5 DE MAYO DEL 2023

Facultad de Estudios Superiores

Acatlán

ESTADÍSTICA II

DÍAZ BARRIGA REYES RODOLFO 100%
GARCIA PEREZ SALOMON 100%
MEJIA ESPINOSA RUBEN ALAN 100%
PINEDA ROGDRIGEZ ALEXIS CRISTIAN 100%
SARMIENTO IBARRA JONATAN 100%
URQUIZA ROBLES DIEGO ISMAEL 100%

Contenido

Tabla de ilustraciones

1 Resumen

El objetivo de este análisis de regresión lineal múltiple es examinar la relación entre la demanda de combustibles específicos, como el gas natural, gas LP, coque de petróleo, diésel y turbosina, respecto las emisiones de gases de efecto invernadero (GEI) en México, y determinar la naturaleza y fuerza de la correlación entre estas variables.

Se recopiló información sobre la demanda de combustibles específicos y las emisiones de GEI en México durante un período de 25 años y se utilizó un modelo de regresión lineal múltiple para examinar la relación entre estas variables.

Los resultados del análisis de regresión lineal múltiple indican que la demanda de gas natural, gas LP, coque de petróleo, diésel y turbosina están positivamente correlacionados con las emisiones de GEI en México. La fuerza de la correlación varía el año, probablemente debido a políticas implementadas durante estos.

El análisis de regresión lineal múltiple indica que la demanda de combustibles específicos es un factor importante que contribuye a las emisiones de GEI en México. Es necesario tomar medidas para reducir la demanda de estos combustibles y mejorar la eficiencia del transporte para reducir la emisión de GEI y mejorar la calidad del aire. Además, se deben considerar otros factores para futuros análisis, que afectan como el tipo de vehículo y las condiciones de manejo, para desarrollar políticas efectivas de reducción de emisiones.

Comentado [JV1]: ¿Y los números de página?

Comentado [JV2]: Buen resumen

2 Introducción

La emisión de gases de efecto invernadero (GEI) es un problema ambiental y de salud pública en México, donde se han registrado niveles alarmantes de contaminación en diversas ciudades del país. El sector transporte es uno de los principales contribuyentes a esta problemática, con emisiones de GEI derivadas del uso de combustibles fósiles como la gasolina y el diésel. A medida que la población urbana sigue creciendo y el número de vehículos aumenta, el impacto de estas emisiones en la calidad del aire y en la salud de la población se vuelve cada vez más significativo.

Para abordar este problema, es fundamental comprender la relación entre las emisiones y el consumo de combustibles. En este contexto, el análisis de regresión lineal se presenta como una herramienta valiosa para examinar esta relación y determinar la naturaleza y la fuerza de la correlación entre las variables.

En el análisis de regresión lineal, se utiliza una ecuación matemática para modelar la relación entre una variable dependiente (en este caso, las emisiones de GEI) y una o varias variables independientes (en este caso, el consumo de combustibles). La ecuación matemática de la regresión lineal se utiliza para predecir el valor de la variable dependiente en función de los valores de las variables independientes.

En el contexto del consumo de combustibles y las emisiones de GEI en México, se espera que exista una relación directa entre estas variables, ya que cuanto mayor sea el consumo de combustible, mayor será la cantidad de GEI emitidos. Sin embargo, la fuerza de esta correlación puede variar dependiendo de factores tanto ambientales, como sociales, es menester realizar un estudio a mas profundidad, considerando estos factores.

3 Marco teórico

3.1 Antecedentes

El termino regresión lineal se atribuye a Francis Galton (s. XIX), quien al describir un fenómeno biológico logro observar que tener ancestros de alta estatura no garantiza la misma, ya que esta tiende a la media. (Soriano, 2016, pág. 10). Existen modelos tales como la regresión línea, cuya principal característica es ser una combinación lineal de los parámetros del modelo

$$\psi(\vec{X}, \vec{\beta}) = \epsilon + \sum_{i=1}^{p} \beta_i Xi$$

donde $\vec{\beta}$ es el vector de parámetros que debemos calcular una vez obtenido nuestros datos. (Soriano, 2016)

3.2 Regresión lineal simple

El modelo mencionado tendrá únicamente dos parámetros β_0 y β_1, será llamado regresión lineal simple. Para obtener los estimadores de nuestros parámetros existen dos métodos: mínimos cuadrados y máxima verosimilitud; En particular para la regresión lineal simple, ambos estimadores serán los mismos. (Soriano, 2016)

Por lo tanto, los estimadores de mínimos cuadrados son aquellos, que como su nombre lo indica, obtendremos minimizando la suma de los cuadrados de los errores $e_i = y_i - \hat{y}_i$ donde:

$$\hat{y}_i = \sum \hat{\beta}_i x_i$$

Para el modelo de regresión lineal simple tenemos que:

$$\min \psi(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Tenemos que encontrar la derivada de esta suma de cuadrados respecto a los parámetros $\vec{\beta}$:

$$\frac{\partial \psi}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \qquad \frac{\partial \psi}{\partial \beta_0} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Procedemos a igualar las derivadas con cero.

$$2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Podemos deducir el siguiente sistema de ecuaciones:

Comentado [JV3]: No pedí marco teórico. Pedí los anexos que para este propósito les envié por correo.

Comentado [JV4]: Si la cita de paráfrasis termina una oración o un párrafo, el punto va después de la cita APA, no antes.

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Cuya solución es:

$$\hat{\beta}_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ahora queda determinar qué tipo de punto crítico obtuvimos, calculamos la matriz Hessiana y la determinante:

$$\det(H) = \begin{vmatrix} \frac{\partial^2 \psi}{\partial \hat{\beta}_0^2} & \frac{\partial^2 \psi}{\partial \hat{\beta}_1 \hat{\beta}_0} \\ \frac{\partial^2 \psi}{\partial \hat{\beta}_1 \hat{\beta}_0} & \frac{\partial^2 \psi}{\partial \hat{\beta}_0^2} \end{vmatrix} = \begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix} = 4nS_{xx}$$

Por el criterio Hessiano, existe un mínimo local si $\det(H) > 0$ y $\frac{\partial^2 \psi}{\partial \hat{\beta}_0^2} > 0$ notese que esto sera cierto siempre que n y el determinante de nS_{xx} sean mayores que 0. (Soriano, 2016)

3.3 Regresión lineal simple

Comentado [JV5]: múltiple

Lo establecido son los cálculos del modelo de regresión lineal simple, dada una variable regresora, es menester realizar el cálculo para p variables, es decir $p + 1$ estimadores $\hat{\beta}_0, \ldots, \hat{\beta}_p$, el modelo será análogo al de regresión lineal simple, donde tenemos $p = 1$ teniendo dos estimadores.

