PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS SOCIALES

Especialidad de Economía

Trabajo Final Grupal

Curso: Laboratorio de R y Python

Nombres y Códigos:

- Daniela Soledad Ordoñez Ascencio (20193426)
- Luzcia Johanna Halanoca Argote (20190972)
- Rodrigo Franco Valle Arguedas (20182657)
- Lorenzo Gabriel Chiroque Camacho (20190022)

Profesor:

Roberto Carlos Mendoza Matos

Lima, Semestre 2022-2

FACULTAD DE CIENCIAS SOCIALES
ESPECIALIDAD DE ECONOMÍA
Laboratorio de R y Python
SEMESTRE 2022-2
Prof. Roberto Mendoza Matos

Grupo 6: Trabajo Final

---

1 Tables y regresiones

Este primer apartado se centrará en el trabajo de investigación *Economics Shocks and Civil Conflict: An Instrumental Variables Approach* de Edward Miguel, Shanker Satyanath y Ernest Sergenti, el cual busca evaluar la relación entre los indicadores económicos y el estallido de conflictos bélicos en África en la segunda mitad del siglo XX.

La base de datos utilizada en aquel trabajo, *mss_repdata.dta*, presenta una estructura de datos panel, y contiene variables relacionadas al PBI, la población, las lluvias, producción agrícola, conflictos civiles, indicadores de democracia, entre otros, para cada país africano entre los años 1981-1999.

1.1 Tabla de estadísticos descriptivos

En primer lugar, se realizó una tabla de estadísticas (media, desviación estándar y total de observaciones) de las variables: $NDVU\_g$ (Tasa de variación del índice de vegetación), tot_100 (términos de intercambio), $trade\_pGDP$ (porcentaje de las exportaciones respecto al PBI), pop_den_rur (Densidad poblacional rural), $land\_crop$ (porcentaje de tierra cultivable en uso), va_agr (Valor agregado del sector agrícola respecto al PBI) y va_ind_manf (Valor agregado del sector manufacturero respecto PBI). Para ello, se utilizaron los lenguajes de programación R y Python.

- Para realizar la tabla de estadísticos descriptivos **en R**, en primer lugar, se instaló el paquete librarian que permitiría trabajar posteriormente según lo solicitado. Posterior a ello, se creó una tabla que incluía las siete variables indicadas, de manera que después se pueda referir en conjunto a dichas variables como "table1". Sin embargo, fue necesario, además, convertir aquella tabla a Dataframe para que esa librería pueda reconocer aquellos datos y transformarlo según las funciones que se indiquen. Luego, se describieron los nombres de cada variable a través de "covariate.label", y se eliminó la columna de mínimos y máximos mediante "$min.max = F$". De ahí, se construyó una función llamada "list_vars" para ordenar la lista de las variables que serían colocadas en la tabla. Por último, se colocó el comando *stargazer* indicando las características de la tabla (las variables, la lista de estadísticos requeridos, una nota, etc.) para así poder obtener el código en Latex.

Table 1: Estadísticos descriptivos

Statistic	Mean	St. Dev.	N
Tasa de variación del índice de vegetación ($NDVI\_g$)	0.01	0.09	646
Términos de intercambio (tot_100)	109.88	34.68	668
Porcentajes de las exportaciones respecto al PBI ($trade\_pGDP$)	64.25	34.29	698
Densidad poblacional rural (pop_den_rur)	324.82	193.09	720
Porcentaje de tierra cultivable en uso ($land\_crop$)	1.98	3.37	701
Valor agregado del sector agrícola respecto al PBI (va_agr)	32.18	15.17	702
Valor agregado del secctor manufacturero respecto al PBI (va_ind_manf)	11.12	6.26	669

Nota: Se tomó las variables indicadas.

