INSTITUTO DE ECONOMÍA
FACULTAD DE CIENCIAS ECONÓMICAS
Y ADMINISTRATIVAS

# Tarea N°1

Econometría I

Valentina Andrade, Rosario Novión y Catalina Baeza

sábado 18, septiembre 2021

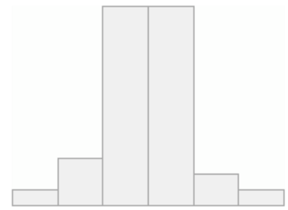
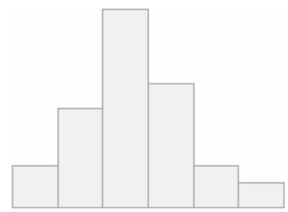
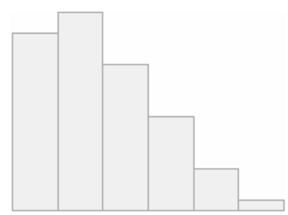
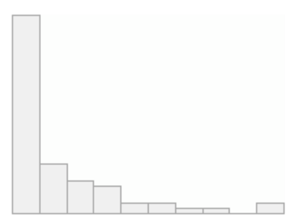
**Resumen**

El siguiente reporte tiene por objetivo presentar los análisis realizados en la Tarea N°1 del ramo Econometría I dictado por el profesor Juan Urquiza. El código del reporte, así como los análisis adicionales se pueden obtener en el siguiente link

# Análisis descriptivo

## 1. Distribuciones univariadas

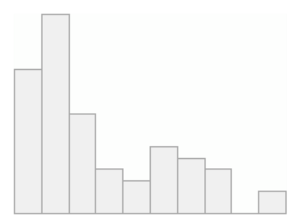
- Distribución de growth, tradeshare, yearscchol
- Describe (media, desviación estándar, mínimo y máximo)

| Variable | Estadísticas / Valores | Frec. (% sobre válidos) | Gráfico |
|---|---|---|---|
| growth<br>[numeric] | Media (d-s) : 1.9 (1.8)<br>min < mediana < max:<br>-2.8 < 2 < 7.2<br>RI (CV) : 2 (1) | 64 valores distintos |  |
| tradeshare<br>[numeric] | Media (d-s) : 0.5 (0.2)<br>min < mediana < max:<br>0.1 < 0.5 < 1.1<br>RI (CV) : 0.3 (0.4) | 64 valores distintos |  |
| yearsschool<br>[numeric] | Media (d-s) : 4 (2.6)<br>min < mediana < max:<br>0.2 < 3.6 < 10.1<br>RI (CV) : 3.5 (0.6) | 62 valores distintos |  |
| rev_coups<br>[numeric] | Media (d-s) : 0.2 (0.2)<br>min < mediana < max:<br>0 < 0.1 < 1<br>RI (CV) : 0.3 (1.3) | 23 valores distintos |  |

Cuadro 2: Matriz de correlaciones (R Pearson)

| Variables | growth | tradeshare | yearsschool | rev_coups | assasinations | rgdp60 |
|---|---|---|---|---|---|---|
| growth | 1 | | | | | |
| tradeshare | 0.211* | 1 | | | | |
| yearsschool | 0.323*** | -0.08 | 1 | | | |
| rev_coups | -0.271** | -0.206* | -0.329*** | 1 | | |
| assasinations | -0.136 | -0.348*** | -0.12 | 0.487*** | 1 | |
| rgdp60 | 0.095 | -0.127 | 0.834*** | -0.389*** | -0.095 | 1 |

*Nota:*
*** \\textit{p} < 0,01; ** \\textit{p} < 0,05; * \\textit{p} < 0,1.

| Variable | Estadísticas / Valores | Frec. (% sobre válidos) | Gráfico |
|---|---|---|---|
| assasinations<br>[numeric] | Media (d-s) : 0.3 (0.5)<br>min < mediana < max:<br>0 < 0.1 < 2.5<br>RI (CV) : 0.2 (1.8) | 26 valores distintos |  |
| rgdp60<br>[numeric] | Media (d-s) : 3.1 (2.5)<br>min < mediana < max:<br>0.4 < 2 < 9.9<br>RI (CV) : 4 (0.8) | 64 valores distintos |  |

## 2. Distribuciones bivariadas

### Matriz de correlaciones de todas las variables

- y luego discutir significancia estadística

## 3. Gráficos de dispersión

Como se ve en la Figura…

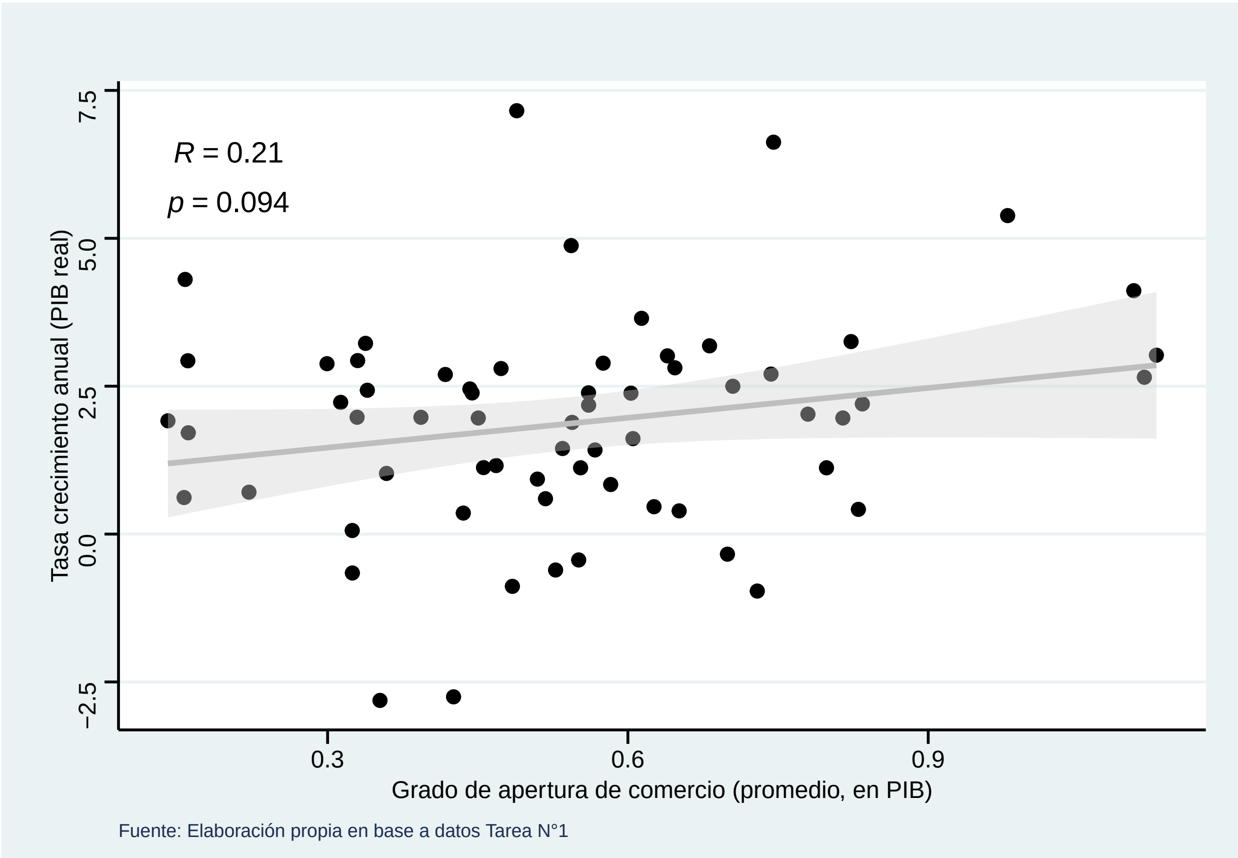



Figura 1: Gráfico de dispersión entre la tasa de crecimiento económico y el grado de apertura de comercio.

