UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

LICENCIATURA EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

Estadística II

Relación entre RSU, Población y PIB en México: Un Análisis Multivariante

Presenta

Arreola Calderón Jesús Enrique 100 %
Cano Verduzco Mónica 100 %
Cortés Mejía Kaleb 100 %
Palomares Olegario Alexis 100 %
Romo Aldaco Luis Angel 100 %

Profesor: Jaime Vergara Prado
08 de mayo de 2024

Índice

Índice de cuadros

Índice de figuras

1. Resumen

Este documento presenta un análisis multivariante para examinar la relación entre la generación de residuos sólidos urbanos (RSU) y diversos factores socioeconómicos en México, como la población, el producto interno bruto (PIB), la incidencia delictiva, la esperanza educativa, el ingreso corriente anual, la densidad poblacional y la pobreza. Se brinda un marco teórico sobre la regresión lineal múltiple, la interpretación de coeficientes y los métodos de selección de variables. Utilizando datos de los estados mexicanos y el software R, se ajusta un modelo de regresión lineal múltiple y se evalúan los supuestos teóricos. Tras aplicar transformaciones logarítmicas y recíprocas a ciertas variables, se encuentra que el PIB es el único factor significativo para predecir la generación de RSU. El documento concluye que, contrario a la intuición, los RSU en México están más relacionados con el modelo económico y la generación de ingresos que con otros factores como la pobreza o el nivel educativo. Se destaca la utilidad de los métodos de selección de variables para ahorrar tiempo en el análisis.

Palabras clave: Regresión lineal, Residuos sólidos urbanos, Transformación logarítmica

2. Introducción

La gestión adecuada de los residuos sólidos urbanos (RSU) es un desafío crucial que enfrentan las ciudades de todo el mundo, especialmente en países en vías de desarrollo como México. La creciente generación de RSU, impulsada por el aumento de la población, la urbanización y el consumo, ejerce una presión significativa sobre los recursos naturales, los ecosistemas y la salud pública (Kaza et al., 2018). En este contexto, comprender los factores que determinan la generación de RSU es fundamental para desarrollar estrategias efectivas de gestión sostenible.

En este estudio, se propone un análisis multivariante para explorar la relación entre la generación de RSU, la población, el producto interno bruto (PIB), incidencia delictiva, esperanza educativa, ingreso corriente anual, número de habitantes por kilómetro cuadrado y pobreza en México. Se pretende identificar los patrones y tendencias que caracterizan esta relación, y evaluar la influencia de estos factores socioeconómicos en la producción de RSU a nivel nacional y estatal.

Una extensión natural del modelo de regresión lineal simple consiste en considerar más de una variable explicativa.

Los modelos de regresión múltiple estudian la relación entre:

- Una variable de interés Y (variable respuesta o dependiente)
- Un conjunto de variables explicativas o regresoras X1, X2, . . . , Xp

En el modelo de regresión lineal múltiple se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

---

3. Marco Teórico

Planteamiento del problema. ¿Existe una relación entre la generación de los residuos sólidos urbanos en las entidades federativas de la República Mexicana y otras variables, tales como población y el producto interno bruto, entre otras?

3.1. Regresión lineal múltiple

El modelo de regresión lineal múltiple es una herramienta fundamental en la estadística y la econometría para analizar las relaciones entre múltiples variables. Sin embargo, su validez y eficacia dependen de una serie de supuestos y condiciones que deben cumplirse para obtener resultados confiables y significativos.

Uno de los supuestos fundamentales del modelo de regresión lineal múltiple es la linealidad de la relación entre las variables independientes y la variable dependiente. Según Gujarati (2003), este supuesto establece que el efecto de un cambio en una variable independiente sobre la variable dependiente es constante, lo que implica que la relación entre estas variables puede ser representada de manera lineal.

Otro supuesto importante es la ausencia de multicolinealidad entre las variables independientes. La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas entre sí, lo que dificulta la estimación precisa de los coeficientes de regresión. Este supuesto fue discutido por Johnston (1963) en su obra clásica sobre regresión lineal.

Además, el modelo de regresión lineal múltiple asume que los errores de la regresión tienen una distribución normal con media cero y varianza constante. Este supuesto, conocido como homocedasticidad, es crucial para garantizar la eficiencia y consistencia de los estimadores de mínimos cuadrados ordinarios (MCO), como señaló Greene (2012) en su libro sobre econometría.

Otro supuesto relevante es la independencia de los errores, que implica que los errores de una observación no están correlacionados con los errores de otras observaciones. Esta condición es esencial para obtener estimaciones no sesgadas de los parámetros del modelo y fue destacada por Wooldridge (2016) en su texto sobre econometría.

Además de estos supuestos fundamentales, el modelo de regresión lineal múltiple también requiere que las observaciones utilizadas en el análisis sean aleatorias y representativas de la población de interés. Este requisito garantiza la validez de las inferencias realizadas a partir de los resultados del modelo.

3.2. Interpretación de los coeficientes de los coeficientes de regresión

La interpretación de los coeficientes en modelos de regresión lineal múltiple es fundamental para comprender cómo las variables independientes afectan a la variable dependiente. Cada coeficiente representa el cambio en la variable dependiente asociado con un cambio unitario en la variable independiente correspondiente, manteniendo todas las demás variables constantes. Sin embargo, esta interpretación puede ser compleja debido a la presencia de múltiples variables independientes.

Uno de los enfoques comunes para interpretar los coeficientes en modelos de regresión lineal múltiple es utilizando el concepto de elasticidad. Según Gujarati (2003), la elasticidad de una variable independiente xi con respecto a la variable dependiente y se define como el cambio porcentual en y asociado con un cambio porcentual en xi. Esta interpretación es útil para comprender la relación relativa entre las variables.

