} - X_{li})\right] \quad (35\text{-}1)$$

Donde X^k, X^l denotan el vector de observaciones para los individuos k y l, es decir que es una medida relativa del riesgo que relaciona un individuo con respecto de otro en términos de los parámetros que se obtuvieron en la regresión logística (Díaz Monroy y Morales Rivera, 2009, p.151).

1.1.7. Elección de las mejores variables predictoras

Se puede utilizar uno de los siguientes principios para la selección de un modelo:

a) Selección hacia adelante

Comienza con un modelo nulo, el procedimiento agrega, en cada paso, la variable que daría el valor p más bajo de la variable que aún no se ha incluido en el modelo. El procedimiento termina cuando todas las variables han sido agregadas, o cuando ninguna variable alcanza el límite preespecificado para el valor p. Es decir, La primera variable que se introduce es la que tiene mayor correlación (+ o -) con la variable dependiente o respuesta. La variable mencionada se introducirá en el modelo solo si cumple el criterio de entrada (Henrik y Thyregod, 2010, p.32).

b) Selección hacia atrás

En este caso comienza con el modelo completo, las variables se van eliminando del modelo en cada iteración, hasta que todas las variables restantes alcancen un límite especificado para sus valores p. En cada paso, se debe eliminar la variable que contenga el valor p más grande. Dado un contraste, se va eliminando una variable y es la menos significativa, se la descarta y se realiza nuevamente algoritmo (Alan, 2015, p.143).

c) Selección escalonada

Para esta selección se basa en una modificación del principio de selección directa. Las variables se agregan al modelo paso a paso, pero en cada paso, el procedimiento también examina si las variables que ya están en el modelo propuesto se pueden eliminar. Es decir, esta técnica es una combinación de los procedimientos anteriores.

d) Mejor selección de subconjunto

Para un k que empieza en 1 hasta un límite especificado por el investigador, este procedimiento se encarga de identificar un número específico de mejores modelos que contienen k variables (Henrik y Thyregod, 2010, p.33).

1.1.8. Comparación y selección de modelos

Para modelizar un conjunto de datos siempre hay una gran variedad de alternativas, pero el objetivo es determinar el tipo de modelo, considerando que las transformaciones sean más adecuadas, identificar las variables relevantes, descartando las innecesarias, y posteriormente abordar la diagnosis y la respectiva validación del modelo ajustado. Si el modelo está no tiene un buen ajuste, las estimaciones de los coeficientes pueden estar sesgadas. Sin embargo, mientras más variables predictoras se incluya en el modelo, se obtiene mejores pues el sesgo se reduce, aunque menos precisión debido a que el número de variables predictoras es proporcional a la varianza. Por lo que se recurre a varios criterios para comparar modelos y escoger el más adecuado.

Es importante enfatizar que en varias ocasiones no todos los criterios dan los mismos resultados por lo que el resultado final lo da el analista acorde a sus prioridades en la investigación (Aparicio, 2013, p.119).

- La significatividad de los predictores
- El coeficiente de determinación ajustado
- El error residual del ajuste s^2
- El estadístico C_p de Mallows
- El estadístico AIC (Akaike Information Criteria)
- El error de predicción PRESS

Comparación de modelos basada en las sumas de cuadrados

Es el criterio de significatividad de los predictores, donde propone las siguientes hipótesis

$$H_0: y = X_p\beta_p + \epsilon$$
$$H_1: y = X_{p+q}\beta_{p+q} + \epsilon$$

Para contrastar estas hipótesis se basa en una prueba de ANOVA basado en el estadístico F, es decir se debe calcular sus sumas cuadradas SSE(p) y SSE(p+q). Además, la diferencia indica la reducción del error que se debe a la implementación de las q variables predictoras adicionales, se nota también que H0 sigue una distribución Chi cuadrado. Por lo que se define el estadístico F para realizar la comparación de los modelos como:

$$F_q = \frac{(SSE(p) - SSE(p+q))/q}{SSE(p)/(n-p)} \sim F_{q,n-p} \quad (36\text{-}1)$$

Si el estadístico F resulta significativo determina que las q variables adicionales son significativas (Aparicio, 2013, p.119).

Los estadísticos AIC y BIC

El criterio de información de Akaike (Akaike, 1973) es basado en la función de verosimilitud que también incluye una penalización que aumenta dado el número de parámetros estimados en el modelo de regresión. Gana el modelo que da un buen ajuste, todo esto en términos de verosimilitud, el AIC está definido como:

$$AIC = -2l(\hat{\beta}) + 2p \quad (37\text{-}1)$$

Existe una versión que también considera el número de datos utilizados en el ajuste, conocido como Schwarz's Bayesian criteration, que se conoce como BIC, y se define:

$$BIC = -2l(\hat{\beta}) + \log(n)p \quad (38\text{-}1)$$

Si existen valores pequeños en ambos modelos estos reflejan mejores modelos (Montesinos, 2011, p.3).

1.1.9. Principales modelos lineales generalizados

Los modelos lineales son el caso más elemental de un GLM es decir de los modelos lineales generalizados, estos modelos estadísticos son herramientas metodológicas que de acuerdo a la situación permiten codificar diferentes situaciones de análisis dentro de un mismo esquema general, en la siguiente table se aprecia como trabaja alguno de los modelos generalizados principales (López Emelina y Ruiz Marcos, 2011).

Tabla 5-1: Principales modelos lineales generalizados

Naturaleza de la Variable Dependiente	**Componentes**		**Función de enlace**	**Modelo lineal**	
	Sistemático	**Aleatorio**			
Numérica/ cuantitativa	Numérico	Normal	Identidad	Regresión lineal	ML
	Categórico/mixtos(cuantitativos/ cualitativos)	Normal	Identidad	ANOVA´s	
Categórica/ binaria					MLG
No Agrupada	Mixto (cuantitativos/cualitativos)	Binomial	Logística/logit	Regre. Logística/logit	
Agrupada	categórico	Bin. generalizada	Logística/logit	Análisis logit/Probit	
Categórica/ politómica					
No agrupada	Mixto (cuantitativos/cualitativos)	Multinomial	Logística G.	Regre. Logística/logit multinomial	
Agrupada	Categórico	Multinomial	Logística G.		
Conteos	Mixto (cuantitativos/cualitativos)	Poisson	Logarítmica/log	Regresión de Poisson	
Frecuentista	categórico	Poisson	Logarítmica/log	Análisis log-lineal	

Fuente: López, E; Ruiz, M. 2011.

1.1.10. Comparación de LM y GLM

Se observa que el Modelo Lineal es el caso más elemental del Modelo Lineal Generalizado, las diferencias y similitudes que tienen los modelos establecidos, hacen que los modelos lineales generalizados sean un modelo más adecuado a las variables que se esté en tratamiento (López Emelina y Ruiz Marcos, 2011).

Tabla 6-1: Comparación de un LM y GLM

Modelo Lineal (ML)	**Modelo Lineal Generalizado (MGL)**
$y_i = \sum_j \beta_j X_{ij} + \varepsilon_i$	$y_i = \sum_j \beta_j X_{ij} + \varepsilon_i$
$\mu_i = E(Y_i)$	$\mu_i = E(Y_i)$
$\eta_i = \sum_j \beta_j X_{ij}$	$\eta_i = \sum_j \beta_j X_{ij}$
$\eta_i = \mu_i$	$\eta_i = g(\mu_i)$
y_i: vector de la variable respuesta	
X_{ij}: matriz de variables predictoras y covariables.	
β_j: vector de parámetros	
η_i: vector del predictor lineal	

Fuente: López, E; Ruiz, M. 2011.

1.2. Bases teóricas.

1.2.1. Poliducto

El poliducto es una red de tubería largo utilizadas para transportar productos derivados del petróleo como combustibles líquidos o cualquier otro producto terminado que proceden de refinerías. El transporte es realizado en lotes de productos conocidos como baches (Echeverría y Jiménez, 2014, p.2).

Considerando que los oleoductos transportan exclusivamente petróleo crudo y los gasoductos productos gaseosos, cabe enfatizar que un poliducto transporta diferentes hidrocarburos líquidos, que poseen características químicas y físicas en periodos que son programados y se los denomina baches (Mieser y Leffler, 2006, p.6).

Los poliductos pueden transportar líquidos o gases de un lugar a otro, los poliductos pueden ser pequeños de diámetros como de 4 pulgadas o de tamaño promedio que varía de 24 a 32 pulgadas de diámetro, y las grandes de 48 a 60 pulgadas; que pueden ser líneas cortas, líneas de reunión o líneas troncales. Además, los poliductos proporcionan la presión que se necesita los equipos de

bombeo y controladores, así también como válvulas, trampas raspadoras y reguladores (Menon, 2014, p.1).

1.2.2. Conceptos básicos

Barril

Para medir el volumen del petróleo y sus derivados, se utiliza la unidad de medida barril, que es equivalente a 158.98 litros los cuales están medido a 15.5°C o también 42 galones americanos (EP Petroecuador, 2018, p.6).

Capacidad de Refinación

Se hace referencia en un día, la cantidad máxima de crudo que se puede procesar relacionándolo con las unidades de destilación que hay en cada refinería (EP Petroecuador, 2018, p.6).

Comercialización

Se le conoce al proceso de compra y venta del crudo o sus derivados, puede ser dentro o fuera del país, por lo que se considera como comercialización interna o internacional; además, rige normas específicas (EP Petroecuador, 2018, p.6).

Crudo

Se encuentra debajo de la superficie de la tierra de forma líquida en reservorios naturales como una mezcla de hidrocarburos (EP Petroecuador, 2018, p.8).

1.2.3. Productos

Diésel 1 o Kerosene

Es un destilado medio que se ha utilizado en mercados de iluminación o de calefacción en algunos lugares de Asia o Japón, pero en general se usa como combustible en determinadas industrias (EP Petroecuador, 2019, p.8).

Diésel 2

Es un destilado medio que es utilizado como combustible para transporte pesado o generación eléctrica e industria (EP Petroecuador, 2019, p.8).

Fuel Oíl # 4

Es un destilado medio, pero también una mezcla de residuos, que es utilizado para satisfacer necesidades del transporte marítimo o el sector eléctrico (EP Petroecuador, 2018, p.6).

Fuel Oíl # 6

Es una mezcla de residuos que contiene diluyente, es utilizado para la generación de calefacción en el hemisferio norte, generación eléctrica, además, como fuente de energía en varias industrias que procesan azúcar, cemento, vidrio y otros elementos industriales. También, es utilizado como combustible marítimo y es conocido como Bunker (EP Petroecuador, 2018, p.6).

Gas Natural Asociado

Es un hidrocarburo de estado gaseoso, que se encuentra en el subsuelo, como solución o en contacto con el crudo de petróleo, por lo que al ser explotados se produce el gas natural, así como también el líquido, la relación debe ser menor a los 100.000 pies cúbicos que hay por barril normal, las mediciones deben ser hechas en una condición atmosférica normal (EP Petroecuador, 2018, p.7).

1.2.4. Generalidades de los procesos de extracción, almacenamiento y transporte de GLP

El GLP se extrae de fuentes fósiles junto con otros hidrocarburos como el gas natural y/o el petróleo crudo. Industrialmente el GLP es obtenido de dos formas: a partir la refrigeración del gas asociado (gas natural disuelto en crudo) mediante procesos de absorción, compresión y adsorción; o a través de las diferentes etapas de refinación del petróleo tales como la destilación del crudo, hydrocracking, reformado de naftas, FCC y la coquización (Antaki, 2003, p.31).

Densidad del GLP

En la fase de vapor el GLP, la densidad va variando directamente proporcional con la presión e inversamente proporcional con la temperatura y esto genera un crecimiento considerable en el volumen y el decremento significativa en su densidad (Warren, 2006, p.908).

Para el proceso que garantiza y regula el comercio justo de hidrocarburos es decir con masa y volumen entre los terminales de almacenamiento es la densidad relativa para los cálculos de la transferencia de custodia. Con los cálculos se obtiene la masa del hidrocarburo. Se obtiene resultados de las dos terminales en estudio y estas deben ser equivalentes con la finalidad de conocer las pérdidas de los hidrocarburos que están siendo transportados mediando el poliducto (Warren, 2006, p.209).

Velocidad de circulación del GLP

Relaciona el caudal de derivados que se transporta mediante el poliducto con el diámetro de las tuberías, y el objetivo es facilitar la lectura de la velocidad de circulación del GLP. Además, para el GLP, la presión de vapor constituye una medida indirecta de la temperatura mínima por debajo de la cual se produce su vaporización (Lluch, 2011, p.90).

Compresibilidad del GLP

La compresibilidad se define como la variación volumétrica de una sustancia por efecto de la presión ejercida sobre la misma. La compresibilidad es mínima en zonas de alta presión (puntos de descarga de las bombas centrífugas) y máxima en zonas de baja presión (puntos de succión de las bombas) (Carson, 2002, pp.47-48).

Viscosidad del GLP

Es la resistencia a fluir de un líquido, que es el resultado de efectos combinados de la cohesión y la adherencia. Se produce la viscosidad por el efecto de deslizamiento o corte que resulta del movimiento de una capa de fluido con respecto a otro y esta es distinta de la atracción molecular. La conocida ley de viscosidad de Newton propone que, para porcentaje de deformación angular del fluido, el esfuerzo cortante incrementa como la viscosidad y se la conoce como viscosidad absoluta o dinámica (μ).

$$\mu = \frac{\tau}{\frac{d\mu}{dy}} \qquad (39\text{-}1)$$

1.2.5. Red de poliductos actualmente operativos

La gerencia de transporte de EP Petroecuador es la entidad pública encargada del control técnico-operativo en las diferentes fases del transporte y almacenamiento de derivados de petróleo. En el Ecuador la red de poliductos, mostrada en la Figura 1, posee aproximadamente 1 630 km de longitud de tubería (EP Petroecuador, 2018, pp.24-30).

Figura 8-1: Diagrama de la red de poliductos en Ecuador

Fuente: (EP Petroecuador, 2021).

CAPÍTULO II

2. MARCO METODOLÓGICO

2.1. Tipo de la Investigación

Por el método de investigación mixta, porque intervienen variables categóricas como el poliducto, tramo o el tipo de producto y variables numéricas como número de barriles por depósito de acopio, entre otros, según el objetivo es aplicada ya que se centra en análisis de las proyecciones obtenidas en el campo de petróleos, según el nivel de profundización en el objeto de estudio descriptiva y explicativa ya que se busca realizar un análisis del número de barriles de derivados por poliductos y ajustar un modelo para obtener pronósticos, según la manipulación de variables no experimental porque la información obtenida es de una fuente secundaria, según el tipo inductiva ya que se busca aproximar el número de barriles de derivados mediante poliductos en periodos tiempo futuros, según el periodo temporal es transversal que no se sigue al objeto de estudio durante el lapso de estudio (Hernández Sampieri et al. 2014).

2.2. Diseño de la investigación no experimental

Si utiliza un método de investigación mixto y según la manipulación de variables es un diseño no experimental (Berger et al. 2018).

2.2.1. *Localización de estudio*

El proyecto se lo realiza en el Empresa Pública Petroecuador en la ciudad de Quito, AlpallanaE8-86 y Av. 6 de Diciembre.

Figura 1-2: Mapa de la Ubicación de EP Petroecuador

Fuente: (Google maps).

2.2.2. *Población de estudio*

Se realiza el análisis con información de cuatro años (2017, 2018, 2019, 2020) es decir 224 individuos objetos de estudio, que contiene el número de barriles que EP Petroecuador distribuyen a cada centro de acopio a Nivel Nacional mediante poliductos.

2.2.3. *Método de muestreo*

La información con la que se va a trabajar es adquirida directamente por la empresa EP Petroecuador, razón por la cual no se aplica un método de muestreo.

2.2.4. *Tamaño de la muestra*

Se realiza el estudio de los 224 individuos objetos de estudio que presentan información obtenida durante 4 años (2017, 2018, 2019, 2020) por EP Petroecuador.

2.2.5. *Técnica de recolección de datos*

La información adquirida proviene del departamento de Servicios de Formación y Capacitación de la empresa EP Petroecuador, que contiene todos los factores de los que depende la distribución del número de barriles mediante poliductos.