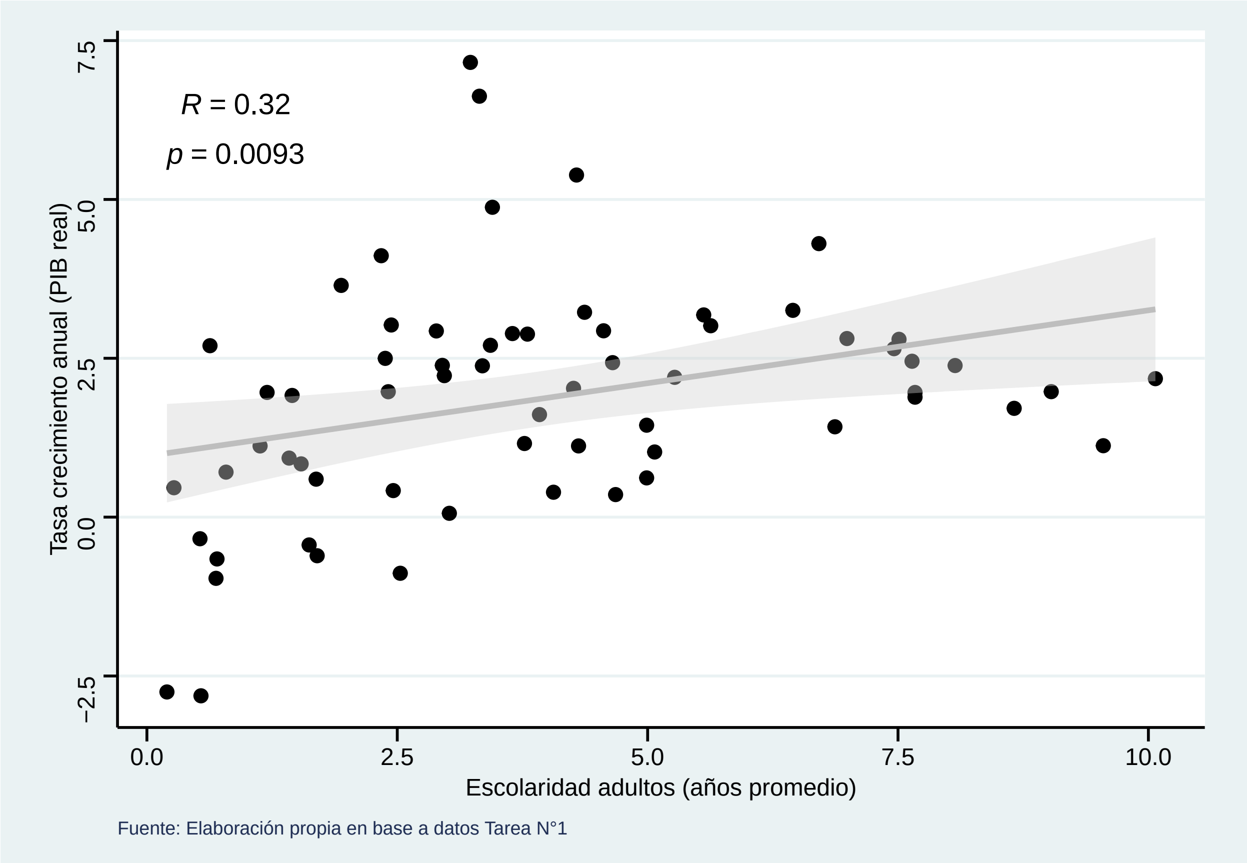



Figura 2: Gráfico de dispersión entre la tasa de crecimiento económico y los años de educación de adultos

# Análisis de regresión

## 4. Análisis de regresión lineal múltiple

Cuadro 3: Modelo de regresión lineal que estima la tasa de crecimiento anual (en PIB)

| | Modelo 1 |
|---|---|
| Intercepto | 0.627 |
| | (0.783) |
| Determinante económico | |
| Grado de apertura comercio (promedio, en PIB) | 1.341 |
| | (0.960) |
| Determinante social | |
| Escolaridad adultos (años promedio) | $0.564^{***}$ |
| | (0.143) |
| Determinantes políticos | |
| Hitos disruptivos (promedio anual) | $-2.150$ |
| | (1.119) |
| Asesinatos políticos (promedio anual) | 0.323 |
| | (0.488) |
| Control | |
| PIB per cápita (base 1960) | $-0.461^{**}$ |
| | (0.151) |
| $R^2$ | 0.291 |
| Adj. $R^2$ | 0.230 |
| Num. obs. | 64 |
| F statistic | 4.764 |
| RMSE | 1.594 |

$^{***}p < 0{,}001$; $^{**}p < 0{,}01$; $^{*}p < 0{,}05$. Coeficientes de regresión no estandarizados y error estándar entre paréntesis

En términos formales tenemos las dos siguientes rectas de regresión, donde *(1)* indica la función de regresión poblacional que se busca estimar y *(2) con continuación en (3)* la función de regresión lineal estimada con Método de Mínimos Cuadrados Ordinarios.

**Función de Regresión Lineal Poblacional**

$$\widehat{\text{growth}} = \beta_0 + (\beta_1 \cdot X_{\text{tradeshare}}) + (\beta_2 \cdot X_{\text{yearsschol}}) + (\beta_3 \cdot X_{\text{revcoups}}) + (\beta_4 \cdot X_{\text{assasinations}}) + (\beta_5 \cdot X_{\text{rgdp60}}) + u \tag{1}$$

**Función de Regresión Lineal Muestral**

$$\widehat{\text{growth}} = 0.627 + (1.341 \cdot X_{\text{tradeshare}}) + (0.564 \cdot X_{\text{yearsschol}}) + (-2.15 \cdot X_{\text{revcoups}}) + \quad (2)$$
$$(0.323 \cdot X_{\text{assasinations}}) + (-0.461 \cdot X_{\text{rgdp60}}) + u \quad (3)$$

Una forma gráfica de verlo en la *Figura*

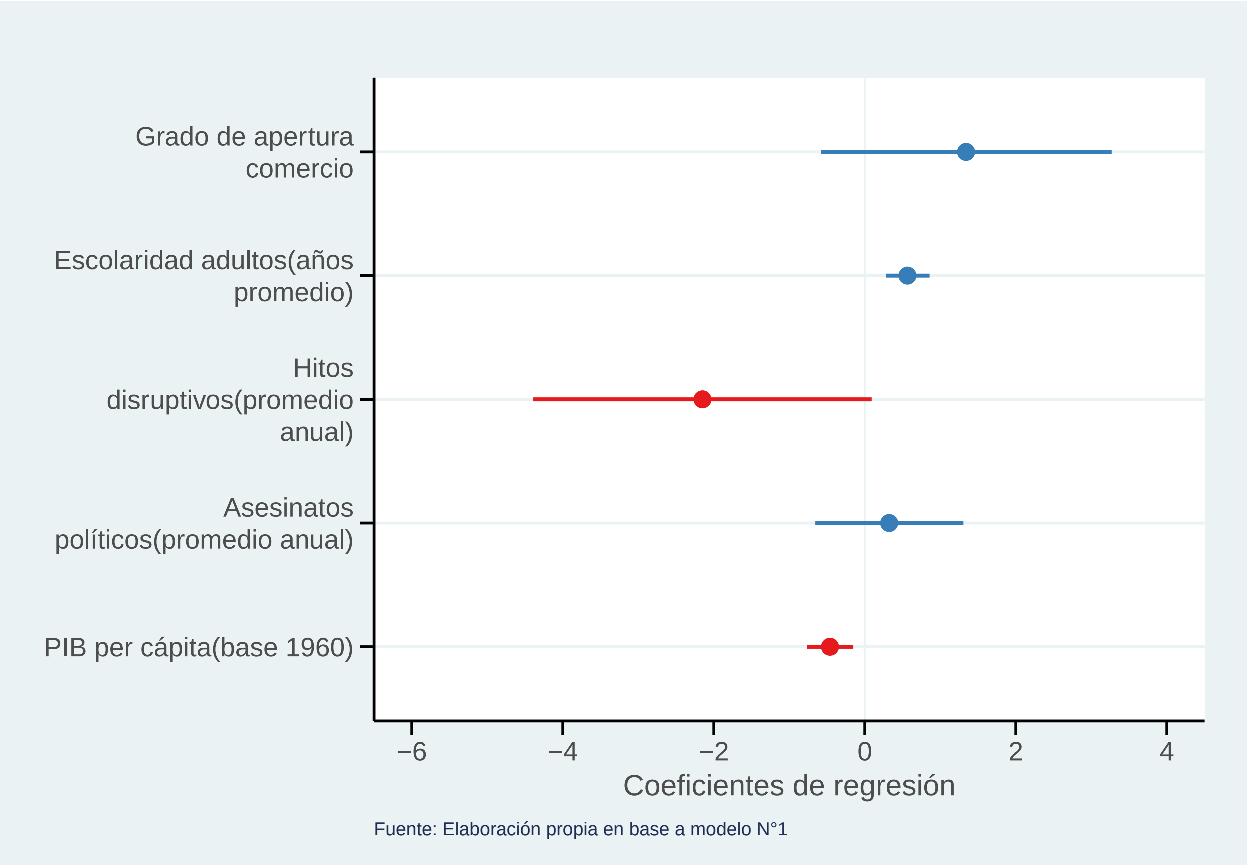



Figura 3: Forest-plot con los coeficientes estimados del modelo de regresión que estima la tasa de crecimiento anual (PIB real)

## 5. Propiedades del Modelo de Regresión

### 5.1 Suma de residuos

Uno de los supuestos importantes de *OLS*[1] es que el valor esperado de los residuos es cero ($E(\hat{u}) = 0$), o en términos matriciales que la suma de los residuos $\hat{u}_i$ es igual a cero[2]

[1] *Ordinary Least Squares*, o en español, Mínimo Cuadrado Ordinarios (*MCO*)

[2] Para ser exactas el valor es $-1.8041124 \times 10^{-15}$, valor muy pequeño que tiende a cero.

Cuadro 4: Media condicional de variables independientes

| Variables independientes (X) | E(u\|X) |
|---|---|
| assasinations | 0 |
| rev_coups | 0 |
| rgdp60 | 0 |
| tradeshare | 0 |
| yearsschool | 0 |

*Nota:*

Los intervalos de confianza fueron calculados a un 95% de confianza

$$E(u) = \sum_{i=0}^{n} \hat{u}_i = 0$$

### 5.2 Ortogonales a variables explicativas

Sea $\hat{u}$ un vector *nx1* que denota los residuos y $X_{nxm}$ matriz de predictores del modelo que contiene a $X_1, X_2, ..., X_k$ parámetros, diremos que la suma de residuos cuadrado (*SSR*) $\hat{u}'\hat{u}$ son ortogonales. Pero además, diremos que la matriz $\hat{u}$ es ortogonal a las variables explicativas, lo que significa que hay una relación de independencia de los valores esperados de $\hat{u}$ y $X$ (supuesto *MLR4* del Teorema de Gauss Markow para la estimación *OLS*)

$$\hat{u}'X = E(\hat{u}|X) = 0$$

Como modo de complementar este análisis, una forma convencional es la visualización de los residuos, dado ciertos parámetros de los modelos.[3] Como se puede ver en las figuras del Apéndice ...(aquí describir)

### 5.4 Promedio de la variable dependiente coincide con promedio de variable dependiente estimada

Por ley de esperanzas iteradas decimos que $E(y_i) = E[E(y_i|X)]$, donde $E(y_i|X)$ representa el valor esperado de $y$ que ha sido estimado a partir de un cambio en $X$, manteniendo lo demas constante (en otras palabras $\hat{y}$ estimado).

