

Grado en Estadística

Título: Determinantes de la Inversión Extranjera Directa

Autor: Valentina Ochoa Chilito

Director: Jordi Suriñach

Departamento: Econometría, Estadística y Economía Aplicada

Convocatoria: Extraordinaria Junio 2021



Resumen

El método de MCO nos permite estimar los parámetros de distintos modelos econométricos a través de la minimización de la suma de las distancias verticales de las observaciones de la muestra y las respuestas del modelo. No obstante, según el modelo utilizado los resultados serán más eficientes o no. Cada modelo tiene unas ciertas condiciones que en el caso de no cumplirse hace que los estimadores estén sesgados, sean inconsistentes y poco fiables. Este trabajo analiza y compara el uso de los MCO para una regresión lineal múltiple con datos de corte transversal y de datos panel. Además, se introduce la aplicación de los MCO para la estimación de los parámetros de un modelo gravitacional sencillo.

Palabras clave: Inversión Extranjera Directa, Mínimos cuadrados ordinarios, Modelo gravitacional

Clasificación AMS:

62-07 Data analysis

62P25 – Applications to economics

Title: Determinants of Foreign Direct Investment

Abstract

The OLS method allows us to estimate the parameters of different econometric models by minimising the sum of the vertical distances of the sample's observations and the model responses. However, depending on the model used, the parameters estimations will be more efficient. Each model has certain conditions that, if not met, would make the estimators biased, inconsistent and unreliable. This paper analyses and compares the use of OLS for multiple linear regression with cross-sectional and panel data. In addition, the application of OLS to the estimation of the parameters of a simple gravity model is introduced.

Key words: Foreign Direct Investment, OECD, Ordinary Least Squares, Gravity Model

ÍNDICE

1. Introducción	4
2. Metodología	6
2.1 Modelo de regresión múltiple	7
2.1.2. Bondad del ajuste	10
2.1. Modelo de regresión lineal con datos panel	11
2.2. Modelo gravitacional	11
3. Base de datos	13
3.1. Análisis descriptivo	16
3.1.2. Estadísticos descriptivos variables de corte transversal 2017	18
3.2. Estadística descriptiva de los datos panel	23
3.3. Estadística descriptiva de la base de datos del modelo gravitacional	25
4. Selección del modelo y estudio de bondad	26
4.1. Resultados modelo de regresión lineal múltiple por MCO para datos de corte transversal	26
4.2. Resultados modelo de regresión lineal múltiple por MCO para datos panel	30
4.3. Resultados del modelo gravitacional	33
6. Bibliografía	36
Anexos	38

Gráficos

Gráfico 1. Matriz de correlación.....	22
Gráfico 2. Evolución de los flujos de entrada de IED 2007-2017	23
Gráfico 3. Evolución de los flujos de entrada de IED 2007-2017	24
Gráfico 4. Evolución de los flujos de entrada de IED 2007-2017	24
Gráfico 5. Relación entre los países en base a su PIB y a la IED recibida	25

Ilustraciones

Ilustración 1. Gráfico de residuos Homoscedasticidad vs. Heteroscedasticidad.	9
Ilustración 2. Histograma para las variables numéricas.....	21
Ilustración 3. Análisis de residuos del modelo 2	28
Ilustración 4. Análisis de residuos del modelo 3	29
Ilustración 5. Resultado de regresión lineal por MCO datos panel 1	30
Ilustración 6. Resultado de regresión lineal por MCO datos panel	31
Ilustración 7. Resultado de regresión lineal por MCO datos panel 3	32
Ilustración 8. Resultados del Gravity Model	33

Tablas

Tabla 1. Variables de la Base de datos 3	16
Tabla 2. Comparativa variables absolutas y variables relativas	17
Tabla 3. Estadísticos descriptivos de las variables sujetas a estudio	18
Tabla 4. Frecuencia absoluta y relativa para la variable dummy pertenecer a la UE	22
Tabla 5. Frecuencia absoluta y relativa para la variable dummy er un país en vías de desarrollo	23
Tabla 6. Resultados de los modelos de regresión lineal por MCO	26

Lista de acrónimos

IED – Inversión Extranjera Directa

MCO – Mínimos Cuadrados Ordinarios

MLR – Regresión lineal múltiple

OCDE – Organización para la Cooperación y el Desarrollo Económicos

PPA - Paridad del Poder Adquisitivo

SPACs – Empresas de adquisición con propósito especial

UE – Unión Europea

1. Introducción

La Inversión Extranjera Directa (IED) es una de las partidas más importantes de la inversión de un país y por consiguiente tiene una relación de causalidad con el crecimiento económico del mismo (Moudatsou & Kyrkilis). Este tipo de inversión toma una gran relevancia a partir de los años 90, momento en el que China comenzó su “*Go out policy*”, y en el que los flujos crecieron de forma exponencial, alcanzando un máximo a nivel mundial de aproximadamente 3 billones USD de entrada neta de capital en 2005 (Banco Mundial, n.d). A pesar de esto, en los últimos años la tendencia se ha revertido. En 2019, última cifra actualizada por el Banco Mundial, la entrada neta de capital por IED fue de 1,63 billones USD (Banco Mundial, n.d).

Esta evolución de flujos de IED nos lleva a plantearnos la siguiente pregunta, ¿cuáles han sido las causas que explican este crecimiento? Este trabajo introduce las principales variables que explican la razón por la cuál hay países que perciben un mayor nivel de IED respecto a su economía en comparación con otros. En este trabajo se analizará una muestra de 29 países pertenecientes a la Organización para la Cooperación y el Desarrollo Económicos (OCDE) y se estudiará la significación estadística de distintas variables macroeconómicas. Por ejemplo, el nivel de inflación o la relevancia de ciertos sectores en la economía como el sector industrial, agrario o sector inmobiliario, al igual que el nivel de impuestos que se imponen a corporaciones. Además, se estudiará las distintas condiciones que pueden darse para poder establecer distintos grupos que se diferencien estadísticamente de otros. Esto último nos ayudará a determinar si hay diferencia significativa entre pertenecer al mercado interior de la Unión Europea o ser un país en vías de desarrollo, entre otras cosas.

Para realizar dicho estudio se llevará a cabo una regresión lineal múltiple, en la cuál se utiliza el método de Mínimos Cuadrados Ordinarios (MCO) para estimar los parámetros del modelo. Este método se basa en minimizar la suma de los cuadrados de los residuos y resolver las condiciones de primer orden respecto a los distintos parámetros a estimar. Este proceso es comúnmente resuelto por el método de los momentos (Faraway, 2004). Posteriormente, se revisará que la estimación por MCO cumpla los supuestos de Gauss Markov y así se pueda considerar el Estimador lineal insesgado óptimo. Esto se explicará y analizará más detalladamente en los apartados posteriores. Este método se aplicará para tres tipos de modelo distintos. El primero es el propio para datos de cortes transversal. El segundo modelo de regresión lineal incluye un análisis temporal de los distintos determinantes de la IED, a través de un modelo de regresión de efectos. Y por último se introduce un modelo gravitacional, comúnmente utilizado en los estudios de comercio internacional para explicar los flujos de entrada y de salida de la IED en nuestro caso.

Este trabajo de fin de grado consta de cinco secciones. En la primera parte se analiza la metodología que ha sido seguida para la realización de los distintos modelos econométricos. En el segundo apartado se introducen las tres bases de datos utilizadas y se realiza el análisis

descriptivo de las variables de estudio. En la tercera sección se realiza el estudio de los distintos modelos y se estudia su bondad introduciendo distintas pruebas estadísticas relevantes, así como los resultados e interpretaciones de cada uno de los tres modelos. Por último, se exponen las conclusiones del trabajo, así como sus limitaciones y posibles estudios posteriores

2. Metodología

La econometría es una de las ramas de la estadística que se centra en estudiar la relación causal de ciertas variables en relación con variables independientes. Es una herramienta que se ha utilizado comúnmente en el ámbito de la economía, aunque también se pueden observar análisis econométricos en otras especialidades como la medicina o ingeniería.

En relación con la economía y concretamente a lo que este trabajo se refiere, la explicación de los distintos niveles de IED que un país recibe viene dado por distintos factores, lo cuáles han de estar incluidos en su mayoría en la regresión. De lo contrario, si estos no se incluyen como variables independientes en la regresión, estarán “incluidos” en el término del error, lo que ocasionará que los estimadores de la regresión estén sesgados y no sean óptimos. Estos conceptos se explicarán más adelante, en el apartado de las implicaciones de las condiciones de Markov.

Teniendo en cuenta lo anteriormente explicado, durante la realización de este trabajo se ha optado por la realización de una regresión lineal múltiple (MLR), ya que, aunque la regresión simple introducida por Gauss en 1809 ofrece una estimación simple y sencilla de realizar a través de los MCO. Esta se basa en la asunción en que los errores han de seguir una distribución normal, la cuál no se satisface si se omiten variables relevantes (Faraway, 2004). Las dos condiciones que se han de dar para que no encontremos ante el caso de omisión de variables relevante es que alguna variable dependiente este correlacionada con la variable omitida y que dicha variable omitida sea un determinante de la variable dependiente y del modelo. Esta es una de las propiedades que ha de cumplir la estimación por MCO de MLR. En nuestro caso un causante de los cambios en las entradas de IED.