2.2.6. *Identificación de variables*

- Número de barriles de derivados refinados e importados que se distribuyen mediante poliductos.
- Nombre del poliducto
- Año en el que se recolecta la información
- Tipo de producto
- tramo
- Capacidad Bombeo Bls/hr
- Caudal Promedio Bls/hr
- Volumen Transportado

2.2.7. *Modelo estadístico*

Los modelos estadísticos a ejecutarse serán técnicas univariantes y multivariantes, dada el tipo de información obtenida. Además, se usará para predicciones del número de barriles de derivados, modelos lineales generalizados y para validar el modelo ejecutado, se probará supuestos de normalidad (Jarque Bera) Homocedasticidad (Bartlett) e independencia (Durbin Watson) así como para los GLM se probará sobredisperción.

2.3. Variables en estudio

2.3.1. Operacionalización de variables.

Tabla 1-2: Operacionalización de variables

Nombre de la variable	Descripción	Tipo de variable	Escala de medición	Categoría o intervalo
NBDRI	Número de barriles de derivados refinados e importados que se distribuyen mediante poliductos.	Cuantitativa	Razón	[0 ,+∞[
CBOMBEO	Capacidad Bombeo Bls/hr.	Cuantitativa	Razón	[0 ,+∞[
CPROMEDIO	Caudal Promedio Bls/hr.	Cuantitativa	Razón	[0 ,+∞[
VTRANSP	Volumen Transportado	Cuantitativa	Razón	[0 ,+∞[
POLIDUCTO	Nombre del poliducto	Cualitativa	Nominal	
AÑO	Año en el que se recolecta la información	Cualitativa	Ordinal	
PRODUCTO	Tipo de producto	Cualitativa	Nominal	
TRAMO	tramo	Cualitativa	Nominal	

Elaborado por: Bastidas, Johanna, 2022.

Teniendo en cuenta las variables categóricas, estas también se codifican para realizar un mejor análisis.

Tabla 2-2: Codificación de la variable poliducto

Poliducto	Codificación
TRES BOCAS - PASCUALES - CUENCA	1
LIBERTAD-PASCUALES-MANTA-MONTEVERDE-CHORRILLO	2
SHUSHUFINDI - QUITO	3
QUITO-AMBATO-RIOBAMBA	4
ESMERALDAS - SANTO DOMINGO - QUITO - MACUL	5

Elaborado por: Bastidas, Johanna, 2022.

Tabla 3-2: Codificación para la variable Tramo

TRAMO	Codificación
PASCUALES - CUENCA	1
TRES BOCAS - FUEL OIL	2
TRES BOCAS - PASCUALES	3
LIBERTAD - MANTA	4
LIBERTAD - PASCUALES	5
MONTEVERDE - CHORRILLO	6
SHUSHUFINDI - QUITO	7
QUITO - AMBATO - RIOBAMBA	8
ESMERALDAS - SANTO DOMINGO	9
SANTO DOMINGO-BEATERIO	10
SANTO DOMINGO-PASCUALES	11

Elaborado por: Bastidas, Johanna, 2022.

Tabla 4-2: codificación de la variable producto

PRODUCTOS	Codificación
DESTILADO1	1
DIESEL 1	2
DIESEL 2	3
DIESEL OIL	4
DIESEL P. IMPORTADO	5
DIESEL PREMIUM	6
FUEL OIL	7
G. BASE	8
G.EXTRA	9
G.EXTRA 85 OCT	10
G.SUPER	11
G.SUPER 90 OCT	12
G.SUPER 92 OCT	13
GAS. IMPORTADA	14
GLP	15
JET A1	16
JET FUEL	17
NAO	18
PREMEZCLA	19

Elaborado por: Bastidas, Johanna, 2022.

CAPÍTULO III

3. MARCO DE RESULTADOS Y DISCUSIÓN DE RESULTADOS

3.1. Análisis Descriptivo

Se realiza un análisis descriptivo tanto para variables cuantitativas como cualitativas, para poder tener un panorama general de las variables en estudio.

3.1.1. Variables cualitativas

Tabla 1-3: Tabla de contingencia de poliductos por cada año en estudio

	AÑO				
POLIDUCTO	**2017**	**2018**	**2019**	**2020**	**Total**
TRES BOCAS - PASCUALES - CUENCA	16	16	16	16	**64**
LIBERTAD-PASCUALES-MANTA MONTEVERDE CHORRILLO	17	17	17	17	**68**
SHUSHUFINDI - QUITO	5	5	5	5	**20**
QUITO-AMBATO-RIOBAMBA	6	6	6	6	**24**
TRES BOCAS - PASCUALES - CUENCA	12	12	12	12	**48**
Total	**56**	**56**	**56**	**56**	**224**

Realizado por: Bastidas, Johanna, 2022.

Se puede observar que en los poliductos que se transporta más productos derivados y refinados fue en el poliducto Libertad-Pascuales-Manta Monteverde Chorrillo, pues 68 productos se transportaron mediante este poliducto, así como también se puede apreciar que en el poliducto Shushufindi-Quito, se transportaron menos productos, 20 derivados y refinados.

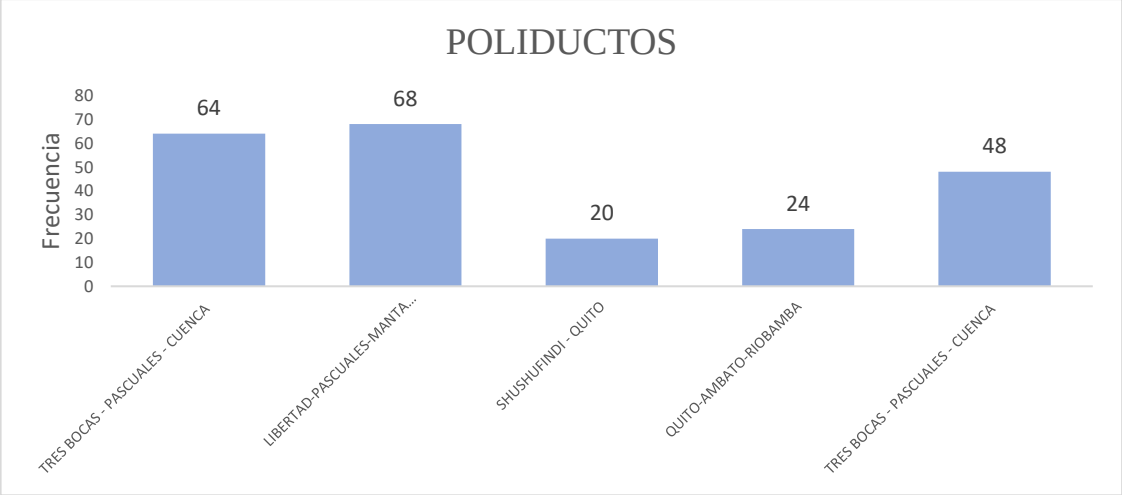


Gráfico 1-3: Diagrama de barras de los poliductos en EP Petroecuador.

Realizado por: Bastidas, Johanna, 2022.

De un total de 224 productos transportados durante el año 2017 al 2020, en el Gráfico 1-3**,** se observa el número de productos derivados y refinados que se han transportado por poliducto, corroborando con mayor incidencia el transporte de la producción del poliducto Libertad-Pascuales-Manta Monteverde Chorrillo.

Tabla 2-3: Tabla de contingencia del tramo versus el año

	AÑO				
TRAMO	**2017**	**2018**	**2019**	**2020**	**Total**
PASCUALES - CUENCA	7	7	7	7	28
TRES BOCAS - FUEL OIL	1	1	1	1	4
TRES BOCAS - PASCUALES	8	8	8	8	32
LIBERTAD - MANTA	7	7	7	7	28
LIBERTAD - PASCUALES	9	9	9	9	36
MONTEVERDE - CHORRILLO	1	1	1	1	4
SHUSHUFINDI - QUITO	5	5	5	5	20
QUITO - AMBATO - RIOBAMBA	6	6	6	6	24
ESMERALDAS - SANTO DOMINGO	5	5	5	5	20
SANTO DOMINGO-BEATERIO	5	5	5	5	20
SANTO DOMINGO-PASCUALES	2	2	2	2	8
Total	56	56	56	56	224

Realizado por: Bastidas, Johanna, 2022.

A partir del tramo que se utiliza de cada poliducto, se transportaron más productos de derivados y refinados mediante el tramo Libertad-Pascuales pues durante el periodo en estudio hubo un total de 36 productos, y los tramos en los que menos se transportó los productos fueron mediante el tramo Tres Bocas – Fuel Oil, y Monteverde-Chorrillo pues durante el periodo en estudio se transportó 4 productos.

Tabla 3-3: Tabla de contingencia del producto versus el año

	AÑO				
PRODUCTO	**2017**	**2018**	**2019**	**2020**	**Total**
DESTILADO1	1	1	1	1	4
DIESEL 1	4	4	4	4	16
DIESEL 2	7	7	7	7	28
DIESEL OIL	1	1	1	1	4
DIESEL P. IMPORTADO	1	1	1	1	4
DIESEL PREMIUM	7	7	7	7	28
FUEL OIL	1	1	1	1	4
G. BASE	4	4	4	4	16
G.EXTRA	5	5	5	5	20
G.EXTRA 85 OCT	1	1	1	1	4
G.SUPER	4	4	4	4	16
G.SUPER 90 OCT	1	1	1	1	4
G.SUPER 92 OCT	1	1	1	1	4
GAS. IMPORTADA	2	2	2	2	8
GLP	3	3	3	3	12
JET A1	6	6	6	6	24
JET FUEL	1	1	1	1	4
NAO	3	3	3	3	12
PREMEZCLA	3	3	3	3	12
Total	**56**	**56**	**56**	**56**	**224**

Realizado por: Bastidas, Johanna, 2022.

Desde el periodo 2017-2020, los productos que más se transportaron fue Diesel 2, y Diesel Premium y 8 productos que menos se transportaron fueron Destilado 1, Diese Oil, Diesel P. importado, Fuel Oil, G. Extra 85 Oct, G. Super Oct, G. Susper 92 Oct, Jet Fuel.

3.1.2. Variables Cuantitativas

Tabla 4-3: Resumen descriptivo de la variable volumen

VOLUMEN (Barriles)	
Media($\bar{x}$)	1851140,58
Mediana($\tilde{x}$)	957238,071
Moda	1851140
Desviación estándar(s)	2495547,12
Varianza de la muestra(s^2)	6,2278E+12
Curtosis	3,22365742
Coeficiente de asimetría	1,88583654
Rango	11671653
Mínimo	0

Máximo	11671653
Suma	414655490
Cuenta	224

Realizado por: Bastidas, Johanna, 2022.

Durante el año 2017 al 2020 se han transportado en promedio aproximadamente 1851141 barriles de productos refinados y derivados, teniendo en cuenta casos en los que se transportaba 0 barriles de productos o hasta 11671653 barriles de productos, presenta una asimetría positiva, es decir, los valores del volumen oscilan con mayor frecuencia por encima de la media. Además, presenta una curtosis positiva de 3.22 tratándose de una distribución leptocúrtica.

Tabla 5-3: Resumen descriptivo de la variable capacidad

CAPACIDAD	
Media($\bar{x}$)	38647,2321Bls/dia
Mediana($\tilde{x}$)	24000Bls/dia
Moda	21600 Bls/dia
Desviación estándar(s)	32669,14 Bls/dia
Varianza de la muestra(s^2)	1067272667
Curtosis	-0,49 Bls/dia
Coeficiente de asimetría	0,92>0
Rango	106840 Bls/dia
Mínimo	1160 Bls/dia
Máximo	108000 Bls/dia
Suma	8656980 Bls/dia
Cuenta	Bls/dia 224

Realizado por: Bastidas, Johanna, 2022.

La capacidad de bombeo de cada centro de acopio en donde llegan los productos derivados y refinados tienen una capacidad en promedio de 38647,23 barriles, teniendo en cuenta que la mayoría de los centros de acopio son tienen una capacidad estimada de 21600 barriles, así como también hay centros de acopio con una capacidad mínima de 1160 barriles o de una capacidad máxima para 10800 barriles. Además, presenta una asimetría positiva, pero una curtosis negativa denotado una curva platicúrtica.

Tabla 6-3: Resumen descriptivo de la variable caudal

CAUDAL	
Media($\bar{x}$)	1629,1 BLS/hr
Mediana($\tilde{x}$)	1000 BLS/hr
Moda	900 BLS/hr
Desviación estándar(s)	1342,4 BLS/hr
Varianza de la muestra(s^2)	1801960,64
Curtosis	-0,4 <0BLS/hr
Coeficiente de asimetría	0,94>0 BLS/hr
Rango	4150 BLS/hr
Mínimo	350 BLS/hr
Máximo	4500 BLS/hr
Suma	364918 BLS/hr
Cuenta	224

Realizado por: Bastidas, Johanna, 2022.

El Caudal Promedio aproximado para el transporte de los productos derivados y refinados es de 1629,1 barriles por hora, aunque en su mayoría el caudal es de 900 barriles por hora, existiendo así caudales mínimos desde 350 barriles por hora o máximos de hasta 4500 barriles por hora. Además, denota una asimetría positiva, y una curtosis negativa denotando una curva platicúrtica.

Tabla 7-3: Resumen descriptivo de la variable Volumen Despachado

VOLUMEN DESPACHADO (Barriles)	
Media($\bar{x}$)	1809239,39
Mediana($\tilde{x}$)	1451662,64
Moda	1828383,16
Desviación estándar(s)	2050714,79
Varianza de la muestra(s^2)	4,2054E+12
Curtosis	4,34077746
Coeficiente de asimetría	2,01260422
Rango	11074193
Mínimo	0
Máximo	11074193
Suma	405269623
Cuenta	224

Realizado por: Bastidas, Johanna, 2022.

El Volumen despachado dentro del periodo en estudio de cada producto fue en promedio de aproximadamente 1809239,39 barriles, considerando la demanda en cada centro de acopio esta fue variando desde 0 barriles hasta 11074193 barriles. Además, el volumen despachado presenta una asimetría positiva y una curtosis positiva tratándose de una curva leptocúrtica.

3.2. Análisis de métodos de regresión

3.2.1. Matriz de varianzas y covarianzas

Tabla 8-3: Matriz de varianzas y covarianzas de variables cuantitativas

	Volumen	**Capacidad**	**Caudal**	**vdespachado**
Volumen	$6{,}20*10^{12}$	$3{,}50*10^{10}$	$1{,}40*10^{9}$	$2{,}62*10^{12}$
Capacidad	$3{,}50*10^{10}$	$1{,}06*10^{9}$	$4{,}32*10^{7}$	$3{,}00*10^{10}$
Caudal	$1{,}40*10^{9}$	$4{,}32*10^{7}$	$1{,}79*10^{6}$	$1{,}3010^{9}$
vdespachado	$2{,}62*10^{12}$	$3{,}00*10^{10}$	$1{,}30*10^{9}$	$4{,}19*10^{12}$

Realizado por: Bastidas, Johanna, 2022.

Con las variables en estudio podemos notar que existen pares de variables que tienen una covarianza muy alta o incluso variables que, al tener una covarianza nula, están son independientes o incorreladas. Así como también las varianzas de cada variable son mínimas, aunque también alcanzan varianzas muy altas como se puede observar en la Tabla 8-3. La varianza de volumen despachado.

3.2.2. Matriz de correlación

Tabla 9-3: Matriz de correlación

	Volumen	**Capacidad**	**Caudal**	**vdespachado**
Volumen	1,00	0,43	0,42	0,51
Capacidad	0,43	1,00	0,99	0,45
Caudal	0,42	0,99	1,00	0,47
vdespachado	0,51	0,45	0,47	1,00

Realizado por: Bastidas, Johanna, 2022.

Con la matriz de correlación podemos ver que tan correlacionadas están las variables, pues al ir de un valor de -1 a 1, indicando el número negativo una correlación inversamente proporcional y el signo positivo de un número una correlación directamente proporcional.

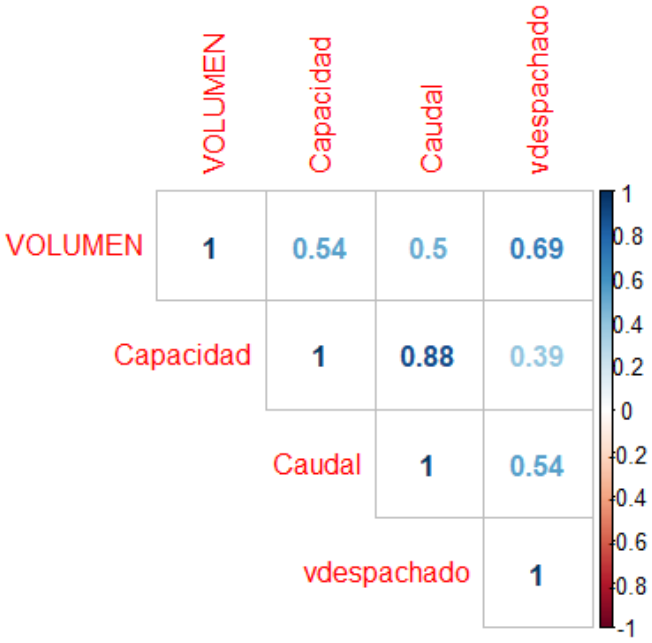


Gráfico 2-3: Correlogrma

Realizado por: Bastidas, Johanna, 2022.

