Taller 4 - Análisis descriptivo

Juliana Andrea Cagua Vega, María Paula Rincón Cepeda, Santiago Rojas Gómez

16/05/2021

## **Activación de paquetes**

library(tidyverse)  
library(readxl)  
library(knitr)  
library(ggplot2)  
library(scales)  
library(stargazer)  
library(ggthemes)  
library(GGally)  
library(ggpubr)

## **Desarrollo del taller**

**Nota:** *Con el código propuesto a continuación, se desarrolló el Punto 1, Punto 2 y Punto 3*

**Punto 1:** *Carguen la(s) base(s) de datos usando la función que corresponda de los paquetes readr, readxl o haven*

**Punto 2:** *De ser necesario, asegúrense que la base de datos esté lista para el análisis, usando la(s) funion(es) del paquete tidyr que correspondan. Sean cuidadosos con la eliminación de los valores faltantes. No los eliminen salvo que lo consideren esencial de manera justificada*

**Punto 3:** *Si están usando datos provenientes de más de una base de datos, unan las bases de datos usando la función que corresponda del paquete dplyr*

BasesWB <- list.files("WB/")  
  
for (i in seq\_along(BasesWB)) {  
 datos <- read\_xlsx(paste0("WB/", BasesWB[i]), skip = 3)  
 datos <- datos %>% select("Country Name", "1998", "2018")  
 datos <- datos %>% rename(Pais = `Country Name`)  
 assign(BasesWB[i],datos)  
}  
  
WB <- left\_join(Industria.xlsx, left\_join(Ingreso.xlsx, Poblacion.xlsx, by = "Pais"), by = "Pais")  
  
BasesWBG <- list.files("WBG/")  
  
for (i in seq\_along(BasesWBG)) {  
 datos <- read\_xlsx(paste0("WBG/", BasesWBG[i]))  
 datos <- datos %>% select("country", "1998", "2018")  
 datos <- datos %>% rename(Pais = `country`)  
 assign(BasesWBG[i],datos)  
}  
  
WBG <- left\_join(Emision.xlsx, IDH.xlsx, by = "Pais")  
Base\_Final <- left\_join(WB, WBG, by = "Pais")  
Base\_Final <- drop\_na(Base\_Final)

**Punto 4:** *De ser necesario, modifiquen las variables de interés usando el paquete dplyr para asegurar lo siguiente*

* Las variables deben ser adecuadamente identificadas por R según su tipo. Por ejemplo, todas las variables numéricas (continuas y discretas), deben identificarse como tal y no como carácteres, y todas las variables categóricas deben identificarse por R como factores ordenados o no ordenados, según el caso.
* La variables deben tener nombres que sean fáciles de entender. Por ejemplo, si la variable de género se llama P6020, deben modificar el nombre para que sea fácil entender cuál es el contenido de la variable
* Las variables binarias deben tener valores de 0 o 1. El valor de 1 debe corresponder con el nombre de la variable. Por ejemplo, si la variable se llama mujer, 1 corresponde a las mujeres y 0 a los hombres, o si la variable se llama bachiller, 1 debe corresponder a quienes son bachilleres y 0 a quienes no

*Renombrar las variables acorde a su información*

Base\_Final <- Base\_Final %>% rename(Industria98 = `1998`, Industria18 = `2018`, Ingreso98 = `1998.x.x`, Ingreso18 = `2018.x.x`, Poblacion98 = `1998.y.x`, Poblacion18 = `2018.y.x`, Emision98 = `1998.x.y`, Emision18 = `2018.x.y`, IDH98 = `1998.y.y`, IDH18 = `2018.y.y`)

*Aplicar logaritmo natural a las variables establecidas en el modelo y seleccionar las variables a utilizar*

Base\_Final <- Base\_Final %>% mutate(Ln\_Ingreso98 = log(Ingreso98), Ln\_Ingreso18 = log(Ingreso18), Ln\_Poblacion98 = log(Poblacion98), Ln\_Poblacion18 = log(Poblacion18))  
  