Con respecto al estudio de los factores explicativos de la IED podemos encontrar distintos estudios ya realizados, por ejemplo, el realizado por Mogrevo, el cuál analiza los determinantes para países de Latinoamérica en los años 90 utilizando un estudio de datos de panel. Se pudo concluir que, para dichos países durante los años 90, variables como los costes laborales no resultaron ser estadísticamente significativos. Otros estudios se enfocan en la realización de modelos gravitacionales que permiten estudiar la relación de inversión que hay entre los países, y explicar cómo se relacionan entre ellos y cuáles son las causas de que haya más o menor flujo intra-países. Este tipo de modelos ha permitido concluir, que al contrario que el estudio de Mogrevo, variables relacionadas con los costes laborales son estadísticamente significativos para determinar los flujos de IED que se dan en la UE (Sánchez & González, 2004).

Estos distintos enfoques hacen que los modelos y resultados obtenidos por los diferentes estudios puedan diferir. Por eso es importante identificar las características de la base de datos de la que se dispone y el objetivo que se pretende alcanzar con el análisis econométrico.

En relación con la estructura de la base de datos, esta tiene un gran impacto en la selección del modelo. En nuestro caso se tratan de tres bases de datos, la primera corresponde a una estructura de corte transversal puesto que la base de datos que se ha diseñado consta inicialmente de 30 países, de los cuáles se tiene la información de distintas variables, en un momento concreto en el tiempo, 2017. La segunda base de datos corresponde a una estructura de datos panel, con las mismas variables que en la primera base de datos, a diferencia de que se dispone información de las variables de los últimos 10 años, lo que permite estudiar los efectos temporales en los flujos de inversión. Por último, la tercera base de datos es la que más dista de las dos anteriores puesto que contiene los datos necesarios para llevar a cabo un modelo gravitacional teniendo en cuenta variables como la distancia entre dos países. Estas variables serán explicadas de una manera más detallada en el próximo capítulo.

2.1 Modelo de regresión múltiple

Tal y como hemos introducido anteriormente, el MLR es un modelo de regresión lineal, el cuál permite incluir k variables independientes explicativas (Wooldridge, 2015). El modelo poblacional puede expresarse de la siguiente manera:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u$$

Dónde,

y es la variable dependiente o variable explicada

$x_1 \dots x_k$ son las variables independientes, o variable explicativa

β_0 es el intercept

β_1 es el parámetro asociado con x_1 (la primera variable independiente)

β_2 es el parámetro asociado con x_2 , y así sucesivamente, hasta el parámetro k asociado con la variable independiente k .

u término del error, ocasionado por las variables no observadas

La principal asunción del MLR es que al igual que en la regresión simple, requiere que los factores o variables independientes no estén correlacionados con el término del error. O lo que es lo mismo:

$$E(u|x_1, x_2, \dots, x_k) = 0$$

Para la obtención de la estimación del modelo, se aplicará el método de MCO. Se basa en la minimización de la suma de los cuadrados de los residuos (Wooldridge, 2015). La suma de los cuadrados de los residuos se puede observar en la siguiente expresión:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik})^2$$

La minimización de esta expresión se basa en la minimización de $k + 1$ ecuaciones lineales, con $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ desconocidos. Para su resolución se aplicará lo que se conoce como condiciones de primer orden. Las condiciones de primer son las siguientes:

$$\sum_i \hat{\varepsilon}_i = 0, \sum_i \hat{\varepsilon}_i X_{1i} = 0, \dots, \sum_i \hat{\varepsilon}_i X_{ki} = 0$$

Estas condiciones implican que la media muestral de los residuos ha de ser 0 y la covarianza muestral entre X_i y el error es 0.

El resultado final de la estimación del modelo daría lugar a la siguiente expresión:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

En relación con la interpretación de la estimación de los parámetros del modelo podemos concluir lo siguiente. La constante se puede interpretar de forma matemática cómo el valor que toma la variable dependiente cuándo el resto de las variables dependientes es 0. A pesar de esta interpretación, no se puede concluir que tenga sentido económico (Hill, Griffiths & Lim, 2018). La interpretación del resto de parámetros del modelo puede ser interpretados como el cambio en el valor de la variable dependiente dado un cambio de una unidad en la variable explicativa, todas las demás variables se mantienen *ceteris paribus* o lo que es lo mismo constantes. Siempre y cuándo las variables tengan una escala constante. En el caso de estar expresadas en logaritmos, hablaremos de cambios porcentuales (en términos de elasticidad).

Por último, el signo nos permite analizar tanto para el caso en el que las unidades están en términos de elasticidades o lineales, la relación que las variables guardan con la variable dependiente, siendo está positiva o negativa. Lo que implica que el aumento del parámetro implicaría el aumento de la variable dependiente o su disminución (Hill, Griffiths & Lim, 2018).

Para que estas interpretaciones sean fiables, es decir que la estimación del modelo sea consistente, insesgada y óptima han de cumplir ciertas condiciones, las cuáles se explican de forma detallada en el siguiente apartado.

1.1.1 Condiciones de las estimaciones por MCO

(1) Los errores tienen una media condicionada a 0:

$$E(u|x_1, x_2, \dots, x_k) = 0$$

(2) Las observaciones han de ser independientes e idénticamente distribuidas

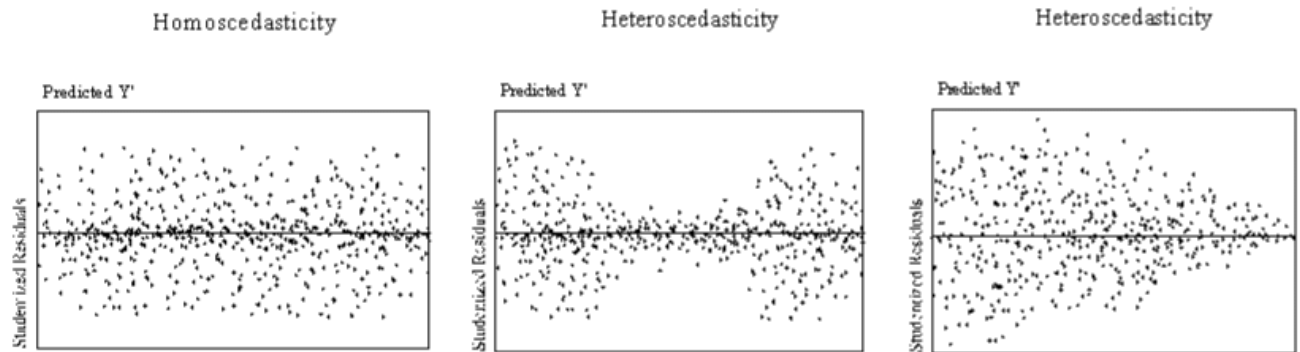
(3) Los valores extremos han de ser evitados y la variable dependiente ha de tener cuatro momentos finitos

(4) No hay multicolinealidad perfecta

(5) Homocedasticidad:

$$\text{var}(u|X_{1i}, \dots, X_{ki}) = \sigma_u^2$$

Ilustración 1. Gráfico de residuos Homoscedasticidad vs. Heteroscedasticidad.



Fuente: https://www.researchgate.net/figure/Example-of-homoscedasticity-Ideally-residuals-are-randomly-scattered-around-0-the_fig2_319091775

En la ilustración 1 se puede observar como ha de ser el gráfico de los residuos estandarizados si se cumple la condición de homocedasticidad (a la izquierda), y cuándo por lo contrario esta condición no se satisface (los dos gráficos restantes).

- (6) Linealidad. Es decir, el estimador se puede escribir como una combinación lineal de los valores de la variable dependiente.
- (7) Los estimadores han de ser insesgados. Es decir, dentro de los posibles estimadores posibles por su error cuadrático, estos han de tener la menor varianza posible. En otras palabras, la media de la distribución del estadístico converge en probabilidad al verdadero valor del parámetro.
- (8) Eficiencia: significa que el estimador es aquel que tiene varianza mínima dentro de la familia de estimadores lineales e insesgados. Se demuestra por el Teorema de Gauss Márkov¹. Las condiciones que se han de cumplir para cumplir el Teorema son las correspondientes a los supuestos (1), (2), (5), (6). En el supuesto en el que las condiciones se cumplan estaríamos ante el mejor estimador lineal insesgado (BLUE: *best linear unbiased estimator*).

En el momento de realizar la regresión habremos de comprobar que se cumplen estas condiciones para comprobar que la inferencia se realiza de manera adecuada, y en efecto los resultados del modelo son fiables.

