UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

LICENCIATURA EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

ESTADÍSTICA II

Análisis de los Residuos Sólidos Urbanos en México mediante regresión lineal múltiple

PRESENTA
CONTRERAS MUÑIZ JORGE EDUARDO - 100%
MARTÍNEZ CAMACHO FERNANDO - 100%
RAMIREZ ALMAZAN EMILIO - 100%
ALFARO CAMPOS EDGAR MAXIMILIANO - 100%
LEÓN RODRÍGUEZ GUSTAVO - 100%

PROFESOR
JAIME VERGARA PRADO

México, 9 de mayo 2024

Índice

Tabla de contenido

Resumen

En este proyecto se llevó a cabo un Análisis de Residuos Sólidos Urbanos en México durante el año 2020, para ello se utilizaron los datos más recientes (2020) sobre los residuos sólidos urbanos en México en dicho año (CIBNOR,2020). Para este fin, también se utilizaron los datos sobre el Producto Interno Bruto (INEGI,2020), Turismo (DATATUR,2020), Total de Población (INEGI,2020) e índice de Pobreza en México (CONEVAL,2020).
Cabe destacar que en este periodo de tiempo se desencadenó una pandemia a nivel global a causa del coronavirus 2019 (COVID-19) donde se esperaba que los datos de turismo en el país tuvieran algún impacto a la hora de crear el modelo de regresión lineal múltiple, sin embargo, como se podrá ver en la realización del proyecto, esto no fue así.

Palabras clave

Regresión lineal múltiple, RSU, relación.

Introducción

Para el Gobierno de México (2017) los residuos sólidos urbanos se generan en las casas habitación, que resultan de la eliminación de los materiales que se utilizan en actividades domésticas, de los productos que se consumen, de envases, embalajes o empaques. Además de los residuos provenientes de cualquier otra actividad dentro de establecimientos o en la vía pública y los resultantes de la limpieza de las vías y lugares públicos. Es necesario conocer qué son los residuos sólidos urbanos, uno de los principales problemas en la gestión de residuos sólidos urbanos en México es la alta dependencia de los vertederos como método de disposición final. Muchos de estos vertederos carecen de medidas adecuadas de control ambiental, lo que puede provocar contaminación del suelo, del agua y del aire, así como impactos negativos en la salud de las comunidades circundantes.

La generación de residuos sólidos en México es considerable, con una estimación de más de 100 millones de toneladas al año, según datos de la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT). Esta cifra se ve impulsada por diversos factores, como el aumento en el consumo, la falta de cultura de separación de residuos y la escasez de infraestructura adecuada para su gestión.

El manejo inadecuado de los RSU conlleva impactos negativos tanto ambientales como sociales. La contaminación del suelo, agua y aire, la proliferación de enfermedades, la degradación del paisaje urbano y la afectación a la flora y fauna son solo algunas de las consecuencias directas de una gestión deficiente de estos residuos.

El proyecto tiene como objetivo abordar el problema de los residuos sólidos urbanos (RSU) en México de manera integral, buscando soluciones que mejoren significativamente la gestión y reduzcan los impactos negativos asociados con su manejo inadecuado.

Marco Teórico

Para la realización de este proyecto fue necesario aplicar los conocimientos adquiridos sobre la situación general de regresión, o regresión lineal múltiple. Los modelos de regresión lineal múltiple son aquellos modelos de regresión en donde intervienen más de una variable regresora (Montgomery et al., 2006).

En un modelo de regresión lineal múltiple es posible relacionar la respuesta con k regresores, o variables predictoras, de tal forma que el modelo de regresión lineal múltiple con k regresores quedaría de la siguiente forma (Montgomery et al., 2006):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \mathcal{E}$$

En donde a β_j, j = 0,1,..., k se le conoce como coeficientes de regresión.

Para estimar los coeficientes de regresión de la ecuación anterior se aplica el método de mínimos cuadrados. Para ello es posible escribir el modelo muestral de regresión de la siguiente forma (Montgomery et al., 2006):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \mathcal{E}_i$$

La función de mínimos cuadrados sería:

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \Sigma \left(y_i - \beta_0 - \Sigma \beta_j x_{ij} \right)^2$$

Para que la función S se minimice, los estimadores de $\beta_0, \beta_1, \ldots, \beta_k$ deben de cumplir lo siguiente Montgomery et al., 2006):

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

Ilustración 1. Condición 1 para minimización

y

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

Ilustración 2. Condición 2 para minimización

Al simplificar estas ecuaciones, se obtienen las ecuaciones normales de mínimos cuadrados (Montgomery et al., 2006):

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

Ilustración 3. Ecuaciones normales de mínimos cuadrados

La solución a estas ecuaciones serán los estimadores por mínimos cuadrados $\beta_0, \beta_1, \ldots, \beta_k$ (Montgomery et al., 2006).
Como se puede observar, los modelos de regresión múltiple pueden llegar a ser complicados de expresar, por lo que se puede utilizar la expresión matricial para manejar fácilmente estos modelos. Con notación matricial el modelo queda expresado de la siguiente forma (Montgomery et al., 2006):

$$y = X\beta + \mathcal{E}$$

en donde:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Ilustración 4. Componentes matriciales

Al igual que la forma anterior, se desea determinar el vector β de estimadores de mínimos cuadrados que minimice lo siguiente (Montgomery et al., 2006):

$$S(\beta) = \Sigma\mathcal{E}_i^2 = \mathcal{E}'\mathcal{E} = (y - X\beta)'(y - X\beta) = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta = y'y - 2\beta'X'y + \beta'X'X\beta$$

Los estimadores de mínimos cuadrados deben de satisfacer lo siguiente:

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

lo que es simplificado como:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Para que la matriz $(X'X)^{-1}$ exista, es necesario que los regresores sean linealmente independientes, es decir, que ninguna columna en la matriz X sea una combinación lineal de las demás (Montgomery et al., 2006).
El modelo ajustado de regresión correspondiente a los niveles de las variables regresoras x' es (Montgomery et al., 2006):

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

La matriz $X(X'X)^{-1}X'$ se le llama matriz de sombrero, la cual desempeña un papel importante en el análisis de regresión, ya sea para detectar la extrapolación oculta en la regresión múltiple o para realizar comprobaciones en la adecuación del modelo (Montgomery et al., 2006).

Para evaluar la adecuación del modelo utilizamos los estadísticos R cuadrada y R cuadrada ajustada. La R cuadrada aumenta siempre al agregar un regresor al modelo, sin importar el valor de contribución de dicha variable, por lo que es difícil juzgar si el aumento de la R cuadrada nos dice algo importante (Montgomery et al., 2006).