- Ahora bien, en **Python** se trabajo primero instalando los paquetes necesarios para desarrollar modelos lineales. Específicamente para obtener estadísticos descriptivos. Se puede destacar el uso de **"statmodels", "linearmodels" y "pysout"** porque sirvieron para concretar la exportación de lo trabajado a **Latex**. **En segundo lugar**, se paso a cargar la base de datos para seleccionar las variables a estudiar específicamente en el presente trabajo. **En tercer lugar**, se obtuvo los estadísticos descriptivos de las variables seleccionadas, sin embargo, existían diferentes categorías como valores mínimo, máximo, y cortes por cuantiles u otros porcentajes (25, 50 y 75). Por ello, se seleccionó determinadas categorías solicitadas como mean (promedio), std (desviación estándar) y count (nro. observaciones. **En cuarto lugar**, se pasó a cambiar los nombres de las columnas, dado que las variables estaban codificadas por abreviaturas. Así, conforme a lo pedido, se cambio de nombre a NDVI_g, tot_100, trade_pGDP, pop_den_rur, land_crop, va_agr y va_ind_manf. **Finalmente**, se paso a exportar a latex la tabla concluida, con las variables solicitadas, en la disposición correcta y con los estadísticos.

Table 2: Estadísticos Descriptivos

	Promedio	Desv. Est.	N° Obs.
Tasa de variación del índice de vegetación	0.01	0.09	646
Términos de intercambio	109.88	34.68	668
Porcentaje de las exportaciones respecto al PBI	64.25	34.29	698
Densidad poblacional rural	324.82	193.09	720
Porcentaje de tierra cultivable en uso	1.98	3.37	701
Valor agregado del sector agrícola respecto PBI	32.18	15.17	702
Valor agregado del sector manufacturero respecto PBI	11.12	6.26	669

1.2 Regresión con MCO

- Al realizar la regresión **en R**, se

	Model 1	Model 2
Growth in rainfall, t	−0.029	−0.098
	(0.085)	(0.072)
Growth in rainfall, t-1	−0.120	−0.089
	(0.086)	(0.073)
Country fixed effects	yes	yes
Country-specific time trends	yes	yes
RMSE	0.442	0.372
R^2	0.003	0.003
Adj. R^2	0.000	0.001
Num. obs.	743	743

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 3: Dependent Variable: Economic Growth Rate, t

- En **Python**, se tiene el siguiente desarrollo: En **primer lugar**, se trabajo analizando la base de datos, mediante: obtención de estadísticos descriptivos e información por años, captura de variables omitidas, concatenación de ambas bases modificadas. De esta manera, en **segundo lugar**, se creó los efectos country_trend mediante la multiplicación de dummies con la variable temporal para capturar variables omitidas variantes en el tiempo por cada país. Además, se cambió los nombres de las variables dependientes que se solicitan para cada modelo. Así, en **tercer lugar** se trabajo simultáneamente creando las regresiones simples del modelo 1 y modelo 2. Donde, se fijo las variables dependientes **any_prio y war_prio** y explicativas **GPCP_g y GPCP_g_l**. En ese sentido, se obtuvo valores y pruebas sobre errores estándar, errores estándar robustos, homocedasticidad, valores predichos, R2 y R2 ajustado. **Finalmente**, se ajusta la regresión de los modelos agregando efectos fijos y el country_trend y, cada regresión se ajusta a las variables exógenas y endógenas indicadas. Así, exportamos la tabla que junta los resultados del modelo 1 y modelo 2, donde se ajusta nombres de variables, orden de los descriptivos y estructura de la tabla general.

Table 4: Rainfall and Civil Conflict (Reduced-Form)

