

# Entrega Final: Proyecto programación

31/05/2021

## Integrantes de grupo.

- *Joan Galeano R*
- *Nicolás González J*
- *Alejandro Guevara H*

## Buen desempeño económico, la clave del éxito para un rendimiento sobresaliente en los juegos olímpicos.

El buen desempeño económico de los países está dado por una multiplicidad de condiciones que resulta difícil mencionar a cabalidad. Para este trabajo de investigación, se explicará el buen desempeño económico desde el análisis de los agregados macroeconómicos fundamentales, como lo son: la Inflación, el PIB (Producto Interno Bruto) y la Tasa de Desempleo. De esta manera, se podrá inferir que un país 1 con un nivel de desempeño óptimo desde la óptica económica será aquel cuyos agregados monetarios presenten valores ajustados a sus rangos objetivo. Por otro lado, se decidió incluir el tamaño poblacional de los países bajo el entendido de que un mayor número de personas pueden ser más productivas, y mejorar el desempeño económico de los países. Teniendo en cuenta lo anterior, se explicará el rendimiento de determinados países en los juegos olímpicos de verano, desde su desempeño económico (Inflación, Desempleo y Producto interno Bruto) y su tamaño poblacional. Para medir el rendimiento de los países en los Juegos Olímpicos, se tendrá en cuenta el total de medallas ganadas durante el periodo de 1950 hasta 2021.

## Paquetes

Para cumplir el objetivo del trabajo hacemos uso de RStudio como herramienta metodológica y de análisis. Se utilizan varios paquetes para el analisis, para el manejo de datos, graficos y modelos entre otros.

```
library(tidyverse)
library(rvest)
library(haven)
library(wbstats)
library(dplyr)
library(naniar)
library(knitr)
library(ggthemes)
library(readxl)
library(GGally)
library(Hmisc)
library(corrplot)
library(PerformanceAnalytics)
library(stargazer)
```

```
library(lmtest)
library(nortest)
library(lindia)
library(car)
```

## Manejo de la base de datos

### Variable explicada

Organizamos la base de datos, eliminamos datos faltantes y renombramos algunos países para evitar problemas (p.e. The Bahamas -> Bahamas, The), de esta manera, todos los nombres coinciden en las diferentes bases de datos sin afectar los datos de interés.

```
pagina <- "http://www.olympedia.org/statistics/medal/country"
pagina_desc <- read_html(pagina)

países <- pagina_desc %>% html_nodes("td:nth-child(1)") %>% html_text()

medallas <- pagina_desc %>% html_nodes("td:nth-child(6)") %>% html_text()
medallas <- as.integer(medallas)
medallas_por_pais <- tibble(países, medallas)

medallas_por_pais[4,1] <- "United Kingdom"
medallas_por_pais[7,1] <- "China"
medallas_por_pais[18,1] <- "Korea, Rep."
medallas_por_pais[43,1] <- "Iran, Islamic Rep."
medallas_por_pais[50,1] <- "Slovak Republic"
medallas_por_pais[59,1] <- "Egypt, Arab Rep."
medallas_por_pais[62,1] <- "Bahamas, The"
medallas_por_pais[83,1] <- "Venezuela, RB"
medallas_por_pais[84,1] <- "Serbia"
medallas_por_pais[97,1] <- "Cote d'Ivoire"
medallas_por_pais[98,1] <- "Hong Kong SAR, China"
medallas_por_pais[113,1] <- "Moldova"
medallas_por_pais[117,1] <- "Tanzania"
medallas_por_pais[121,1] <- "Kyrgyz Republic"
medallas_por_pais[122,1] <- "Saudi Arabia"
```

### Variables explicativas

En este caso, se presenta la Tasa de crecimiento del PIB como primera variable explicativa.

```
growth_gdp <- wb_data("NY.GDP.PCAP.KD.ZG", start_date = 1950, end_date = 2016)
growth_gdp <- tibble(growth_gdp$country, growth_gdp$date, growth_gdp$NY.GDP.PCAP.KD.ZG)
growth_gdp <- growth_gdp %>%
  rename(países = `growth_gdp$country`,
         fecha = `growth_gdp$date`,
         growth = `growth_gdp$NY.GDP.PCAP.KD.ZG`
  )
growth_gdp <- drop_na(growth_gdp)
```

```

growth_gdp_prom <- aggregate(growth_gdp$growth, list(growth_gdp$países), FUN=mean)
growth_gdp_prom <- growth_gdp_prom %>%
  rename(
    países = Group.1, GDP=x
  ) %>%
  mutate(GDP2=GDP^2)

```

A continuación, se usará la Tasa de crecimiento de la población.

```

growth_pob <- wb_data("SP.POP.GROW", start_date = 1950, end_date = 2016)
growth_pob <- tibble(growth_pob$country, growth_pob$date, growth_pob$SP.POP.GROW)
growth_pob <- growth_pob %>%
  rename(
    países = "growth_pob$country",
    fecha = "growth_pob$date",
    growth_p = "growth_pob$SP.POP.GROW"
  )
growth_pob <- drop_na(growth_pob)

growth_pob_prom <- aggregate(growth_pob$growth_p, list(growth_pob$países), FUN=mean)
growth_pob_prom <- growth_pob_prom %>%
  rename(
    países = Group.1, POB = x
  ) %>%
  mutate(POB2=POB^2)

```

También se decidió incluir la Tasa promedio de paro.

```

desempleo_total <- wb_data("SL.UEM.TOTL.ZS", start_date = 1950, end_date = 2016)
desempleo_total <- tibble(desempleo_total$country, desempleo_total$date, desempleo_total$SL.UEM.TOTL.ZS)
desempleo_total <- desempleo_total %>%
  rename(
    países = "desempleo_total$country",
    fecha = "desempleo_total$date",
    desempleo_t = "desempleo_total$SL.UEM.TOTL.ZS"
  )
desempleo_total <- drop_na(desempleo_total)
desempleo_total_prom <- aggregate(desempleo_total$desempleo_t, list(desempleo_total$países), FUN=mean)
desempleo_total_prom <- desempleo_total_prom %>%
  rename(
    países = Group.1, DESP=x
  )