Luego, calculamos la media de $\hat{y}$ en los datos del modelo, obteniendo 1.86912. Mientras que, tal como reportamos en el *Cuadro 1* la media de la tasa de crecimiento en la muestra es 1.86912

[3] Recomendación de visualización de la relación entre residuos y valores predichos en STATA

### 5.5 Comprobar coeficiente determinación $R^2$ es igual al cuadrado del coeficiente de correlacion entre variable dependiente y estimada.

Según Wooldrige (2019) $R^2$ se puede mostrar como el cuadrado del coeficiente de correlación entre las $y_i$ reales y los valores ajustados de $\hat{y}_i$ ($\rho_{y_i,\hat{y}}$). Formalmente

$$R^2 = \frac{(\sum_{i=1}^{n}(y_i - \hat{y})(\hat{y}_i - \bar{\hat{y}}))^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} = (\rho_{y_i,\hat{y}})^2 \tag{4}$$

Primero, se debe notar que se ha utilizado la notación del promedio de las $\hat{y}_i$ para ser fiel a la fórmula del coeficiente de correlación, pero sabemos que el promedio es igual a las $\bar{y}$ debido a que el promedio muestral de los residuos es cero y $y_i = \hat{y}_i + \hat{u}_i$.

Luego,

$$\sum_{i=1}^{n}(y_i - \hat{y})(\hat{y}_i - \bar{\hat{y}}) = cov(y_i, \hat{y}) = 0.9602752 \tag{5}$$

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sigma_{y_i} = 1.8161889 \tag{6}$$

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 = \sigma_{y_i} = 0.9799363 \tag{7}$$

Si reemplazamos

$$R^2 = \frac{(0.9602752))^2}{1.8161889 \cdot 0.9799363} = (0.5181226)^2 = 0.2911211$$

Como podemos ver, el resultado indicado se corresponde con el $R^2$ reportado en el *Cuadro 2*, esto es, que el $R^2$ es 0.2911

## 6. Efectos marginales

Vuelva a graficar las relaciones de la pregunta 3, pero ahora incluya en sus gráficos las predicciones del modelo 1 ¿diría que son significativas?

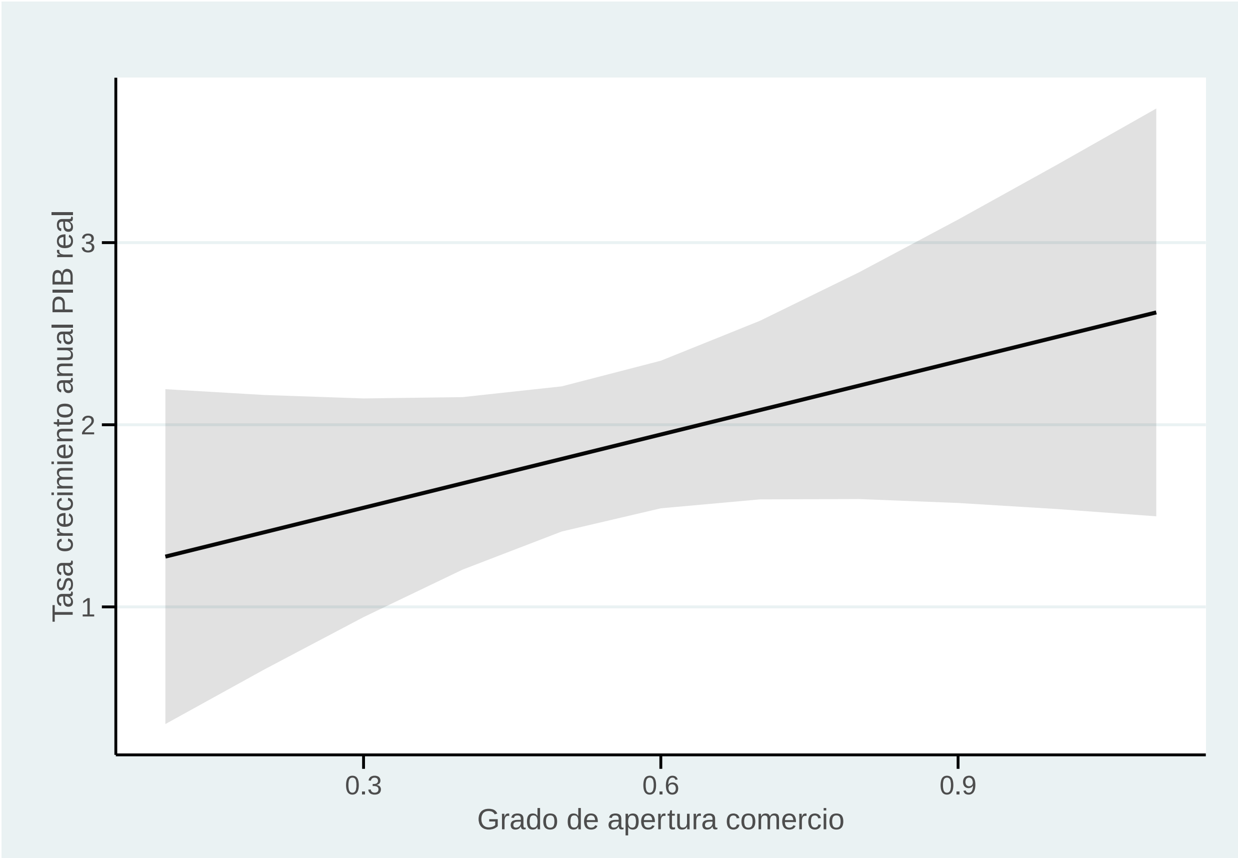



Figura 4: Valores predichos para la tasa de crecimiento anual (en PIB) según grado de apertura del comercio (PIB promedio)

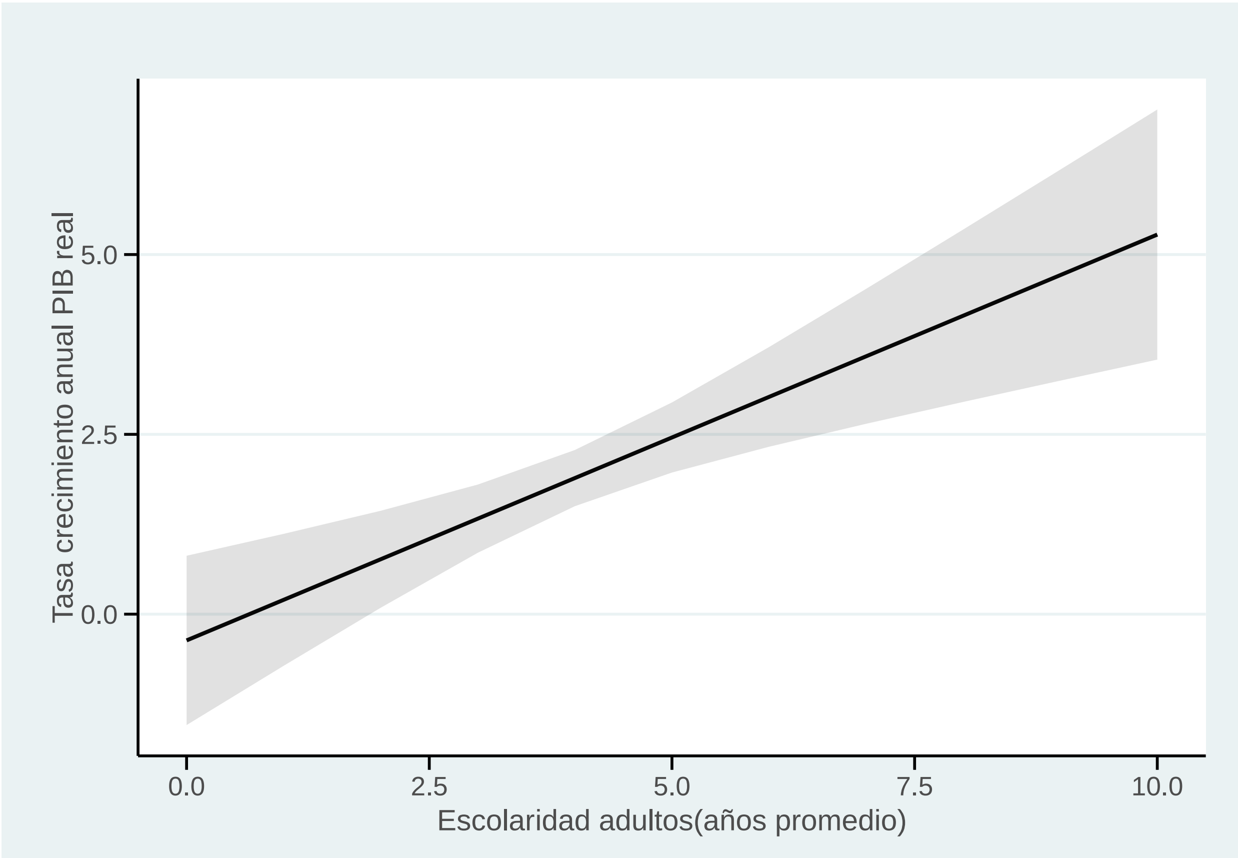



Figura 5: Valores predichos para la tasa de crecimiento anual (en PIB) segúnaños de escolaridad promedio de la población adulta)

## 7. Significancia global del modelo

Sabemos que el estadístico F (aquí describir), se define como

$$F_{(F,k,gl)} = \frac{(SSC_r - SSC_{nr})/q}{SSC_{nr}/(n-k-1)} \tag{8}$$

Donde $SSR$ indica la suma de residuos cuadrados, $n$ el número de observaciones del modelo, $k$ el número de parámetros ($gl = n - k - 1$) y $q$ el número de restricciones. Los súbíndices $r$ y $nr$ indican el modelo con restricciones $q$ en parámetros y sin restricciones, respectivamente.