¹ Teorema de Gauss Markov: establece los supuestos para que un estimador por MCO sea lineal e insesgado óptimo (BLUE)

2.1.2. Bondad del ajuste

Por último, hemos de comprobar cuán de sensible es nuestro modelo y como se ajusta a la realidad. Para estudiar la bondad del ajuste del modelo se suelen analizar los siguientes estadísticos:

Error estándar de los residuos (RSE): es la forma que se conoce para estimar la varianza del error, ya que normalmente se desconoce. Es la desviación típica residual.

$$RSE = \sqrt{RSS/(n - 2)}, \text{ donde}$$

RSS: es la estimación sobre la desviación promedio de cualquier punto respecto a la verdadera recta de regresión (Gil, 2018)

R cuadrado o coeficiente de determinación del modelo, este valor se interpreta como el porcentaje de la varianza de la variable dependiente que se explica por el modelo. Se expresa de la siguiente forma:

$$R^2 = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2}$$

Este valor ha de estar entre 0 y 1. Para valores de 0 el modelo sería nulo, y para el valor de 1, el modelo sería perfecto. En la realidad un valor del R cuadrado de alrededor del 0.80, sería bueno. En caso de tomar un valor más elevado, puede ser un indicador de autocorrelación entre las variables o de una elevada multicolinealidad. Este estadístico puede llevar a conclusiones erróneas ya que mide la correlación que hay entre ciertas variables en un momento determinando, aunque no refleja si en efecto hay una relación causal entre ambas o no.

Otra de las limitaciones es que este estadístico no se ve alterado al introducir nuevas variables independientes al modelo, incluso si estas variables son irrelevantes. Para corregir este problema se suele analizar el estadístico **R² ajustado**, el cuál esta corregido por los grados de libertad. Se expresa de la siguiente forma:

$$\bar{R}^2 = 1 - \frac{n - 1}{n - k} (1 - R^2)$$

Además, es conveniente analizar los criterios de información para poder comprar modelos anidados y así escoger aquel que tome el valor inferior de estos estadísticos. Los más conocidos son:

Criterio de Akaike (AIC):

$$AIC(k) = \log[\hat{\sigma}_k^2] + \frac{2k}{n}$$

Criterio de Schwarz (SIC):

$$SIC(k) = \log[\hat{\sigma}_k^2] + \frac{k \log[n]}{n}$$

2.1. *Modelo de regresión lineal con datos panel*

El segundo modelo que se calcula en este trabajo corresponde a un modelo de regresión lineal múltiple utilizando el método de los MCO explicado en el apartado anterior con la diferencia de que nuestra segunda base de datos es propia de datos panel. Puesto que a diferencia que en el primer data set no sólo contiene los datos de 2017 sino los datos de los 10 últimos años. Aunque las variables y países estudiados son las mismas.

Este modelo se define de la siguiente forma:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

Dónde Z_i nos cambios que se producen en el tiempo. Para tenerlos en cuenta se calculará los efectos fijos de la regresión, para de esta forma poder estimar β_1 de tal forma que estudie el efecto de Y_i al cambiar la variable explicativa X_i . Por lo que definimos $\alpha_i = \beta_0 + \beta_2 Z_i$, y obtenemos el siguiente modelo de regresión de efectos fijos:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

Los diferentes valores que tomas los específicos valores de α_i es lo que denominamos efectos fijos de la entidad i . Por lo que el modelo teniendo en cuenta estos efectos fijos puede escribirse de forma equivalente, en la que se incluye el modelo y $n-1$ dummy regresores ($D2, D3, \dots, Dn$), los cuáles miden el efecto de cada i . También se incluye la constante, de esta forma se evita un escenario de perfecta multicolinealidad.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \dots + \gamma_n D3_n + u_{it}$$

2.2. *Modelo gravitacional*

Este es un modelo que también utiliza el método de cuadrados ordinarios para ser resuelto. Aunque también puede utilizar otras distintas aproximaciones como por ejemplo la aplicación de *Pseudo Poisson* de la máxima verosimilitud para adaptarse de una mejor forma a los datos de la muestra (Wölwer, Burgard, Kunst & Vargas, 2018). No obstante, en este trabajo se utilizará la versión más sencilla que es la que aplica el MCO. Ya que con el trabajo se pretende analizar las principales aplicaciones del método de MCO para el estudio de flujos de IED.

Este modelo fue introducido por Newton en 1729 con su ley gravitacional, aunque ha resultado de gran relevancia su aplicación en los estudios de comercio internacional (Wölwer, Burgard, Kunst & Vargas, 2018). El modelo se puede definir de la siguiente manera:

$$X_{ij} = G \frac{Y_i^{\beta_1} Y_j^{\beta_2}}{D_{ij}^3}$$

Este modelo recoge los flujos tanto de entrada ($Y_i^{\beta_1}$) como de salida ($Y_j^{\beta_2}$) de un país, en los que normalmente estandarizados utilizando el PIB para evitar que los efectos estructurales de tamaño distorsionen los resultados. La expresión también incluye la información relativa a la distancia D_{ij} entre los países en los que se produce el intercambio de flujos, en nuestro caso de IED.

3. Base de datos

La fuente de la base de datos de las distintas variables pertenece a los datos publicados por la OCDE. Debido a que se pretende estudiar la relevancia que tiene esta organización como referente en asuntos económicos y analizar su fuente de datos así cómo cuán de fácil es acceder a sus publicaciones. La estructuración que ha dado lugar a la base de datos utilizada para la realización de este trabajo ha sido de elaboración propia, ya que se han seleccionado las variables que a priori según la revisión literaria de los estudios realizados mostraban tener una mayor significación estadística y económica para la explicación de la IED.

Base de datos 1 – Corte transversal

La base de datos es una base de datos transversal. Puesto que contiene los datos correspondientes a 30 países para el año 2017, por lo que los datos corresponden a un momento puntual en el tiempo. A diferencia de algunas variables como el Producto Interior Bruto (PIB), o el flujo de exportaciones e importaciones de las cuáles se tienen valores actualizados al último año 2020, hay variables como los impuestos a corporaciones, en que el último año disponible es el 2017. Esto muestra la complejidad y la dificultad para acceder a datos publicados por gobiernos, y cómo, aunque estos estén obligados a proporcionarlos, muchos países optan por la confidencialidad de estos como es el caso de Holanda. Esta dificultad de acceso a datos también muestra el problema generalizado que tienen las agencias estadísticas en el momento de recopilar la información de una forma rápida sin que se incurran a varios años de retraso.

Las variables seleccionadas para la realización del análisis estadístico y su posible inclusión al modelo econométrico son las siguientes:

VARIABLES NUMÉRICAS:

Flujos de Inversión Extranjera: se dispone tanto de la información de los flujos de entrada (**IED_in**) y los flujos de IED de salida (**IED_out**). Es importante tener en cuenta que las estadísticas que reflejan estos estudios no incluyen los flujos de inversión correspondientes a empresas de adquisición con propósito especial (SPACs) debido a que los flujos de inversión ocasionados por este tipo de entidades sólo son flujos de puente y no representan la inversión real en el país (OECD, nd).

Otra consideración que es necesaria tener en cuenta para la interpretación de estos flujos, es el hecho de que estos pueden tomar valores negativos, cómo es el caso para países como Bélgica, Islandia o Luxemburgo. Esto puede ser debido a tres razones, La primera razón es que se deba a un proceso de desinversión en el cuál el interés de la inversión vuelva a la empresa de origen. La segunda razón es que el dinero de la inversión proceda de un préstamo a una parte de la empresa afiliada. Y la última razón es que los beneficios de la reinversión sean negativos (OECD, nd).

No obstante, desde un punto estadístico, tener valores negativos en una de las principales variables del modelo, hace que no se puedan tomar logaritmos, y por consiguiente que no se pueda hacer un estudio de las elasticidades de dichas variables. No obstante, se podría hacer cambios en la misma para evitar esto como es un proceso de normalización, aunque esto último dificultaría la interpretación directa de los parámetros estimados por el modelo.

PIB: es uno de los indicadores que se utiliza para medir el tamaño de una economía a nivel nacional. No obstante, puede llevar a una mala interpretación si se interpreta en términos absolutos debido a los efectos del tamaño intrínseco del país, cómo área geográfica o población. Por esta razón, esta variable será de ayuda para expresar distintas variables del modelo en términos del PIB, y así relativizarlas, como por ejemplo para las variables relativas a los sectores económicos. En relación con la unidad de esta variable están expresados a precios constantes y tipo de cambio constante, en el cuál el año base de referencia es 2015.

Sectores económicos (*Ind, Agri, Cons, Real_est*): son variables numéricas que contienen el PIB correspondiente al sector de la industria, agricultura, construcción y bienes raíces. Normalmente hay sectores que tengan una mayor relevancia en el momento de atraer IED. Por este

Indicadores de **comercio exterior:** desde un punto de vista económico se espera que aquellos países con una mayor apertura al comercio internacional perciban una mayor inversión extranjera directa. Las variables *Exp* e *Imp* contiene la información relativa a las exportaciones e importaciones respectivamente, realizadas por cada país en 2017 en términos del PIB.

