ESTUDIO SOBRE LOS DETERMINANTES ECONÓMICOS DE LA TASA DE MIGRACIÓN EN 2015

TRABAJO FIN DE GRADO

FACULTAD DE ECONOMÍA
UNIVERSIDAD DE VALENCIA

TUTOR

Carles BRETO MARTÍNEZ

carles.breto@uv.es

AUTOR

Teodoro MOUNIER TEBAS

temou@alumni.uv.es

JUNIO 2020

Resumen

El siguiente trabajo consiste en averiguar, cuáles son los determinantes económicos responsables de las diferencias en las tasas de inmigración entre los países del mundo. El estudio resume el marco teórico a partir de los estudios previos ligados a este tema. Selecciona con justificación las variables necesarias para una estimación econométrica. Desarrolla un análisis econométrico completo recogiendo información de 109 países en 2015 a partir de la base de datos del Banco Mundial y de la OCDE. Para la manipulación de datos y la modelización econométrica se utiliza el programa Rstudio. El código R necesario para el análisis, está presente a lo largo del trabajo para más transparencia en los resultados. De las nueve variables potencialmente explicativas, cinco de ellas fueron seleccionadas. Se estima por mínimos cuadrados ordinarios cuatro modelos, pero solo tres son válidos. Se concluye con la última estimación donde se observan resultados coherentes con la teoría. El trabajo no encontrará un modelo convincente porque sólo conseguirá explicar un poco más del 30 % de la varianza de la tasa de inmigración.

Palabras claves : Inmigración internacional, Análisis econométrico, Modelización Rstudio.

Índice

I. INTRODUCCIÓN

Las migraciones no es algo nuevo sino que nos podemos remontar a más o menos 2.5 millones de años antes JC, con el primer migrante llamado Homo erectus. Se asentó en África entre -1.9 millones de años y -1.5 millones de años, después en Asia (-1.3millones de años, -400 000 años) y por fin en Europa (-800 000 años). Seguido por migraciones de Homo sapiens en los mismos lugares entre (-100 000 años y -10 000 años) [1]. No creo que podamos decir que las razones por las cuales Homo erectus migró fuesen económicas. Pero podemos intuir la búsqueda de una mejora de su bienestar cada vez que migraba.

Si queremos definir la inmigración de hoy en día, nos podemos quedar con la definición sencilla de la enciclopedia Larousse[1]que define la inmigración como la entrada en un país de extranjeros venidos a instalarse y que suele ser por motivos de búsqueda de un empleo o para mejorar su calidad de vida. He elegido esta definición porque solo trata de las migraciones con determinantes económicos y es lo que vamos a estudiar en este análisis.

Antes de todo, me parece importante recordar la diferencia entre un inmigrante y un refugiado[2]. La diferencia se encuentra en las causas del porqué huyen. La confusión se hace porque el resultado es el mismo. Un inmigrante es aquel individuo que deja su país para establecerse en otro por razones económicas. El refugiado huye de su país por situaciones que pudieran poner en peligro su vida ya sean persecuciones políticas, desastres naturales o guerras. Además, se pueden complicar las estadísticas cuando un inmigrante se convierte en un refugiado durante la migración. O al revés, cuando un refugiado no quiere volver a su país porque ha encontrado en el país de acogida una situación económica mayor. Ahora que la diferencia entre un refugiado y un inmigrante está clara volvemos con un poco de historia.

A lo largo de la vida[3] el ser humano se asentó poco a poco pero las migraciones nunca pararon y fueron evolucionando a través del tiempo. Europa fue durante el siglo XIX la mayor fuente de emigración hacia los "países nuevos". Entre 1800 y 1900, cerca de 60 millones de Europeos se instalan en otros continentes con la esperanza de mejorar su vida huyendo las crisis endémicas, las persecuciones políticas y el empobrecimiento. Hoy en día los continentes fuente de migración serían África y América Latina siendo Europa y América del norte los continentes "acogedores" de esa migración. Como lo podéis adivinar, nuestro tema tratará de migraciones. Entonces antes de todo vamos a proceder a unas especificaciones importantes.

[1]https://www.larousse.fr/encyclopedie/divers/immigration/60005 (consultado el 20/01/2020)
[2]https://difiere.com/diferencia-entre-inmigrante-y-refugiado/ (consultado el 20/01/2020)
[3]https://www.britannica.com/topic/immigration https://fr.wikipedia.org/wiki/Immigration (consultado el 20/01/2020)

En nuestro estudio, no haremos una diferencia entre un inmigrante y un refugiado porque la Tasa de inmigración[4], es decir el porcentaje de inmigrantes y refugiados en la población total, no recoge esta diferencia. Esto nos quita el problema de diferenciación entre un inmigrante y un refugiado a la hora de recoger los datos. Pero tendremos que tomar en cuenta esta información en nuestras conclusiones.

A diferencia de los análisis clásicos que intentan explicar los determinantes de los flujos migratorios, nosotros vamos a intentar explicar el stock de migrantes, es decir, ¿por qué existen países con una mayor tasa de inmigración que otros? Y, ¿cuáles son los determinantes económicos que lo pueden explicar? Nuestro análisis no consistirá en encontrar el mejor modelo econométrico que pudiera explicar cuales son los determinantes de los flujos migratorios, como lo hacen los analisis de Moreno Torres [2] y Péridy [3]. Sino que intentaremos estudiar la significatividad de unas variables económicas elegidas previamente que pudieran explicar la persistencia de los migrantes en unos países más que en otros. La elección de las variables explicativas, se harán gracias a los análisis citados anteriormente. Dicho de otra forma vamos a ver si las mismas variables económicas que suelen explicar los flujos migratorios siguen siendo válidas a la hora de explicar el stock de migrantes.

Nuestro estudio consistirá también en desarrollar un análisis econométrico explicado. Es decir, justificando y explicando cada etapa del análisis paso a paso para que uno que no sepa de econometría pudiera seguir el proceso y replicar el análisis. Utilizaremos como programa Rstudio, conocido en el mundo del análisis de datos. En el Anexo os proporcionaré los comandos que me hicieron falta para que podáis replicar el análisis y comprobar los resultados. El script para el análisis econométrico se descompondrá en cuatro etapas. Empezaremos con la recogida de los datos brutos descargados directamente desde las páginas web de entes públicas internacionales que proporcionan datos seguros y controlados. Después procederemos a una limpieza de los datos antes de poder trabajar con ellos de manera más eficiente. Por fin, realizaremos una metodología econométrica con el objetivo de contrastar si las teorías económicas sobre las migraciones se verifican en nuestro modelo. A lo largo del análisis se les propondrá gráficas y resúmenes de resultados con dos objetivos esenciales: el primero, proporcionar un análisis que ponga en valor el potencial de Rstudio y el segundo que el análisis no sea algo aburrido para los ojos.

[4]https://fr.wikipedia.org/wiki/Taux de migration (consultado el 17/01/2020)

Queremos averiguar cuáles son los determinantes económicos de la tasa de inmigración hoy en dia. Para ello realizaremos un análisis de datos cruzados (cross-section data) con datos del año 2015. Es un año que nos proporciona la mayor información actualizada para el mayor número de países. Escogeremos los países que proporcionan la información suficiente para realizar el análisis. Empezaremos con 109 países de los 260 que proporcionaba la base de datos del Banco mundial, pero se podrán ir reduciendo el número de países a lo largo del estudio. En primer lugar procederemos a una justificación de las variables económicas elegidas para explicar la Tasa de inmigración. Vamos a resumir las teorías que existen sobre las relaciones entre la variable explicada y cada una de las variables explicativas elegidas. Comprobaremos las correlaciones con nuestros datos con gráficas. En una segunda parte procederemos a un análisis econométrico empezando por la especificación de nuestro modelo econométrico. Explicaremos la obtención de datos. Realizaremos una exploración de datos, explicando la distribución de cada variable con las estadísticas descriptivas. Una vez las variables conocidas y limpiadas de sus valores atípicos empezaremos la estimación de modelos por Mínimos Cuadrados Ordinarios comprobando las hipótesis básicas y, por último, iremos comparando las estimaciones para quedarnos con el mejor. Por fin, vamos a resumir todos los resultados encontrados y concluiremos sobre el mejor modelo interpretando los resultados y mostrando los límites de nuestro trabajo.

II. ANÁLISIS ECONÓMICO

II.1. Marco teórico

El trabajo de Iván Moreno Torres y Guillem López Casasnovas junto con el análisis de Nicolas Péridy, son las principales fuentes que vamos a usar para resumir las teorías tradicionales y actuales de los determinantes de las migraciones. En este apartado, solamente haremos un resumen de las teorías que nos interesan es decir las teorías económicas. En un segundo apartado, explicaremos porqué elegimos el stock y no los flujos de inmigración apoyándonos en el trabajo de Péridy. Por fin, vamos a justificar la elección de cada variable explicativa ligada a la teoría desarrollada anteriormente y al mismo tiempo iremos contrastando las teorías con los datos de los 109 países seleccionados para nuestro análisis en 2015. Se presentarán gráficas con las correlaciones de cada variable explicativa con el stock de migrantes.

Nicolas Péridy[5] hace una diferencia entre las **teorías tradicionales**, es decir, anteriores a los 90, y las nuevas. Resumiendo, las teorías tradicionales, defienden que los determinantes de las migraciones dependen del diferencial de salarios entre el país de origen y destino netos de los costes migratorios. Esas teorías fueron desarrolladas por Larry A. Sjaastad [4]; Harris y Todaro [5]. Heckscher Ohlin y Samuelson[6], con su teoría sobre el comercio internacional, añadieron que si los países difieren únicamente en su dotación de factores productivos (trabajo y capital), entonces el comercio internacional tiene que igualar el precio de los factores, lo que convierte el comercio en una alternativa a las migraciones. **Los modelos de autoselección** de Borjas [6] permitieron encontrar nuevos determinantes de las migraciones, como por ejemplo, la desigual repartición de los salarios (gini) o las cualificaciones de los migrantes. Borjas completará su teoría de autoselección con la del Welfare Magnets en 1999 [7]. Una vez el migrante se auto-selecciona, elegirá el país donde haya mayores prestaciones sociales para instalarse. Teorías sobre la huida de cerebros vendrán a completar estos modelos.

Iván Moreno Torres y Guillem López Casasnovas[7] resumen de otra forma las teorías sobre las migraciones. Empiezan con **la teoría económica neoclásica** que defiende que habrá migración siempre y cuando el beneficio de emigrar será mayor que los costes, considerando el salario como el único beneficio y el transporte o el hecho de dejar a la familia como costes. Los trabajadores de los países donde los salarios son bajos, es decir donde el factor trabajo es mayor que el factor capital, se desplazarán hacia los países donde los salarios son altos, es decir donde el factor trabajo es menor que el factor capital. Se llegará a un equilibrio donde las diferencias salariales entre países corresponden al coste de migrar (Massey et al., 1993) [8]. **La**

[5]https://www.cairn.info/revue-economique-2010-6-page-981.htm (consultado el 10/01/2020)
[6]https://en.wikipedia.org/wiki/Heckscher%E2%80%93Ohlin$_t heorem$ (consultado el 15/01/2020)
[7]https://old.aecr.org/images//ImatgesArticles.pdf (consultado el 20/01/2020)

teoría económica keynesiana solo añade que el trabajador será atraído por el salario nominal y no el real, lo que pone en tela el equilibrio neoclásico. **La teoría del mercado de trabajo dual** nos explica que las migraciones dependen de la demanda de trabajadores por parte de los países industrializados que necesitan cubrir los puestos de trabajos más difíciles del mercado de trabajo (factores «pull»). Y no son tanto los salarios bajos o el desempleo de los países menos desarrollados que favorecen esta migración (factores «push»). **La nueva economía de la migración**, defiende que la decisión de emigrar no tiene porque ser una decisión individual sino que puede ser una decisión conjunta de los miembros de un hogar para diversificar los riesgos que pueden provocar los fallos de mercado y así asegurar una entrada de dinero. Uno de los ejemplos es el acceso al crédito. Esta decisión permite a las familias reducir su dependencia de la situación económica de un único país. La decisión de emigrar no tiene porqué ser siempre hacia otro país sino que puede ser interna. Esta teoría de Stark y Taylor [9] nos dice que cuando no hay una valoración del capital humano por parte del país de acogida, el individuo podrá elegir quedarse en su país para acceder a trabajos que valoran su formación. Acabaremos nuestro resumen de teorías con **La teoría del sistema mundial**. Esta teoría afirma en general que hay una mayor migración entre países que tienen relaciones comerciales debido a la globalización. Cuanto más cerca estén los dos paises (distancia, antigua colonia, cultura parecida, idioma, etc.), mayor será el número de migrantes.

Se podrían desarrollar mucho más todas esas teorías pero no es nuestro sujeto principal aquí. Les sugiero leer los trabajos de los autores citados anteriormente con una prevalencia para los trabajos mencionados en mi introducción con los cuales se realizó este resumen.

II.2. Especificación del modelo teórico

II.2.1. Variable explicada

Para explicar cuales son los determinantes de las migraciones internacionales se puede elegir entre el flujo de migración o el stock de migrantes. A pesar de que la mayoría de los análisis suelen escoger el flujo de migración, lo que se ajusta más a los supuestos teóricos, nosotros explicaremos el stock de migrantes. La primera razón para justificar esta elección fue que teníamos una información más abundante eligiendo el stock. Además, la base de datos del Banco mundial proporciona solo el stock de migrantes y es la fuente que hemos preferido para la obtención de nuestros datos. La segunda razón es que la variable stock, solo se ve influenciada por movimientos estructurales de migrantes y no migraciones temporales como lo puede hacer el flujo. Esta elección se hizo también en el análisis de Péridy. Llamaremos "tasa_migrantes" a los volúmenes internacionales de migrantes en porcentaje de la población del país de destino. El Banco Mundial define el volumen internacional de migrantes[8] como: *"la cantidad de personas nacidas en un país en el que no viven. También incluye a los refugiados."*. En cuanto al tamaño de la muestra, trabajaremos con datos de 109 países en el año 2015 para que el análisis sea, a nivel mundial, pertinente en la actualidad.

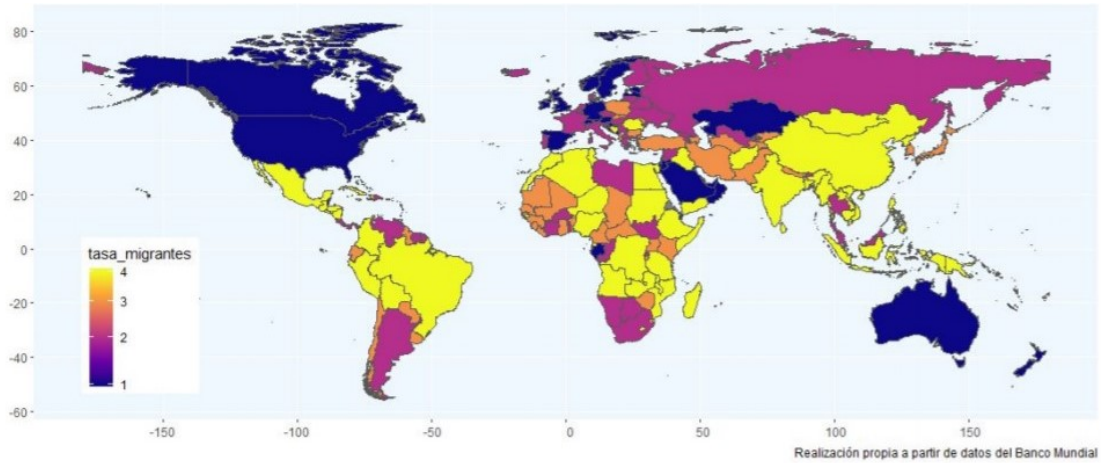


Figura 1: LA TASA DE MIGRACIÓN EN EL MUNDO EN 2015

El planisferio anterior representa a los países del mundo con un color diferente en función del volumen de migrantes de cada país en 2015. Los países se dividen en cuatro colores, los más oscuros, de color azul, son los que tienen un stock de migrantes en porcentaje de la población mayor. Al revés, los países más claros, de color amarillo, son los que tienen un stock de migrantes más bajo. Con este mapa

[8]https://datos.bancomundial.org/indicador/SM.POP.TOTL.ZS (consultado el 17/03/2020)

observamos rápidamente cuales son los países que tienen un stock de migrantes superior, países desarrollados como son los de la OCDE. Y cuales son los países con un stock de migrantes inferior al resto del mundo. Países de África y América Latina y Asia en general.

A simple vista podemos ver a países que son excepciones en su continente. Es el caso de Australia[9], en Oceanía. Australia es un país de migración desde su origen hasta hoy en día ya que se sitúa en el octavo país con mayores números de migrantes (29 % de la población). La razón es que el gobierno australiano sigue desde los años setenta una política de multiculturalismo lo que provoca, desde entonces, olas importantes y continuas de migrantes. Eso sí, suele ser una "migración de cerebros", con la acogida de trabajadores cualificados.

En Asia, también tenemos a varios países con tasas de migrantes por encima de la media del continente, como por ejemplo Arabia Saudí[10]. Casi la totalidad de la mano de obra de este país es extranjera: pakistaníes, indios, yemeníes, por citar algunos. Esa migración masiva empezó con el descubrimiento de recursos petroleros en los años 30. A pesar de que al principio esa migración era de personal técnico cualificado, ahora la mayoría de los trabajadores extranjeros ocupan puestos de empleados en el sector agrícola, de limpieza y de servicios domesticos. El Líbano[11] también se encuentra con un stock de migrantes superior a lo que se podría esperar (% de su población), pero esta vez no son trabajadores extranjeros sino que son refugiados que vienen de los países vecinos. Palestinos después de su expulsión por parte del estado israelí en 1948, irakies huyendo de la violencia en Irak, Sirios huyendo de la guerra y, por fin, sudaneses.

Finalmente, acabaremos con el único país de África que se ve en azul oscuro que es el de Gabón. Nos podemos preguntar, ¿Por qué este país tiene más migrantes que los demás países africanos? El estudio de Claude Bouet [10] sobre el problema de la mano de obra en el Gabón describe a un país "dramáticamente subpoblado, subdesarrollado, insuficientemente abastecido", que se contrasta con su potencial de riqueza proveniente del petróleo sobre todo. El bajo crecimiento natural con la necesidad de cubrir los trabajos podría ser una de las razones por las cuales Gabón tiene una mayor proporción de migrantes que los demás países africanos.