En el *Cuadro 2* podemos ver el resultado del estadístico $F$ en términos globales, es decir, se compara el *modelo 1* con un modelo sin predictores (*modelo nulo*). El valor reportado es *F(4.7638665, 5, 58)*, donde la primera coordenada indica el valor observado de F (*F, k, gl*). A partir de estos valores se puede obtener su *valor-p* que es *0.001028* el que nos indica que (...)

A su vez, sabemos que si el estadístico $F$ es calculado a partir de la suma de residuos cuadrados, y el coeficiente de determinación $R\hat{}2$ lo podemos calcular como $R^2 = 1 - \frac{SSR}{SST}$, donde $SST$ es la suma total de los residuos, podemos decir que

$$F_{(q,k,gl)} = \frac{(R^2_{nr} - R^2_r)/q}{1 - R^2_{nr}/(n-k-1)} \tag{9}$$

Recordemos que dado que el $R^2_r$ no tiene predictores, este es cero

$$F_{(q,k,gl)} = \frac{(R^2_{nr}/k}{1 - R^2_{nr}/(n-k-1)} \tag{10}$$

$$F_{(q,k,gl)} = \frac{(0.2911)/5}{1 - 0.2911/(64-5-1)} = 4.763$$

Se puede concluir de $R^2$ que (...)

## 8. Predicción

Se calcularán dos predicciones de la tasa de crecimiento ($\widehat{growth}$) a partir de los resultados obtenidos en el *modelo 1*. Luego de ello, se discutirá si estas predicciones son significativamente distintas

### 8.1 Predicción en base a la media de los predictores

La primera predicción considera los valores promedio ($\bar{X}$) de cada uno de los parámetros incorporados en el modelo. Formalmente tenemos que

$$\widehat{\text{growth}} = 0.627 + (1.341 \cdot \bar{X}_{tradeshare}) + (0.564 \cdot \bar{X}_{yearsschol}) + (-2.15 \cdot \bar{X}_{revcoups}) \tag{11}$$
$$+(0.323 \cdot \bar{X}_{assasination}) + (-0.461 \cdot \bar{X}_{rgdp60}) + u \tag{12}$$

Donde valores esperados fueron tomados del *Cuadro 1*

$$\begin{aligned} \bar{X}_{tradeshare} &= 0.542 \\ \bar{X}_{yearsschol} &= 3.959 \\ \bar{X}_{revcoups} &= 0.17 \\ \bar{X}_{assasination} &= 0.282 \\ \bar{X}_{rgdp60} &= 3.131 \end{aligned}$$

Con ello obtenemos

$$\widehat{\text{growth}}_1 = 0.627 + 1.341 \cdot 0.542 + 0.564 \cdot 3.959 + -2.15 \cdot 0.17 + 0.323 \cdot 0.282 + -0.461 \cdot 3.131 = 1.869$$

### 8.2 Predicción en base a la media de predictores y dos desviaciones estándar de *tradeshare*

Segundo, se realizará una predicción del crecimiento ($\widehat{growth_2}$) con los valores promedios para todas las variables, excepto para *tradeshare* que toma un valor igual a dos desviaciones estándar por encima de la media. Esto significa que $\bar{X}_{tradeshare} + 2\sigma_{tradeshare}$

$$\widehat{\text{growth}}_2 = 0.627 + 1.341 \cdot 0.999 + 0.564 \cdot 3.959 + -2.15 \cdot 0.17 + 0.323 \cdot 0.282 + -0.461 \cdot 3.131 = 2.4818414$$

### 8.3 Comparación de predicciones

Ahora veremos si estas estimaciones son significativamente distintas. Primero, podemos evaluar esto con el siguiente cuadro resumen en donde claramente se evidencia que los intervalos de confianza se *"solapan"*

En resumen entre ambas predicciones tenemos la siguiente tabla

Cuadro 5: Predicciones lineales ($xb$) del MRL de tasa de crecimiento

| Predicción | $\hat{y}$ | *se* | t | $P > \|t\|$ | lim. inferior | lim. superior |
|---|---|---|---|---|---|---|
| $\hat{growth}_1$ | 1.86912 | .199211 | 9.38 | 0.000 | 1.470356 | 2.267884 |
| $\hat{growth}_2$ | 2.481425 | .4815638 | 5.15 | 0.000 | 1.517471 | 3.44538 |

*Nota*: *se* (error estándar) calculado con método delta y los límites del intervalo de predicción con un 95% de confianza.

Para estar más seguras de esta afirmación, testeamos la $H_0$ que indica que no hay diferencias significativas entre la predicción con los valores promedio y aquella que considera para *tradeshare* un valor sobre dos desviaciones estándar de su media. Como podemos ver en el resultado del análisis de comparación entre esas dos combinaciones lineales,[4] no hay evidencia suficiente para rechazar $H_0$, por lo que **no** se puede plantear de manera certera que a nivel poblacional ambas predicciones sean estadísticamente distintas ($p > 0.05$)

Cuadro 6: Test de combinación lineal (*mlincom*)

| lincom | pvalue | ll | ul |
|---|---|---|---|
| -0.612 | 0.168 | -1.490 | 0.265 |

## 9. Estimación del cambio en el crecimiento en base a cambio educativo

Estime el cambio en la tasa de crecimiento que predice el MODELO 1 para un país que en 1960 implementó una política educativa que permitió aumentar la escolaridad promedio de 4 a 6 años, y luego refiérase a la significancia estadística y económica.

[4] Multiple Linear Combinations with *margins* in STATA

Cuadro 7: Valores predichos para la tasa de crecimiento (en PIB) según años de escolaridad promedio 4 y 6 años

| Tasa crecimiento | Grado apertura | Hitos disruptivos | Asesinatos políticos | PIB 1960 | Años de escolaridad promedio | Intervalo superior | Intervalo inferior |
|---|---|---|---|---|---|---|---|
| 1.892130 | 0.5423919 | 0.1700666 | 0.281901 | 3.130813 | 4 | 2.291066 | 1.493195 |
| 3.020619 | 0.5423919 | 0.1700666 | 0.281901 | 3.130813 | 6 | 3.728292 | 2.312946 |

*Nota:*
Los intervalos de confianza fueron calculados a un 95% de confianza

También lo podemos ver gráficamente (...)

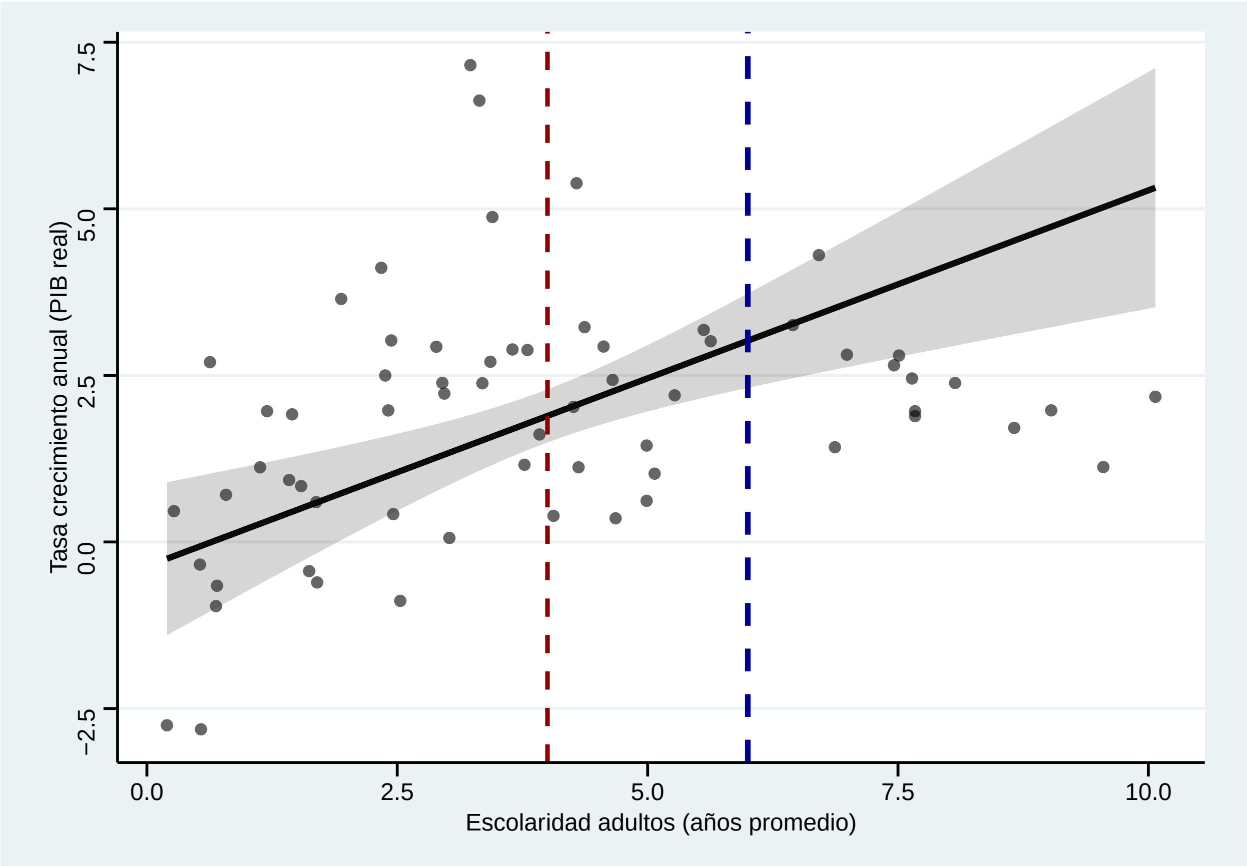



Figura 6: Gráfico de valores predichos para la tasa de crecimiento según los años de escolaridad promedio. *Nota*: La línea roja marca los años de escolaridad en 4 y la azul en 6

$$\triangle \widehat{growth} = \triangle \bar{X}_{yearsschool} \cdot E(growth|yearsschool = 6)$$