Productividad: *Prod*: mide la productividad de un país. Esta productividad se mide como el PIB producido por hora trabajada. Esta expresado en USD a precios constantes con 2015 como año base y a paridad del poder adquisitivo (PPs). Esta variable es normalmente incluida en los estudios de IED ya que, desde un punto de vista económica, se puede entender que aquellos países en los que la productividad es elevada, es susceptible a una mayor IED (Denisia, 2010). Debido a que mayor serán los rendimientos de los inversores si los salarios se mantienen constantes.

Nivel de precios: *Infl*: contiene la información de la inflación en cada país en 2017. La inflación, al igual que las anteriores variables también ha sido estudiada por diversos autores como posible variable determinante de la IED (Economou, Hassapis, Philippas & Tsionas, 2017). Una de las razones económicas es por su relación con los tipos de interés y por consiguiente con la balanza de pagos de un país. Esto tiene una relación directa desde el punto de vista económico, ya que una subida elevada de los niveles de inflación genera expectativas negativas por parte de los inversores, lo que hace que estos inviertan menos. Pese a esto, en los

estudios realizados no se ha conseguido probar estadísticamente que esta variable es estadísticamente significativa.

Índice del efecto de la **regulación restrictiva** sobre la IED (**Reg**). Este índice es diseñado por la OCDE. Introduce distintos tipos de restricciones regulatorias que pueden afectar a la IED. Entre las cuáles destaca las limitaciones a activos extranjeros, mecanismos de supervisión o restricciones a empleados extranjeros, entre otras cosas. Cuanto mayor sea el índice, implica que mayor es el nivel de restricciones a la IED (OECD, nd).

Salar: los datos corresponden a la **media de los salarios anuales** para un empleado que trabaje a tiempo completo en la economía. Los datos están expresados como precios constantes en USD y por la paridad del poder adquisitivo² (PPPs).

Impu: contiene la información de los **impuestos a corporaciones** recaudados por parte del gobierno. Esta variable parece relevante desde un punto de vista económico ya que la presión fiscal puede ser una de las razones que conlleve a una mayor atracción de IED si esta es baja o a una menor IED si por el contrario la presión ejercida a empresas es elevada. Esta variable está normalmente incluida en la mayoría de los estudios y también se tiene en cuenta en distintas publicaciones de referencia, por ejemplo, en *Doing Business*³ del Banco Mundial.

Variables categóricas:

UE: es una variable dicotómica, la cuál toma valor 1 en caso de que el país, sea un Estado Miembro de la Unión Europea, y 0 de lo contrario.

desarrollo: es una variable dicotómica, la cuál toma el valor 1 si el país se considera como un país en vías de desarrollo, y 0 de lo contrario. Para la referencia de un país en vías de desarrollo se ha tomado la referencia que utiliza el Banco Mundial, el cuál determina que un país está en vías de desarrollo según sea su PIB per cápita. En nuestra base de datos son Chile, México y Turquía.

Base de datos 2 – Datos panel

Esta base de datos introduce las mismas variables que en la Base de Datos 1 con la diferencia de que en vez de incluir estos datos sólo para el año 2017, incluye todos los datos observados en el intervalo de tiempo de 2007-2017. En el estudio del modelo se optará por eliminar ciertos países del estudio, puesto que para alguno de ellos no se dispone de los datos de la variable dependiente (flujos de entrada de IED).

² Paridad del poder adquisitivo (PPA), mide la comparación de adquisición de un país

³ *Doing business*, es una publicación anual que se encarga de analizar cuáles son los países en los que por razones regulatorias y del propio país es más recomendable establecer una actividad económica. Entre los indicadores destacan: acceso a la electricidad, permisos de construcción y pago de impuestos.

Base de datos 3 – Datos para el modelo gravitacional

Para la realización del estudio gravitacional de los flujos de IED hemos utilizado la base de datos disponible por el Centro de investigación y especialización en economía mundial (CEPII). Esta base de datos nos proporciona los datos relevantes al factor de distancia entre los países, factor imprescindible para la realización del modelo. Esta base de datos también proporciona variables comúnmente utilizadas por los modelos gravitacionales de comercio exterior como es el tipo de moneda y si ha habido relación colonial entre los países o no, al igual que los acuerdos internacionales de comercio existentes entre los mismos, como uniones aduaneras o zonas de libre comercio. No obstante, en este trabajo se simplificará la versión del modelo ya que el enfoque principal de este trabajo es mostrar una aplicación de los MCO en diversas regresiones lineales.

En lo que respecta a las otras variables que conforman la base de datos, la fuente utilizada sigue siendo la base de datos proporcionada por la OCDE, la cuál cabe destacar que tiene muchos valores *missing*, por lo que el marco temporal del estudio se verá afectado. Esto se explicará con más detalle en el apartado de resultados.

Un resumen de las variables que componen esta base de datos se puede observar en la siguiente tabla:

Tabla 1. Variables de la Base de datos 3

Variable	Nombre Variable	Explicación
Flujos de entrada de IED en el país	IED_in	Flujos de IED realizados de un país i a j. Está expresado en US million dollar. Los flujos de IED no incluyen los hechos por SPACs y puede tener valores negativos al tratarse de flujos netos
Año	Year	Año al que corresponden las distintas variables numéricas observadas
País receptor	pais_in	País que recibe el flujo de inversión
País emisor	pais_out	País del que se realiza la IED
Distancia	Dist	Distancia geográfica entre país i y j
Producto Interior Bruto 1	PIB1	Producto Interior Bruto del país que recibe IED
Producto Interior Bruto 2	PIB2	Producto Interior Bruto del país del cuál se emite la IED

3.1. Análisis descriptivo

Para el análisis de este trabajo se ha realizado una transformación de alguna de las variables numéricas que estaban en términos absolutos, a términos relativos en base al PIB, como es el caso de las variables relativas a los distintos sectores de la economía.

En primer lugar, se tienen en cuenta los valores *missing* que tiene la base de datos. Es importante recordar que entre las múltiples opciones que hay para tratarlos, una de ellas es excluir la observación del estudio lo cuál puede resultar bastante perjudicial en nuestro caso al tratarse de una muestra pequeña. Sustituir el valor no observado por la media si se trata de una variable numérica o por la moda si la variable fuese categórica. También hay diferentes algoritmos que se utilizan en la minería de datos para imputar valores, como el *k-nearest-neighbors* que nos permite hacer una aproximación más precisa del valor no observado.

En nuestro caso, tenemos 3 observaciones que faltan. Dos de ellas corresponden a los distintos flujos de IED tanto de entrada como de salida para Holanda y Estonia. Para el caso de Holanda, se ha decidido eliminarlo de la muestra al considerarlo un valor extremo que alteraba los resultados. Esto es debido a que sabemos que Holanda es uno de los países con mayor recepción y emisión de IED a nivel mundial, por lo que una aproximación utilizando la media daría lugar a conclusiones erróneas. (UNCTAD, 2019). Por otro lado, para el caso de Estonia se ha optado por una aproximación utilizando la media, debido a que estamos ante una variable numérica. Si esta fuese categórica, el estadístico utilizado sería la moda. El último valor del cuál no se tiene información es para la variable salarios en Turquía, puesto que por motivos de confidencialidad no permite que estos datos se hagan públicos a través de la OCDE. En este caso se ha optado por una aproximación por la media de la variable, puesto que se trata de una variable numérica. Esto implica que la muestra final sobre la cuál se estimará el modelo econométrico esta conformada por 29 países.

Transformación de las variables

Las variables que eran perceptibles a sufrir efectos de tamaño, ocasionados por las diferencias estructurales de los países se han transformado a variables relativas en relación con su PIB respectivo. Dichas variables son las correspondientes a los diferentes sectores económicos (Industria, Agricultura, Construcción y bienes raíces). También se ha relativizado la variable de impuestos por la misma razón.

Tabla 2. Comparativa variables absolutas y variables relativas

Variables	Máximo_absoluto	V1	Máximo_relativo	V2	Mínimo_absoluto	V3	Mínimo_relativo	V4
IED_in	United States	314977.0	Estonia	0.9505454	Luxembourg	-23157.479	Luxembourg	-0.3765540
IED_out	United States	353356.0	Estonia	1.0184408	Norway	-6865.796	Norway	-0.0172077
Ind	United States	2733245.2	Czech Republic	0.2948630	Iceland	2702.883	Luxembourg	0.0639462
Agri	United States	198260.9	Turkey	0.0631950	Luxembourg	110.563	Luxembourg	0.0017978
Cons	United States	742178.9	Turkey	0.0838685	Iceland	1212.331	Greece	0.0149863
Real_est	United States	2265575.3	Greece	0.1556077	Iceland	1837.419	Poland	0.0428486
Exp	United States	2361530.7	Luxembourg	2.1475913	Iceland	10548.746	United States	0.1244014
Imp	United States	2970702.4	Luxembourg	1.7978981	Iceland	9915.492	United States	0.1564915
Impu	United States	305978.6	Luxembourg	0.0531845	Estonia	412.070	Turkey	0.0151041

Tal y cómo se puede observar para la mayoría de las variables, difieren entre si toman los valores relativos o los valores absolutos. Para los estudios posteriores las variables se estudiarán en término relativos al tamaño de su economía, es decir de su PIB. Ya que tal y cómo se ha podido observar en la Tabla 3.1, hay cambios significativos entre tomar valores absolutos relativos, debido a sufrir efectos de tamaño estructurales de cada país. Estos estadísticos corresponden a los datos de 2017.