Base\_Final <- Base\_Final %>% select(c("Pais", "Industria98", "Industria18", "Emision98", "Emision18", "IDH98", "IDH18", "Ln\_Ingreso98", "Ln\_Ingreso18", "Ln\_Poblacion98", "Ln\_Poblacion18"))

*Verificar la clase de cada una de las variables*

str(Base\_Final)

## tibble[,11] [118 x 11] (S3: tbl\_df/tbl/data.frame)  
## $ Pais : chr [1:118] "Albania" "United Arab Emirates" "Argentina" "Australia" ...  
## $ Industria98 : num [1:118] 15.3 40 26.7 25.4 28.5 ...  
## $ Industria18 : num [1:118] 21.3 47 23.1 24.2 25.5 ...  
## $ Emision98 : num [1:118] 0.563 28.7 3.87 18 8.35 3.97 0.0403 12.7 0.185 0.189 ...  
## $ Emision18 : num [1:118] 1.59 21.4 4.41 16.9 7.75 3.7 0.0467 8.69 0.623 0.531 ...  
## $ IDH98 : num [1:118] 0.649 0.769 0.752 0.892 0.828 0.627 0.287 0.866 0.385 0.453 ...  
## $ IDH18 : num [1:118] 0.791 0.866 0.83 0.938 0.914 0.754 0.423 0.919 0.52 0.614 ...  
## $ Ln\_Ingreso98 : num [1:118] 8.01 11.44 9.39 10.09 10.19 ...  
## $ Ln\_Ingreso18 : num [1:118] 9.54 11.14 10.06 10.85 10.98 ...  
## $ Ln\_Poblacion98: num [1:118] 4.74 3.68 2.58 0.89 4.57 ...  
## $ Ln\_Poblacion18: num [1:118] 4.65 4.91 2.79 1.18 4.67 ...

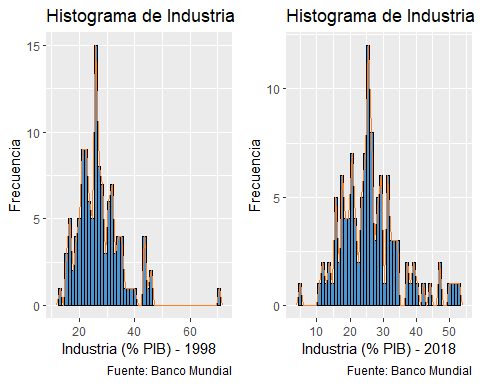
**Punto 5:** *Usen alguna de las siguientes funciones para realizar una tabla de estadísticas descriptivas para las variables que van a incluir en su estudio, según sus necesidades: summarize, statdesc(), summary(), stargazer(), datasummary() o datasummary\_balance. Si su variable dependiente o la principal variable independiente es una variable categórica, realice una distribución de frecuencias que incluya las frecuencias absoluta y relativa, y de ser el caso, las frecuencias absoluta y relativa acumuladas. Incluya esta tabla en el RMarkdown usando la función kable del paquete knitr, o stargazer directamente, según sea el caso*

stargazer(as.data.frame(Base\_Final), type = "text", out = "tabla1.tex", title = "Tabla 1 - Estadísticas Descriptivas")