- Importante: no olvidar hablar de control estadístico

## 10. Significancia conjunta

Para evaluar la significancia conjunta de determinados predictores, se compara la diferencia entre el modelo original (modelo sin restricciones) y el modelo sin las variables de interés (modelo restringido). Sea $\hat{\beta}_{tradeshare}$, $\hat{\beta}_{revcoups}$, $\hat{\beta}_{assasinations}$ los coeficientes de las variables *tradeshare*, *rev_coups* y *assasinations*, diremos que

$$H_0 : \hat{\beta}_{tradeshare} \vee \hat{\beta}_{revcoups} \vee \hat{\beta}_{assasinations} = 0 H_1 : \hat{\beta}_{tradeshare} \vee \hat{\beta}_{revcoups} \vee \hat{\beta}_{assasinations} \neq 0$$

El estadístico que nos permitió hacer ese contraste es el *estadístico F*, cuyos resultados indican que *F(3,58) = 2.38* (*p = 0.0787*). Ahora bien, como habíamos definido antes este también se puede calcular como

$$F_{(q,k,gl)} = \frac{(SSC_r - SSC_{nr})/q}{SSC_{nr}/(n-k-1)}$$

Donde, cada parámetro indica

- $SSC$, suma de residuales cuadrados
- Subíndices *nr* y *r* que indican estadísticos sin restricciones y con restricciones, respectivamente
- $q$, número de restricciones que son **3**
- $k$, número de parámetros del modelo sin restringir que son **5**
- $n$, número de observaciones del modelo, *64*
- *gl*, grados de libertad que se calculan como *n-k-1* e indica el denominador del estadístico.

Luego, volviendo a estimar un segundo modelo restringido (*modelo2*) y obtenemos que $SSC_r =$ 165.462195. En consecuencia como podemos ver obtenemos el mismo valor que en el *software*

$$\hat{F}_{(3,5,58)} = \frac{(165.462195 - 147.310823)/(3)}{147.310823/(58)} = 2.38$$

Cuando comparamos este $\hat{F}_(3, 5, 58)$ observado con el valor crítico o teórico a un 95% de confianza podemos ver que el valor va entre 2.76 ($F_{(3,60)}$) y 2.79 ($F_{(3,50)}$). Luego,

$$|\hat{F}_{(3,5,58)}| < \hat{F}_{(3,60)} \Longrightarrow \hat{\beta}_{tradeshare} \vee \hat{\beta}_{revcoups} \vee \hat{\beta}_{assasinations} \neq 0$$

Lo que indica que no se puede rechazar la hipótesis nula que el modelo restringido es estadísticamente distinto al modelo restringido. A partir de la fórmula que considera la suma de residuos al cuadrado podemos recoger que la suma de residuos entre el modelo con esos predictores y sin esos predictores no es lo suficientemente grande como para producir una diferencia significativa entre los modelos. En otras palabras, la incorporación de estas variables al modelo original (o sin restringir) no produce un mejor ajuste entre los valores observados y predichos.

En segunda instancia, es importante recordar que este tipo de pruebas estadísticas refieren a un *conjunto* de predictores. Por ello, no estimamos *qué predictor **específicamente** no aporta de manera sustantiva al modelo. A su vez, este tipo de pruebas tienen un significado estadístico relevante cuando existe una alta correlación entre variables, algo que no se cumple entre las variables incorporadas en la restricción pero* sí* en aquellas que fueron conservadas (recordemos lo que aparece en la matriz de correlaciones donde años de educación y PIB en 1960 correlacionan alta y significativamente en *r = 0.83*).

# Referencias

# Apéndice

## Apéndice A

En este apartado se encuentra incorporado el código de análisis en STATA y los links de archivos para la construcción del reporte en *LateX*

### Código en STATA (*.dofile*)

En el siguiente apartado se puede encontrar el archivo dofile con el cuál fue generado este reporte. También se puede acceder en el siguiente link a un repositorio en GitHub que contiene el .dofile y .tex para construir el reporte el LateX.