Por otro lado, la R cuadrada ajustada solo aumentará al agregar una variable al modelo si esa adición reduce el cuadrado medio residual. La R cuadrada ajustada penaliza la adición de variables que no son útiles, lo que nos permite evaluar y comparar los posibles modelos de regresión (Montgomery et al., 2006).

Retomaremos ahora conceptos discutidos de la práctica anterior:

Un tema importante para ver es la gráfica de dispersión en el cual Juan Pablo de la Guerra (2015) define "Los Diagramas de Dispersión o Gráficos de Correlación permiten estudiar la relación entre 2 variables. Dadas 2 variables X y Y, se dice que existe una correlación entre ambas si cada vez que aumenta el valor de X aumenta proporcionalmente el valor de Y (Correlación positiva) o si cada vez que aumenta el valor de X disminuye en igual proporción el valor de Y (Correlación negativa)." (Las siete herramientas de la calidad, pág. 6)

"El *ANOVA* es un conjunto de técnicas estadísticas de gran utilidad y ductilidad. Es útil cuando hay más de dos grupos que necesitan ser comparados, cuando hay mediciones repetidas en más de dos ocasiones, cuando los sujetos pueden variar en una o más características que afectan el resultado y se necesita ajustar su efecto o cuando se desea analizar simultáneamente el efecto de dos o más tratamientos diferentes" (Dagnino, 2014, p. 1).

De acuerdo con Paula Elosua (2011) "R es un entorno de programación, análisis estadístico y software gráfico derivado del lenguaje de programación S (Becker, Chambers y Wilks, 1988; Chambers, 1998; Chambers y Hastie, 1992; Venables y Ripley, 2000)." "R dispone de funciones básicas relacionadas con los análisis descriptivos de datos, y de los modelos más complejos y actuales concernientes con los últimos avances en el campo de la estadística, la psicometría, la econometría o el análisis de datos." (Introducción al entorno R, pág 13)

Montgomery et al. (2002) aclara que "el análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Un aspecto esencial del análisis de regresión es la recolección, recopilación o adquisición de datos. Un buen esquema de recolección de datos puede asegurar un análisis simplificado y un modelo de aplicación más general.
Los modelos de regresión se usan con varios fines, que incluyen los siguientes: descripción de datos, estimación de parámetros, control, predicción y estimación". (Introducción al análisis de regresión lineal, pp. 1-9).

Alfonso Novales (2010) define a una variable dependiente como: "variable cuyo comportamiento se pretende explicar, Y_i. También tome los nombres de endógena o variable a explicar(respuesta)" (Análisis de regresión, pág 14)

Alfonso Novales (2010) dice "mientras que la variable denotada por X_i recibe el nombre de variable independiente, exógena o explicativa" y permite controlar el modelo. (Análisis de regresión, pág 14).

Se estará definiendo dos tipos de regresión lineal: simple y múltiple, sin embargo, para este caso particular sólo se estará trabajando con la regresión lineal simple o RLS, y añadiendo el tema de mínimos cuadrados para el ajuste de rectas que posteriormente se estará calculando en R.

En el ámbito del análisis de regresión lineal múltiple, el supuesto de normalidad de los errores es fundamental para garantizar la validez de las inferencias estadísticas. **La prueba de Anderson-Darling** se erige como una herramienta robusta para evaluar el cumplimiento de este supuesto, ya que es una prueba de bondad de ajuste que compara la distribución empírica de los residuos del modelo de regresión lineal múltiple con una distribución normal teórica. El estadístico de Anderson-Darling (A2) cuantifica la divergencia entre ambas distribuciones, donde valores elevados del estadístico indican una mayor discrepancia.

Hipótesis de la Prueba:

- H0 (Hipótesis Nula): Los residuos del modelo de regresión provienen de una población con una distribución normal.
- H1 (Hipótesis Alternativa): Los residuos del modelo de regresión no provienen de una población con una distribución normal.

El proceso empieza extrayendo los residuos, que son la discrepancia entre los valores observados de la variable dependiente y los valores predichos por el modelo. Se ordenan estos residuos de menor a mayor para hacer más fácil el análisis subsiguiente.

Luego, la distribución empírica de los residuos se establece, la cual describe la proporción acumulada de datos dentro de cada intervalo especificado. Al suponer que los residuos deben seguir una distribución normal, se procede a realizar la estimación de una distribución normal teórica. Se caracteriza por una media de cero y una desviación estándar que se calcula a partir de los residuos observados, esta distribución teórica.

Se calcula este estadístico y luego se determina el valor p asociado, que se obtiene a través de software estadístico avanzado que en este caso el que usamos fuer R o tablas de distribución especializadas. El valor p proporciona una medida de la significancia estadística, permitiendo evaluar si es posible rechazar o no la hipótesis de normalidad dentro del contexto del modelo aplicado.

Es esencial realizar un análisis meticuloso para garantizar la integridad y confiabilidad de las inferencias derivadas del modelo de regresión lineal múltiple, siendo parte fundamental del marco teórico que respalda la investigación.

(Landa G, 2011; Tapia & Cevallos, 2021)

Prueba de Breusch-Pagan

La hipótesis nula de esta prueba dice que los errores son homocedásticos, es decir

$$H_0: Var(u_t|X_t) = E(u^2{}_t|X_t) = \sigma^2$$

Por tanto, si la hipótesis nula no es cierta, E (u²_t |Xt) no será constante, sino que será función de al menos una de las variables explicativas del modelo.

$$H_1: Var(u_t|X_t) = E(u^2{}_t|X_t) = g(Z_t) = \sigma^2{}_t$$

donde $Z_t = (Z_{1t}, \ldots, Z_{pt})'$ son algunas (o todas) las variables explicativas del vector X_t y/o funciones conocidas de ellas. Por tanto, una forma de comprobar si la media de $u^2{}_t$ depende de ese subconjunto $Z_t = (Z_{1t}, \ldots, Z_{pt})'$ de las variables explicativas es mediante la regresión

$$u^2{}_t = \delta_0 + \delta_1 Z_{1t} + \ldots. + \delta_p Z_{pt} + v_t$$

donde v_t es un término de error que suponemos tiene esperanza cero dadas $Z_{1t}, \ldots, Z_{pt}$
La hipótesis nula de homocedasticidad es

$$H_0: \delta_1 = \ldots. = \delta_p = 0$$

Los pasos para realizar el contraste Breusch-Pagan de heterocedasticidad son:
(1) Se estima por MCO $Y_t = X_t'\beta + u_t$ y se calculan los residuos.
(2) A continuación se estima por MCO la ecuación

$$e^2{}_t = \delta_0 + \delta_1 Z_{1t} + \ldots. + \delta_p Z_{pt} + error_t$$

(3) El estadístico de contraste y su distribución bajo H_0 son

$$T\,R^2 \simeq X^2{}_p$$

siendo R² el coeficiente de determinación de la regresión del paso (2) y p el número variables explicativas (excluyendo la constante) en la regresión del paso (2).