```

Se incluirá la Tasa promedio de inflación para explicar su efecto parcial sobre el desempeño deportivo en los juegos olímpicos.

```

inflacion <- wb_data("NY.GDP.DEFL.KD.ZG", start_date = 1950, end_date = 2016)
inflacion <- tibble(inflacion$country, inflacion$date, inflacion$NY.GDP.DEFL.KD.ZG)
inflacion <- inflacion %>%
  rename(
    países = "inflacion$country",
    fecha = "inflacion$date",

```

```

    inflacion_t ="inflacion$NY.GDP.DEFL.KD.ZG"
  )
inflacion <- drop_na(inflacion)

inflacion_prom <- aggregate(inflacion$inflacion_t, list(inflacion$países), FUN=mean)
inflacion_prom <-inflacion_prom %>%
  rename(
    países = Group.1,INF=x
  ) %>%
  mutate(INF2=INF^2)

```

Finalmente, se hace un proceso de adjunción de todas las variables en un mismo objeto.

```

datos <- left_join(medallas_por_pais, growth_pob_prom,by = "países")
datos1 <- left_join(datos,growth_gdp_prom,by = "países")
datos2 <- left_join(datos1, desempleo_total_prom,by = "países")
tidy_data<- left_join(datos2, inflacion_prom,by = "países")

tidy_data <- drop_na(tidy_data)
tidy_data <- tidy_data %>%
  relocate(países,medallas,GDP,POB,DESP,INF)
head(tidy_data)

```

```

## # A tibble: 6 x 9
##   países      medallas  GDP  POB  DESP  INF  POB2  GDP2  INF2
##   <chr>      <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 United States    2847  2.00  1.04  6.07  3.40  1.08    4.00  11.6
## 2 Germany          1019  1.88  0.220  7.72  2.55  0.0483  3.52  6.51
## 3 United Kingdom    898  2.03  0.402  6.74  5.61  0.161  4.13  31.4
## 4 France           863  2.18  0.640  9.88  4.28  0.410  4.74  18.3
## 5 Italy            716  2.11  0.337  9.90  6.65  0.114  4.47  44.3
## 6 China           608  6.85  1.30   3.88  3.60  1.68  46.9  13.0

```

## Estadísticas descriptivas

```

resumen_países <- tidy_data %>%
  summary()
resumen_países

```

```

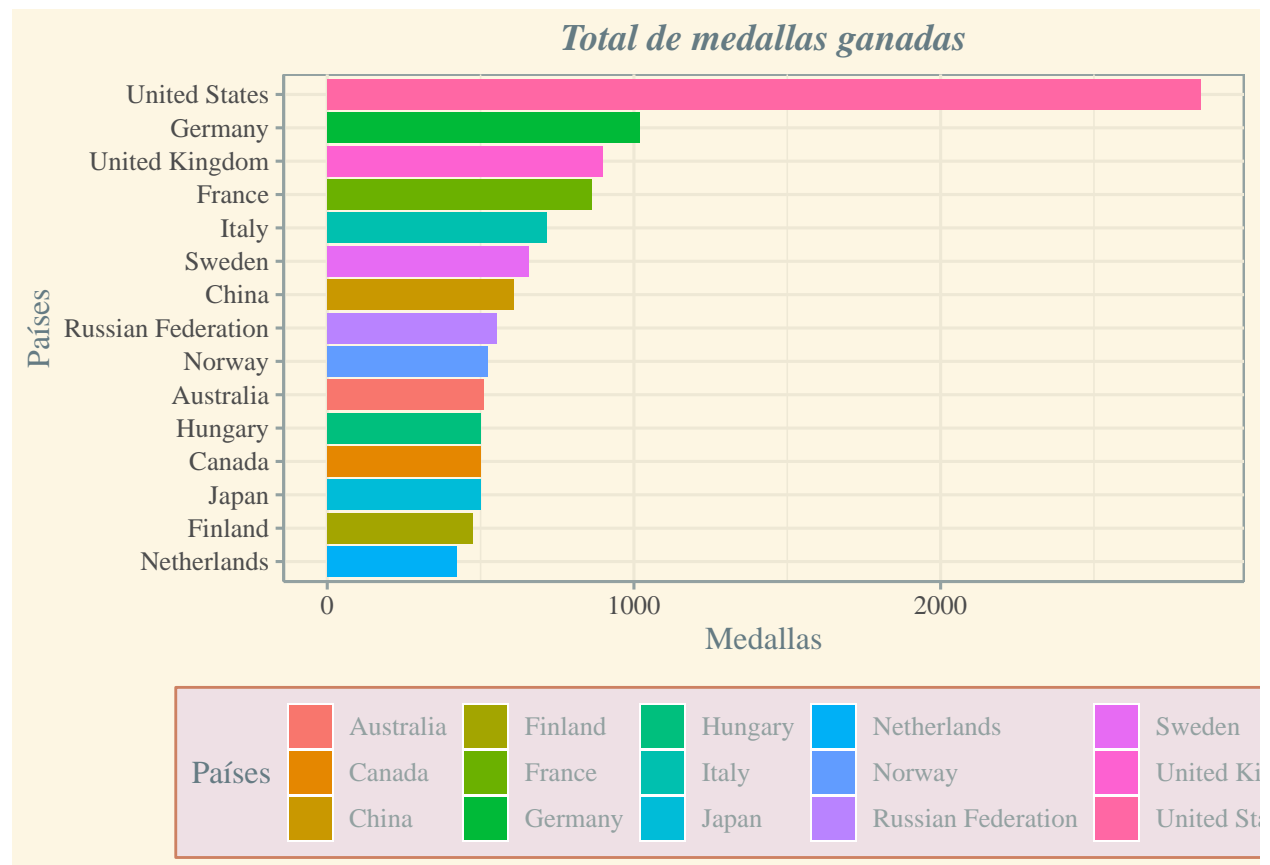
##   países      medallas      GDP      POB
## Length:120    Min.   : 1.00    Min.   : -1.976    Min.   : -0.1763
## Class :character 1st Qu.: 2.00    1st Qu.: 1.453    1st Qu.: 0.5564
## Mode  :character Median : 14.50    Median : 2.210    Median : 1.6006
##              Mean   : 136.15    Mean   : 2.274    Mean   : 1.6837
##              3rd Qu.: 97.75    3rd Qu.: 2.866    3rd Qu.: 2.3548
##              Max.   :2847.00    Max.   : 6.848    Max.   : 8.2464
##   DESP      INF      POB2      GDP2
## Min.   : 0.4885    Min.   : 1.201    Min.   : 0.00022    Min.   : 0.02327
## 1st Qu.: 4.0865    1st Qu.: 4.682    1st Qu.: 0.30972    1st Qu.: 2.13591
## Median : 7.2510    Median : 7.670    Median : 2.56410    Median : 4.88603