3.1.2. Estadísticos descriptivos variables de corte transversal 2017

A continuación, nos disponemos a realizar la estadística descriptiva de las variables. En la Tabla 3.2 se pueden observar los principales estadísticos descriptivos propios para variables numéricas.

Tabla 3. Estadísticos descriptivos de las variables sujetas a estudio

Variabes	Mínimo	Media	Mediana	Máximo	DT	Kurtosis
IED_in	-0.38	0.04	0.02	0.95	0.19	14.98
IED_out	-0.02	0.06	0.01	1.02	0.19	20.47
Industria	0.06	0.19	0.19	0.29	0.05	-0.25
Agricultura	0.00	0.02	0.02	0.06	0.01	0.31
Construcción	0.01	0.05	0.05	0.08	0.01	0.58
Bienes raíces	0.04	0.09	0.09	0.16	0.03	0.06
Exportaciones	0.12	0.58	0.48	2.15	0.38	7.84
Importaciones	0.16	0.54	0.43	1.80	0.32	5.98
Productividad	25.84	54.93	52.43	96.54	19.16	-0.81
Inflación	0.53	2.31	1.88	11.14	2.00	10.67
Índice de restricción	0.00	0.06	0.04	0.23	0.06	1.31
Salario	16984.12	41085.82	39551.53	69541.03	15172.36	-1.15

La primera columna de la Tabla 3.2, indica el valor mínimo que toma cada variable. La segunda columna el valor medio de la variable, el cuál se define con la siguiente expresión:

$$\bar{x} = \frac{\sum_{i=1}^N X_i}{N}$$

Es importante notar que para algunas variables difiere considerablemente de los valores de la columna de la mediana. Esto normalmente se debe a que las variables contienen valores extremos, puesto que la media es un estadístico que es sensible a dichos valores. Por ejemplo, para el caso de los flujos de entrada y de salida de la IED. Cómo para el nivel de exportaciones. Este fenómeno se puede observar en distintos tipos de gráficos como los diagramas de cajas o histogramas, los cuáles se introducirán a continuación.

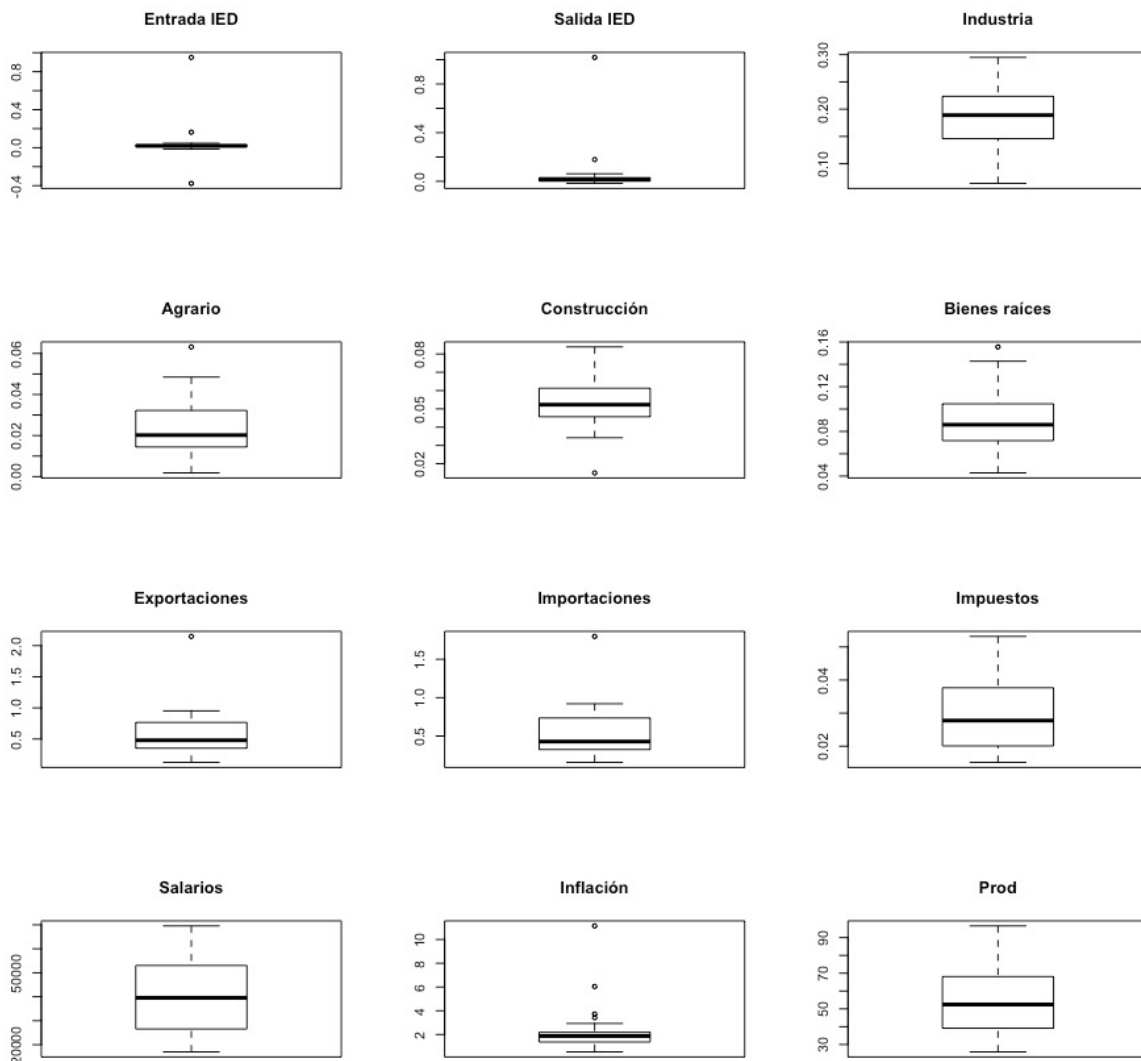
En esta tabla también podemos observar la desviación típica, la cuál se expresa de la siguiente forma:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

La desviación típica es una medida que nos permite medir la separación normal que hay entre los valores observados y la media. Nos permite estudiar la precisión de la media, y cuán de centrados están los datos respecto a la media. Entre mayor es la desviación típica mayor es la dispersión. Aunque es importante tener en cuenta la escala de la variable y la amplitud de los datos. Es decir, entre el valor más pequeño y el valor más alto. Esto nos indica que una de las variables con una mayor dispersión es la inflación.

Estos estadísticos de localización, así como de dispersión se pueden observar de forma visual con los diagramas de cajas tal y como se muestra a continuación:

Ilustración 3.1. Diagrama de cajas para las variables numéricas



Podemos observar cómo la mayoría de los gráficos muestran una gran sensibilidad a valores extremos (puntos extremos con gran distancia respecto a la mayoría de los datos representados por la caja).