Además, la interpretación de los coeficientes también puede implicar la comparación de los efectos marginales de las variables independientes sobre la variable dependiente. Wooldridge (2016) destaca que el efecto marginal de una variable independiente xi se refiere al cambio en la variable dependiente y cuando aumenta en una unidad, manteniendo constantes todas las demás variables independientes.

Esta interpretación ayuda a entender el impacto individual de cada variable independiente en la variable dependiente.

Es importante considerar la escala de las variables al interpretar los coeficientes en modelos de regresión lineal múltiple. Greene (2012) menciona que los coeficientes pueden variar en magnitud dependiendo de la escala de las variables involucradas. Por lo tanto, es fundamental estandarizar las variables antes de interpretar los coeficientes para facilitar la comparación de sus efectos.

Otro aspecto importante en la interpretación de los coeficientes es su significancia estadística. La significancia estadística indica si el coeficiente estimado es diferente de cero, lo que sugiere si la variable independiente correspondiente tiene un efecto significativo en la variable dependiente. Este aspecto fue discutido por Johnston (1963) en su obra sobre métodos de regresión.

En el modelo de regresión lineal múltiple se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

β_0 es el término independiente. Es el valor esperado de Y cuando $X_1, \ldots, X_p$ son cero y $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes parciales de la regresión:

- β_1 mide el cambio de Y por cada cambio unitario en X_1 manteniendo $X_2, X_3, \ldots, X_p$ constantes
- β_2 mide el cambio de Y por cada cambio unitario en X_2 manteniendo $X_2, X_3, \ldots, X_p$ constantes
- β_p mide el cambio de Y por cada cambio unitario en X_p manteniendo $X_1, \ldots, X_p$ a -1
- ε es el error de observación debido a variables no controladas

De la expresión matemática del modelo de regresión lineal generalmente se deduce:

$$y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon_i \text{ , donde } i = 1, 2, \ldots, n$$

- Asumimos que los errores $\varepsilon_1, \varepsilon_2, \ldots, n$ tienen distribución normal de media cero y varianza σ^2 y que son independiente
- Las variables explicativas son linealmente independientes entre sí

Para poder obtener a partir los estimadores:

- De los coeficientes $\beta^0, \beta^1, \ldots, \beta^p$
- De la varianza del error $\hat{\sigma}^2$

Podemos plantear el modelo en forma matricial de la siguiente manera:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Figura 1: Modelo en forma matricial

Podríamos escribir la expresión:

$$Y = X\beta + \varepsilon$$

Para estimar el vector de parámetros β podemos aplicar el método de mínimos cuadrados, igual que en el modelo lineal simple, y como resultado se obtiene el siguiente estimador:

$$\hat{\beta} = X_t X - 1X_t$$

Donde X_t denota la matriz transpuesta de X.
La variabilidad de toda la muestra se denomina variabilidad total (VT).

$$VT = \sum_{1=1}^{n} (y_i - \bar{y})^2$$

Al igual que en el modelo de regresión lineal simple, podemos descomponer la variabilidad total de Y en dos sumandos.

- La variabilidad explicada (VE)

$$VE = \sum_{1=1}^{n} (\hat{y}_i - \bar{y})^2$$

- La variabilidad no explicada (VNE) por la regresión

$$VNE = \sum_{1=1}^{n} (y_i - \hat{y})^2$$

Descomposición de la variabilidad

$$VT = VE + VNE$$

Factor de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrado medio	Estadístico F_0	p-valor
Regresión (VE)	$SC_R = \beta_1 S_{XY}$	1	CM_R	$\frac{CM_R}{CM_E}$	$Pr(F > F_0)$
Error o residual (VNE)	$SC_E = S_{YY} - \beta_1 S_{XY}$	$n-2$	CM_E		
Total (VT)	S_{YY}	$n-1$			

Cuadro 1: Tabla ANOVA teoría

El coeficiente de determinación (R^2) se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión

$$R^2 = \frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

- El coeficiente de determinación presenta el inconveniente de aumentar siempre que aumenta el número de variables regresoras (algunas veces de forma artificial)
- Por ello y para penalizar el número de variables regresoras que se incluyen en el modelo de regresión, es conveniente utilizar el coeficiente de determinación corregido por el número de grados de libertad

Coeficiente de determinación ajustado:

$$R^2 \text{ ajustado} = 1 - \frac{\frac{VNE}{(n-(p+1))}}{\frac{VT}{(n-1)}}$$

$$\hat{\beta} = X_t X - 1 X_t$$

Como estimador de la varianza del error se puede emplear:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Figura 2: Fórmula de la varianza del error

3.3. Métodos de selección de variables

3.3.1. Eliminación hacia atrás (Backwards)

Se introducen todas las variables en la ecuación y después se van excluyendo una tras otra.

En cada etapa se elimina la variable menos influyente según el contraste individual (de la t o de la F).

3.3.2. Selección hacia adelante (Fordward)

Las variables se introducen secuencialmente en el modelo.

La primera variable que se introduce es la de mayor correlación (+ o -) con la variable dependiente.

Dicha variable se introducirá en la ecuación solo si cumple el criterio de entrada.

A continuación se considera la variable independiente cuya correlación parcial sea la mayor y que no esté en la ecuación.

El procedimiento termina cuando ya no quedan variables que cumplan el criterio de entrada.

3.3.3. Pasos sucesivos (Stepwise Regression)

Este método es una combinación de los procedimientos anteriores.

En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga la probabilidad para F más pequeña (i.e. hacia adelante).

Las variables ya introducidas en la ecuación de regresión pueden ser eliminadas del modelo (i.e. hacia atrás).