Todos esos países son "atípicos" en sus continentes propios y son interesantes de estudiar por esto. Pero aquí, como trataremos de encontrar determinantes de la inmigración en general, es decir, para todo los países del mundo, procederemos a la eliminación de esas observaciones en el análisis econométrico.

[9] https://en.wikipedia.org/wiki/AustraliaAncestry$_{a}nd_{i}mmigration$ (consultado el 18/03/2020)
[10] https://en.wikipedia.org/wiki/Foreign$_{w}orkers_{i}n_{S}audi_{A}rabia$ (consultado el 18/03/2020)
[11] https://fr.wikipedia.org/wiki/Liban (consultado el 18/03/2020)

II.2.2. Variables explicativas

Con el objetivo de explicar los niveles de las tasas de inmigración de los países observados, hemos elegido variables que en teoría tienen una incidencia sobre las migraciones a nivel global de un país. Algunas de estas variables hacen referencias a las teorías tradicionales donde la decisión de migrar se hace con un análisis beneficio-coste. El beneficio siendo el nivel de salario recogida por la variable PIB per cápita y los costes siendo el desempleo (tasa de desempleo). A esas variables hemos añadido otras como por ejemplo la balanza comercial y las inversiones directas extranjeras que hacen referencia a la apertura del país con el exterior. Otras siendo más ligadas a características internas del país como son, el porcentaje de población urbana, la población en edad de trabajar o el gasto público en educación. Finalmente, hemos elegido variables que recogían al nivel de pobreza de los individuos de un país con el índice de pobreza y el coeficiente de Gini.

Las gráficas de correlación entre las variables explicativas y la tasa de migrantes nos permite estudiar visualmente las correlaciones y comprobar si están en acuerdo con lo que la teoría predice.
➢ Una correlación positiva, donde la recta de regresión es creciente, significa que un aumento de la variable explicativa aumenta también la tasa de inmigrantes.
➢ Una correlación negativa, donde la recta de regresión es decreciente, significa que un aumento de la variable explicativa provoca una disminución de la tasa de inmigrantes.
Además, hemos distribuido a los países por continentes atribuyendo colores diferentes para cada continente respectivamente. Será una forma pertinente para darnos cuenta de las disparidades entre ellos.

El problema de endogeneidad de algunas variables es un problema importante cuando tratamos de hacer una estimación de nuestro modelo. En efecto, una de las hipótesis que nos permite estimar por Mínimos Cuadrados Ordinarios (MCO) es la exogeneidad de las variables explicativas. Eso significa que las variables explicativas tienen que ser independientes con el término de error del modelo. Este problema puede surgir por varias razones y una de ellas es cuando la variable que queremos explicar (dependiente), está explicando las variables explicativas (independientes). En nuestro caso, sospechamos en la variable "población desempleada" de tener un efecto feedback en la tasa de inmigrantes: la tasa de inmigración en un país podría influenciar la tasa de desempleo. Un aumento en el número de inmigrantes aumentaría la demanda de trabajo para la misma oferta lo que supondría un aumento de la tasa de desempleo. No entraremos en los detalles para resolver este problema de endogeneidad, pero podemos decir que es la prueba de Hausman [11], la que permite contrastar la endogeneidad. Dependiendo del número de variables endógenas, se estima por Variables Instrumentales (VI) o por Mínimos Cuadrados en 2 Etapas (MC2E) para que los estimadores sean consistentes.

1. **PIB per cápita, PPA ($ a precios internacionales actuales) año 2015:**

   La variable PIB per cápita[12] tiene como objetivo, dar un panorama del nivel de renta medio de un individuo en los diferentes países. El producto interno bruto por habitante corresponde al valor de la producción de todo el país dividido por el número de habitantes de este. Por lo tanto, es una medida más adecuada que el PIB para comparar el nivel de desarrollo entre países, pero siendo solo una media, no nos dé información sobre la desigual repartición de la renta. Citando la definición del Banco Mundial: *"El PIB por paridad del poder adquisitivo (PPA) es el producto interno bruto convertido a dólares internacionales utilizando las tasas de paridad del poder adquisitivo. Un dólar internacional tiene el mismo poder adquisitivo sobre el PIB que el que posee el dólar de los Estados Unidos en ese país"*. Eso significa que convirtieron en dólares todas las monedas para que tengan el mismo poder adquisitivo de tal forma que podamos comparar el PIB per cápita de cada país.

   La primera hipótesis sería que cuanto más PIB per cápita tiene un país, mayor será su tasa de inmigración. Suponiendo que los individuos toman decisiones racionales en sus elecciones, un individuo con capacidad de elegir entre dos países donde trabajar, sin que entraran en cuestión otros factores, elegirá el país donde obtendrá el mayor salario. Dicho de otra forma, el país donde se consiguen rentas superiores es el país que atraerá a más trabajadores y con el tiempo se encontrará con un volumen mayor de migrantes. Se espera entonces una correlación positiva entre el PIB per cápita y la Tasa de inmigración.

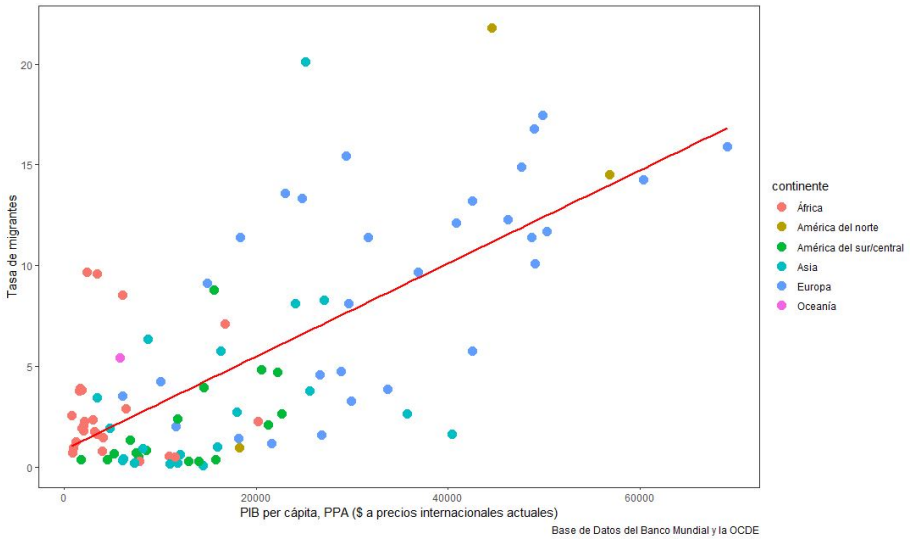


Figura 2: Relación entre el PIB per cápita y la Tasa de inmigración

[12]https://datos.bancomundial.org/indicador/NY.GDP.PCAP.PP.CD (consultado el 20/03/2020)

La gráfica 2 confirma con los datos lo que la teoría predice: cuando el PIB per cápita crece, es decir, cuando la renta percibida por los individuos crece, entonces la tasa de inmigración aumenta en los países en cuestión. Es interesante comparar dónde se sitúa cada continente en la gráfica. Nos damos cuenta que casi todos los países en rojo, es decir, de África se encuentran en la parte baja-izquierda de la gráfica lo que significa que los niveles de renta de estos países son inferiores al resto y entonces las tasas de inmigración también. Para América Latina, en verde, tienen rentas per cápita ligeramente superiores a África, pero con las mismas tasas de inmigración. En cuanto a Europa, en azul, y América del Norte es decir México, Estados Unidos y Canada, en caqui, sin sorpresa, se encuentran con rentas per cápita por encima de los otros continentes, pero también con tasas de inmigración superiores (salvo México). De momento solo se puede confirmar que hay una correlación entre las dos variables. Verificaremos si es significativa a la hora de estimar el modelo econométrico.

2. **Desempleo total (% de la población activa total):**

   La segunda variable es el desempleo[13] en porcentaje de la población activa total. Según la Organización Internacional del Trabajo[14] (OIT): *"El desempleo es la proporción de la población activa que no tiene trabajo pero que busca trabajo y está disponible para realizarlo. Las definiciones de población activa y desempleo difieren según el país"*. En general, un desempleado es un individuo en edad de trabajar y que cumple las tres condiciones siguientes,estar sin empleo, es decir, no haber trabajado ni una hora en una semana de referencia. Estar disponible para coger un empleo en los 15 días, y por fin, haber buscado activamente un empleo en el mes anterior. Corresponden entonces a activos inocupados. La gráfica siguiente los resume.

[13]https://datos.bancomundial.org/indicador/SL.UEM.TOTL.ZS (consultado el 20/03/2020)
[14]Organización Internacional del Trabajo, base de datos sobre estadísticas de la OIT (ILOSTAT)

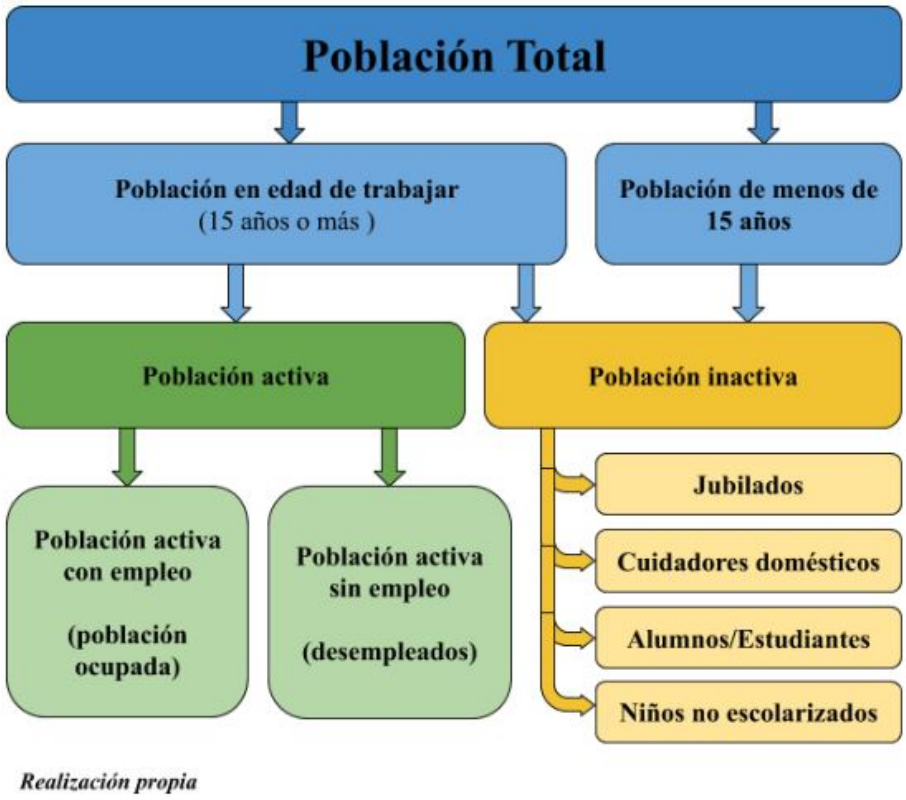


Figura 3: Esquema del mercado laboral

Ahora que tenemos en mente lo que es el desempleo podemos preguntarnos ¿cuál es la relación entre desempleo y tasa de migración? Aquí viene la segunda hipótesis: A mayor tasa de desempleo, menor tasa de inmigración y, por tanto, cuanto menor es la tasa de desempleo, significa que hay oferta de trabajo, lo que atrae a los migrantes en busca de trabajo. Es importante recordar que suponíamos endogeneidad de la variable desempleo, es decir no descartamos que también el desempleo podría ser explicado por la tasa de inmigración. Aquí podemos estar frente a dos situaciones:
➢ Si consideramos que el desempleo explica la tasa de inmigración, se tendría que observar una recta decreciente que justificaría que cuando mayor es la tasa de desempleo de un país, menor es el número de inmigrantes queriendo instalarse en aquel país y por consiguiente menor es la tasa de inmigración.
➢ Si consideramos que es la tasa de inmigración que explica el desempleo se tendría que observar una recta creciente que nos confirmaría la endogeneidad del desempleo junto con la hipótesis de que un país con volúmenes de inmigrantes mayores se encuentra con tasas de desempleo superiores.

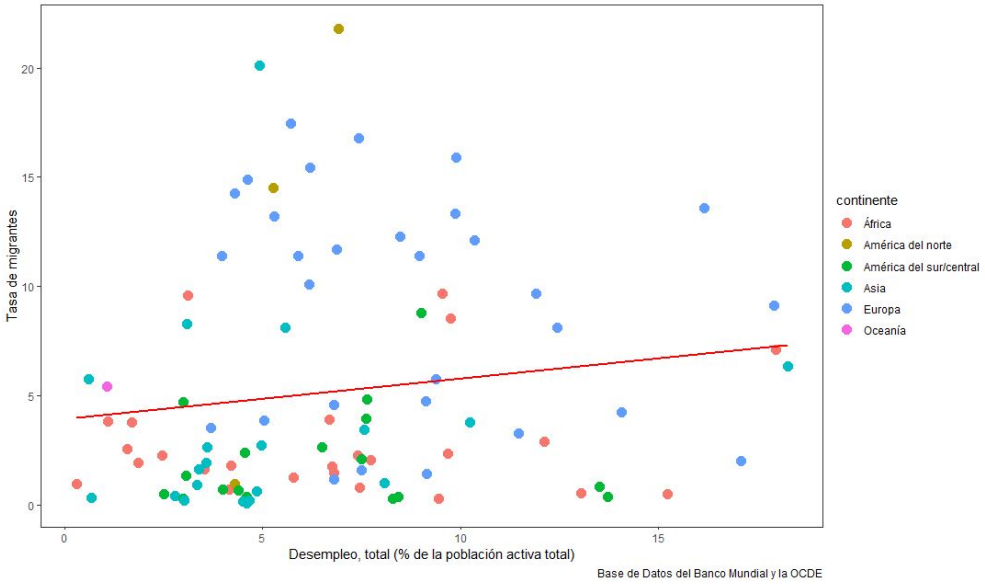


Figura 4: Relación entre el Desempleo y la Tasa de migrante

La gráfica 4, fue realizada a partir de los datos del Banco Mundial en el año 2015 y pone en relación la variable desempleo en abscisa y la tasa de inmigración en ordenada. Podemos observar una relación creciente entre el desempleo y la tasa de inmigración, lo que significa que al menos gráficamente y para el año 2015, no parece que se confirme la hipótesis de que a mayor desempleo menor en la tasa de inmigración. Frente a esta relación se confirmaría más la hipótesis de que cuanto más migrantes hay en un país, mayor es su tasa de desempleo. Tampoco podemos concluir rápidamente porque, si observamos bien, la relación creciente no es muy clara. Primero, porque la recta de regresión no crece de una manera muy significativa. Segundo, por el hecho de que las observaciones están muy dispersas alrededor de la recta y el error de la estimación es seguramente importante. Aquí lo sorprendente es que no hay ningún continente que destaque por tasas de desempleo iguales en sus países. Es más, cada continente recoge a países con tasas de desempleo importantes y a otros países que no.

3. **Tasa de incidencia de la pobreza:**

   La tasa de pobreza nacional[15], según el Banco Mundial: *"es el porcentaje de personas que vive debajo de la línea de pobreza nacional. Las estimaciones nacionales se basan en estimaciones de subgrupos ponderados según la población, obtenidas a partir de encuestas de los hogares."*. Existen otros indicadores que pretenden estimar el nivel de pobreza de un país como por ejemplo el Human Poverty Index (HPI) o el Multidimensional Poverty Index. Pero aquí, hemos elegido solo trabajar con datos provenientes del Banco Mundial, no solo porque son datos con los cuales podemos confiar, sino que también por la sencillez para descargarlos directamente desde R. La variable POV es una variable que viene a completar la variable PIB per cápita. Cuando los ingresos en el país de origen caen por debajo de un cierto umbral, las migraciones se paran, ya que los migrantes no disponen de recursos suficientes para huir de sus países. Esperamos un signo negativo del coeficiente de la recta de regresión.

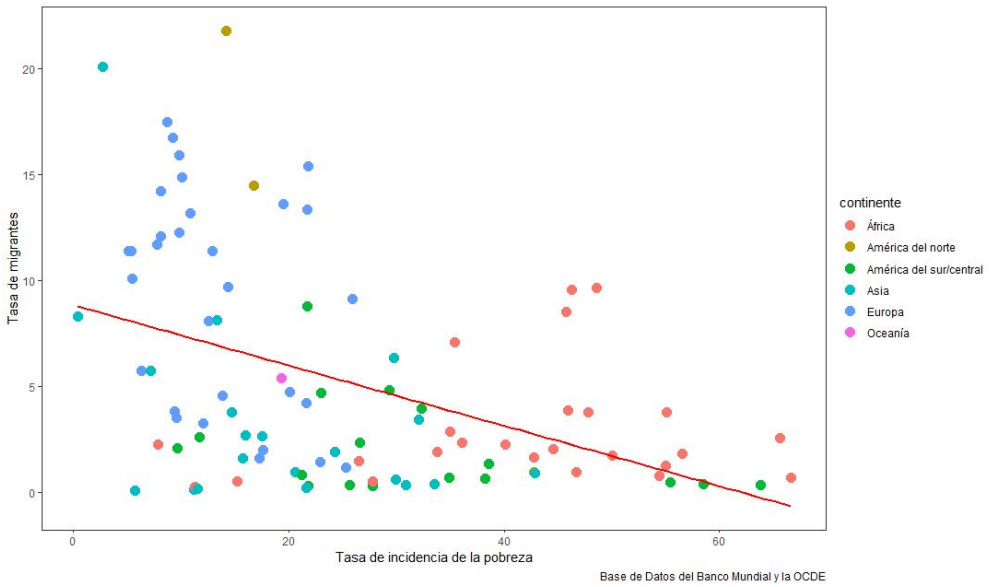


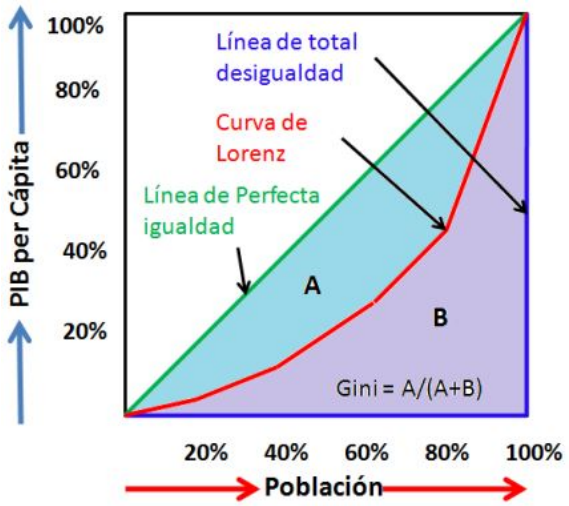
Figura 5: Relación entre Tasa de pobreza y la Tasa de migrantes

La gráfica 5 confirma la hipótesis de que cuando la pobreza es tal en el país de origen, las migraciones disminuyen por el hecho de que no haya suficientes recursos para pagar el viaje y todos los gastos que pueden suponer una migración por razones económicas (no se trata de refugiados).

