Tarea 2

Fecha de entrega: 21-oct

Nota: Deberán subir a *Canvas* un archivo de texto con sus respuestas. Pueden utilizar el formato de su preferencia (e.g. LaTeX, Word u hojas escritas a mano y escaneadas). Además, deberán subir otro archivo que genere todos sus resultados de las preguntas prácticas. Puede ser un R-script, Do-File o script de algún otro software similar.

En esta tarea continuaremos explorando la relación de distintas variables asociadas al COVID-19 y algunas características de una muestra de países. En *Canvas* encontrarás la base T2_covid (disponible en formato .dta y .csv). Dicha base contiene datos sobre la pandemia de COVID-19,[1] así como datos sobre ingreso, salud y algunas características demográficas para un conjunto de 89 países. La base únicamente contiene 89 países debido a que para dichos países contamos con información completa para las variables que analizaremos en esta Tarea. La base de datos contiene información recopilada por el sitio "Our World in Data" proveniente de instituciones como el Centro Europeo para el Control y Prevención de Enfermedades, reportes de autoridades sanitarias a nivel nacional, Banco Mundial, División de Población de las Naciones Unidas, OCDE y el PNUD. Al final del documento encontrarás el Cuadro 1, el cual contiene la descripción de cada una de las variables que integran la base T2_covid y el Cuadro 2 en el que se indican las claves para cada continente.

1. Iniciemos explorando la base de datos. Elabora una tabla con estadísticas descriptivas básicas para todas las variables de la base de datos (exceptuando la variable *gdp_pc_er*). Incluye en tu tabla número de observaciones, media, desviación estándar, mínimo y máximo. [Tip de R: utiliza el comando `summary` o `stargazer`.]

2. Al final de la *Tarea 1* estudiamos la relación entre el número de *personas a las que se le ha realizado pruebas* y el *total de casos confirmados*, ambos como porcentaje de la población total. En esta ocasión nos interesará la relación entre el *total de pruebas realizadas por millón de habitantes* y el *total de casos confirmados por millón de habitantes*. La primera de estas variables la tendrás que generar y en adelante la llamaremos (*tests_per_mil*)[2]. [Tip de R: Puedes crear variables con el comando `mutate` de la librería `dplyr`.]

   (a) Para explorar la relación entre estas variables empieza realizando un diagrama de dispersión (scatterplot) entre ambas variables. Ubica en el eje X *tests_per_mil* y en el eje Y *confirmed_per_mil*. Agrega a dicha gráfica el resultado que obtendrías de llevar a cabo una estimación de MCO simple (sin controles) entre ambas variables. Reporta en formato de ecuación la estimación del modelo y reporta la interpretación del coeficiente de *tests_per_mil*. [Tip de R: Puedes realizar *scatterplots*

[1] Fecha de corte: 14 de septiembre de 2020.

[2] $tests\_per\_mil = \frac{(tests\_performed)(confirmed\_per\_mil)}{confirmed}$.

con `geom_point` y agregar el ajuste de MCO con `geom_smooth(method="lm")` de la librería `ggplot2`. Para realizar estimaciones por MCO (simples y múltiples) utiliza el comando `lm`.]

(b) Repite todas las indicaciones del inciso (a) pero ahora utilizando en el eje X el logaritmo natural de *tests_per_mil*.

(c) Una vez más repite las indicaciones del inciso (a), pero ahora utiliza el logaritmo natural para AMBAS variables. Recuerda producir la gráfica, agregar la estimación de MCO y reportar la interpretación del coeficiente relevante.

(d) (*Puntos extras*) Considerando el modelo del inciso (c), estima regresiones cuantílicas considerando el primer y tercer cuartil y la mediana. Reporta los resultados de dichas regresiones en formato de ecuación, grafica en un diagrama de dispersión las tres regresiones cuantílicas y la regresión de MCO. ¿Qué puedes concluir de esta gráfica? [Tip de R: Puedes estimar regresiones cuantílicas con el comando `rq` de la librería `quantreg`].