**Otro recursos del informe**

## Apéndice B

### Valor esperado de residuos condicional a predictores

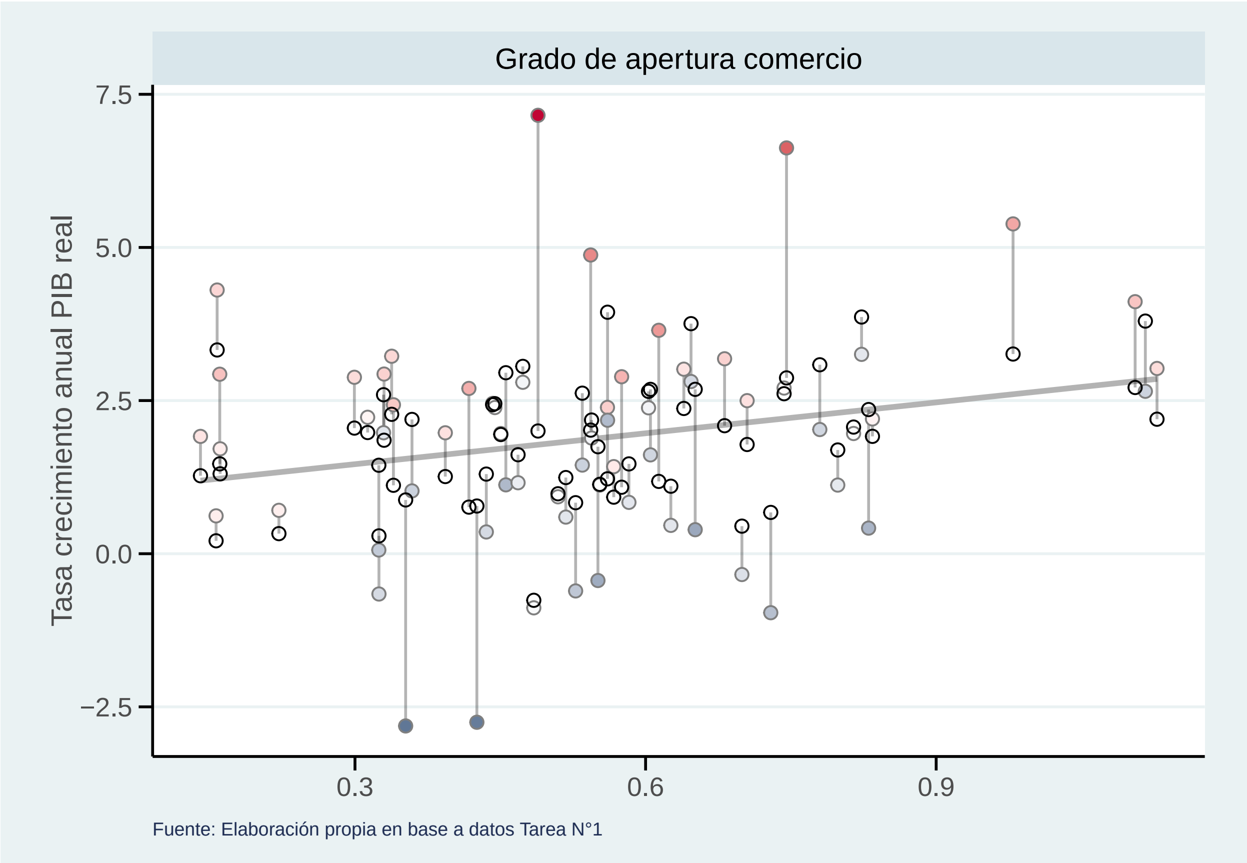



Figura 7: Distribución de residuos para cada predictor del modelo

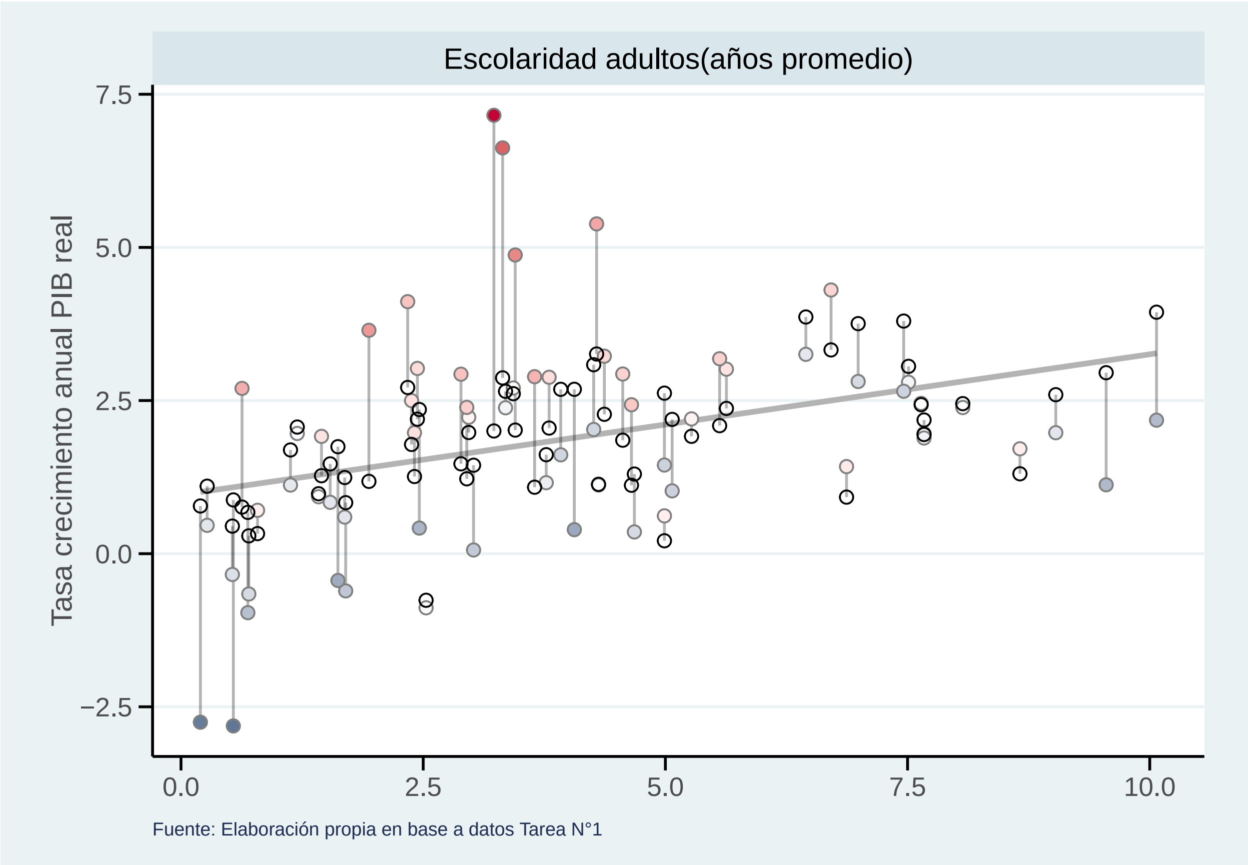



Figura 8: Distribución de residuos para cada predictor del modelo