[15]https://datos.bancomundial.org/indicator/SI.POV.NAHC (consultado el 23/03/2020)

4. **Índice de Gini:**

   El índice de Gini[16] (o coeficiente de Gini) según el Banco mundial: *"mide hasta qué punto la distribución del ingreso entre individuos u hogares dentro de una economía se aleja de una distribución perfectamente equitativa. Una curva de Lorenz muestra los porcentajes acumulados de ingreso recibido total contra la cantidad acumulada de receptores, empezando a partir de la persona o el hogar más pobre. El índice de Gini mide la superficie entre la curva de Lorenz y una línea hipotética de equidad absoluta, expresada como porcentaje de la superficie máxima debajo de la línea. Así, un índice de Gini de 0 representa una equidad perfecta, mientras que un índice de 100 representa una inequidad perfecta."*. A continuación, encontrará un esquema que les enseña cómo se calcula el coeficiente de gini a partir de la curva de Lorenz.



fuente: *https://baconomia2010.blogspot.com*

Figura 6: Curvas de Lorenz

En general, este coeficiente nos informa sobre la repartición de las riquezas entre individuos de un territorio dado. Hemos decidido utilizar este indicador como variable explicativa de la tasa de inmigración, porque se supone que la repartición desigual de la renta está fuertemente ligada a las emigraciones. En efecto, según la teoría, a mayor desigualdad, mayor será la emigración del país, es lo mismo que decir que menor será su tasa de inmigración. Entonces se observa una relación decreciente en la representación gráfica si la teoría cumple con los datos. Por supuesto, existen otras formas de desigualdad y no solo la desigualdad monetaria que recoge el coeficiente de Gini pero nos centramos en determinantes económicos.

[16]https://datos.bancomundial.org/indicador/SI.POV.GINI (consultado el 24/03/2020)

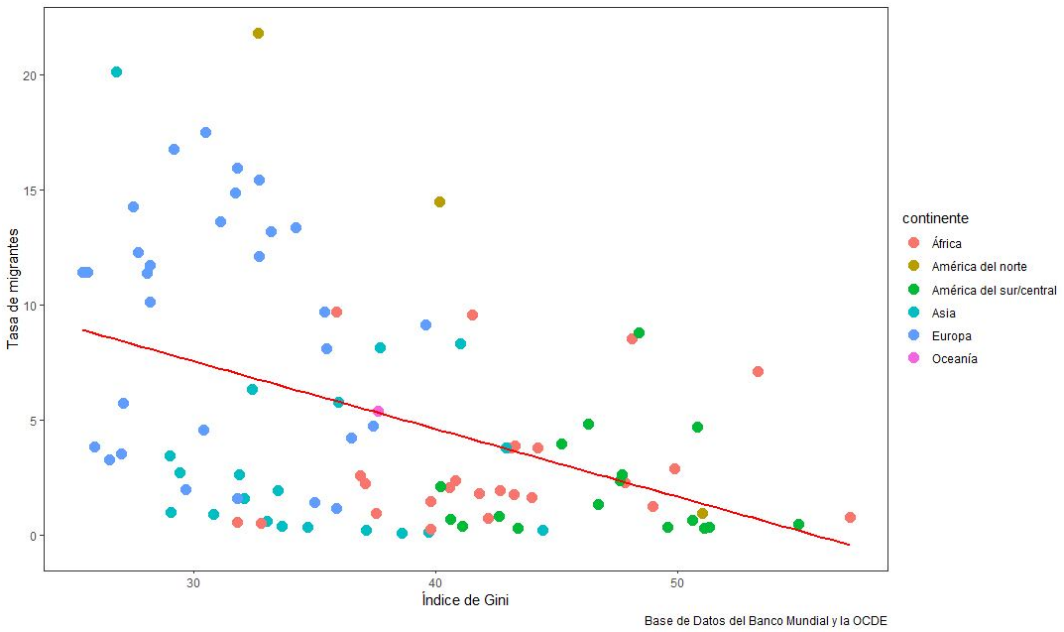


Figura 7: Relación entre el Índice de Gini y la Tasa de migrantes

La gráfica 7 nos confirma la hipótesis de que en un país donde la desigualdad está presente, los migrantes no quieren estar y esos países se encuentran con tasas de inmigración inferiores. Como lo podíamos imaginar, son los países europeos los que tienen los índices de gini más bajos y los países latinoamericanos y africanos los índices superiores.

5. **Población entre 15 y 64 años de edad (% del total):**

   La variable pob_edad_trab[17] es el porcentaje de la población total del país que tiene entre 15 años y 64 años. Nos permite estimar cuál es la proporción de activos en un país, es decir, individuos trabajando o en búsqueda de trabajo. Esta variable vendría de una teoría que dice que, a corto y medio plazo, un país que se está desarrollando, se va industrializando, por lo que su población aumenta como su renta per cápita, entonces su tasa de migración va a crecer en el corto y medio plazo. Aquí explicamos la tasa de inmigración es decir el porcentaje de migrantes en el país y no el de emigrantes. Entonces supondremos aquí, que a mayor porcentaje de activos, en mejor estado está el mercado de trabajo y la economía en general, lo que atraería a los trabajadores extranjeros en búsqueda de trabajo. Podremos encontrar problemas de endogeneidad si suponemos que es la propia tasa de inmigración del país, lo que infla las cifras de activos del país ya que los inmigrantes, por razones económicas, buscan trabajos.

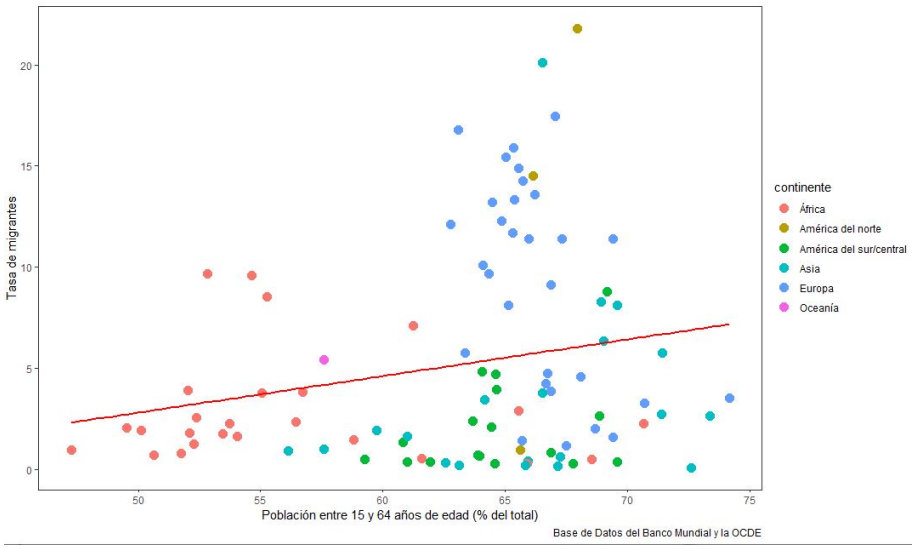


Figura 8: Relación entre la población en edad de trabajar y la tasa de migrantes

A partir de la gráfica 8, podemos observar que si existe una correlación positiva entre la proporción de activos en un país y su tasa de inmigración. Entonces podemos confirmar a la vista de los datos que en un principio, la teoría está cumpliéndose. Pero tendremos que esperar la estimación de los modelos para confirmar esa relación positiva observada.

[17] https://datos.bancomundial.org/indicador/SP.POP.1564.TO.ZS (consultado el 24/03/2020)

6. **Gasto público en educación, total (% del PIB):**

   El gasto público en educación[18] comprende el gasto público total (corriente y de capital) en educación expresado como porcentaje del Producto Interno Bruto (PIB) en un año determinado, 2015 para este análisis. Incluye el gasto del Gobierno en instituciones educativas (públicas y privadas), administración educativa y subsidios o transferencias para entidades privadas (estudiantes/hogares y otras entidades privadas). Esta variable viene de la teoría del Welfare Magnets, desarrollada por Borjas en 1999 que avanza que un inmigrante elegirá el país con mayores prestaciones sociales si tiene la oportunidad. Entonces un país que dedica una parte importante de sus ingresos en la inversión en educación será un país que atraerá a inmigrantes. Aquí podríamos añadir que un inmigrante con familia no huye solo para mejorar su propia condición de vida sino la de su familia incluyendo sus hijos y eso pasa por la educación. La hipótesis es, que un aumento del porcentaje del PIB dedicado a la educación tendrá un efecto positivo en la tasa de inmigración de los países acogedores.

   La gráfica 9, que se encuentra a continuación, confirma visualmente la teoría del Welfare Magnets. Los países con un gasto público en educación mayor, suelen tener una tasa de inmigración superior a los demás también.

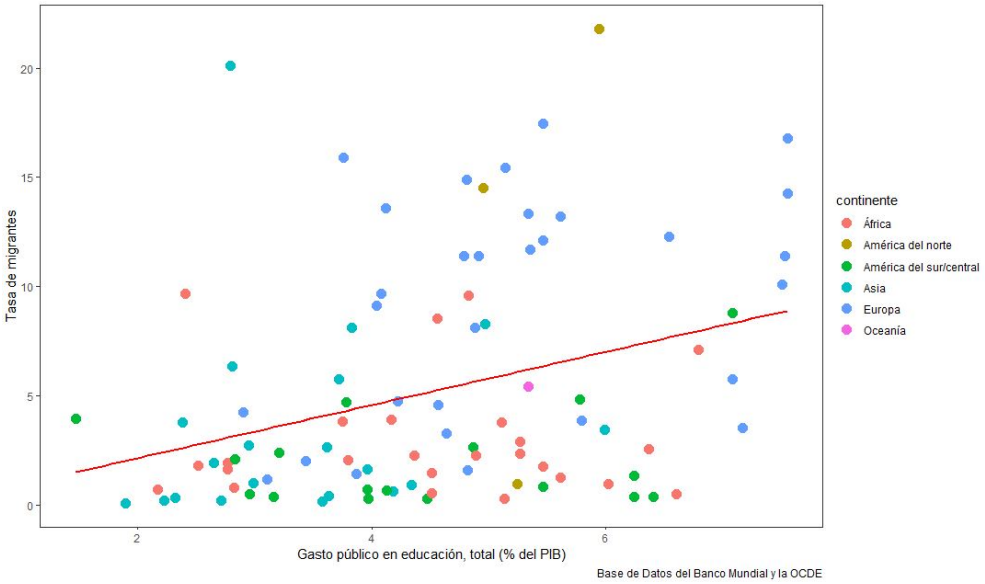


Figura 9: Relación entre el gasto público en educación y la tasa de migrantes

[18]https://datos.bancomundial.org/indicador/SE.XPD.TOTL.GD.ZS (consultado el 24/03/2020)

7. **Balanza comercial de bienes y servicios (% del PIB):**

   La balanza comercial de bienes y servicios[19] es igual a las exportaciones de bienes y servicios menos las importaciones de bienes y servicios. Es una variable que nos informa sobre el grado de apertura del país con el resto del mundo. La hemos añadido como variable explicativa porque el comercio entre países podría explicar también los movimientos migratorios. En efecto, cuanto más abierto comercialmente es un país con otros, mayor será la tasa de inmigración siempre y cuando el país con el cual comercia vaya peor. Dos países que comercian pueden tener o un pasado común (colonias) o las mismas características culturales (países desarrollados entre ellos). En ambos casos el hecho de tener más o menos la misma cultura o idioma, en el caso de las colonias, reducen los costes de emigrar. Eso nos lleva a la conclusión de que la recta de regresión de la gráfica tendrá que ser positiva. A mayor comercio, mayor tasa de inmigración.

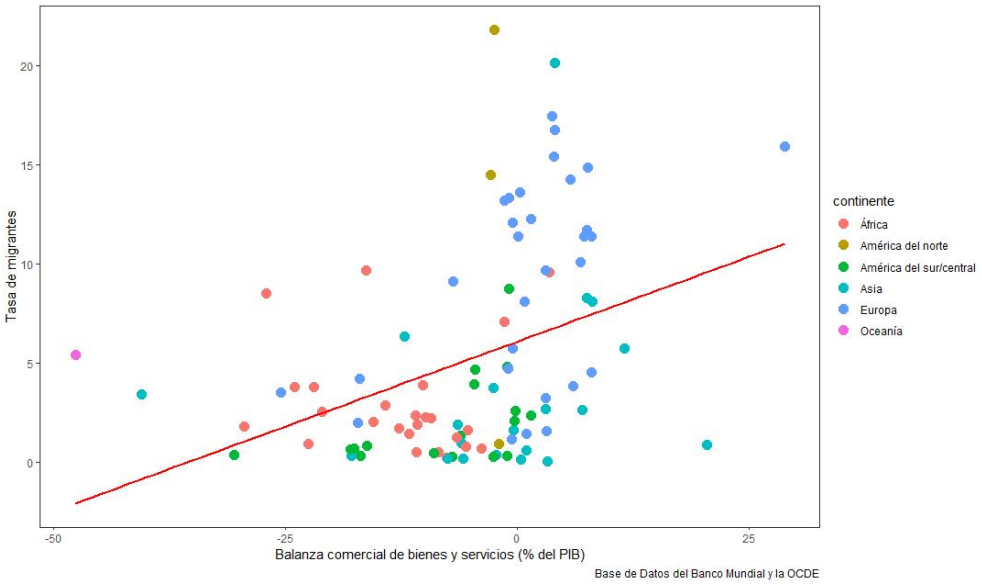


Figura 10: Relación entre la balanza comercial de bienes y servicios y la tasa de migrantes

Podemos observar una relación positiva entre la balanza comercial y la tasa de inmigración como lo enseña la gráfica 10. Esa gráfica apoya la teoría explicada anteriormente. Como para todas las variables anteriores, no se puede concluir con unas correlaciones que observamos a simple vista, pero sí que nos indica más o menos la relación que tienen las dos variables entre ellas.

[19]https://datos.bancomundial.org/indicador/NE.RSB.GNFS.ZS (consultado el 24/03/2020)

8. **Población urbana (% del total):**

   La población urbana[20] es la proporción de individuos que viven en zonas urbanizadas. Cada país define zona urbana como lo hacen sus oficinas de estadística nacional. La fuente de esta variable es de las Naciones Unidas, Perspectivas de la urbanización mundial. En el trabajo de Iván Moreno Torres y Guillem López Casanovas se añade esta variable para explicar los determinantes de las migraciones hacia España y los justifican de la siguiente forma: *"a medida que un país poco desarrollado se va industrializando y urbanizando, aumenta su población y aumenta su renta per cápita, su tasa de migración va a crecer a corto y medio plazo."*. Pero aquí hay que tener cuidado porque no tratamos de explicar la tasa de migración de un país que se está desarrollando, sino su tasa de inmigración. Aquí podemos suponer que a mayor desarrollo del país, mayor será el número de migrantes en el país. Queríamos averiguar si la variable pob_urb podía ser una variable relevante a la hora de explicar las tasas de inmigración a nivel global esta vez.

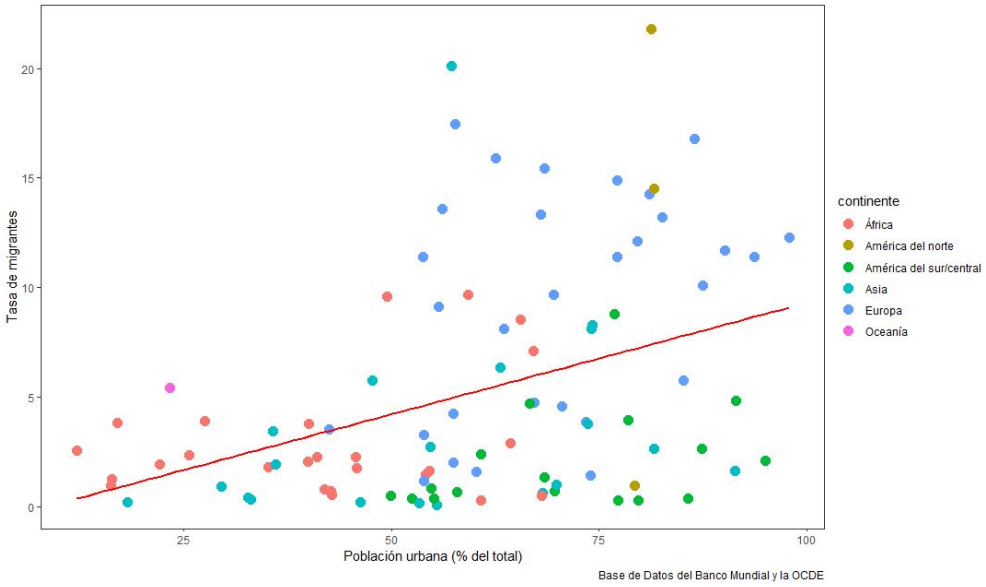


Figura 11: Relación entre la población urbana y la tasa de migrantes

La gráfica 11 pone en relación la población urbana con la tasa de inmigración de cada país observado. Observamos que cuando aumenta la proporción de población urbana de un país, aumenta también su tasa de inmigración. El signo positivo que se esperaba con la teoría se ha confirmado con los datos de 2015.