Dicho esto, definimos:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p} \end{pmatrix} \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} Y = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

El modelo será reescrito como:

$$Y = X\beta + \epsilon$$

Análogamente:

$$\hat{Y} = \hat{X}\beta$$

Continuamos con el cálculo de mínimos cuadrados para regresión múltiple:

$$\psi(\beta) = \left(Y - \hat{Y}\right)^t \left(Y - \hat{Y}\right) = (Y^2 - \beta^t X^t)^t (Y - X\beta)$$
$$= Y^t Y - Y^t X\beta - \beta^t X^t Y + \beta^t X^t X\beta$$
$$= Y^t Y - 2\beta^t X^t Y + \beta^t X^t X\beta$$

Nótese que $Y^tX\beta\ y\ \beta^tX^tY$ son simtricas, dada su dimensión $1X1$, seguimos con el cálculo de las derivadas parciales de ψ respecto de β e igualamos a 0:

$$\frac{\partial\psi}{\partial\beta} = -2X^tY + 2X^tX^t\beta$$
$$-2X^tY + 2X^tX^t\beta = 0 \quad \Leftrightarrow \quad \hat{\beta} = (X^tX)^{-1}X^tY$$

Hemos obtenido los estimadores del modelo, siempre que exista $(X^tX)^{-1}$ lo que implica necesariamente que las variables son linealmente independientes. Es menester mencionar que no es suficiente lo ya establecido, pues necesitamos verificar los supuestos que sustentan el modelo:

- $\epsilon \sim N(0, \sigma^2 I)$
- Homocedasticidad
- No multicolinealidad
- No presencia de valores atípicos
- No correlación de los errores

Sin embargo, después de verificar que se cumplen los supuestos, es necesario asegurar que el modelo propuesto es justificable y que no existe mejor modelo que se ajuste a los datos, esto lo realizaremos a través del teste de falta de ajuste, el cual deberá ser realizado antes de corroborar todos los supuestos o de lo contrario habremos perdido tiempo en un modelo incorrecto, lo cual puede solucionarse realizando transformaciones en las variables. (Darper & Smith, 1998)

Comentado [JV6]: test

3.4 Estadística inferencial

El primer supuesto del modelo de regresión lineal tiene como fin realizar el análisis inferencial sobre los estimadores $\hat{\beta}$ y sobre posibles predicciones de la variable Y, Recordemos que el modelo tiene la forma $\hat{Y} = \hat{X}\beta$ en su forma multivariable, dado el mismo podemos realizar implicaciones bajo el primer supuesto:

$$E[Y] = E[X\beta + \epsilon]$$
$$= X\beta$$

$$Var[Y] = Var[x\beta + \epsilon]$$
$$= Var[\epsilon]$$
$$\sigma^2 I$$

Que implica:

$$Var[\hat{\beta}] = Var[(X^tX)^{-1}X^tY]$$
$$= (X^tX)^{-1}X^tVar[Y][(X^tX)^{-1}X^t]^t$$
$$= (X^tX)^{-1}X^t\sigma^2 I[X(X^tX)^{-1}]$$
$$= \sigma^2(X^tX)^{-1}$$

Sabemos que Y tiene una distribución normal, dado que ϵ también lo es. A su vez como los estimadores de mínimos cuadrados (como los máximos verosímiles) son combinaciones lineales de Y_i, esto implica necesariamente que $\hat{\epsilon} \sim N$. Adicionalmente sabemos que:

$$c_i = \frac{X_i - \bar{X}}{S_{XX}} \qquad \sum_{i=1}^{n} c_i = 0 \qquad \sum_{i=1}^{n} X_i c_i = 1$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \qquad \hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{X}\right) Y_i$$

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} c_i E(Y_i)$$
$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i$$
$$= \beta_1$$

$$E(\hat{\beta}_0) = \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{X}\right) E(Y_i)$$
$$= \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{X}\right)(\beta_0 + \beta_1 X_i)$$
$$= \sum_{i=1}^{n} \left(\frac{\beta_0}{n} + \frac{\beta_1 X_i}{n} - c_i \bar{X}\beta_0 - \beta_1 \bar{X} c_i X_i\right)$$
$$= \beta_0 + \beta_1 \bar{X} + 0 - \beta_1 \bar{X} = \beta_0$$

$$Var(\hat{\beta}_1) = \sum_{i}^{n} c_i^2 Var(Y_i)$$
$$= \frac{\sigma^2}{S_{XX}^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$
$$= \frac{\sigma^2}{S_{XX}}$$

$$Var(\hat{\beta}_0) = \sum_{i}^{n} \left(\frac{1}{n} - c_i \bar{X}\right) Var(Y_i)$$
$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{2\bar{X} c_i}{n} + c_i^2 \bar{X}^2\right)$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$$

Que en conjunto con todo lo establecido anteriormente, implica que:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)\right) \qquad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

(Montgomery, 2021)

Notemos que existe un parámetro del modelo σ^2 constante para toda Y_i . El estimador máximo verosímil es:

$$\hat{\sigma}_{MV}^2 = \frac{(Y - \hat{Y})^t (Y - \hat{Y})}{n}$$

Donde tenemos p parámetros regresores; cuando $p = 2$, el caso simple esto se reduce a que:

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Sin embargo, este estimador es sesgado, puesto que $n\frac{\hat{\sigma}^2_{MV}}{\sigma^2} \sim X^2_{n-p}$, lo que implica que $E[\hat{\sigma}^2_{MV}] = \sigma^2\frac{n-p}{n}$, para evitar esto utilizamos el estimador insesgado:

$$\hat{\sigma}^2 = \frac{(Y-\hat{Y})^t(Y-\hat{Y})}{n-p} = \frac{n}{n-p}\sigma^2_{MV}$$

(Soriano, 2016)

Queda realizar intervalos de confianza, estimaciones puntuales y pruebas de hipótesis a partir de que:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1}_{i+1,i+1}}} \sim t_{n-p}$$

Es de interés realizar el análisis inferencial sobre la media Y dada x_0, es decir $E[Y_0|X = x_0]$, la cual bajo el modelo propuesto es $X_0\beta$. Para ello usaremos el estimador insesgado $\hat{Y}_0 = X_0\beta$

$$\begin{aligned}
&E[\hat{Y}_0] = E[X_0\hat{\beta}] \\
&= X_0 E[\hat{\beta}] \\
&= X_0\beta \\
&Var[\hat{Y}_0] = Var[X_0\hat{\beta}] \\
&= X_0^t Var[\hat{\beta}] X_0 \\
&= \sigma^2 X_0^t (X\hat{X})^{-1} X_0
\end{aligned}$$

(Soriano, 2016)

Como $\hat{Y}_0$ es una combinación lineal de las variables distribuidas normalmente, se distribuye normal y el estadístico que ocupamos para todo el análisis inferencial es:

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - E[Y_0|X = x_0]}{\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0}} \sim t_{n-p}$$

El estimador para el caso simple seria:

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - E[Y_0|X = x_0]}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-p}$$

Queda de interés encontrar el intervalo de confianza para la variable $\hat{Y}_0$ dado un valor x_0, también llamado intervalo de predicción. Como $Y_0 \sim N$ y $\hat{Y}_0 \sim N$, entonces $Y_0 - \hat{Y}_0 \sim N$, junto con la media y la varianza:

$$\begin{aligned}
[Y_0 - \hat{Y}_0] &= E[Y_0] - E[\hat{Y}_0] \\
&= (X\beta) - (X\beta) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
Var[Y_0 - \hat{Y}_0] &= Var[Y_0] + Var[\hat{Y}_0] \\
&= \sigma^2 + \sigma^2 x_0^t (X^t X)^{-1} x_0
\end{aligned}$$

$$= \sigma^2(1 + x_0^t(X^tX)^{-1}x_0)$$

Generalizando el estadístico:

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2(1 + x_0^t(X^tX)^{-1}x_0)}} \sim t_{n-p}$$

En particular para regresión simple es:

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

(Soriano, 2016)

3.5 Multicolinealidad

"Dos variables estadísticas son **estadísticamente independientes** cuando el comportamiento estadístico de una de ellas no se ve afectado por los valores que toma la otra." ("Estadística Descriptiva - MATERIA. ESTADISTICA DESCRIPTIVA ... - Studocu") La correlación se refiere a la **relación lineal** entre dos variables aleatorias, en otras palabras, si dos variables están correlacionadas, el valor de una variable tiende a afectar directamente el valor de la otra variable. La correlación está íntimamente relacionada con la covarianza, la cual, al igual que la correlación, mide la relación lineal en un par de muestras X y Y de tamaño n. La covarianza está definida como

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Una covarianza positiva indica que el producto los términos $(x_i - \bar{x})$ y $(y_i - \bar{y})$ en la sumatoria suelen ser más positivos que negativos, lo cual indica que, en conjunto, los valores de X y Y suelen estar por encima de su media correspondiente. Por otro lado, una covarianza negativa indica que los valores de X y Y suelen estar por debajo de su media correspondiente. Ya que, para el cálculo de la covarianza se tienen que restar los valores de la muestra directamente, se trata de un valor afectado por las unidades que se emplearon al tomar la muestra, por lo tanto, no hay una escala que indique si se tiene una relación lineal alta entre las variables o no. Para solucionarlo, se puede dividir cada diferencia entre la desviación estándar de dicha variable para normalizar su valor, con lo que se obtiene la medida de correlación $\rho_{X,Y}$

$$\rho_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_X}\right)\left(\frac{y_i - \bar{x}}{\sigma_Y}\right) = \frac{Cov(X, Y)}{\sigma_X\sigma_Y}$$

Se dice que no hay correlación entre dos variables X y Y si $\rho_{X,Y} = 0$.