El método termina cuando ya no hay más variables candidatas a ser incluidas o eliminadas.

3.4. Transformación logarítmica de una distribución asimétrica

La transformación logarítmica significa tomar un conjunto de datos y tomar el logaritmo natural de las variables. A veces, es posible que sus datos no se ajusten del todo al modelo que está buscando y una transformación logarítmica puede ayudar a ajustar una distribución muy sesgada a un modelo más normal. Como resultado, podrá ver más fácilmente patrones en sus datos. La transformación de registros no "normaliza"sus datos; su propósito es reducir el sesgo.

Si está ejecutando una prueba estadística paramétrica con sus datos (por ejemplo, un ANOVA), el uso de datos muy sesgados hacia la derecha o hacia la izquierda puede generar resultados de prueba engañosos. Por lo tanto, si desea realizar una prueba con este tipo de datos, ejecute una transformación de registro y luego ejecute la prueba con los números transformados.

Existen muchas transformaciones posibles. Sin embargo, solo debes utilizar una transformación de registro si:

- Sus datos están muy sesgados hacia la derecha (es decir, en la dirección positiva).
- La desviación estándar del residual es proporcional a los valores ajustados.
- La relación de los datos se acerca a un modelo exponencial.
- Crees que los residuos reflejan errores multiplicativos que se han acumulado durante cada paso del cálculo.

3.5. Transformación recíproca

En la regresión lineal múltiple, el término ”recíproco” se refiere a la inclusión de la variable independiente como su recíproco en el modelo de regresión. Esto se hace cuando se sospecha que la relación entre la variable independiente y la variable dependiente es no lineal y se desea explorar una relación inversa.

Por ejemplo, si tenemos una variable independiente X y una variable dependiente Y, y sospechamos que la relación entre X y Y es no lineal, podríamos incluir $\frac{1}{X}$ como una variable independiente en el modelo de regresión para explorar una posible relación recíproca.