[20]https://datos.bancomundial.org/indicator/SP.URB.TOTL.IN.ZS (consultado el 24/03/2020)

9. **Inversión extranjera directa, salida neta de capital (% del PIB):**

   Acabaremos nuestro análisis económico con la variable Inversión extranjera directa[21]. Según el Fondo Monetario Internacional: *"La inversión extranjera directa constituye la entrada neta de inversiones para obtener un control de gestión duradero (por lo general, un 10 % o más de las acciones que confieren derecho de voto) de una empresa que funciona en un país que no es el del inversionista"*. La variable ide tendría que explicar más o menos lo mismo que hace la balanza comercial porque la información recogida por las inversiones son los movimientos de capital. Como lo suponíamos con el comercio de bienes y servicios, los movimientos de capital también podrían explicar la tasa de inmigración. Se esperaría un coeficiente positivo.

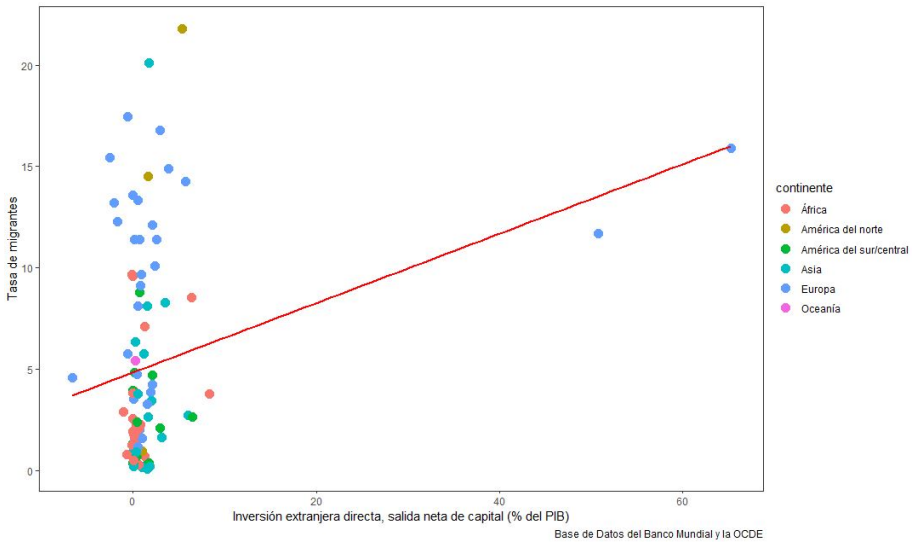


Figura 12: Relación entre la inversión extranjera directa y la tasa de migrantes

En la gráfica 12, podemos observar que todos los datos están más o menos alrededor de 0 y la recta se construyó gracias a dos outliers que se alejan totalmente del resto de países. Esta variable no nos aporta información interesante al no observar ninguna relación clara entre ide y tasa de inmigrantes. Aquí no entraremos en los detalles y las razones de porque los países se distribuyen así. Sino que como tenemos otra variable explicativa (bc) recogiendo la misma información, no necesitamos a la variable ide para contrastar las teorías ligadas al comercio y la inmigración. Entonces, no seleccionaremos ide para el desarrollo econométrico.

[21] https://datos.bancomundial.org/indicador/BM.KLT.DINV.WD.GD.ZS (consultado el 24/03/2020)

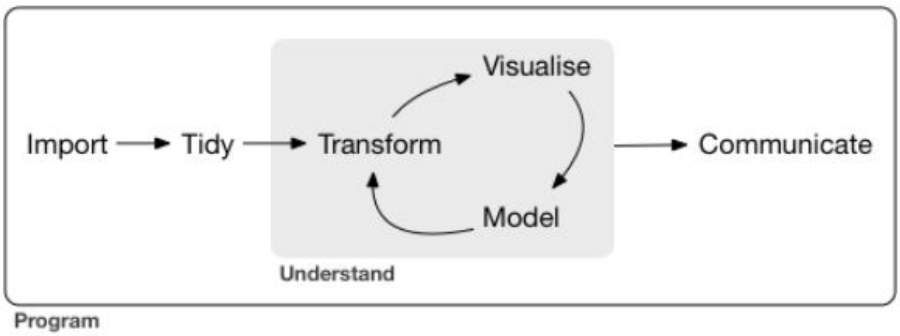
Tabla 1: Resumen de las variables explicativas

Variables	Nombre	Descripción	Medida	Signo esperado	Signo observado	Fuente
Volúmenes internacionales de migrantes	tasa_ migrantes	Porcentaje de la cantidad de inmigrantes.	% de la pob	-	-	Banco Mundial
Pib per cápita	pib_pc	PIB per cápita ajustado por la paridad del poder adquisitivo ($ a precios internacionales actuales) en 2015.	$/habitantes	>0	>0	Banco Mundial
Desempleo, total	pob_ desemp	Porcentaje de la fuerza de trabajo desempleada.	% de la pob activa total	<0	<0	Banco Mundial
Tasa de incidencia de la pobreza	pov	Porcentaje de la población por debajo de la línea de pobreza nacional.	% de la pob	<0	<0	Banco Mundial OCDE
Índice de Gini	gini	Índice de Gini aplicado a la distribución de la renta.	0<gini<100 %	<0	<0	Banco Mundial
Población entre 15 y 64 años de edad	pob_edad _trab	Porcentaje de la población total con edades comprendidas entre los 15 y los 64 años.	% de la pob	-	>0	Banco Mundial
Gasto público en educación, total	gastopub _educ	Porcentaje de gasto público en educación sobre el PIB.	% del PIB	>0	>0	Banco Mundial
Balanza comercial de bienes y servicios	bc	Porcentaje del PIB que representan las importaciones y exportaciones.	% del PIB	>0	>0	Banco Mundial
Población urbana	pob_urb	Porcentaje de la población total en zonas urbanas.	% de la pob	>0	>0	Banco Mundial
Inversión extranjera directa, salida neta de capital	ide	Porcentaje del PIB que representa el flujo neto de inversión directa extranjera.	% del PIB	>0	-	Banco Mundial

III. METODOLOGÍA ECONOMÉTRICA

III.1. Introducción

La modelización econométrica implica un modelo econométrico con unos supuestos. En esta segunda parte trataremos de desarrollar una metodología econométrica con las variables que hemos seleccionado y explicado en la primera parte. La metodología econométrica implica unas etapas previas de análisis de los datos. Para ello, vamos a seguir las etapas que desarrollan Hadley Wickham y Garrett Grolemund en su libro **for Data Science.** [12]



Como lo explica el esquema previo, en primer lugar, descargaremos ("**Import**") los datos que nos interesan del Banco mundial con los paquetes necesarios para ello. Después, limpiaremos ("**Tidy**") la base de datos quitando las observaciones que no nos aportan la información suficiente. Para algunos países observaremos que no tenemos los valores de algunas variables para el año 2015. Entonces, procederemos a una transformación ("**Transform**") de los datos, en el sentido que completaremos la información con una predicción de los valores para el año 2015. Esa predicción se hará calculando las medias de estas variables con los años anteriores. Además, completaremos unas observaciones con datos de la OCDE. En la etapa de visualización ("**Visualise**"), trataremos de observar cómo son los datos, si existen valores atípicos, si las variables siguen una ley normal o si están correlacionadas entre ellas. Todo eso, para que las estimaciones de los modelos ("**Model**"), cumplan con las hipótesis básicas y que sean válidos. Estimaremos varios modelos para buscar aquel que cumpla la hipótesis y al mismo tiempo explique la mayor parte de la varianza de la tasa de inmigración de los países del mundo en 2015.

III.2. Exploración y análisis de la base de datos

III.2.1. Obtención y manipulación de datos

Como expliqué en la introducción, encontrarán el código R que escribí para cada etapa del análisis, de los anexos 1 a 4. Este código no pretende, ni mucho menos, ser el único posible y tampoco el más eficiente para realizar este análisis. Pero creo que es interesante dejar el código para aquel que tiene conocimientos básicos de R y le interese comprobar, mejorar o inspirarse del siguiente trabajo.

1. **Obtención de datos:**

   Para empezar, hemos descargado para el año 2015 cada variable que nos interesaba directamente desde R con el paquete "wbstats". Después hemos creado un dataframe juntando, con "left_join", todas las variables. Un dataframe es como una tabla donde las columnas son las variables y las líneas son las observaciones (países). Teníamos 9 variables con 264 observaciones. Pasamos de 264 observaciones a 179 quitando a los países que no iban a ser relevantes en el análisis por su desconocimiento y la falta de información. Quitamos países como: "*Aruba*", "*Barbados*" o "*Kiribati*". Además hemos quitado 45 observaciones que correspondían a continentes o partes del mundo y no a países específicos. como por ejemplo: "*el mundo árabe*" o "*estados pequeños del Caribe*". Eso nos hizo pasar de 179 a 135 observaciones.

2. **Manipulación de datos:**

   Para el coeficiente de Gini, el índice de pobreza y el gasto público en educación, no teníamos todos los valores para muchas observaciones para el año 2015 pero sí que el Banco Mundial tenía esa información para los años anteriores. Elegimos predecir esos valores calculando la media de esas variables para cada país en los años anteriores. Una vez hecha la predicción fuimos reemplazando los NA's con esos valores. Al ver que seguían faltando valores del índice de pobreza para países importantes como lo son, "*Australia*" o "*Estados Unidos*" decidimos completar la base con datos de la OCDE. Eso supone un sesgo en el análisis al ser dos fuentes con procesos distintos en el tratamiento de datos. Pero también, añade información que necesitábamos para seguir adelante con el trabajo. Por fin quitamos los NA's y pasamos de 135 a 109 observaciones. Podemos empezar el análisis de nuestra base de datos. Primero, comentaremos las estadísticas descriptivas que nos proporcionan las funciones "summary" o "summarytools". Después limpiaremos la base, identificando y eliminando los valores atípicos porque pueden sesgar los resultados de las estimaciones. Observaremos si las variables siguen una distribución normal para poder elegir las buenas técnicas en el cálculo de la matriz de correlación. Nos quedaremos solo con variables que no sean correlacionadas de manera muy significativa para no sesgar tampoco la estimación.

III.2.2. Estadísticas descriptivas

Las estadísticas descriptivas son la información que contiene cada variable. Pueden estar obtenidas gracias a la función "summary". Este comando nos da acceso a los valores mínimos y máximos de cada variable, la media, la mediana y los cuartiles también.

1. **Summary:**

   En el anexo 5, podemos observar que existen disparidades importantes entre los 109 países de la muestra. En efecto, la amplitud, es decir la diferencia entre los valores mínimos y máximos de la mayoría de variables, es importante. La mayor diferencia está en el PIB per cápita con un mínimo de 764,2$ PPA y un máximo de 103.750,8$ PPA, pero también el índice de pobreza, la balanza comercial o la población urbana carecen de disparidades importantes.

2. **Homogeneidad:**

   Podemos observar también que para una gran mayoría de las variables la media y la mediana son muy cercanas. La mediana se diferencia de la simple media aritmética porque divide la muestra en dos, es decir con el 50 % por debajo de ella y el 50 % restante por encima. Por tanto, cuando está cerca de la media, significa que las variables son relativamente uniformes, sin valores falseando la media. Dicho de otra forma, si la mediana se aleja mucho de su media, podemos decir que la distribución de la variable no es homogénea. Observando las medias y medianas podemos concluir que para nuestra variable explicada "*tasa_migrantes*" y las variables explicativas siguientes "*pib_pc*" y "*pov*", las medianas se alejan de sus medias. Eso sí para el resto de variables explicativas las observaciones se distribuyen de manera relativamente homogénea. Lo que

   puede provocar esa diferencia entre las medias y medianas son las observaciones atípicas que tienen un peso significativo en la media a pesar de que solo son un par de países y no la mayoría. Por tanto, procederemos a la eliminación de esos países porque no son representativos y pueden comprometer la estimación si no se retiran.

III.2.3. Valores atípicos

Para determinar si existen valores atípicos para cada una de las variables explicativas utilizaremos la función "Boxplot". Es una representación gráfica que nos permite a simple vista averiguar cuáles son los valores potencialmente atípicos, serán los valores que se encuentran fuera del bigote. A partir de esos resultados, realizaremos unos test para confirmar si esos mismos países, en nuestro caso, son efectivamente atípicos. Los países atípicos se verán retirados de la base.

1. **Boxplot (visualmente):**

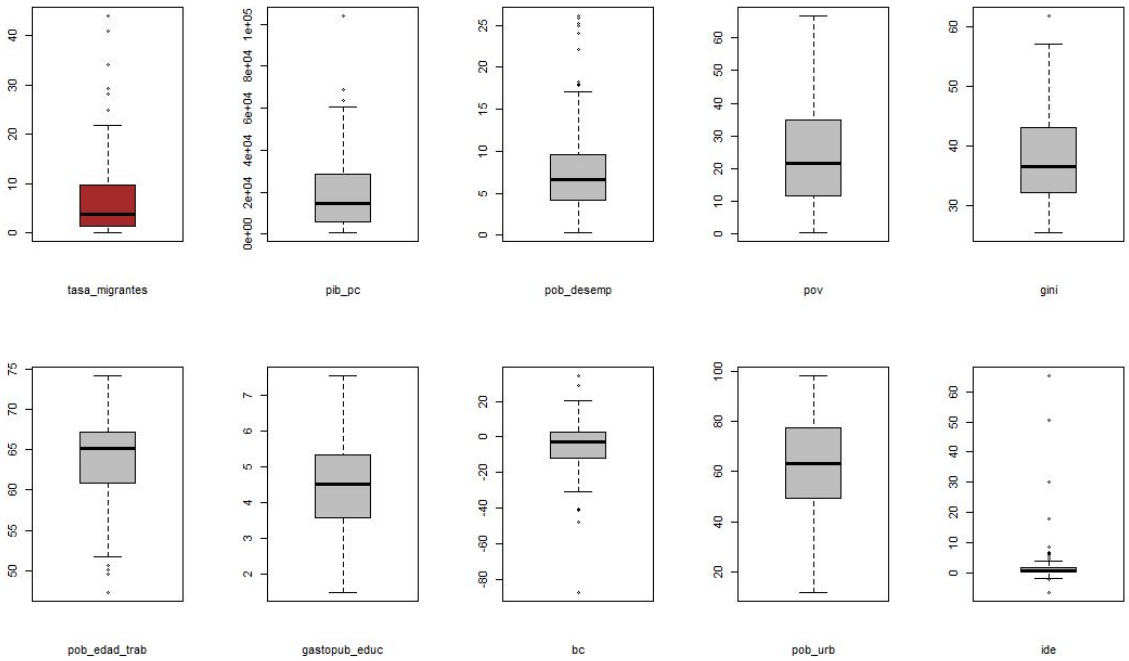


Figura 13: Diagramas de caja para cada variable

La gráfica anterior nos permite segmentar las variables en tres categorías, las que tienen un valor atípico, las que tienen varios valores atípicos y por fin las que no tienen. Dos test se pueden aplicar en función del número de valores atípicos. Aplicaremos el test de Grubbs a la variable "*gini*" porque visualmente podemos ver un único punto fuera del bigote. Para todas las demás menos "*pov*", "*gastopub_educ*" y "*pob_urb*" donde no se observan puntos sueltos, aplicaremos el test de Rosner.

2. **Test de Grubbs para "gini":**

   La prueba de Grubbs [13] consiste en considerar una hipótesis nula H0 donde no hay valores atípicos, y su alternativa, H1 que sí existe un valor atípico. En el anexo 6 les proporcione el código y resultado del test. Considerando un nivel de significatividad del 5 %, el p-valor igual a 0.1059 es superior a 0.05 entonces aceptamos la hipótesis nula de que no haya un valor atípico para la variable "*gini*"

3. **Test de Rosner para las demás variables:**

   La prueba de Rosner [14] es parecida a la de Grubbs pero esta vez la hipótesis alternativa H1 es que hay entre dos y diez valores atípicos. Tenemos que aplicar el test de Rosner para todas las variables donde observábamos en los boxplot muchos puntos atípicos. Desarrollaremos el test de Rosner para la variable "*pib_pc*" únicamente, porque las demás variables siguen el mismo proceso. En el anexo 7, podemos observar que la prueba de Rosner nos indica que de las tres observaciones que suponíamos atípicas solo una lo es, con un nivel de significatividad del 5 %. Gracias a las funciones "sort" y "order" se puede averiguar cuál es la observación atípica. En nuestro caso es la observación 64, que corresponde al Luxemburgo. En la tabla siguiente, resumimos todos los países atípicos que quitaremos antes de empezar las estimaciones econométricas.

Tabla 2: Resumen de los valores atípicos detectados

Variables	Número de valores potencialmente atípicos	Test aplicado	Número de valores efectivamente atípicos	Observaciones atípicas retiradas de la base
tasa_migrantes	6	Rosner	6	50/4/97/61/54/64
pib_pc	3	Rosner	1	64
pob_desemp	9	Rosner	6	76/108/92/39/95/93
gini	1	Grubbs	0	X
pob_edad_trab	4	Rosner	0	X
bc	7	Rosner	1	62
ide	>10	X	X	X

La última columna de la Tabla 2 nos indica las observaciones atípicas. Es decir, que después de haber aplicado las pruebas adecuadas a cada variable nos encontramos con 14 observaciones atípicas entre todas las variables. De esas 14 observaciones solo retiraremos 13 porque la observación 64 es atípica tanto para "*tasa_migrantes*" como para "*pib_pc*". Además, retiraremos a la variable "*ide*" de nuestras variables explicativas por tres razones. La primera es que su relación con la variable explicada "*tasa_migrantes*" no fue concluyente (cf: parte económica de este estudio). La segunda razón es que constatamos que la variable "*ide*" tiene más de 10 valores atípicos y que no podemos aplicar el test de Rosner que es eficiente solo para un número inferior a 10. La tercera razón es que la balanza comercial tiene que contrastar la misma teoría que la inversión directa extranjera entonces no perderemos información al retirarla de la base. Después de este apartado nos quedamos entonces con 96 países y 8 variables explicativas.