Como se observa en la Gráfico 2-3, existen pares de variables que están altamente correlacionadas como Capacidad y Caudal con una correlación de 0.88 es decir, tienen una correlación directa y se presenta con un color azul más intenso que los demás, Así también, volumen despachado con una correlación de 0.39, pero las variables como volumen y volumen despachado presentan una correlación positiva de 0.69 que no es tan pronunciada.

3.2.3. Variables redundantes

Se realiza un análisis de variables redundantes solamente para variables cuantitativas, y ver si existe alguna combinación lineal entre estas ya que, si existiese sería como usar información redundante.

Se realiza el cálculo de autovalores propios asociados a sus autovectores propios para este análisis, obteniéndose los siguientes resultados:

Tabla 10-3: Autovalores de las variables cuantitativas en estudio

$4{,}11*10^{18}$	0	0	0
0	$8{,}99*10^{14}$	0	0
0	0	$3{,}99*10^{10}$	0
0	0	0	$1{,}27*10^{6}$

Realizado por: Bastidas, Johanna, 2022.

Tabla 11-3: Autovectores de las variables cuantitativas

$-4{,}60*10^{-5}$	1,02	$1{,}35*10^{2}$	$1{,}00*10^{6}$
$6{,}22*10^{3}$	$-9{,}99*10^{5}$	$-3{,}99*10^{4}$	6,40
$2{,}72*10^{2}$	$-3{,}99*10^{4}$	$9{,}99*10^{5}$	$-1{,}35*10^{2}$
$1{,}00*10^{6}$	$6{,}22*10^{3}$	$-2{,}38*10$	$-3{,}10*10^{3}$

Realizado por: Bastidas, Johanna, 2022.

Se observa que no existen autovalores nulos o cercanos a 0, por lo que el $rango(S_x) = 4$

Con $rango(S_x) = 4\ y\ p = 4$

Se tiene:

$$rango(S_x) \leq p$$
$$r \leq p \qquad (1\text{-}3)$$
$$4 \leq 4$$

Dado el teorema de la dimensión, no existe variable que sea combinación lineal de otra.

Para el análisis del número de barriles de petróleo derivados e importados mediante poliductos a nivel nacional, entran en estudio 4 variables cuantitativas, las cuales están incorrelados, por lo que no existe variable que sea combinación lineal de las otras, siendo así; que se continua el estudio con todas las variables estadísticas.

3.3. Variable respuesta: volumen transportado

Dada la variable respuesta Volumen Transportado, veamos su distribución

Tabla 12-3: Resumen estadístico de la variable respuesta Volumen Transportado

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	103058	957238	1851141	2451752	11671653

Realizado por: Bastidas, Johanna, 2022.

Como la media es mayor a la mediana, la distribución del volumen transportado es sesgada a la derecha. Podemos corroborarlo visualmente:

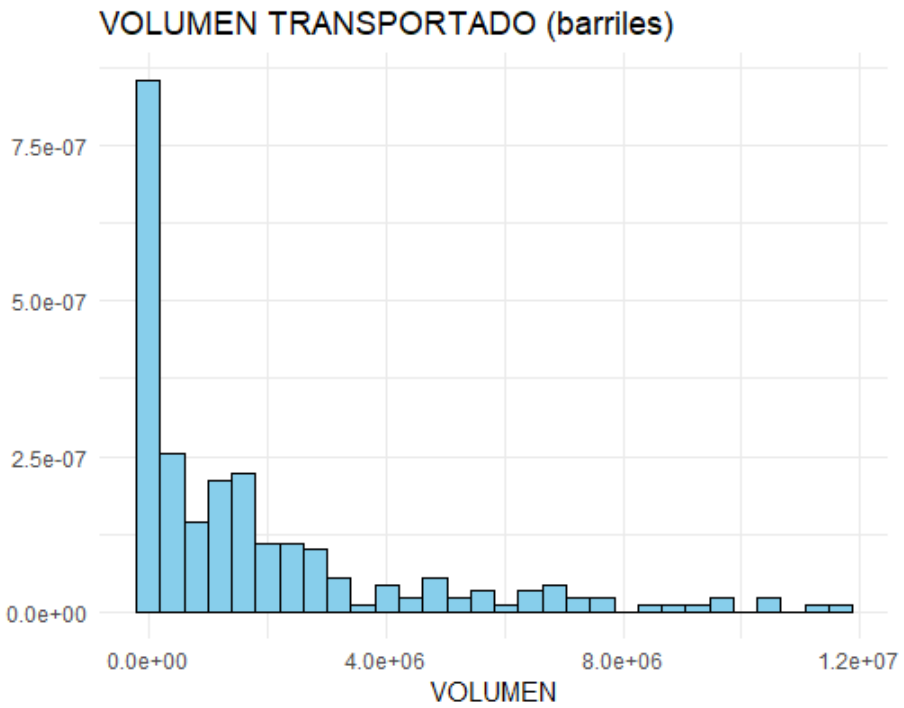


Gráfico 3-3: Histograma del número de barriles transportado

Realizado por: Bastidas, Johanna, 2022.

Se observa en el histograma realizado para la variable cuantitativa "volumen transportado" que no sigue una distribución normal, es decir hablamos de una distribución asimétrica, con un sesgo a la derecha. Una distribución asimétrica positiva y curtosis positiva que denota una curva leptocúrtica.

3.3.1. Modelo 1: Regresión lineal múltiple

Se genera la ecuación de Regresión Lineal múltiple de VOLUMEN sobre Xi

$$\widehat{Volumen} = \hat{\beta}_0 + \hat{\beta}_1 poliducto + \hat{\beta}_2 tramo + \hat{\beta}_3 producto + \hat{\beta}_4 \text{año} + \hat{\beta}_5 capacidad + \hat{\beta}_6 caudal + \hat{\beta}_7 vdespachado \quad (2\text{-}3)$$

Obteniendo mediante el software R, los siguientes coeficientes:

Tabla 13-3: Modelo 1: Regresión lineal múltiple

	Modelo de regresión lineal múltiple							
	Intercept	**poliducto**	**tramo**	**productos**	**año**	**Capacidad**	**Caudal**	**vdespachado**
Coef	759200000	1930000000	-815600000	33300000	-805000	92650	-1800000	639,40

Realizado por: Bastidas, Johanna, 2022.

Obteniendo un modelo:

$$\widehat{Volumen} = \hat{\beta}_0 + \hat{\beta}_1 poliducto + \hat{\beta}_2 tramo + \hat{\beta}_3 productos + \hat{\beta}_4 \text{año} + \hat{\beta}_5 Capacidad + \hat{\beta}_6 Caudal + \hat{\beta}_7 Vdespachado$$

$$\widehat{Volumen} = 759200000 + 1930000000\, poliducto - 815600000\, tramo + 33300000\, productos - 805000\, \text{año} + 92650\, Capacidad - 1800000\, Caudal + 639.40\, Vdespachado \quad (3\text{-}3)$$

Evaluar el rendimiento del modelo 1:

Tabla 14-3: Coeficiente de determinación, modelo 1

Coeficiente de determinación	0,4435

Realizado por: Bastidas, Johanna, 2022.

El modelo ajustado al tener un coeficiente de determinación del 44.35% no es tan bueno, es muy bajo para considerar un ajuste adecuado y obtener predicciones con un error pequeño que sería lo óptimo. Ahora, es coherente realizar un análisis más a fondo de la información, es decir detectar posibles datos atípicos que puedan estar distorsionando los resultados, o variables que no sean significativas en el estudio.

Tabla 15-3: ANOVA modelo 1

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**	
poliducto	1	$5{,}38*10^{17}$	$5{,}38*10^{17}$	150.260	0,0001407	(***)
tramo	1	$2{,}21*10^{16}$	$2{,}21*10^{16}$	0,6165	0,4332248	
productos	1	$1{,}13*10^{16}$	$1{,}13*10^{16}$	0,315	0,5751966	
año	1	6,56E+15	$6{,}56*10^{15}$	0,1833	0,668953	
Capacidad	1	3,03E+18	$3{,}03*10^{18}$	848.037	$2{,}2010^{-16}$	(***)
Caudal	1	9,30E+16	$9{,}30*10^{16}$	25.989	0,1084014	
V_despachado	1	2,46E+18	$2{,}46*10^{18}$	686.147	$1{,}25*10^{-11}$	(***)
Residuales	216	7,73E+18	$3{,}58*10^{16}$			

Realizado por: Bastidas, Johanna, 2022.

Se observa en la Tabla 15-3. que las variables poliductos, capacidad, y volumen despachado son significativas en el ajuste del modelo, y las 4 variables restantes no son significativas por lo que una posible solución sería, eliminar dichas variables, del modelo, tomando en cuenta que esto sería perder información necesaria para la interpretación de la variable a predecir.

Supuestos: Modelo 1

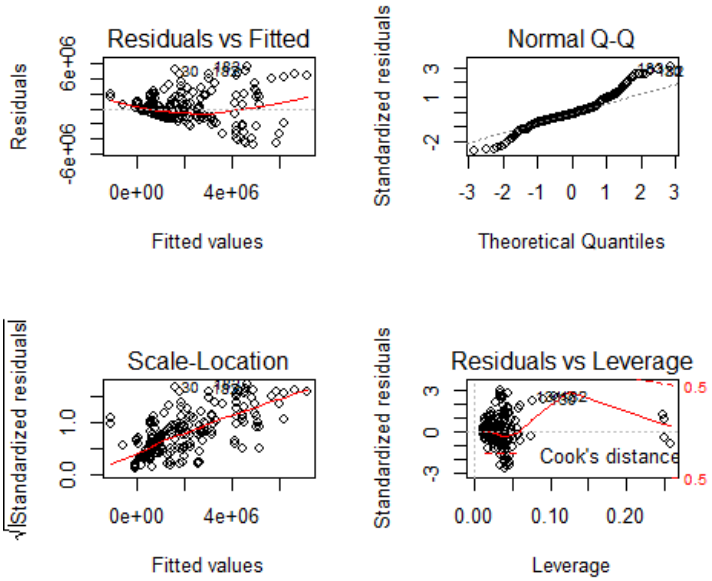


Gráfico 4-3: Supuestos del modelo 1

Realizado por: Bastidas Johanna, 2022.

En el Gráfico 4-3, se evidencia a priori que no cumple la normalidad, ni homocedastidad, y tampoco la independencia. Sin embargo, es pertinente recurrir a test teóricos, para corroborar lo observado en la gráfica.

Tabla 16-3: Valores p de los supuestos para el modelo 1

Valores P		
Normalidad	**Homocedasticidad**	**Independencia**
7,34E-02	0,2566	0,0000022

Realizado por: Bastidas, Johanna, 2022.

En la Tabla 16-3, podemos observar que con los valores obtenidos luego de haber aplicado la prueba teórica Jarque bera se obtiene un p valor de 7.34E-02 el cual es menor al nivel de significancia considerado 0.05 por lo que se rechaza H0 y se concluye que los residuos no siguen una distribución normal. Además, para probar Homocedasticidad se recurre al test de Goldfeld-Quant con el cual se obtiene un p valor de 0.256 el cual es mayor que el nivel de significancia por lo que no se rechaza H0 y podemos concluir que los residuos cumplen con el supuesto de homocedastidad. También, para probar la Independencia se recurre al test de Durbin Watson con el que se obtiene un valor aproximado de 0.0000022 el cual es menor a 0.05 por lo que se rechaza H0, y se concluye que no cumple con el supuesto de independencia. Por lo tanto, el modelo 1

ajustado cumple con uno de los supuestos, pero al no cumplir con la normalidad ni independencia en los residuos, el modelo ajustado no se lo puedo considerar completamente adecuado.

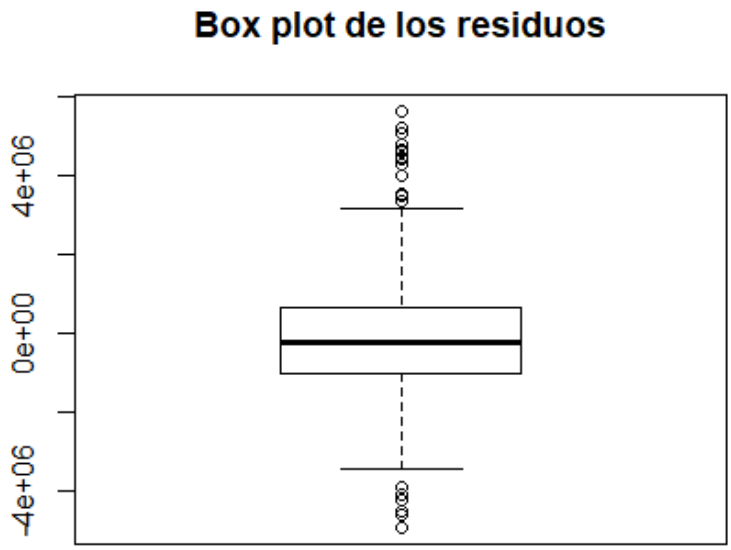


Gráfico 5-3: Diagrama de caja para los residuos del modelo 1

Realizado por: Bastidas, Johanna, 2022.

Se realiza un diagrama de cajas para los residuos del modelo 1 ajustado por lo que se puede evidenciar la presencia de datos atípicos que posiblemente sean causantes de que el modelo ajustado no sea tan bueno, por lo que se recurre a técnicas estadísticas para poder mejorarlo.

3.4. Detección de datos atípicos

Sea $X = (X_{1,}X_{2,}X_{3,}X_{4})^{t}$ vector estadístico 4-variante con componentes cuantitativas.

Se tiene una distribución estadística unitaria 4-variante; $x = \{x_i\}_{i=1}^{n=224}$, con ***n*** la numerosidad del colectivo.

Se eliminan todos los valores sospechosos detectados y se vuelve al primer paso para analizar los datos restantes, hasta que no se detecten más datos atípicos, o se haya considerado hasta el 40 % de los datos.

Realizando el procedimiento correspondiente de identificación de datos atípicos sospechosos se llevó a cabo **dos corridas** en las cuales obtuvimos 129 datos identificados como sospechosos.