Es útil también mencionar las posibles consecuencias de la presencia de heterocedasticidad en el modelo de regresión múltiple. Pueden incluir estimaciones ineficientes de los coeficientes, intervalos de confianza incorrectos y pruebas de significancia sesgadas.

(De Alicante Departamento de Fundamentos del Análisis Económico, 2011)

Prueba de Durbin-Watson

La prueba de Durbin-Watson se emplea en estadística para identificar si hay autocorrelación, la cual es una medida estadística que se encarga de medir la relación lineal entre los valores de una misma variable a lo largo de un intervalo de tiempo, en este caso es la autocorrelación de primer orden en los residuos de un modelo de regresión, aunque inicialmente se usaba solo con modelos de regresión lineal simple, ahora también se utiliza en modelos de regresión múltiple. Dado que la presencia de autocorrelación puede invalidar muchas de las inferencias estadísticas estándar basadas en la regresión, esta prueba es particularmente importante.
Sea Y una variable dependiente y $X_1, X_2, \ldots, X_K$ las variables independientes en un modelo de regresión múltiple. Supongamos que el modelo está dado por:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_i + \epsilon_i \quad , i = 1,2,\ldots,n$$

donde ϵ_i son los términos de error asociados con cada observación, que idealmente deberían

ser independientes y distribuidos normalmente con media cero y varianza constante.
Después de estimar los coeficientes del modelo, $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k$, se calculan los residuos

$\hat{\epsilon}_i = Y_i - \hat{Y}_i$, donde $\hat{Y}_i = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_i$

El estadístico de Durbin-Watson se define como:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Ilustración 5. Estadístico Durbin-Watson

La forma de interpretar esta prueba es la siguiente:

- Valor ideal: Si el valor de d es cercano a 2, no hay autocorrelación significativa entre los residuos.
- Autocorrelación positiva: Valores por debajo de 2 sugieren fuertemente autocorrelación positiva.
- Autocorrelación negativa: Indican autocorrelación negativa valores que son significativamente mayores a 2.

  Es esencial asegurar la validez de las inferencias realizadas a partir del modelo de regresión
  mediante la correcta aplicación e interpretación de la prueba de Durbin-Watson en el contexto
  de la regresión múltiple.

(Alonso, M. 1994. Análisis de Residuales en Regresión, pág. 14)

Prueba de Levene

El estadístico de esta prueba se encuentra definido de la siguiente forma:

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i\left(\overline{Z}_{i.} - \overline{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \overline{Z}_{i.}\right)^2}$$

Ilustración 6. Prueba de Levene

De donde Z_{ij} puede ser de alguna de las siguientes formas:

1. $Z_{ij} = \left|X_{ij} - \overline{X}_{i.}\right|$ donde $\overline{X}_{i.}$ es la media del i-ésimo subgrupo.
2. $Z_{ij} = \left|X_{ij} - \widetilde{X}_{i.}\right|$ donde $\widetilde{X}_{i.}$ es la mediana del i-ésimo subgrupo.
3. $Z_{ij} = \left|X_{ij} - \overline{X'}_{i.}\right|$ donde $\overline{X'}_{i.}$ la media recortada al 10 % del i-ésimo subgrupo.

Ilustración 7. Formas que puede tomar Z

La prueba de Levene rechaza la hipótesis de que las varianzas son iguales con un nivel de significancia α si $W > F_{\alpha,k-1,N-k}$ donde $F_{\alpha,k-q,N-k}$ es el valor crítico superior de la distribución F con $k-1$ grados de libertad en el numerador y $N-k$ grados de libertad en el denominador a un nivel de significancia α. (Estudio de potencia de pruebas de homogeneidad de varianza, pp 59)

La prueba de Levene ofrece una alternativa más robusta que el procedimiento de Bartlett, ya que es poco sensible a la desviación de la normalidad. Eso significa que será menos probable que rechace una verdadera hipótesis de igualdad de varianzas sólo porque las distribuciones de las poblaciones muestreadas no son normales. (Estudio de potencia de pruebas de homogeneidad de varianza, pp 59)

Prueba Kolmogórov-Smirnov

La prueba de Kolmogórov-Smirnov es una prueba de bondad de ajuste ampliamente utilizada para probar la normalidad de los datos muestrales, siendo particularmente útil en procesos físicos no lineales e interactivos, por cuanto éstos conducen, generalmente, a distribuciones no gaussianas y, por lo tanto, el mecanismo generador de los procesos puede entenderse mejor al examinar la distribución de las variables seleccionadas. Además, para implementar pruebas de normalidad algunas pruebas estadísticas requieren o son óptimos bajo el supuesto de normalidad y, por lo tanto, constituye un prerrequisito determinar si este supuesto se cumple (Steinskog et al., 2007).

La prueba Kolmogórov-Smirnov se aplica para contrastar la hipótesis de normalidad de la población, la siguiente ecuación la representa y se muestra a continuación.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \quad \{ \begin{matrix} 1: \ si\ y_i \leq x, \\ 0: aleternativa \end{matrix}$$

Ilustración 8. Prueba de Kolmogórov-Smirnov

Para dos colas el estadístico viene dado por las ecuaciones:

$$D_n^+ = \max\left(F_n\,(x) - F\,(x)\right)$$
$$D_n^- = \max(F(x) - F_n(x))$$

Ilustración 9. Estadístico para dos colas

T - Test

Una prueba t (también conocida como prueba t de Student) es una herramienta para evaluar las medias de uno o dos grupos mediante pruebas de hipótesis. Una prueba t puede usarse para determinar si un único grupo difiere de un valor conocido (una prueba t de una muestra), si dos grupos difieren entre sí (prueba t de muestras independientes), o si hay una diferencia significativa en medidas pareadas (una prueba t de muestras dependientes o pareada) (JMP Statistical Discovery, 2024).

Asunciones de la prueba t
Aunque las pruebas t resisten relativamente bien las desviaciones de la hipótesis, al hacer una prueba t se asume que:

- Los datos son continuos.
- La muestra de datos se ha tomado aleatoriamente de la población.
- Hay homogeneidad en la varianza (por ejemplo, la variabilidad de datos de cada grupo es similar).
- La distribución es aproximadamente normal.

Para pruebas t de dos muestras, debemos tener muestras independientes. Si las muestras no son independientes, puede ser más adecuada una prueba t pareada (JMP Statistical Discovery, 2024).