III.2.4. Distribución de las variables (Normalidad)

Antes de proceder a las estimaciones nos falta comprobar una de las hipótesis básicas para poder estimar un modelo por Mínimos Cuadrados Ordinarios (MCO) es la hipótesis de exogeneidad de las variables explicativas. Esa hipótesis supone dos cosas, primero que el número de observaciones sea mayor que el número de variables explicativas, lo que se verifica en nuestro caso. La segunda es que no haya multicolinealidad entre las variables explicativas. Es decir que no haya correlación entre ellas. Para comprobar esa correlación podemos utilizar el coeficiente de Pearson o el de Spearman en función de la distribución de las variables. Como el coeficiente de Pearson supone una distribución normal de la variable, tendremos que verificar la normalidad de las variables para saber si utilizaremos el coeficiente de Pearson o Spearman en el cálculo de la matriz de correlación. Una distribución normal se puede comprobar visualmente con un histograma y con una prueba, Shapiro Wilk normality test.

1. **Histograma (visualmente):**

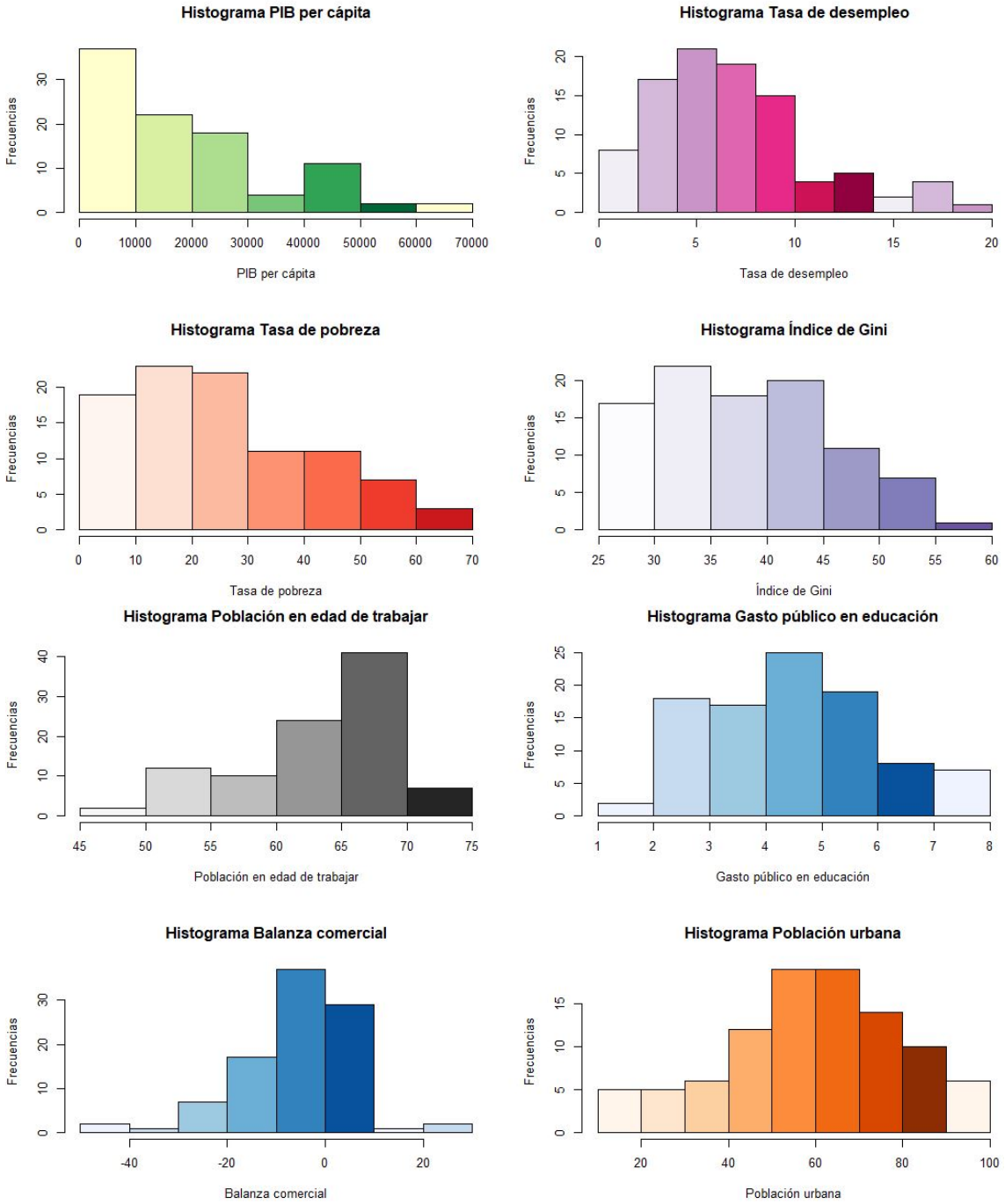


Figura 14: Histogramas de las variables explicativas

Mirando a los histogramas de la figura 14, podemos constatar a simple vista, que las variables no siguen una distribución normal. Siendo el gasto público en educación y la población urbana las dos variables acercándose más a esa distribución. Podemos completar nuestro análisis buscando la skewness y la curtosis para cada variable. Esos coeficientes nos dan unas indicaciones a propósito del aplanamiento y la simetría de la distribución de las variables. La ley normal se caracteriza con un coeficiente de aplanamiento y de asimetría igual a cero. Cuando el coeficiente de Skewness es positivo significa una asimetría a la izquierda y al revés, un coeficiente negativo, una asimetría a la derecha. En el caso del aplanamiento, una curtosis positiva significa que la distribución es más plana que una distribución normal. El caso contrario es cuando la curtosis es negativa.

En el anexo 8, tenemos los coeficientes de Skewness y curtosis para cada una de nuestras variables. Como lo suponíamos, es el gasto público en educación la variable que se acerca más a una distribución normal si miramos a la simetría con un coeficiente de Skewness igual a 0.2. Por lo general, estamos frente a asimetrías hacia la izquierda mayores del 0.7 excepto para la población en edad de trabajar, la balanza comercial y la población urbana. En cuanto al aplanamiento, es la distribución del PIB per cápita la que se acerca más a una distribución normal con una curtosis de 0.18 seguido de la variable población urbana, con un coeficiente de -0.2. Aquí las únicas variables que se distribuyen de forma más plana que una normal, es decir con un coeficiente de curtosis positivo, son el PIB per cápita, la población desempleada y por fin la balanza comercial.

Los histogramas y los coeficientes de Skewness y curtosis son herramientas visuales para ver si una distribución sigue más o menos una ley normal. Para confirmar estas hipótesis sobre la normalidad de las variables es necesario efectuar el test de Shapiro-Wilk.

2. **Test de Shapiro:**

   El test de Shapiro-Wilk consiste en considerar una hipótesis nula H0 de que la variable sigue una ley normal y la alternativa de que no con un nivel de significación que decidimos. El anexo 9 proporciona todos los test de normalidad para todas las variables explicativas. Con un nivel de significatividad del 5 %, rechazamos la hipótesis nula de que las variables pib_pc, pob_desemp, pov, gini, pob_edad_trab, y bc siguen una distribución normal. Para el gastopub_educ y pob_urb, se confirma estadísticamente lo que suponíamos visualmente. Fijándose en el p-valor, no se rechaza la hipótesis nula de que sigan una ley normal con un nivel de significatividad del 5 %.

III.2.5. Correlación

Cuando las variables no siguen una ley normal, el cálculo de la correlación de Spearman es recomendada. Este tipo de correlación se dice robusta porque no depende de la distribución de los datos como lo hace la correlación de Pearson. Existen una multitud de formas para representar una matriz de correlación con Rstudio. Queremos proporcionarles al menos dos formas de representarla. La primera forma, es la figura 15 que sigue. La segunda forma está en el anexo 10.

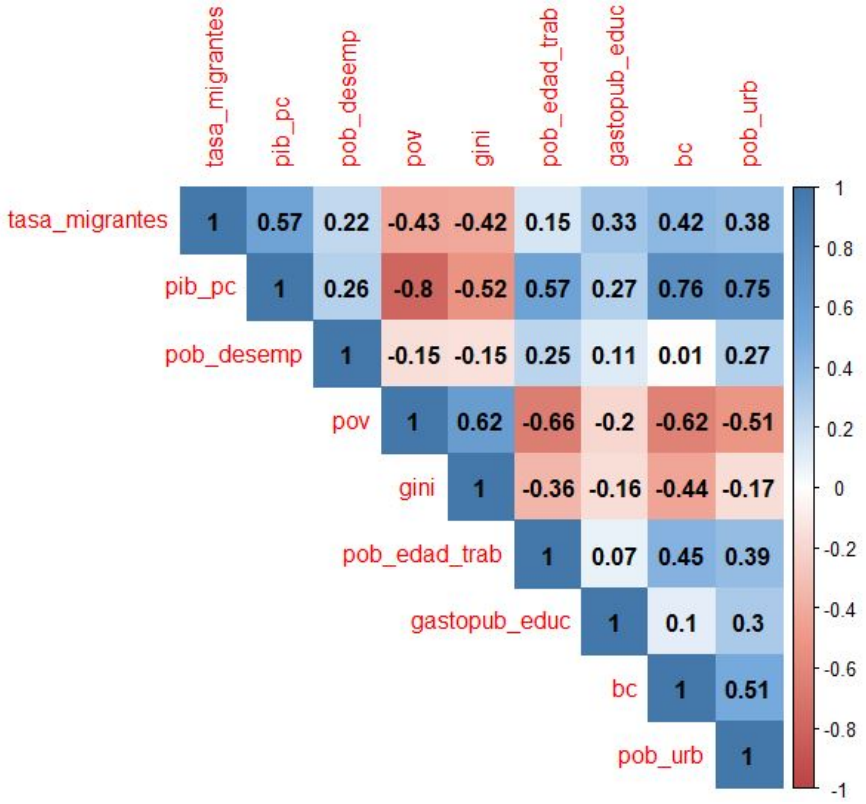


Figura 15: Matriz de correlación 1

Si analizamos las correlaciones entre las variables explicativas, podemos observar que el PIB per cápita y el índice de pobreza tienen coeficientes superiores al 0.5 en valor absoluto, con casi todas las demás variables. Para la balanza comercial, al ser de 0.51 la correlación con la población urbana, la conservaremos para las estimaciones. A la hora de estimar iremos quitando la variable "*pov*" porque recoge más o menos la misma información que "*gini*". En cuanto "*pib_pc*" tendremos que quitarla también porque está correlacionada con todas las demás variables lo que podría falsificar los resultados del análisis.

III.3. Estimaciones econométricas

III.3.1. Especificación del modelo econométrico general

El modelo de regresión lineal múltiple es el siguiente:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + \beta_4 \cdot X_{4i} + \beta_5 \cdot X_{5i} + \beta_6 \cdot X_{6i} + \beta_7 \cdot X_{7i} + u_i$$

Donde :

- Y_i : **"tasa_migrantes"** nuestra variable explicada, dependiente, es la que vamos a analizar usando algunas de las 9 variables explicativas que hemos ido analizando en la primera parte. Representa el volumen de migrantes en 2015 en porcentaje de la población.
- X_{1i} : **"pib_pc"** es el producto interior bruto dividido por el número de habitantes del país ajustado por la paridad del poder adquisitivo ($ a precios internacionales actuales) en 2015.
- X_{2i} : **"pob_desemp"** es el porcentaje de la fuerza de trabajo desempleada (% de la pob activa total).
- X_{3i} : **"gini"** es el Índice de Gini aplicado a la distribución de la renta (0<gini<100 %)
- X_{4i} : **"pob_edad_trab"** es el porcentaje de la población total con edades comprendidas entre los 15 y los 64 años (% de la pob).
- X_{5i} : **"gastopub_educ"** es el porcentaje de gasto público en educación sobre el PIB (% del PIB).
- X_{6i} : **"bc"** es el porcentaje del PIB que representan las importaciones y exportaciones (% del PIB).
- X_{7i} : **"pob_urb"** es el porcentaje de la población total en zonas urbanas (% de la pob).

Ademas :

- i es el número de observaciones, es decir el número de países de la muestra → i=1,2,3,...,96
- $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4,\ \beta_5,\ \beta_6,$ son los parámetros del modelo.
- u_i es el término de error.

III.3.2. Hipótesis de un modelo estimado por MCO

Todo modelo estimado por el método de los mínimos cuadrados ordinarios tiene que cumplir unos criterios específicos que describimos a continuación antes de poder validar el modelo y sacar conclusiones.

El modelo tiene que ser **bien especificado**, es decir que existe una relación lineal entre las variables, entre los parámetros y que no haya omisión o redundancia de variables. La prueba de Reset Ramsey nos permitirá verificar si nuestro modelo está bien especificado o no. Tampoco tiene que haber **multicolinealidad** perfecta entre las variables exógenas, dicho de otra forma, el coeficiente de correlación simple entre dos variables explicativas tiene que ser distinto de la unidad. Lo verificaremos con el Factor de Inflación de la Varianza, VIF.

Las últimas hipótesis tienen que ver con la perturbación aleatoria. En primer lugar, los residuos tienen que seguir una **distribución normal**. Lo que se puede verificar con la prueba de Jarque-Bera pero solo se tiene que verificar cuando la muestra es inferior a 30 observaciones. En nuestro caso, tenemos 96 observaciones, por lo que no será necesario verificar esta hipótesis. En segundo lugar, se verificará la **homocedasticidad** de los residuos. La varianza del término de error tiene que mantenerse constante a lo largo de todo el periodo muestral. Varias pruebas existen para esta hipótesis, pero aquí utilizaremos el test de Breusch-Pagan. Por fin acabaremos nuestra serie de pruebas observando la distancia de Cook que nos informa sobre las **observaciones influyentes**. La eliminación de un país no tendría, en principio, que afectar la precisión de la regresión. Si todas esas condiciones no se verifican entonces el modelo no será válido.

Para cada modelo que estimaremos, verificaremos si cumple o no las hipótesis básicas, y si no las cumple, tomaremos medidas para que la estimación siguiente cumpliera cada una de esas hipótesis.

III.3.3. Estimación del modelo 1

Estimamos nuestro primer modelo suponiendo que todas las variables explicativas son exógenas, no hay omisión de variables, ecuaciones simultáneas ni tampoco un efecto feedback. En la introducción y en el apartado dedicado a justificar la elección de las variables explicativas, teníamos dudas sobre la endogeneidad de la variable "*pob_desemp*" entonces no la incluiremos en las estimaciones por MCO para cumplir con exogeneidad de las variables explicativas.

La especificación del primer modelo es la siguiente:

$$tasa\_migrantes_i = \beta_0 + \beta_1 \cdot gini_i + \beta_2 \cdot pob\_edad\_trab_i + ... \\ ... + \beta_3 \cdot gastopub\_educ_i + \beta_4 \cdot bc_i + \beta_5 \cdot pob\_urb_i + u_i$$

```
Call:
lm(formula = tasa_migrantes ~ gini + gastopub_educ + pob_edad_trab +
    bc + pob_urb, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5863 -2.8987 -0.5175  2.1769 13.3084

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.33596    6.68765   2.443 0.016531 *
gini          -0.24374    0.06322  -3.855 0.000217 ***
gastopub_educ  0.75827    0.32832   2.310 0.023202 *
pob_edad_trab -0.14255    0.09215  -1.547 0.125414
bc             0.09127    0.04544   2.009 0.047584 *
pob_urb        0.07019    0.02890   2.428 0.017159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.349 on 90 degrees of freedom
Multiple R-squared:  0.3699,    Adjusted R-squared:  0.3349
F-statistic: 10.57 on 5 and 90 DF,  p-value: 5.274e-08
```

Figura 16: MODELO n°1: Especificación lineal

El modelo estimado es el siguiente :

$$\widehat{tasa\_migrantes_i} = 16{,}34 - 0{,}244 \cdot \widehat{gini_i} - 0{,}142 \cdot \widehat{pob\_edad\_trab_i} + ... \\ ... + 0{,}758 \cdot \widehat{gastopub\_educ_i} + 0{,}091 \cdot \widehat{bc_i} + 0{,}07 \cdot \widehat{pob\_urb_i}$$

Podemos empezar mirando la significatividad global del modelo con el test de Fisher. Como el $p - valor = 5{,}274e - 0{,}8 < 0{,}5$ rechazamos la hipótesis nula de que el modelo no es significativo en su conjunto, es decir que existe al menos una variable explicativa que es significativa en la explicación de la tasa de inmigrantes. En nuestro caso, la única variable no significativa al menos al 10 % es la variable población en edad de trabajar. La calidad de ajuste del modelo se puede observar con el valor del R^2, que en este caso es igual al 0.3699. Eso significa que el modelo explica más o menos el 37 % de la varianza de la tasa de inmigración.

III.3.4. Pruebas de hipótesis

El primer criterio consiste en verificar si el modelo fue estimado con la buena forma funcional. Para ello tenemos el test de Ramsey. La prueba consiste en verificar que el modelo no sufre una omisión de variables pertinentes, introduciendo una variable ficticia y mirando su significatividad. Si no es significativa, entonces el modelo está bien especificado, si no se tendrá que modificar la forma funcional para que el modelo se ajuste mejor.

1. **De la forma functional (Test de Ramsey):** En la hipótesis nula la forma

   funcional del modelo es lineal, y en la hipótesis alternativa no lo es. El p-valor de la prueba de Ramsey es igual a 0.09146. Podríamos aceptar la hipótesis nula de que el modelo esté bien especificado solo si consideramos un nivel de significatividad del 10 %. En nuestro análisis vamos a estimar un segundo modelo que tenga una forma funcional más ajustada a la muestra. Entonces rechazamos la hipótesis nula de que el modelo está bien especificado con un nivel de significación del 10 %. Para resolver este problema de mala especificación

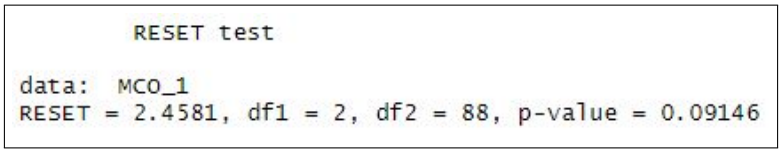


Figura 17: Test de Reset Ramsey

   podemos cambiar el modelo poniendo la variable explicada en logaritmos por ejemplo, se transformaría en un modelo log-lin. Si además, se añade logaritmos a variables explicativas sería un modelo log-log. En nuestro caso vamos a empezar estimando un nuevo modelo añadiendo logaritmos solo en nuestra variable explicada, tasa migrantes dejando las variables explicativas sin cambios.