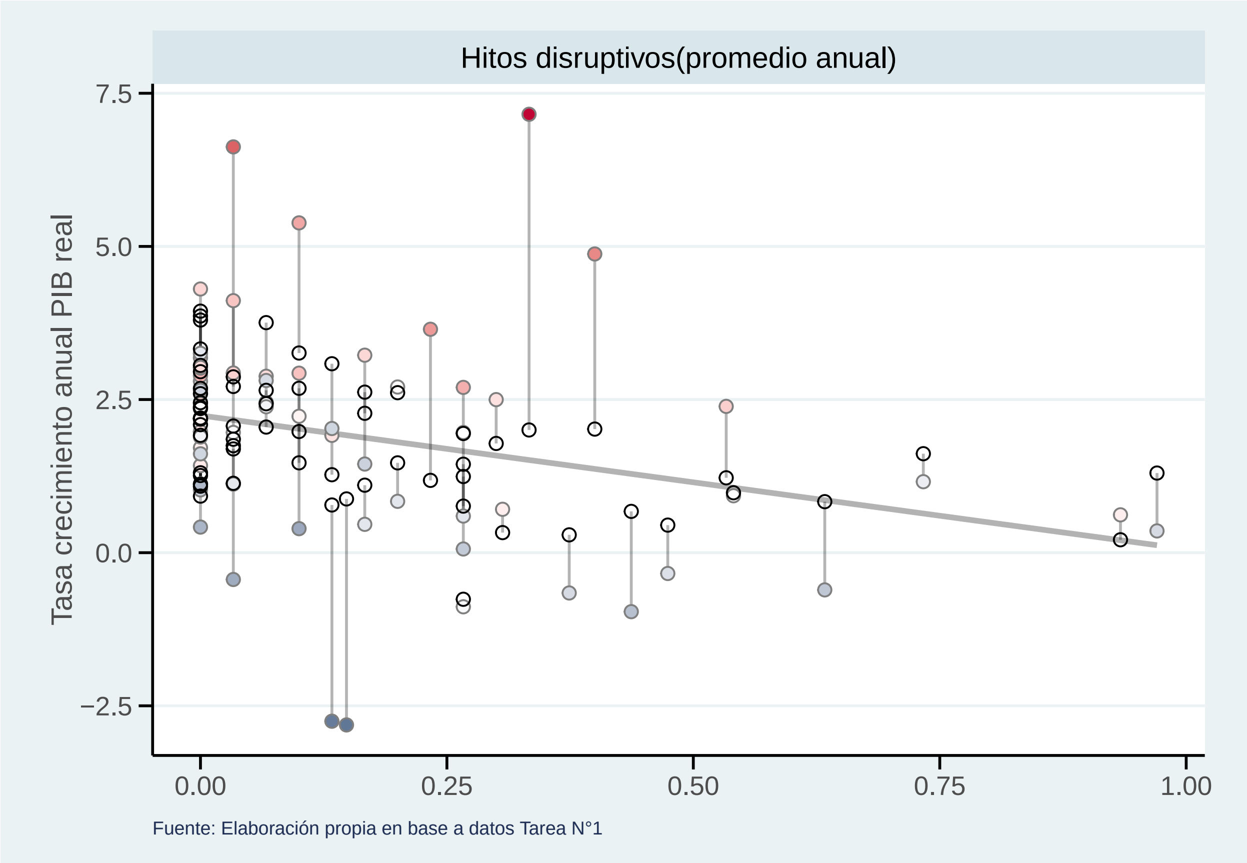



Figura 9: Distribución de residuos para cada predictor del modelo

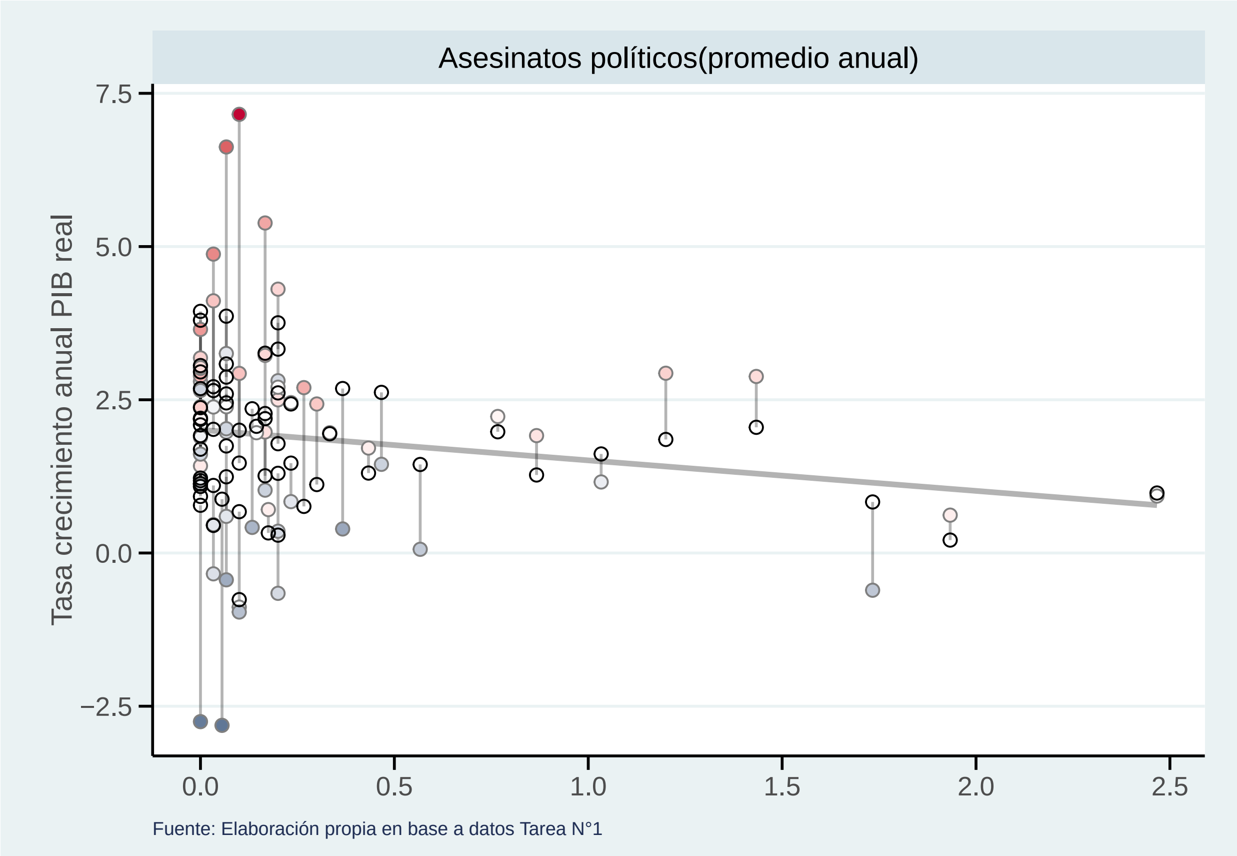



Figura 10: Distribución de residuos para cada predictor del modelo

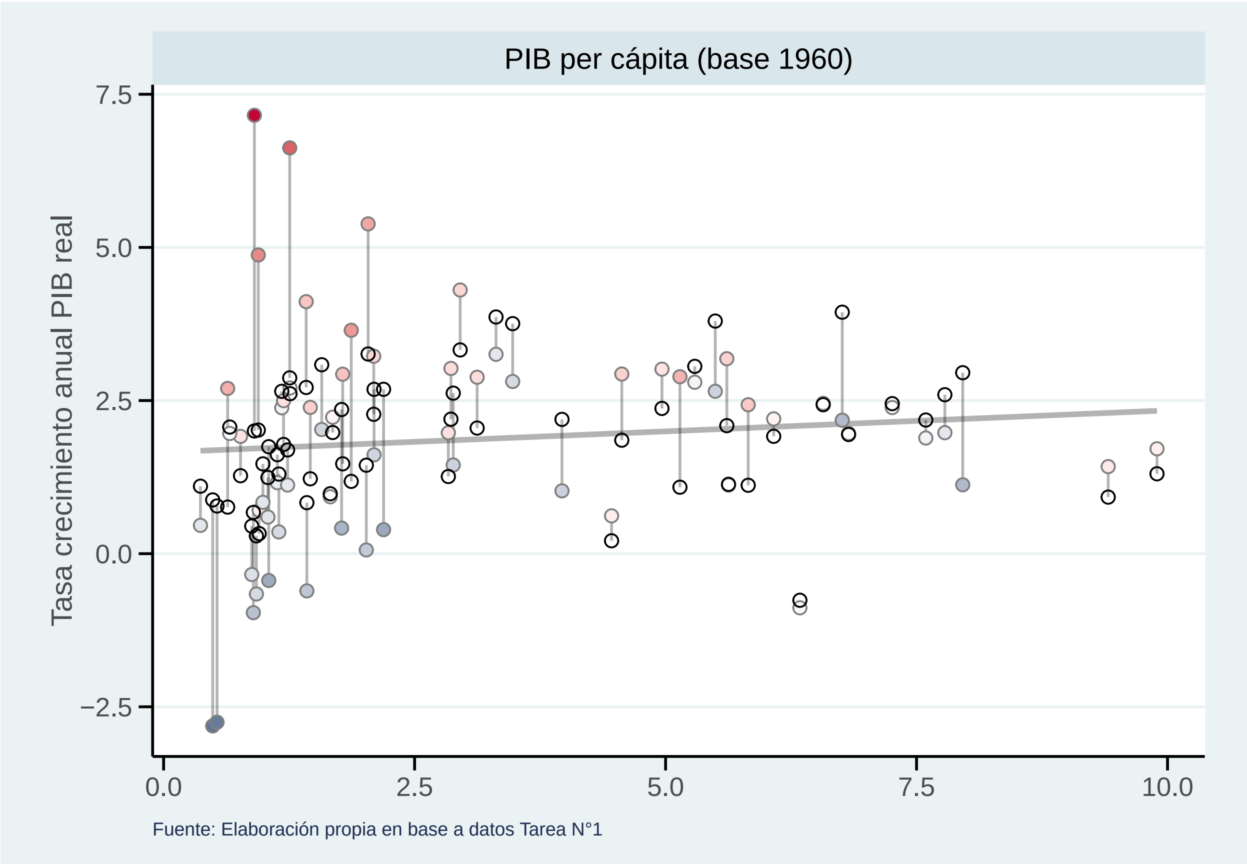
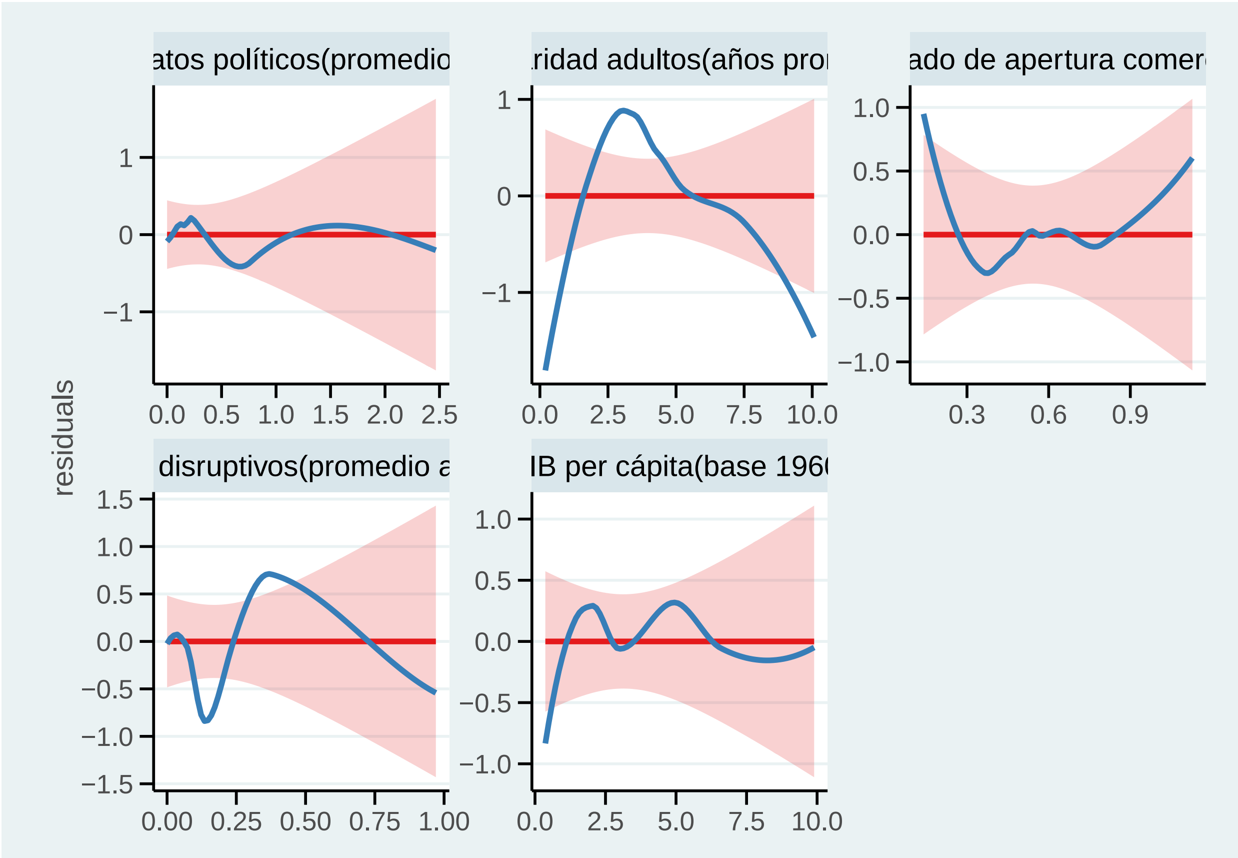
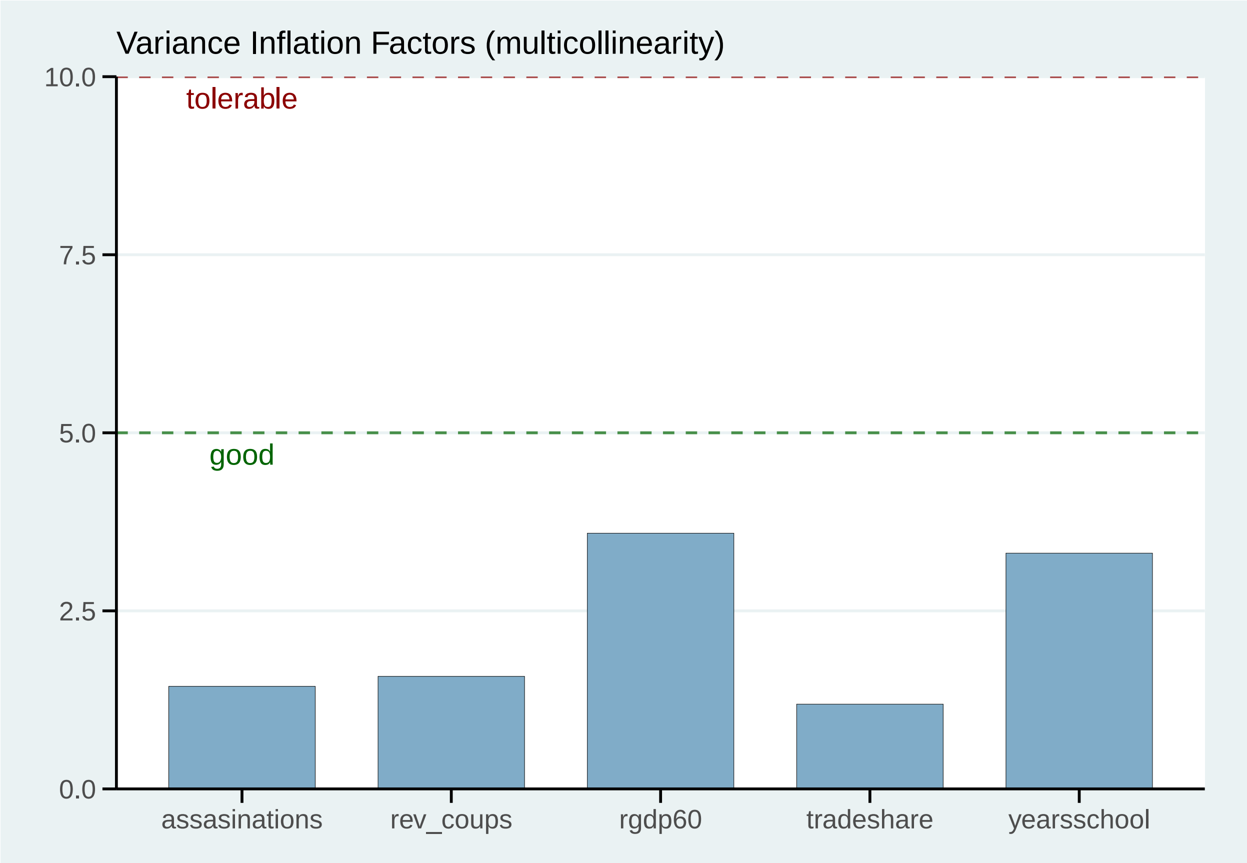