Tabla 17-3: Datos Sospechosos

X_i	volumen	Capacidad	Caudal	V_despachado	X_i	volumen	Capacidad	Caudal	V_despachado
X_1	2483815	96000	4000	2483815	X_{66}	1348390	96000	4000	1348390
X_2	226203	30800	1283	136422	X_{67}	6467634	96000	4000	6467634
X_3	0	30800	1283	0	X_{68}	1523047	108000	4500	1523047
X_4	40982	30800	1283	40982	X_{69}	4886440	108000	4500	4886440
X_5	181741	30800	1283	192043	X_{70}	6868361	96000	4000	6868361
X_6	1049435	14400	600	490606	X_{71}	5395678	108000	4500	5395678
X_7	690277	14400	600	0	X_{72}	3362487	48000	2000	2029493
X_8	2642321	24000	1000	2642321	X_{73}	2892702	48000	1670	1788109
X_9	3025977	21600	900	3025977	X_{74}	2710265	48000	1670	1602609
X_{10}	2666602	24000	1000	2666602	X_{75}	2588384	48000	2000	1517763
X_{11}	2720019	21600	900	2720019	X_{76}	486093	46500	1800	257470
X_{12}	964528	10800	450	244021	X_{77}	1773292	46500	1800	1083769
X_{13}	903018	9600	400	253935	X_{78}	900	46500	1800	900
X_{14}	791291	9600	400	264292	X_{79}	264236	46500	1800	379596
X_{15}	678976	10800	450	239320	X_{80}	4056491	46500	1800	1894468
X_{16}	201201	11700	488	201201	X_{81}	4132007	14400	600	1852114
X_{17}	145547	11700	488	145547	X_{82}	5060871	14400	600	2304192
X_{18}	1788402	11700	488	1753278	X_{83}	3834950	30800	1283	1868425
X_{19}	2367436	14040	585	1846922	X_{84}	1027667	46500	1800	432649
X_{20}	2103233	11700	488	1453688	X_{85}	1212368	30800	1283	0
X_{21}	2711310	11700	488	2711310	X_{86}	1818073	46500	1800	160736
X_{22}	2765522	14040	585	2765522	X_{87}	3689971	14400	600	96669
X_{23}	3033992	14040	585	3033992	X_{88}	4527299	14400	600	346037
X_{24}	2534205	11700	488	2534205	X_{89}	3243207	30800	1283	353850
X_{25}	277875	11700	488	277875	X_{90}	9733098	81000	3375	8878125
X_{26}	0	11700	488	0	X_{91}	10502995	70000	2768	9548916
X_{27}	0	11700	488	0	X_{92}	11671653	70000	2768	11074193
X_{28}	104325	11700	488	104325	X_{93}	11307471	81000	3375	10410837
X_{29}	6485	11700	488	6485	X_{94}	1983998	60504	2521	263085
X_{30}	4697	11700	488	4697	X_{95}	2106205	74400	3100	536077
X_{31}	1911759	60504	2521	1911759	X_{96}	1906449	74400	3100	528355
X_{32}	1987128	74400	3100	1987128	X_{97}	1582505	60504	2521	299099
X_{33}	2363057	74400	3100	2363057	X_{98}	9083464	60504	2521	2449747
X_{34}	1301473	60504	2521	1301473	X_{99}	9669817	74400	3100	2540238
X_{35}	1615870	60504	2521	1560479	X_{100}	10280219	74400	3100	2652507

X_{36}	1507582	52800	2200	1459441	X_{101}	8676753	60504	2521	2396700
X_{37}	1368049	52800	2200	1360961	X_{102}	3381960	60504	2521	2080509
X_{38}	1429029	60504	2521	1429029	X_{103}	4934824	74400	3100	2246889
X_{39}	1490613	52800	2200	1490613	X_{104}	6447334	74400	3100	2347497
X_{40}	1492956	52800	2200	1492956	X_{105}	4810308	60504	2521	2102954
X_{41}	0	108000	4500	5776034	X_{106}	7716293	60504	2521	272753
X_{42}	0	96000	4000	6247186	X_{107}	6825060	74400	3100	313850
X_{43}	0	96000	4000	6730498	X_{108}	5579855	74400	3100	144200
X_{44}	109645	108000	4500	5329381	X_{109}	4636894	60504	2521	95999
X_{45}	23470	108000	4500	38426	X_{110}	6624018	60504	2521	4964787
X_{46}	21755	96000	4000	36230	X_{111}	6994421	52800	2200	5232680
X_{47}	18371	96000	4000	32293	X_{112}	7407848	52800	2200	5181910
X_{48}	7311	108000	4500	18961	X_{113}	5963168	46752	1948	4003192
X_{49}	1770970	108000	4500	3717139	X_{114}	1295558	60504	2521	6096640
X_{50}	1487176	96000	4000	3841092	X_{115}	2733715	52800	2200	6559105
X_{51}	1018388	96000	4000	3490169	X_{116}	4183620	52800	2200	6852991
X_{52}	1626430	108000	4500	3350457	X_{117}	2700191	46752	1948	5132485
X_{53}	6739002	108000	4500	6739002	X_{118}	7055659	60504	2521	1538312
X_{54}	8340007	96000	4000	8340007	X_{119}	4690232	52800	2200	1433600
X_{55}	7544387	96000	4000	7544387	X_{120}	5433903	52800	2200	1252555
X_{56}	5644878	108000	4500	5644878	X_{123}	7055659	60504	2521	1538312
X_{57}	858249	108000	4500	1410869	.	.	.	.	.
X_{58}	355086	96000	4000	1419301	.	.	.	.	.
X_{64}	739546	108000	4500	5329381					
X_{65}	2467602	108000	4500	2467602	X_{129}	375795	36480	1600	375795

Fuente: (EP Petroecuador, 2020).

Realizado por: Bastidas, Johanna, 2022.

Seguidamente se calcula la distancia de Mahalanobis con la matriz de datos sospechosos.

$$d_M^2(x_i, \bar{x}_R) > p + 3\sqrt{2p} \qquad \underline{x}_i \text{ sospechoso es atípico}$$

$$d_M^2(x_i, \bar{x}_R) \leq p + 3\sqrt{2p} \qquad \underline{x}_i \text{ sospechoso no es atípico}$$

(4-3)

$$p = 4$$

$$p + 3\sqrt{2p}$$

$$5 + 3\sqrt{2(4)} = 12.49$$

Distancias

$$d_M^2 = (x_i - \bar{x}_R)^t S_R^{-1}(x_i - \bar{x}_R) \quad (5\text{-}3)$$

Tabla 18-3: Criterio de decisión para datos sospechosos atípicos

Distancia^2	$p + 3\sqrt{2p}$	Decisión	Distancia^2	$p + 3\sqrt{2p}$	Decisión
95,553	12,49	atípico	98,402	12,49	atípico
0,006	12,49	sospechoso no atípico	140,845	12,49	atípico
0,002	12,49	sospechoso no atípico	161,494	12,49	atípico
0,003	12,49	sospechoso no atípico	182,986	12,49	atípico
0,006	12,49	sospechoso no atípico	145,054	12,49	atípico
280,194	12,49	atípico	194,796	12,49	atípico
426,452	12,49	atípico	1633,384	12,49	atípico
10,485	12,49	sospechoso no atípico	9,029	12,49	sospechoso no atípico
14,466	12,49	atípico	10,944	12,49	sospechoso no atípico
10,682	12,49	sospechoso no atípico	1053,484	12,49	atípico
11,215	12,49	sospechoso no atípico	10,413	12,49	sospechoso no atípico
467,854	12,49	atípico	0,364	12,49	sospechoso no atípico
379,813	12,49	atípico	1,312	12,49	sospechoso no atípico
249,691	12,49	atípico	8,392	12,49	sospechoso no atípico
172,601	12,49	atípico	0,048	12,49	sospechoso no atípico
0,015	12,49	sospechoso no atípico	4770,981	12,49	atípico
0,010	12,49	sospechoso no atípico	6987,037	12,49	atípico
0,001	12,49	sospechoso no atípico	0,000	12,49	sospechoso no atípico
254,100	12,49	atípico	9,104	12,49	sospechoso no atípico
0,001	12,49	sospechoso no atípico	0,002	12,49	sospechoso no atípico
0,000	12,49	sospechoso no atípico	1,340	12,49	sospechoso no atípico
12,273	12,49	sospechoso no atípico	0,000	12,49	sospechoso no atípico
15,335	12,49	atípico	0,000	12,49	sospechoso no atípico
0,000	12,49	sospechoso no atípico	0,000	12,49	sospechoso no atípico
0,026	12,49	sospechoso no atípico	908,558	12,49	atípico
0,005	12,49	sospechoso no atípico	1,396	12,49	sospechoso no atípico
0,005	12,49	sospechoso no atípico	1,073	12,49	sospechoso no atípico
0,008	12,49	sospechoso no atípico	1053,820	12,49	atípico
0,006	12,49	sospechoso no atípico	2679,451	12,49	atípico
0,006	12,49	sospechoso no atípico	2256,912	12,49	atípico
30,145	12,49	atípico	1747,865	12,49	atípico
58,456	12,49	atípico	1501,737	12,49	atípico
55,669	12,49	atípico	40240,948	12,49	atípico
27,740	12,49	atípico	46458,386	12,49	atípico
33,042	12,49	atípico	53197,955	12,49	atípico
23,476	12,49	atípico	36057,379	12,49	atípico
16,244	12,49	atípico	1560,781	12,49	atípico
34,716	12,49	atípico	6614,244	12,49	atípico

22,205	12,49	atípico	15360,703	12,49	atípico
22,231	12,49	atípico	6699,140	12,49	atípico
0,000	12,49	sospechoso no atípico	50510,956	12,49	atípico
35831,214	12,49	atípico	38610,304	12,49	atípico
41550,617	12,49	atípico	26876,826	12,49	atípico
25180,883	12,49	atípico	18744,990	12,49	atípico
159,280	12,49	atípico	2606,344	12,49	atípico
92,889	12,49	atípico	2937,400	12,49	atípico
92,810	12,49	atípico	4636,176	12,49	atípico
158,811	12,49	atípico	3571,357	12,49	atípico
3682,864	12,49	atípico	21085,430	12,49	atípico
5213,007	12,49	atípico	13378,590	12,49	atípico
5745,075	12,49	atípico	6539,923	12,49	atípico
2946,013	12,49	atípico	5428,251	12,49	atípico
214,118	12,49	atípico	27781,159	12,49	atípico
188,941	12,49	atípico	9667,782	12,49	atípico
182,489	12,49	atípico	15942,187	12,49	atípico
174,640	12,49	atípico	11923,212	12,49	atípico
444,916	12,49	atípico	2,531	12,49	sospechoso no atípico
1153,177	12,49	atípico	7,044	12,49	sospechoso no atípico
2116,084	12,49	atípico	10,018	12,49	sospechoso no atípico
434,015	12,49	atípico	0,038	12,49	sospechoso no atípico
11727,966	12,49	atípico	10,053	12,49	sospechoso no atípico
21520,483	12,49	atípico	4,759	12,49	sospechoso no atípico
30677,132	12,49	atípico	0,104	12,49	sospechoso no atípico
19508,935	12,49	atípico	0,002	12,49	sospechoso no atípico
160,653	12,49	atípico			

Realizado por: Bastidas, Johanna, 2022.

Aplicando el criterio de decisión de si los datos sospechosos son atípicos, los mismos se retiran de la matriz original, mientras que los individuos considerados como datos sospechosos no atípicos regresan o se mantienen en la matriz original para seguir con el estudio.

Luego de ejecutar el algoritmo que permite la identificación de grupos atípicos se llevó a cabo 2 corridas hasta obtener 0 datos sospechosos, obteniendo así un total de 129 datos sospechosos en las 2 corridas, y luego de hallar la distancia de Mahalanobis de la matriz de datos identificados como sospechosos y haber aplicado el criterio respecto a $p + 3\sqrt{2p}$, que en este caso fue de 12.485 se concluye que 42 sospechosos identificados son sospechosos no atípicos ya que su distancia calculada correspondiente es menor a 12.485 por lo que se incorporan a la matriz original y teniendo así un total de 87 datos atípicos. Finalmente obteniéndose una matriz libre de datos atípicos 8-variante con componentes cuantitativas y cualitativas con d.e.u. $x = \{x_i\}_{i=1}^{137}$, teniendo en cuenta que, con una numerosidad de 137, se evidencia que se ha retirado

aproximadamente el 40% de datos de la matriz original, por lo que con la nueva matriz libre de atípicos continuamos con el estudio.

Ahora, que se tiene una matriz de datos libre de atípicos que podrían haber estado distorsionando los resultados, el comportamiento de los datos aun podrían mejorar, una opción es una transformación logarítmica el cual facilita el análisis de un conjunto de datos Multivariantes sea más simple pues su distribución se hace más simétrica y las relaciones entre variables son lineales. Por eso la transformación logarítmica es útil para transformar distribuciones con sesgo positivo. Sin embargo, se aplica para variables cuantitativas positivas, lo que no sucede con los datos en estudio, por lo que no se aplica.

3.5. Modelo 2: Regresión lineal múltiple, nueva matriz

Se genera la ecuación de Regresión Lineal múltiple de VOLUMEN sobre Xi, con la matriz de datos libres de atípicos.

$$\begin{aligned} \widehat{Volumen} = \hat{\beta}_0 + \hat{\beta}_1 polid. + \hat{\beta}_2 tramo + \hat{\beta}_3 prod. + \hat{\beta}_4 \text{año} \\ + \hat{\beta}_5 Capacidad + \hat{\beta}_6 Caudal \\ + \hat{\beta}_7 Volumen\ despachado \end{aligned} \tag{6-3}$$

Obteniendo mediante el software R, los siguientes coeficientes:

Tabla 19-3: Modelo 2: Regresión lineal múltiple

	Modelo de regresión lineal múltiple 2							
	Intercept	**poliducto**	**tramo**	**productos**	**año**	**Capacidad**	**Caudal**	**vdespachado**
Coef	-74970000000	334900000	-268000000	6377000	37220000	67530	-1321000	796,60

Realizado por: Bastidas, Johanna, 2022.

Obteniendo un modelo:

$$\begin{aligned} \widehat{Volumen} &= \hat{\beta}_0 + \hat{\beta}_1 poliducto + \hat{\beta}_2 tramo + \hat{\beta}_3 productos + \hat{\beta}_4 \text{año} \\ &\quad + \hat{\beta}_5 Capacidad + \hat{\beta}_6 Caudal + \hat{\beta}_7 Vdespachado \\ \widehat{Volumen} &= -74970000000 + 334900000\ poliducto - 268000000\ tramo + \\ &\quad 6377000\ productos + 37220000\ \text{año} + 67530\ Capacidad - \\ &\quad 1321000\ Caudal + 796.60\ Vd.espachado. \end{aligned} \tag{7-3}$$

Evaluar el rendimiento del modelo 2:

Tabla 20-3: Coeficiente de determinación, modelo 2

Coeficiente de determinación	0,635

Realizado por: Bastidas, Johanna, 2022.

Con el coeficiente de determinación del modelo 2, indica que el 63.5% de la variabilidad de los datos es explicada por el modelo, a priori se puede considerar como un ajuste "bueno". Sin embargo, este resultado es preliminar, ya que el objetivo es obtener un buen ajuste.

Supuestos: Modelo 2

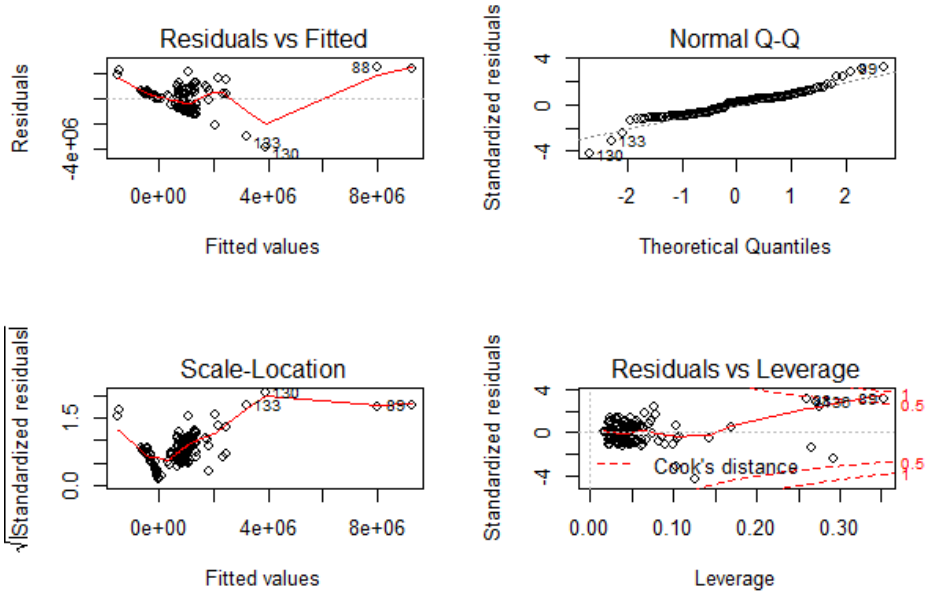


Gráfico 6-3: Supuestos modelo 2

Realizado por: Bastidas, Johanna, 2022.

En el Gráfico 6-3, se evidencia a priori que no cumple la normalidad, ni homocedastidad, y posiblemente cumple la independencia. Sin embargo, se recurre a test teóricos, para probar los resultados obtenidos visualmente.

Tabla 21-3: Valores p de los supuestos para el modelo 2

Valores P		
Normalidad	**Homocedasticidad**	**Independencia**
1,62E-05	0,002112	3,292E-12

Realizado por: Bastidas, Johanna, 2022.

En la Tabla 21-3 podemos observar que con los valores obtenidos luego de haber aplicado la prueba teórica Jarque bera se obtiene un p valor de 1.62E-02 el cual es menor al nivel de significancia considerado 0.05, por lo que se rechaza H0 y se concluye que los residuos no siguen una distribución normal. Además, para probar Homocedasticidad se recurre al test de Goldfeld-Quant con el cual se obtiene un p valor de 0.002112, el cual es menor que el nivel de significancia por lo que se rechaza H0 y podemos concluir que los residuos no cumplen con el supuesto de

homocedastidad. También, para probar la Independencia se recurre al test de Durbin Watson con el que se obtiene un valor aproximado de 3,292E-12 el cual no es mayor a 0,05 y se concluye que no cumple con el supuesto de independencia. Por lo tanto, el modelo 2 ajustado cumple no cumple con ninguno de los supuestos, por lo que se debe ajustar un nuevo modelo.