Ya que la correlación es una medida que no se ve afectada por las unidades de la muestra, se puede asumir fácilmente si dos variables están correlacionadas o no, ya que el coeficiente de correlación anterior tiene un rango fijo de $-1 \leq \rho_{X,Y} \leq 1$. Para obtener la correlación en R se emplea la función $\boldsymbol{cor}(\boldsymbol{X}, \boldsymbol{Y})$. No requiere ninguna librería.

Un coeficiente de **correlación** cercano a cero indica que no existe una relación lineal entre dos variables, pero eso no significa que no pueda haber otro tipo de relación entre dichas variables. La definición de **independencia** nos indica que, si existe cualquier tipo de relación entre dos variables, se dice que son variables dependientes,

Comentado [JV7]: Hay un montón de plagio en su marco teórico: párrafos completos sin cita a la fuente de donde tomaron las ideas. Debería rebotarles el trabajo en cero, pero mejor descontaré cinco puntos.

de tal manera que rechazar la relación lineal entre dos variables no implica que las variables sean independientes, ya que puede haber otro tipo de relación entre dichas variables.

En otras palabras, se puede resumir independencia y correlación con los siguientes dos puntos:

- **Independencia**: No existe ningún tipo de relación entre dos variables
- **No correlación**: No existe una relación lineal entre dos variables

(Vega Ayala, Diaz Barriga, & Garcia Totozintle, 2023)

3.6 Homocedasticidad

"Se dice que la varianza del término de perturbación del modelo de regresión lineal es heterocedástica cuando no es constante para todas las observaciones" (Universidad del País Vasco [EHU], 2017). En el modelo lineal general, $y = X\beta + u$, se supone que la perturbación aleatoria es tal que $E[u] = 0_{n\times 1}$ y $Var(u) = E[uu^t] = \sigma^2 I_{n\times n}$, lo cual implica que:

- $E[u_t] = 0, \forall t \in \{1, \dots, n\}$
- $E[u_t^2] = Var(u_t) = \sigma^2 \ \forall t \in \{1, \dots, n\}$ (varianza constante = homocedasticidad)
- $E[u_i u_j] = Cov(u_i, u_j) = 0, \forall i \neq j \in \{1, \dots, n\}$ (incorreción)

"Cuando se incumple el supuesto de homocedasticidad, es decir, la varianza no es constante se dice que hay heterocedasticidad".

Las principales causas de heterocedasticidad de un modelo son: situaciones en las que se disponen de datos de sección cruzada, cuando las observaciones de la variable dependiente pueden subdividirse en grupos y se usan como datos los promedios proporcionados por tales grupos, si se omite una variable relevante en el modelo, es esperable que la perturbación aleatoria dependa de dicha variable omitida.

(Hernandez Perez & Urquiza Robles, 2023)

3.7 Independencia y no correlación

Existen diversos métodos que nos permiten detectar la multicolinealidad en un modelo de regresión; algunos de ellos son:

- Analizar la matriz de correlaciones R, esto es, teniendo la matriz de correlaciones entre variables tenemos los siguientes criterios:
  - Se determina que existen serios problemas de multicolinealidad si sus correlaciones son mayores o iguales a 0.9.
  - Si las correlaciones están entre 0.7 y 0.9 entonces existen problemas moderados de colinealidad.
- Basarnos en el coeficiente de determinación múltiple de cada variable x_j con las restantes y los coeficientes de correlación parcial de las variables x_j y x_k.
- En función a los valores y vectores propios de la matriz de correlación.
- A través del error estándar del j-ésimo coeficiente de regresión que puede expresarse como el producto del error estándar residual de la regresión por el factor de inflación de varianza (VIF).

Ahora que tenemos la base de la colinealidad debemos de enfocarnos en las razones por las que es necesario este análisis. En primera instancia: tenemos los problemas que esto nos ocasiona a nuestro modelo de regresión lineal, tal como:

Comentado [JV8]: No solamente nos interesa que las variables regresoras no correlacionen. También nos interesa que los residuos no correlacionen; si falla este supuesto, indica que hay algún patrón que nuestro modelo no está recogiendo. Tendrían que haber mencionado aquí la prueba de Durbin-Watson.

- La interpretación de los coeficientes ya no es posible.
- Los coeficientes de las variables pueden ser incorrectos en cuanto a su magnitud y signo.
- El aumento del error estándar puede llevar a rechazar las pruebas de hipótesis individuales.

(Ballesteros Torres, Perez Acosta, & Sarmiento Ibarra, 2023)

3.8 Prueba de bondad de ajuste - Kolmogorov-Smirnov y Anderson-Darling

Para realizar la prueba de hipótesis H_0 : Una muestra aleatoria $X_1, X_2, \dots, X_n,$ viene de una población con distribución $F_0(x, \theta)$, una distribuci´on continua completamente especificada, donde θ es el vector de par´ametros de la distribución F0 (que se deben estimar que en caso de que sean desconocidos), se puede usar el siguiente procedimiento:

1. Ordenar los valores x_i para $i = 1,2, \dots, n$, de manera ascendente, $X_1 < X_2 < \cdots < X_n$.
2. Calcular $Z_i = F_0\left(x_{(i)}, \theta\right)$ para $i = 1,2, \dots, n$.
3. En caso de querer realizar la prueba de Kolmogorov-Smirnov, se debe calcular el siguiente:

$$D^+ = max\left(\frac{i}{n} - Z_{(i)}\right);\ D^- = max\left(Z_{(i)} - \frac{i-1}{n}\right)$$
$$D = max(D^+, D^-)$$

En caso de querer realizar la prueba de Anderson-Darling, se debe calcular el siguiente estadístico:

$$A^2 = -\frac{\Sigma\{(2i-1)[\ln(Z_i) + \ln(-Z_{n+1-i})]\}}{n}$$

Realizando los ajustes para la cola superior en caso de ser necesario.
(Delgado Rivera & Ortega Vergara, 2023)

3.9 ANOVA

El análisis de varianza consiste en buscar una media de cuanta variabilidad esta expresada en nuestro modelo para lo cual descompone la varianza de la variable respuesta en dos partes:

$$Y_i - \bar{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)$$
$$\Longrightarrow \underset{n}{(Y_i - \bar{Y})^2} = \underset{n}{\left(Y_i - \hat{Y}_i\right)^2} + \underset{n}{\left(\hat{Y}_i - \bar{Y}\right)^2} + 2\underset{n}{\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right)}$$
$$\Longrightarrow \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=i}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right)$$
$$\Longrightarrow \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=i}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}\left(\hat{Y}_i - \bar{Y}\right)^2$$
$$\Longrightarrow SST = SSE + SSR$$

Donde las siglas SS significan "*Sum of squares*" y el termino cruzado vale cero (Darper & Smith, 1998).