Mapa conceptual

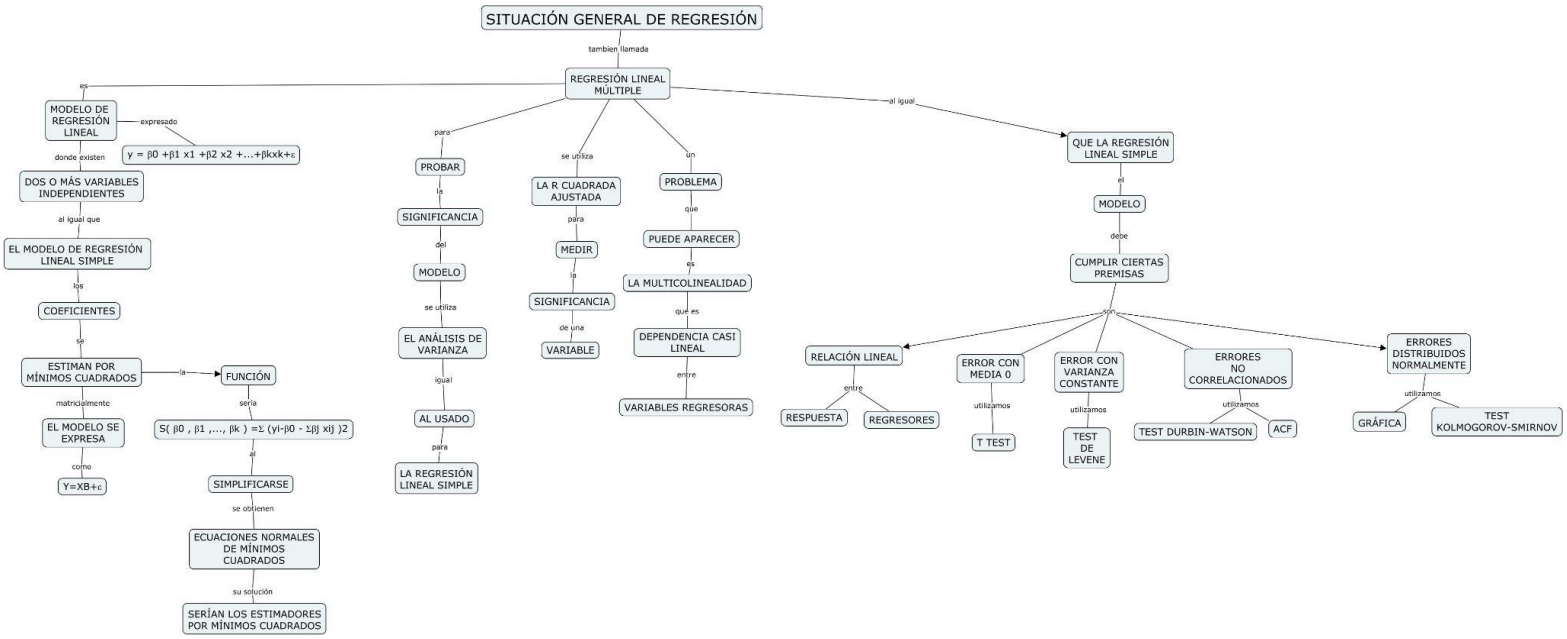

Ilustración 10. Mapa conceptual del Marco Teórico

Desarrollo

Para el desarrollo se utilizó la herramienta R para calcular el modelo de Regresión Lineal y sus cálculos subsecuentes.

Cargamos los datos necesarios y creamos un data frame.

```r
# Cargar los datos
datos <- data.frame(
  Entidad = c("Aguascalientes", "Baja California", "Baja California Sur", "Campeche", "Coahuila de Zaragoza", "Colima", "Chiapas",
  "Chihuahua", "Ciudad de México", "Durango", "Guanajuato", "Guerrero", "Hidalgo", "Jalisco", "México", "Michoacan de Ocampo", "Morelos", "Nayarit",
  "Nuevo León", "Oaxaca", "Puebla", "Queretaro", "Quintana Roo", "San Luis Potosí", "Sinaloa", "Sonora", "Tabasco", "Tamaulipas", "Tlaxcala",              # nolint
  "Veracruz de Ignacio de la Llave", "Yucatán", "Zacatecas"),

  Total_Poblacion_2020 = c(1425607, 3769020, 798447, 928363, 3146771, 731391, 5543828, 3741869, 9209944, 1832650, 6166934, 3540685, 3082841, 8348151, 16992418,
  4748846, 1971520, 1235456, 5784442, 4132148, 6583278, 2368467, 1857985, 2822255, 3026943, 2944840, 2402598, 3527735, 1342977, 8062579, 2320898, 1622138),
  Total_Residuos_2020 = c(1330, 3535, 737, 888, 3032, 743, 4964, 3638, 9552, 1767, 6031, 3421, 2694, 7961, 16739, 4459, 1878, 1146, 5310, 3538, 5991, 2085,
  1546, 2640, 3068, 2916, 2471, 3591, 1123, 7813, 2016, 1505),

  Total_PIB_2020 = c(20482, 554008, 121902, 48197, 535908, 101199, 254447, 539798, 2856972, 189953, 65082,
  218674, 244145, 1126206, 1488204, 397577, 179753, 10734, 1268066, 237528, 530958, 365965, 222079, 341565, 369694, 550906, 464414, 473823, 91087, 738747, 242552, 145444),

  Pobreza_Miles_en_personas = c(396500, 851700, 223400, 472400, 812100, 196000, 4218000, 952500, 3009400, 715500, 2649600, 2363200,
  1570600, 2633400, 83425, 2133700, 1006700, 376600, 1425000, 2569800, 4136600, 750400, 892900, 1214000, 853900, 885000, 1316100, 1233900, 800400, 4749600, 1156900, 745700),

  Turismo = c(463242, 2445832, 1808822, 86877, 1359239, 856038, 2046603, 2656757, 4294361, 467576, 2396576, 4139521, 960516, 4474467, 1243345, 1326169, 1166146, 1353066,
  1271653, 1598486, 1996146, 1046373, 7181292, 992345, 4049352, 720752, 720752, 1424826, 138766, 3286824, 1115818, 602600)
)
```

0.0s

Ilustración 11. Carga de los datos

Para asegurarnos que las operaciones con los datos anteriores serán correctas comprobamos la integridad y formato de estos.

```r
str(datos)
head(datos)
tail(datos)
```

[25] 0.0s

```
'data.frame':   32 obs. of  6 variables:
 $ Entidad                 : Factor w/ 32 levels "Aguascalientes",..: 1 2 3 4 8 9 5 6 7 10 ...
 $ Total_Poblacion_2020    : num  1425607 3769020 798447 928363 3146771 ...
 $ Total_Residuos_2020     : num  1330 3535 737 888 3032 ...
 $ Total_PIB_2020          : num  20482 554008 121902 48197 535908 ...
 $ Pobreza_Miles_en_personas: num  396500 851700 223400 472400 812100 ...
 $ Turismo                 : num  463242 2445832 1808822 86877 1359239 ...
```

