Econometría 2 - Taller 1

17/5/2020

Contents

PRIMERA PARTE: Datos asignados, grupo número 21.

1) Elabore un análisis descriptivo de los datos:

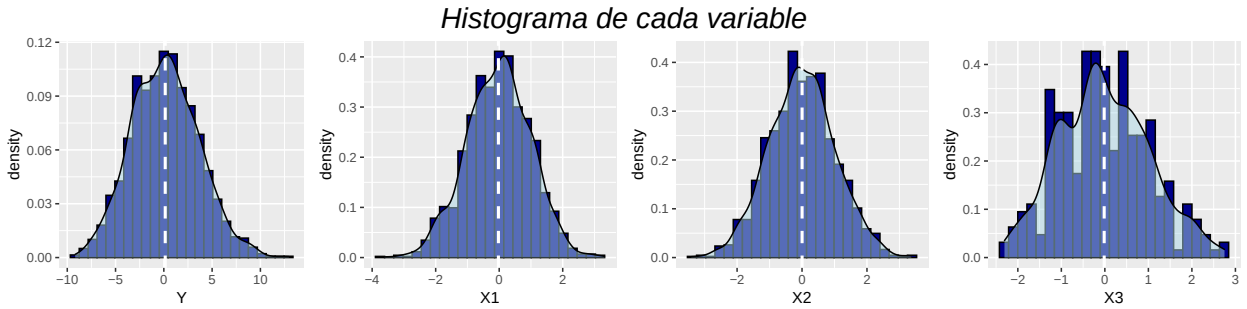
¿Cuantos individuos y periodos de tiempo hay en los 1500 datos de nuestra muestra?

Cantidad de individuos muestreados: 300 Cantidad de fechas muestreadas: 5

¿Hay Valores faltantes o registros duplicados en los datos?

0 = No hay valores faltantes (na). 0 = No hay valores duplicados.

- **Histograma**



Como se puede notar en los histogramas presentados, las cuatro variables, a simple vista, presentan una distribución de frecuencia bastante similar y en donde cada una de ellas asemeja una distribución normal y cuasi simétrica.

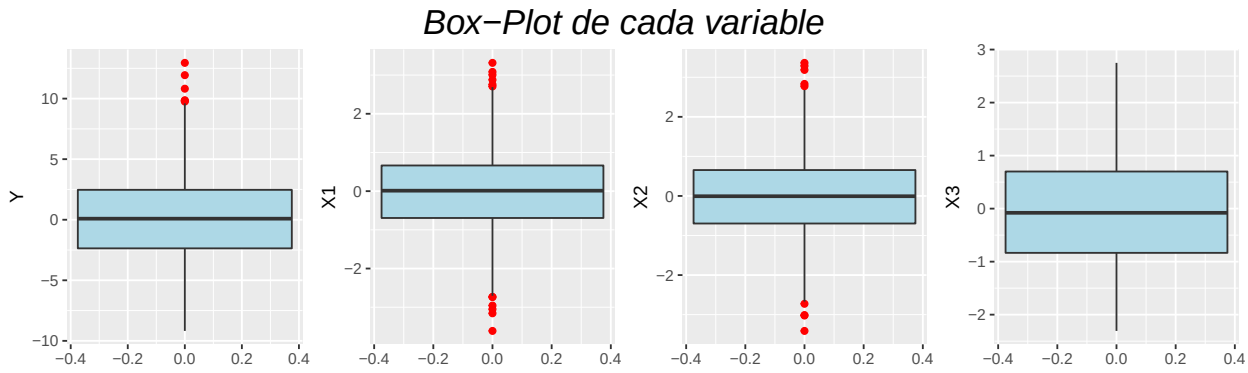
- **Estadística Descriptiva**

  Valores mínimos, máximos, rango intercuartil, mediana y media aritmética de las variables.

Table 1: Medidas Representativas

	Mínimo	Media	Mediana	Máximo	Rango Inter Cuartil
Y	-9.18	0.145	0.087	12.952	4.834
X1	-3.612	-0.024	0.012	3.316	1.357
X2	-3.412	0.002	-0.007	3.364	1.351
X3	-2.307	-0.016	-0.079	2.749	1.535

- Representación gráfica de las anteriores medidas:



Es posible observar también en las gráficas de caja o "boxplot" para Y, X1, X2 y X3 que las distancias entre Q1-Q2 y Q2-Q3 son para los 4 casos, casi iguales, simétricas y que las cuatro poseen una distribución o dispersión bastante similar. También es posible concluir que, a excepción de X3, existen datos atípicos en la zona inferior y superior para X1 y X2 y únicamente en la zona superior para Y.

- Los cuatro momentos estadísticos (Media, Varianza, Asimetría y Curtosis):

Table 2: Momentos estadísticos

	Media	Varianza	Asimetria	Curtosis
Y	0.145	3.463	0.19	2.916
X1	-0.024	1.001	-0.076	2.983
X2	0.002	1.023	0.03	2.966
X3	-0.016	1.032	0.188	2.63

- **Y**

Como puede ser observado en los resultados, la curtosis en Y es 2.91620200746014, con lo cual podría decirse que posee una distribución platicúrtica dado que es <3, fácilmente podría decirse que su distribución posee una concentración normal, pues sus valores se acercan a 3, con lo cual se hablaría de una distribución mesocúrtica.

- **X1**

En cuanto a X1, la curtosis es 2.9825740252592, luego es factible afirmar que aunque posee una concentración muy cercana a una normal (3). Luego su forma de distribución es platicúrtica y cuasi mesocúrtica. Además de esto, observamos que su media es muy cercana a cero (0) y su varianza es muy cercana a uno (1), señales de que su distribución es muy semejante a una Normal Estándar.

- **X2**

Al igual que en la anterior variable, para X2, la distribución es platicúrtica y se acerca mucho a tener una distribución normal mesocúrtica ya que el valor de su curtosis es de 2.96594510002024. Así mismo su media y su varianza la asemejan a una distribución Normal Estándar.

- **X3**

Por último, tenemos que X3 posee una menor distribución de sus datos en torno a su media en comparación con las demás variables estudiadas, a saber, su curtosis es de 2.63019269806822, lo cual corresponde a una forma platicúrtica. Sin embargo, su media y su varianza son, de nuevo, semejantes a las de una distribución Normal Estándar.

- **Asimetría** Por otro lado, mediante la Tabla 2., es posible respaldar que nnuestras variables Y, X1, X2 y X3 poseen una dispersión casi simétrica. Pues su asimetría es 0.19, -0.076, 0.03 y 0.188 respectivamente con valores muy cercanos al 0.

Con estos resultados, teniendo en cuenta más que todo las tres variables explicativas, se podría asegurar la distribución Normal Estándar para la variable dependiente Y. Dado que en la estimación del siguiente modelo:

$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$

Se cumple lo que menciona Damodar N. Gujarati en su libro *"Econometrics":* "cualquier funcion lineal de variables normalmente distribuidas estará también normalmente distribuida" (p. 99). Sin embargo, los análisis de normalidad en las variables los realizaremos en breve.

2) Estimación del modelo bajo el método de Mínimos Cuadrados Combinados (Pooling):

- **Estimación del modelo:**

  Para realizar la estimación de los Datos de tipo longitudinal o Panel, el cual es un Panel de tipo *Corto* en este caso; pues, como se mencionó inicialmente tenemos 300 individuos y 5 periodos de tiempo muestreados. Así, bajo el método de Mínimos Cuadrados Combinados (**MCOC**) tenemos la siguiente ecuación:

  $Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$

Table 3: Mínimos Cuadrados Combinados (Pooled)

	Dependent variable:
	Y
X1	2.067^{***}
	(0.028)
X2	-1.525^{***}
	(0.028)
X3	1.982^{***}
	(0.028)
Constant	0.230^{***}
	(0.028)
Observations	1,500
R^2	0.899
F Statistic	$4{,}445.708^{***}$ (df = 3; 1496)

Como puede ser evidenciado en la ***Tabla 3: Mínimos Cuadrados Combinados (Pooled)***, todos los coeficientes del modelo son estadísticamente significativos bajo un alpha del 1% ($\alpha = 0.01$). Según el coeficiente de determinación R^2, el modelo se ajusta muy bien a las variables consideradas, pues posee un valor del 89%. Finalmente, los errores estándar son bajos en relación a la magnitud de los coeficientes estimados.

3) Inclusión de efectos individuales y efectos temporales al modelo anterior:

- **Comparación del modelo bajo el método "Pooled" cuando se incluyen efectos fijos vs efectos individuales.**

  A nuestro anterior modelo, podemos añadirle con fines investigativos el tiempo o los individuos como coeficientes por estimar con el fin de observar si hay efectos inobservables que varien por individuo o en el tiempo. Si llega a encontrarse la presencia de dichos efectos el modelo sufrirá del llamado sesgo por variable omitida. A continuación se presentan ambas estimaciones y luego se realizan las respectivas pruebas de hipótesis.

Table 4: Mínimos Cuadrados Combinados (Pooled)

	Dependent variable:	
	Y	
	Efectos Temporales	Efectos Individuales
X1	2.067***	2.019***
	(0.028)	(0.029)
X2	−1.526***	−1.529***
	(0.028)	(0.028)
X3	1.982***	1.910***
	(0.028)	(0.264)
as.factor(t)2		
Constant	0.206***	0.725**
	(0.063)	(0.352)
Observations	1,500	1,500
R^2	0.900	0.935
F Statistic	1,916.951*** (df = 7; 1492)	57.138*** (df = 301; 1198)

¿Es significativa la presencia de Efectos Individuales, Efectos Temporales o ambos?

Para las siguientes tres pruebas bajo el **Test de Multiplicadores de Lagrange** las hipótesis a contrastar son:

$H_0 : Efecto\ evaluado\ no\ significativo\ en\ el\ modelo$

$H_1 : Efecto\ evaluado\ significativo\ en\ el\ modelo$

- **Prueba para presencia de Efectos Temporales**

```
Lagrange Multiplier Test - time effects (Breusch-Pagan) for balanced
panels

data:  Y ~ X1 + X2 + X3
chisq = 4.6216, df = 1, p-value = 0.03157
alternative hypothesis: significant effects
```

Se concluye que el modelo sufre de endogeneidad entre alguna(s) de la(s) variable(s) explicativa(s) y el error ideosincrático o temporal el cual está comprendido dentro del término de error compuesto u_{it}.