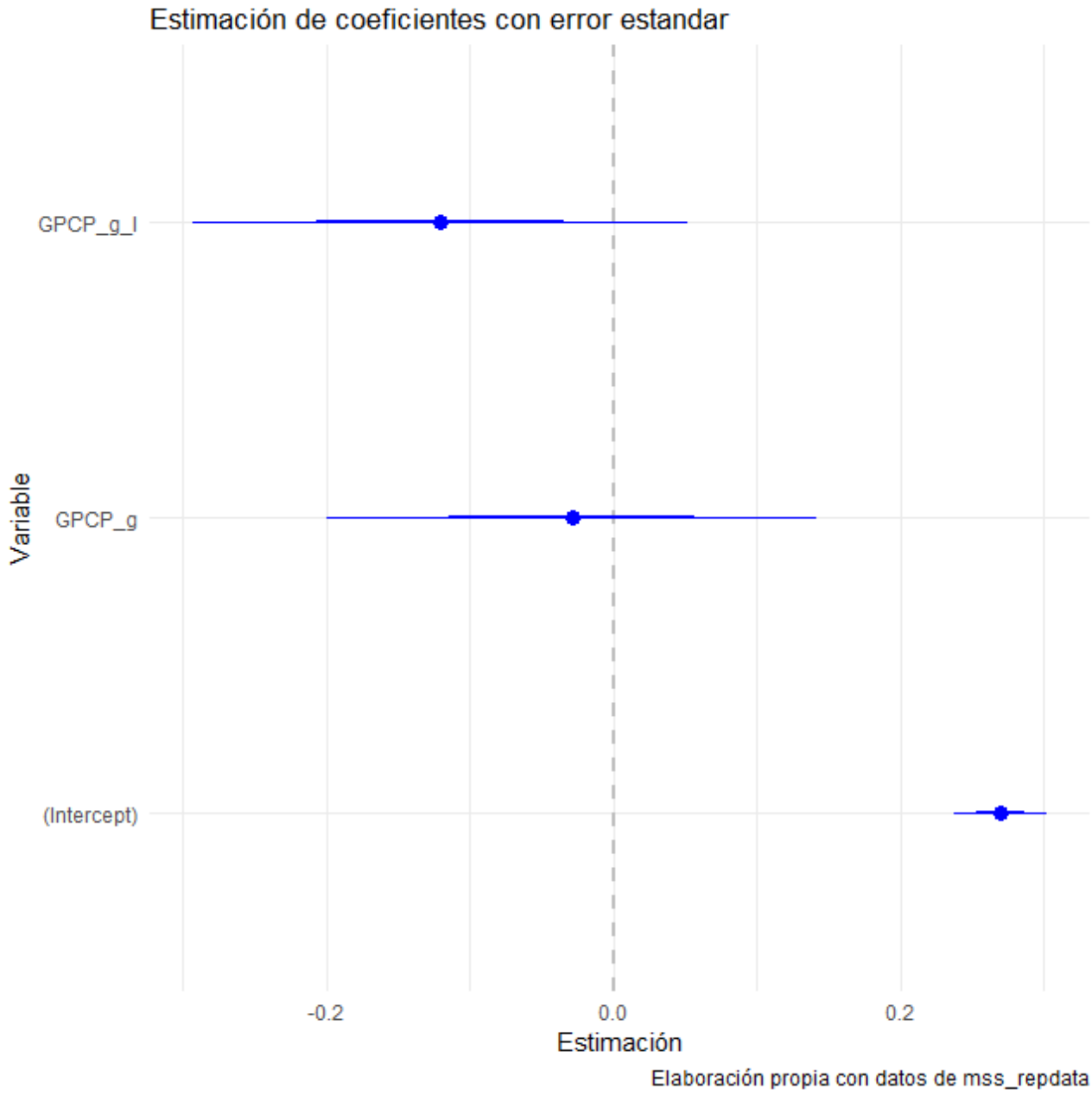
	Dependent Variable	Dependent Variable
	Civil Conflict ≥ 25	Civil Conflict ≥ 100
Explanatory	Death (OLS)	Death (OLS)
Variable	(1)	(2)
Growth in rainfall, t	-0.024	-0.062**
	(0.043)	(0.030)
Growth in rainfall, t-1	-0.122**	-0.069**
	(0.052)	(0.032)
Country fixed effects	yes	yes
Country-specific time trends	yes	yes
R^2	0.708	0.699
Root mean square error	0.24	0.2
Observations	743	743

Note.—Huber robust standard errors are in parentheses.
Regression disturbance terms are clustered at the country level.
A country-specific year time trend is included in all specifications (coefficient estimates not reported).
* Significantly different from zero at 90 percent confidence.
** Significantly different from zero at 95 percent confidence.
*** Significantly different from zero at 99 percent confidence.

1.3 Gráfico del Coeft plot de la variable $GPCP\_g$

- Al realizar la regresión **en R**, se utilizó el comando *coefplot*, se indicó los dos modelos realizados en el item 1.2 como componentes del comando y posteriormente se indicaron el título del gráfico, el nombre de los ejes y la nota. Finalmente, se guardó la imagen correspondiente apartir del comando *ggsave*, indicando la altura, el ancho y la resolución de la imagen.

Coefplot



- En **Python**, se realizó el mismo procedimiento a través de la librería "plt".

2 Web Scraping

Finalmente, en este segundo apartado, encontramos la extracción de data de sitios web. Nos centramos en descargar la información brindada por el programa **Juntos** desde la página de estadísticos InfoJuntos.