```

```
## Mean : 8.3105 Mean : 26.795 Mean : 4.69046 Mean : 7.23571
## 3rd Qu.:11.2228 3rd Qu.: 17.474 3rd Qu.: 5.54519 3rd Qu.: 8.21238
## Max. :33.1550 Max. :455.599 Max. :68.00310 Max. :46.89291
## INF2
## Min. : 1.44
## 1st Qu.: 21.92
## Median : 58.85
## Mean : 4315.70
## 3rd Qu.: 305.50
## Max. :207570.06
```

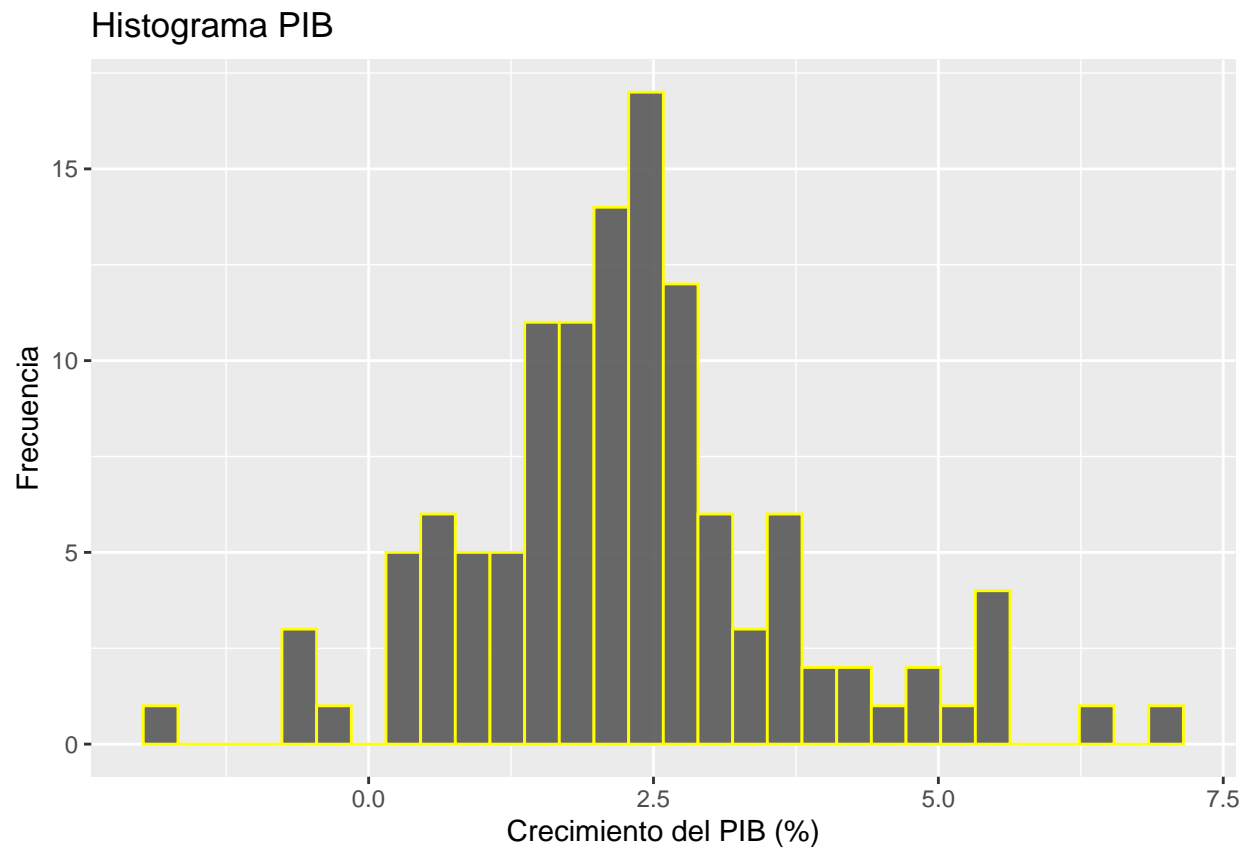
## Medallas y PIB

```
tidy_data %>%
  group_by(países,medallas) %>%
  head(15)%>%
  ggplot(aes(x=reorder(países,medallas),y=medallas, fill=países)) +
  geom_col() + coord_flip() + labs(title = "Total de medallas ganadas", x= "Países", y= "Medallas")+
  theme_solarized(light = T) + scale_colour_solarized('green')+theme(text = element_text(family = "serif",
  plot.title = element_text(face = "bold.italic", hjust = 0.5))+ theme(legend.position = "bottom")+theme
```