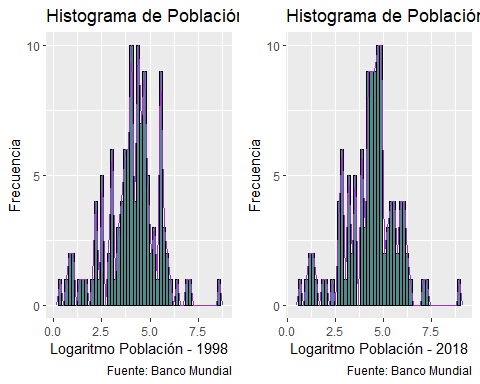
##   
## Tabla 1 - Estadísticas Descriptivas  
## ==================================================================  
## Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max   
## ------------------------------------------------------------------  
## Industria98 118 27.202 8.166 12.566 22.069 31.143 70.147  
## Industria18 118 26.119 8.942 5.270 20.150 30.335 53.327  
## Emision98 118 4.292 4.940 0.037 0.698 6.800 28.700  
## Emision18 118 4.349 4.937 0.024 1.053 6.240 31.300  
## IDH98 118 0.629 0.167 0.248 0.511 0.758 0.906   
## IDH18 118 0.734 0.148 0.377 0.629 0.857 0.954   
## Ln\_Ingreso98 118 8.608 1.159 6.063 7.806 9.404 11.436  
## Ln\_Ingreso18 118 9.491 1.143 6.659 8.601 10.403 11.514  
## Ln\_Poblacion98 118 4.048 1.401 0.416 3.407 4.819 8.676   
## Ln\_Poblacion18 118 4.305 1.361 0.711 3.585 4.977 8.981   
## ------------------------------------------------------------------

**Punto 6:** *Realicen un histograma y/o gráfico de barras, según se trate de una variable cuantitativa o categórica, para la variable dependiente y para la principal variable independiente. Las gráficas las deben realizar usando el paquete ggplot, y deben modificarlas en cuanto a número de bins, títulos, subtítulos, etiquetas de los ejes, escalas numéricas de los ejes y colores de una manera que permitan visualizar con claridad la información*

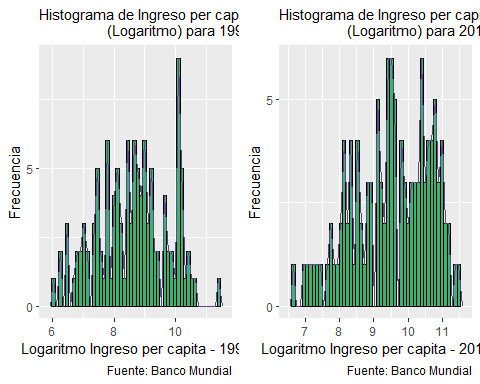
a1 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Industria98)) +  
 geom\_histogram(bins = 50, fill = "steelblue3", color = "black") +   
 geom\_freqpoly(bins = 50, color = "chocolate1") +  
 labs(title = "Histograma de Industria para 1998", x = "Industria (% PIB) - 1998", y = "Frecuencia", caption = "Fuente: Banco Mundial") +  
 scale\_y\_continuous(breaks = seq(from = 0, to = 15, by = 5))  
  
  
a2 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Industria18)) +  
 geom\_histogram(bins = 50, fill = "steelblue3", color = "black") +   
 geom\_freqpoly(bins = 50, color = "chocolate1") +  
 labs(title = "Histograma de Industria para 2018", x = "Industria (% PIB) - 2018", y = "Frecuencia", caption = "Fuente: Banco Mundial") +   
 scale\_y\_continuous(breaks = seq(from = 0, to = 15, by = 5))  
  
ggarrange(a1, a2)



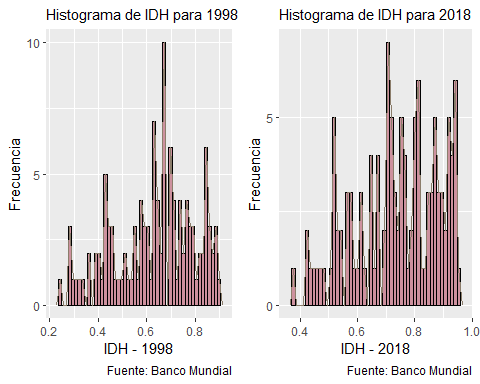
b1 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Ln\_Poblacion98)) +  
 geom\_histogram(bins = 50, fill = "darkslategray4", color = "black") +   
 geom\_freqpoly(bins = 50, color = "darkorchid3") +  
 labs(title = "Histograma de Población (Logaritmo) para 1998", x = "Logaritmo Población - 1998", y = "Frecuencia", caption = "Fuente: Banco Mundial") +  
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
  
  
b2 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Ln\_Poblacion18)) +  
 geom\_histogram(bins = 50, fill = "darkslategray4", color = "black") +   
 geom\_freqpoly(bins = 50, color = "darkorchid3") +  
 labs(title = "Histograma de Población (Logaritmo) para 2018", x = "Logaritmo Población - 2018", y = "Frecuencia", caption = "Fuente: Banco Mundial") +   
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
ggarrange(b1, b2)