- **Prueba para presencia de Efectos Individuales**

```
Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels

data:  Y ~ X1 + X2 + X3
chisq = 109.38, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Se concluye que el modelo sufre de endogeneidad entre alguna(s) de la(s) variable(s) explicativa(s) y el efecto fijo α_i el cual está comprendido dentro del término de error compuesto u_{it}.

- **Prueba para presencia de Ambos Efectos**

```
    Lagrange Multiplier Test - two-ways effects (Breusch-Pagan) for
    balanced panels

data:  Y ~ X1 + X2 + X3
chisq = 114, df = 2, p-value < 2.2e-16
alternative hypothesis: significant effects
```

En concordancia con las anteriores dos pruebas, esta última prueba nos confirma que el modelo sufre de endogeneidad entre sus variables explicativas y el término de error compuesto.

Dada la ***Tabla 4: Mínimos Cuadrados Combinados (Pooled)*** y teniendo en cuenta los resultados de la prueba del multiplicador de Lagrange, las estimaciones cambian significativamente tanto para Efectos Individuales como para Efectos Temporales, pues en ambos casos al ser el p-valor menor o igual al nivel de significancia $\alpha = 0.05$ correspondiente, se rechaza la hipótesis nula.

4) Estimación bajo el modelo de Primeras Diferencias (FD):

Para eliminar dichos efectos inobservados en el método de estimación de Datos Panel se puede realizar dicha estimación bajo el modelo de Primeras Diferencias, el cual se expresa de la siguiente manera:

$y_{it} - y_{it-1} = \beta_0 + \beta_1(x_{it1} - x_{i(t-1)1}) + \beta_2(x_{it2} - x_{i(t-1)2}) + \beta_3(x_{it3} - x_{i(t-1)3}) + (u_{it} - u_{it-1})$

Table 5: Primeras Diferencias

	Dependent variable:
	Y
X1	2.011^{***}
	(0.028)
X2	-1.533^{***}
	(0.028)
Constant	0.034
	(0.040)
Observations	1,200
R^2	0.870
F Statistic	$4{,}002.255^{***}$ (df = 2; 1197)

Partiendo de la ***Tabla 5: Primeras Diferencias***, se obtiene que: en la eliminación de efectos inobservables bajo el modelo de Primeras Diferencias, los coeficientes estimados son significativos al 1%, a excepción del intercepto β_0. La bondad de ajuste no cambia en relación a la estimación hecha por MCO Combinados (pooling) y los errores estándar no cambian mucho pese a la reducción del tamaño de la muestra. Además, es importante notar que los signos de los coeficientes no cambian. Por otro lado, el cambio en los coeficientes estimados no sorprenden si se tiene en cuenta que los estimadores de Efectos Fijos y de Primeras Diferencias se acercan más cuanto más pequeña es la dimensión temporal de la muestra.

5) Estimación del modelo bajo los métodos de Efectos Fijos (Within) y Variables Binarias (VB):

Otra manera para eliminar dichos efectos inobservados en el método de estimación de Datos Panel se puede realizar bajo el modelo de Efectos Fijos y su variante de Variables Binarias, los cuales se expresan de la siguiente manera:

- Efectos Fijos:

$$y_{it} - \overline{y}_i = \beta_0 + \beta_1(x_{it1} - \overline{x}_{i1}) + \beta_2(x_{it2} - \overline{x}_{i2}) + \beta_3(x_{it3} - \overline{x}_{i3}) + (u_{it} - \overline{u}_i)$$

- Variables Binarias:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \lambda_1 n_1 + ... + \lambda_N n_N + u_t \qquad \textit{donde } n : 1...N \textit{ numero de individuos}$$

- **Comparación de ambos modelos:**

  Para fines de comodidad en la observación se ocultaron los 299 estimadores individuales de Variables Binarias ya que no son relevantes dentro del actual contexto.

Table 6: Efectos Fijos vs. Variables Binarias

	Dependent variable:	
	Y	
	panel *linear* Efectos Fijos	*OLS* Variables Binarias
X1	2.019^{***}	2.019^{***}
	(0.029)	(0.029)
X2	-1.529^{***}	-1.529^{***}
	(0.028)	(0.028)
X3		1.910^{***}
		(0.264)
as.factor(id)2		
Observations	1,500	1,500
R^2	0.867	0.935
F Statistic	$3{,}896.582^{***}$ (df = 2; 1198)	57.138^{***} (df = 301; 1198)

Como puede observarse en la ***Tabla 6: Efectos Fijos vs. Variables Binarias***, los estimadores son idénticos para Efectos Fijos y para Variables Binarias y en ambos casos son significativos al 1%. Dada su naturaleza, no se puede decir si en realidad el modelo de Variables Binarias ofrece una mayor bondad de ajuste simplemente por tener un R2 mayor, pues ambos coeficientes de determinación no son comparables. Además, los grados de libertad son identicos, pues con N número de individuos en la muestra, T momentos en el tiempo y k variables regresoras del modelo:

Cuando tenemos Efectos Fijos se pierde un grado de libertad por cada efecto fijo eliminado:

Grados de Libertad de Efectos Fijos $= NT - k - N$

Mientras que, Variables binarias es simplemente MCOC al se le incluye una variable dummy por cada observación de corte transversal N, entonces sus grados de libertad son los mismos de MCOC restándole N:

Grados de Libertad de Variables Binarias $= NT - k + 1 - N - 1 = NT - k - N$

6) Estimación e interpretación de thetha para el modelo bajo el método de Efectos Aleatorios (Random):

Para eliminar dichos efectos inobservados en el método de estimación de Datos Panel se puede realizar dicha estimación bajo el modelo de Efectos Aleatorios, el cual se expresa de la siguiente manera:

$y_{it} - \theta\overline{y}_i = \beta_0(1-\theta) + \beta_1(x_{it1} - \theta\overline{x}_{i1}) + \beta_2(x_{it2} - \theta\overline{x}_{i2}) + \beta_3(x_{it3} - \theta\overline{x}_{i3}) + (u_{it} - \theta\overline{u}_i)$

El estimador de θ puede ser obtenido según la fórmula:

$\hat{\theta} = 1 - (\frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + T\widehat{\sigma}_a})^{\frac{1}{2}}$

Al revisar la anterior ecuación, es notable que cuando $\theta = 0$ se trata del modelo simple por MCOC (donde no hay efectos temporales ni individuales inobservados; cuando $\theta = 1$ se trata del modelo bajo Efectos Fijos. Al no obtener ninguno de estos dos resultados, se infiere o por lo menos se intuye que nuestro modelo deberá ser estimado bajo Efectos Aleatorios. Obtener, pues, el valor de este θ es especialmente necesario para nuestro ejercicio. Sin embargo, no calcularemos el valor poblacional de dicho θ sino, como lo muestra la ecuación, su valor estimado. Obtener el valor estimado $\widehat{\theta}$ hará que nuestro método de estimación sea identificado bajo el nombre de **Mínimos Cuadrados Generalizados Factibles (*MCGF*)** y nuestros estimadores para los $\hat{\beta}$ se encontrarán bajo el modelo de **Efectos Aleatorios**. Cabe anotar que también es posible obtener dicho theta por medio del summary que arroja R para un modelo aleatorio de datos panel. Luego por su simplicidad y precisión, se decidió tomar esta estimación de theta $\hat{\theta}$ para nuestro estudio.

```
$ercomp
                 var std.dev share
idiosyncratic 0.9773  0.9886  0.81
individual    0.2292  0.4788  0.19
theta: 0.3216
```

Ahora, como ya fue mencionado, se intuía (bajo las pruebas anteriormente realizadas) que el valor para nuestro theta no podría ser 0 (ya que hay presencia de efectos inobservados bajo ese método), y ya su valor tampoco es igual a 1 se entiende que el modelo de Efectos Fijos tampoco es eficaz para nuestros datos. Lo que señala que el mejor modelo será el de Efectos Aleatorios.

7) Prueba de hipótesis: Random vs. Within.

Recordando que en el punto 5 se determinó que Efectos Fijos (Within) es preferible a Variables Binarias (VB) corresponde ahora validar entre Efectos Fijos y Efectos Aleatorios ¿Cuál es el mejor modelo? Además, es pertinente revisarlo ya que en el punto anterior se dejó claro que el valor de theta nos deja una intuición que vale la pena probar. Los dos modelos pueden verse a continuación:

Table 7: Efectos Fijos vs. Efectos Aleatorios

	Dependent variable:	
	Y	
	Efectos Fijos	Efectos Aleatorios
X1	2.019^{***}	2.044^{***}
	(0.029)	(0.027)
X2	-1.529^{***}	-1.527^{***}
	(0.028)	(0.027)
X3		1.983^{***}
		(0.037)
Constant		0.229^{***}
		(0.038)
Observations	1,500	1,500
R^2	0.867	0.887
F Statistic	$3{,}896.582^{***}$ (df = 2; 1198)	$11{,}776.300^{***}$

- **Test de Hausman**

  La prueba de Hausmann es en esencia una prueba de endogeneidad, en este caso concreto. Se quiere probar la endogeneidad del término de error (heterogeneidad inobservable), que **teóricamente** es eliminado en el modelo de Efectos Fijos. Esta endogeneidad suele darse por la omisión de variables relevantes que en efecto están correlacionadas con las variables que sí se tienen en cuenta, el efecto de estas variables omitidas sobre la estimación del modelo es absorbida entonces por el término de error.

  La regla de decisión será, pues, esta **Prueba de Hausman** donde las hipótesis a contrastar son:

  $H_0 : No\ hay\ diferencias\ significativas\ entre\ EF\ y\ EA.$

  $H_1 : Hay\ diferencias\ significativas\ entre\ EF\ y\ EA.$

```
    Hausman Test

data:  Y ~ X1 + X2 + X3
chisq = 7.5846, df = 2, p-value = 0.02254
alternative hypothesis: one model is inconsistent
```

Como se puede observar se cae en zona de RH_0 ya que el p-value es menor a nuestro nivel de significancia $\alpha = 0.05$, con lo que se concluye que hay diferencias significativas entre ambos modelos.

8) Comparación de todos los modelos estimados hasta ahora.

A continuación se pueden observar los modelos trabajados hasta ahora:

Table 8: Comparación General

	Dependent variable:				
	Y				
	Pooling	*panel linear* F.D.	Within	*OLS* VB	*panel linear* Random
X1	2.067^{***}	2.011^{***}	2.019^{***}	2.019^{***}	2.044^{***}
	(0.028)	(0.028)	(0.029)	(0.029)	(0.027)
X2	-1.525^{***}	-1.533^{***}	-1.529^{***}	-1.529^{***}	-1.527^{***}
	(0.028)	(0.028)	(0.028)	(0.028)	(0.027)
X3	1.982^{***}			1.910^{***}	1.983^{***}
	(0.028)			(0.264)	(0.037)
as.factor(id)2					
Observations	1,500	1,200	1,500	1,500	1,500
R^2	0.899	0.870	0.867	0.935	0.887

En definitiva, como ya se ha mencionado, estos modelos no son comparables bajo los datos que son visibles en la ***Tabla 8: Comparación General***, pues sus coeficientes de determinación R^2 no nos sirven para la selección del mejor modelo. Para esta decisión se puede seguir lo desarrollado hasta ahora en los puntos anteriores.

En primera instancia, la estimación bajo el método de MCOC resulta insuficiente pues se detectó la presencia de efectos temporales e individuales inobservados, violando el supuesto de endogeneidad de las regresoras respecto al término de error. Para la corrección de este problema se realizaron las pruebas detalladas para Primeras Diferencias, Efectos Fijos, Variables Binarias y Efectos Aleatorios. Donde poco a poco se fueron descartando las opciones y, en el punto anterior, se llegó a que mediante la interpretación del θ estimado, sumado a la prueba de Hausman, el mejor modelo es el de Efectos Aleatorios.

9) Validación de supuestos del modelo seleccionado (Random):

- **Distrubución normal de las variables**

  Como se mencionó anteriormente, confirmar la revisión de nuestra regresión cuando goza de las propiedades MELI, principalmente enfocándonos en la normalidad en las regresoras, asegurará que la variable regresada también gozara, en su estimación, una distribución normal. Para la validación de Normalidad utilizaremos la prueba de **Shapiro - Wilk** para muestras grandes.

Las hipótesis a contrastar son:

$H_0 : Normalidad$

$H_1 : No\ Normalidad$

- *Para la variable Y:*