3. Ahora analizaremos la relación entre el *PIB per cápita* y el total de *casos confirmados por cada millón de habitantes*. Como recordarán se ha mencionado en distintas ocasiones que los países de mayores ingresos han tenido mayor número de contagios y que en general el Covid es una enfermedad parece ser una enfermedad con una incidencia mayor para ingresos altos. Para explorar esta creencia realiza lo siguiente:

(a) Realiza un diagrama de dispersión en el que en el eje X esté el logaritmo natural de *gdp_pc* y en el eje Y esté el logaritmo natural de *confirmed_per_mil*. Agrega a dicha gráfica el resultado que obtendrías de llevar a cabo una estimación de MCO simple entre ambas variables. Reporta en formato de ecuación la estimación del modelo y reporta la interpretación del coeficiente del logaritmo de *gdp_pc*.

(b) **Errores de medición**. En la base de datos encontrarás la variable *gdp_pc_er*, la cual se define como $gdp\_pc\_er_i = gdp\_pc_i + 1000\nu_i$, donde ν_i es un error aleatorio con $\nu_i \sim N(0,1)$ y $Cov(gdp\_pc, \nu) = 0$. Considerando dicha variable realiza lo siguiente:

(i) Sin realizar ningún cálculo, explica qué crees que sucedería con el coeficiente β_1 del inciso (a) si utilizáramos *gdp_pc_er* en lugar de *gdp_pc* para realizar la estimación anterior.

(ii) Estima nuevamente el modelo del inciso (a) pero utilizando como variable explicativa *gdp_pc_er*. Compara tus resultados con los del inciso (a). ¿Cómo llamamos al sesgo obtenido? ¿Se confirma lo que hubieras esperado que sucediera de acuerdo a la teoría vista en clase?

(c) Considerando tu respuesta del inciso (a) y la relación que observaste en la pregunta 2, ¿consideras que existe un problema de sesgo por variables omitidas en tu estimación de la relación entre *gdp_pc* y *confirmed_per_mil*? En caso de que

consideres que existe dicho problema, señala cuál crees que sea la dirección del sesgo y explica tu respuesta.

(d) Ahora realiza lo siguiente:

(I) Realiza una regresión simple en la que el logaritmo natural de *gdp_pc* sea la variable dependiente y el logaritmo natural de *tests_per_mil* sea la única variable explicativa. Reporta dicha estimación en formato de ecuación.

(II) Genera una nueva variable que contenga los residuales de la regresión que calculaste en el paso anterior. Ahora, realiza una regresión con el logaritmo natural de *confirmed_per_mil* como variable dependiente y los residuales que estimaste como variable explicativa. Reporta tu estimación en formato de ecuación y lleva a cabo un diagrama de dispersión con las observaciones y la estimación de MCO. [Tip de R: Los residuales de un modelo estimado por MCO los puedes encontrar dentro de la estructura de datos en que se guarda la estimación de la siguiente forma: `mod$residuals`, donde `mod` es el nombre que le asignaste a la variable que guarda tu estimación de MCO.]

Lo que acabas de realizar es conocido como *partial-out*. Este procedimiento consiste en obtener el coeficiente de cierta variable en una regresión múltiple, utilizando una regresión simple.

(e) Para ver lo que esto quiere decir, estima el siguiente modelo:

$$\ln(confirmed\_per\_mil_i) = \beta_0 + \beta_1 \ln(gdp\_pc_i) + \beta_2 \ln(tests\_per\_mil_i) + U_i$$

Reporta tu estimación en formato de ecuación. ¿Qué notas del coeficiente que obtuviste en este inciso para $\ln(gdp\_pc)$ en comparación con el que obtuviste en la regresión simple del inciso anterior? Intuitivamente, ¿por qué crees que obtuviste este resultado?

4. Ahora analizaremos las muertes por COVID-19 y su relación con otras condiciones médicas como el sobrepeso, la diabetes y enfermedades cardiovasculares. Para esto, deberás realizar las estimaciones que se incluyen en la Tabla 3, que se encuentra al final del documento. Ojo: las líneas horizontales en algunas variables (—) significan que NO debes incluir esta variable en la estimación de dicha columna. Utiliza errores heterocedásticos y agrega los asteriscos que indiquen nivel de significancia: * 10 %, ** 5 % y *** 1 %. [Tip de R: Para estimar errores heterocedásticos, utiliza el comando `vcovHC` de la librería `sandwich`. En dicho comando deberás indicar el nombre con el que guardaste tu estimación `lm` y errores del tipo `HC1` (i.e. `vcovHC(nombre_modelo,type='HC1')`). Para crear tablas con resultados de regresiones puedes utilizar el comando `stargazer` de la librería homónima.]
Nota: Las siguientes variables no aparecen en la base de datos que se te proporcionó, por lo que debes de generarlas en tu base:

- cfr=$\frac{deaths_i}{confirmed_i}$, tasa de fatalidad a la fecha de corte.
- *eur* es una variable dummy que toma el valor de 1 si el país se localiza en Europa.
- *asia* es una variable dummy que toma el valor de 1 si el país se localiza en Asia.
- *nam* es una variable dummy que toma el valor de 1 si el país se localiza en América del Norte.
- *std_hdi* es la variable *hdi* estandarizada.
- *hm_cfr* es una variable dummy que toma el valor de 1 si la tasa de fatalidad acumulada del país es mayor a 0.019. Dicho valor corresponde a la mediana poblacional de *cfr*.

5. Considerando tus respuestas de la pregunta anterior, interpreta los siguientes coeficientes. (Nota: lleva a cabo la interpretación del valor estimado de los coeficientes a pesar de que algunos resulten no significativos.):

   (a) *overwgh_prev* en la columna (1).

   (b) $\ln(gdp\_pc)$ en la columna (1).

   (c) $\ln(tests\_per\_mil)$ en la columna (2).

   (d) *eur* en la columna (2).

   (e) *median_age* en la columna (2)

   (f) *diab_prev* en la columna (3).

   (g) $\ln(tests\_per\_mil)$ en la columna (3).

   (h) std_hdi^2 en la columna (3).

   (i) *overwgh_prev* en la columna (4).

   (j) $\ln(hosp\_beds\_per\_thou)$ en la columna (4).

   (k) *nam* en la columna (4).

6. Utilizando tus resultados de la Tabla 3, contesta las siguientes preguntas:

   (a) Utilizando los resultados de la columna (2) y los datos de México, ¿cuál sería la predicción del número de muertes por millón de habitantes? ¿Cómo se compara esta predicción con el dato real?

   (b) Utilizando la columna (3): calcula un intervalo de confianza de 95 % para el cambio de *cfr* que estaría asociado a un incremento de 0.1 desviaciones estándar en *hdi* para el caso del país promedio.

   (c) Imagina que un gobierno hubiera podido haber dedicado recursos para incrementar el número de pruebas de COVID-19 por cada millón de habitantes en 15 %, pero esto hubiera disminuido los recursos de un programa para la prevención del

sobrepeso y la diabetes que hubiera provocado un aumento de 1.5 puntos porcentuales en la incidencia de sobrepeso (*overwgh_prev*) y un aumento de 2 puntos porcentuales en la incidencia de diabetes (*diab_prev*). Utilizando la columna (4), ¿qué cambio en la predicción de la probabilidad de tener un *cfr* por encima de la mediana hubiera tenido? Evalua la significancia estadística de este cambio usando el valor-p. [Tip de R: Puedes realizar pruebas de hipótesis unidimensionales y multidimensionales con el comando `linearHypothesis` de la librería `car`. Para esto debes indicar `linearHypothesis(modelo,matriz,white.adjust='hc1')`, donde `modelo` es el modelo a utilizar, `matriz` es la matriz o vector de restricciones y `white.adjust='hc1'` le indica que realice la prueba considerando heterocedasticidad.]

7. Ahora llevaremos a cabo modificaciones de las especificaciones de la Tabla 3 para contestar las siguientes preguntas:

   (a) **Diferencias entre Europa y Asia.**

   (i) Utilizando los resultados de la columna (2) quisiéramos hacer una comparación entre el país promedio de Europa y de Asia (i.e. creando un *país promedio europeo* y un *país promedio asiático*). Utilizando los valores promedio de las variables explicativas para Europa y Asia, calcula para cada continente una predicción de muertes por millón de habitantes. Evalua si esta diferencia es significativa al 5 %.

   (ii) Repite el inciso anterior, pero volviendo a estimar la especificación (2), pero solo usando países de Europa y Asia. ¿Cambia tu conclusión?