3.6. Modelo 2_1: modelo 2 mejorado

Una forma de poder mejorar el modelo es aumentando interacciones en el modelo.

Tabla 22-3: Coeficientes del modelo 2 mejorado

Coeficientes del Modelo 2 mejorado			
(Intercept)	**poliducto**	**tramo**	**productos**
71800000000,0	2563000000,0	-1374000000,0	150000000,0
Capacidad	**vdespachado**	**Caudal**	**año**
-5308000,0	-11820,0	-2680000,0	-34640000,0
poliducto: tramo	**poliducto: productos**	**poliducto: Capacidad**	**poliducto: vdespachado**
226700000,0	-457900000,0	-1231000,0	9192,0
tramo: productos	**tramo: Caudal**	**productos: Capacidad**	**productos: Caudal**
169600000,0	-13410000,0	-51000,0	889200,0
productos: vdespachado	**Capacidad: año**	**Capacidad: Caudal**	**tramo: Capacidad**
805,4	2655,0	116,5	589100,0
tramo: vdespachado	**poliducto: Caudal**	**vdespachado: Caudal**	**Capacidad: vdespachado**
-252,8	23440000,0	13,9	-0,5
poliducto: tramo: productos	**poliducto: tramo: Capacidad**	**poliducto: tramo: vdespachado**	**poliducto: vdespachado: Caudal**
-11140000,0	4163,0	-829,3	1,8
tramo: productos: Capacidad	**productos: Capacidad: vdespachado**	**productos: vdespachado: Caudal**	**poliducto: productos: vdespachado**
-7683,0	0,1	-1,9	-644,5
tramo: productos: vdespachado	**poliducto: productos: Caudal**	**poliducto: tramo: productos: vdespachado**	**poliducto: productos: vdespachado: Caudal**
53,2	763100,0	55,6	-0,2

Realizado por: Bastidas, Johanna, 2022.

$$\begin{aligned}\widehat{Volumen} = \hat{\beta}_0 &+ \hat{\beta}_1 poliducto + \hat{\beta}_2 tramo + \hat{\beta}_3 productos \\ &+ \hat{\beta}_4 Capacidad + \hat{\beta}_5 Vdespachado + \hat{\beta}_6 Caudal \\ &+ \hat{\beta}_7 \text{año} + \hat{\beta}_8 poliducto * tramo + \cdots + \hat{\beta}_{35} poliducto \\ &* productos * vdespachado * Caudal\end{aligned} \tag{8-3}$$

$$\widehat{Volumen} = 71800000000 + 2563000000\ poliducto - 1374000000\ tramo + 150000000\ productos - 5308000\ Capacidad - 11820\ Vdespachado - 2680000\ Caudal - 34640000\ año + 226700000\ poliducto * tramo + \cdots - 0.2\ poliducto * productos * vdespachado * Caudal$$

Evaluar el rendimiento del modelo 2 mejorado:

Tabla 23-3: Coeficiente de determinación, modelo 2 mejorado

Coeficiente de determinación	0,9134

Realizado por: Bastidas, Johanna, 2022.

El coeficiente de determinación del modelo 2 mejorado indica que el 91.34% de la variabilidad de los datos de la matriz sin datos atípicos es explicada por el modelo, lo cual es muy bueno. Ahora, es pertinente realizar un análisis de residuos, para poder considerar este un modelo adecuado.

3.7. Selección de los mejores predictores

Para la selección de las mejores variables predictoras se usa la técnica de pasos sucesivos (stepwise), el cual es una combinación de las estrategias: Eliminación de variables hacia atrás (Backward Stepwise Regression) y Selección de variables predictoras hacia adelante (Fordward Stepwise Regression). Para determinar la calidad del modelo ajustado se toma en cuenta el valor de AIC (Akaike).

Tabla 24-3: Selección escalonada de variables predictoras (Stepwsise)

Variable respuesta:	VOLUMEN
Start	**Variables predictoras**
AIC=3642.52	poliducto + tramo + productos + Capacidad + vdespachado + poliducto * tramo + poliducto * productos + poliducto * Capacidad + poliducto * vdespachado + tramo * productos + tramo * Caudal + productos * Capacidad + productos * Caudal + productos * vdespachado + año * Capacidad + Capacidad * Caudal + poliducto * tramo * productos + poliducto * tramo * Capacidad + poliducto * tramo * vdespachado + poliducto * Caudal * vdespachado + tramo * productos * Capacidad + productos * Capacidad * vdespachado + productos * Caudal * vdespachado + poliducto * tramo * productos * vdespachado + poliducto * productos * Caudal * vdespachado
AIC=3640.62	poliducto + tramo + productos + Capacidad + vdespachado + Caudal + año + poliducto*tramo + poliducto*productos + poliducto*Capacidad + poliducto*vdespachado + tramo*productos + tramo*Caudal + productos*Capacidad + productos*Caudal +

	productos*vdespachado + Capacidad*año + Capacidad*Caudal + tramo*Capacidad + tramo*vdespachado + poliducto*Caudal + vdespachado*Caudal + Capacidad*vdespachado + poliducto*tramo*productos + poliducto*tramo*vdespachado + poliducto*vdespachado*Caudal + tramo*productos*Capacidad + productos*Capacidad*vdespachado + productos*vdespachado*Caudal + poliducto*productos*vdespachado + tramo*productos*vdespachado + poliducto*productos*Caudal + poliducto*tramo*productos*vdespachado + poliducto*productos*vdespachado*Caudal
AIC=3639.16	poliducto + tramo + productos + Capacidad + vdespachado + Caudal + año + poliducto*tramo + poliducto*productos + poliducto*Capacidad + poliducto*vdespachado + tramo*productos + tramo*Caudal + productos*Capacidad + productos*Caudal + productos*vdespachado + Capacidad*Caudal + tramo*Capacidad + tramo*vdespachado + poliducto*Caudal + vdespachado*Caudal + Capacidad*vdespachado + poliducto*tramo*productos + poliducto*tramo*vdespachado + poliducto*vdespachado*Caudal + tramo*productos*Capacidad + productos*Capacidad*vdespachado + productos*vdespachado*Caudal + poliducto*productos*vdespachado + tramo*productos*vdespachado + poliducto*productos*Caudal + poliducto*tramo*productos*vdespachado + poliducto*productos*vdespachado*Caudal

Realizado por: Bastidas, Johanna, 2022.

Después de haber realizado la selección de variables predictoras mediante la técnica de Stepwise o selección escalonada, se obtiene un modelo final ajustado con menos variables que el propuesto al inicio, y que además presenta un AIC reducido de 3639.16; es decir se va mejorando el modelo. El mejor modelo 2 ajustado resultante, dada la selección de variables predictoras significativas es:

Tabla 25-3: Coeficientes del mejor modelo 2 con variables predictoras significativas

Mejor modelo 2 ajustado con variables predictoras significativas			
(Intercept)	**poliducto**	**tramo**	**productos**
-16900000000,00	1700000000,00	-1269000000,00	120400000,00
Capacidad	**vdespachado**	**Caudal**	**año**
57370,00	-11920,00	-4402000,00	9688000,00
poliducto*tramo	**poliducto*productos**	**poliducto* Capacidad**	**poliducto*vdespachado**
269900000,00	-429900000,00	-1381000,00	9308,00
tramo*productos	**tramo*Caudal**	**productos* Capacidad**	**productos*Caudal**
166600000,00	-15640000,00	-52590,00	956100,00
productos*vdespachado	**Capacidad*Caudal**	**tramo*Capacidad**	**tramo*vdespachado**
885,80	119,70	675600,00	-275,30
poliducto*Caudal	**vdespachado*Caudal**	**Capacidad* vdespachado**	**poliducto*tramo* productos**
28740000,00	14,04	-0,47	-11870000,00

poliducto*tramo* vdespachado	**poliducto*vdespachado* Caudal**	**tramo*productos* Capacidad**	**productos*Capacidad* vdespachado**
-831,00	1,73	-7645,00	0,08
productos*vdespachado* Caudal	**poliducto*productos* vdespachado**	**tramo*productos* vdespachado**	**poliducto*productos* Caudal**
-1,98	-703,40	53,94	737600,00
poliducto*tramo*productos *vdespachado	**poliducto*productos* vdespachado*Caudal**		
59,89	-0,21		

Realizado por: Bastidas, Johanna, 2022.

$$\widehat{Volumen} = \hat{\beta}_0 + \hat{\beta}_1 poliducto + \hat{\beta}_2 tramo + \hat{\beta}_3 productos + \hat{\beta}_4 Capacidad + \hat{\beta}_5 Vdespachado + \hat{\beta}_6 Caudal + \hat{\beta}_7 año + \hat{\beta}_8 poliducto * tramo + \cdots + \hat{\beta}_{33} poliducto * productos * vdespachado * Caudal$$

$$\widehat{Volumen} = -16900000000 + 1700000000\ poliducto - 1269000000\ tramo + 120400000\ productos + 57370\ Capacidad - 11920 Vdespachado - 4402000 Caudal + 9688000 año + 269900000\ poliducto * tramo + \cdots - 0.21\ poliducto * productos * vdespachado * Caudal \quad (9\text{-}3)$$

Intervalos de confianza

Intervalos de confianza de los coeficientes del último modelo estimado

Tabla 26-3: Intervalos de confianza de los coeficientes del modelo ajustado

Intervalos de confianza al 95%		
	2,50%	**97,50%**
(Intercept)	-195725800000000,0	161918600000000,0
poliducto	-2481966000000,0	5882704000000,0
tramo	-2404740000000,0	-132726600000,0
productos	-465088000000,0	705828500000,0
Capacidad	-431910100,0	546643800,0
vdespachado	-18067560,0	-5770158,0
Caudal	-16939380000,0	8135466000,0
anio	-78955370000,0	98332000000,0
poliducto*tramo	-17027430000,0	556763700000,0

poliducto*productos	-954998700000,0	95162480000,0
poliducto*Capacidad	-2156876000,0	-604647900,0
poliducto*vdespachado	5356449,0	13258940,0
tramo*productos	56206560000,0	276914700000,0
tramo*Caudal	-24449980000,0	-6836823000,0
productos*Capacidad	-105365700,0	185136,5
productos*Caudal	-569733600,0	2482030000,0
productos*vdespachado	367536,2	1404122,0
Capacidad*Caudal	54336,4	185121,5
tramo*Capacidad	326443200,0	1024671000,0
tramo*vdespachado	-988916,7	438325,5
poliducto*Caudal	9250571000,0	48220430000,0
vdespachado*Caudal	6825,7	21247,5
Capacidad*vdespachado	-770,7	-164,9
poliducto*tramo*productos	-46344080000,0	22601840000,0
poliducto*tramo*vdespachado	-1240010,0	-421959,3
poliducto*vdespachado*Caudal	425,8	3023,8
tramo*productos*Capacidad	-14081330,0	-1208012,0
productos*Capacidad*vdespachado	27,7	127,8
productos*vdespachado*Caudal	-3179,0	-775,9
poliducto*productos*vdespachado	-1055800,0	-351040,7
tramo*productos*vdespachado	-24564,7	132440,5
poliducto*productos*Caudal	20-1088100,0	1274181000,0
poliducto*tramo*productos*vdespachado	25707,4	94077,4
poliducto*productos*vdespachado*Caudal	-386,2	-36,3

Realizado por: Bastidas, Johanna, 2022.

Cada pendiente del modelo ajustado, es decir los coeficientes parciales de regresión de las variables predictores se define como: asumiendo que las variables predictoras permanecen constantes, por cada unidad que aumente la variable predictora considerada, la variable respuesta Y varía positivamente o negativamente como indique el coeficiente del valor estimado de la variable predictora. En la Tabla 26-3, muestran al 95 % de confianza el intervalo en el que la pendiente o el coeficiente del modelo ajustado puede variar.

Supuestos: Modelo 2 mejorado con variables predictoras significativas

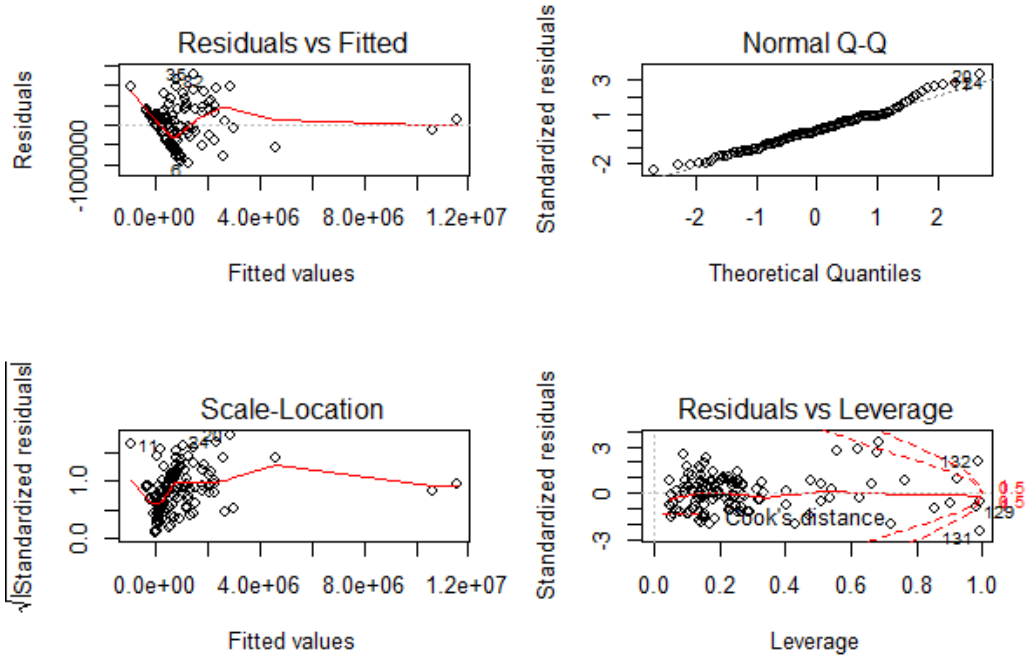


Gráfico 7-3: Supuestos modelo 2 mejorado

Realizado por: Bastidas, Johanna, 2022.

En la Gráfico 7-3, se puede apreciar visualmente como resultados preliminares que los residuos del modelo 2 mejorado siguen una distribución normal, en cuanto a la homocedasticidad también se cumple, y con relación al supuesto de independencia, es parece cumplirse también, sin embargo, se opta por test teóricos para corroborar lo dicho.

Tabla 27-3: Valores p de los supuestos del modelo 2 mejorado

Valores P		
Normalidad	**Homocedasticidad**	**Independencia**
0.5548	0.9994	3.443e-08

Realizado por: Bastidas, Johanna, 2022.

En la Tabla 27-3, se observa con la prueba teórica Jarque bera se obtiene un p valor de 0.5548 el cual es mayor al nivel de significancia, considerado 0.05, por lo que no se rechaza H0 y se concluye que los residuos del modelo 2 mejorado, siguen una distribución normal. Además, para probar Homocedasticidad en los residuos, se recurre al test de Goldfeld-Quant con el cual se obtiene un p valor de 0.9994, siendo este un valor mayor al nivel de significancia de 0.05, por lo que no se rechaza H0 y se puede concluir que los residuos del modelo, cumplen con el supuesto de homocedastidad. También, para probar la Independencia se opta al test de Durbin Watson con el que se obtiene un valor aproximado de 3.443e-08, que es menor a 0,05 significativamente, por

lo que se rechaza H0 y se concluye que no cumple con el supuesto de independencia. Por lo tanto, el modelo 2 mejorado ajustado cumple con los supuestos de normalidad y homocedastidad, pero no independencia, por lo que el modelo propuesto podría ser adecuado.

3.8. Modelo 3: Modelo lineal generalizado

Dada la variable respuesta "Volumen" correspondiente al número de barriles por hora que se transportan mediante poliductos, se considera como un variable de conteo o recuento por lo que es pertinente aplicar un modelo lineal generalizado, específicamente un modelo de regresión de Poisson.