2.1 Info-Juntos

En primer lugar accedimos al sitio web, y para descargar la información solicitada explicaremos algunos puntos. En primer lugar vamos a tener que seleccionar la población residente del VRAEM, esto lo vamos a encontrar en el apartado gris de la izquierda (1). Luego de seleccionar la población, vemos en (2) las variables que contiene la base de datos:

- **Departamentos atendidos**
- **Provincias atendidas**
- **Distritos atendidos**
- **Hogares afiliados**
- **Hogares abonados**
- **CCPP en CIAP con hogares afiliados**
- **Establecimientos de Salud**
- **Instituciones Educativas**
- **Transferencia S/.**

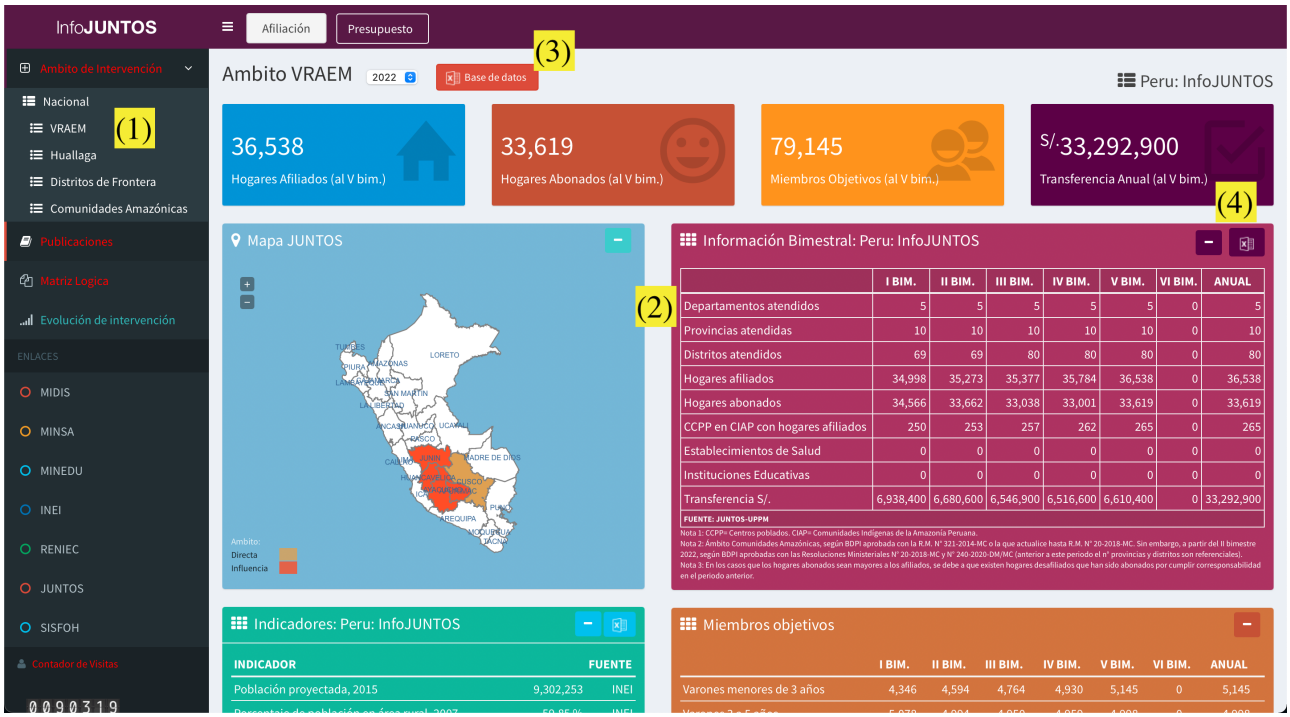
Continuando con la familiarización de la página web, encontramos en (3) los años que buscamos trabajar (2014-2021) y también la información completa; toda la data bajo el formato de tabla de Excel. Finalmente, podemos hacer click en (4) para exportar las tablas de información por bimestres. Ilustrando lo anterior mencionado, podemos comprender de mejor manera con la siguiente captura de pantalla:

Luego de descargar lo solicitado para los años solicitados, recopilamos toda la información y procedemos de inmediato a la elaboración de una sola tabla. Para ello, introducimos la información obtenida en nuestro programa trabajado, en este caso **Jupyter Notebook**; y definimos la primera variable que leerá los "archivos xls" descargados de la base de datos de INFO-juntos.