4. Mapa Conceptual

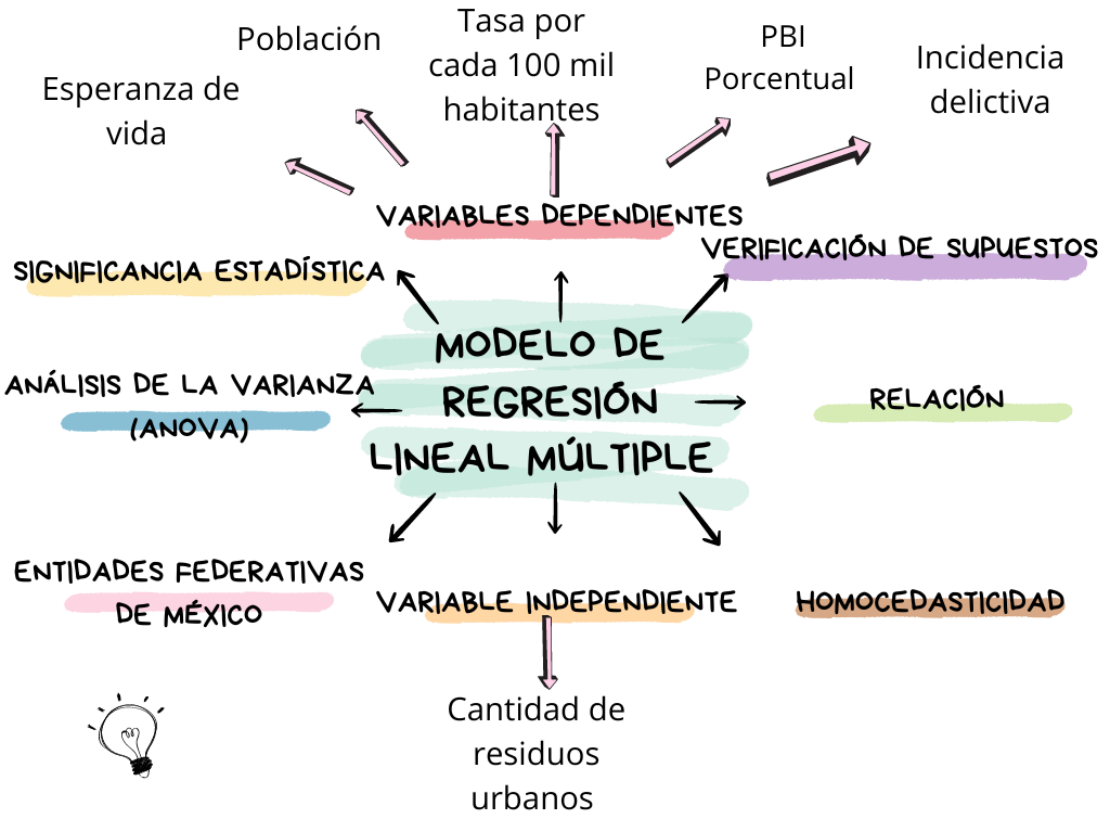


Figura 3: Mapa Conceptual

5. Desarrollo

5.1. Descripción de variables y unidades de medición

Para nuestro modelo de regresión lineal múltiple, las variables que se utilizaron en son:
RSU t/diax100milH: cálculo hecho para medir la generación de residuos sólidos urbanos por cada 100 mil habitantes al día en cada estado, calculado de la siguiente manera, de la población total se obtiene el cociente de dividir entre 100 mil (para así obtener los resultados en 100 mil habitantes). Seguidamente se divide la generación de RSU toneladas por día entre la población en 100 mil habitantes, de esta forma, se asegura que estemos usando los RSU de toneladas al día medidos por cada 100 mil habitantes en cada estado de la república Mexicana.
PIB: muestra el PIB porcentual de cada estado (con respecto a la república Mexicana).
Incidencia delictiva: indica cuántos delitos se cometen por cada 100 mil habitantes.
Esperanza educativa: cuántos años se espera que una persona de entre 5 y 29 años de edad esté inscrita en algún nivel educativo, independientemente del nivel educativo que se curse.
Ingreso corriente anual: indica el total de ingreso de un hogar (en pesos mexicanos) que todos sus miembros puedan proveer en un año.
Número de habitantes por km cuadrado: indica la cantidad de habitantes por cada kilómetro cuadrado de territorio.
Pobreza por 100 mil habitantes: indica cuántos habitantes de cada 100 mil están en alguna situación de pobreza.
Todas nuestras variables cuando se tratan de población, se manejarán en cientos de miles. A diferencia de aquellas que sean por tiempo (como el ingreso corriente anual o la incidencia delictiva), dichas variables se manejarán por año, ya que, estamos estudiando el año 2020.

5.2. Uso de software R

Una vez que tenemos los datos a utilizar en sus respectivas medidas (cientos de miles o años), procedemos a trabajar con el software R. Primeramente, se realizó un modelo de regresión lineal con todas las variables posibles, para así tener una perspectiva más acertada de cuales serán o no significativas.

```r
model.backward<-lm(formula = hoja$`RSU t/diax100milH` ~ hoja$PIB +
                     hoja$`Incidencia delictiva`+
                     hoja$`Esperanza educativa`+
                     hoja$`Ingreso corriente anual`+
                     hoja$`Número de habitantes por km cuadrado`+
                     hoja$`Pobreza por 100 mil habitantes`, data = hoja)
summary(model.backward)
```

Figura 4: Modelo de regresión con todas las variables en R

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              8.096e+01  1.363e+01   5.940 3.36e-06 ***
hoja$PIB                                 1.026e+00  9.773e-01   1.049    0.304
hoja$`Incidencia delictiva`              1.165e-05  1.920e-04   0.061    0.952
hoja$`Esperanza educativa`               8.480e-01  1.001e+00   0.847    0.405
hoja$`Ingreso corriente anual`          -4.703e-06  4.414e-05  -0.107    0.916
hoja$`Número de habitantes por km cuadrado` -4.019e-04  1.784e-03  -0.225    0.824
hoja$`Pobreza por 100 mil habitantes`   -4.623e-02  1.146e-01  -0.403    0.690
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 5: Resultados del modelo de regresión con todas las variables

Observamos que las variables PIB y Esperanza educativa son las que más se acercan a ser significativas, utilizaremos ahora los tres métodos de selección de variables backward, forward y stepwise en ese orden.

```r
#selección de variables por backward
model.backward<-lm(formula = hoja$`RSU t/diax100milH` ~ hoja$PIB +
                     hoja$`Incidencia delictiva`+
                     hoja$`Esperanza educativa`+
                     hoja$`Ingreso corriente anual`+
                     hoja$`Número de habitantes por km cuadrado`+
                     hoja$`Pobreza por 100 mil habitantes`, data = hoja)
library(MASS)  # Para poder usar la funcion stepAIC que es el metodo de seleccion
modback <- stepAIC(model.backward, trace=TRUE, direction="backward")
#seleccion de variables por forward
#modelo vacio para el forward
model.forward <- lm(hoja$`RSU t/diax100milH`~ 1, data=hoja)
#objetivo que tiene el forward
horizonte <- formula(hoja$`RSU t/diax100milH` ~ hoja$PIB +
                       hoja$`Incidencia delictiva`+
                       hoja$`Esperanza educativa`+
                       hoja$`Ingreso corriente anual`+
                       hoja$`Número de habitantes por km cuadrado`+
                       hoja$`Pobreza por 100 mil habitantes`, data = hoja)
modforw <- stepAIC(model.forward, trace=FALSE, direction="forward", scope=horizonte)
#seleccion de variables para stepwise
#modelo vacio para stepwise
model.stepwise <- lm(hoja$`RSU t/diax100milH` ~ 1, data=hoja)
modboth <- stepAIC(model.stepwise, trace=FALSE, direction="both", scope=horizonte)
```

Figura 6: Código para la selección de variables en R

```
> #Resultado backward
> summary(modback)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ hoja$PIB, data = hoja)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7333 -3.3594  0.1715  2.6543  9.1897

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92.0115     1.2665  72.652   <2e-16 ***
hoja$PIB      0.7156     0.2893   2.474   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.023 on 30 degrees of freedom
Multiple R-squared:  0.1694,    Adjusted R-squared:  0.1417
F-statistic:  6.12 on 1 and 30 DF,  p-value: 0.01925
```

Figura 7: Resultado de selección backward

```
> #Resultado forward
> summary(modforw)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ hoja$PIB, data = hoja)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7333 -3.3594  0.1715  2.6543  9.1897

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92.0115     1.2665  72.652   <2e-16 ***
hoja$PIB      0.7156     0.2893   2.474   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.023 on 30 degrees of freedom
Multiple R-squared:  0.1694,    Adjusted R-squared:  0.1417
F-statistic:  6.12 on 1 and 30 DF,  p-value: 0.01925
```

Figura 8: Resultado de selección forward

```
> #Resultado stepwise
> summary(modboth)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ hoja$PIB, data = hoja)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7333 -3.3594  0.1715  2.6543  9.1897

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92.0115     1.2665  72.652   <2e-16 ***
hoja$PIB      0.7156     0.2893   2.474   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.023 on 30 degrees of freedom
Multiple R-squared:  0.1694,    Adjusted R-squared:  0.1417
F-statistic:  6.12 on 1 and 30 DF,  p-value: 0.01925
```