El termino SST es la varianza sin ningún modelo aplicado en el contexto de regresión lineal simple, y es la varianza del modelo reducido en el contexto de análisis de varianza; en regresión lineal simple el modelo completo es $Y = \beta_0 + \beta_1 + \epsilon$ y el modelo reducido es $Y = \beta_0 + \epsilon$. El siguiente termino SSE es la suma de cuadrado de los errores; En todo el modelo que contemple error se encontrará presente, en el caso de que $SSE = 0$, quiere decir que el modelo se ajustará perfectamente y no tendrá error. Por último $SSR = SST - SST$ se puede interpretar como la cantidad de información que se puede explicar por el modelo completo, pero no por el modelo reducido (Soriano, 2016).

Fuente	SS	DF	MS=SS/DF	F
Modelo completo	$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	$p-1$	$\frac{SSR}{p-1}$	$\frac{MSR}{MSE}$
Error	$SSE = \sum_{i=i}^{n} \left(Y_i - \hat{Y}_i\right)^2$	$n-p$	$\frac{SSE}{n-p}$	
Modelo reducido	$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$n-1$	$\frac{SST}{n-1}$	

F sirve para determinar la significancia de la regresión; en términos de pruebas de hipótesis esto seria: $H_0: \beta_i = 0 \quad i = 1, \dots, p$ contra $H_1: \beta_i \neq 0$ para algún i. Esta prueba es llamada prueba de significancia y como vemos, compara un modelo contra otro (Soriano, 2016).

3.10 Error puro

Según (Draper, 1998), la suma de los cuadrados de los errores residuales es igual a la suma de los cuadrados de las desviaciones de los valores observados de su media, menos la suma de los cuadrados de las desviaciones de los valores predichos de su media. Es decir, si llamamos a los errores residuales e_i, a los valores observados y_i, a los valores predichos $\hat{y}_i$ y a la media general de los valores observados y, la suma de cuadrados de los errores residuales es:

$$SCE = (e_i) = \sum (y - \hat{y}_i)^2 = \sum (y_i - y)^2 - \sum (\hat{y}_i - y)^2$$

- SCE = Suma de Cuadrados del Error
- y_i son los valores observados de la variable dependiente
- $\hat{y}_i$ son los valores predichos por el modelo de regresión

(Mateo Cardenas, Garcia Perez, & Alvarez Borja, 2023)

3.11 Prueba de falta de ajuste

Según (Peón, 2011) está diseñada para evaluar si una relación curvilíneal podría ajustar mejor a los datos que un modelo lineal. Para ello la SCE se descompone en dos partes (Mateo Cardenas, Garcia Perez, & Alvarez Borja, 2023):

- El componente de error puro.
- El componente de falta de ajuste.

Estos dos componentes son utilizados para construir un estadístico de prueba F particular con el fin de contrastar la hipótesis siguiente:H_o: $la\ relación\ es\ lineal\ vs\ H_1$: $la\ relación\ no\ es\ lineal$, (Peón, 2011).

La fórmula para calcular está prueba es:

$$F = \frac{(SCE_R - SCE_F) \cdot gl_F}{SCE_F \cdot (gl_{R-}gl_F)}$$

Dónde:

SCE_R= Suma de cuadrados del error del modelo reducido.

SCE_F= Suma de cuadrados del error del modelo completo.

gl_R= Grados de libertad del modelo reducido-

gl_F = Grados de libertad del modelo completo.

3.12 Medidas de forma y diagramas explicativos

Toda esta sección se tomó de (Bembibre, 2009) excepto la implementación en R.

En estadística, un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. En el eje vertical se representan las frecuencias, y en el eje horizontal los valores de las variables, normalmente señalando las marcas de clase, es decir, la mitad del intervalo en el que están agrupados los datos.

Interpretación:

Comentado [JV9]: Una interpretación de gráficas sin gráficas es suigéneris.

- Sesgo <0, indica que existen más valores agrupados a la derecha de la distribución.
- Sesgo=0, indica una distribución normal.
- Sesgo>0, indica que existen más valores agrupados a la izquierda de la distribución.
- Curtosis<3, implica una distribución platicúrtica.
- Curtosis=3, implica una distribución mesocúrtica.
- Curtosis>3, implica una distribución leptocúrtica.

Implementación en R

```
library(moments)   # Para hallar las medidas de forma
skewness(x)    #A) Sesgo
kurtosis(x)    #B) Curtosis
```

(Llinás Solano, 2014)

Diagrama caja bigote: los diagramas de Caja-Bigotes (boxplots o box and whiskers) son una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría. Para su realización se representan los tres cuartiles y los valores mínimo y máximo de los datos, sobre un rectángulo, alineado horizontal o verticalmente. Una gráfica de este tipo consiste en una caja rectangular, donde los lados más largos muestran el recorrido intercuartílico. Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana y por lo tanto su relación con los cuartiles primero y tercero(recordemos que el segundo cuartil coincide con la mediana (Jiménez J. 28 de noviembre de 2019).

Comentado [JV10]: Otra explicación de una gráfica sin una mínima ilustración.

4 Metodología

Para ello, se llevará a cabo la recopilación y presentación de datos sobre la demanda interna de combustibles, así como la emisión anual de gases de efecto invernadero. Posteriormente, se construirá un modelo de regresión lineal múltiple que permita analizar la relación entre estas variables, siendo los combustibles los regresores y las emisiones nuestra variable de respuesta. Se verificarán los supuestos del modelo y, en su caso, se realizarán correcciones para obtener un modelo ajustado y válido.

Este proyecto se enmarca en un contexto actual de creciente preocupación por el medio ambiente, donde se busca promover la reducción de emisiones de gases de efecto invernadero.

Algunos aspectos que resaltar son que las unidades de demanda de Gas L.P., gas natural, gasolina, combustóleo y turbosina se miden en miles de barriles diarios, mientras que el coque de petróleo y las emisiones de CO_2 se miden en miles de toneladas.

4.1 Planteamiento

En México, la emisión de gases de efecto invernadero (GEI) es un problema ambiental importante que afecta la calidad del aire y contribuye al cambio climático global. La quema de combustibles fósiles es una de las principales fuentes de emisiones de CO_2, uno de los gases de efecto invernadero más importantes. En este contexto, se plantea como objetivo desarrollar un modelo de regresión lineal múltiple que relacione la demanda interna de diferentes tipos de combustibles (turbosina, gas, gas l.p., combustóleo, gasolina, etc.) con la emisión anual de CO_2 en México. Para alcanzar este objetivo, es necesario recolectar datos precisos y confiables sobre la demanda interna de combustibles en México y la cantidad de CO_2 emitida por cada tipo de combustible.

4.2 Definición de la población

Definimos a la población como los miles de barriles diarios demandados en México anualmente de Gas Natural, Gas L.P., Combustóleo, Diesel, Turbosina. Para el Coque de Petróleo y las Emisiones de CO_2 la población consiste en miles de toneladas demandadas y emitidas en territorio nacional.

4.3 Determinación del grado de precisión deseado

Utilizaremos un grado de precisión del 95%, como media estándar para el análisis estadístico para equilibrar la precisión y la confianza en los resultados de la prueba. Este rango proporciona una probabilidad adecuada de que los resultados sean precisos y, al mismo tiempo, permite una detección efectiva de diferencias estadísticas significativas.

4.4 Recogida de datos

Todos los datos fueron recogidos de la página de la Secretaría de Energía (Secretaria de Energia, 2021) , en lo que respecta a la demanda interna de combustibles, los datos relacionados a las emisiones fueron obtenidos desde “Our Wolrd in Data” (Our World in Data, 2022), que recopila y aglomera la información de distintos organismos tanto gubernamentales y como organizaciones civiles, en particular solicitamos la información de México.

4.5 Estadística descriptiva

Antes de comenzar el análisis de la relación entre el consumo de combustibles y las emisiones de gases en México, es importante tener una comprensión clara de los datos disponibles. En este sentido, la estadística

Comentado [JV11]: ¿Qué es ello?

Comentado [JV12]: Un problema se plantea por medio de una pregunta. Se establece un contexto geográfico (eso sí lo hacen) y un contexto temporal; aquí habría sido importante indicar el periodo de tiempo en el que tomaron sus datos.