III.3.5. Estimación del modelo 2

Estimamos un nuevo modelo con una forma funcional log-lin para resolver la mala especificación del primer modelo. La especificación del segundo modelo es la siguiente:

$$log(tasa\_migrantes_i) = \beta_0 + \beta_1 \cdot gini_i + \beta_2 \cdot pob\_edad\_trab_i + ... \\ ... + \beta_3 \cdot gastopub\_educ_i + \beta_4 \cdot bc_i + \beta_5 \cdot pob\_urb_i + u_i$$

```
Call:
lm(formula = log(tasa_migrantes) ~ gini + pob_edad_trab + gastopub_educ +
    bc + pob_urb, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.54051 -0.80593  0.06627  0.99421  1.82598

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.756998   1.745393   2.725 0.007716 **
gini          -0.061891   0.016500  -3.751 0.000311 ***
pob_edad_trab -0.055272   0.024051  -2.298 0.023875 *
gastopub_educ  0.209911   0.085688   2.450 0.016232 *
bc             0.012108   0.011859   1.021 0.309972
pob_urb        0.019607   0.007544   2.599 0.010920 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.135 on 90 degrees of freedom
Multiple R-squared:  0.3208,    Adjusted R-squared:  0.283
F-statistic:   8.5 on 5 and 90 DF,  p-value: 1.271e-06
```

Figura 18: MODELO n°2: estimación de un modelo log-lin

El modelo estimado es el siguiente :

$$\widehat{log(tasa\_migrantes_i)} = 4{,}757 - 0{,}062 \cdot \widehat{gini_i} - 0{,}055 \cdot \widehat{pob\_edad\_trab_i} + ... \\ ... + 0{,}21 \cdot \widehat{gastopub\_educ_i} + 0{,}012 \cdot \widehat{bc_i} + 0{,}02 \cdot \widehat{pob\_urb_i}$$

Como lo hemos hecho para el modelo anterior, empezamos con la significatividad global del modelo. El $p - valor = 1{,}271e - 0{,}6 < 0{,}5$ entonces el modelo sigue siendo significativo en su conjunto. Si nos fijamos en la significatividad individual, la población en edad de trabajar pasó a ser significativa al 10 %. La balanza comercial dejó de ser significativa. Eso sí la calidad de ajuste del modelo no se conservó pasando del 0.3699 al 0.3208. Ahora el modelo explica un poco más del 32 % de la varianza de la tasa de inmigración. Tenemos que verificar las hipótesis básicas que implican la estimación por MCO, antes de poder aceptar el modelo.

III.3.6. Pruebas de hipótesis

1. **De la forma functional (Test de Ramsey):**

   Hemos estimado este segundo modelo para resolver el problema de mala especificación del primer modelo. Aquí tenemos un p-valor de la prueba de Ramsey igual a 0.4537 superior a cualquier nivel de significación. El modelo está bien especificado, podemos pasar a las otras pruebas de hipótesis.

```
        RESET test

data:  MCO_2
RESET = 0.79755, df1 = 2, df2 = 88, p-value = 0.4537
```

Figura 19: Test de Reset Ramsey

2. **Multicolinealidad (Test VIF):**

   El segundo criterio consiste en verificar la existencia de multicolinealidad entre las variables explicativas. Para ello utilizaremos el criterio del Factor de Inflación de la Varianza (VIF). No vamos a explicar aquí cómo se calcula, pero la regla de decisión es la siguiente: si el VIF es superior a 10 habrá problema de multicolinealidad. En el caso presente, los VIF de cada una de las variables son inferiores al 10, entonces no existe ningún problema de multicolinealidad.

```
> vif(MCO_2)
        gini pob_edad_trab gastopub_educ           bc       pob_urb
    1.223972      1.569565      1.133070     1.440636      1.758562
```

Figura 20: VIF test

3. **Homocedasticidad (Test de Breusch Pagan):**

   El segundo criterio consiste en comprobar la homocedasticidad de los residuos, es decir que los residuos del modelo tengan una varianza constante en el tiempo. En caso contrario decimos que hay un problema de heterocedasticidad. Los residuos del modelo pueden ser heterocedásticos por cuatro razones. La primera sería que la forma funcional no conviene, lo que podemos rechazar en nuestro caso al comprobarlo con la prueba de Reset Ramsey. En segundo lugar, los valores atípicos podrían tener la culpa, pero antes de empezar el análisis fuimos quitándolos para que no provocaran este tipo de problemas. Por último, podrían ser las variables explicativas, la causa de heterocedasticidad y si no es así, consideramos la causa no identificable.

La primera etapa es comprobar la existencia de heterocedasticidad gracias a la prueba de Breusch Pagan. La prueba supone una hipótesis nula que considera los residuos como homocedásticos si se acepta, y a una hipótesis alternativa donde los residuos son heterocedásticos.

```
> bptest(MCO_2)

        studentized Breusch-Pagan test

data:  MCO_2
BP = 7.8568, df = 5, p-value = 0.1643
```

Figura 21: Test de Breusch Pagan

A la vista de los resultados, es decir mirando el p-valor igual a 0.1643, aceptamos la hipótesis nula de homocedasticidad de los residuos del modelo con cualquier nivel de significación del 1 %, 5 % y 10 %. El hecho de que aceptamos la homocedasticidad de los residuos no significa que no haya cierta heterocedasticidad. A pesar de que no sea relevante para nuestro análisis en concreto, el apartado que sigue estudia cuáles son las causas de la heterocedasticidad de los residuos, solo a nivel informativo.

El anexo 11 proporciona una representación gráfica de los residuos de cada variable explicativa estimada. Podemos suponer con estas gráficas que las variables responsables de la heterocedasticidad son la población en edad de trabajar y la balanza comercial al observar curvaturas redondeadas y no planas. No obstante, se tiene que verificar la significatividad de estas curvaturas con una prueba estadística. El resultado de esta prueba nos indica que la única variable culpable de la heterocedasticidad es la balanza comercial al ser significativa al test de Tukey. En el caso de que el modelo no hubiera pasado el test de Breusch Pagan, una de las soluciones hubiera sido eliminar la balanza comercial en la estimación del nuevo modelo y comprobar si se resuelve el problema.

```
> residualPlots(MCO_2)
               Test stat Pr(>|Test stat|)
gini              0.0469          0.96273
pob_edad_trab     0.0116          0.99080
gastopub_educ    -1.1441          0.25564
bc                1.8630          0.06576 .
pob_urb          -0.7835          0.43539
Tukey test       -1.2629          0.20662
```

Figura 22: Culpable de la heterocedasticidad

4. **Datos influyentes (Distancia de Cook):**

   Lo último que nos queda hacer es estudiar las observaciones influyentes verificando que no hay observaciones que influyen demasiado en la regresión. Se puede estudiar la influencia de una observación calculando la distancia de Cook. Si la distancia es mayor que la unidad entonces se tendrá que eliminar esa observación de la muestra porque es demasiado influyente. El plot siguiente nos permite tomar esta decisión de manera sencilla. El eje de ordenadas muestra la distancia de Cook y podemos observar que ninguna observación alcanza una distancia mayor del 0.09. Podemos concluir que no estamos frente a un problema de datos influyentes.

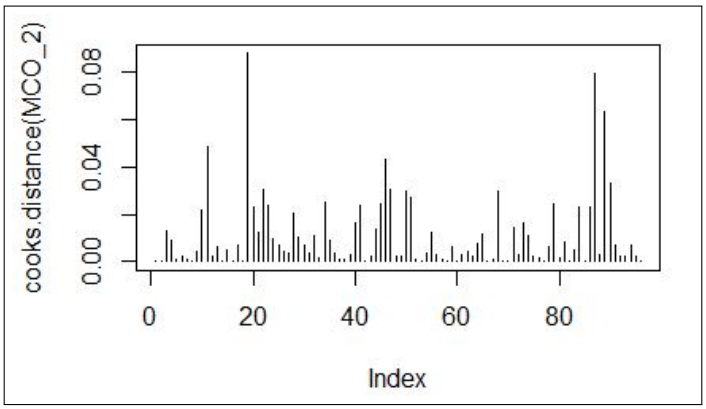


Figura 23: Distancia de Cook

III.3.7. Funciones útiles

Rstudio proporciona muchos paquetes para la estimación y la validación de los modelos. El modelo de regresión lineal está basado en hipótesis y la validación de estas hipótesis pasa especialmente con el diagnóstico de los residuos que se articulan alrededor de 6 gráficas. La función autoplot() del paquete fortify permite obtener esas gráficas con una sola línea de código. A continuación, podéis observar las gráficas que proporciona la aplicación de esa función a nuestro segundo modelo estimado. Iremos comentando la hipótesis que se trata en cada gráfica.

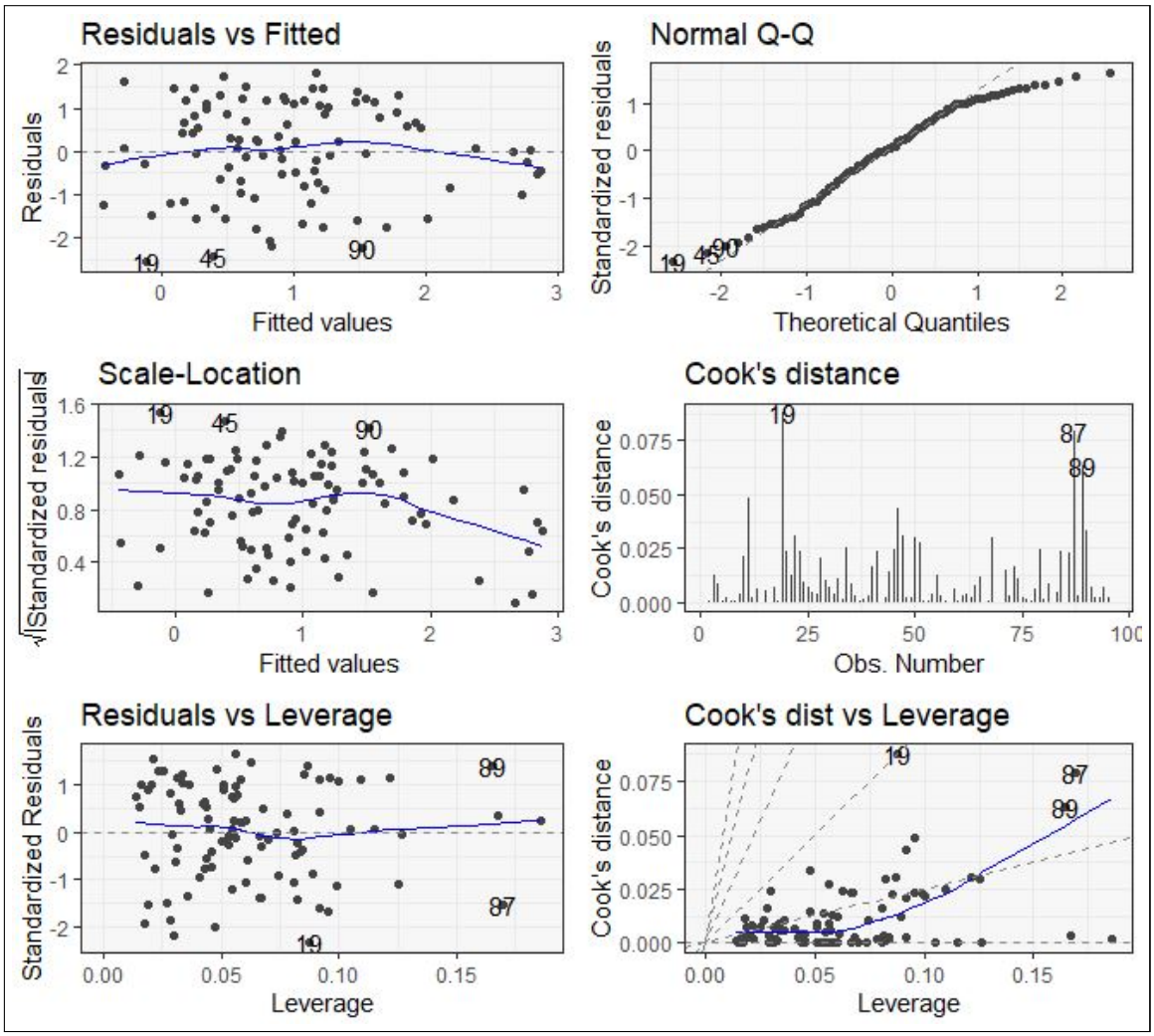


Figura 24: Validación gráfica de las hipótesis básicas de una regresión lineal

La hipótesis de homocedasticidad o constancia de la varianza está puesta de manifiesto en las gráficas **Residuals vs Fitted** y **Scale-Location**. En efecto, se acepta la hipótesis de homocedasticidad cuando no se observa ninguna tendencia particular en la dispersión de los puntos representando a los residuos. La gráfica Scale-Location persigue el mismo objetivo, pero con los residuos estandarizados. La hipótesis de normalidad de los residuos se verifica analizando la gráfica **Normal Q-Q**. Los residuos seguirán una ley normal siempre y cuando los puntos siguen el trazo oblicuo punteado. Por fin, las 3 últimas gráficas tratan de estudiar si existen puntos influyentes **Cook's distance**, provocando un efecto palanca como se puede observar sencillamente en la gráfica **Cook's dist vs Leverage**.

Una validación de esas hipótesis analíticamente es importante para confirmar de manera robusta lo que se observa visualmente en las gráficas. En vez de proceder con pruebas individuales el paquete gvlma, abreviación de Global Validation of Linear Models Assumptions informa automáticamente de la validez de las hipótesis de un modelo lineal en general. La captura siguiente nos enseña el resultado para la estimación de nuestro segundo modelo y, como lo pudimos prever, las hipótesis se cumplen para nuestro modelo MCO_2.

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = .)

                     Value p-value                Decision
Global Stat        6.8341  0.1449 Assumptions acceptable.
Skewness           2.5339  0.1114 Assumptions acceptable.
Kurtosis           2.5073  0.1133 Assumptions acceptable.
Link Function      1.6901  0.1936 Assumptions acceptable.
Heteroscedasticity 0.1029  0.7484 Assumptions acceptable.
```

Figura 25: gvlm() para MCO_2

III.3.8. Estimación del modelo 3

Vamos a estimar un tercer modelo para ver si podemos mejorar el anterior. En este modelo cambiaremos la especificación porque añadiremos logaritmos a todas las variables explicativas, entonces será una forma funcional log-log. Además, quitaremos la variable balanza comercial porque era la única variable no significativa anteriormente y era una de las explicaciones de la heterocedasticidad de los residuos.

La especificación del tercer modelo es la siguiente:

$$log(tasa\_migrantes_i) = \beta_0 + \beta_1 \cdot log(gini_i) + \beta_2 \cdot log(pob\_edad\_trab_i) + ... \\ ... + \beta_3 \cdot log(gastopub\_educ_i) + \beta_5 \cdot log(pob\_urb_i) + u_i$$

```
Call:
lm(formula = log(tasa_migrantes) ~ log(gini) + log(pob_edad_trab) +
    log(gastopub_educ) + log(pob_urb), data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4089 -0.8107  0.1323  1.0166  1.8045

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        18.9675     6.7231   2.821  0.00587 **
log(gini)          -2.5493     0.6150  -4.145  7.6e-05 ***
log(pob_edad_trab) -3.4681     1.4970  -2.317  0.02276 *
log(gastopub_educ)  0.9600     0.3463   2.772  0.00675 **
log(pob_urb)        1.0445     0.3276   3.188  0.00196 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.133 on 91 degrees of freedom
Multiple R-squared:  0.3154,    Adjusted R-squared:  0.2853
F-statistic: 10.48 on 4 and 91 DF,  p-value: 4.984e-07
```

Figura 26: MODELO n°3: Estimación de un modelo log-log

El modelo estimado es el siguiente :

$$\widehat{log(tasa\_migrantes_i)} = 18{,}97 - 2{,}549 \cdot \widehat{log(gini_i)} - 3{,}468 \cdot \widehat{log(pob\_edad\_trab_i)} + ... \\ ... + 0{,}96 \cdot \widehat{log(gastopub\_educ_i)} + 1{,}044 \cdot \widehat{log(pob\_urb_i)}$$

El modelo sigue siendo significativo conjuntamente con un $p-valor = 4{,}984 < 0{,}5$. Podemos observar una mejora de la significatividad individual de todas las variables explicativas. Hablando de la calidad de ajuste, tenemos que utilizar el R^2 ajustado para poder comparar MCO_3 con MCO_2. A pesar de que los dos modelos

tengan la misma variable endógena *log(tasa_migrantes)* y un término independiente, no tienen el mismo número de regresores al haber quitado la balanza comercial en esta última regresión. Podemos observar una mejora muy leve del ajuste pasando de 0.283 a 0.285. Ahora tenemos que verificar si el hecho de eliminar a una variable explicativa y cambiar de forma funcional no altera el cumplimiento de las hipótesis básicas de un modelo lineal.

III.3.9. Pruebas de hipótesis

Para que el análisis fuera más accesible hemos explicado con "detalle" varias pruebas de hipótesis con los modelos anteriores. Ahora solo iremos comentando los resultados de esas pruebas para los modelos siguientes. En los anexos 12 y 13 tenéis las pruebas a las que vamos a hacer referencia a continuación.

Empezando por la prueba de Ramsey sobre la forma funcional, con una especificación en logaritmos tanto para la variable explicada como para las variables explicativas, volvemos a aceptar la hipótesis nula de una buena forma funcional. El p-valor es igual a 0.4763 mayor que cualquier nivel de significación. En cuanto al Factor de Inflación de la Varianza es decir la prueba VIF para comprobar la presencia de multicolinealidad, observamos que ningún coeficiente es mayor que 10 entonces no hay multicolinealidad. La prueba de Breusch-Pagan para verificar la homocedasticidad de los residuos nos da un p-valor del 0.06169 lo que nos permite aceptar la hipótesis nula de homocedasticidad solo con un nivel de significación del 1 % y 5 % pero la rechazaremos para el 10 %. Acabamos analizando la distancia de Cook, donde no se observa ningún dato particularmente influyente en la estimación.

A la vista de estos resultados, podemos aceptar la validez de la estimación del modelo MCO_3. Pero podríamos mejorar la modelización, estimando un último modelo dejando solo algunas variables explicativas en logaritmos y otras no, y estudiar las diferencias.

III.3.10. Estimación del modelo 4

Para la última estimación decidimos dejar en logaritmos solamente, el gasto público en educación junto con la tasa de inmigración que tratamos de explicar.

$$log(tasa\_migrantes_i) = \beta_0 + \beta_1 \cdot gini_i + \beta_2 \cdot pob\_edad\_trab_i + ... \\ ... + \beta_3 \cdot log(gastopub\_educ_i) + \beta_5 \cdot pob\_urb_i + u_i$$