Entidad	Total_Poblacion_2020	Total_Residuos_2020	Total_PIB_2020	Pobreza_Miles_en_personas	Turismo
Aguascalientes	1425607	1330	20482	396500	463242
Baja California	3769020	3535	554008	851700	2445832
Baja California Sur	798447	737	121902	223400	1808822
Campeche	928363	888	48197	472400	86877
Coahuila de Zaragoza	3146771	3032	535908	812100	1359239
Colima	731391	743	101199	196000	856038

	Entidad	Total_Poblacion_2020	Total_Residuos_2020	Total_PIB_2020	Pobreza_Miles_en_personas	Turismo
27	Tabasco	2402598	2471	464414	1316100	720752
28	Tamaulipas	3527735	3591	473823	1233900	1424826
29	Tlaxcala	1342977	1123	91087	800400	138766
30	Veracruz de Ignacio de la Llave	8062579	7813	738747	4749600	3286824
31	Yucatán	2320898	2016	242552	1156900	1115818
32	Zacatecas	1622138	1505	145444	745700	602600

Ilustración 12. Comprobación de integridad en los datos

Como vemos se imprimió el tipo de dato por columna y sus primero y últimos datos de data-frame para visualizar su integridad.

Modelo

Pasamos al cálculo de la regresión lineal por mínimos cuadrados. En este caso utilizaremos como variable Y al total de residuos en 2020 y procedemos a aplicar Stepwise **Selection** que combina el método de forward selection y backward elimination, considerando agregar o eliminar variables en cada paso basándose en algún criterio predefinido, en nuestro caso se usa el valor p de las variables para seleccionar las variables predictoras más relevantes para incluir en el modelo.

```r
# Ajustar el modelo inicial
modelo_inicial <- lm(Total_Residuos_2020 ~ Total_Poblacion_2020 +
Total_PIB_2020 + Pobreza_Miles_en_personas + Turismo, data = datos)

# Aplicar eliminación hacia atrás usando step()
modelo_final <- step(modelo_inicial, direction = "both", trace = TRUE)
```

Ilustración 13. Creación del modelo

En el código anterior se nota que se aplicó el argumento “trace = TRUE” en el modelo final, para que en nuestra salida nos muestre los pasos de selección de las variables.

```
Start:  AIC=334.4
Total_Residuos_2020 ~ Total_Poblacion_2020 + Total_PIB_2020 +
    Pobreza_Miles_en_personas + Turismo

                          Df Sum of Sq       RSS    AIC
- Turismo                  1      1713    810346 332.46
<none>                                    808634 334.40
- Pobreza_Miles_en_personas  1    158076    966710 338.11
- Total_PIB_2020           1    272194   1080828 341.68
- Total_Poblacion_2020     1 138007707 138816340 497.05

Step:  AIC=332.46
Total_Residuos_2020 ~ Total_Poblacion_2020 + Total_PIB_2020 +
    Pobreza_Miles_en_personas

                          Df Sum of Sq       RSS    AIC
<none>                                    810346 332.46
+ Turismo                  1      1713    808634 334.40
- Pobreza_Miles_en_personas  1    162690    973036 336.32
- Total_PIB_2020           1    296726   1107072 340.45
- Total_Poblacion_2020     1 138360466 139170813 495.14
```

Ilustración 14. Selección de variables

Vemos que la variable “Turismo” no pasó la prueba por lo que no fue incluida en el modelo.

Una vez calculada la regresión lineal múltiple procedemos a imprimir el summary para poder analizar su salida.

```
# Mostrar los resultados del modelo final
summary(modelo_final)
```

```
Call:
lm(formula = Total_Residuos_2020 ~ Total_Poblacion_2020 + Total_PIB_2020 +
    Pobreza_Miles_en_personas, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-432.67 -123.10   22.50   87.55  297.52

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -7.515e+01  5.325e+01  -1.411  0.16919
Total_Poblacion_2020       9.673e-04  1.399e-05  69.143  < 2e-16 ***
Total_PIB_2020             2.488e-04  7.771e-05   3.202  0.00339 **
Pobreza_Miles_en_personas -6.519e-05  2.750e-05  -2.371  0.02486 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 170.1 on 28 degrees of freedom
Multiple R-squared:  0.9975,    Adjusted R-squared:  0.9972
F-statistic:  3747 on 3 and 28 DF,  p-value: < 2.2e-16
```

Ilustración 15. Resumen del modelo

Coeficientes

El coeficiente para "Total_Poblacion_2020" es 9.673e-04. Esto significa que, manteniendo constante el efecto de las otras variables, por cada unidad adicional en la población en 2020, se espera un aumento de aproximadamente 0.0009673 en los residuos de 2020.

El coeficiente para "Total_PIB_2020" es 2.488e-04. Esto sugiere que, manteniendo constantes las otras variables, por cada unidad adicional en el PIB en 2020, se espera un aumento de aproximadamente 0.0002488 en los residuos de 2020.

El coeficiente para "Pobreza_Miles_en_personas" es -6.519e-05. Esto indica que, manteniendo constantes las otras variables, por cada unidad adicional en la cantidad de personas en pobreza en miles, se espera una disminución de aproximadamente 0.00006519 en los residuos de 2020.

Valores p

Un valor p menor que 0.05 indica que el coeficiente es estadísticamente significativo.

Nuestro summary nos da los resultados para las variables “Total_Poblacion_2020" y "Total_PIB_2020" en donde tienen valores p muy pequeños ($p < 0.05$), lo que sugiere que son estadísticamente significativos en la predicción de los residuos de 2020. Por otro lado, "Pobreza_Miles_en_personas" tiene un valor p de aproximadamente 0.025, lo que indica que también es significativo, pero con un nivel de significancia ligeramente más alto.

Bondad de ajuste del modelo

El coeficiente de determinación (R-cuadrado) es 0.9975, lo que significa que aproximadamente el 99.75% de la variabilidad en los residuos de 2020 es explicada por las variables predictoras en el modelo. Esto nos sugiere que el modelo ajustado tiene un buen ajuste a los datos.