En relación con la tabla, podemos ver que la última columna se muestra el estadístico de Kurtosis, el cuál nos permite estudiar la asimetría de los datos de cada variable. En concreto cuán encarpada o achatada está la distribución de los datos. Se puede expresar de la siguiente forma:

$$\text{Exceso de Kurtosis} = \frac{\sum_{i=1}^N (X_i - \bar{x})^4 \cdot n_i}{NS_x^4} - 3, \text{ donde}$$

n_i es la frecuencia absoluta del dato x_i

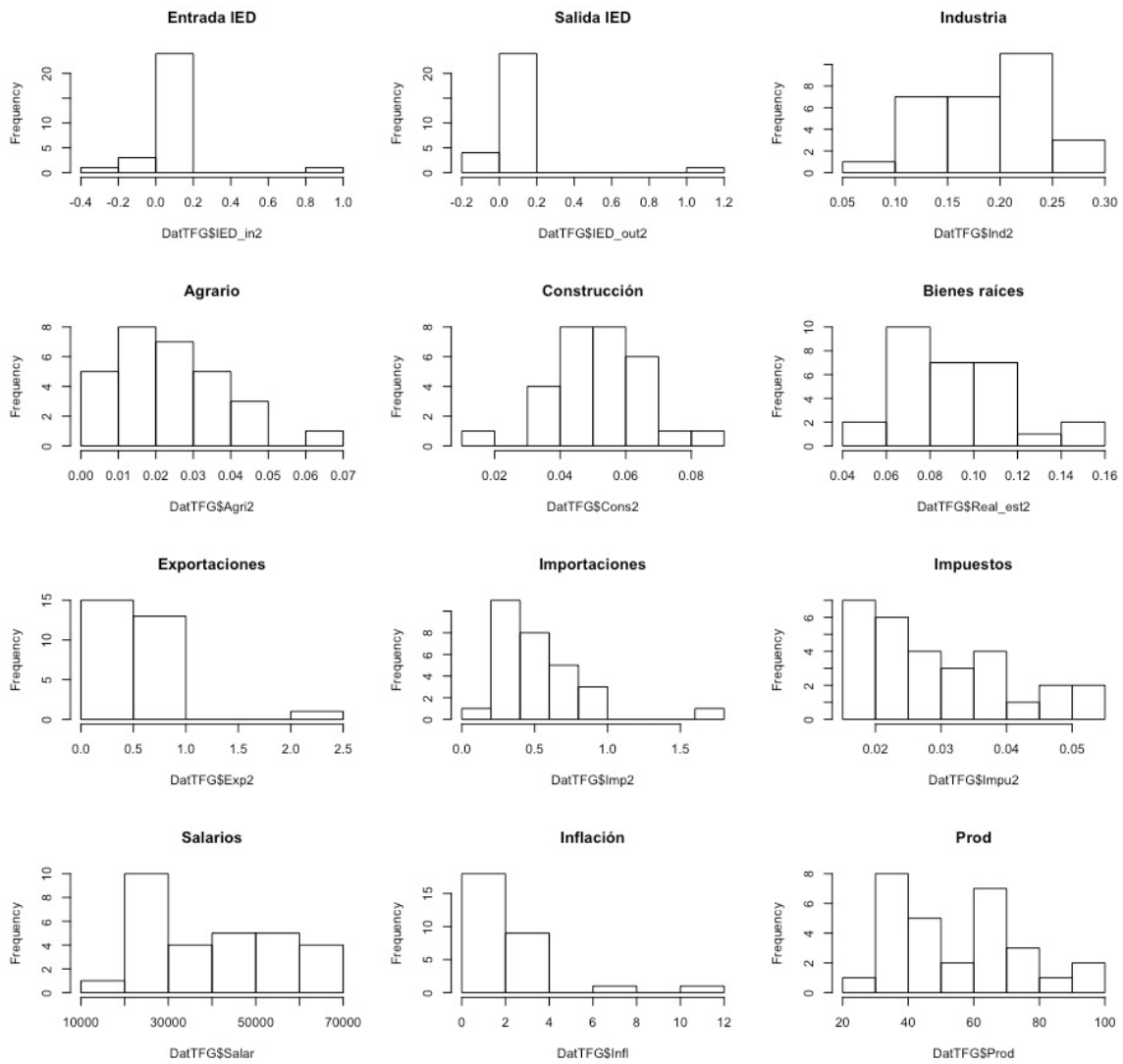
El estadístico nos permite concluir que para valores que son mayores a 0, estamos ante una forma de distribución leptocúrtica en la que los valores están centrados en la media. En nuestro caso las variables son las pertenecientes a los flujos de entrada y de salida de IED con un estadístico de 14.98 y 20.47 respectivamente. Bajo este mismo supuesto se encuentran los datos correspondientes a las exportaciones, importaciones e inflación.

Los valores cercanos a 0, se puede concluir que son propios a una distribución normal, entre los que destacan los datos correspondientes a el peso en la economía de los distintos sectores económicos y de la productividad.

Los valores que son inferiores a 0 implican que la forma de la distribución es platicúrtica, lo que implica que los datos no están concentrados alrededor de la media, el cuál en nuestro caso son los datos correspondientes a los salarios medios anuales.

Esta relación respecto a la media, al igual que la forma de la distribución que siguen los datos de las distintas variables se puede apreciar de manera visual a través de histogramas, tal y cómo se muestra a continuación.

Ilustración 2. Histograma para las variables numéricas

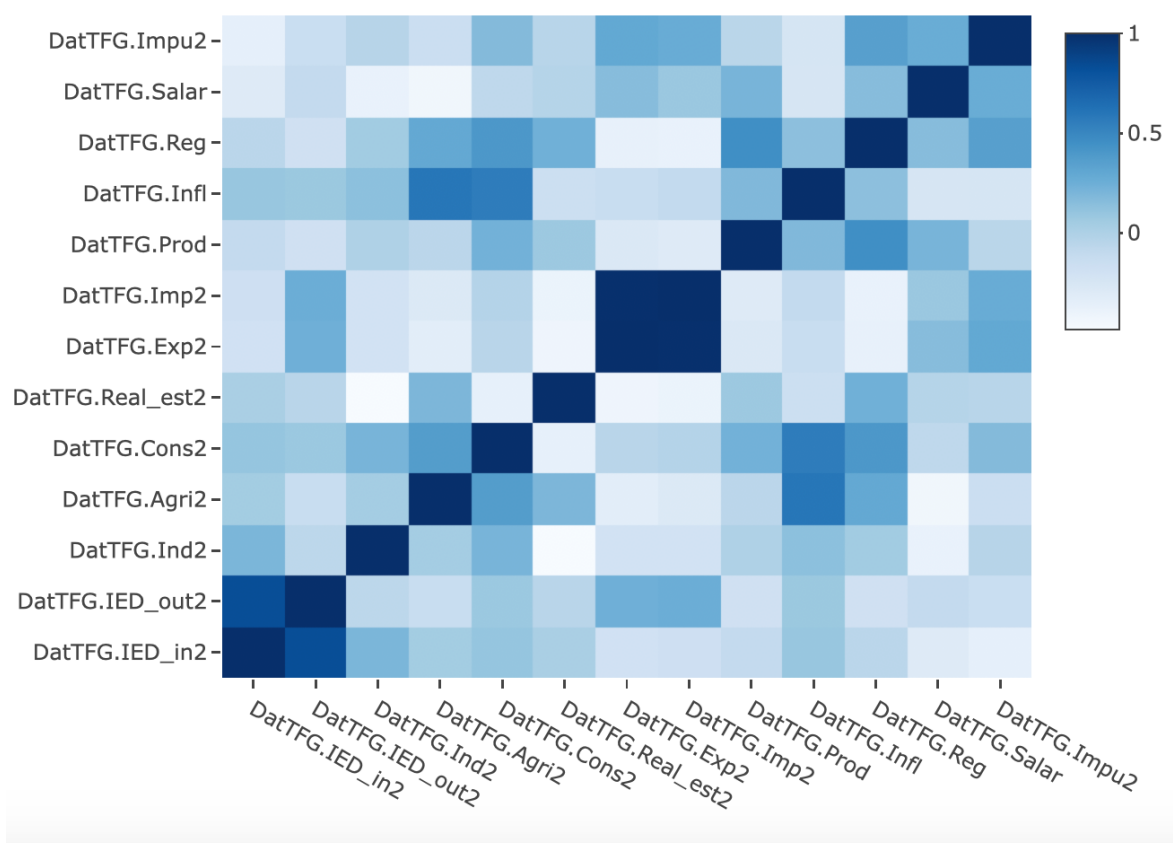


A continuación, podemos observar la matriz de correlación de las variables numéricas de la base de datos. Esto nos permitirá identificar que variables están correlacionadas y así evitar que se incluya la misma información en forma de dos variables independientes en la regresión.

La correlación se puede expresar de la siguiente forma:

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

Gráfico 1. Matriz de correlación



En el anterior gráfico podemos observar que para aquellos cuadrados que tomen una tonalidad de azul más intensa indican que mayor es el nivel de correlación entre dichas variables. En nuestro caso podemos observar como los flujos de entrada y de salida de IED respecto al PIB están altamente correlacionados, con un coeficiente de correlación del 83.37% (Anexo 1). Otra de las variables que muestran tener una correlación elevada son las variables son entre la inflación y los sectores de la agricultura y la construcción, con un coeficiente de correlación del 59% y 56% respectivamente (Anexo 1).

Este estadístico es relevante para poder anticiparnos a posibles problemas de multicolinealidad en la regresión econométrica.

Por parte de las variables categóricas es importante tener en cuenta que los estadísticos presentados anteriormente no son relevantes. Y en su caso los estadísticos descriptivos más apropiados son tanto la frecuencia absoluta como la frecuencia relativa.

Tabla 4. Frecuencia absoluta y relativa para la variable dummy pertenecer a la UE

	Frecuencia Absoluta	Frecuencia Relativa
No UE	9	0.31
UE	20	0.69

Podemos observar como el 69% de los datos son Estados Miembros de la UE, mientras que el 31% no pertenecen al mercado interior, siendo 20 y 9 países respectivamente.