```
Call:
lm(formula = log(tasa_migrantes) ~ gini + pob_edad_trab + log(gastopub_educ) +
    pob_urb, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3825 -0.7609  0.1453  0.9662  1.9092

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.191493   1.758135   2.384  0.01920 *
gini              -0.065702   0.016001  -4.106 8.77e-05 ***
pob_edad_trab     -0.053303   0.023774  -2.242  0.02739 *
log(gastopub_educ) 0.880275   0.350428   2.512  0.01377 *
pob_urb            0.022825   0.006861   3.327  0.00127 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.13 on 91 degrees of freedom
Multiple R-squared:  0.3191,    Adjusted R-squared:  0.2892
F-statistic: 10.66 on 4 and 91 DF,  p-value: 3.941e-07
```

Figura 27: MODELO n°4: Estimación del MCO_4

El modelo estimado es el siguiente :

$$\widehat{log(tasa\_migrantes\_i)} = 4{,}191 - 0{,}066 \cdot \widehat{gini_i} - 0{,}053 \cdot \widehat{pob\_edad\_trab_i} + ... \\ ... + 0{,}88 \cdot \widehat{log(gastopub\_educ_i)} + 0{,}023 \cdot \widehat{pob\_urb_i}$$

A primera vista podemos observar que las variables explicativas siguen siendo todas significativas, pero con un nivel de significación individual inferior al modelo anterior. Hay una mejora del R^2 que explica un 31.91 % de la varianza de la tasa de inmigración. Nos queda comprobar que esta última estimación sea válida antes de empezar la interpretación de los resultados.

III.3.11. Pruebas de hipótesis

El modelo MCO_4 tiene una buena forma funcional a la vista de la prueba de Ramsey. Tampoco hace frente a un problema de multicolinealidad cuando se observan los coeficientes VIF del anexo 14. Hablando de homocedasticidad de los residuos, el p-valor de la prueba de Breusch-Pagan se mejora ligeramente en comparación a la modelización anterior. Aceptamos entonces, la hipótesis de homocedasticidad de los residuos con un nivel de significación del 1 % y 5 %. Por fin, no estamos frente a un problema de datos influyentes en esta estimación.

IV. PRESENTACIÓN DE LOS RESULTADOS

IV.1. Introducción

En este apartado procederemos a una presentación y un resumen de los resultados de las cuatro estimaciones del apartado anterior. Seleccionaremos el mejor modelo e interpretaremos sus coeficientes mirando si las teorías, están en acuerdo con la evidencia empírica. Concretamente, observaremos el p-valor y el signo de los coeficientes para comprobar la significatividad y la coherencia del signo con lo anunciado por la teoría. En primer lugar, tenemos que elegir el mejor modelo. La tabla que sigue es un resumen de los cuatro modelos estimados con los resultados de las pruebas de hipótesis necesarias para aceptar el modelo. Esta tabla nos permitirá comparar los modelos de manera sencilla proporcionándonos toda la información necesaria para elegir el mejor.

TABLA 3: Resumen de las estimaciones por MCO

	MCO_1	MCO_2 (log-lin)	MCO_3 (log-log)	MCO_4
Variable explicada	*tasa_migrantes*	*log(tasa_migrantes)*	*log(tasa_migrantes)*	*log(tasa_migrantes)*
Variables explicativas	Gini gastopub_educ pob_edad_trab bc pob_urb	Gini gastopub_educ pob_edad_trab bc pob_urb	log(gini) log(gastopub_educ) log(pob_edad_trab) log(pob_urb)	Gini log(gastopub_educ) pob_edad_trab pob_urb
Ramsey	X	$\surd^{***}$	$\surd^{***}$	$\surd^{***}$
VIF	$\surd$	$\surd$	$\surd$	$\surd$
Breusch-Pagan	X	$\surd^{***}$	$\surd^{**}$	$\surd^{**}$
Culpables	0	$\simeq (bc)$	0	0
Cook	$\surd$	$\surd$	$\surd$	$\surd$
R^2	0.3699	0.3208	0.3154	0.3191
$R^2 ajustado$	0.3349	0.283	0.2853	0.2892
Número de variables significativas	4/5	4/5	4/4	4/4

Mirando la tabla 3, de entrada, podemos eliminar a MCO_1 como mejor modelo porque no cumple con las hipótesis básicas de buena forma funcional y de homocedasticidad de los residuos. Nos quedan entonces solo tres modelos que tienen una especificación en logaritmo lo que resolvió el problema de mala forma funcional. Además, como lo podemos observar, estos tres modelos cumplen con todas las hipótesis de una estimación por MCO. Entonces solo nos queda seleccionar, el modelo que tiene la mayor calidad de ajuste. Mirando al R^2 ajustado, porque la especificación lo permite, nos quedamos con la última estimación MCO_4.

IV.2. Interpretación de los coeficientes

IV.2.1. Para el modelo MCO_4

Antes de empezar la interpretación de los coeficientes recordamos la estimación del modelo MCO_4:

$$\widehat{log(tasa\_migrantes\_i)} = 4{,}191 - 0{,}066 \cdot \widehat{gini_i} - 0{,}053 \cdot \widehat{pob\_edad\_trab_i} + ... \\ ... + 0{,}88 \cdot \widehat{log(gastopub\_educ_i)} + 0{,}023 \cdot \widehat{pob\_urb_i}$$

Recordamos que es un modelo que intenta explicar la tasa de inmigración con cuatro variables explicativas que son el coeficiente de Gini, la población en edad de trabajar, el gasto público en educación y por fin la población urbana. Este modelo es significativo en su conjunto lo que significa que fue pertinente estimarlo. Su R^2 es igual a 0.3191 por lo que, con este modelo, solo el 32 % de las variaciones de la tasa de inmigración de los países está explicado. En cuanto a la significatividad individual de los regresores, todos son significativos al menos al 10 %. Ahora vamos a interpretar cada coeficiente estimado empezando con el coeficiente de Gini.

El coeficiente de Gini venía aportándonos información sobre la desigual repartición de la renta y la teoría suponía una relación decreciente entre la tasa de inmigración y el coeficiente de Gini (cf: apartado variables explicativas). En la estimación MCO_4, el coeficiente de Gini es significativo para cualquier nivel de significación y además tiene el signo negativo esperado. El modelo predice que, para un país, cuando su coeficiente de Gini aumenta de un punto porcentual, su tasa de inmigración disminuye de un 6.6 %. El resultado es coherente con la teoría.

La población en edad de trabajar, es una variable que recoge el porcentaje de la población activa de cada país con el objetivo de considerar un país como desarrollado, aquel que tiene una población activa mayor. Queríamos observar si el hecho de que haya más activos en un país podía influir en la tasa de inmigración de aquel país. En el apartado de justificación de las variables explicativas, se observaba una relación positiva entre ambas variables. Los países con un mayor porcentaje de activos

tenían un porcentaje de inmigrantes por encima de los demás. En nuestro modelo, la evidencia empírica nos proporciona un coeficiente negativo. Si la población activa aumenta de un punto porcentual la tasa de inmigración se reduce de un 5,3 % con un nivel de significación del 10 %. Ese resultado va en contra de la teoría desarrollada.

La tercera variable es el gasto público en educación especificada en logaritmos en aquella estimación. El modelo predice que, si un país aumenta de 1 % el gasto público en educación, su tasa de inmigración aumentará un 0.88 % con un margen de error del 10 %. El signo positivo del coeficiente es coherente con la teoría del Welfare Magnets (cf: apartado variables explicativas).

Por fin acabamos con la variable que recoge el porcentaje de población que vive en zonas urbanas. El coeficiente igual a 0.023 es significativo al 5 % y además tiene un signo positivo lo que esperábamos por la teoría. Si el porcentaje de población urbana aumenta de un punto porcentual en el país, el modelo predice que la tasa de inmigración aumentaría un 2.3 %.

Para confirmar si los signos de los coeficientes no dependen solo de esta estimación particular, vamos a resumir en una última tabla los resultados para todos los modelos estimados válidos.

IV.2.2. En general para todos los modelos estimados

De las nueve variables explicativas que tenían que explicar la tasa de inmigración, solo cinco fueron incluidas en las estimaciones. Tuvimos que eliminar "*pib_pc*" y "*pov*" por la correlación, pero también "*pob_desemp*" y "*ide*" por endogeneidad supuesta y valores atípicos respectivamente. De las cinco estimadas, cuatro son significativas y sólo "*gini*", "*gastopub_educ*" y "*pob_urb*" son coherentes con lo previsto por la teoría.

IV.3. Tabla resumen

TABLA 4: Resumen de los principales resultados

Hipótesis teórica	**Especificación (variables)**	**Evidencia empírica (significación y signo)**	**Resultado**
Un aumento del PIBpc, supone un aumento de la tasa de inmigración	**Pib per cápita**	no incluida (correlación fuerte)	-
A mayor tasa de desempleo, menor tasa de inmigración	**Desempleo, total**	no incluida (endogeneidad)	-
A mayor tasa de pobreza, menor es la tasa de inmigración	**Tasa de incidencia de la pobreza**	no incluida (correlación fuerte)	-
Una repartición desigual de la renta, disminuye la tasa de inmigración	**Índice de Gini**	Significativa y con signo negativo *(MCO_2/3/4)*	Coherente
A medida que un país se va desarrollando y aumenta su fuerza de trabajo, mayor es su tasa de inmigración	**Población entre 15 y 64 años de edad**	Significativa y con signo negativo *(MCO_2/3/4)*	Incoherente
El aumento del porcentaje del PIB dedicado a la educación aumenta la tasa de inmigración	**Gasto público en educación, total**	Significativa y con signo positivo *(MCO_2/3/4)*	Coherente
A mayor comercio entre los países, mayor será la tasa de inmigración	**Balanza comercial de bienes y servicios**	No significativa *(MCO_2)*	-
A mayor desarrollo y urbanización del país, mayor es la tasa de inmigrantes	**Población urbana**	Significativa y con signo positivo *(MCO_2/3/4)*	Coherente
A mayor movimiento de capital entre los países, mayor será la tasa de inmigración	**Inversión extranjera directa, salida neta de capital**	no incluida (outliers/correlación)	-

V. CONCLUSIÓN Y DISCUSIÓN DE LOS RESULTADOS

Con este análisis, tratábamos de explicar las diferencias en las tasas de inmigración de los países del mundo en 2015. Nos hemos enfocado en variables económicas para estudiar si determinantes económicos podrían afectar las migraciones. Hemos utilizamos el método de regresión lineal múltiple para estimar los modelos. De las cuatro estimaciones, encontramos un modelo explicando más o menos 30 % de la varianza de la tasa de inmigración para 109 países. Las variables influyendo a la tasa de inmigración de manera significativa son el índice de Gini, la población entre 15 y 64 años de edad, el gasto público en educación y por fin la población urbana.

El R^2 de nuestro mejor modelo acerca solo el 32 %. En este apartado trataremos de averiguar cuáles son las razones de este R cuadrado con el fin de justificar la mala cifra del coeficiente de ajuste. La primera razón podría ser que tratábamos de estudiar la inmigración con un enfoque solamente económico, entonces decidimos dejar a un lado a otras variables que explican también las diferencias entre países frente a las migraciones, ya sean culturales, históricas, ambiental, etcétera. La segunda razón podría venir directamente de la especificación de nuestra variable explicada, la tasa de inmigración. En efecto, las estadísticas sobre inmigrantes, suelen mezclar los inmigrantes de los refugiados. Es complejo, a la hora de recoger los datos, diferenciar el uno del otro, y sin embargo no es lo mismo. Los determinantes de la huida son distintos entre un refugiado y un inmigrante. Entonces, al tener una variable que recoge más información de la que queremos explicar, el análisis pudo alterarse. La tercera razón sería que, al estimar los modelos por MCO, tuvimos que dejar a un lado, a la variable desempleo, sospechadas de endogeneidad para cumplir con las hipótesis básicas y que los modelos estimados sean válidos. Estimar por Variables Instrumentales o, Mínimos Cuadrados en 2 Etapas, hubiera resuelto la endogeneidad. No lo hicimos en este trabajo y es uno de los límites del análisis. La cuarta razón es que tuvimos problemas de datos faltantes para algunas observaciones para el año 2015. Resolvimos el problema estimando la variable del año 2015 con una media de años anteriores. No es la mejor forma de estimar para el año 2015 y eso pudo dañar el resultado. Por fin, no incluimos como variable explicativa el flujo de inmigración para tratar de explicar el stock como lo hacen en otros trabajos. Todo esto, son pistas para explicar el flojo ajuste de nuestro modelo a la realidad.

Mi trabajo viene a completar otros estudios hechos sobre este tema, confirmando la significatividad de algunas variables en la explicación de la tasa de inmigración. Además de ello, comprobamos que las teorías se ajustan a la evidencia empírica del año 2015. De las cuatro variables estimadas, tres tienen un signo coherente con la teoría. La población entre 15 y 64 años, es decir la población activa, es la única variable con un signo opuesto a lo previsto. Habíamos supuesto que un país con una población activa importante era un país en desarrollo, por lo que atraía a una inmigración extranjera en búsqueda de trabajo. Podemos intentar explicar la relación negativa entre población activa y la tasa de inmigración de la siguiente forma. Los países que consideramos en vías de desarrollo en 2015 son países que intentan salir de la pobreza, y suelen tener una población joven. De allí viene el porcentaje mayor en población activa. Pero, una población activa, no significa una población con trabajo y oportunidades.

Si me permiten, me pareció importante mencionar las dificultades encontradas a lo largo del trabajo. La primera, sería en la búsqueda de la información justa y adecuada en relación con el tema cuando tenemos una información masiva gracias a internet. Segundo, sería el tratamiento de los datos con la herramienta Rstudio, desde la importación, pasando por la limpieza y toda la metodología econométrica. Las dificultades se encontraban en el desarrollo del RScript para la realización del análisis. Por fin, y por supuesto, la búsqueda del mejor modelo posible no es algo sencillo cuando tienes que estimar muchos modelos antes de encontrar el que se ajusta mejor sin saber cuándo parar.

En un principio, la construcción de un modelo econométrico permite hacer previsiones sobre el futuro, con un margen de error por supuesto. Ahora bien, me parece importante recordar que existen factores y fenómenos exteriores, imprevisibles, que pueden comprometer todas esas previsiones. Es el caso de todas las crisis que conoce este mundo ya sean las guerras, el cambio climático o las crisis sanitarias como la que estamos enfrentando con la covid-19.

VI. BIBLIOGRAFÍA

Referencias

[1] Jean-Christophe Delmas. *Dico Atlas de l'Histoire du Monde*. Belin: éducation, 2018.

[2] Iván M. Torres and Guillem L. Casasnovas. Los determinantes de la inmigración internacional en españa/the determinants of international immigration in spain. *Investigaciones Regionales*, (9):23, 2006.

[3] Nicolas Péridy. Un modèle généralisé des déterminants des migrations internationales: Application aux migrations des pays méditerranéens vers l'ue. *Revue économique*, 61(6):981–1010, 2010.

[4] Larry A. Sjaastad. The costs and returns of human migration. *Journal of Political Economy*, 70(5):80–93, 1962.

[5] John R. Harris and Michael P. Todaro. Migration, unemployment and development: A two-sector analysis. *The American Economic Review*, 60(1):126–142, 1970.

[6] George J. Borjas. Self-selection and the earnings of immigrants. *The American Economic Review*, 77(4):531–553, 1987.

[7] George J. Borjas. Immigration and welfare magnets. *Journal of Labor Economics*, 17(4):607–637, 1999.

[8] Douglas S. Massey, Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J. Edward Taylor. Theories of international migration: A review and appraisal. *Population and Development Review*, 19(3):431–466, 1993.

[9] Oded Stark and J. Edward Taylor. Migration incentives, migration types: The role of relative deprivation. *The Economic Journal*, 101(408):1163–1178, 1991.

[10] Claude Bouet. Problèmes actuels de main-d'œuvre au gabon. conditions d'une immigration contrôlée. 1978.

[11] J. A. Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

[12] Hadley Wickham y Garrett Grolemund. R for data science. `https://r4ds.had.co.nz/index.html`, 2014.

[13] Frank E. Grubbs. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1):27–58, 1950.

[14] Bernard Rosner. Percentage points for a generalized esd many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.

VII. ANEXOS

Anexo 1: Código de obtención de datos

```
#----------1.Cargamos las librerias necesarias:
library("wbstats")
library("tidyverse")
library("rio")

#----------2.descargamos nuestras variables para el analisis:

#primero nuestra variable explicada Y : Volúmenes internacionales
de migrantes (% de la población) en el ano 2015
dfY <- wb(country = "all",indicator = "SM.POP.TOTL.ZS",lang = "es",
startdate = 2000,enddate = 2018,removeNA = FALSE) %>% filter(date
== 2015)
dfY <- dfY %>% select(1,pais = country,tasa_migrantes = value)

#nuestras variables explicativas X1 : PIB per cápita, PPA ($ a
precios internacionales actuales) en el ano 2015.
dfX1 <- wb(country = "all",indicator = "NY.GDP.PCAP.PP.CD",lang =
"es", startdate = 2000,enddate = 2018,removeNA = FALSE)%>% filter
(date == 2015)
dfX1 <- dfX1 %>% select(1,pais = country, pibpc= value)

#X2 :Desempleo, total (% de la población activa total)
dfX2 <- wb(country = "all",indicator = "SL.UEM.TOTL.ZS",lang = "es"
, startdate = 2000,enddate = 2018,removeNA = FALSE)%>% filter(date
== 2015)
dfX2 <- dfX2 %>% select(1,pais = country, pob_des= value)

#X3:Tasa de incidencia de la pobreza
dfX3 <- wb(country = "all",indicator = "SI.POV.NAHC",lang = "es",
startdate = 2000,enddate = 2018,removeNA = FALSE)%>% filter(date ==
2015)
dfX3 <- dfX3 %>% select(1,pais = country, pov= value)

#Hacemos lo mismo con las 9 variables ...