Tabla 28-3: Coeficientes del modelo 3 GLM

Coeficientes del modelo 3 ajustado GLM							
(Intercept)	**poliducto**	**tramo**	**productos**	**año**	**Capacidad**	**Caudal**	**vdespachado**
6293,00	468,40	-233,60	16,79	-3,56	0,05	-1,08	-0,00002

Realizado por: Bastidas, Johanna, 2022.

$$\log(\mu_i) = \log(n_i) + x_i^T \beta$$
$$\log(\text{Volumen}) = \log(\text{total}) + 6293 + 468.4\ \text{poliducto} - 233.6\ \text{tramo} + 16.79\ \text{productos} - 3.56\ \text{año} + 0.05\ \text{Capacidad} - 1.08\ \text{Caudal} - 0.00002\ \text{Vdespachado} \quad (10\text{-}3)$$

El modelo 3 ajustado presenta un AIC de 74277254.113, por lo que se tratará de mejorar el modelo.

Intervalos de confianza al 95%, para los coeficientes del modelo 3 ajustado

Tabla 29-3: Intervalos de confianza del modelo 3 al 95%

Intervalo de confianza para el modelo 3			
	exp(coef)	**2.5%**	**97.5%**
(Intercept)	5.406.088.336	4.531.634.569	6.449.283.428
poliducto	15.974.398	15.969.844	15.978.953
tramo	0.7916491	0.7915238	0.7917744
productos	10.169.356	10.169.145	10.169.567
año	0.9964476	0.9963605	0.9965347
Capacidad	10.000.483	10.000.483	10.000.483
Caudal	0.9989224	0.9989214	0.9989235
vdespachado	10.000.000	10.000.000	10.000.000

Realizado por: Bastidas, Johanna, 2022.

Lo cual indica que para una unidad de caudal que se incremente el volumen aumentará en 10000000 unidades, teniendo en cuenta las demás variables predictoras fijas, así puede ir variando la interpretación con cada coeficiente estimado.

Evaluación del modelo 3

Sobredispersión

La distribución de Poisson es caracterizada por su esperanza y su varianza coinciden; por lo que cuando se ajusta un modelo de Poisson o de recuento a un conjunto de datos, este puede presentar valores que difieran significativamente.

Tabla 30-3: Test de sobredispersión del modelo 3

test de sobredispersión	
p valor	2,20E-16
dispersión	285934,3

Realizado por: Bastidas, Johanna, 2022.

Se observa que con la prueba de sobredispersión basada en la prueba de Wald, se arroja un p valor de 2.20E-16 que es menor al nivel de significancia de 0.05, por lo que se rechaza H0, y se concluye al 95% de confianza que hay evidencia de sobredispersión con un estimado de 285934.3.

Estadística chi-cuadrado de Pearson

La estadística se usa como una medida de bondad de ajuste y para el modelo 3 propuesto nos da un p valor de aproximadamente 0, lo cual indica que el modelo ajustado no es adecuado y corrobora los resultados anteriores.
Por lo que el modelo lineal generalizado ajustado presenta sobredispersión, con un AIC de 74277254.113.

3.9. Escoger el mejor modelo

Para la selección del mejor modelo:

Comparación de modelos basada en suma de cuadrados

Para ver la significancia de las variables predictoras en el modelo, resulto que al comparar el modelo 2 mejorado y el modelo 3.

$$H_0: el\ mejor\ modelo\ 2$$

$$H_1: modelo\ 3$$

Tabla 31-3: Estadístico F para escoger el mejor modelo

	Estadístico F
p valor	0.99

Realizado por: Bastidas, Johanna, 2022.

Al obtener un p valor significativamente mayor al nivel de significancia de 0.05, no se rechaza H0, y se concluye al 95% de confianza que el mejor modelo 2 es el más adecuado.

El estadístico AIC

Dado que tenemos calculados lo AIC tanto para el mejor modelo 2 y el modelo 3 ajustados se los compara y se tiene que:

Tabla 32-3: Criterio AIC para escoger el mejor modelo

	AIC
Modelo 2	6.346.167
Modelo 3	74.277.254.113

Realizado por: Bastidas, Johanna, 2022.

Se visualiza que el modelo 2 presenta un AIC significativamente menor al del modelo 3, por lo que el modelo más adecuado es el modelo 2.

Predicciones

Tabla 33-3: Predicciones dadas por el modelo 2

poliducto	**tramo**	**productos**	**Capacidad**	**V_despachado**	**Caudal**	**año**	**$\widehat{\text{VOLUMEN}}$**
1	1	9	108000	5800000	4600	2022	7107540.2
1	2	7	49000	1605000	2500	2022	2022387.4
1	3	9	108000	5800000	4600	2022	6485059.0
2	4	9	8500	16000	500	2022	19852.0
2	5	19	25000	300000	1000	2022	323410.1
2	6	15	81000	9000000	4000	2022	9550110.9
3	7	3	10800	25000	500	2022	32767.8

4	8	6	12000	2000000	500	2022	2348928.9
5	10	3	75000	300000	3500	2022	279776.8
5	10	6	55000	1500000	2500	2022	2102251.7
5	11	6	37000	6000000	1700	2022	6101844.6

Realizado por: Bastidas, Johanna, 2022.

Se observa que al pasar los años si la capacidad, el Volumen despachado y el Caudal aumentan pues el volumen que se transportará, también lo hará, pero depende el poliducto el tramo y el producto ya que puede hacerlo significativamente o no como en el poliducto 2 correspondiente al poliducto Libertad Pascuales Manta Monteverde Chorrillo, tomando el tramo 5, Libertad Pascuales, transportando el producto 15, es decir GLP, se observa que el volumen despachado será similar al volumen estimado a transportar.

3.10. Regresión logística nominal

Análisis exploratorio de las variables

Tabla 34-3: Resumen numérico de poliducto vs. tramo

	tramo									
poliducto	1	2	4	5	6	7	8	10	11	Sum
ESM_STO_DMG_QUI_MAC	0	0	0	0	0	0	0	2	8	10
LIB_PASC_MAN	0	0	28	35	2	0	0	0	0	65
QUI_AMB_RIOB	0	0	0	0	0	0	22	0	0	22
SHUSHUF_QUITO	0	0	0	0	0	16	0	0	0	16
TRES_BOCAS_PASC_CUEN	22	2	0	0	0	0	0	0	0	24
Sum	22	2	28	35	2	16	22	2	8	137

Realizado por: Bastidas, Johanna, 2022.

EP Petroecuador tiene 5 poliductos y en total 11 tramos por los cuales se transporta los derivados del petróleo, pero podemos observar que mediante el poliducto Libertas Pascuales Manta transportan más producto por el tramo 4 y 5 durante los 3 años en estudio.

Tabla 35-3: Resumen numérico de poliducto vs. producto

	productos																		
poliducto	1	2	3	4	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum
ESM_STO_DMG_QUI_MAC	0	0	1	0	4	0	0	0	0	0	0	0	0	0	5	0	0	0	10
LIB_PASC_MAN	0	8	8	0	7	0	8	4	0	4	0	0	4	2	4	0	8	8	65
QUI_AMB_RIOB	0	0	0	4	3	0	0	0	3	0	4	4	0	0	4	0	0	0	22
SHUSHUF_QUITO	4	0	4	0	0	0	4	0	0	0	0	0	0	0	0	4	0	0	16
TRES_BOCAS_PASC_CUEN	0	4	4	0	2	2	0	4	0	4	0	0	0	2	0	0	0	2	24

Sum	4	12	17	4	16	2	12	8	3	8	4	4	4	4	13	4	8	10	137

Realizado por: Bastidas, Johanna, 2022.

EP Petroecuador disribuye a nivel nacional mediante poliductos los derivados del petroleo, pero no todos pueden transportar todos los productos, así es como se observa en la Tabla 35-3 que el producto que transportan casi todos los poliductos a excepción del poliducto Quito Ambato Riobamba es Diesel 2, aunque también se observa que se distribuye Diesel Premiun en la mayoría de poliductos de la empresa.

Tabla 36-3: Media y desv. estándar del Volumen respecto al tramo

	Volumen	
tramo	M	SD
1	893408,7	1273041,9
2	2801483,5	129002,4
4	546545,5	669904,1
5	827084,6	881724
6	11087324	826366
7	418568,1	732875,6
8	694767	1027165
10	1067209,5	243549,5
11	339546,4	267196,5

Realizado por: Bastidas, Johanna, 2022.

Se evidencia que durante los años en estudio 2017-2020, en promedio se transportó más barriles mediante el tramo 3 (Tres Bocas Pascuales) aunque una variación notoria de volumen durant el periodo en estudio fue en el tramo 1 (Pascuales Cuenca).

Tabla 37-3: Media y desv. estándar del Volumen respecto al producto

	Volumen	
producto	M	SD
1	4029	1.028.797
2	63553	57.758.966
3	871152,2	916.098.905
4	182381,5	25.717.887
6	1675489,8	1.243.203.007
7	2801483,5	129.002.440
8	883292,1	661.521.290
9	238906,9	621.861.991
10	2670345,7	120.977.048

11	212874,6	217.094.347
12	119489,2	141.582.248
13	73473	66.687.776
14	296070	243.943.009
15	6103670,8	5.774.863.255
16	345517,7	321.769.871
17	5056,5	5.839.498
18	508533,2	566.816.256
19	1261209,9	1.028.230.077

Realizado por: Bastidas, Johanna, 2022.

En promedio se ha transportado aproxiamadamente 6103671 barriles de deivados del producto 15 (GLP), siendo el producto que más se transportó en este periodo, aunque cabe recalcar que también es el que mayor variación de volumen transportado presentó.

3.10.1. Variable respuesta: poliductos

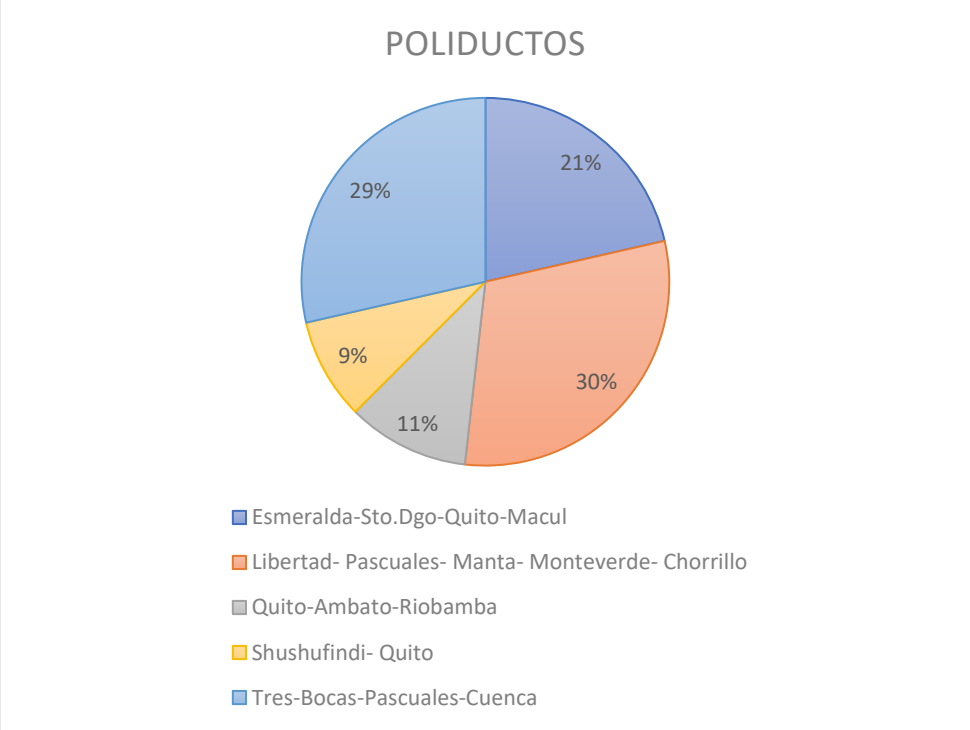


Gráfico 8-3: Diagrama de pastel para la variable poliductos

Realizado por: Bastidas, Johanna, 2022.

Mediante el diagrama de pastel, notemos que los poliductos que transportan más volumen de derivados en un 30% es La libertad-Pascuales-Manta-Monteverde-Chorrillo y un 29% de Tres-Bocas-Pascuales-Cuenca, pero no en todos los poliductos se transporta los mismos productos.

Dada la variable respuesta Poliducto es categórica nominal, se aplica una regresión logística multinomial respecto al producto y volumen como variables predictoras, y tomando en cuenta que se lo realizará con referencia al poliducto Esmeraldas-Santo Domingo-Quito-Macul.

Ajustar el modelo de regresión logística nominal

Tabla 38-3: Coeficientes del modelo de regresión logística nominal

Modelo de regresión logístico nominal			
Coeffic ients:	**(Intercept)**	**productos**	**VOLUMEN**
LIB_PASC_MAN	1,695593	-0,011276662	0,000000437
QUI_AMB_RIOB	0,732615	-0,009471519	0,000000270
SHUSHUF_QUITO	1,46906	-0,105932688	-0,000000117
TRES_BOCAS_PASC_CUEN	1,331082	-0,083739065	0,000000470

Realizado por: Bastidas, Johanna, 2022.

Con los coeficientes que resultan al aplicar el modelo de regresión multinomial, resulta:

$$\frac{\ln\left(P(poliducto = LIB_{PASC_{MAN}})\right)}{\ln\left(P(poliducto = ESM - STO.DGO - QTO - MAC)\right)} = 1.696 - 0.011 productos + 0.000000437\, VOLUMEN$$

$$\frac{\ln\left(P(poliducto = QUI - AMB - RIOB)\right)}{\ln\left(P(poliducto = ESM - STO.DGO - QTO - MAC)\right)} = 0.73 - 0.0095\, productos + 0.00000027\, VOLUMEN \quad (11\text{-}3)$$

$$\frac{\ln\left(P(poliducto = Shushufindi\ Quito)\right)}{\ln\left(P(poliducto = ESM - STO.DGO - QTO - MAC)\right)} = 1.47 - 0.106\, productos - 0.00000012\, VOLUMEN$$

$$\frac{\ln\left(P(poliducto = TRES - BOCAS - PASC - CUEN)\right)}{\ln\left(P(poliducto = ESM - STO.DGO - QTO - MAC)\right)} = 1.33 - 0.084\, productos + 0.00000047\, VOLUMEN$$

Tabla 39-3: Criterio de información AKAIKE, regresión logística

AIC:	394.962

Realizado por: Bastidas, Johanna, 2022.

El criterio de información akaike arroja un valor de 394.96.

Tabla 40-3: Exponencial de los coeficientes del modelo logit

exp(coeficientes del modelo)			
	(Intercept)	productos	VOLUMEN
LIB_PASC_MAN	5,449874	0,9887867	1,0000004
QUI_AMB_RIOB	2,080514	0,9905732	1,0000003
SHUSHUF_QUITO	4,345149	0,8994852	0,9999999
TRES_BOCAS_PASC_CUEN	3,785138	0,9196712	1,0000005

Realizado por: Bastidas, Johanna, 2022.

3.10.2. Interpretaciones del riesgo relativo

- la razón de riesgo relativo para un aumento de una unidad en el volumen es de 1.000004 por estar en el poliducto Libertad-Pascuales-Manta versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el Producto es de 0.989 por estar en el poliducto Libertad-Pascuales-Manta versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el volumen es de 1.000003 por estar en el poliducto Quito-Ambato-Riobamaba versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el producto es de 0.991 por estar en el poliducto Quito-Ambato-Riobamba versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el volumen es de 0.999999 por estar en el poliducto Shushufindi-Quito versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el producto es de 0.8994852 por estar en el poliducto Shushufindi-Quito versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el volumen es de 1.0000005 por estar en el poliducto Tres-Bocas-Pascuales-Cuenca versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.
- la razón de riesgo relativo para un aumento de una unidad en el producto es de 0.9197 por estar en el poliducto Tres-Bocas-Pascuales-Cuenca versus el poliducto Esmeraldas-Santo-Domingo-Quito-Macul.