Tabla 5. Frecuencia absoluta y relativa para la variable dummy er un país en vías de desarrollo

	Frecuencia Absoluta	Frecuencia Relativa
País desarrollado	26	0.9
País en vías de desarrollo	3	0.1

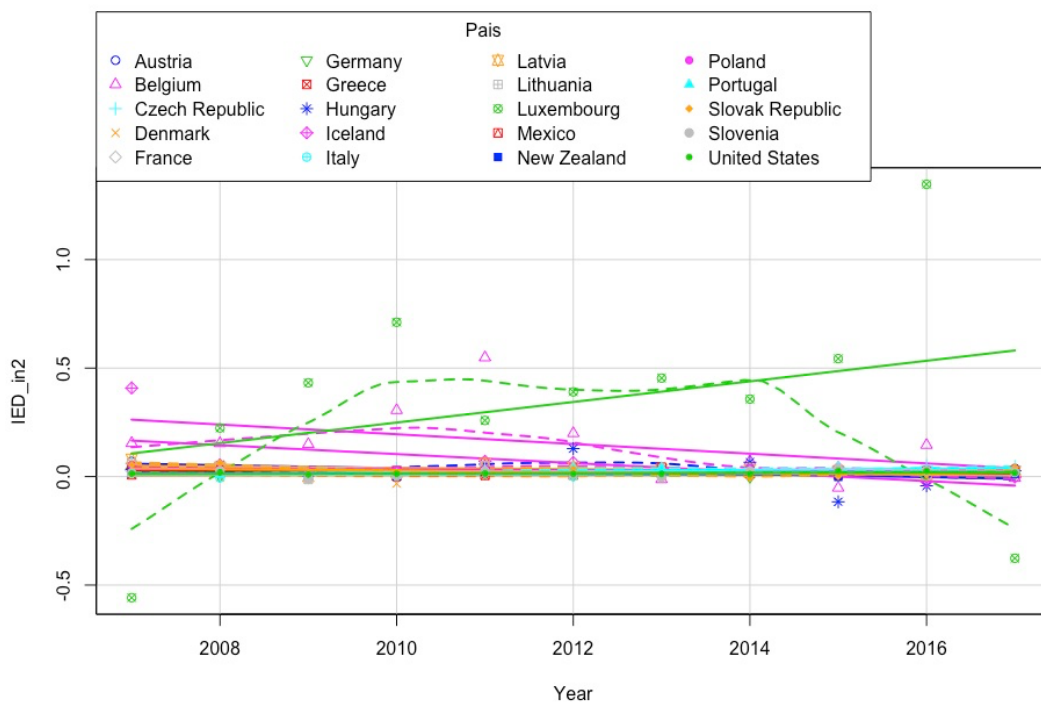
Podemos observar como el 90% de los países que conforman la base de datos son países desarrollados, mientras que el 10 % se consideran países en vías de desarrollo.

3.2. Estadística descriptiva de los datos panel

En este apartado estudiamos los principales gráficos de dispersión que nos permite anticipar ciertas tendencias en la variable dependiente, es decir flujos de entrada de inversión en relación con los distintos países que conforman la muestra o los efectos anuales.

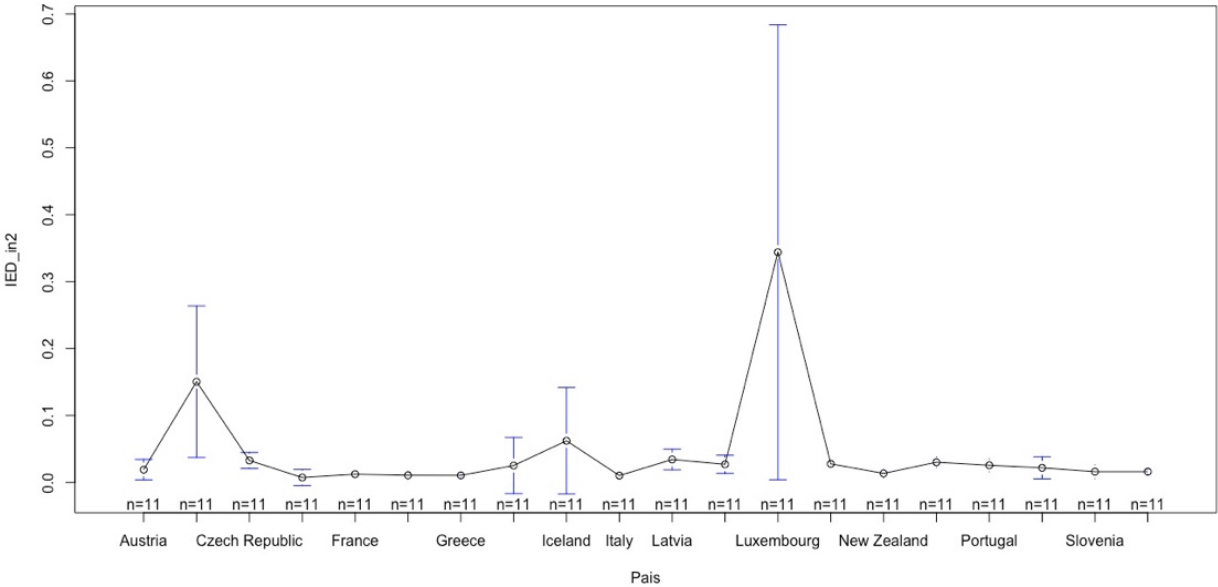
En el Gráfico 3.2 podemos observar la evolución de los flujos de entrada de IED en cada uno de los países de la muestra en relación con su PIB. De forma visual podemos concluir que la mayoría de los países siguen una tendencia similar a excepción de ciertos países como Luxemburgo. En este gráfico también se puede observar como hay ciertos países como Luxemburgo que tiene ciertos valores extremos en comparación con el resto de la muestra.

Gráfico 2. Evolución de los flujos de entrada de IED 2007-2017



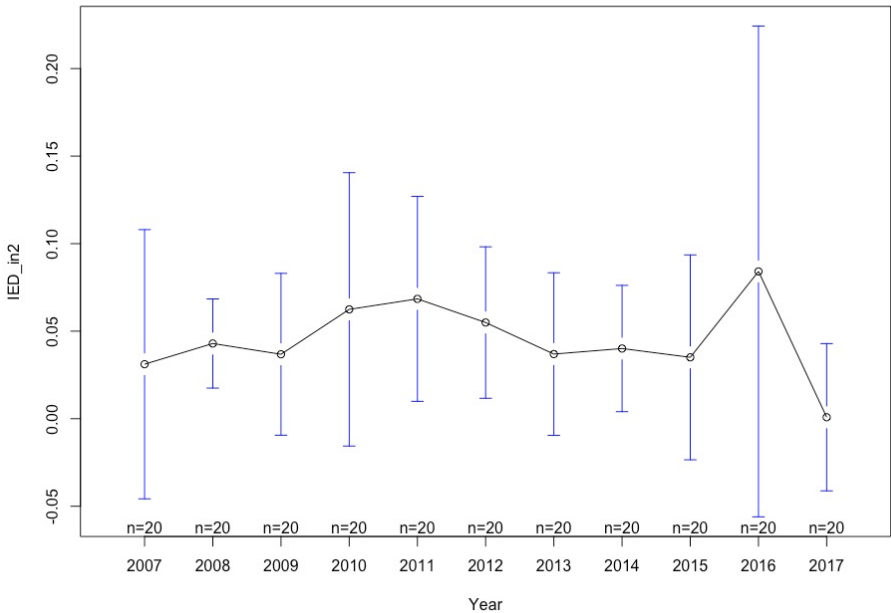
Prosiguiendo con el análisis visual de la base de datos panel. Podemos observar como el Gráfico 3.3 reafirma las sospechas de que Luxemburgo es el país con el mayor valor extremos de la muestra de los datos. Al igual que su tendencia generalizada es superior al resto de los países que conforman la muestra. Este gráfico nos permite estudiar la descriptiva de cada variable respecto al país y ver dónde se localiza su media en relación con los demás países.

Gráfico 3. Evolución de los flujos de entrada de IED 2007-2017



Y el último gráfico del análisis de los datos temporales del Gráfico 3.4. Podemos ver como se distribuyen los datos tomando como valor de referencia los distintos periodos de tiempo en lugar de los países como hemos visto en el anterior gráfico. Podemos ver que la dispersión de cada año es relativamente parecida, aunque los años que se caracterizan por tener una mayor dispersión dentro los países que conforman la muestra es en 2007 y 2016.