Figura 11: Distribución de residuos para cada predictor del modelo

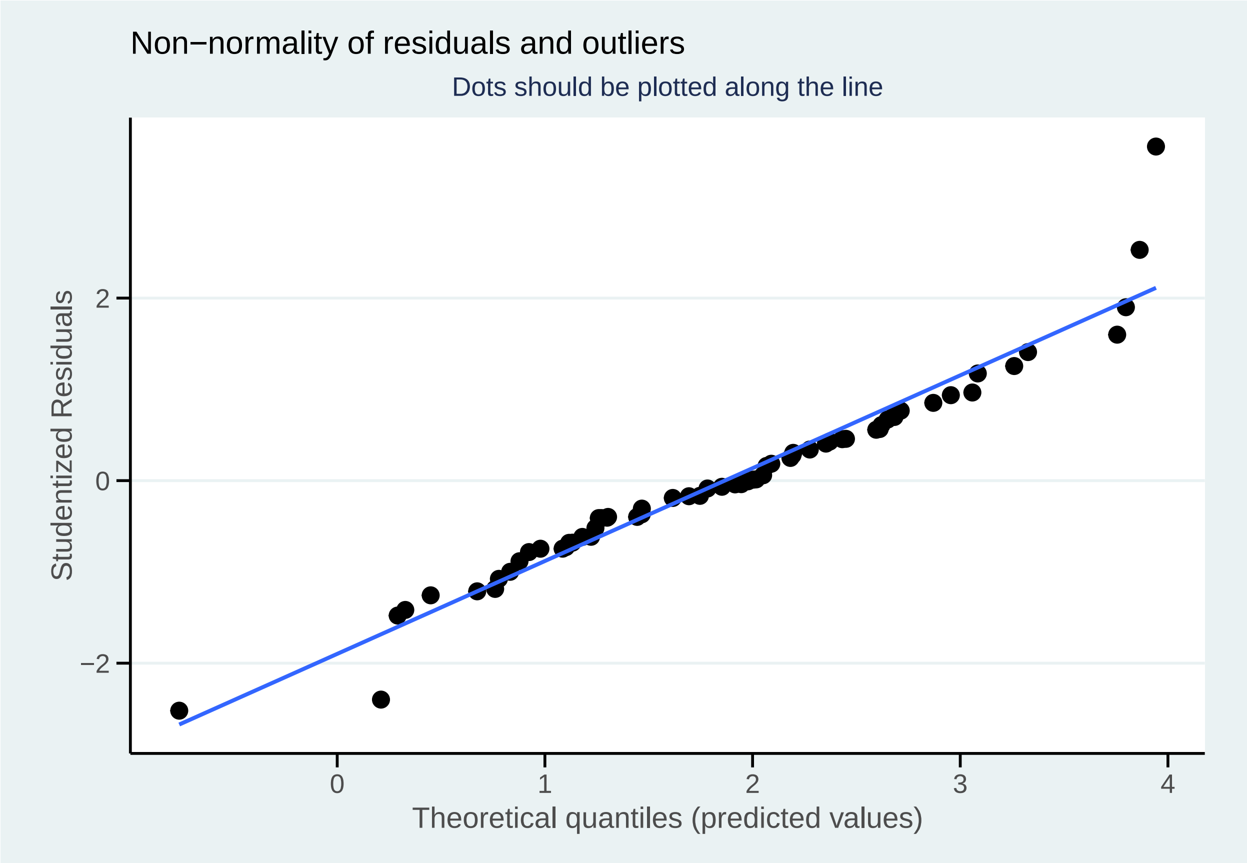
**Visualización de propiedades de Gauss Markow y MCL**



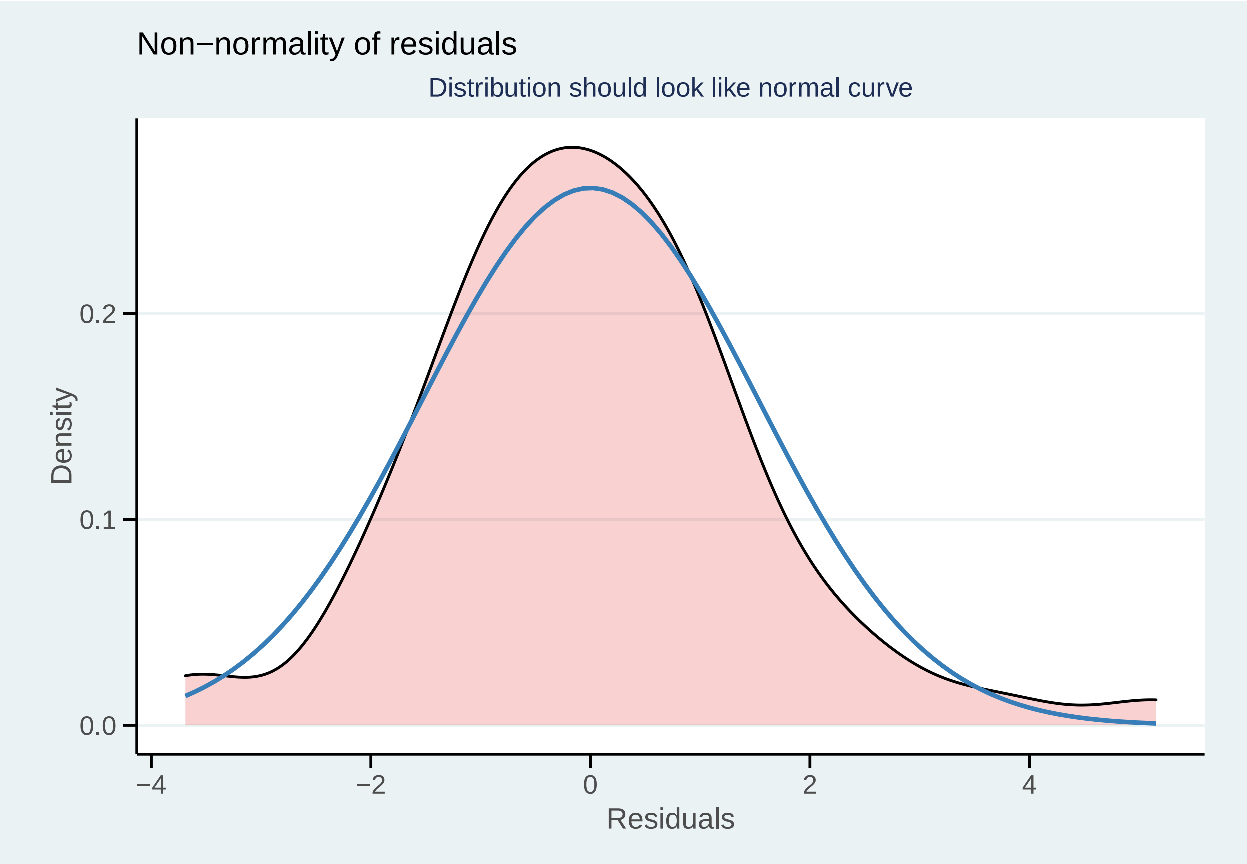

[[1]]


Variance Inflation Factors (multicollinearity)
10.0
7.5
5.0
2.5
0.0
tolerable
good
assasinations
rev_coups
rgdp60
tradeshare
yearsschool

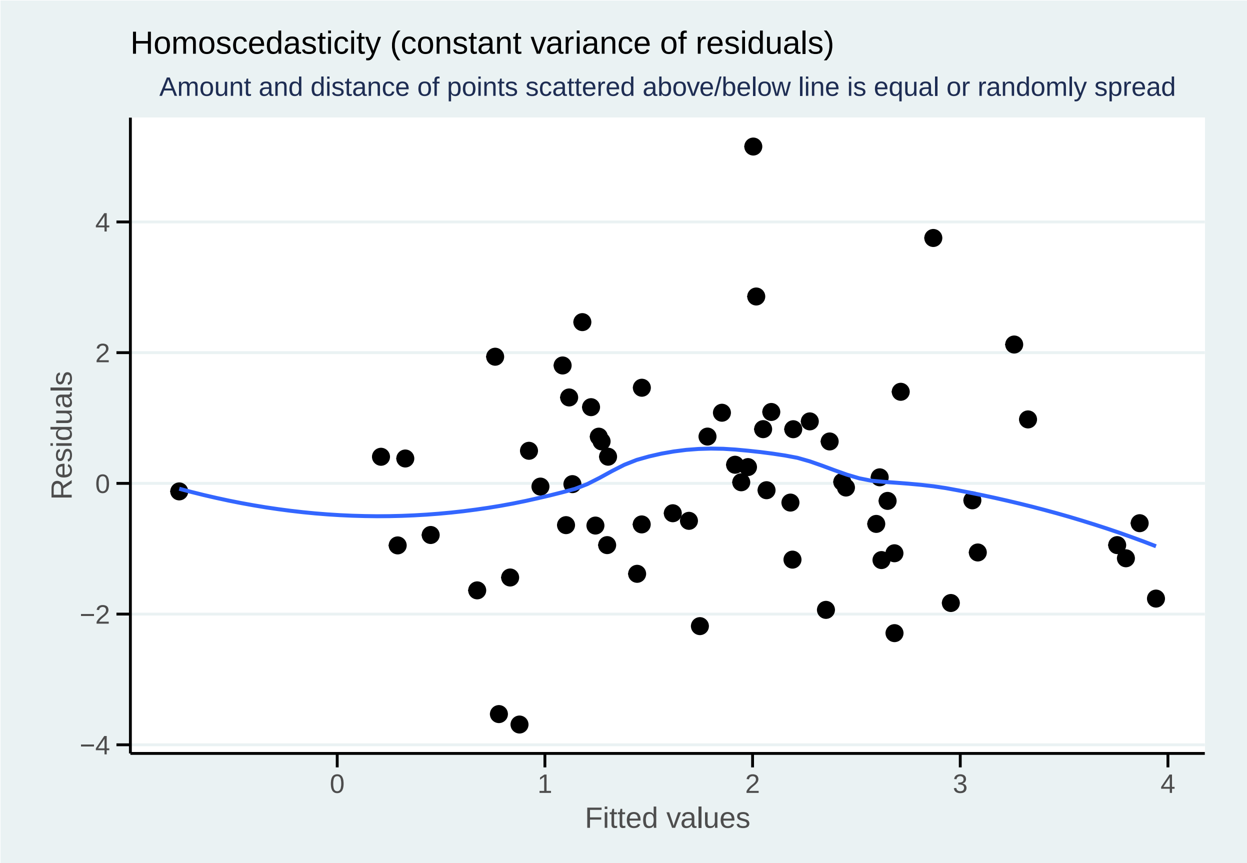
[[2]]


Non−normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
2
0
−2
0
1
2
3
4
Theoretical quantiles (predicted values)

[[3]]


Non−normality of residuals
Distribution should look like normal curve
Density
0.2
0.1
0.0
−4
−2
0
2
4
Residuals

[[4]]


Homoscedasticity (constant variance of residuals)
Amount and distance of points scattered above/below line is equal or randomly spread
Residuals
4
2
0
−2
−4
0
1
2
3
4
Fitted values