ANOVA

```
anova_table <- anova(modelo_final)
print(anova_table)
```

✓ 0.0s

```
Analysis of Variance Table

Response: Total_Residuos_2020
                         Df    Sum Sq   Mean Sq    F value    Pr(>F)
Total_Poblacion_2020      1 324828421 324828421 11223.8358 < 2.2e-16 ***
Total_PIB_2020            1    310383    310383    10.7247  0.002815 **
Pobreza_Miles_en_personas 1    162690    162690     5.6214  0.024860 *
Residuals                28    810346     28941
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ilustración 16. Tabla ANOVA del modelo

Considerando el Pr(>F) (Valor p)

El valor p asociado con el valor F indica la probabilidad de obtener un valor F igual o más extremo que el observado, bajo la hipótesis nula de que no hay efecto de la variable independiente en la variable dependiente.Los valores p bajos (que técnicamente son 0) indican que la variable independiente es significativa en el modelo. Observemos los valores de cada una de nuestras variables.

- Total_Poblacion_2020:

En esta variable obtenemos un valor p: < 2.2e-16 (esencialmente cero) Esto significa que hay una evidencia muy fuerte en contra de la hipótesis nula de que el coeficiente de "Total_Poblacion_2020" es igual a cero. Con esto concluimos que hay una fuerte evidencia de que la variable "Total_Poblacion_2020" tiene un efecto significativo en la cantidad total de residuos.

- Total_PIB_2020:

En la variable del PIB el valor obtenido es p: 0.002815 por esto indica que hay una evidencia significativa en contra de la hipótesis nula de que el coeficiente de "Total_PIB_2020" es igual a cero.

- Pobreza_Miles_en_personas:

Por último, en esta variable obtenemos un valor p: 0.024860. Esto sugiere que hay una evidencia moderada en contra de la hipótesis nula de que el coeficiente de "Pobreza_Miles_en_personas" es igual a cero.

Considerando el valor F value:

El valor F compara la cantidad de variabilidad explicada por cada variable independiente con la variabilidad no explicada por el modelo en su conjunto. Valores F más altos indican que la variable independiente correspondiente es más importante en el modelo. En este caso, "Total_Poblacion_2020" tiene el valor F más alto, seguido por "Total_PIB_2020" y luego "Pobreza_Miles_en_personas".

Pruebas

Anderson-Darling

La prueba de normalidad (Anderson-Darling) nos da el siguiente valor:

```
        Anderson-Darling normality test

data:  residuos
A = 0.34999, p-value = 0.4511
```

Ilustración 17. Prueba Anderson-Darling

Dado que el valor p es mayor que 0.05, no hay suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto los residuos del modelo parecen seguir una distribución normal. Esto sugiere que el supuesto de normalidad de los residuos puede considerarse razonablemente válido en este caso.

T-test

```
t.test(residuos)
✓ 0.0s

        One Sample t-test

data:  residuos
t = -3.1076e-17, df = 31, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -58.29162  58.29162
sample estimates:
    mean of x
-8.881784e-16
```

Ilustración 18. Prueba T

Dado que el valor p es 1, no hay suficiente evidencia para rechazar la hipótesis nula de que la media de la muestra es igual a cero. Esto sugiere que la media de los residuos no es significativamente diferente de cero, lo que es consistente con el supuesto de que los residuos tienen una media de cero en un modelo de regresión lineal.

Breusch-Pagan

```
bptest(modelo_final)
✓ 0.0s

        studentized Breusch-Pagan test

data:  modelo_final
BP = 4.3684, df = 3, p-value = 0.2243
```

Ilustración 19. Prueba de Breusch-Pagan

El valor p obtenido es 0.2243, que es mayor que 0.05. Por lo tanto, no hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad. Esto sugiere que la varianza de los errores en el modelo es constante, lo que cumple con el supuesto de homocedasticidad.

Durbin-Watson