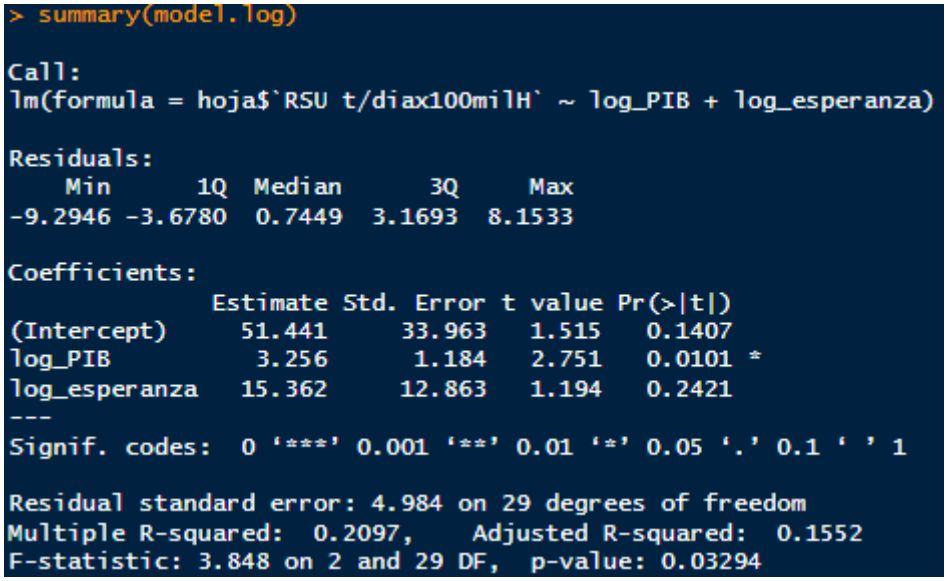
Figura 9: Resultado de selección stepwise

5.2.1. Tratamiento de datos

En todos los métodos de selección solo tenemos como variable significativa al PIB, aunque aún trataremos la variable PIB y Esperanza educativa (con transformación logarítmica y su recíproco) para comprobar que solo el PIB es significativo. Cabe aclarar que usamos Esperanza educativa por ser la segunda variable más cercana a ser significativa.

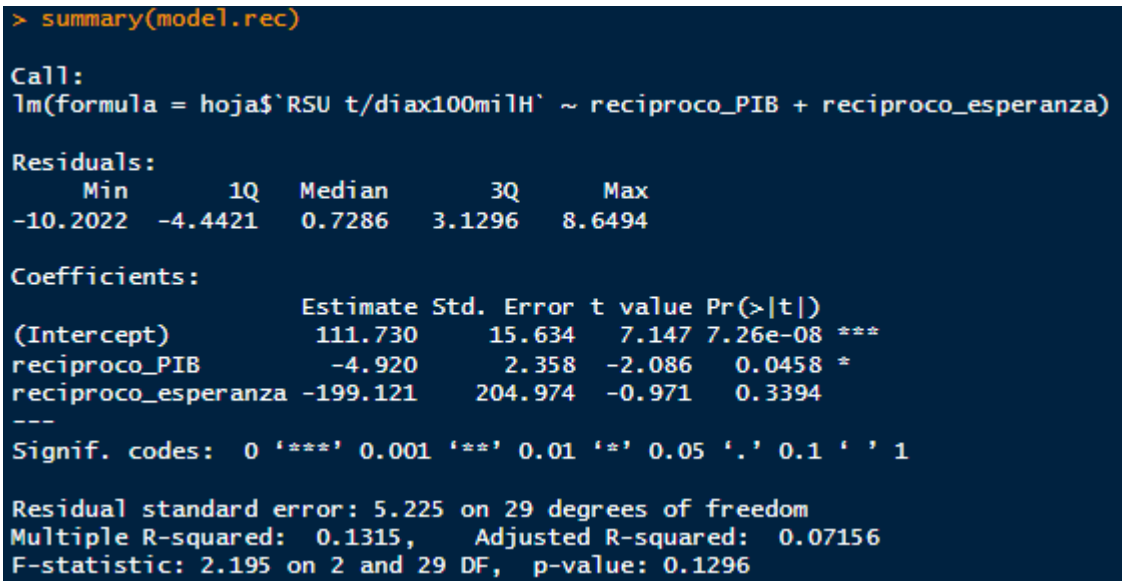
```
#transformar con logaritmos las variables más cercanas a ser significativas
log_PIB <- log(hoja$PIB)
log_esperanza <- log(hoja$`Esperanza educativa`)
#transformar a reciprocos las variables más cercanas a ser significativas
reciproco_PIB <- 1 / hoja$PIB
reciproco_esperanza <- 1 / hoja$`Esperanza educativa`
#hacer el modelo de regresion con los nuevos logaritmos
model.log<-lm(formula = hoja$`RSU t/diax100milH` ~ log_PIB + log_esperanza)
#hacer el modelo de regresion con los reciprocos
model.rec<-lm(formula=hoja$`RSU t/diax100milH` ~ reciproco_PIB + reciproco_esperanza)
#combinar reciprocos y logaritmos
model.logrec<-lm(formula=hoja$`RSU t/diax100milH` ~ log_PIB + reciproco_esperanza)
model.reclog<-lm(formula=hoja$`RSU t/diax100milH` ~ reciproco_PIB + log_esperanza)
```

Figura 10: Código en R para logaritmo y recíproco



```
> summary(model.log)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ log_PIB + log_esperanza)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2946 -3.6780  0.7449  3.1693  8.1533

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.441     33.963   1.515   0.1407
log_PIB         3.256      1.184   2.751   0.0101 *
log_esperanza  15.362     12.863   1.194   0.2421
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.984 on 29 degrees of freedom
Multiple R-squared:  0.2097,    Adjusted R-squared:  0.1552
F-statistic: 3.848 on 2 and 29 DF,  p-value: 0.03294
```