Comentado [JV13]: Aunque describen estos elementos, en realidad no hay un ejercicio de muestreo.

Comentado [JV14]: ¿De qué periodo hablan?

descriptiva es una herramienta valiosa para resumir y presentar los datos de manera significativa. En este apartado, se utilizará la estadística descriptiva para examinar las características fundamentales de los datos, incluyendo medidas de tendencia central, dispersión y forma de la distribución. De esta manera, podremos obtener una visión general de las principales tendencias y patrones en los datos, lo que nos permitirá avanzar con mayor confianza en nuestro análisis de regresión lineal.

Empezamos contrastando los datos de demanda anual de todos los combustibles con el paso de los años, añadiendo también el gráfico de líneas de las emisiones, todo esto con las funciones dentro de la librería ggplot2 en R.

Para efectos de visualización se convierten los valores de las demandas a su logaritmo base 10, como modo de reducir el rango de distribución de los valores.

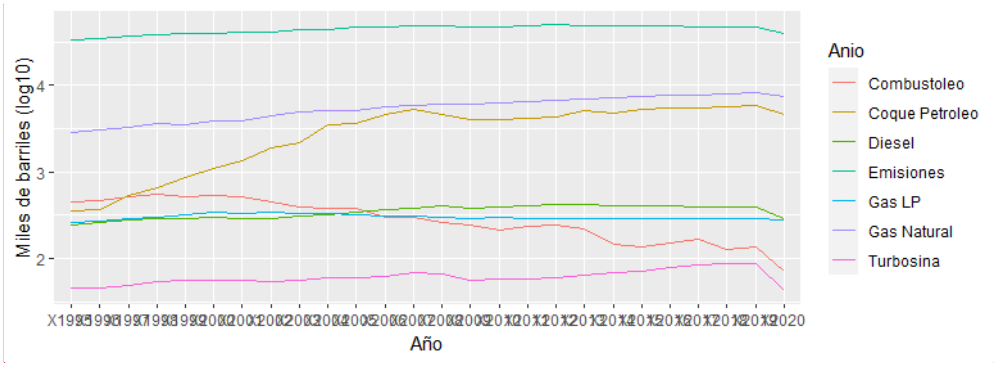


Tabla 1 Demanda de combustibles

A excepción del combustóleo, todos los combustibles muestran un crecimiento progresivo en su demanda. Del mismo modo se percibe un aumento de las emisiones de gases de efecto invernadero. Al tiempo que se confirman las razones para investigar una posible relación lineal entre estos (combustibles y emisiones), también se evidencia el impacto de la pandemia, pues en el 2020 hay un cambio brusco en la tendencia de crecimiento, llegándose inclusive a un mínimo histórico en el caso de la turbosina.

Ahora se obtienen otras medidas descriptivas ahora sobre cada uno de los combustibles

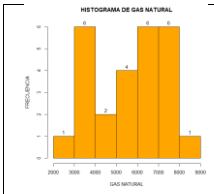
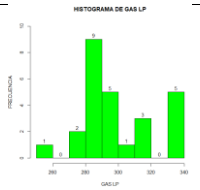
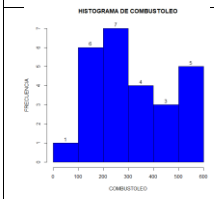
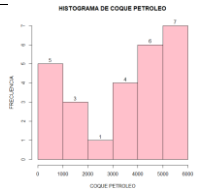
Comentado [JV15]: El título debe ir en la gráfica. Y no sería Tabla, sino Figura.

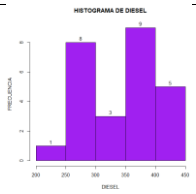
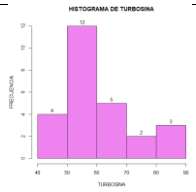
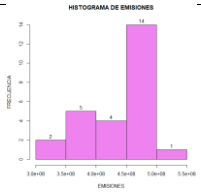
```
 Gas.Natural       Gas.LP        Combustoleo
Min.   :2872   Min.   :259.4   Min.   : 69.53
1st Qu.:4043   1st Qu.:282.8   1st Qu.:175.89
Median :6015   Median :290.6   Median :274.75
Mean   :5677   Mean   :298.2   Mean   :309.67
3rd Qu.:7145   3rd Qu.:316.7   3rd Qu.:448.44
Max.   :8158   Max.   :337.4   Max.   :541.81
Coque.Petroleo       Diesel        Turbosina       Emisiones
Min.   : 347.6   Min.   :239.5   Min.   :40.98   Min.   :331596600
1st Qu.:1471.3   1st Qu.:288.4   1st Qu.:54.41   1st Qu.:399724265
Median :4101.0   Median :367.7   Median :56.95   Median :464807055
Mean   :3439.0   Mean   :345.3   Mean   :60.41   Mean   :442410277
3rd Qu.:4976.5   3rd Qu.:396.4   3rd Qu.:66.12   3rd Qu.:479720045
Max.   :5817.5   Max.   :420.3   Max.   :86.69   Max.   :501568830
```

Se perciben diferencias entre la media y mediana entre todos los combustibles, sugiriendo diversas tendencias y la posibilidad de datos extremadamente altos o bajos en los datos, siendo el caso más evidente el del coque petróleo, pues al tener una mediana mayor que la media, nos podría indicar que hay valores extremadamente altos que están sesgando los datos hacia la derecha, lo cual también podemos constatar con el grafico de líneas anterior, confirmando que fue el del crecimiento más violento. Por otro lado, los máximos y mínimos revelan la gran variabilidad en la demanda de combustóleo y coque petróleo, contrarios al Gas LP, siendo este el más constante con los años.

Comentado [JV16]: Hablaron en su marco teórico del diagrama de cajas, aquí era una inmejorable oportunidad de usarlo.

Histogramas. Dado que, los datos de cada muestra (variables independientes y variable dependiente) son muy semejantes, se construyen histogramas para visualizar la distribución de cada variable.

	Se logra observar que existe una mayor concentración de datos entre 6000 y 8000 barriles.
	Se logra observar que existe una gran concentración de datos entre 280 y 300.
	Se logra observar que existe una mayor concentración de datos entre 200 y 300 barriles.
	Se logra observar que existe una mayor concentración de datos entre 5000 y 6000 barriles.

	Se logra observar que existe una mayor concentración de datos entre 350 y 400 barriles.		Se logra observar que existe una mayor concentración de datos entre 50 y 60 barriles.
	Se logra observar que entre 4.5e+08 y 5.0e+08 de GEI existe una mayor concentración de datos.		

De acuerdo con cada histograma, se logra visualizar que la distribución del gas natural, diésel, coque petróleo y las emisiones de GEI tienen un sesgo negativo, implica que los datos se concentran a la derecha de la distribución. Por otro lado, el gas LP, el combustóleo y la turbosina tienen un sesgo positivo, implica que los datos se concentran a la izquierda de la distribución. Para verificar dichas afirmaciones obtenemos el sesgo de cada distribución en R. Se tiene que instalar e importar la librería moments y se emplea la función skewness(variable).