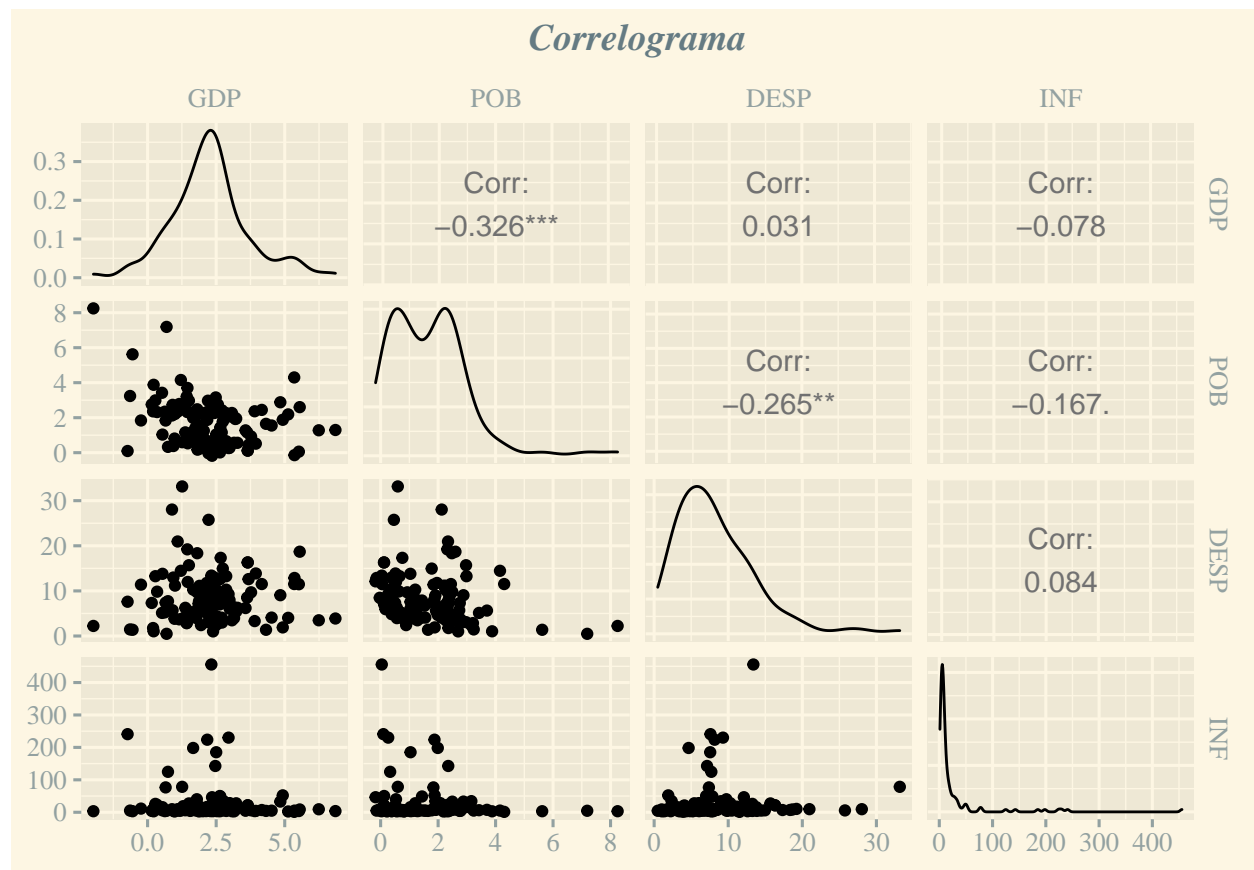
```
ggplot(tidy_data, aes(x = tidy_data$GDP)) +
  geom_histogram(position = "identity", color = "Yellow",alpha = 0.9)+
  theme(legend.position = "top") +
```

```
ggtitle("Histograma PIB") +
  ylab("Frecuencia") +
  xlab("Crecimiento del PIB (%)")
```



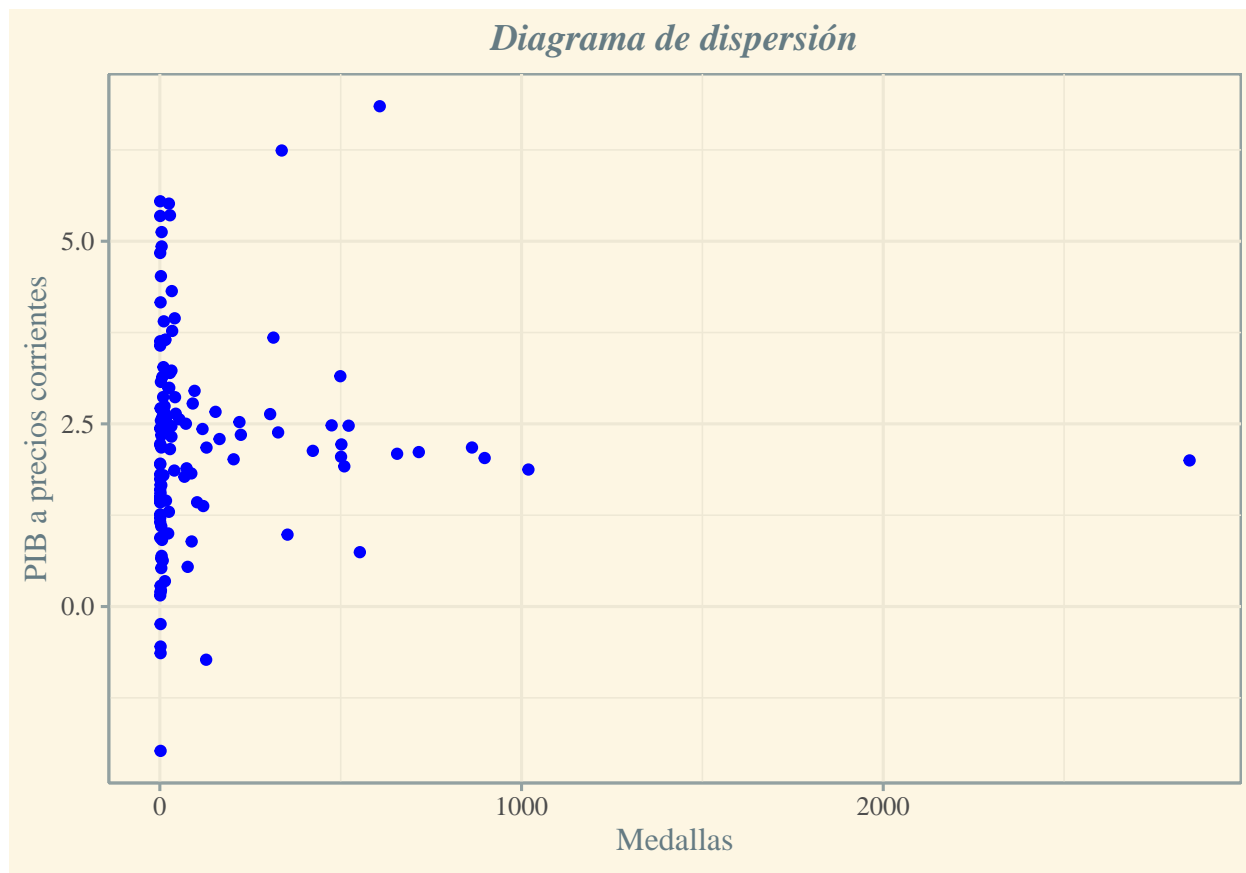
## Correlaciones

```
ggpairs(tidy_data, columns = 3:6, method = c("everything", "pearson"), title="Correlograma", color="red",
  plot.title = element_text(face = "bold.italic", hjust = 0.5)) + theme(legend.position = "bottom") + theme
```