```
Shapiro-Wilk normality test
```

```
data:  BD$Y
W = 0.99698, p-value = 0.005313
```

Para el caso de la variable dependiente Y, el p-value para la prueba realizada es menor al $\alpha = 0.05$, por lo tanto se cae en zona de RH_0 concluyendo que esta variable no goza de normalidad.

- *Para la variable X1:*

```
    Shapiro-Wilk normality test

data:  BD$X1
W = 0.99892, p-value = 0.5162
```

Para el caso de la variable explicativa X1, el p-value para la prueba realizada es mayor al $\alpha = 0.05$, por lo tanto se cae en zona de NRH_0 concluyendo que esta variable goza de normalidad.

- *Para la variable X2:*

```
    Shapiro-Wilk normality test

data:  BD$X2
W = 0.9994, p-value = 0.934
```

Para el caso de la variable explicativa X2, el p-value para la prueba realizada es mayor al $\alpha = 0.05$, por lo tanto se cae en zona de NRH_0 concluyendo que esta variable goza de normalidad.

- *Para la variable X3:*

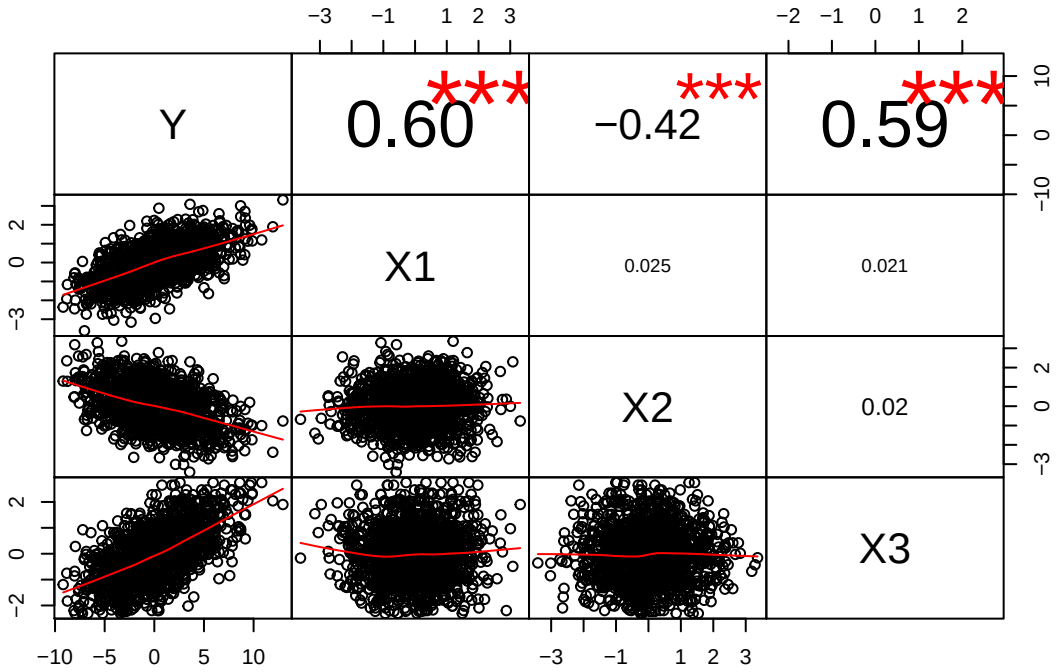
```
    Shapiro-Wilk normality test

data:  BD$X3
W = 0.99216, p-value = 3.649e-07
```

Para el caso de la variable explicativa X3, el p-value para la prueba realizada es menor al $\alpha = 0.05$, por lo tanto se cae en zona de RH_0 concluyendo que esta variable no goza de normalidad.

- **Correlación entre las variables**

Se espera que las variables del modelo no sufran de multicolinealidad perfecta (correlación = -1 o 1). Para esta propiedad del modelo podemos ver la siguiente gráfica:



Es posible apreciar el anterior gráfico en donde, a simple vista, y dada la nube de puntos formada entre las variables, hay una relación directa entre éstas; dentro de la gráfica se puede encontrar bajo qué nivel de significancia resultan significativas las correlaciones bajo el Test de Correlación de Pearson para cada variable con respecto a Y. Los asteriscos indican que las variables explicativas son significativas bajo un alpha de 0.01% $\alpha = 0.001$ donde X1 posee un valor de 0.5983464 según el test de Pearson, lo cual traduciría a una correlación positiva entre las variables. Esto implica que la relación entre X1 y Y es directa (cuando aumenta X1, aumenta Y). En cuanto a Y y X2, según el test de Pearson, se da una correlación de -0.4237041, es decir una relación inversa, tal y como puede ser observado en la nube de puntos. Por último, Y y X3 poseen una correlación de 0.5939949, casi tan acentuada como Y y X1, pero directa también.

- **Correcta especificación del modelo**

Mediante la prueba de Ramsey se determina si el modelo está correctamente especificado o no. Dicha prueba se realiza estimando alternativamente:

$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \gamma_1 \widehat{Y}_{it}^2 + \gamma_2 \widehat{Y}_{it}^3 + u_{it}$

Las hipótesis a contrastar son:

$H_0 : \gamma_1 = \gamma_2 = 0$

$H_1 : al\ menos\ uno\ de\ los\ \gamma \neq 0$

```
Linear hypothesis test

Hypothesis:
cuadra = 0
cubica = 0

Model 1: restricted model
Model 2: Y ~ X1 + X2 + X3 + cuadra + cubica
```

```
  Res.Df Df  Chisq Pr(>Chisq)
1   1496
2   1494  2 2.1953     0.3336
```

Como se puede observar, el p-value asociado (aquí mencionado como $Pr(> Chisq)$) es mayor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de NRH_0 y se concluye que no es necesario añadir exponentes al modelo original; es decir, que el modelo está correctamente especificado.

- **Homoscedasticidad**

Para comprobar si el modelo goza de varianza constante en el término de error se pueden realizar varias pruebas de hipótesis.

En todas ellas las hipótesis a contrastar son:

$H_0 : Var(u) = \sigma_u^2$

$H_1 : Var(u) \neq \sigma_u^2$

- *Prueba de BREUSCH - PAGAN:*

```
    studentized Breusch-Pagan test

data:  RE
BP = 1.859, df = 3, p-value = 0.6022
```

Ya que el p-value para la prueba realizada es mayor al $\alpha = 0.05$, se cae en zona de NRH_0 concluyendo que el modelo goza de Homoscedasticidad en el término de error.

- *Prueba de GOLFED - QUANDT:*

```
    Goldfeld-Quandt test

data:  RE
GQ = 0.99766, df1 = 746, df2 = 746, p-value = 0.5128
alternative hypothesis: variance increases from segment 1 to 2
```

Ya que el p-value para la prueba realizada es mayor al $\alpha = 0.05$, se cae en zona de NRH_0 reafirmando que el modelo goza de Homoscedasticidad en el término de error.

- **No Auto-Correlación serial**

Para comprobar si el término de error del modelo (u) se correlaciona con sus rezagos en un proceso autoregresivo de orden 1 ($AR[1]$) o de un orden mayor a 1 ($AR[2+]$) se utiliza la prueba de Durbin & Watson para el primer caso y la prueba de Breusch & Godfrey para el segundo caso. El modelo a evaluar para el primer caso ($AR[1]$) es:

$u_t = \rho u_{t-1} + \epsilon_t$

Y para el segundo caso ($AR[2+]$) es:

$u_t = \rho_1 u_{t-2} + \rho_2 u_{t-3} + ... + \rho_p u_{t-p} + \epsilon_t$

- *Prueba de DURBIN - WATSON:*

Las hipótesis a contrastar son:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

```
    Durbin-Watson test for serial correlation in panel models

data:  Y ~ X1 + X2 + X3
DW = 1.9854, p-value = 0.3829
alternative hypothesis: serial correlation in idiosyncratic errors
```

Ya que el p-value para la prueba DW realizada es mayor al $\alpha = 0.05$, se cae en zona de NRH_0 con lo que se concluye que no hay autocorrelación bajo un proceso autoregresivo de 1 periodo $AR[1]$ en el término de error. Ahora, hay que revisar si sucede con periodos mayores a 1.

- *Prueba de BREUSCH - GODFREY:*

Las hipótesis a contrastar son:

$H_0 : \rho_1 = \rho_2 = \rho_p = 0$

$H_1 : Al\ menos\ uno\ de\ los\ \rho \neq 0$