Predicciones

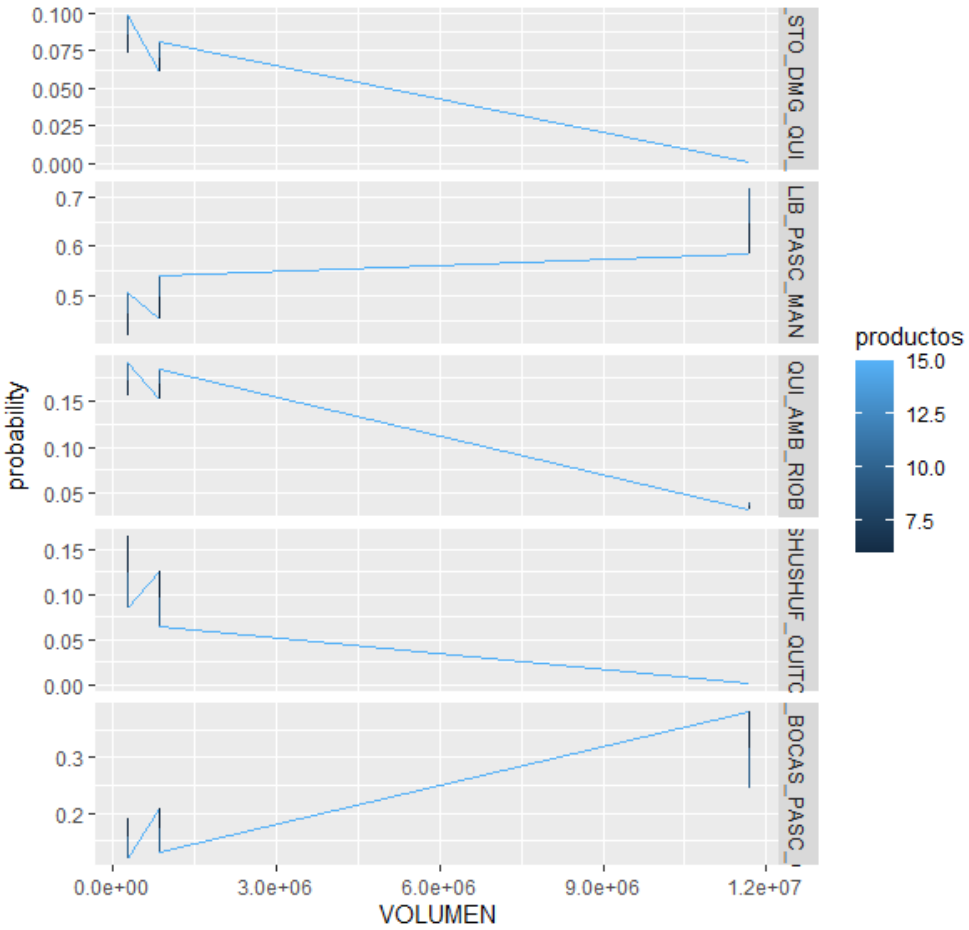


Gráfico 9-3: Gráfica de predicciones del modelo de regresión logística

Realizado por: Bastidas, Johanna, 2022.

Se toma en consideración tras el análisis exploratorio que los productos con más demanda como el Diésel, Diesel Premium y Gasolina Super, y notemos que el Poliducto Santo Domingo Quito Macul, Quito Ambato Riobamba y Shushufindi Quito tienen una alta probabilidad de transportar un mínimo número de barriles de derivados, mientras que los Poliductos Tres Bocas Pascuales Cuenca, mientras más volumen de derivados se desea transportar la probabilidad aumenta, lo que a diferencia del Poliducto Libertad Pascuales Manta la probabilidad de que transporte los productos mencionados se mantiene aproximadamente constante.

Tabla 41-3: Pronósticos dado un volumen constante

		Poliducto				
	Producto	**ESM_STO_DMG_QUI_MAC**	**LIB_PASC_MAN**	**QUI_AMB_RIOB**	**SHUSHUF_QUITO**	**TRES_BOCAS_PASC_CUEN**
1	Destilado 1	0,049	0,389	0,129	0,175	0,258
2	Diesel 1	0,052	0,402	0,133	0,164	0,248
3	Diesel 2	0,054	0,415	0,138	0,154	0,239
4	Diesel Oil	0,056	0,428	0,142	0,145	0,229
5	Diesel P importado	0,059	0,440	0,147	0,135	0,219
6	Diesel Premiunm	0,061	0,452	0,151	0,127	0,209
7	Fuel Oil	0,063	0,464	0,155	0,118	0,200
8	G. Base	0,065	0,475	0,159	0,110	0,190
9	G. EXTRA	0,068	0,486	0,163	0,102	0,181
10	G. EXTRA 85 OCT	0,070	0,496	0,167	0,095	0,172
11	G.Super	0,072	0,506	0,171	0,088	0,163
12	G. Super 90 OCT	0,074	0,516	0,174	0,082	0,154
13	G.Super 92 Oct	0,076	0,525	0,177	0,076	0,146
14	Gas. Importada	0,078	0,533	0,181	0,070	0,138
15	GLP	0,081	0,541	0,184	0,065	0,130
16	Jet A1	0,083	0,548	0,186	0,060	0,123
17	Jet Fuel	0,085	0,556	0,189	0,055	0,116
18	NAO	0,087	0,562	0,192	0,051	0,109
19	Premezcla	0,089	0,568	0,194	0,046	0,102

Realizado por: Bastidas, Johanna, 2022.

La probabilidad de que cada poliducto transporte el promedio del número de barriles en el periodo de estudio 2017-2020, es decir que se transporte aproximadamente de 865085 barriles de derivados por cada poliducto de cada producto se muestra en la tabla expuesta, de lo que se destaca que el Poliducto Libertad Pascuales Monteverde Chorrillo hay más posibilidades de que se transporte tal cantidad de volumen de derivados, y dada la cantidad a transportar tiene baja probabilidad de transportar tal cantidad el Poliducto Esmeralda Santo Domingo Quito Macul.

CONCLUSIONES

- La información proporcionada por EP Petroecuador dado factores externos como el mantenimiento del poliducto, falta de insumos, entre otros hace posible una contaminación en los datos es decir existencia de datos atípicos, por lo que se debe hacer un tratamiento adecuado. Tras realizar un análisis descriptivo de detección de datos atípicos, se efectúa dos corridas del algoritmo en Excel detectando 87 datos atípicos.
- Para ajustar un modelo con variable respuesta cuantitativa se usa en principio un modelo de regresión lineal múltiple, pero para validar el modelo se prueba supuestos de normalidad, homocedastidad e independencia y al no cumplir independencia se ajusta un modelo lineal generalizado, dada el tipo de variable respuesta se aplica el modelo de regresión de Poisson, el cual al ser validado presenta sobredispersión.
- Para mejorar el modelo lineal múltiple se aplica la selección de los mejores predictores siendo así que mediante la técnica stepwise, se reduce el estadístico de criterio de información Akaike 3642.52 a 3639.16, lo cual se logró al reducir algunas iteraciones de las variables predictoras del modelo propuesto.
- Al tener dos modelos tentativos como buenos ajustes se procede a tomar criterios de comparación de modelos para poder escoger el mejor, como el método de significancia de variables con suma de cuadrados o el Criterio de información de Akaike, siendo así el mejor modelo es el modelo lineal.
- Se concluye que mientras el tiempo pase el volumen aumentará si la capacidad, caudal y volumen despachado lo hacen, teniendo en cuenta que la intensidad del aumento dependerá fuertemente del producto y tramo que se considere para cada poliducto.
- Para realizar un ajuste de un modelo con variable cualitativa se hace mediante un análisis de regresión logística multinomial, considerando como variables predictoras a Volumen y Producto, con el cual resulta ser el poliducto con más potencia para transportar derivados el Poliducto libertad Pascuales Monteverde Chorrillo, mientras que el Poliducto Esmeraldas Santo Domingo Quito Macul, a pesar de que transporta varios productos el volumen en que lo hace, es poco.

RECOMENDACIONES

- En principio sería lo ideal trabajar con toda la información para ajustar un modelo, sin embargo, hay que preparar siempre la base de datos antes del análisis a aplicarse, como posible existencia de datos atípicos que puedan contaminar la información.
- En el caso de la variable respuesta cuantitativa Volumen, el modelo que se ajusta y resulta ser el mejor ajuste podría mejorarse con alguna transformación de la variable a predecir, Sin embargo, hay que considerar que las interpretaciones cambiarían por lo que depende de la investigación que se quiera realizar.
- El personal de EP Petroecuador debería implementar varios escenarios que se pueden presentar en la empresa para la posterior toma de decisiones. También, proponer mejoras en cada refinería o poliducto y de esta forma decidir cuál sería el cambio conveniente con los pronósticos a obtener.
- EP Petroecuador debe considerar la repotenciación de las refinerías para que poliductos que pasan por ciudades grandes o llegan a centros de acopio grandes sean abastecidos totalmente, ya que el Poliducto Esmeraldas Santo Domingo Quito Macul, tiene probabilidades bajas de transportar más cantidad de barriles de petróleo de lo que ya maneja.

BIBLIOGRAFÍA

ACOSTA A. *Breve Historia Económica del Ecuador.* Primera. Quito, Ecuador : Corporción, 2006.

ALAN, Agresti. *foundations of linear and generalized linear models.* Canada : John Wiley & Sons, Inc. All rights reserved, 2015.

ANTAKI, George. *Piping and Pipeline Engineering: Design, Construction, Maintenance, Integrity, and Repair.* s.l. : CRC Press, 2003.

APARICIO, Juan. *Modelos lineales aplicados en R.*, Centro de investigación Operativa, 2018, p.229.

BROSA, Jaume. *El diagnóstico de la sobredispersión de la sobredispersión en modelos de análisis de datos de recuento.* Universitat Autónoma de Barcelona, s.l. : 2002.

CARSON, P.A. *Hazardous Chemicals Handbook.* Inglaterra : Elsevier-Butterworth-Heinemann, 2002.

CONGACHA, Jorge. *Estadística Aplicada a la educación con actividades de aprendizaje.* Segunda. Riobamba, Ecuador : editorial académica española, 2016.

CUBILLOS, Adela; & ESTENSSORO, Fernando. *Los desafíos del crecimiento y desarrollo en el contexto del cambio climático.* Santiago de Chile : s.n., 2011, idea, Vol. 2, p.103.

DÍAZ MONROY, Luis; & MORALES RIVERA, Guillermo. *Análisis estadístico de datos categóricos.* Primera. Bogotá, Colombia : Universidad Nacional de Colombia, 2009.

DOBSON, Adrian. *An introduction to generalized linear models.* London, New York : Taylor & Francis Group CRC Press, 2018.

ECHEVERRÍA, Johanna; & JIMÉNEZ, Fátima. 2014. *Estudio técnico económico para mejorar la distribución de productos limpios del poliducto Quito-Ambato.* Escuela Politécnica Nacional, s.l. :2014.

EP Petroecuador. *El petróleo en el Ecuador la nueva era petrolera.* 2013, EP Petroecuador Empresa Pública de Hidrocarburos del Ecuador, Vol. 1, p.145.

EP Petroecuador. *Informe estadístico.* 2018, Jefatura corporativa de planifiación de EP Petroecuador.

EP Petroecuador. *Informe estadístico.* 2019, Jefatura corporativa de planificación de EP Petroecuador.

FIGUEROA, Giulianna. *La fecundidad y su relación con variables socioeconómicas, demografía y educativas aplicando Modelos de regresión de Poisson.* Universidad Nacional Mayor de San Marcos, Lima, Perú : 2005.

FOX, John. *Applied regression analysis generalized linear models.* California, United States : SAGE Publications, 2016.

HENRIK, Madsen; & THYREGOD, Poul. *Introduction to general and generalized linear models.* London, New York : Taylor & Francis Group, 2010.

LLUCH, José. *Tecnología y margen de refino del petróleo.* Madrid- España : Díaz de Santos S.A., 2011.

LÓPEZ-ONZÁLEZ, Emelina; & RUIZ SOLER, Marcos. *Análisis de datos con el Modelo Lineal generalizado.* 2011, revista española de pedagogía rep, pp.59-80.

MACÍAS, Daniel; & MARTÍNEZ, Jorge. *Modelación para la programación del transporte de productos refinados en la red nacional de poliducto de Ecopetrol S.A.* Pontificia Universidad Javeriana, s.l. : 2012.

MENON, Shashi. *Transmission Pipeline Calculations and Simulations Manual.* s.l. : Gulf Proffesional Publishing, 2014.

MIESER, Thomas; & LEFFLER, William. *OIl and Gas PIpelines in Nontechnical language.* s.l. : Tulsa : PennWell Corp., 2006.

MONROY, Saldívar. *Estadística Descriptiva.* México : Instituto Politécnico Nacional, 2008.

MONTESINOS, Abelardo. *Estudio del AIC y BIC en la selección de modelos de vida con datos censurados.* Centro de Investigación en Matemáticas A.C., s.l. : 2011.

MOSQUERA, Vanessa; & SIMBAÑA, Jonathan. *Análisis de las exportaciones de los derivados del petróleo y su incidencia en la balanza comercial del Ecuador Periodo 2010-2017.* Universidad de Guayaquil, s.l. : 2019.

MUIRRAGUI, Viena; & GUILLÍN, Carlos. *Análisis del impacto económico que tendrá la refinería del Pacífico Eloy Alfaro en la economía ecuatoriana.* Universidad Estatal de Milagro, s.l. : 2013.

MURRAY, Spiegel; & LARRY, Stephens. *Estadística Schaum.* Cuarta. México : MCGraw-Hill/Interamericana Editores, S.A. de C.V., 2009.

PEÑA, Daniel. *Análisis de datos multivariantes.* 2002.

PULIDO, Humberto; & SALAZAR, Román. *Análisis y Diseño de experimentos.* Segunda. México : McGraw-Hill/Ineramericana editores, S.A. de C.V, 2008.

SUPURRIER, Walter. *Impacto sobre el gasto público,* 1996, ILDIS, Vol. 1, p.155.

TAPIA, Oswaldo. *OPEC Annual Statistical Bulltin.* 2017, Organization of petroleum exporting countries, Vol. 52, p.133.

VIEDMA, Carlos. *Estadística descriptiva e inferencial.* Madrid, España : Ediciones IDT, 2018.

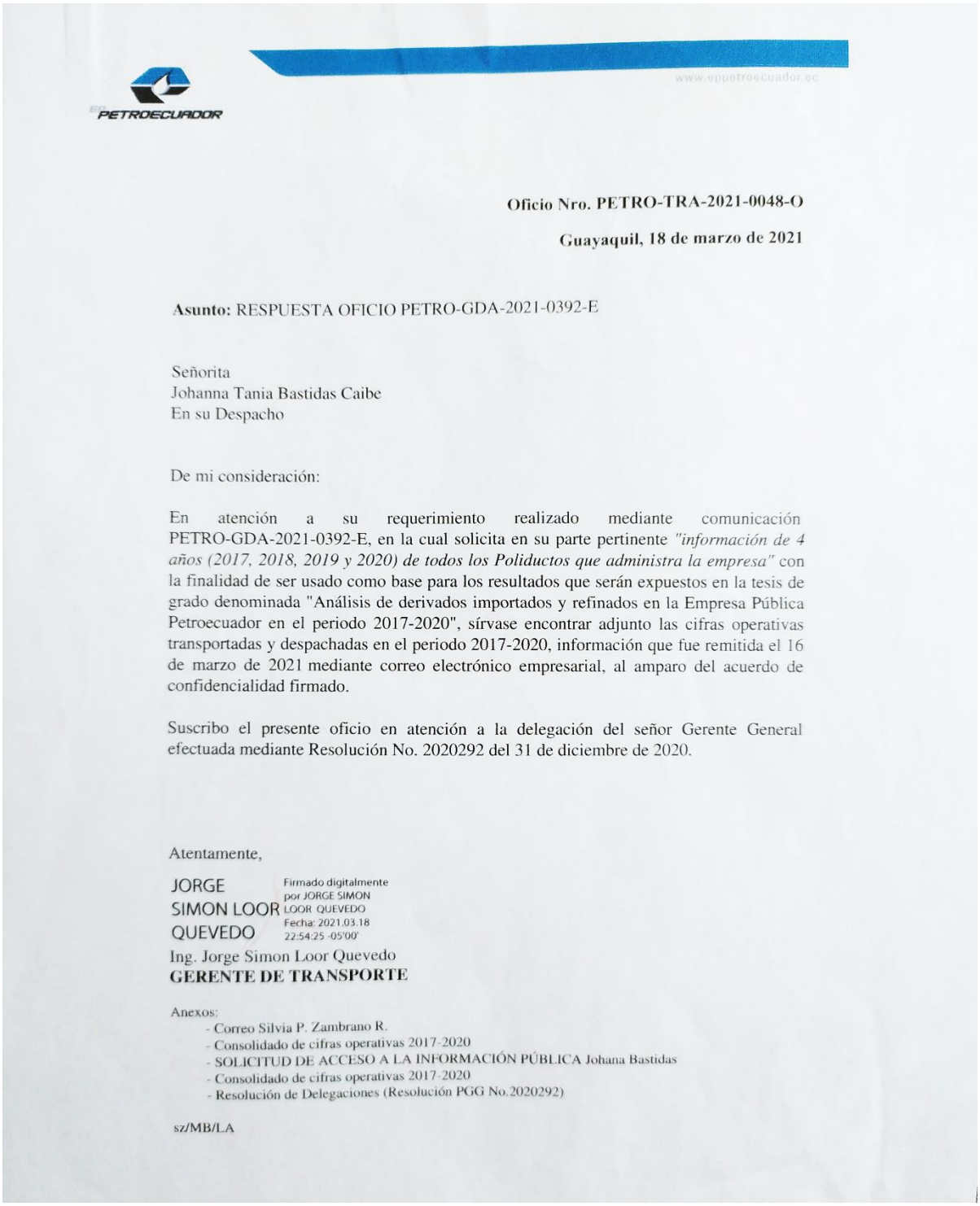
WARREN, John K. *Evaporites:Sediments, Resources and Hydrocarbons.* s.l. : Springer Science y Business Media, 2006.