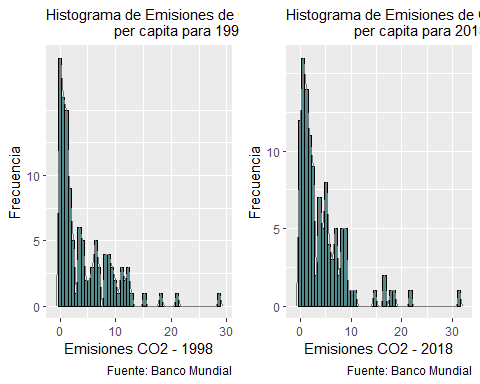
c1 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Ln\_Ingreso98)) +  
 geom\_histogram(bins = 50, fill = "mediumseagreen", color = "black") +   
 geom\_freqpoly(bins = 50, color = "mediumpurple4") +  
 labs(subtitle = "Histograma de Ingreso per capita   
 (Logaritmo) para 1998", x = "Logaritmo Ingreso per capita - 1998", y = "Frecuencia", caption = "Fuente: Banco Mundial") +  
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
  
  
c2 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Ln\_Ingreso18)) +  
 geom\_histogram(bins = 50, fill = "mediumseagreen", color = "black") +   
 geom\_freqpoly(bins = 50, color = "mediumpurple4") +  
 labs(subtitle = "Histograma de Ingreso per capita  
 (Logaritmo) para 2018", x = "Logaritmo Ingreso per capita - 2018", y = "Frecuencia", caption = "Fuente: Banco Mundial") +   
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
ggarrange(c1, c2)



d1 <- ggplot(data = Base\_Final,   
 mapping = aes(x = IDH98)) +  
 geom\_histogram(bins = 50, fill = "pink3", color = "black") +   
 geom\_freqpoly(bins = 50, color = "bisque4") +  
 labs(subtitle = "Histograma de IDH para 1998", x = "IDH - 1998", y = "Frecuencia", caption = "Fuente: Banco Mundial") +  
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
  
  
d2 <- ggplot(data = Base\_Final,   
 mapping = aes(x = IDH18)) +  
 geom\_histogram(bins = 50, fill = "pink3", color = "black") +   
 geom\_freqpoly(bins = 50, color = "bisque4") +  
 labs(subtitle = "Histograma de IDH para 2018", x = "IDH - 2018", y = "Frecuencia", caption = "Fuente: Banco Mundial") +   
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
ggarrange(d1, d2)



e1 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Emision98)) +  
 geom\_histogram(bins = 50, fill = "darkslategray4", color = "black") +   
 geom\_freqpoly(bins = 50, color = "gray48") +  
 labs(subtitle = "Histograma de Emisiones de CO2  
 per capita para 1998", x = "Emisiones CO2 - 1998", y = "Frecuencia", caption = "Fuente: Banco Mundial") +  
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
  
  
e2 <- ggplot(data = Base\_Final,   
 mapping = aes(x = Emision18)) +  
 geom\_histogram(bins = 50, fill = "darkslategray4", color = "black") +   
 geom\_freqpoly(bins = 50, color = "gray48") +  
 labs(subtitle = "Histograma de Emisiones de CO2  
 per capita para 2018", x = "Emisiones CO2 - 2018", y = "Frecuencia", caption = "Fuente: Banco Mundial") +   
 scale\_y\_continuous(breaks = seq(from = 0, to = 10, by = 5))  
  
ggarrange(e1, e2)