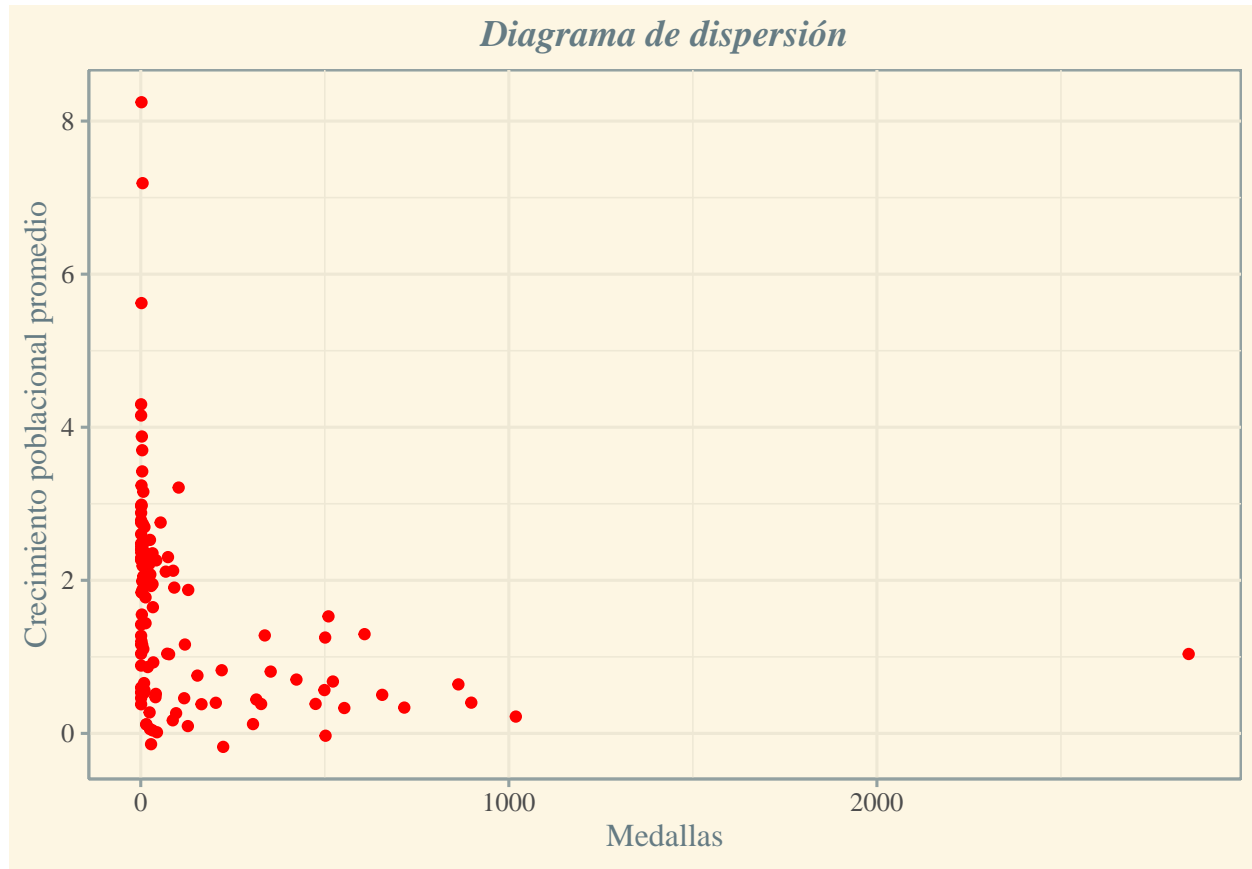
### Dispersión de los datos

```
ggplot(data = tidy_data, aes(x =medallas,y =GDP)) +
  geom_point(color="blue") +
  labs(title = "Diagrama de dispersión", x= "Medallas", y= "PIB a precios corrientes")+theme_solarized(li
  plot.title = element_text(face = "bold.italic", hjust = 0.5))+ theme(legend.position = "bottom")+theme
```

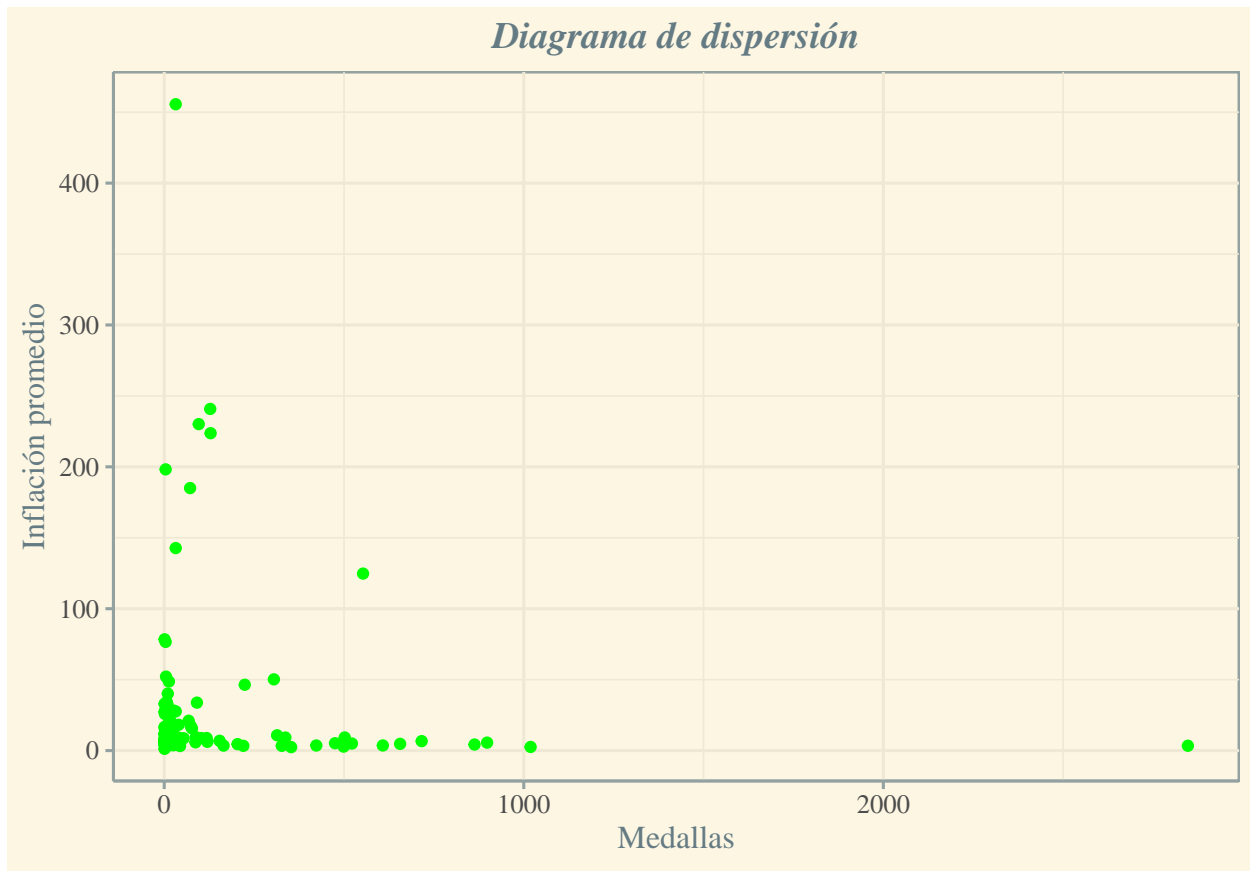


```
ggplot(data = tidy_data, aes(x =medallas,y =POB)) +
geom_point(color="red") +
labs(title = "Diagrama de dispersión", x= "Medallas", y= "Crecimiento poblacional promedio")+ theme_sol
plot.title = element_text(face = "bold.italic", hjust = 0.5))+ theme(legend.position = "bottom")+theme
```