Figura 11: Resultados regresión con logaritmos



```
> summary(model.rec)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ reciproco_PIB + reciproco_esperanza)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2022  -4.4421   0.7286   3.1296   8.6494

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         111.730     15.634   7.147 7.26e-08 ***
reciproco_PIB        -4.920      2.358  -2.086   0.0458 *
reciproco_esperanza -199.121    204.974  -0.971   0.3394
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.225 on 29 degrees of freedom
Multiple R-squared:  0.1315,    Adjusted R-squared:  0.07156
F-statistic: 2.195 on 2 and 29 DF,  p-value: 0.1296
```

Figura 12: Resultados regresión con reciproco

```
> summary(model.logrec)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ log_PIB + reciproco_esperanza)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3667 -3.7246  0.6799  3.1559  8.1562

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         105.736     13.774   7.676 1.83e-08 ***
log_PIB               3.182      1.190   2.674   0.0122 *
reciproco_esperanza -191.744    191.085  -1.003   0.3239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.018 on 29 degrees of freedom
Multiple R-squared:  0.1987,    Adjusted R-squared:  0.1434
F-statistic: 3.595 on 2 and 29 DF,  p-value: 0.04028
```

Figura 13: Resultados regresión con PIB logaritmo y Esperanza recíproco

```
> summary(model.reclog)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ reciproco_PIB + log_esperanza)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1262  -4.3922   0.8662   3.1638   8.1155

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     54.173     35.755   1.515   0.1406
reciproco_PIB   -5.161      2.358  -2.189   0.0368 *
log_esperanza   16.479     13.867   1.188   0.2444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.184 on 29 degrees of freedom
Multiple R-squared:  0.1448,    Adjusted R-squared:  0.08586
F-statistic: 2.456 on 2 and 29 DF,  p-value: 0.1035
```

Figura 14: Resultados regresión con PIB recíproco y Esperanza logaritmo

Se aprecia como el valor con de PIB con la transformación a logaritmo es más significativo que incluso el recíproco de PIB, por ende se hará un

modelo de regresión con el logaritmo de PIB y nuestra variable objetivo que es **RSU t/diax100milH**.

```
> #modelo de regresión lineal
> model.logPIB<-lm(formula=hoja$`RSU t/diax100milH`~ log_PIB)
> summary(model.logPIB)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ log_PIB)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5044 -3.7434  0.7306  3.1201 11.0536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   91.973      1.274  72.214   <2e-16 ***
log_PIB        2.818      1.134   2.486   0.0187 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.019 on 30 degrees of freedom
Multiple R-squared:  0.1709,    Adjusted R-squared:  0.1432
F-statistic: 6.182 on 1 and 30 DF,  p-value: 0.0187
```

Figura 15: Modelo de regresión lineal

5.2.2. Comprobación de supuestos teóricos

Para afirmar que existe un modelo de regresión lineal, debemos comprobar sus supuestos teóricos.

```
> #comprobación de supuestos teóricos
> residuales <- residuals(model.logPIB)
> #media cero (independencia)
> t.test(residuales)

        One Sample t-test

data:  residuales
t = 3.2e-16, df = 31, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.780053  1.780053
sample estimates:
   mean of x
2.792905e-16
```