Obtención del sesgo en R

```
> print("Sesgo")
[1] "Sesgo"
> skewness(Datos_CyE$`Gas Natural`)
[1] -0.2091705
> skewness(Datos_CyE$`Gas LP`)
[1] 0.5068541
> skewness(Datos_CyE$`Combustoleo`)
[1] 0.1683902
> skewness(Datos_CyE$`Diesel`)
[1] -0.2899633
> skewness(Datos_CyE$`Coque Petroleo`)
[1] -0.5015389
> skewness(Datos_CyE$`Turbosina`)
[1] 0.7298217
> skewness(Datos_CyE$`Emisiones`)
[1] -0.7650698
```

Obtenemos el coeficiente de curtosis de cada R.

```
> print("Curtosis")
[1] "Curtosis"
> kurtosis(Datos_CyE$`Gas Natural`)
[1] 1.719053
> kurtosis(Datos_CyE$`Gas LP`)
[1] 2.123821
> kurtosis(Datos_CyE$`Combustoleo`)
[1] 1.621293
> kurtosis(Datos_CyE$`Diesel`)
[1] 1.546348
> kurtosis(Datos_CyE$`Coque Petroleo`)
[1] 1.686792
> kurtosis(Datos_CyE$`Turbosina`)
[1] 2.962379
> kurtosis(Datos_CyE$`Emisiones`)
[1] 2.331936
```

De acuerdo con (Llinás Solano, 2014), el coeficiente de curtosis de todas las variables es menor que tres, implica que cada distribución es “platicúrtica”. Es decir, existe una baja concentración de los datos entorno a la media de cada distribución.

Tabla 2 Histogramas

Diagrama caja-bigote. Toda la sección de caja-bigote se tomó de Jiménez J. (28 de noviembre de 2019)

El diagrama de caja-bigote nos ayuda a identificar al igual que los histogramas el sesgo de una distribución y, además, gracias a este diagrama podemos identificar si existen o no datos atípicos en la distribución. Para obtener el diagrama de caja-bigote en R, se emplea la función “boxplot(variable)”, no se necesita instalar ninguna librería.

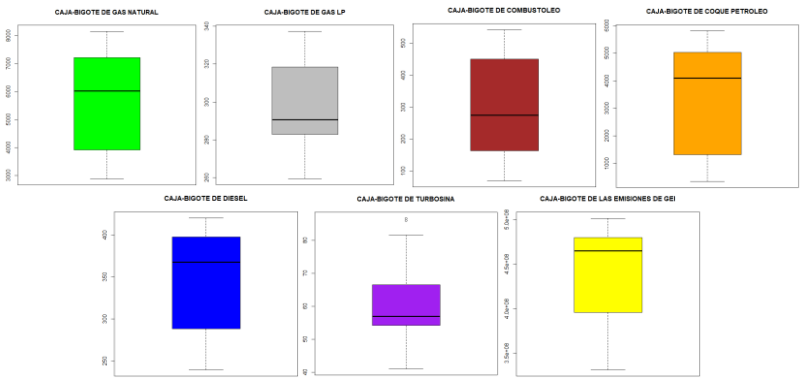


Tabla 3 Diagramas caja-bigote

Dado los diagramas caja-bigote respecto a cada muestra, se tiene que sólo en la muestra de "turbosina" existen datos atípicos. Al igual que en los histogramas anteriores, se logra visualizar el sesgo de las distribuciones: sesgo negativo, distribución de gas natural, coque petróleo, diésel y emisiones; sesgo positivo, distribución de gas LP, combustóleo y turbosina. Según Cano E. (7 de diciembre de 2022): "Para identificar los datos atípicos en R empleamos la función boxplot.stats(variable), no se necesita instalar ninguna librería".

```
> boxplot.stats(Datos_CyE$Turbosina)
$stats
[1] 40.97726 54.22564 56.94744 66.51153 81.59048

$n
[1] 26

$conf
[1] 53.14049 60.75439

$out
[1] 86.05886 86.69035
```

Por ende, se logra observar que los datos atípicos de la distribución "turbosina" son: 86.05886 y 86.69035

4.6 Modelo de regresión lineal múltiple

Procedemos a cargar nuestros datos en R y a realizar nuestro modelo de regresión lineal múltiple con todas nuestras variables predictoras.

Comentado [JV17]: Queda un vacío aquí. ¿Usaron el logaritmo o los datos brutos?

```
cvse <- read.csv("Combustibles vs Emisiones CO2.csv")
cvse$Año <- NULL
cvse
```

Gas.Natural	Gas.LP	Combustóleo	Coque.Petróleo	Diesel	Turbosina	Emisiones
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
2872.419	259.4498	449.68284	347.5632	239.4692	44.36860	331596600
3057.779	270.0435	470.31951	368.2673	256.0856	45.14273	345766100
3284.491	282.2953	508.83670	528.7979	275.1892	47.22758	368644800
3584.940	294.2495	541.81320	656.4874	288.4666	52.45232	388356900
3541.574	318.8341	518.01446	853.1305	286.1616	55.34574	390516450
3888.525	337.3719	534.05845	1101.5017	295.8674	55.53471	396066370
3912.968	331.5780	512.56184	1327.0524	288.3203	55.27270	410697950
4434.455	336.8760	444.71725	1904.2355	282.3807	53.31633	411967680
4858.553	332.5130	396.97875	2201.4005	307.1355	54.22564	437757540
5167.396	334.3367	377.34001	3550.9720	318.5484	57.75833	438856700
5087.474	318.4820	383.05341	3623.2222	336.5445	58.65999	463993180
5672.861	311.2469	301.29343	4623.1163	359.8478	61.17364	476565220
5925.819	306.9824	293.75041	5183.8607	375.5268	67.89052	479786800
6109.812	297.1739	255.75183	4603.8798	399.5350	64.95528	492979680
6104.015	286.6893	242.19503	3968.5971	378.9105	54.96069	475903040
6340.863	292.8504	213.40890	3989.8176	390.1899	55.81887	463782500
6512.228	290.3949	231.01727	4212.2324	401.1500	56.13654	484164860
6678.438	290.8867	238.39011	4358.5383	420.3089	59.30342	501568830
6952.351	286.4753	215.18090	5026.0339	413.9497	62.23857	495485200
7209.332	287.1899	146.30470	4827.7213	410.1563	66.51153	484113700
7504.091	282.9856	134.32854	5260.1219	404.5850	70.77395	479519780
7618.664	282.5046	147.55486	5421.3366	397.7493	76.24776	479789760
7611.943	282.8156	163.38205	5539.2788	392.1659	81.59048	465620930
7968.286	282.8858	126.49194	5550.7151	387.3856	86.05886	475268930
8158.406	282.2458	135.34590	5817.5252	386.1490	86.69035	472191500
7539.008	274.9539	69.52969	4568.8400	286.0782	40.97726	391706200

Tabla 3 Carga de datos en R

Así, nuestro modelo es descrito en la siguiente tabla.

```
modelo_completo <- lm(Emisiones ~ ., data = cvse)
summary(modelo_completo)
```

```
Call:
lm(formula = Emisiones ~ ., data = cvse)

Residuals:
      Min        1Q    Median        3Q       Max
-16477557  -3649518    582766   3716140  14085530
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    44074734   36647942   1.203  0.24388
Gas.Natural       -3244       6981  -0.465  0.64742
Gas.LP           452361     131999   3.427  0.00283 **
Combustóleo       47272      81917   0.577  0.57066
Coque.Petróleo    13186       3978   3.315  0.00364 **
Diesel           707034      61438  11.508 5.24e-10 ***
Turbosina       -368944     271482  -1.359  0.19006
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7926000 on 19 degrees of freedom
Multiple R-squared:  0.9806,    Adjusted R-squared:  0.9745
F-statistic: 160.3 on 6 and 19 DF,  p-value: 3.125e-15
```

Tabla 4 Modelo con todas las variables

Se observa que las variables Diesel, Gas LP y Coque Petróleo tienen una relación estadísticamente significativa con la emisión de gases de efecto invernadero, ya que tienen *p-values* menores que 0.05. En cambio, las variables Combustóleo Gas Natural y Turbosina no son significativas, ya que tienen *p-values* mayores que 0.05. Además, el estadístico R (coeficiente de Pearson) indica que hay una relación positiva fuerte. El estadístico F (p-value: 3.125e-15) indica que el modelo ajustado es significativo en su conjunto.

Comentado [JV18]: Debieron hablar de R cuadrada ajustada. Nos dice el porcentaje de variabilidad alrededor de la media que recogió nuestro modelo.