Una vez con las bases de datos descargadas, eliminamos las columnas y las filas que no usaremos para el ejercicio con el comando ".drop".; y cambiamos de nombre a las filas restantes para facilidad del ejercicio, para luego invertir la matriz de datos a través del cambio del orden de las filas con las columnas. Este proceso lo realizaremos de igual manera para los años del 2014 al 2021, para el caso del VRAEM y seleccionaremos la posición cero de la base para poder especificar en la tabla a trabajar.

Realizamos el mismo procedimiento para los años 2014, 2015 ,2016, 2017, 2018, 2019, 2020 y 2021, para poder obtenerlo dentro de una sola tabla, realizamos un merge outer,

Captura del sitio Web "InfoJUNTOS"



	I BIM.	II BIM.	III BIM.	IV BIM.	V BIM.	VI BIM.	ANUAL
Departamentos atendidos	5	5	5	5	5	0	5
Provincias atendidas	10	10	10	10	10	0	10
Distritos atendidos	69	69	80	80	80	0	80
Hogares afiliados	34,998	35,273	35,377	35,784	36,538	0	36,538
Hogares abonados	34,566	33,662	33,038	33,001	33,619	0	33,619
CCPP en CIAP con hogares afiliados	250	253	257	262	265	0	265
Establecimientos de Salud	0	0	0	0	0	0	0
Instituciones Educativas	0	0	0	0	0	0	0
Transferencia S/.	6,938,400	6,680,600	6,546,900	6,516,600	6,610,400	0	33,292,900

FUENTE: JUNTOS-UPPM

Nota 1: CCPP= Centros poblados. CIAP= Comunidades Indígenas de la Amazonía Peruana.
Nota 2: Ámbito Comunidades Amazónicas, según BDPI aprobada con la R.M. N° 321-2014-MC o la que actualice hasta R.M. N° 20-2018-MC. Sin embargo, a partir del II bimestre 2022, según BDPI aprobadas con las Resoluciones Ministeriales N° 20-2018-MC y N° 240-2020-DM/MC (anterior a este periodo el n° provincias y distritos son referenciales).
Nota 3: En los casos que los hogares abonados sean mayores a los afiliados, se debe a que existen hogares desafiliados que han sido abonados por cumplir corresponsabilidad en el periodo anterior.

INDICADOR		FUENTE
Población proyectada, 2015	9,302,253	INEI

	I BIM.	II BIM.	III BIM.	IV BIM.	V BIM.	VI BIM.	ANUAL
Varones menores de 3 años	4,346	4,594	4,764	4,930	5,145	0	5,145

que permita unir variables de una base de datos con otra base, bajo el nombre de una variable. Este proceso lo repetiremos cuatro veces (merge1, merge2, merge3 y merge4) ya que existen ocho bases de datos en total con respecto a cada año de análisis, y el merge unicamente nos permite unir variables de dos bases de datos. Una vez culminado el primer proceso de merge realizado, empleamos nuevamente merge para unir los merges hallados y les asignaremos nombres (merge5, merge6). Por último, del segundo proceso de unir las bases de datos, utilizamos una vez más merge para unir los últimos dos merges hallados, bajo el nombre de una variable (merge total), la cual contendrá la unión de todos las bases de datos. Finalmente, añadimos una lista llamada "Year" que lo convertiremos en data frame y lo añadiremos al merge final (merge total), para definirlo bajo el nombre de "Imagen2" y llegar al resultado de este ejercicio.

tabla 5

	Departamentos atendidos	Provincias Atendidas	Distritos Atendidos	CPP con hogares afiliados	Hogares afiliados	CCNN con hogares afiliafos	CCPP con hogares abonados	Hogares abonados	Transferencia S/	Year
0	5	10	51	2450	52022	135	2396	46506	55488149.04	2014
1	5	10	59	2451	50354	141	2451	47431	58021852.26	2015
2	5	10	68	141	46487	203	0	40778	50809264.28	2016
3	5	10	69	0	42878	208	0	38896	48194288.38	2017
4	5	10	69	0	39132	212	0	36557	45530749.81	2018
5	5	10	69	0	39137	178	0	35864	42119689	2019
6	5	10	69	0	36385	175	0	36237	43859094.75	2020
7	5	10	69	0	35606	251	0	34299	41799197.88	2021