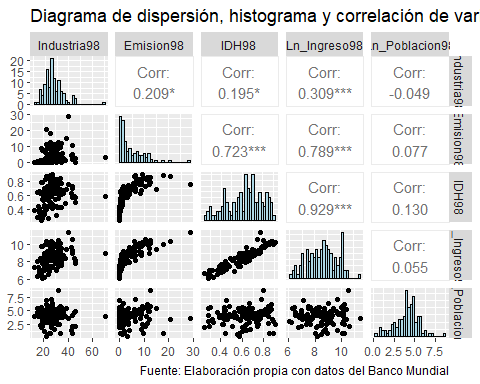
**Nota:** *Con el código propuesto a continuación, se desarrolló el Punto 7 y Punto 8*

**Punto 7:** *Para las variables cuantitativas (continuas y discretas), calculen la correlación entre cada pareja de variables y presenten estas correlaciones en una matriz de correlaciones. Para las variables categóricas hagan una tabla cruzada para cada par de variables. En ambos casos, comente sobre los resultados más relevantes*

**Punto 8:** *Realicen una gráfica que les permita relacionar la variable dependiente y la principal variable independiente de interés. Si tienen varias variables independientes de interés, realicen una gráfica para mostrar la relación entre cada una de estas variables y la variable dependiente. Realicen las transformaciones en las variables que consideren necesarias para mostrar de mejor manera la información en estas gráficas. Las gráficas las deben realizar usando el paquete ggplot, y deben modificar los títulos, etiquetas de los ejes, escalas numéricas de los ejes y colores de una manera que permitan visualizar con claridad la información.*

Base\_Final98 <- Base\_Final %>%   
 select(ends\_with("98"))  
ggpairs(Base\_Final98,diag = list(continuous = wrap("barDiag", fill = "lightblue", colour= "black")))+ labs(title = "Diagrama de dispersión, histograma y correlación de variables de estudio para 1998", caption = "Fuente: Elaboración propia con datos del Banco Mundial")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

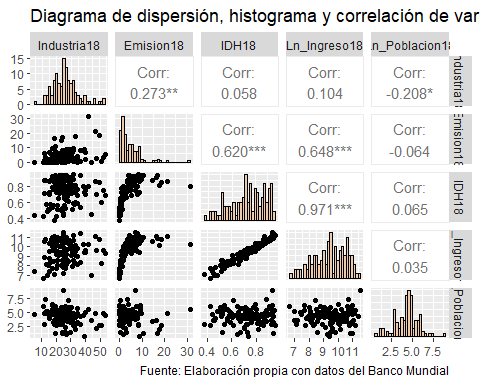


Estadísticamente se usa la correlación entre variables para estudiar similitudes en su comportamiento. Cuando el coeficiente de correlación es positivo cercano a 0.5, indica una relación positiva moderada y entre más cercano a 1.0, indica una relación positiva perfecta entre las variables. Por otro lado, entre más cercano sea el coeficiente de correlación a 0.0, indica una relación nula entre estas, mientras que si el coeficiente es negativo, indica una relación inversa entre las variables.

Teniendo en cuenta lo mencionado anteriormente, respecto a la correlación entre la variable dependiente (Emisión98) y las variables independientes (IDH98, Ln\_Ingreso98,Ln\_Población98 y la industria98) se observa que la relación con el logaritmo natural de la población y con la industria es baja presentando un coeficiente de 0.077 y 0.209, respectivamente. No obstante, la relación con el IDH y el logaritmo natural de los ingresos es alta representando valores de 0.723 y 0.789, respectivamente.

Base\_Final18 <- Base\_Final %>%   
 select(ends\_with("18"))  
ggpairs(Base\_Final18,diag = list(continuous = wrap("barDiag", fill = "peachpuff2", colour= "black")))+ labs(title = "Diagrama de dispersión, histograma y correlación de variables de estudio para 2018", caption = "Fuente: Elaboración propia con datos del Banco Mundial")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Respecto a la correlación entre la variable dependiente (Emisión18) y las variables independientes (IDH18, Ln\_Ingreso18,Ln\_Población18 y la industria18) se observa que la relación con logaritmo natural de la población es baja e inversa (-0.064) y para la industria es baja y positiva con un coeficiente de 0.273. Adicionalmente, la correlación con el IDH y el logaritmo natural de los ingresos es alta representando valores de 0.620 y 0.648, respectivamente.