LEONARDO FABIO MEDINA NUSTE

Firmado digitalmente por LEONARDO FABIO MEDINA NUSTE
Nombre de reconocimiento (DN): c=EC, o=BANCO CENTRAL DEL ECUADOR, ou=ENTIDAD DE CERTIFICACION DE INFORMACION-ECIBCE, l=QUITO, serialNumber=0000621485, cn=LEONARDO FABIO MEDINA NUSTE
Fecha: 2022.01.24 13:18:58 -05'00'

ANEXOS

ANEXO A: AVAL DE EP PETROECUADOR



PETROECUADOR

www.eppetroecuador.ec

Oficio Nro. PETRO-TRA-2021-0048-O

Guayaquil, 18 de marzo de 2021

Asunto: RESPUESTA OFICIO PETRO-GDA-2021-0392-E

Señorita
Johanna Tania Bastidas Caibe
En su Despacho

De mi consideración:

En atención a su requerimiento realizado mediante comunicación PETRO-GDA-2021-0392-E, en la cual solicita en su parte pertinente *"información de 4 años (2017, 2018, 2019 y 2020) de todos los Poliductos que administra la empresa"* con la finalidad de ser usado como base para los resultados que serán expuestos en la tesis de grado denominada "Análisis de derivados importados y refinados en la Empresa Pública Petroecuador en el periodo 2017-2020", sírvase encontrar adjunto las cifras operativas transportadas y despachadas en el periodo 2017-2020, información que fue remitida el 16 de marzo de 2021 mediante correo electrónico empresarial, al amparo del acuerdo de confidencialidad firmado.

Suscribo el presente oficio en atención a la delegación del señor Gerente General efectuada mediante Resolución No. 2020292 del 31 de diciembre de 2020.

Atentamente,

JORGE SIMON LOOR QUEVEDO
Firmado digitalmente por JORGE SIMON LOOR QUEVEDO
Fecha: 2021.03.18 22:54:25 -05'00'

Ing. Jorge Simon Loor Quevedo
GERENTE DE TRANSPORTE

Anexos:
- Correo Silvia P. Zambrano R.
- Consolidado de cifras operativas 2017-2020
- SOLICITUD DE ACCESO A LA INFORMACIÓN PÚBLICA Johana Bastidas
- Consolidado de cifras operativas 2017-2020
- Resolución de Delegaciones (Resolución PGG No.2020292)

sz/MB/LA

Anexo B: CÓDIGO EN R

Regresión lineal múltiple

```
library(readxl)
library(dplyr)
library(ggplot2)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
library(psych)
library(tseries)
library(nortest)## data
library(scales)
library(lmtest)
leer la base de datos

datam2<- read.table("DOCS/petro1.txt",header = TRUE, dec = ","); datam2

# Anpalisis descriptivo
#variables cualitativas
addmargins(table(datam2$poliducto,datam2$anio))
addmargins(table(datam2$tramo, datam2$anio))
addmargins(table(datam2$productos, datam2$anio))
#variables cuantitativas
summary(datam2[, -(1:4)])
summary(datam2$VOLUMEN)
# GRÁFICAS
dev.off()
barplot(datam2$poliducto)
hist(datam2$VOLUMEN)
density(datam2$VOLUMEN)

#matriz de varianzas y cov
var <- var(datam2[,-(1:4)])

  aut <- eigen(var)
round(aut$values,2)
```

```
resu<-round(cor(datam2),2)
#solo corrplot resu
corrplot(resu, method = "number", type = "upper")

#histograma de la variable VOLUMEN
ggplot(data = datam2) +
  geom_histogram( bins = 30, fill = "skyblue", color = "black", mapping = aes(x = VOLUMEN,
y = ..density..)) +
  stat_function(fun = dnorm,
                args = list(mean = mean(datam2$VOLUMEN), sd = sd(datam2$VOLUMEN)),
                color = "red",
                size = 1, xlim = c(9, 45))+ labs(title = "VOLUMEN TRANSPORTADO (barriles)",
                                                 caption = "Johanna Bastidas C", y=" ")+  theme_minimal()

#  datam<- read.table("DOCS/petro_g.txt",header = TRUE, sep = " "); datam
ggplot(data = poli, mapping = aes(x = poliducto)) +
  geom_bar(fill = "skyblue", color = "black")+ labs(title = "POLIDUCTOS",
                                                    caption = "Johanna Bastidas C", y="Frecuencia")+
  theme_minimal()

#pie

ggplot(poli2,aes(x="",y=porcentaje, fill= poliducto))+
  geom_bar(stat = "identity",
           color="white")+
  geom_text(aes(label=percent(porcentaje/100)),
            position=position_stack(vjust=0.5),color="white",size=6)+
  coord_polar(theta = "y")+
  scale_fill_manual(values=c("salmon","steelblue","orange","gray","skyblue" ))+
  theme_void()+
  labs(title="POLIDUCTOS")+ theme(plot.title = element_text(hjust = 1))
pairs.panels(datam2)
#egresion multiple
rem2<-lm(VOLUMEN~.,datam2);rem2

  #coeficiente de determinacion
summary(rem2)
```

```
#significancia
anova(rem2)
#supuestos
#normalidad
jarque.bera.test(rem2$residuals)
#homocedasticidad
gqtest(rem2)
  # independencia

dwtest(rem2)

dev.off()

hist(rem2$residuals)
boxplot(rem2$residuals, main= "Box plot de los residuos")
lillie.test(rem2$residuals)
boxplot(datam2$VOLUMEN)
#datos atipicos
boxplot(datam2)
par(mfrow=c(2,2))
plot(rem2)

confint(rem2,level = 0.95)
```

Ajuste del modelo tras ejecutar un análisis descriptivo de detección de datos atípicos multivariado

```
rem3_1 <- lm(VOLUMEN~poliducto + tramo + productos+ Capacidad+ vdespachado+
                poliducto*tramo + poliducto*productos + poliducto*Capacidad+
                poliducto*vdespachado+ tramo*Caudal + tramo*Caudal+ productos*Capacidad+
                productos*Caudal, datam2)
summary(rem3_1)
anova(rem3_1)

rem_4_1 <- lm(VOLUMEN~poliducto + tramo + productos+ Capacidad+ vdespachado+
                 poliducto*tramo + poliducto*productos + poliducto*Capacidad+
                 poliducto*vdespachado+ tramo*Caudal+ productos*Capacidad+
                 productos*Caudal+ productos*vdespachado +Capacidad *Caudal+
                 poliducto*tramo*productos + poliducto*tramo*vdespachado+
                 tramo*productos*Capacidad+ productos*Capacidad*vdespachado+
                 productos*Caudal*vdespachado , datam2 )
summary(rem_4_1)
anova(rem_4_1)

rem_4_2 <- lm(VOLUMEN~.^4 , datam2 )
summary(rem_4_2)
anova(rem_4_2)
rem_4_3 <- lm(VOLUMEN~poliducto + tramo + productos+ Capacidad+ vdespachado+
                 poliducto*tramo + poliducto*productos + poliducto*Capacidad+
                 poliducto*vdespachado+ tramo*productos+tramo*Caudal+
productos*Capacidad+
                 productos*Caudal+ productos*vdespachado +anio*Capacidad+
                 Capacidad *Caudal+poliducto*tramo*productos + poliducto*tramo*Capacidad+
              poliducto*tramo*vdespachado+ poliducto*Caudal*vdespachado+
                 tramo*productos*Capacidad+ productos*Capacidad*vdespachado+
                 productos*Caudal*vdespachado+ poliducto*tramo*productos*vdespachado +
                 poliducto*productos*Caudal*vdespachado, datam2 )
rem_4_3;
summary(rem_4_3)
anova(rem_4_3)
```

```
par(mfrow=c(2,2))

plot(rem_4_3)
#significancia
#supuestos
#normalidad
jarque.bera.test(rem2$residuals)
#homocedasticidad
      gqtest(rem2)
# independencia

dwtest(rem2)
#SUPUESTOS REM4_3

#normalidad
jarque.bera.test(rem_4_3$residuals)
#homocedasticidad
gqtest(rem_4_3)
# independencia

dwtest(rem_4_3)

a<-step(object = rem_4_3, direction = "both", trace = 1);a

rem5 <- lm(VOLUMEN~poliducto + tramo + productos + Capacidad +
           vdespachado + Caudal + anio + poliducto*tramo +
           poliducto*productos + poliducto*Capacidad +
           poliducto*vdespachado + tramo*productos + tramo*Caudal +
           productos*Capacidad + productos*Caudal + productos*vdespachado +
           Capacidad*Caudal + tramo*Capacidad + tramo*vdespachado +
           poliducto*Caudal + vdespachado*Caudal + Capacidad*vdespachado +
           poliducto*tramo*productos + poliducto*tramo*vdespachado +
           poliducto*vdespachado*Caudal + tramo*productos*Capacidad +
           productos*Capacidad*vdespachado + productos*vdespachado*Caudal +
           poliducto*productos*vdespachado + tramo*productos*vdespachado +
           poliducto*productos*Caudal + poliducto*tramo*productos*vdespachado +
           poliducto*productos*vdespachado*Caudal, datam2)
```

```
rem5
rem6 <- lm(VOLUMEN~.^4, datam2[,-2])
rem6
a<-step(object = rem6, direction = "both", trace = 1);a
summary(rem6)
anova(rem6)
#normalidad
jarque.bera.test(rem6$residuals)
#homocedasticidad
gqtest(rem6)
# independencia

dwtest(rem6)
confint(rem5)
AS <- predict(rem5, data.frame( poliducto=c(1:5), tramo= c(6:10), productos=c(14:18),
                    Capacidad=c(95000,96000,109000,110000,115000), vdespachado=
c(18000,5700000, 6300000,7500000, 9000000),
                    Caudal= c(4000, 4500, 5000, 5500, 6000), anio= c(2021, 2022, 2023,2024,
2025)))
a <- data.frame( poliducto=c(1:5), tramo= c(6:10), productos=c(14:18),
              Capacidad=c(95000,96000,109000,110000,115000), vdespachado=
c(18000,5700000, 6300000,7500000, 9000000),
              Caudal= c(4000, 4500, 5000, 5500, 6000), anio= c(2021, 2022, 2023,2024, 2025),
VOLUMEN=-AS)

#mejorar el modelo

AS <- predict(rem6, data.frame( poliducto=c(1:5), productos=c(14:18),
                        Capacidad=c(95000,96000,109000,110000,115000), vdespachado=
c(18000,5700000, 6300000,7500000, 9000000),
                        Caudal= c(4000, 4500, 5000, 5500, 6000), anio= c(2021, 2022, 2023,2024,
2025)))
a <- data.frame( poliducto=c(1:5), tramo= c(6:10), productos=c(14:18),
              Capacidad=c(95000,96000,109000,110000,115000), vdespachado=
c(18000,5700000, 6300000,7500000, 9000000),
              Caudal= c(4000, 4500, 5000, 5500, 6000), anio= c(2021, 2022, 2023,2024, 2025),
VOLUMEN=-AS)
```

Modelo lineal generalizado: Regresión de Poisson

```
rem3 <- glm(VOLUMEN~ poliducto+ tramo + productos + anio + Capacidad+
            Caudal+ vdespachado+ offset(log(total)), family = poisson); rem3
summary(rem3)
a <-(1- 74270000)/(1.1e+08)
# evaluacion de la bondad de ajuste del modelo
library(AER)
dispersiontest(rem3)
pchisq(rem3$deviance, df=rem3$df.residual, lower.tail=FALSE)
cbind(exp(coef(rem3)),exp(confint(rem3)))
```

Escoger el mejor modelo

```
anova(rem2, rem3)
AIC(rem2, rem3)
AIC(rem3)
anova(rem2, rem3)
```

Modelo de regresión logística multinomial

```
library(ISLR)
library(tidyverse)
library(readxl)
library(dplyr)
library(ggplot2)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
library(psych)

library(tseries)
library(nortest)## data
library(scales)
library(lmtest)
library(tidyverse)
library(caret)
library(foreign)
library(nnet)
```

```
library(ggplot2)
library(reshape2)

#leer la base de datos
x1<- read.table("DOCS/petrologit.txt",header = TRUE, dec = ","); x1
names(x1)
dim(x1)
r <- table(x1$poliducto, x1$tramo, dnn = c("poliducto","tramo "))
addmargins(r)
r <- table(x1$poliducto, x1$productos, dnn = c("poliducto","productos "))
addmargins(r)
  with(x1, do.call(rbind, tapply(VOLUMEN, tramo, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(VOLUMEN, productos, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(VOLUMEN, anio, function(x)c(M=mean(x), SD=sd(x)))))

 with(x1, do.call(rbind, tapply(Capacidad, tramo, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(Capacidad, productos, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(Capacidad, anio, function(x)c(M=mean(x), SD=sd(x)))))

 with(x1, do.call(rbind, tapply(Caudal, tramo, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(Caudal, productos, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(Caudal, anio, function(x)c(M=mean(x), SD=sd(x)))))

 with(x1, do.call(rbind, tapply(vdespachado, tramo, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(vdespachado, productos, function(x)c(M=mean(x), SD=sd(x)))))
 with(x1, do.call(rbind, tapply(vdespachado, anio, function(x)c(M=mean(x), SD=sd(x)))))
 class(x1$poliducto)

names(x1)
#View(x1)
test <- multinom(poliducto~productos+VOLUMEN, x1);test
summary(test)
##############################################################################
#######
#######################              PREDICT
exponentes <- exp(coef(test)); exponentes
dataf <- data.frame(productos= c(1:19),
```

```
                VOLUMEN=mean(x1$VOLUMEN) )



pred <- predict(test, newdata = dataf, type = "probs"); pred
#########################################################
####      GRAFICOS
#########################################################

# dwrite <- data.frame(productos=rep(c(1:19),2) ,
VOLUMEN=c(rep(c(median(datam2$VOLUMEN), mean(datam2$VOLUMEN),
#
max(datam2$VOLUMEN)),6),c(median(datam2$VOLUMEN))))

dwrite <- data.frame(productos=rep(c(6,10,15),3) ,
VOLUMEN=c(rep(median(datam2$VOLUMEN), 3),rep(mean(datam2$VOLUMEN),3),
                                        rep(max(datam2$VOLUMEN), 3)))

ppwrite <- cbind(dwrite, predict(test, newdata= dwrite, type= "probs", se=T))
head(ppwrite)

# poner la transpuesta
xx <- melt(ppwrite, id.vars = c("productos","VOLUMEN"), value.name = "probability")
head(xx)

#grafica
ggplot(xx, aes(x = VOLUMEN, y= probability, color= productos))+ geom_line()+
facet_grid(variable~., scales = "free")
```

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

DIRECCIÓN DE BIBLIOTECAS Y RECURSOS DEL APRENDIZAJE

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: *14 / 09 / 2021*

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: *Johanna Tania Bastidas Caibe*
INFORMACIÓN INSTITUCIONAL
Facultad: *Ciencias*
Carrera: *Estadística*
Título a optar: *Ingeniera en Estadística*
f. Analista de Biblioteca responsable: *Ing. Leonardo Medina Ñuste MSc.*

LEONARDO FABIO MEDINA NUSTE

Firmado digitalmente por LEONARDO FABIO MEDINA NUSTE
Nombre de reconocimiento (DN): c=EC, o=BANCO CENTRAL DEL ECUADOR, ou=ENTIDAD DE CERTIFICACION DE INFORMACION-ECIBCE, l=QUITO, serialNumber=0000621485, cn=LEONARDO FABIO MEDINA NUSTE
Fecha: 2021.09.14 17:05:23 -05'00'