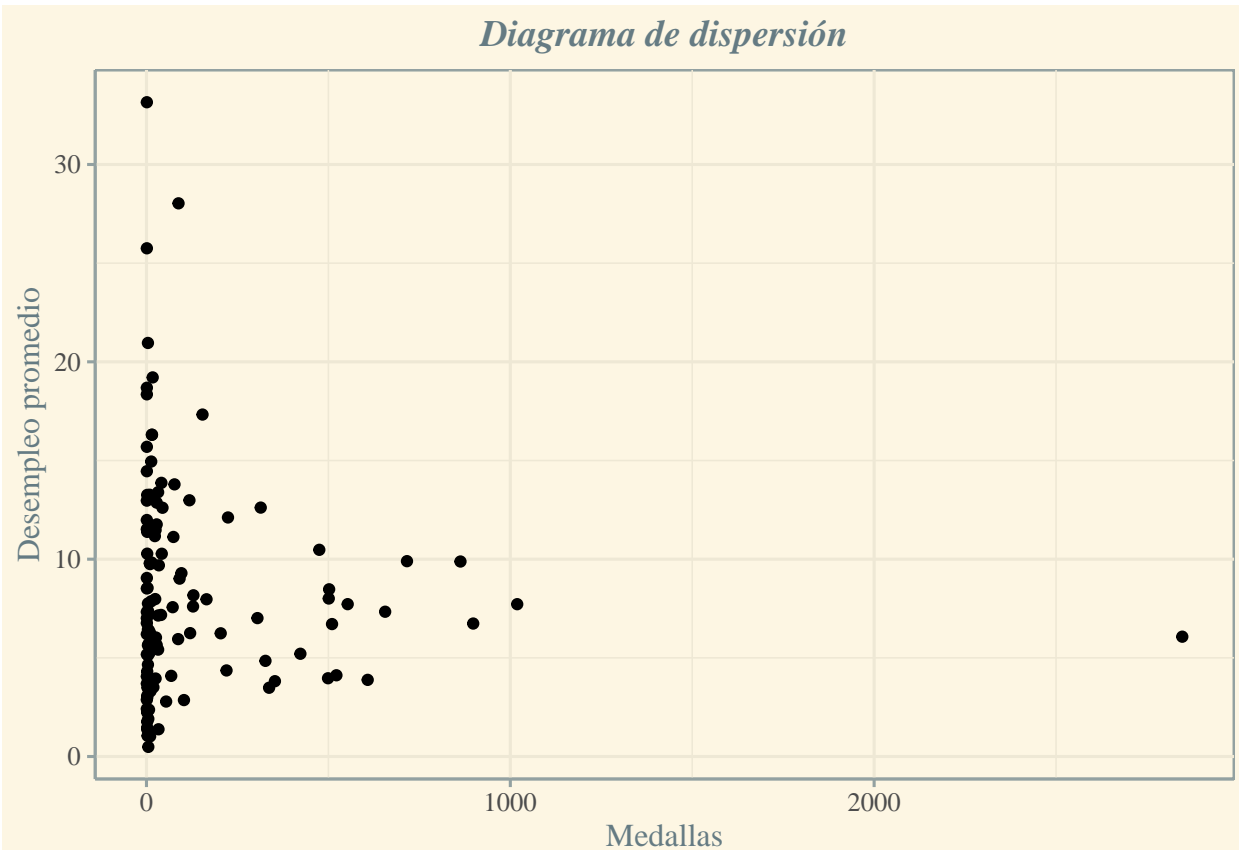




```
ggplot(data = tidy_data, aes(x =medallas,y =INF)) +
  geom_point(color="green") +
  labs(title = "Diagrama de dispersión", x= "Medallas", y= "Inflación promedio")+theme_solarized(light = '
  plot.title = element_text(face = "bold.italic", hjust = 0.5))+ theme(legend.position = "bottom")+theme
```



```
ggplot(data = tidy_data, aes(x =medallas,y =DESP)) +
  geom_point(color = "black") +
  labs(title = "Diagrama de dispersión", x= "Medallas", y= "Desempleo promedio")+theme_solarized(light = '
  plot.title = element_text(face = "bold.italic", hjust = 0.5))+ theme(legend.position = "bottom")+theme
```



Al haber analizado las estadísticas descriptas presentadas en los puntos anteriores, se logra observar sin lugar a duda que Estados Unidos es el país que mayor número de medallas (entre ellas; oro, plata y bronce) ha sumado a lo largo de la historia. Esto podría ser explicado parcialmente por el tamaño de la población, dado que es uno de los países participantes con mayor número de habitantes. Con respecto al análisis agregado junto con los demás países, se encuentra que la media de medallas ganadas por país desde 1950 hasta la actualidad corresponde a 132, sin embargo, este valor es bastante sensible ante valores atípicos, como lo indica la participación de Estados Unidos en los juegos internacionales.