No obstante, para evitar problemas con la interpretación de cada variable en el modelo, es importante verificar la no colinealidad de las variables. Por lo tanto, se procede a detectar la multicolinealidad de cada variable con el factor de inflación de varianza. Si el factor de inflación de varianza es mayor a 5, se retira dicha variable. El objetivo de eliminar la multicolinealidad en el modelo es mejorar la precisión de los coeficientes y asegurar que los resultados sean precisos e interpretables.

```
print(vif(modelo_completo))
```

```
  Gas.Natural        Gas.LP    Combustóleo Coque.Petróleo        Diesel 
    54.906329      3.481484      61.272298      22.937941       4.985642 
     Turbosina 
     4.250012 
```

Tabla 5 Detección de multicolinealidad con FIV

De acuerdo con el criterio establecido, retiramos de nuestro modelo las variables Gas Natural, Combustóleo y Coque de Petróleo de nuestro modelo.

Comentado [JV19]: Tendrían que haber explicado el efecto de eliminar del modelo al coque de petróleo, ya que era una variable significativa. ¿Por qué no probar qué pasaba con la colinealidad eliminando solamente a las otras dos variables?

```
modelo_reducido <- lm(Emisiones ~ Diesel + Gas.LP + Turbosina, data = cvse)
summary(modelo_reducido)
```

```
Call:
lm(formula = Emisiones ~ Diesel + Gas.LP + Turbosina, data = cvse)
Residuals:
     Min       1Q   Median       3Q      Max 
-24139676  -4608126    348172   5101036  18410378 
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -14740320   32575287  -0.453    0.655    
Diesel         874096      49654  17.604 1.88e-14 ***
Gas.LP         530736      92236   5.754 8.67e-06 ***
Turbosina      -49077     233374  -0.210    0.835    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 10070000 on 22 degrees of freedom
Multiple R-squared:  0.9638,   Adjusted R-squared:  0.9589 
F-statistic: 195.3 on 3 and 22 DF,  p-value: 5.309e-16
```

Tabla 6 Modelo sin multicolinealidad

En el modelo ‘modelo_reducido’, se aprecia que variables, ‘Diesel’ y ‘Gas.LP’ son significativas dado que su *p-value* es menor a 0.05, mientras que la variable ‘Turbosina’ y el intercepto no son significativos. El *p-value* de F indica que este modelo también es significativo y el estadístico. El estadístico *Ajusted R-squared* indica que el **95.89%** de la variación de la variable dependiente es explicado por las variables Diesel, Gas L.P. y Diesel. (Sanjuán, 2018)

Ahora procedemos con una función en R llamada step que seleccionará el mejor modelo basado en las fórmulas AIC. (RDocumentation, s.f.)

```
step (
   modelo_reducido
)-> mejor_modelo
summary(mejor_modelo)
```

```
Call:
lm(formula = Emisiones ~ Diesel + Gas.LP, data = cvse)
Residuals:
      Min        1Q    Median        3Q       Max
-24216461  -4497291    138718   4414908  18454330
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14892227   31883423  -0.467    0.645
Diesel         866865      35069  24.719  < 2e-16 ***
Gas.LP         529677      90164   5.875 5.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9856000 on 23 degrees of freedom
Multiple R-squared:  0.9637,    Adjusted R-squared:  0.9606
F-statistic: 305.6 on 2 and 23 DF,  p-value: < 2.2e-16
```

Nuestro modelo final, al que más adelante le haremos pruebas, queda descrito por las variables *Diesel* y *Gas.LP*. Ambas variables son significativas en nuestro modelo, no así el intercepto. El estadístico *Ajusted R-squared* indica que el **96.06%** de la variabilidad de la emisión de CO_2 es explicada por las variables independientes del modelo. Finalmente, el *p-value* del estadístico *F* indica que el modelo es significativo. Procederemos a realizar los supuestos que debe cumplir nuestro modelo.

4.7 ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Diesel	1	5.601902e+16	5.601902e+16	576.70454	8.664619e-18
Gas.LP	1	3.352226e+15	3.352226e+15	34.51049	5.483915e-06
Residuals	23	2.234138e+15	9.713644e+13	NA	NA

En este caso, el análisis indica que tanto la variable Diesel como la variable Gas LP son fuentes significativas de variación en el modelo, ya que ambos tienen valores *F* altos y *p-values* muy bajos (< 0.05). Es decir, existe una relación estadísticamente significativa entre estas variables y la variable de respuesta (Darper & Smith, 1998). La suma de los cuadrados de los residuales representa la variabilidad no explicada por el modelo (Marco Sanjuán, 2018), y en este caso es muy alta.

Comentado [JV20]: Es Draper

Comentado [JV21]: Pues sí, pero va a acorde con la escala de los datos.

4.8 Residuos con media cero

Realizaremos la prueba *t* para verificar que la media de los residuales es cero.

```
t.test(residuales)
```

```
        One Sample t-test


data:  residuals
t = -3.4778e-16, df = 25, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -3818284  3818284
sample estimates:
```

```
   mean of x
-6.447618e-10
```

De la prueba obtenemos que:

- La media de los residuales es -6.447618e-10 (aproximadamente cero). El valor t de la prueba es -3.4778e-16.
- El *p-value* es 1, lo que indica que no hay evidencia significativa para rechazar la hipótesis nula de que la media de los residuales es cero.
- El intervalo de confianza del 95% para la media de los residuales es (-3818284, 3818284).

En otras palabras, los resultados sugieren que no hay evidencia significativa para concluir que la media de los residuales es diferente de cero, lo que sugiere que la suposición de que la media de los residuales es cero es cierta.

4.9 Heterocedasticidad

```
bptest(mejor_modelo)
```

```
        studentized Breusch-Pagan test
data:  mejor_modelo
BP = 4.0436, df = 2, p-value = 0.1324
```

El resultado de la prueba de Breusch-Pagan es una prueba para la heterocedasticidad. El resultado muestra un valor de estadístico de prueba de BP=4.0436 con 2 grados de libertad y un p-valor de 0.1324. Como el *p-value* es mayor que 0.05, no se rechaza la hipótesis nula de homocedasticidad, lo que sugiere que no hay evidencia suficiente para afirmar que hay heterocedasticidad en los residuos del modelo, en otras palabras, no hay suficiente evidencia estadística para rechazar la hipótesis nula de que la varianza es constante.

4.10 Residuos no correlacionados

Realizaremos la prueba de *Durbin-Watson* para probar que los residuos no están correlacionados.

```
dwtest(mejor_modelo)
```

```
        Durbin-Watson test
data:  mejor_modelo
DW = 1.2328, p-value = 0.006097
alternative hypothesis: true autocorrelation is greater than 0
```

Tabla 7 Prueba Durbin-Watson

En este caso, el valor del estadístico *Durbin-Watson* es de 1.2328, lo que sugiere que hay cierta evidencia de correlación positiva en los residuos. El *p-value* asociado a la prueba es de 0.006097, lo que indica que esta evidencia es estadísticamente significativa. En general, se desea que los residuos no presenten correlación, ya que esto puede indicar que el modelo no está capturando correctamente la estructura de los datos.

Comentado [JV22]: Aquí procedía ya hacer alguna corrección al modelo. Quizás alguna de las variables eliminadas no era estadísticamente significativa, pero en conjunto ayudaba al buen ajuste del modelo.

4.11 Residuos normales

```
ad.test(residuales)
```

```
        Anderson-Darling normality test
data:  residuals
A = 0.19924, p-value = 0.8715
```

La prueba de normalidad de *Anderson-Darling* se utiliza para evaluar si los residuos de un modelo siguen una distribución normal. En este caso, los resultados indican que el valor de la estadística A es 0.19924 y el *p-value* es 0.8715. Dado que el *p-value* es mayor que el nivel de significancia de 0.05, no hay suficiente evidencia para rechazar la hipótesis nula de que los residuos siguen una distribución normal. Por lo tanto, podemos concluir que los residuos se distribuyen normalmente.