Figura 16: Supuesto media cero

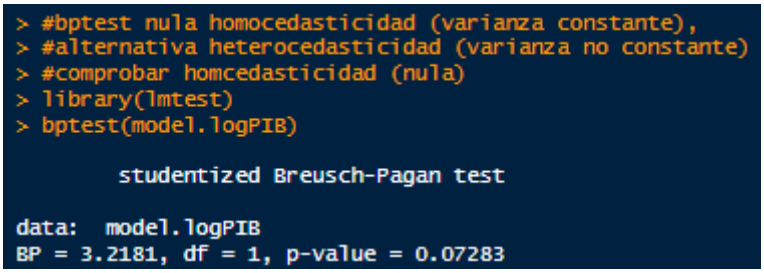


Figura 17: Supuesto homocedasticidad

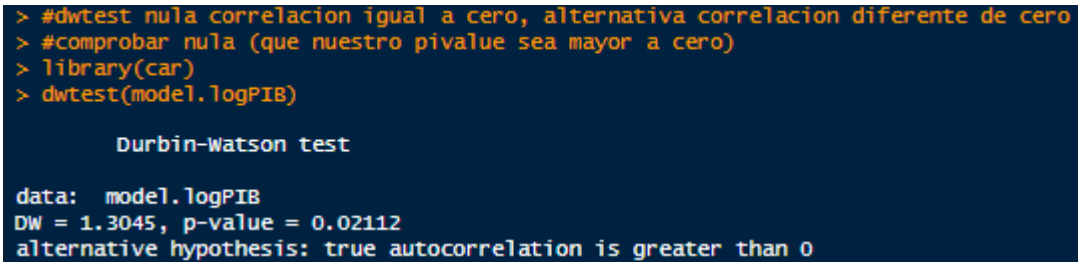


Figura 18: Supuesto correlación

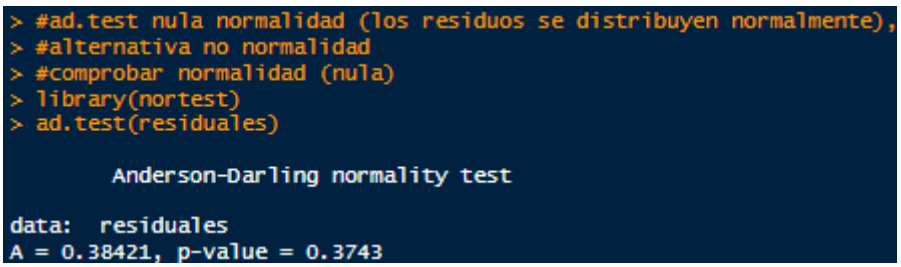


Figura 19: Supuesto residuos se distribuyen normalmente

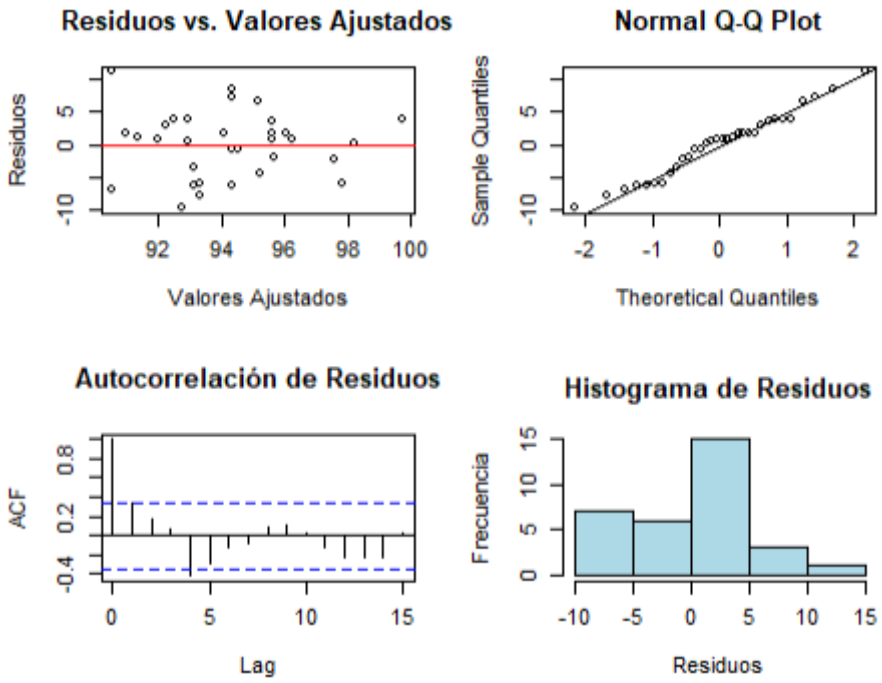


Figura 20: Supuestos de manera gráfica

Los supuestos se cumplen, con la prueba t.test la media es cero. No se rechaza la hipótesis nula de Breusch-Pagan (que indica homocedasticidad por el pivalue>.05). Con la prueba Durbin-Watson la hipótesis alternativa plantea que existe almenos una relación lineal entre mis datos (cosa que se comprueba al rechazar la hipótesis nula). En la prueba Anderson-Darling la hipótesis nula indica que los datos se distribuyen normalmente, con lo cual no existe suficiente evidencia estadística para descartarla. Gráficamente nuestros residuos no están oscilando en el cero, por ende corregiremos esta anomalía aplicando una transformación logarítmica a nuestra variable dependiente.

5.2.3. Corrección de supuestos

```
> #como los supuestos no se cumplen vamos a transformar la variable objetivo
> logRSU<- log(hoja$`RSU t/diax100milH`)
> model.loglog<-lm(formula=logRSU~ log_PIB)
> summary(model.logPIB)

Call:
lm(formula = hoja$`RSU t/diax100milH` ~ log_PIB)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5044 -3.7434  0.7306  3.1201 11.0536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   91.973      1.274  72.214   <2e-16 ***
log_PIB        2.818      1.134   2.486   0.0187 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.019 on 30 degrees of freedom
Multiple R-squared:  0.1709,    Adjusted R-squared:  0.1432
F-statistic: 6.182 on 1 and 30 DF,  p-value: 0.0187
```

Figura 21: Modelo de regresión con variable dependiente transformada

```
> #comprobación de supuestos teóricos
> residualeslog <- residuals(model.loglog)
> #media cero (independencia)
> t.test(residualeslog)

        One Sample t-test

data:  residualeslog
t = -1.9983e-16, df = 31, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01908774  0.01908774
sample estimates:
    mean of x
-1.870249e-18
```

Figura 22: Media cero con variable dependiente transformada

```
> #bptest nula homocedasticidad (varianza constante),
> #alternativa heterocedasticidad (varianza no constante)
> #comprobar homcedasticidad (nula)
> library(lmtest)
> bptest(model.loglog)

        studentized Breusch-Pagan test

data:  model.loglog
BP = 3.7774, df = 1, p-value = 0.05195
```

Figura 23: Homocedasticidad con variable dependiente transformada

[H]

```
> #dwtest nula correlacion igual a cero, alternativa correlacior
> #comprobar alternativa (que nuestro pivalue sea mayor a cero)
> library(car)
> dwtest(model.loglog)

        Durbin-Watson test

data:  model.loglog
DW = 1.3226, p-value = 0.02414
alternative hypothesis: true autocorrelation is greater than 0
```

Figura 24: Correlación con variable dependiente transformada

```
> #ad.test nula normalidad (los residuos se distribuyen normalmente),
> #alternativa no normalidad
> #comprobar normalidad (nula)
> library(nortest)
> ad.test(residualeslog)

        Anderson-Darling normality test

data:  residualeslog
A = 0.40512, p-value = 0.3334
```