En cuanto al PIB corriente expresado en miles de millones de dólares, se evidencia que la media corresponde a 7829mm de dólares norteamericanos, además, el grado de variabilidad es bastante alto debido a que el nivel mínimo de PIB corresponde a 159mm de dólares norteamericanos, y el máximo a 40313mm. Al haber observado el histograma de la variable anterior, se encuentra que la gran mayoría de los datos están concentrados entre el mínimo y el promedio, lo cual deja en evidencia que gran parte de los países participantes poseen bajos niveles de PIB. Además, la distribución de la variable de acuerdo con el histograma llega a presentar asimetría positiva, o hacia la derecha, lo cual establece que hay diversos niveles de producto al interior las naciones partícipes de los juegos olímpicos, donde pocos países poseen ingresos muy altos, y la gran mayoría posee ingresos relativamente bajos. Por otro lado, se descarta que el comportamiento de la variable se comporte normal, debido a la asimetría positiva.

De acuerdo con los diagramas de dispersión presentados en las tablas cruzadas, se logra establecer que todas las distribuciones presentan asimetría positiva, de manera que se rechaza parcialmente el hecho de que alguna variable logre presentar un comportamiento normal. Asimismo, algunos de los cruces presentan mayor cohesión entre las observaciones, por lo que la variabilidad de los datos es más baja para aquellas gráficas cuyas observaciones estén más juntas. Finalmente, al haber analizado el diagrama de dispersión de la variable dependiente e independiente, se establece que se logra gestar una relación positiva, donde mayores niveles de ingreso (PIB), podrían explicar mayor número de medallas ganadas con el paso del tiempo.

## Modelo

Con el fin de cuantificar el efecto parcial de agregados macroeconómicos en el número de medallas ganadas por los países participantes de los Juegos Olímpicos, en un periodo que comprende desde 1950 hasta 2020, se decidieron establecer 3 modelos básicos. El primero explicará el número de medallas ganadas controlando por la tasa de crecimiento del GDP, la población, la inflación y el desempleo. Teniendo en cuenta los rendimientos crecientes a escala, se establece que el aumento intensivo en algunos de los factores de control podrían llegar a generar reducciones en el número de medallas ganadas, por lo tanto, se decide incluir los mismos factores de control elevados al cuadrado como otros potenciales regresores. El segundo modelo elimina las variables independientes al cuadrado, y se encarga de medir los efectos de los agregados macroeconómicos en el número de medallas ganadas. Finalmente, y siguiendo la lógica de la teoría económica, el modelo 3 pondera únicamente los efectos de inflación y desempleo a la hora de explicar el número de medallas ganadas. Lo anterior tiene en cuenta el postulado de la Curva de Phillips, donde inflaciones potencialmente altas estarían asociadas con niveles de desempleo altos. Se espera que el efecto parcial de variables como el crecimiento del GDP y el tamaño de la población, estén asociados con incrementos en el número de medallas por país. Sin embargo, variables como el desempleo y la inflación estarían asociadas con reducciones parciales en el número de medallas ganadas por país.

```
mod1 <- lm(tidy_data$medallas~tidy_data$POB+tidy_data$POB2+tidy_data$GDP+tidy_data$GDP2+tidy_data$DESP+
mod2 <- lm(tidy_data$medallas~tidy_data$POB+tidy_data$GDP+tidy_data$DESP+tidy_data$INF)
mod3 <- lm(tidy_data$medallas~tidy_data$POB+tidy_data$DESP)
resumen <- stargazer(mod1,mod2,mod3, type="text", dep.var.labels = "Numero de medallas olímpicas obtenidas")
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               Numero de medallas olímpicas obtenidas
##                               (1)                (2)                (3)
## -----
## POB                -155.205***                -89.416***                -77.074***
##                   (47.184)                (23.302)                (21.652)
##
## POB2                11.717
##                   (7.943)
##
## GDP                -23.938                -21.890
##                   (57.153)                (21.132)
##
## GDP2                1.192
##                   (9.425)
##
## DESP                -9.415*                -9.384*                -9.344*
##                   (5.345)                (5.345)                (5.344)
##
## INF                -1.160                -0.648
##                   (1.210)                (0.485)
##
## INF2                0.001
##                   (0.003)
##
```

```
## Constant          491.418***          431.823***          343.572***
##                  (118.537)          (97.825)          (70.476)
##
## -----
## Observations          120          120          120
## R2                    0.145          0.122          0.103
## Adjusted R2           0.091          0.091          0.087
## Residual Std. Error  310.860 (df = 112)    310.874 (df = 115)    311.563 (df = 117)
## F Statistic          2.709** (df = 7; 112) 3.987*** (df = 4; 115) 6.686*** (df = 2; 117)
## =====
## Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

## Discusión de resultados

Como se esperaba, aumentos marginales en la inflación y el desempleo estaban asociados con reducciones en el número de medallas ganadas por país. Sin embargo, el crecimiento del GDP sorprendió, dado que en el modelo 1; un aumento de un punto porcentual en la tasa del crecimiento del GDP, está asociado con una reducción de 24 medallas en promedio, manteniendo lo demás constante. La variable de GDP al cuadrado presenta un coeficiente positivo, por lo que se establece que cuando la variación en el GDP es positiva y alta, un aumento de un punto porcentual estaría asociado con el incremento de 2 medallas ganadas por país en promedio, controlando por los demás factores.

El tamaño de la población discrepó de la intuición, dado que a mayor número de habitantes, menor número de medallas obtenidas por país. Sin embargo, y teniendo en cuenta el análisis de la variable población al cuadrado, aumentos muy significativos de población, están asociados con el aumento de 11 medallas en promedio por país, controlando por las demás variables. Con respecto al análisis de la inflación y el desempleo, se encuentra que aumentos marginales porcentuales en cada una de estas tasas, están asociados con reducciones promedio de 1 y 9 medallas respectivamente, manteniendo lo demás constante. La variable de inflación al cuadrado presenta un resultado insignificante en términos de magnitud, dado que cuando la inflación está desanclada o es de carácter galopante, no hay aumentos de medallas ganadas por país. El comportamiento de los regresores es el mismo en los tres modelos, al igual que la significancia estadística. En los tres modelos se afirma que la variable de población es significativa a un nivel de 1%, por lo tanto, es considerado como un regresor que se debe incluir fundamentalmente al ejecutar el análisis de regresión. Por otro lado, la tasa de desempleo llega a ser significativa al 10%, por lo que se considera como un buen regresor. Analizando la significancia conjunta, se obtiene evidencia suficiente para rechazar el hecho de que al menos uno de los coeficientes del modelo sea completamente insignificante. Finalmente, se encuentra que el modelo 1 es aquel cuya proporción de la variación explicada por los regresores es la más alta, siendo así un buen modelo en términos de bondad de ajuste.

## Pruebas

A continuación se presentan los criterios de Akaike y el Bayesiano con el fin de aislar el mejor modelo.