```
dwtest(modelo_final)
✓ 0.1s

        Durbin-Watson test

data:  modelo_final
DW = 1.3408, p-value = 0.02601
alternative hypothesis: true autocorrelation is greater than 0
```

Ilustración 20. Prueba de Durbin-Watson

Dado que el valor p es menor que 0.05, hay evidencia suficiente para rechazar la hipótesis nula de que no hay autocorrelación positiva en los residuos. Esto sugiere que existe cierta correlación positiva entre los residuos adyacentes en el modelo de regresión.

Gráficas

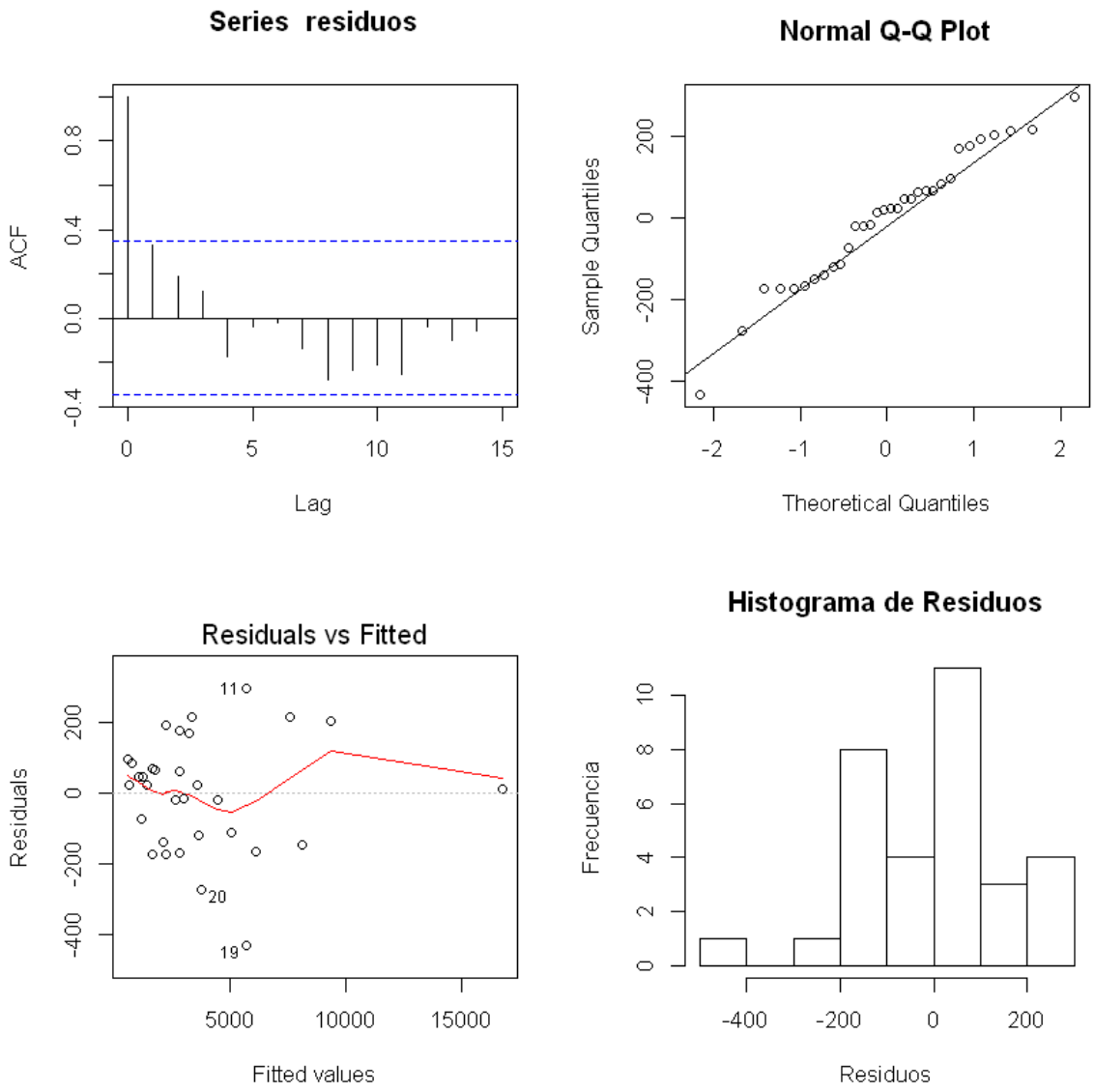

Ilustración 21. Gráficos

ACF

Vemos que todos los picos del gráfico entran en las bandas de confianza, esto sugiere que no hay autocorrelación significativa.

Q-Q Plot (Quantile-Quantile Plot)

En esta gráfica vemos que los puntos se apegan a la diagonal teórica por ello podemos decir que se asemejan al modelo.

Histograma de Residuos

Vemos que los residuos muestran cierto sesgo negativo, pero sigue manteniendo una forma de campana apegada a una normal.

Gráfico de Residuos vs. Valores Ajustados

Vemos que los puntos dentro del gráfico tienen cierta tendencia alrededor de cero lo que sugiere que la varianza de los residuos es constante en diferentes niveles de los valores ajustados (homocedasticidad).

Conclusiones

En general, el modelo creado, que tiene como variables predictoras al producto interno bruto, el índice de pobreza y a la población total, cumple con todos los supuestos sobre los residuos. Por lo que podemos concluir que existe evidencia estadística para decir que la cantidad generada de residuos sólidos urbanos en México está relacionada directamente al PIB, al índice de pobreza y a la cantidad de personas que hay en México.

Por otro lado, a pesar de que en el año 2020 ocurrió la pandemia de COVID-19, que afectó en gran medida al turismo, esperábamos que los datos sobre el turismo tuvieran un impacto directo en los residuos sólidos urbanos, sin embargo, como ya lo vimos en el desarrollo del proyecto, esto no fue así y fue una variable que se tuvo que quitar del modelo por no ser significativa.

Referencias

Alonso, M. (1994). Análisis de Residuales en Regresión. Trabajo para obtener diploma en: Especialidad en Métodos Estadísticos. Universidad Veracruzana.

Correa, Juan Carlos. Iral, René. Rojas, Lucinia.(2006). Estudio de potencia de pruebas de homogeneidad de varianza. Recuperado de https://www.redalyc.org/pdf/899/89929104.pdf

De la Guerra, J. P. (2015). Las siete herramientas de la calidad. CORE. https://core.ac.uk/download/pdf/356676474.pdf

Dagnino, J. (2014). Análisis de Varianza. Bioestadística y Epidemiología. Revista Chilena de Anestesia. pág.1.

De Alicante Departamento de Fundamentos del Análisis Económico, U. (2011, 1 septiembre). Econometría II (Curso 2010-2011). http://hdl.handle.net/10045/18469

Datatur3 - InfTurxEdo. (s/f). Gob.mx. Recuperado el 3 de mayo de 2024, de https://www.datatur.sectur.gob.mx/SitePages/InfTurxEdo.aspx.

Elousa, P. (2011) . Introducción al entorno R. Bilbao: Euskal Herriko Unibertsitateko Argitalpen Zerbitzua/Servicio Editorial de la Universidad del País Vasco.

Gobierno de México. (1 de marzo de 2017). *Residuos Sólidos Urbanos y de Manejo Especial*. https://www.gob.mx/semarnat/acciones-y-programas/residuos-solidos-urbanos-y-de-manejo-especial.

JMP Statistical Discovery(2024). La prueba t. Recuperado de https://www.jmp.com/es_mx/statistics-knowledge-portal/t-test.html#:~:text=%C2%BFQu%C3%A9%20es%20una%20prueba%20t,grupos%20mediante%20pruebas%20de%20hip%C3%B3tesis. (Fecha de acceso: 8 de mayo de 2024).

Montgomery, D., Peck, E. y Vining, G. (2006). *Introducción al Análisis de Regresión Lineal*. Compañía Editorial Continental.

Novales. A. (20 de septiembre de 2010). Análisis de Regresión. Universidad Complutense. https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf

Steinskog, D., Tjøstheim, D., & Gunnar, N. (2007). A cautionary note on the use of the Kolmogorov–Smirnov test for normality. Monthly Weather Review, 135(3), 1151–1157.

https://www.researchgate.net/publication/249621733_A_Cautionary_Note_on_the_Use_of_the_Kolmogorov-Smirnov_Test_for_Normality

Tapia, C. E. F., & Cevallos, K. L. F. (s. f.). PRUEBAS PARA COMPROBAR LA NORMALIDAD DE DATOS EN PROCESOS PRODUCTIVOS: ANDERSON-DARLING, RYAN-JOINER, SHAPIRO-WILK y KOLMOGÓROV-SMIRNOV. http://portal.amelica.org/ameli/jatsRepo/341/3412237018/html/index.html

Anexos teóricos

Se incluye en este espacio la información obtenida de las presentaciones anteriormente vistas:

ANOVA

Conjunto de técnicas estadísticas que sirven para analizar la variación entre muestras y la variación interior de las mismas. Es de utilidad cuando hay más de dos grupos que necesitan ser comparados.

Mediante los resultados del ANOVA es posible evaluar la significancia del modelo, mediante el estadístico F es posible medir qué tan efectiva será la predicción del modelo. Si el p-value asociado a F es menor al valor de alfa propuesto, se rechaza la hipótesis nula que plantea que no existe relación entre las variables.

Error puro

Discrepancia entre el valor observado de una variable y su valor verdadero o teórico, ya sea por errores de medición o imprecisión inherente en los datos.

El error puro se compone de:

- Error sistemático: error que no puede ser explicado por el modelo de regresión.
- Error aleatorio: inherente del fenómeno que se está estudiando.

Diferencia entre independencia y no correlación

Dos variables son independientes cuando al conocer el valor que toma una de ellas, no se tienen ninguna información sobre la otra. No existe relación lineal, cuadrática, logarítmica, etc.

La correlación se utiliza para examinar la dirección y la fuerza de la asociación lineal entre dos variables. La no correlación se observa en el diagrama de dispersión como un grupo de puntos sin ninguna tendencia lineal.

La presencia de no correlación no implica independencia, mientras que la presencia de independencia implica que no hay correlación.

Chi-cuadrada

Método que compara los resultados observados con los datos teóricos para una muestra. La hipótesis nula de esta prueba es que las variables son independientes, mientras que la hipótesis alternativa es que las variables están asociadas.

Tao de Kendall

Se encarga de evaluar el grado de similitud entre dos conjuntos de rangos dado el mismo conjunto de objetos. El coeficiente depende del número de inversiones de pares de objetos que serían necesarias para transformar un rango en el otro.

Correlación lineal

Francis Galton en 1888 acuñó el término de correlación para referirse a un conjunto de variables que influyen en ambas simultáneamente de manera positiva, con 0 si las variables no se asocian y con 1 si las variables están fuertemente relacionadas.