Figura 25: Distribucion normal con variable dependiente transformada

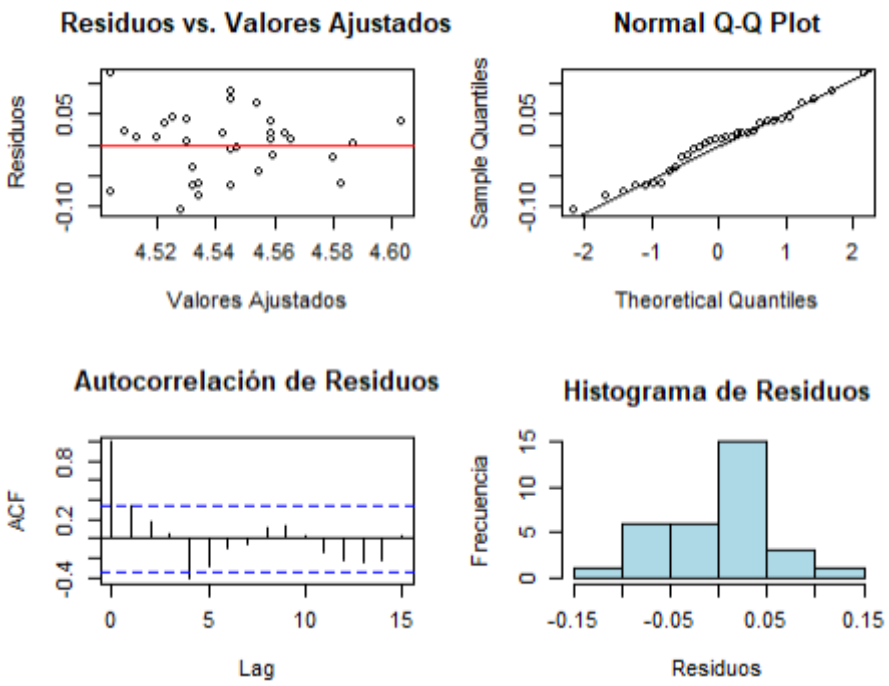


Figura 26: Supuestos gráficamente con variable dependiente transformada

Todos los supuestos se siguen cumpliendo por las razones que en el modelo anterior fueron especificadas, lo que se corrigió fue la anomalía de los residuos, ahora gráficamente sí están alrededor del cero. Por ende tenemos modelo de regresión lineal dado por la ecuación $y = 91{,}973 + 2{,}818x$ donde x = PIB.

6. Conclusión

A través del análisis multivariante encontramos que los métodos de selección de variables son útiles para el ahorro de tiempo. Esto se debe a que el mismo software R encontró que las variables significativas solo eran la independiente y el PIB sin tener que haber hecho todas las transformaciones de los datos. Con respecto a la generación de residuos sólidos urbanos (RSU) en México, encontramos que con el PIB de cada estado, podemos predecir cuantos RSU van a generar.
A diferencia de lo que se puede llegar a creer con intuición (que los RSU están relacionados con la pobreza por familia, nivel educativo, entre todas las otras variables estudidadas), realmente lo que está aportando a la generación de RSU es el modelo social que manejamos basado en la economía y en generar más ingresos.

7. Referencias

- Kaza, S., Yao, L., Bhada-Tata, P., Van der Leven, F., & Wetzler, K. (2018). What a waste 2.0: A global snapshot of solid waste management to 2050. World Bank Group.
- Gujarati, D. N. (2003). Basic Econometrics. McGraw-Hill Education.
- Johnston, J. (1963). Econometric Methods. McGraw-Hill.
- Greene, W. H. (2012). Econometric Analysis. Pearson Education.
- Wooldridge, J. M. (2016). Ïntroductory Econometrics: A Modern Approach". Cengage Learning.
- Andrei. (2024, 3 febrero). Transformations: log, reciprocal, vector, linear. . . - Statistics How to. Statistics How To. `https://www.statisticshowto.com/transformations/#reciprocal`
- De Anda Trasviña, A., García Galindo, E., Peña Castañón, A., Seminario-Peña, J., Nieto Garibay, A. (2021). Residuos orgánicos, ¿basura o recurso? Recursos Naturales y Sociedad, 7(3), 22. Variable RSU recuperada de: https://www.cibnor.gob.mx/revista-rns/pdfs/vol1num3EE/3$_R ESIDUOS.pdf$
- INEGI. (2021). Producto interno bruto por entidad federativa 2020 (Boletín No. 727/21).Variable PBI recuperada de: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/pibe/PIBEntFed2020.pdf
- de Series de Tiempo. Variable incidencia delictiva recuperada de: https://inegi.org.mx/tablerosestadisticos/series$_d e_t iempo/$
- Instituto Nacional de Estadística y Geografía. (2020). Tabulado interactivo sobre educación, [Tabla de esperanza de escolaridad por entidad federativa]. variable de esperanza escolar recuperada de: https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Educacion$_E ducacion_1 3_c 457f93a-1497-43b9-8c16-962c4cf3af40$
- Instituto Nacional de Estadística y Geografía. (2020). Tabulado interactivo sobre hogares [Población en situación de pobreza por entidad federativa . Pobreza por 100 mil habitantes recuperado de : https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Hogares$_H ogares_1 5_9 954f9c6-9512-40c5-9cbf-1b2ce96283e4idrt=54opc=t$
- INEGI. (2020). Presentación de resultados de la Encuesta Nacional de Ingresos y Gastos de los Hogares 2020:[Archivo PDF]. Variable ingreso corriente recuperado de : https://www.inegi.org.mx/contenidos/programas/enigh/nc/2020/doc/enigh2020$_n s_p resentacion$

- Instituto Nacional de Estadística y Geografía. (2020). Tabulado interactivo sobre población [Tabla de densidad de población por entidad federativa]. Variable densidad poblacional recuperada de: https://en.www.inegi.org.mx/app/tabulados/interactivos/?pxq=Poblacion$_Poblacion_07fb7d5132$
$39f0-4a6c-b6f6-4cbe440e048d$