```
    Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  Y ~ X1 + X2 + X3
chisq = 2.3587, df = 5, p-value = 0.7976
alternative hypothesis: serial correlation in idiosyncratic errors
```

El p-value para la prueba Breusch-Godfrey realizada es mayor al $\alpha = 0.05$, se cae en zona de NRH_0 con lo que se concluye que no hay autocorrelación bajo un proceso autoregresivo de dos o más periodos $AR[1+]$ en el término de error.

Con ambas pruebas se concluye que no hay ningún tipo de autocorrelación serial en el término de error del modelo.

- **Normalidad**

La siguiente prueba ya no es aplicada a las variables presentes para la estimación del modelo, sino que ahora se trata de la evaluación del supuesto de *Normalidad* para los **residuos** del modelo.

Nuevamente, las hipótesis a contrastar son:

$H_0 : Normalidad$

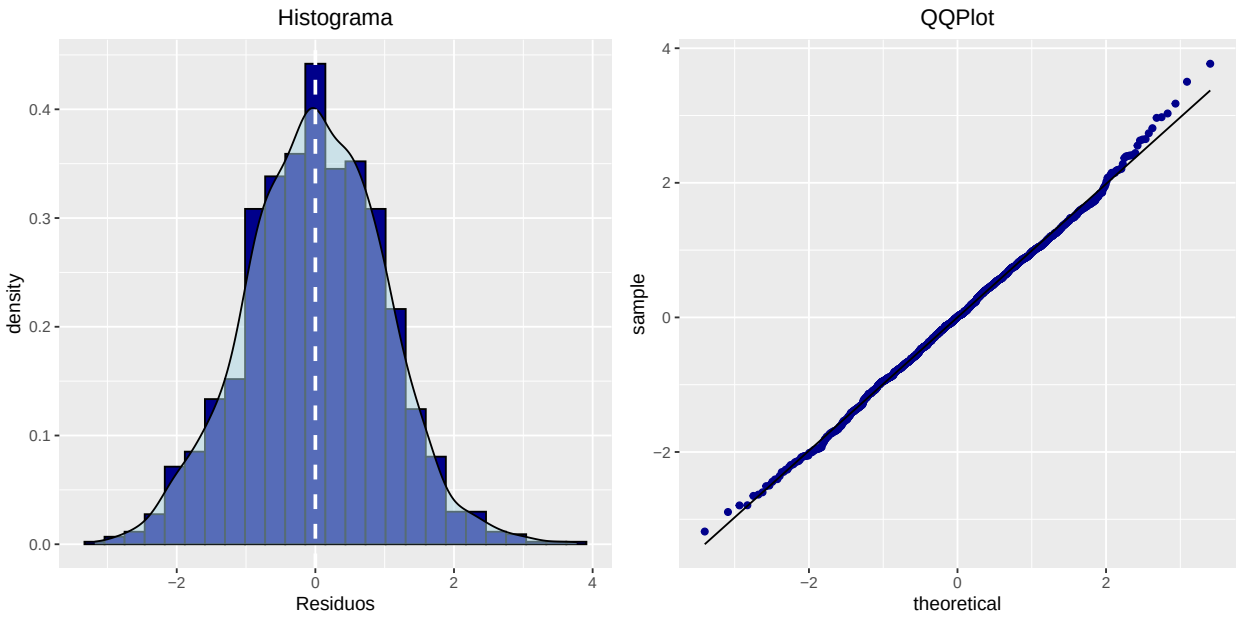
$H_1 : No\ Normalidad$

```
    Shapiro-Wilk normality test

data:  RE$residuals
W = 0.99849, p-value = 0.208
```

Ya que el p-value para la prueba realizada es mayor al $\alpha = 0.05$, se cae en zona de NRH_0 con lo que se concluye que los residuos del modelo gozan de una distribución normal.

- *Visualización*



SEGUNDA PARTE: Análisis sobre datos del coronavirus covid-19.

- **Variables seleccionadas:**