#---------3.juntamos por la variable "country" para formar mi base
de datos.
df <- left_join(dfY,dfX1)
df <- left_join(df,dfX2)
df <- left_join(df,dfX3)  #Hacemos lo mismo con las 9 variables ...
```

Anexo 2: Limpieza de datos

```
#-----------4.Quitamos a los paises que no son revelantes: AND, INX, ASM,
ATG, BMU, VGB, CYM, CHI, CUB, CUW, DJI, DMA, ERI, FRO, FJI, PYF, GIB, GRL
, GRD, GUM, IMN, KIR, PRK, XKX, LIE, MHL, FSM, MCO, NRU, NCL, MNP, PLW,
PNG, SYC, SLE, SXM, SOM, KNA, MAF, SDN, SUR, SYR, TZA, TTO, TKM, TCA, TUV
, VEN, VIR, YEM, CSS, PSS. porque no tienen la informacion suficiente.

df <- df %>% filter(iso3c != "AND" & iso3c != "INX" & iso3c !="ASM") #
etc

#-----------5.Ademas las 45 primeras lineas recogen observaciones que no
son paises sino continentes o zonas del mundo.
df <- df %>% slice(45:n())

#-----------6.Redondeamos a dos digitos para facilitar el trabajo con los
datos.
df[3:12]=round(df[3:12], digits = 2)
```

Anexo 3: Manipulación de datos

```
#-----------1.Prediccion de gini, pov y gastopub-educ en el año 2015:

#Volvemos a descargar cada una de las tres variables pero sin filtrar el
año 2015 esta vez.(aqui para pov)
dfx3 <- wb(country = "all",indicator = "SI.POV.NAHC",lang = "es",startdat
e = 1980,enddate = 2018,removeNA = TRUE)
dfx3 <- dfx3 %>% select(1,ano = date, pais = country, pov= value)

#Calculamos la media para cada país.
dfx3 <- dfx3 %>% group_by(pais) %>% summarise(pov1 = mean(pov))
dfx3$pov1=round(dfx3$pov1, digits = 2)

#Reemplazamos los NA's por las medias:
a <- c(4,9,12,13,14,17,20,23,24,25,28,29,30,34,35,36,37,39,42,43,51,56,65
,66,69,79,85,87,89,90,95,97,98,99,100,101,102,105,106,111,115,121,122,124
,126,127,129,133,135)

df$pov1[a]=df$pov[a]
```

Anexo 4: Observaciones agregadas

```
#---------2.Valores faltantes buscado en la base de datos de la OCDE y
añadidos de la siguiente forma :

df$pov[5] <- 10.5
df$pov[6] <- 8.7
df$pov[10] <- 9.8
df$pov[16] <- 25.7
df$pov[22] <- 14.2
df$pov[33] <- 5.5  #y algunos más
```

Anexo 5: Estadísticas descriptivas de la base de datos df

```
> summary(df)
    iso3c               pais            continente        tasa_migrantes       pib_pc
 Length:109         Length:109         Length:109         Min.   : 0.070   Min.   :   764.2
 class :character   class :character   class :character   1st Qu.: 1.330   1st Qu.:  6082.8
 Mode  :character   Mode  :character   Mode  :character   Median : 3.790   Median : 14455.0
                                                          Mean   : 6.891   Mean   : 19991.0
                                                          3rd Qu.: 9.680   3rd Qu.: 28823.8
                                                          Max.   :43.960   Max.   :103750.8
   pob_desemp          pov              gini       pob_edad_trab   gastopub_educ         bc
 Min.   : 0.300   Min.   : 0.40   Min.   :25.40   Min.   :47.26   Min.   :1.47   Min.   :-87.080
 1st Qu.: 4.210   1st Qu.:11.70   1st Qu.:32.10   1st Qu.:60.84   1st Qu.:3.58   1st Qu.:-11.670
 Median : 6.670   Median :21.60   Median :36.50   Median :65.15   Median :4.51   Median : -2.470
 Mean   : 7.851   Mean   :25.38   Mean   :38.00   Mean   :63.41   Mean   :4.49   Mean   : -5.917
 3rd Qu.: 9.540   3rd Qu.:35.00   3rd Qu.:43.10   3rd Qu.:67.17   3rd Qu.:5.35   3rd Qu.:  2.990
 Max.   :26.070   Max.   :66.60   Max.   :61.71   Max.   :74.16   Max.   :7.55   Max.   : 34.030
    pob_urb            ide
 Min.   :12.08   Min.   :-6.62
 1st Qu.:49.44   1st Qu.: 0.10
 Median :63.09   Median : 0.59
 Mean   :61.23   Mean   : 2.50
 3rd Qu.:77.36   3rd Qu.: 1.74
 Max.   :97.88   Max.   :65.21
```

Anexo 6: Test de Grubbs

```
> grubbs.test(`df`$"gini",type=10, two.sided = FALSE)

        Grubbs test for one outlier

data:  df$gini
G = 3.03736, U = 0.91379, p-value = 0.1059
alternative hypothesis: highest value 61.71 is an outlier
```

Anexo 7: Test de Rosner para el PIB per cápita

```
> ##Rosner test (X1)
> rosnerTest(`df`$pib_pc, k = 3, alpha = 0.05)

Results of Outlier Test
-------------------------

Test Method:                     Rosner's Test for Outliers

Hypothesized Distribution:       Normal

Data:                            df$pib_pc

Sample Size:                     109

Test Statistics:                 R.1 = 4.566033
                                 R.2 = 3.013860
                                 R.3 = 2.844388

Test Statistic Parameter:        k = 3

Alternative Hypothesis:          Up to 3 observations are not
                                 from the same Distribution.

Type I Error:                    5%

Number of Outliers Detected:     1

  i    Mean.i      SD.i     Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 19991.02 18344.10 103750.78      64 4.566033   3.413225    TRUE
2 1 19215.47 16537.12  69056.05      49 3.013860   3.410133   FALSE
3 2 18749.67 15887.19  63938.99      97 2.844388   3.407006   FALSE
```

Anexo 8: Coeficientes de Skewness y Kurtosis

```
> round(sapply(df1[,5:12],skewness),2)
      pib_pc  pob_desemp         pov        gini pob_edad_trab gastopub_educ
        1.00        0.89        0.70        0.34       -0.78        0.20
          bc     pob_urb
       -0.75       -0.37
> round(sapply(df1[,5:12],kurtosis),2)
      pib_pc  pob_desemp         pov        gini pob_edad_trab gastopub_educ
        0.18        0.45       -0.46       -0.78       -0.20       -0.60
          bc     pob_urb
        1.80       -0.49
```

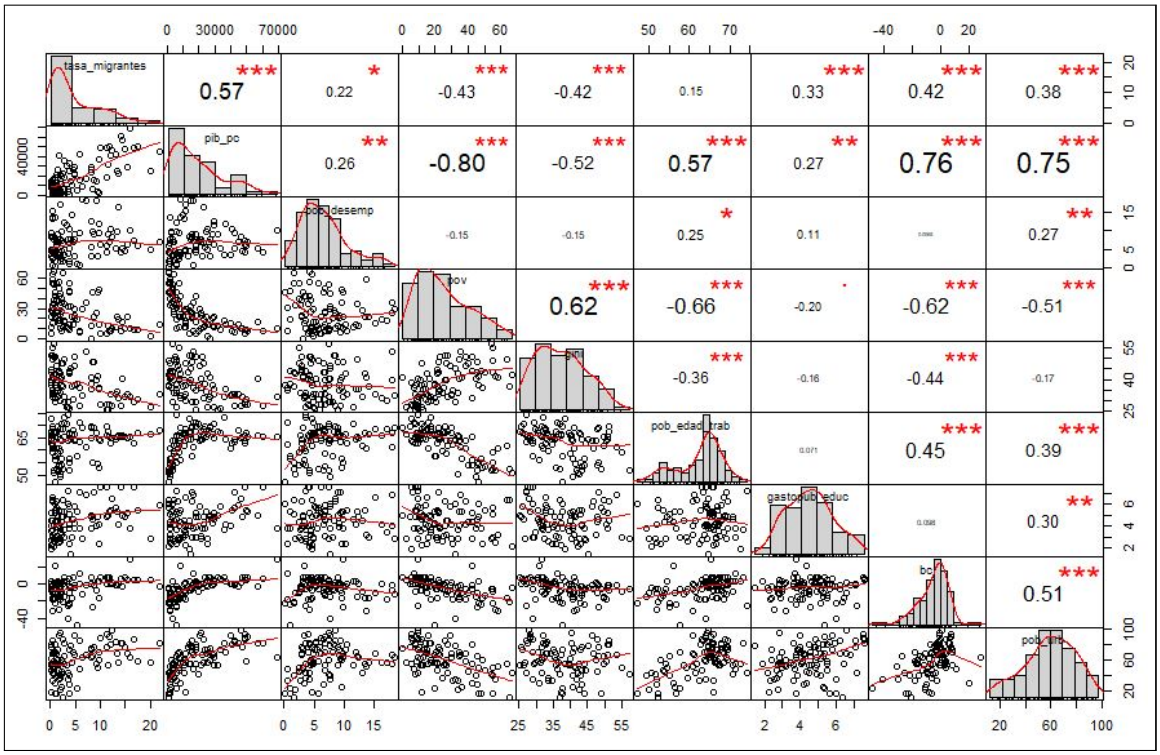
Anexo 9: Test de Shapiro-Wilk

```
> shapiro.test(df1$pib_pc)

        Shapiro-Wilk normality test

data:  df1$pib_pc
W = 0.88841, p-value = 6.493e-07

> shapiro.test(df1$pob_desemp)

        Shapiro-Wilk normality test

data:  df1$pob_desemp
W = 0.93711, p-value = 0.0001779

> shapiro.test(df1$pov)

        Shapiro-Wilk normality test

data:  df1$pov
W = 0.93144, p-value = 8.434e-05

> shapiro.test(df1$gini)

        Shapiro-Wilk normality test

data:  df1$gini
W = 0.96741, p-value = 0.0172
```

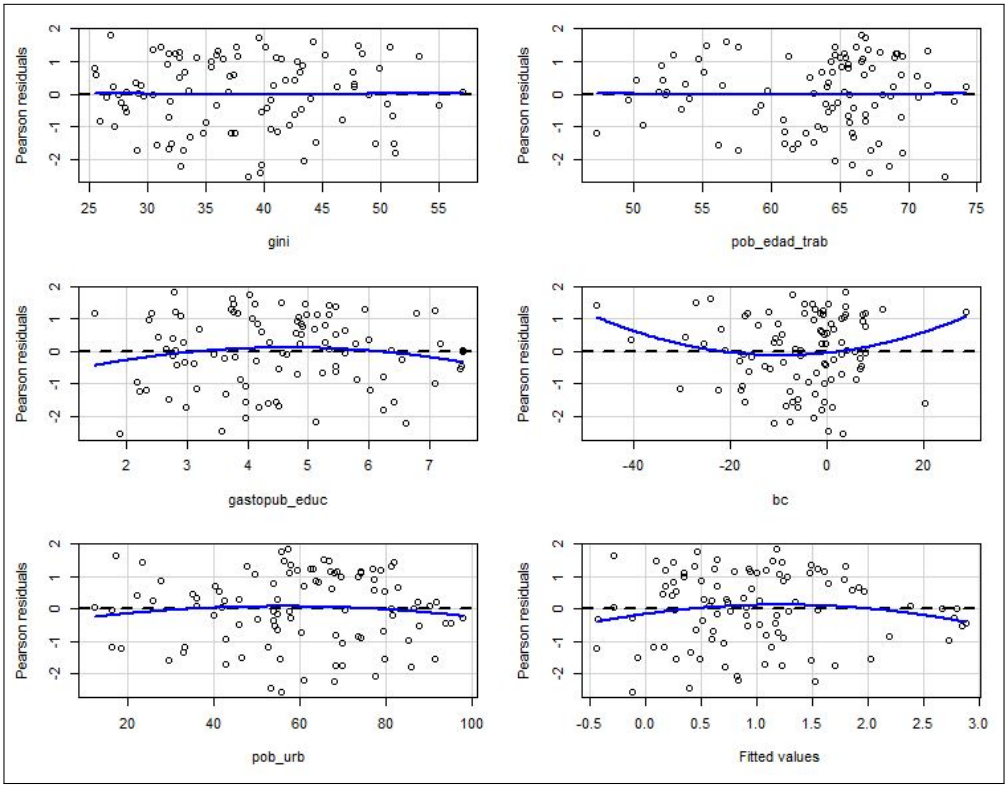
```
> shapiro.test(df1$pob_edad_trab)

        Shapiro-Wilk normality test

data:  df1$pob_edad_trab
W = 0.92375, p-value = 3.207e-05

> shapiro.test(df1$gastopub_educ)

        Shapiro-Wilk normality test

data:  df1$gastopub_educ
W = 0.98112, p-value = 0.1823

> shapiro.test(df1$bc)

        Shapiro-Wilk normality test

data:  df1$bc
W = 0.94654, p-value = 0.0006613

> shapiro.test(df1$pob_urb)

        Shapiro-Wilk normality test

data:  df1$pob_urb
W = 0.97572, p-value = 0.07156
```

Anexo 10: Matriz de correlación 2

```
> library(PerformanceAnalytics)
> mydata <- df1[,c("tasa_migrantes","pib_pc", "pob_desemp", "pov", "gini",
 "pob_edad_trab", "gastopub_educ", "bc", "pob_urb")]
> chart.Correlation(mydata, histogram=TRUE, pch=19,method = c("spearman"))
```



Anexo 11: Modelo 2 residualplot



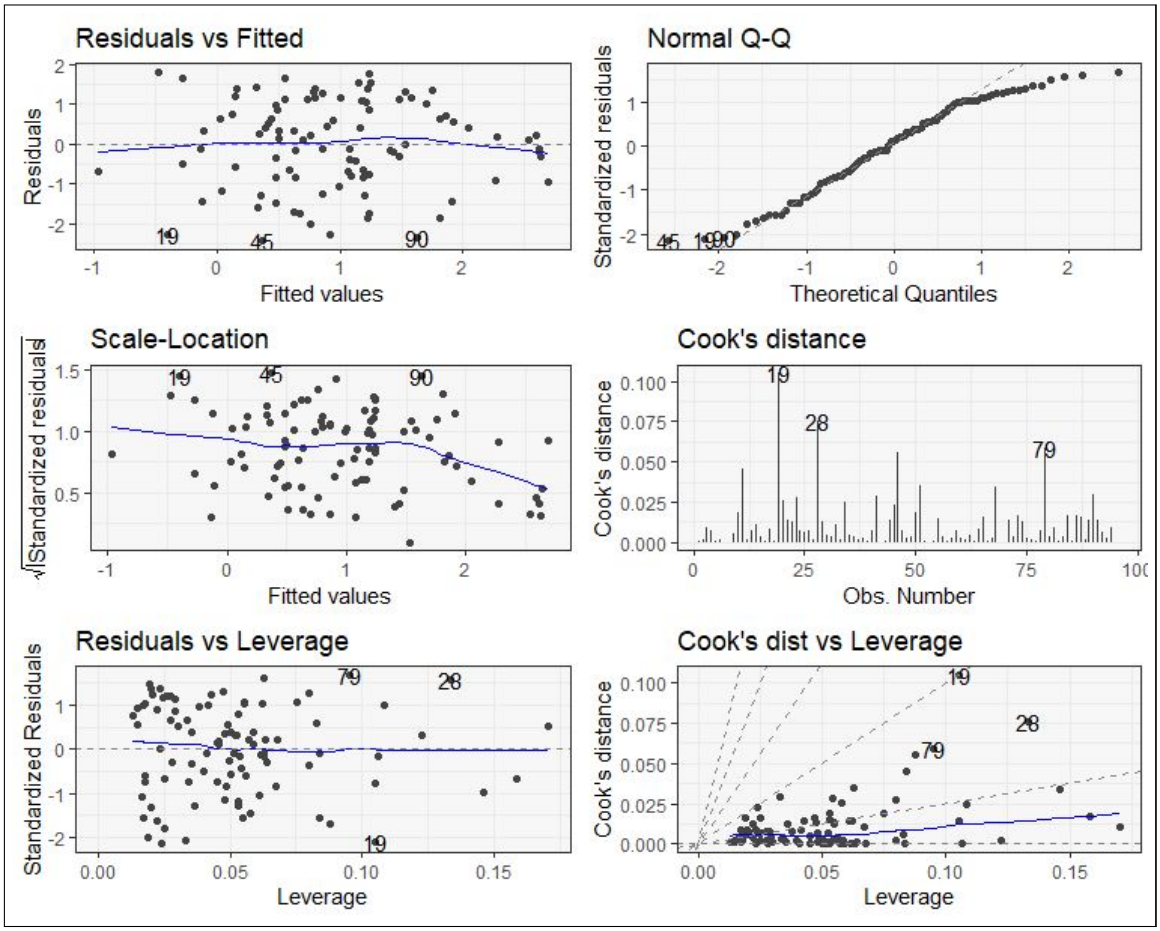
Anexo 12: Modelo 3: pruebas de hipótesis básicas analíticamente

```
> reset(MCO_3)

        RESET test

data:  MCO_3
RESET = 0.27359, df1 = 2, df2 = 89, p-value = 0.7613
```

```
> bptest(MCO_3)

        studentized Breusch-Pagan test

data:  MCO_3
BP = 8.9764, df = 4, p-value = 0.06169
```

```
> vif(MCO_3)
        log(gini) log(pob_edad_trab) log(gastopub_educ)        log(pob_urb)
         1.180221          1.676079          1.059354          1.530206
```

Anexo 13: Modelo 3: pruebas de hipótesis básicas gráficamente



Anexo 14: Modelo 4: pruebas de hipótesis básicas analíticamente

```
> reset(MCO_4)

        RESET test

data:  MCO_4
RESET = 0.74799, df1 = 2, df2 = 89, p-value = 0.4763
```

```
> bptest(MCO_4)

        studentized Breusch-Pagan test

data:  MCO_4
BP = 8.2983, df = 4, p-value = 0.08124
```

```
> vif(MCO_4)
             gini      pob_edad_trab log(gastopub_educ)           pob_urb
         1.160983           1.546924           1.090639          1.467508
```

Anexo 14: Modelo 4: pruebas de hipótesis básicas gráficamente