4.12 Intervalos de confianza

```
confint(mejor_modelo, level = 0.95)
```

	2.5 %	97.5 %
(Intercept)	-80848112.4	51063658.6
Diesel	794318.1	939411.3
Gas.LP	343157.4	716196.1

5 Conclusiones

El modelo de regresión lineal múltiple que se ha presentado parece ser razonablemente bueno. El modelo completo tiene un alto coeficiente de determinación (R-cuadrado) de 0.9806, lo que sugiere que el modelo explica una gran parte de la variabilidad en la variable de respuesta. Además, el estadístico F sugiere que el modelo completo es significativo en su conjunto.

Sin embargo, es importante tener en cuenta que el modelo ha sido reducido eliminando algunas variables debido a la multicolinealidad. Esto significa que el modelo reducido es más simple, pero puede no ser tan preciso como el modelo completo. Además, la prueba de *Durbin-Watson* nos puede llevar a pensar que el modelo no es muy bueno.

Comentado [JV23]: Había que corregir el modelo. Por otro lado, ¿para qué sirve su modelo? Un mínimo interés por pronosticar algo debería plasmarse en su reporte.

6 Bibliografía

Ballesteros Torres, A., Perez Acosta, A., & Sarmiento Ibarra, J. (Abril de 2023). Pruebas de independencia y no correlacion. Mexico: UNAM.

Bembibre, V. (Febrero de 2009). *Definicion de Histograma*. Obtenido de Definicion ABC: https://www.definicionabc.com/tecnologia/histograma.php

Darper, N. R., & Smith, H. (1998). *Applied regression analysis (Vol. 326).* John Wiley & Sons.

Delgado Rivera, A., & Ortega Vergara, A. F. (Abril de 2023). Pruebas de bondad de ajuste. Mexico: UNAM.

Draper, N. R. (1998). Applied regression analysis. En N. R. Draper, *Applied regression analysis*. Nueva York: Wiley.

Hernandez Perez, E. F., & Urquiza Robles, D. I. (Abril de 2023). Homocedasticidad y Heterocedasticidad. *Homocedasticidad y Heterocedasticidad*. Mexico: UNAM.

Llinás Solano, H. (2014). *Introduccion a la estadistica matematica.* Universidad del Norte.

Marco Sanjuán, J. F. (6 de diciembre de 2018). *economipedia*. Obtenido de Suma de cuadrados de los residuos (SCE): https://economipedia.com/definiciones/suma-de-cuadrados-de-los-residuos-sce.html

Mateo Cardenas, J. P., Garcia Perez, S., & Alvarez Borja, E. P. (Abril de 2023). ANOVA y prueba de linealidad. Mexico: UNAM.

Montgomery, D. C. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Our World in Data. (9 de Febrero de 2022). *Our World in Data*. (H. Ritchie, & P. Arriagada, Editores) Obtenido de Annual CO_2 emissions: https://ourworldindata.org/grapher/annual-co2-emissions-per-country?tab=table&time=1891..latest&country=~MEX

Peón, F. V. (2011). *Prueba de falta de ajuste (Lack of fit Test)*. Obtenido de Universidad Autónoma Metropolitana: https://mregresion.files.wordpress.com/2011/10/prueba-falta-de-ajuste.pdf

RDocumentation. (s.f.). Obtenido de step function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step

Sanjuán, F. J. (7 de noviembre de 2018). *economipedia*. Obtenido de economipedia: https://economipedia.com/definiciones/r-cuadrado-ajustado-coeficiente-de-determinacion-ajustado.html#:~:text=En%20palabras%20m%C3%A1s%20simples%2C%20el%20R%20cuadrado%20ajustado,coeficiente%20de%20determinaci%C3%B3n%20sin%20ajustar%20tiende%20a%20aumentar

Secretaria de Energia. (2021). *SIE Informacion Estadistica.* Obtenido de Sistema de informacion Energetica: https://sie.energia.gob.mx/bdiController.do?action=temas

SEMER. (s.f.). *Sistema de Información Energética*. Recuperado el mayo de 2023, de SIE: https://sie.energia.gob.mx/

Soriano, A. (2016). *Regresión múltiple y otras técnicas multivariadas.* Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas.

Vega Ayala, L., Diaz Barriga, R., & Garcia Totozintle, S. (Abril de 2023). No colinealidad y variables influyentes. *No colinealidad y variables influyentes*. Mexico: UNAM.

Walpole, R. E., & Myers, R. H. (2012). *Probabilidad y estadística para ingeniería y ciencias.* Mexico: PEARSON EDUCACIÓN.

Wooldridge, J. M. (2010). *Introducción a la Econometria: Un enfoque moderno* (4 ed.). Michigan, Michigan, Estados Unidos: Michigan State University. Obtenido de https://herioscarlanda.files.wordpress.com/2018/10/wooldridge-2009-introduccic3b3n-a-la-econometrc3ada-un-enfoque-moderno.pdf

Jiménez J. (28 de noviembre de 2019). Detección de outliers.

Criterio	Aspecto por evaluar	Puntaje posible	Puntaje obtenido
Calidad de forma y secciones iniciales del reporte de proyecto			
A0	El reporte está libre de plagio porque incluye citas a fuentes documentales en cada párrafo del marco teórico y tabla automática de referencias, ambas de conformidad con las normas APA.	**Requisito**	-5
A1	El reporte está correctamente redactado y sin faltas de ortografía.	-0.5/0	0
A2	El reporte incluye carátula, tabla de contenido automático, tabla automática de tablas, tabla automática de figuras y numeración de páginas.	-0.5/0	-0.5
A3	El título del proyecto es descriptivo con hasta 12 palabras.	0/0.5	0
A4	El reporte de proyecto cuenta con un resumen de no más de 250 palabras que incluye objetivo, metodología, resultados y conclusiones básicas.	0/0.5	0.5
A5	Se incluyen tres palabras clave.	0/0.5	0
A6	La introducción del escrito incluye tema, objetivo y estructura del reporte. Responde a las preguntas ¿qué? (concepto de lo que se quiere hacer y antecedentes), ¿por qué? (problema a resolver) y ¿para qué? (objetivo del proyecto).	0/0.5	0
Metodología			
A7	Se incluye el planteamiento del problema de forma clara, precisa y accesible, con detalles del contexto en el que se enmarca.	0/2	2
A8	Se describen con detalle los elementos de muestreo: población, nivel de error, datos a incluir en la recolección, definición de la unidad de muestreo, elección del método de muestreo, cálculo del tamaño de muestra, recogida y presentación de datos.	0 a 5	1
A9	El reporte de proyecto incorpora en anexos un marco teórico de máximo 15 cuartillas con los elementos conceptuales y de aplicación del modelo de regresión lineal múltiple. Dicho marco teórico es citado en el desarrollo cada vez que se requiere un resultado teórico para soportar los procedimientos y ecuaciones utilizadas.	0 a 10	8
Resultados			
A10	El reporte argumenta con claridad y objetividad el desarrollo del proyecto, incluye tablas de síntesis y figuras (ambas con título) que se comentan y relacionan en la redacción de la sección.	0 a 2	1.8
Conclusiones			
A11	Hay una conclusión por cada resultado presentado y en conjunto responden al planteamiento del problema.	0 a 2	1.8
Dominio del contenido			
A12	El desarrollo del proyecto refleja el dominio del estudiante sobre los aspectos conceptuales y de aplicación del modelo de regresión lineal múltiple, incluyendo la comprobación de los supuestos que lo sustentan, las pruebas de hipótesis involucradas y la adecuación de incorrecciones.	0 a 10	8
A13	El desarrollo del proyecto se apoya en el software R en todos los aspectos estadísticos incluidos.	0 a 2	2