   (iii) Lleva a cabo 200 simulaciones en las cuales realices lo siguiente: (primero) elige al azar un país de Europa o de Asia; (segundo) utilizando los modelos de los incisos anteriores realiza una predicción del número de muertes por cada millón de habitantes para cada país; (tercero) compara las predicciones del segundo paso con el número de muertes por cada millón de habitantes observado en dicho país (i.e. estás calculando residuales); (cuarto) grafica dos histogramas para mostrar la distribución de los residuales calculados en el punto anterior e indica la media en cada uno de tus histogramas [Tip de R: La función `sample_n` de la librería `dplyr` te permite extraer de forma aleatoria observaciones de un data frame].

   (iv) Utilizando la evidencia anterior, ¿cuál de los dos modelos consideras que es mejor para llevar a cabo la predicción de muertes por cada millón de habitantes para un país de Europa o Asia? Da una intuición de por qué crees que obtienes este resultado en menos de 200 palabras.

   (b) **Comparación entre distintos continentes de la asociación entre la mediana de la edad y las muertes por millón de habitantes.**

(i) Modifica la especificación (2) para no incluir NINGUN control, excepto dummies de continentes y *median_age*. Utilizando las dummies adecuadas muestra qué especificación generaría rectas paralelas en la asociación entre *deaths_per_mil* y *median_age*. Cada recta debe de corresponder a cada continente mencionado en el *Cuadro 2*. Muestra la gráfica y reporta los resultados de esta estimación en la primera columna de una nueva tabla que siga el mismo formato que la *Tabla 3*.

(ii) Modifica la especificación del inciso anterior para evaluar si existe evidencia estadística para rechazar la hipótesis de que la asociación entre *median_age* y muertes por millón de habitantes es paralela para todos los continentes. Agrega los resultados de la especificación que utilizaste como segunda columna de tu nueva tabla. Indica claramente la prueba de hipótesis que estas evaluando y tu conclusión.

(iii) Modifica la especificación anterior para evaluar si la asociación entre *median_age* y muertes por millón de habitantes es igual para Norteamerica que para el resto de los continentes. Agrega los resultados de la especificación que utilizaste como tercera columna de tu nueva tabla y haz una gráfica para ilustrar tu resultado. Indica claramente la prueba de hipótesis que estas evaluando y tu conclusión.

Descripción de la base de datos:

Variable	Descripción
iso_code	Combinación de tres letras asignada a cada país.
country	Nombre del país
continent	Continente en el que se ubica el país
confirmed	Total de casos confirmados (Fecha de corte).
confirmed_per_mil	Total de casos confirmados por millón de hab. (Fecha de corte).
deaths	Total de muertes (Fecha de corte).
deaths_per_mil	Total de muertes por millón de hab. (Fecha de corte).
tests_performed	Total de pruebas realizadas (Fecha de corte).
gdp_pc	PIB per cápita, PPA ($ dólares internacionales 2017) (Año 2019).
median_age	Mediana de edad (Proyección ONU para 2020).
aged_65_older	% población con 65 años o más (Año 2019).
diab_prev	% población entre 20 y 79 años con diabetes tipo I o II (Año 2019).
cardio_dr	Número de muertes anuales por cada cien mil habitantes provocadas por enfermedades cardiovasculares (Año 2017).
hosp_beds_per_thou	Camas de hospital por cada mil habitantes (Año más reciente).
hdi	Índice de Desarrollo Humano (Año 2018).
overwgh_prev	% población adulta con sobrepeso (Año 2016).
gdp_pc_er	Es la variable *gdp_pc* más un error aleatorio.

Cuadro 1: Descripción de Variables.

Código	Continente
AF	África
AS	Asia
EU	Europa
NAM	América del Norte
OC	Oceanía
SAM	América de Sur

Cuadro 2: Claves para cada continente.

	Variable dependiente:			
	ln(*deaths*)	*deaths_per_mil*	*cfr*	*hm_cfr*
	(1)	(2)	(3)	(4)
overwgh_prev				
ln(*cardio_dr*)				
diab_prev				
ln(*hosp_beds_per_thou*)	—	—	—	
ln(*tests_per_mil*)				
aged_65_older	—	—	—	
ln (*gdp_pc*)		—	—	
eur	—		—	—
asia	—		—	—
nam	—	—	—	
std_hdi	—	—		—
std_hdi^2	—	—		—
median_age	—			—
Constante				
Observaciones				
R^2				

$^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01
Errores heterocedasticos entre paréntesis

Cuadro 3: Modelos de regresión lineal