```
a <- AIC(mod1)
b <- AIC(mod2)
c <- AIC(mod3)

e <- BIC(mod1)
f <- BIC(mod2)
g <- BIC(mod3)

AIC <- matrix(c(a,b,c))
AIC <- round(AIC,1)
```

```

AIC <- as.data.frame(AIC)
AIC <- AIC %>%
  mutate(Modelo = c("Mod1", "Mod2", "Mod3"),
         Criterio = "AIC") %>%
  relocate(Modelo) %>%
  rename(AIC = "V1")

BIC <- matrix(c(e,f,g))
BIC <- round(BIC,1)
BIC <- as.data.frame(BIC)
BIC <- BIC %>%
  mutate(Modelo = c("Mod1", "Mod2", "Mod3"),
         Criterio = "BIC") %>%
  relocate(Modelo) %>%
  rename(BIC = "V1")

Pruebas <- bind_rows(AIC, BIC)
Pruebas[4:6,2] <- Pruebas[4:6,4]
Pruebas <- Pruebas[, -4]
Pruebas <- Pruebas %>%
  rename(Resultados = "AIC")

Pruebas <- Pruebas %>% pivot_wider(names_from = Modelo, values_from = Resultados)

Pruebas

```

```

## # A tibble: 2 x 4
##   Criterio Mod1 Mod2 Mod3
##   <chr>    <dbl> <dbl> <dbl>
## 1 AIC      1728. 1725. 1724.
## 2 BIC      1753. 1742. 1735.

```

Definimos el tercer modelo como el mejor modelo, pues es el que tiene menos pérdida de información según los criterios de AIC y BIC.

Ahora, se analizará la homosedasticidad en el modelo.

```
bptest(mod3)
```

```

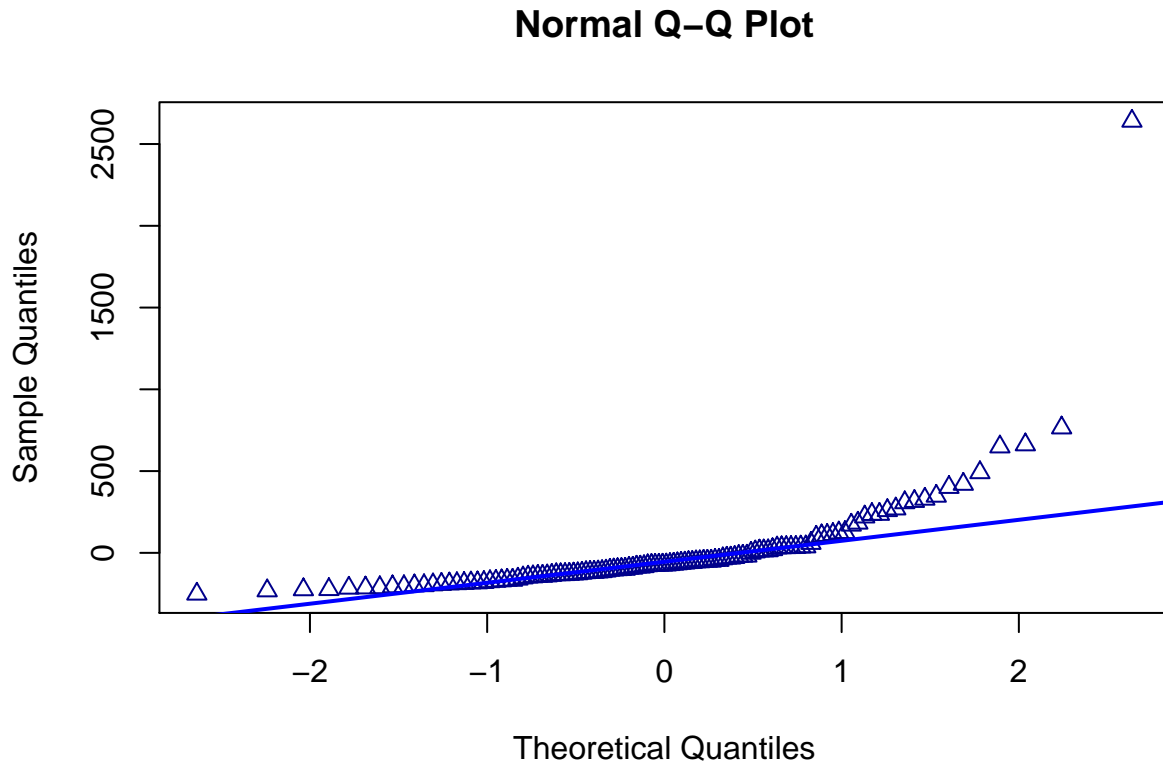
##
## studentized Breusch-Pagan test
##
## data: mod3
## BP = 1.1374, df = 2, p-value = 0.5663

```

Según el valor-p obtenido en la prueba de Breusch-Pagan, no se rechaza el hecho de los errores posean varianza constante.

Un tercer paso será analizar el supuesto de Normalidad en los errores.

```
qqnorm(mod3$residuals, pch = 24, col = "darkblue")
qqline(mod3$residuals, col = "blue", lwd = 2)
```



```
mean(mod3$residuals)
```

```
## [1] -4.558934e-14
```

Al analizar los resultados anteriores, se encuentra que la media es casi 0, y la gráfica muestra una tendencia adherente entre la línea de tendencia y la muestras por cuantil, por lo que se establece que la distribución de los errores es de carácter normal.

Finalmente, se considera pertinente evaluar la autocorrelación con los errores.

```
dwtest(mod3)
```

```
##
## Durbin-Watson test
##
## data: mod3
## DW = 0.52335, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Dado que el valor-p es inferior al nivel de significancia, se encuentra evidencia para rechazar la no autocorrelación entre los errores, por lo tanto, los errores están correlacionados entre sí.

## Conclusiones