```
Rows: 363
Columns: 18
$ Country             <fct> Algeria, Algeria, Algeria, Algeria, Algeria, Ar...
$ Week                <int> 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5,...
$ Death.per.1m        <dbl> 0.14, 0.24, 0.97, 3.48, 3.10, 0.02, 0.02, 0.20,...
$ Lagged.death.per.1m <dbl> 0.02, 0.14, 0.24, 0.97, 3.48, 0.02, 0.02, 0.02,...
$ Cases.per.1m        <dbl> 1.35, 4.31, 13.81, 17.17, 13.92, 0.81, 2.54, 9....
$ Lagged.cases.per.1m <dbl> 0.31, 1.35, 4.31, 13.81, 17.17, 0.20, 0.81, 2.5...
$ Weakly.Tests.per.1m <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, 106, 290, 2...
$ Stringency.Index    <dbl> 47.62, 85.71, 90.48, 95.24, 95.24, 38.10, 95.24...
$ FL.Stringency.Index <int> 33, 48, 86, 90, 95, 14, 38, 95, 95, 95, 100, 29...
$ SL.Stringency.Index <int> 14, 33, 48, 86, 90, 14, 14, 38, 95, 95, 95, 29,...
$ Population          <int> 42228429, 42228429, 42228429, 42228429, 4222842...
$ UHC.sc.index        <int> 78, 78, 78, 78, 78, 76, 76, 76, 76, 76, 76, 87,...
$ Population65        <dbl> 6.4, 6.4, 6.4, 6.4, 6.4, 11.1, 11.1, 11.1, 11.1...
$ Doctors             <dbl> 120.7, 120.7, 120.7, 120.7, 120.7, 390.7, 390.7...
$ Nurses              <dbl> 194.7, 194.7, 194.7, 194.7, 194.7, 421.2, 421.2...
$ Beds                <int> 190, 190, 190, 190, 190, 500, 500, 500, 500, 50...
$ isolate.patients    <int> 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
$ Government.expend   <int> 676, 676, 676, 676, 676, 1140, 1140, 1140, 1140...
```

Con miras a poder explicar la tasa de fatalidad por el coronavirus COVID-19 en múltiples países, de manera rigurosa se escogieron una serie de variables regresoras que poseen una relación y explican en cierta medida la tasa de fatalidad por coronavirus. Así, las variables regresoras escogidas fueron: país, semana, muertes por millón, rezagos de muertes por millón, casos por millón, rezagos de casos por millón, pruebas semanales por millón, Índice de rigurosidad, primer rezago del índice de rigurosidad, segundo rezago del índice de rigurosidad, población, índice UHC de cobertura de salud, población de personas con 65 o más años, doctores, enfermeras, camas de hospital, capacidad de aislar pacientes, y por último la variable regresora de gasto gubernamental en salud en cada país. A continuación, se hará una breve explicación del porqué se decidió añadir estas variables.

- **País (Country)** Esta variable será usada como ID en nuestro modelo, de manera que no hay duda sobre añadirla o no a éste, pues será usado como referente.
- **Semana (Week)** Finalmente fue decidido añadir esta variable y no otra ya que posee información lo suficientemente generalizada sobre el tiempo como para ser adecuada; a diferencia de la variable, por ejemplo, "Fecha de la primera muerte" que nos habla de algo mucho más específico; sin embargo, esta última será añadida al informe presentado para que pueda ser evidenciada la fecha como un marco de referencia al cual acudir temporalmente.
- **Muertes por millón (Death.per.1.million)** Esta es la variable Y que será explicada en nuestro modelo, para lo cual fue pensado en un principio una transformación de tipo logarítmica con el ánimo de ser medida como una tasa y que no dependa del número de población. Sin embargo fue allí cuando nos enfrentamos a uno de los problemas con base en los datos, pues para esta variable, se poseen datos equivalentes a 0, lo que implica que no sea válida la transformación, pues el logaritmo de 0 no existe. Para resolver este problema y según varios autores, era posible cambiar los valores de 0 por 0.00000001, lo cual no fue una opción para el grupo de trabajo ya que sería modificar los valores de los datos por unos que no son reales y con carencia de argumentos; lo que implicaría, a nuestro juicio, que nuestro modelo se aleja drásticamente de la realidad. Finalmente se decidió no realizar la transformación logarítmica, sino trabajar con el panel desbalanceado, pero sin omitir los datos o 0 encontrados.

- **Rezagos (Lagged—)** A nuestro modelo también se le añadieron cuatro variables rezagadas, a saber, los rezagos de las variables muertes por millón, casos por millón, y los dos rezagos de la variable de índice de rigurosidad. Poniendo en contexto y para casos generales, los valores de una variable para el periodo T2 para la variable rezagada son los valores de la variable "normal" siendo rezagada para el periodo T1; luego es crucial que dichas variables sean incluidas en el modelo, pues, por ejemplo, la cantidad de muertes por millón depende de la cantidad de muertes por millón de la semana inmediatamente anterior y exclusivamente de esta, sino también de la anterior a la anterior, así, estas variables rezagadas inciden considerablemente en lo estudiado y no incluirlas podría resultar en ser un error. Dada la naturaleza de las variables rezagadas tal y como fue explicado, es normal que dichas variables posean una alta correlación con la variable original no rezagada, aunque ello no implica que haya sido un error de correlación.

- **Casos por millón (Cases.per.1.million)** Creemos que existe una fuerte relación entre la tasa de mortalidad y el número de casos por millón, y dicha relación se crea de manera natural, a simple vista es posible deducir que a mayor cantidad de casos, habrá más muertos; luego se decidió incluir esta variable con la unidad de medida de "por millón" ya que brinda información apropiada al ser un concepto homólogo al de porcentaje.

- **Pruebas semanales por millón (Tests.per.1.million)** Por otro lado, se decidió también añadir la variable que indica el número de pruebas semanales realizadas en cada país no solo por lo que implica explícitamente, sino porque también podría hablar de cuán preparado se encuentra el país para afrontar la pandemia, pues a mayor número de pruebas, se espera que se encuentren más pacientes con el virus, y sus muertes, eventualmente puedan ser relacionadas esta enfermedad. Esta variable fue escogida sobre las otras similares bajo el mismo argumento de una unidad de medida homólogo al porcentaje.

- **Índice de rigurosidad (Stringency Index)** Según lo encontrado, el Stringency Index tiene como objetivo principal rastrear y comparar las medidas políticas tomadas en diferentes países; lo que implica que un mayor número de índice de rigurosidad, corresponde a una calificación positiva de un grupo de políticas y medidas tomadas; luego se espera que entre mayor sea el índice de rigurosidad, menor sea la tasa de mortalidad, y es por esta razón que creímos desde un principio incluirla en nuestro modelo como una variable regresora que explica dicha tasa mencionada.

- **Población (Population)** Sin duda alguna creímos pertinente incluir la variable la cual nos habla del número de población, pues permite comparar directamente los países con respecto a su tasa de mortalidad y con respecto a las otras variables regresoras. No es comparable un país que tenga una tasa de mortalidad del 1% con 10 habitantes a un país con una tasa de mortalidad del 1% y con 1000 habitantes. Para el correcto análisis y dado que no se tenía una unidad de medida homóloga al de porcentaje como en variables regresoras anteriores, entonces se decidió aplicarle una transformación logarítmica; más adelante se profundizará un poco más.

- **Índice UHC de cobertura de salud (UHC.service.coverage.index)** Según la OMS, el índice UHC de cobertura de salud califica dentro de un rango de 0-100 precisamente lo que su nombre indica, la cobertura en salud que tiene un país a lo largo de su territorio; dentro de los datos es posible por ejemplo encontrar que el valor mínimo es de 40 y su valor máximo es 89, lo que implica que ese país en específico posee una muy buena cobertura en salud. Se consideró también que esta variable podría influir de manera significativa en la tasa de mortalidad de un país y es por esta razón que se incluyó.