Por otro lado, Karl Pearson plantea en el año 1896 el primer coeficiente de correlación.
El término correlación se refiere a la asociación entre dos variables. Se dice que dos variables están asociadas cuando una variable nos da información sobre la otra. Es necesario tener en cuenta el signo y la magnitud con la que las variables están relacionadas; el primero indica la dirección de la relación (puede ser positivo o negativo) y la segunda nos indica la fuerza de la relación. Cuánto más cercano sea el valor a los extremos del intervalo (-1,1) más fuerte será la tendencia de las variables.

Coeficiente de correlación de Pearson
Medida que se encarga de cuantificar la correlación lineal entre dos variables, denotado por la letra r. Se caracteriza por ser muy fácil de interpretar, ser adimensional. El resultado de la aplicación de la prueba es un valor definido entre -1 y 1.
Una desventaja es que solo puede calcular relaciones lineales, para otro tipo de relaciones es necesario aplicar otras pruebas.
Para poder usar el coeficiente de correlación de Pearson es necesario que: las variables deben ser de intervalo o de razón, que los datos sean pareados, que la distribución conjunta sea normal, la relación entre las variables debe ser lineal, las variables deben ser independientes y que cada variable debe tener la misma distribución de probabilidad.

Coeficiente de correlación de Spearman
Prueba, que al igual que el coeficiente de correlación de Pearson, mide el nivel de correlación entre dos variables, pero, a diferencia de este, es un método más robusto y adecuado para datos que no siguen una distribución normal, es fácil de interpretar y es útil para analizar variables ordinales. Sin embargo, es posible que no pueda detectar diferencias en la relación entre variables y sólo puede identificar la fortaleza de correlación en variables que son crecientes o decrecientes.
A diferencia del coeficiente de correlación de Pearson, el coeficiente de correlación de Spearman puede medir asociaciones monótonas, es decir, puede detectar relaciones más generales, no sólo lineales.

Test de Durbin-Watson
Test que indica la autocorrelación entre los residuos de un modelo de regresión lineal. La hipótesis nula indica que si el estadístico es cero, no existe autocorrelación, mientras que la hipótesis alternativa indica que si el estadístico es mayor a 0, existe autocorrelación.

Relación no lineal
Cuando dos variables se relacionan según una curva se habla de una regresión no lineal. Por lo tanto, es necesario buscar la función que describe la dependencia entre las variables.

Kolmogorov-Smirnov
Prueba que permite verificar si los valores de una muestra siguen o no una distribución normal. La prueba compara la distribución acumulada empírica de los datos con la distribución acumulada teórica.

Shapiro-Wilk

Prueba que determina si un conjunto de datos proviene de una distribución normal. Fue desarrollada por Samuel Shapiro y Martin Wilk en 1965 y presenta un enfoque diferente a la prueba de Kolmogorov-Smirnov. A diferencia de Kolmogorov-Smirnov, con Shapiro-Wilk es posible realizar la prueba con muestras menores a cincuenta elementos. Además, Shapiro-Wilk tiene un enfoque basado en la correlación entre los valores ordenados de la muestra y los valores esperados bajo la hipótesis de normalidad.

Significado de requisito de normalidad para los residuales del modelo de regresión

En los modelos de regresión lineal, se asume que los residuos siguen una distribución normal. Esto se debe a que si no se comportan como una normal, habría indicios de errores en el modelo, como errores de especificación o datos atípicos.

Homocedasticidad y heterocedasticidad

La existencia de homocedasticidad indica que la varianza es constante, de no ser así, se dice que hay heterocedasticidad.

Identificación gráfica en los residuales del modelo de regresión

Mediante un gráfico de los residuales es posible encontrar datos atípicos: un dato atípico es un valor alejado del resto. También es posible observar la autocorrelación de los residuos mediante la ACF; si todas las líneas se encuentran dentro del intervalo, podemos decir que estadísticamente no existe autocorrelación en los residuales.

Mediante un Q-Q plot es posible analizar la normalidad en los errores, si los valores se agrupan en una línea, se podría observar la normalidad de los residuos.

Test de Bartlett

Prueba que se usa para comprobar que la varianza sea igual para todas las muestras, aunque es necesario que los datos provengan de una distribución normal. La hipótesis nula es que las varianzas son iguales para todas las muestras, mientras que la hipótesis alternativa es que las varianzas no son iguales para un par o más.

Colinealidad

Se refiere a la alta correlación entre dos o más variables predictoras. Existen dos tipos: la colinealidad exacta, en donde una variable es combinación lineal de otra y la colinealidad aproximada, donde la relación es muy cercana, pero no perfecta. Cuando existe colinealidad, las variables predictoras no son independientes, lo que dificulta la estimación precisa de los coeficientes del modelo.

Los principales problemas de la presencia de colinealidad es que hay mayor variabilidad en los coeficientes, reduciendo la precisión de las estimaciones y que los coeficientes sean inestables con cambios mínimos en el modelo, lo que dificulta la interpretación.

La colinealidad se puede deber a la inclusión excesiva de variables que no son relevantes para el modelo, que se haya tomado una muestra pequeña con relación a las variables o errores de medición.
Las técnicas que se pueden utilizar en el momento que se ha detectado colinealidad en el modelo son: eliminar variables correlacionadas, aumentar el tamaño de la muestra, usar técnicas de regularización como Ridge o Lasso o realizar alguna transformación a las variables.
Para evitar la colinealidad es importante realizar un análisis exploratorio de correlaciones antes de construir el modelo, seleccionar cuidadosamente las variables predictoras, usar técnicas de regularización y considerar métodos alternativos de estimación.

Detección mediante factor de inflación de la varianza
El factor de inflación de la varianza es un estadístico basado en la correlación muestral entre variables explicativas que cuantifica la intensidad de la colinealidad de las variables en un análisis de regresión lineal.

Distancia de Cook
Utilizado para detectar valores atípicos. Se calcula mediante la diferencia entre el modelo completo con todas las muestras, menos el modelo generado, eliminando la muestra bajo estudio. La distancia de Cook crea un nuevo modelo sin una variable y verifica si el modelo ha cambiado mucho o no, después pasa a otro y verifica lo mismo.