Comparado con el año anterior, se observa un aumento mínimo en la correlación de la Emisión y la industria pero una disminución de casi un punto con el IDH y el logaritmo natural de los ingresos. Por otro lado, para el año 2018, la relación con el logaritmo natural de la población sigue siendo baja pero ahora es inversa.

**Punto 9:** *Usando la información presentada en las preguntas 6 a 8, ¿qué concluyen sobre la distribución de sus variables de interés y sobre la relación entre estas variables? Responda esta pregunta en al menos tres párrafos.:*

El histograma presenta las frecuencias de cada variable y sirve para determinar su distribución, en otros términos, sirve para conocer si se presentan sesgos en la variables o si están distribuidos de forma normal.De acuerdo con los histogramas del Punto 6 que nos permiten realizar una comparación de la distribución de las variables en dos periodos de tiempo diferentes (1998 y 2018), podemos concluir lo siguiente:

* Industria:Para el año 1998 se evidencia un sesgo positivo o a la derecha, presentando datos atípicos. Por otro lado, el comportamiento de la variable industria para el año 2018, es mas cercano a una distribución normal pues no hay un fuerte sesgo en la gráfica.
* Logaritmo de la población: Esta variable presenta una distribución cercana a la normal para ambos años pero se evidencia una mayor agrupación de datos en el lado izquierdo. Adicionalmente se presentan datos atípicos en la parte derecha del gráfico.
* Logaritmo del ingreso per cápita: Se observa que no presentan una distribución acampanada sino que esta es plana, lo cual podría indicar la presencia de errores en la medición.
* IDH: Se evidencia un leve sesgo hacia la izquierda para el año 1998, mientras que para el año 2018, la distribución no presenta forma acampanada por lo cual se deduce que puede existir un problema en la medición de la variable.
* Emisión de CO2: El año 1998 y 2018 presentan un sesgo a la derecha, es decir, tienen un sesgo positivo.

Con respecto al diagrama de dispersión que muestra la relación entre las variables;la distribución de los puntos indica la independencia de una variable con respecto a la otra, es decir, si existe relación fuerte, débil o intermedia dependiendo del comportamiento de la variable. Para el año 1998, se observa que La variable Industria98 presenta una mayor agrupación de la nube de puntos en el lado izquierdo con respecto a todas las variables (Emisión98, IDH98, Ln\_Ingreso98, Ln\_Poblacion98), mientras que la variable Ln\_Poblacion98 no muestra una relación con respecto al IDH98 y Ln\_Ingreso98, dada la alta dispersión que se observa, así como tampoco se evidencia una relación débil con Industria98 y Emisión98. Respecto a la Emisión se muestra una relación exponencial con las variables IDH y el Ingreso Per cápita. Es decir, a medida que aumentan estas últimas variables, se espera que la emisión de co2 aumente. Por otro lado la dispersión del IDH98 con Ln\_Ingreso98 presenta una relación fuerte y positiva, esta puede ser explicada por que el IDH tiene en cuenta los ingresos per cápita.

Para el año 2018, se evidencia una mayor dispersión entre variables con respecto al año 1998; en términos de la variable Industria18, no se evidencia una relación entre Ln\_Población18, IDH18 y Ln\_Ingreso18.Por otro lado, la variable Emision18 presenta una relación con todas las variables. Respecto a la relación con la Industria18 se observa una dispersión centrada en el límite inferior; con IDH18 y el Ln\_Ingreso18, se presenta una relación exponencial y, con Ln\_Población18 los valores están concentrados en la parte izquierda de la gráfica. Por último, el IDH18 con Ln\_Ingreso18, así como en el año presentado anteriormente, tienen una fuerte relación lineal.