- **Población de personas con 65 o más años (Population.65.and.above)** Esta variable en específico, aunque su estudio es exigido en uno de los apartados del taller, consideramos también incluirla ya que creemos que incide de manera significativa la cantidad de personas con 65 años o más y la tasa de mortalidad, luego a mayor tasa de este tipo de población, también habrá mayor tasa de muertes por coronavirus dado las bajas defensas que estas personas poseen en su microorganismo y la capacidad de respuesta ante este virus. Esta variable originalmente se encuentra dada como tasa en un porcentaje, de manera que no fue necesaria aplicar una transformación logarítmica con este fin sino trabajar con ella tal y como fue recopilada.

- **Doctores, enfermeras y camas de hospital (Doctors – Nurses.and.midwives – Hospital.beds)** Se cree que la cantidad de doctores, enfermeras y camas de hospital puede incidir significativamente con la tasa de mortalidad, pues a mayor cantidad de doctores, enfermeras y camas de hospital, habrá mayor facilidad para enfrentar el virus en pacientes contagiados. De la misma manera y bajo los mismos argumentos anteriormente mencionados, se espera trabajar estas variables en términos de porcentaje puesto que depende de cada país, para ello se aplicará a cada una de las variables y por separado, una transformación logarítmica que podrá ser apreciada eventualmente en el documento; por último se evaluará si cada una de estas transformaciones logarítmicas aplican como se espera para cada una de las variables mencionadas.

- **Capacidad de aislar pacientes (Capacity.to.isolate.patients)** Por otro lado, se decidió incluir esta variable binaria con la cual se mide la capacidad que tiene el sistema de salud de cada país para aislar o no los pacientes que poseen el virus o que poseen alto riesgo de poseerlo. Sin duda alguna creemos que el que un país no tenga la capacidad de este aislamiento aumentará significativamente la tasa de mortalidad ya que implicaría también una alta tasa de contagio. Al ser una variable binaria y por ende con poca variación en sus datos, se decidió no hacer un análisis descriptivo de ésta como si se hizo con otras variables.

- **Gasto gubernamental en salud (Government.health.expenditure)** Por último, la variable regresora de gasto gubernamental en salud fue considerada como una variable con alta significancia que podría incidir considerablemente en la tasa de mortalidad ya que dependiendo de este gasto gubernamental, el área de la salud tendrá o no los suficientes recursos para hacerse cargo del control de una pandemia como lo es la del virus COVID-19; así, entre más gasto gubernamental en salud haya, se espera que haya menos tasa de mortalidad dados todos los instrumentos para disminuir esta tasa.

- **¿Hay valores faltantes (NA) en los datos?**

  Hay 114 valores faltantes (NA) en la variable 'Weakly Test per 1 Million'.

¿Qué hacer con esos valores faltantes?

Para el actual caso de estudio **NO** se balanceará el panel, pues esto implicaría estandarizar valores faltantes (NA) para distintos países en distintas circunstancias, también implicaría asemejar todos los países bajo un estudio temporal donde la cantidad de semanas sea equivalente para todos. Esto último es lo que presenta mayor complejidad, pues no podemos ignorar el hecho de que el virus no llegó al mismo tiempo para todos los lugares del mundo, además que las condiciones que hicieron que la primera persona muriera debido al covid-19 también varían, estandarizar el panel de datos, en definitiva, consideramos que sería una inclusión de sesgo bastante arriesgada y muy seguramente no tan fructífera para el análisis, propiamente dicho, de los datos presentados.

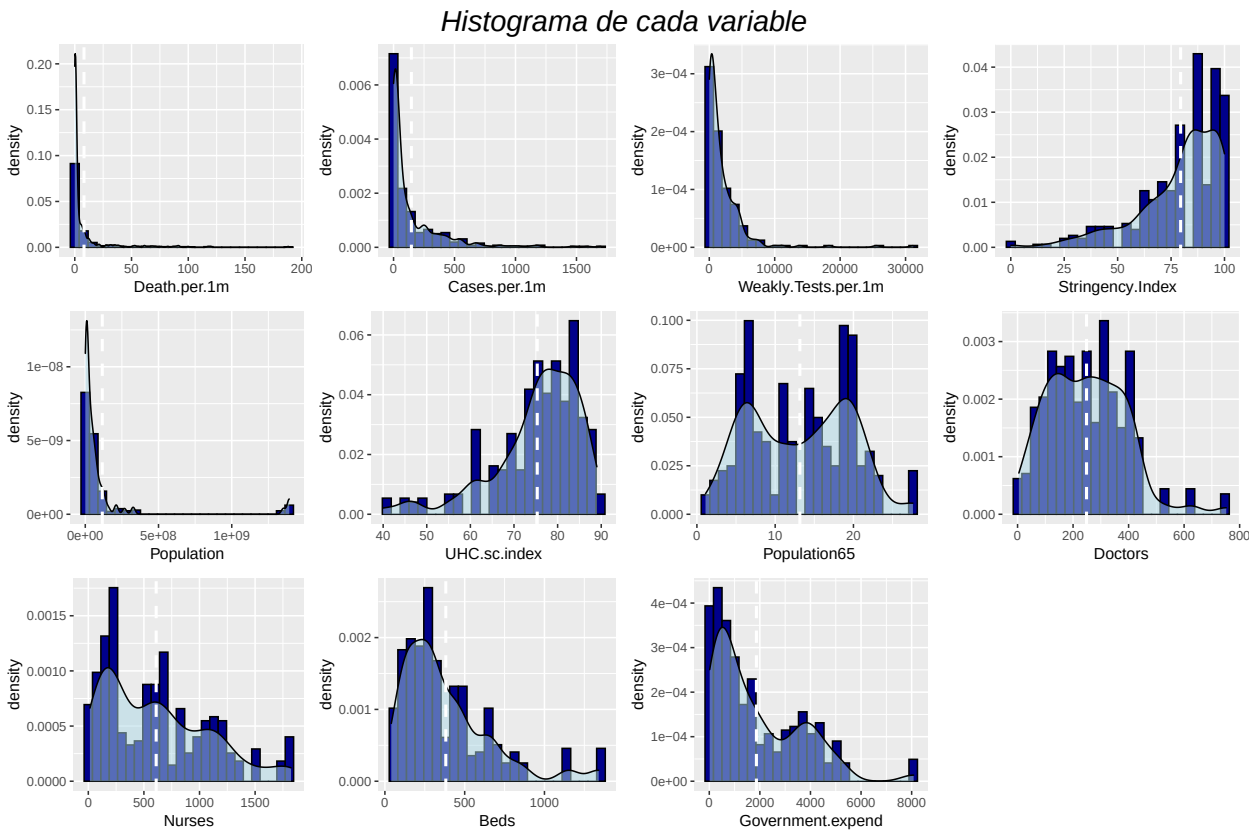
- **Formato de las variables:**

La variable "Week" está siendo tomada bajo un formato erroneo, hay que transformar el formato a variable cualitativa.

Finalmente, se puede convertir los datos en un panel data frame para obtener un mayor provecho a los datos que son de tipo longitudinal.

1) Estadística Descriptiva de las variables

- **Histogramas**



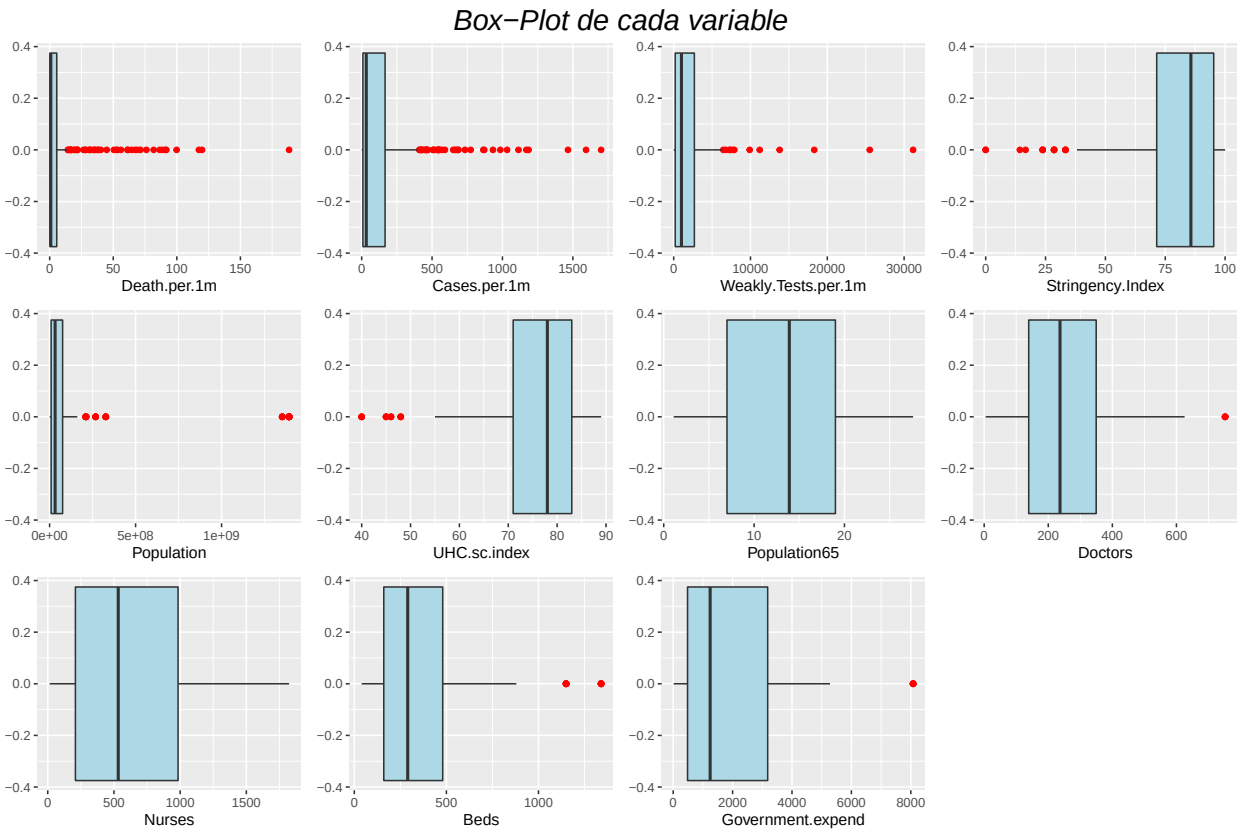
Como puede ser observado los histogramas de cada una de las variables, a simple vista es posible afirmar que ninguna presenta una distribución normal, y en algunas de estas la concentración de datos hacia un lado de la tabla es notorio, por ejemplo en la variable "Death per 1m" que posee una distribución hacia el lado izquierdo de la tabla, esta información podría ser también comparada con el valor de su asimetría (4.323). De manera contraria también es posible encontrar asimetrías negativas como la de la variable "Stringency index" (-1.361) que corresponder a una asimetría a derecha donde su distribución se concentra hacia este lado del histograma como puede ser evidenciado y respaldado. En cuanto a la curtosis de las variables, es factible afirmar que la mayoría posee una distribución de tipo leptocúrtica ya que su valor de curtosis es mayor a 3, luego su forma será un poco más puntiaguda en comparación con otras formas de distribución como la platicúrtica para el caso de la variable de "Población de 65 o más años" donde su valor de curtosis $(1.927) < 3$.

- Valores mínimos, máximos, rango intercuartil, mediana y media aritmética de las variables.

- Gráficos para los anteriores valores:

Table 9: Medidas Representativas

	Mínimo	Media	Mediana	Máximo	Rango Inter Cuartil
Death.per.1m	0	8.125	0.89	188.31	5.36
Cases.per.1m	0	146.361	32.71	1700.8	157.515
Weakly.Tests.per.1m	0	2009.361	1012	31172	2495
Stringency.Index	0	79.463	85.71	100	23.81
Population	352721	116239918.865	31989256	1392730000	66731596.5
UHC.sc.index	40	75.331	78	89	12
Population65	1.1	13.154	13.9	27.6	12
Doctors	4.7	248.462	236.7	751.9	210.6
Nurses	14.5	610.217	532.4	1823	775.3
Beds	40	381.625	290	1340	320
Government.expend	16	1863.193	1244	8078	2698



Por último, y con respecto a los Boxplot realizados, es posible observar que muchas de las gráficas no ofrecen información valiosa dado que o no vale la pena por ser variables dummy o binarias o por el contrario poseen tantos datos atípicos que se hace imposible obtener información a simple vista sobre estas variables. Como un ejemplo de este problema, es posible redimirse a los boxplot realizados para la variable "Population", "Doctors", "Nurses" "Beds" y "Government.expend"

- Los cuatro momentos estadísticos (Media, Varianza, Asimetría y Curtosis):

- **Forma funcional de las variables:**

Table 10: Momentos estadísticos

	Media	Varianza	Asimetria	Curtosis
Death.per.1m	8.125	20.715	4.323	26.229
Cases.per.1m	146.361	247.89	3.013	14.356
Weakly.Tests.per.1m	2009.361	3325.175	5.044	37.685
Stringency.Index	79.463	19.393	-1.361	4.787
Population	116239918.865	296208575.947	3.866	16.638
UHC.sc.index	75.331	10.031	-1.345	4.836
Population65	13.154	6.471	0.072	1.927
Doctors	248.462	141.176	0.628	3.706
Nurses	610.217	473.603	0.734	2.706
Beds	381.625	281.94	1.524	5.352
Government.expend	1863.193	1739.234	1.131	3.989

Como fue mencionado con anterioridad en la sección de selección de variables, para el modelo a estimar consideramos dos razones específicas por las cuales aplicar una transformación a las variables: 1. por la naturalidad de los datos, no es coherente hablar de la cantidad de doctores en un país con 10 habitantes que con 100 habitantes; y 2. se desea disminuir la cantidad de datos atípicos en las muestras. Así pues, se realizaron las transformaciones logarítmicas para las variables población, doctores, enfermeras, camas de hospitales y gasto gubernamental, y se realizó un nuevo análisis descriptivo con dicha transformación aplicada.

```
pcovid$LogPopulation = log(pcovid$Population)
pcovid$LogDoctors = log(pcovid$Doctors)
pcovid$LogNurses = log(pcovid$Nurses)
pcovid$LogBeds = log(pcovid$Beds)
pcovid$LogGovernment.expend = log(pcovid$Government.expend)
```

Tras realizar esta transformación podemos revisar nuevamente las estadísticas y las gráficas anteriores:

- **Histogramas**

Histograma Comparativo

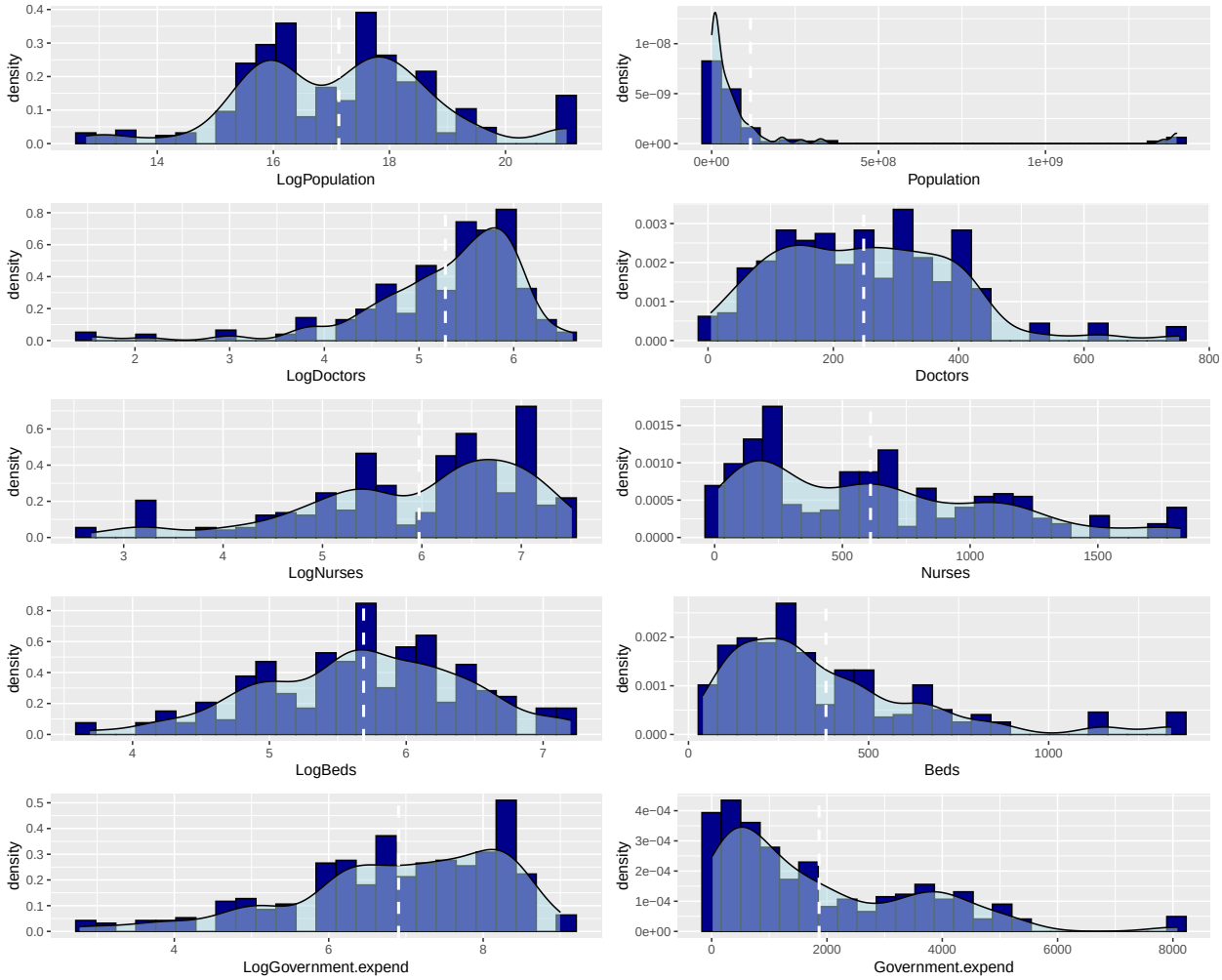
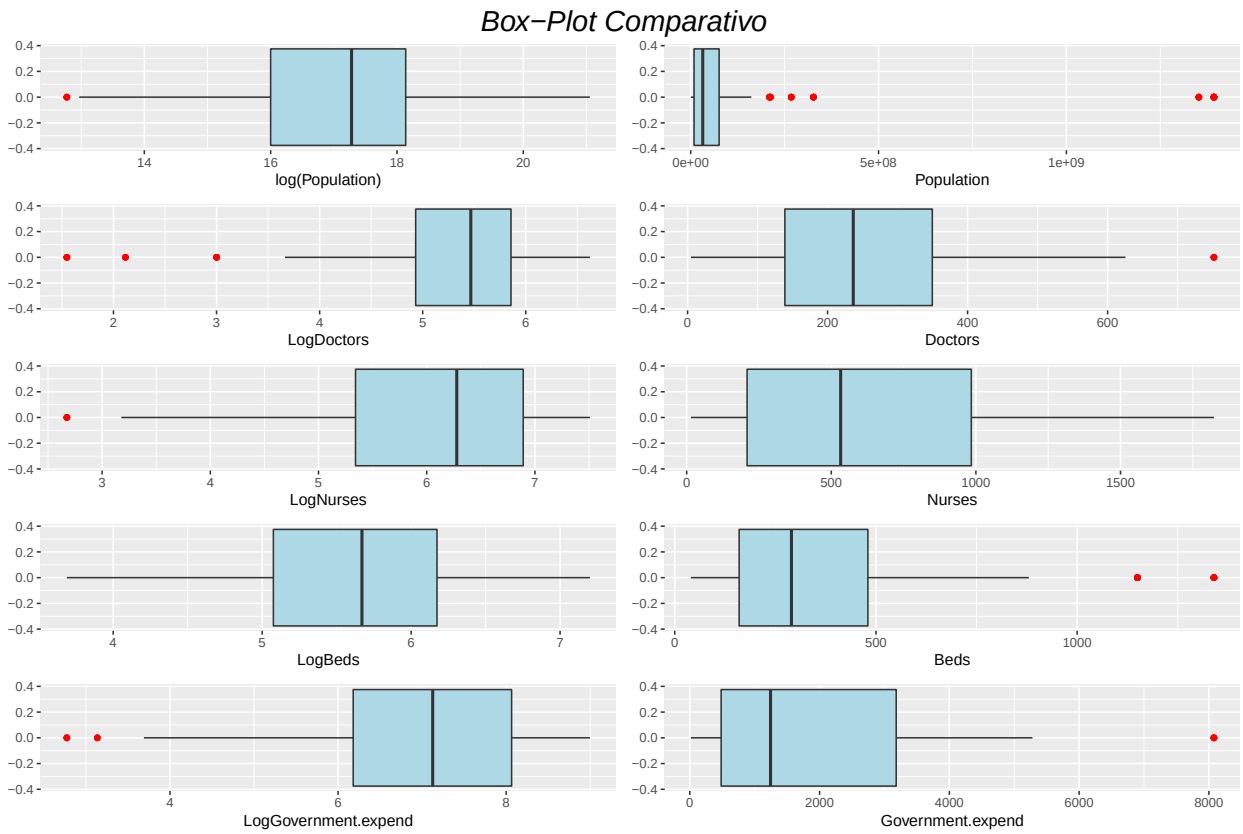


Table 11: Medidas Representativas Comparativos

	Mínimo	Media	Mediana	Máximo	Rango Inter Cuartil
Population	352721	116239918.865	31989256	1392730000	66731596.5
LogPopulation	12.773	17.128	17.281	21.055	2.138
Doctors	4.7	248.462	236.7	751.9	210.6
LogDoctors	1.548	5.279	5.467	6.623	0.925
Nurses	14.5	610.217	532.4	1823	775.3
LogNurses	2.674	5.971	6.277	7.508	1.55
Beds	40	381.625	290	1340	320
LogBeds	3.689	5.688	5.67	7.2	1.099
Government.expend	16	1863.193	1244	8078	2698
LogGovernment.expend	2.773	6.905	7.126	8.997	1.885



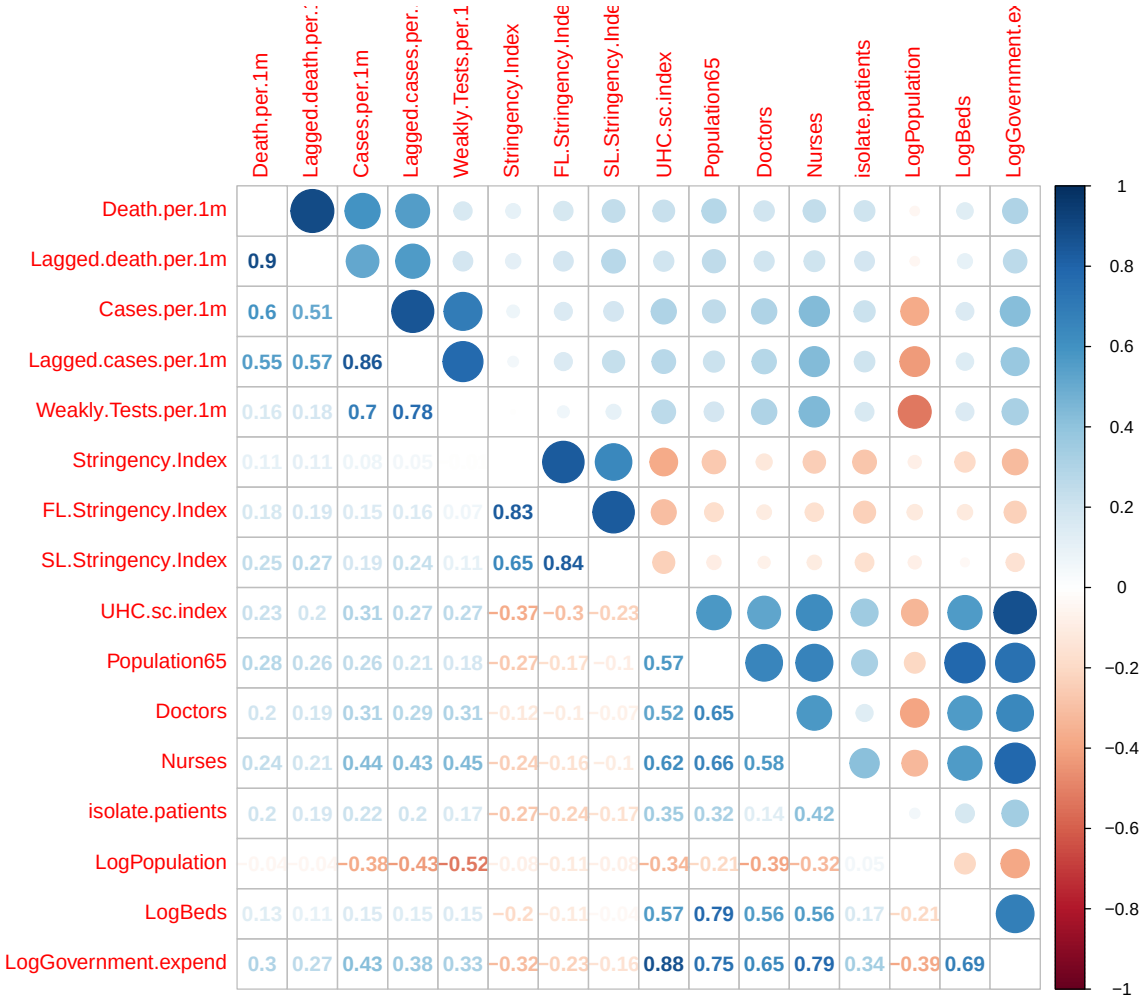
Como puede ser evidenciado en los histogramas comparativos, boxplot comparativos y la Tabla 11 sobre las medidas representativas comparativas entre las variables transformadas en primer lugar y las que no lo están después, algunas de las que tienen la transformación logarítmica presentan una disminución de datos atípicos como era de esperarse, lo cual hace que la forma distribución en el histograma luzca más asimétrica que lo que era antes; lo mismo sucede con su curtosis, la cual es más cercana a 3 (mesocúrtica) que lo que era antes. Ahora bien, no sucede lo mismo para todas las variables, puesto que, si se observa "Doctors" y "Nurses", es posible afirmar que ocurrió exactamente lo contrario a lo esperado: aumentaron la cantidad de datos atípicos y su asimetría se alejó drásticamente, con lo cual su distribución estaba distanciada de ser simétrica. Dados los resultados de las transformaciones logarítmicas en las variables, y comparando el análisis descriptivo, es posible afirmar que se cumplió con los dos objetivos anteriormente planteados y la mejor decisión para trabajar de manera adecuada con estos datos fue realizar la transformación logarítmica para todas las variables consideradas a excepción de las variables "Doctors" y "Nurses".

Table 12: Momentos estadísticos Comparativos

	Media	Varianza	Asimetria	Curtosis
Population	116239918.865	296208575.947	3.866	16.638
LogPopulation	17.128	1.66	0.093	3.409
Doctors	248.462	141.176	0.628	3.706
LogDoctors	5.279	0.833	-1.813	7.775
Nurses	610.217	473.603	0.734	2.706
LogNurses	5.971	1.112	-0.924	3.348
Beds	381.625	281.94	1.524	5.352
LogBeds	5.688	0.739	-0.193	2.709
Government.expend	1863.193	1739.234	1.131	3.989
LogGovernment.expend	6.905	1.346	-0.863	3.331

- **Correlación entre las variables:**

Con aras de presentar una gráfica legible y que proporcionará la información de una manera adecuada para la correlación entre cada una de las 16 variables aquí presentadas, se decidió incluir una gráfica de calor de correlación, pues la función cor en donde se muestran múltiples nubes de puntos y el coeficiente de correlación de Pearson, al ser una cantidad de variables considerable no se podía extraer información a simple vista de dicha nube de puntos. Por el contrario, esta gráfica de color presentada a continuación presenta también el coeficiente de correlación de Pearson, pero con la diferencia que no mostrará una nube de puntos sino unos círculos de color azul y rojo, donde a mayor tamaño del círculo será más acentuada la correlación entre ese par de variables; teniendo en cuenta también que el color azul corresponde a una correlación positiva y el color rojo a una correlación negativa.



Como fue argumentado anteriormente, es claro que los rezagos de las variables pueden presentar altos valores de correlación con respecto a sus mismas variables originales sin rezago o incluso con otras que pueden hablar de algo similar implícitamente, luego se omitirá el análisis de correlación para estos casos. Por otro lado, resulta bastante interesante comprobar por medio de la correlación y sin estudios o análisis de mayor profundidad, que una variable posee bastante influencia la una sobre la otra. por ejemplo, el gasto gubernamental en salud posee una alta correlación con el índice UHC que corresponde al índice de cobertura en un país (0.88), así como el número de doctores y enfermeras también posee un alto valor de correlación según Pearson con respecto al índice UHC. De la misma manera es posible extraer correlaciones negativas esperadas, por ejemplo, entre la variable "LogPopulation" y "Doctors" o "Nurses" con valores de correlación de Pearson de -0.39 y -0.32 respectivamente, lo cual tiene bastante sentido.

2) Estimación del mejor modelo

Tras definir las variables que se consideran importantes para los fines de nuestro análisis, revisar sus medidas representativas y ver cómo se comportan estos datos antes de transformarlos y después de trasnformar unas pocas variables que se consideraron pertinentes para optimizar la regresión, corresponde ahora definir qué modelo será preferido para ello.

Table 13: Comparación de Modelos

	Dependent variable:			
	Death.per.1m			
	MCOC	Efectos Fijos	Primeras Diferencias	Efectos Aleatorios
Lagged.death.per.1m	1.010***	0.790***	0.700***	0.978***
	(0.050)	(0.056)	(0.075)	(0.051)
Cases.per.1m	0.040***	0.028***	0.031***	0.038***
	(0.005)	(0.005)	(0.006)	(0.005)
Lagged.cases.per.1m	−0.006	0.032***	0.023***	−0.002
	(0.006)	(0.009)	(0.008)	(0.006)
Weakly.Tests.per.1m	−0.002***	−0.002***	−0.002***	−0.002***
	(0.0003)	(0.001)	(0.0005)	(0.0003)
Stringency.Index	−0.001	0.050	0.046	0.010
	(0.057)	(0.062)	(0.067)	(0.057)
FL.Stringency.Index	0.001	−0.020	−0.005	−0.004
	(0.057)	(0.051)	(0.042)	(0.055)
SL.Stringency.Index	0.013	0.016	0.051	0.019
	(0.035)	(0.033)	(0.035)	(0.034)
LogPopulation	0.093			0.235
	(0.465)			(0.518)
UHC.sc.index	0.184			0.209
	(0.133)			(0.151)
Population65	0.187			0.259
	(0.185)			(0.212)
Doctors	−0.002			−0.002
	(0.005)			(0.006)
Nurses	0.002			0.001
	(0.002)			(0.002)
LogBeds	0.058			−0.124
	(1.304)			(1.501)
LogGovernment.expend	−1.716			−1.782
	(1.390)			(1.574)
isolate.patients	0.191			0.162
	(1.430)			(1.594)
Constant	−7.663		−0.443	−11.948
	(12.801)		(0.925)	(14.102)
Observations	249	249	192	249
R^2	0.862	0.828	0.534	0.849

- **Prueba para MCOC**

La regla de decisión será una prueba de **Breusch-Pagan** para detectar Homoscedasticidad donde las hipótesis a contrastar son:

$H_0 : Var(u) = \sigma_u^2$

$H_1 : Var(u) \neq \sigma_u^2$

Y si la prueba arroja No Rechazo de la Hipótesis Nula (p-value>0.05) se concluye que el modelo bajo Mínimos Cuadrados Combinados (MCOC) es un buen candidato en términos de eficacia.

```
    studentized Breusch-Pagan test

data:  pooled
BP = 81.995, df = 15, p-value = 3.01e-11
```

Como se puede observar, el p-value asociado a la prueba es menor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de RH_0 y se concluye que MCOC no es una buena selección estadísticamente hablando. Para confirmar esta intuición se debe revisar ahora si hay efectos inobservados en el modelo y de ser así, buscar cómo eliminarlos.

- **Efectos Temporales**

```
    Lagrange Multiplier Test - time effects (Breusch-Pagan) for unbalanced
    panels

data:  Death.per.1m ~ Lagged.death.per.1m + Cases.per.1m + Lagged.cases.per.1m +  ...
chisq = 0.024007, df = 1, p-value = 0.8769
alternative hypothesis: significant effects
```

El p-value asociado a la prueba es mayor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de NRH_0 y se concluye que en el Panel de datos **no** hay efectos temporales significativos.

- **Efectos Individuales**

```
    Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels

data:  Death.per.1m ~ Lagged.death.per.1m + Cases.per.1m + Lagged.cases.per.1m +  ...
chisq = 1.3413, df = 1, p-value = 0.2468
alternative hypothesis: significant effects
```

El p-value asociado a la prueba es mayor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de NRH_0 y se concluye que en el Panel de datos **no** hay efectos individuales significativos.

- **Ambos Efectos**

```
    Lagrange Multiplier Test - two-ways effects (Breusch-Pagan) for
    unbalanced panels

data:  Death.per.1m ~ Lagged.death.per.1m + Cases.per.1m + Lagged.cases.per.1m +  ...
chisq = 1.3653, df = 2, p-value = 0.5053
alternative hypothesis: significant effects
```

El p-value asociado a la prueba es mayor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de NRH_0 y se concluye, en concordancia con las anteriores dos pruebas, que en el Panel de datos **no** hay efectos inobservados significativos.

- **Prueba para Efectos Fijos vs. MCOC**

A pesar de no encontrar dichos efectos inobservables, cabe revisar si el modelo podría estar, estadísticamente hablando, mejor estimado bajo el modelo de Efectos Fijos que bajo MCOC. La regla de decisión será una prueba de **Multiplicadores de Lagrange por Breusch-Pagan** para detectar si el modelo el modelo Fixed sería mejor a MCOC considerando su utilidad eliminando efectos inobservables por insignificantes que hayan resultado. Las hipótesis a contrastar son respecto a la la varianza de los efectos fijos (α_i) dentro del término de error compuesto (u_{it}):

$H_0 : Var(a_i) = 0$

$H_1 : Var(a_i) \neq 0$

```
    Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels

data:  Death.per.1m ~ Lagged.death.per.1m + Cases.per.1m + Lagged.cases.per.1m +  ...
chisq = 1.3413, df = 1, p-value = 0.2468
alternative hypothesis: significant effects
```

El p-value asociado a la prueba es mayor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de NRH_0 y se concluye que los efectos fijos (α_i) dentro del termino de error no son significativos, por lo que se elige MCOC.

- **Prueba para Efectos Aleatorios vs. MCOC**

La misma prueba del caso anterior, aplicada ahora para contrastar entre Efectos Aleatorios y MCOC.

```
    Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels

data:  Death.per.1m ~ Lagged.death.per.1m + Cases.per.1m + Lagged.cases.per.1m +  ...
chisq = 1.3413, df = 1, p-value = 0.2468
alternative hypothesis: significant effects
```

Nuevamente, el p-value asociado a la prueba es mayor al nivel de significancia $\alpha = 0.05$, con lo que se cae en zona de NRH_0 y se concluye que los efectos ideosincráticos dentro del termino de error no son significativos, por lo que se elige MCOC.

En conclusión, no hay motivos para pensar que se deba realizar más pruebas y se acepta que la estimación bajo MCOC es suficiente brindando consistencia y mayor eficacia a la regresión.

3) ¿En cuánto cambia el número de muertes cuando el número de contagios aumenta en 100 personas?

Teniendo en cuenta el modelo seleccionado (Pooled) bajo MCOC:

Table 14: Covid-19

	Dependent variable:
	Death.per.1m
	MCOC
Lagged.death.per.1m	1.010***
	(0.050)
Cases.per.1m	0.040***
	(0.005)
Lagged.cases.per.1m	−0.006
	(0.006)
Weakly.Tests.per.1m	−0.002***
	(0.0003)
Stringency.Index	−0.001
	(0.057)
FL.Stringency.Index	0.001
	(0.057)
SL.Stringency.Index	0.013
	(0.035)
LogPopulation	0.093
	(0.465)
UHC.sc.index	0.184
	(0.133)
Population65	0.187
	(0.185)
Doctors	−0.002
	(0.005)
Nurses	0.002
	(0.002)
LogBeds	0.058
	(1.304)
LogGovernment.expend	−1.716
	(1.390)
isolate.patients	0.191
	(1.430)
Constant	−7.663
	(12.801)
Observations	249
R^2	0.862
Adjusted R^2	0.853
F Statistic	97.290*** (df = 15; 233) (p = 0.000)

Para determinar cuánto se estima que cambien la muertes por Covid-19 cuando el número de contagios aumenta en 100, manteniendo todo lo demás constante (ceteris paribus), el proceso a realizar sería, primero cuantificar nuestra variable **Cases.per.1m** de manera que no sea relativo a un millon de personas, sino cuando sea relativo a solo 100 personas. De esa manera la variable **Cases.per.1m** se considerará en un 0.0001 de su valor.

Ahora, para revisar el cambio se debe multiplicar el β estimado (0.04) por 100 (aumento de 100 en la cantidad de contagios). De manera que:

$DeathPer1m = (0.04 CasesPer1m) * 0.0001$

$DeathPer1m = (0.04 * 100) * 0.0001$

$DeathPer1m = 0.0004 * 1.000.000$

$DeathPer1m = 400$

Table 15: Displaying records 1 - 10

Country	Population65	DeathPer1m	FinalWeek	FirstDeath
Japan	27.6	0.00	9	2020-02-13
Italy	22.8	0.45	8	2020-02-22
Portugal	22.0	2.14	4	2020-03-17
Finland	21.7	1.09	4	2020-03-21
Greece	21.7	0.37	5	2020-03-12
Germany	21.5	0.12	5	2020-03-09
Bulgaria	21.0	0.28	5	2020-03-12
Croatia	20.4	1.22	3	2020-03-25
Sweden	20.1	1.67	5	2020-03-15
France	20.0	0.00	9	2020-02-15

4) ¿Existen diferencias en la tasa de fatalidad entre países con un alto porcentaje de población vieja?¿A qué se debe dicha diferencia?

Para iniciar, veamos algunos países que tengan alto porcentaje de su población que sea mayor a 65 años con su respectiva tasa de fatalidad.

Como puede ser evidenciado en la Tabla 14, se escogieron los países con el mayor porcentaje de población de adultos con 65 o más años y se organizaron con orden un descendente en la segunda columna, junto a una columna con sus respectivas tasas de mortalidad. Se pensaría según la teoría y con cierta especulación que habría una clara relación entre el porcentaje de población de adultos con 65 o más años y la tasa de mortalidad, esto porque en principio se pensaría que éstos no poseen ya las mismas defensas en su cuerpo que una persona de menos edad, con lo cual, se esperaría que esta población incidiera de una manera más significativa o directa con la tasa de mortalidad. Sin embargo los datos estadísticos encontrados no especifican esta relación y por el contrario, parece que no existiera ninguna relación si se analizan estos países con mayor porcentaje de este tipo de población según los datos proporcionados. Además, la información proporcionada por estos datos son respaldados con los resultados arrojados por el coeficiente de correlación de Pearson que anteriormente fue presentado, donde si bien existe una correlación, no es un valor tan grande como se esperaría (0.28).

Luego de múltiples ideas abordadas en el grupo de trabajo, se llegó a la conclusión que esta variable del porcentaje de población de adultos con 65 o más años no es más incidente en la tasa de mortalidad porque depende de manera simultánea de si ésta población de adultos mayores sufría antes de alguna enfermedad o si por el contrario se encontraba totalmente sano y con las defensas suficientes para combatir de manera efectiva el virus COVID-19.

Por último, y con respecto a los siguientes puntos a abordar, las tablas 14, 15 y 16 fueron realizadas con ayuda del lenguaje SQL para una mayor facilidad, visualización y tratamiento de consulta.

5) ¿La incidencia de la tasa de contagio sobre la tasa de fatalidad depende del porcentaje de población mayor a 65 años?

Podemos tener una aproximación con los anteriores resultados, utilizando aquellos países con el mayor porcentaje de población mayor a 65 años observando ahora la tasa de contagios asociada a ellos. Esto con el fin de tener una aproximación a la hipótesis que se busca sostener viendo si los datos la respaldan. Para que se encuentre dicho respaldo en los datos, se esperaría que en aquellos países donde la cantidad de contagios es mayor y a su vez, el porcentaje de la población mayor a 65 años es mayor, también se encontraría mayor cantidad de muertes por el virus. La preposición lógica se podría expresar así:

$Si\ (\uparrow CasesPer1m \quad y \quad \uparrow Population65) \longrightarrow \quad \uparrow DeathPer1m$

Table 16: Displaying records 1 - 10

Country	Population65	CasesPer1m	DeathPer1m
Estonia	19.6	283.67	3.03
Belgium	18.8	202.57	6.12
Norway	17.0	186.75	1.13
Austria	19.0	167.64	0.45
Portugal	22.0	156.75	2.14
Czech Republic	19.4	156.54	1.41
Croatia	20.4	118.64	1.22
Finland	21.7	91.38	1.09
New Zealand	15.7	81.80	0.00
Slovenia	19.6	80.52	0.00

En los datos de la ***Tabla 15*** es notable que dicha condicional se viola. Pues, por ejemplo, Estonia, el país muestreado con mayor cantidad de contagios por cada millon de personas, con 19.6% de su población mayor a 65 años; tiene **menor** cantidad de muertes por millón de personas frente a Bélgica, quien tiene menos cantidad de contagios que Estonia (29% menos), menor porcentaje de su población mayor a 65 años (0.8% menos) **pero** su cantidad de muertes por coronavirus es mayor a la de Estonia, una diferencia mayor al 100% de hecho. Así se podrían revisar algunos contraejemplos más, pero ciertamente con uno basta para derrumbar la hipótesis.

Finalmente, solo por mencionar otra manera posible de revisar dicha relación podría hacerse mediante la estimación bajo el método de Variables Instrumentales donde:

$DeathPer1m = \beta_0 + \beta_1 CasesPer1m + u$

Y además:

$CasesPer1m = \Gamma_0 + \Gamma_1 Population65 + v$

6) Comparación basada en el porcentaje de población mayor a 65 años (15% vs. 25%)

Para revisar la estimación podemos emplear el modelo seleccionado en el primer punto (MCOC) considerando un cambio en los estimadores solicitados manteniendo todo lo demás constante, de manera que:

- **Poblation65 = 15%**

$DeathPer1m = 0.04 CasesPer1m + 0.187 Population65$

$DeathPer1m = (0.04 * 1) + (0.187 * 15)$

$DeathPer1m = 0.04 + 2805 = 112.2$

De manera que, en un país con 15% de su población mayor a 65 años, al aumentar en 1 caso de contagio por cada millon de habitantes, se estima que aumentará 112.2 la cantidad de muertes por cada millón de habitantes.

- **Poblation65 = 25%**

$DeathPer1m = 0.04 CasesPer1m + 0.187 Population65$

$DeathPer1m = (0.04 * 1) + (0.187 * 25)$

$DeathPer1m = 0.04 + 4675 = 187$

De manera que, en un país con 25% de su población mayor a 65 años, al aumentar en 1 caso de contagio por cada millon de habitantes, se estima que aumentará 187 la cantidad de muertes por cada millón de habitantes.