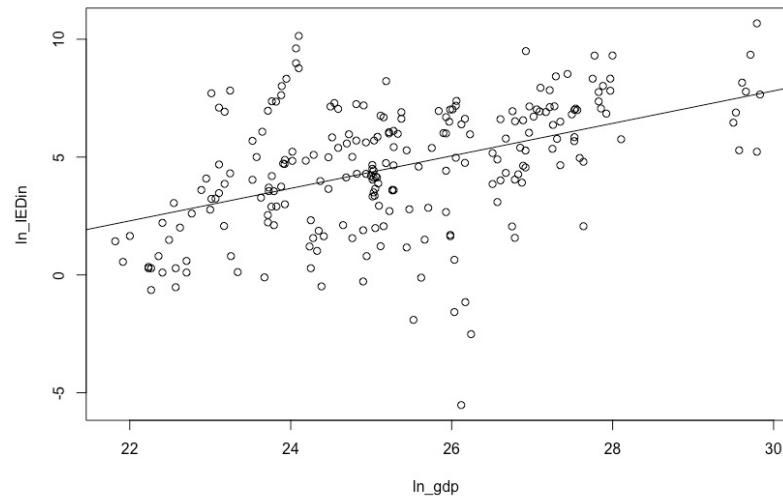
Gráfico 4. Evolución de los flujos de entrada de IED 2007-2017



3.3. Estadística descriptiva de la base de datos del modelo gravitacional

Con relación a la visualización de los datos que conforman la tercera base de datos. Es relevante estudiar la dispersión de el PIB tanto del país emisor de la IED como del país receptor, respecto a la variable dependiente que en nuestro caso es la IED, tal y cómo se muestra en el Gráfico 5.

Gráfico 5. Relación entre los países en base a su PIB y a la IED recibida



En este gráfico podemos observar la diferencia entre los PIB en relación con la IED. Y como estos se distribuyen. En el gráfico el número de puntos no es elevado puesto que el número de países que está siendo evaluado es reducido.

4. Selección del modelo y estudio de bondad

En este apartado se podrá observar las distintas conclusiones que se han podido extraer con la realización de los tres modelos. También se analizarán las limitaciones de cada uno de los modelos, así como la complejidad del estudio estadístico de ciertas variables macroeconómicas como la IED.

4.1. Resultados modelo de regresión lineal múltiple por MCO para datos de corte transversal

La información de los distintos modelos así como los distintos modelos realizados se encuentran en la siguiente tabla resumen. En las distintas columnas podemos observar los resultados para cada modelo. En las columnas de la tabla nos muestras los coeficientes de los estimadores en caso de que dicha variable haya sido incluida en el modelo, como su significación estadística en caso de tener un* acompañando el valor y debajo del mismo se muestra la desviación estándar del estimador. También se incluye el numero de observaciones para cada modelo, es importante resaltar que para los modelos robustos no es correcto calcular los estadísticos de eficiencia del modelo, como es el caso del R^2 ajustado, ya que estos no pueden ser interpretados.

Tabla 6. Resultados de los modelos de regresión lineal por MCO

Regresores	(1)	(2)	(3)
intercept	0.0243 (0.1038)	0.0340 (0.0776)	0.0135 (0.0203)
IED_out2	0.9254 (0.0331)	0.9247 (0.0303)	0.2300 (0.0917)
Prod	-0.0001 (0.0005)	-0.0001 (0.0003)	-0.0001 (0.0001)
Infl	-0.0011 (0.0051)	-0.0014 (0.0044)	-0.0006 (0.0012)
Reg	0.0532 (0.2327)		
Salar	0 (0)	0 (0)	0 (0)
Ind2	0.2350 (0.1773)	0.2173 (0.1473)	0.0200 (0.0392)
Agri2	0.8431 (0.6776)	0.8130 (0.5875)	-0.1846 (0.1745)
Cons2	0.3193 (0.6422)	0.3273 (0.5212)	0.2167 (0.1353)
Real_est2	-0.0609 (0.3938)	-0.0507 (0.3298)	0.1138 (0.0872)
Imp2	-0.0961 (0.0261)	-0.0890 (0.0223)	0.0168 (0.0094)
UE	0.0105 (0.0320)		
desarrollo	-0.0425 (0.0345)	-0.0470 (0.0250)	-0.0013 (0.0073)
Grados de libertad	16	18	15

Numb.			
Observaciones	29	29	26
<i>Significance codes : 0 '***', 0.001 '**', 0.01 '*'</i>			

Para la realización de dicho estudio se ha procedido de la forma que se explica a continuación. En primer lugar, hemos realizado la prueba de normalidad Shapiro-Wilk Test de la variable dependiente. Este estadístico se expresa de la siguiente forma:

$$W = \frac{(\sum_{t=2}^n a_t y_t)^2}{\sum_{t=1}^n (x_t - \bar{y})^2}$$

Dónde:

n : número de observaciones

y_t : valores ordenados de la muestra

a_t : valores tabulados

x_t : es el valor de la observación

Dado a que el p-valor de la prueba (Anexo 2) es inferior a 0.05 hemos podido concluir que los datos de flujos de entrada de IED no siguen una distribución normal.

A continuación, se ha procedido a la realización del modelo 1, sobre el cuál hemos realizado una prueba para descartar la multicolinealidad en el modelo. Para ello hemos calculado el factor de inflación de la varianza, el cuál cuantifica la intensidad de la multicolinealidad en los MCO. El valor de referencia que hemos usado es 5, por lo que la variable dicotómica de la UE y el índice regulatorio de restricción a la IED han sido retirados del modelo debido a que tenían un VIF de 7.82 y 6 respectivamente.

Hemos llevado a cabo una prueba de autocorrelación de los residuos de Durbin Watson, el cuál se expresa de la siguiente forma:

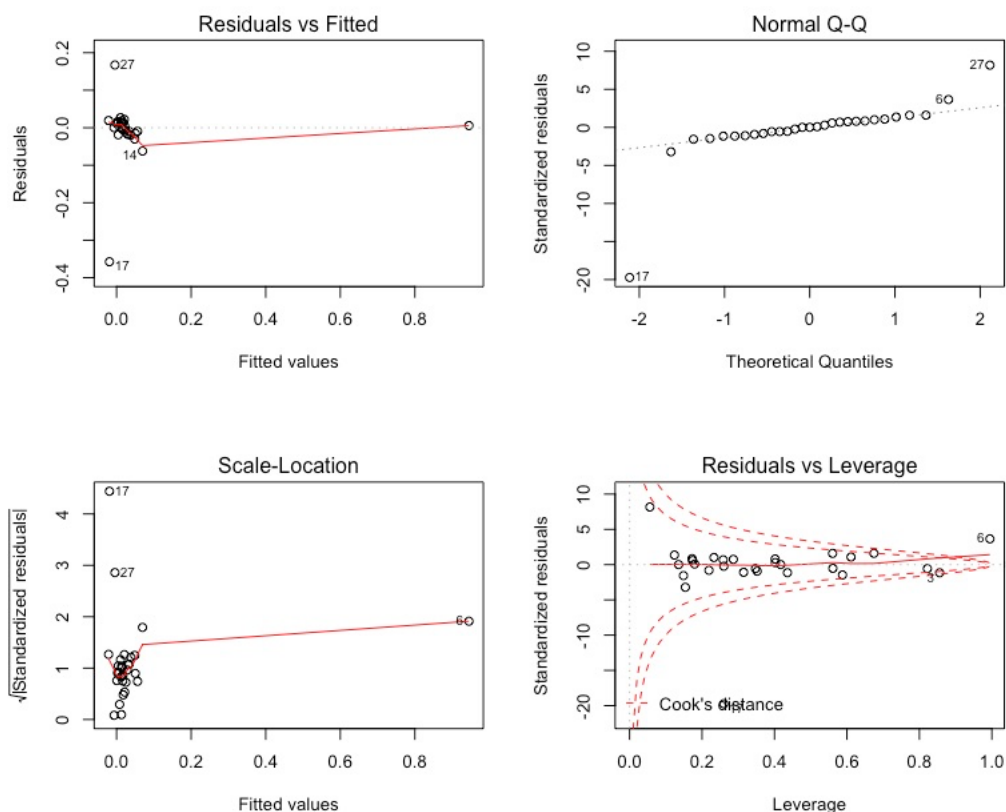
$$DW = \frac{\sum_{t=2}^n (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^n \hat{u}_t^2}$$

El p-valor de la prueba (Anexo 3) nos muestra un p-valor superior a 0.05, por lo que no podemos concluir que los residuos estén autocorrelacionados, por lo que podemos confirmar que el modelo cumple con la condición de no autocorrelación de los residuos.

Después hemos realizado el modelo 2, también utilizando métodos robustos, ya que nos permite realizar una inferencia que corrija los problemas provocados por los valores extremos.

En el modelo 2 podemos observar como ninguna variable muestra ser estadísticamente significativa. El análisis de residuos de este modelo es el siguiente:

Ilustración 3. Análisis de residuos del modelo 2



Al observar la forma de los residuos en los gráficos anteriores, hemos realizado una prueba de heterocedasticidad y así poder concluir si hay homocedasticidad en la varianza de los residuos o no. Para ello hemos llevado a cabo la prueba Breusch-Pagan también conocida como LM test. El cuál se expresa de la siguiente forma:

$$LM = nR^2$$

dónde es el coeficiente de determinación.

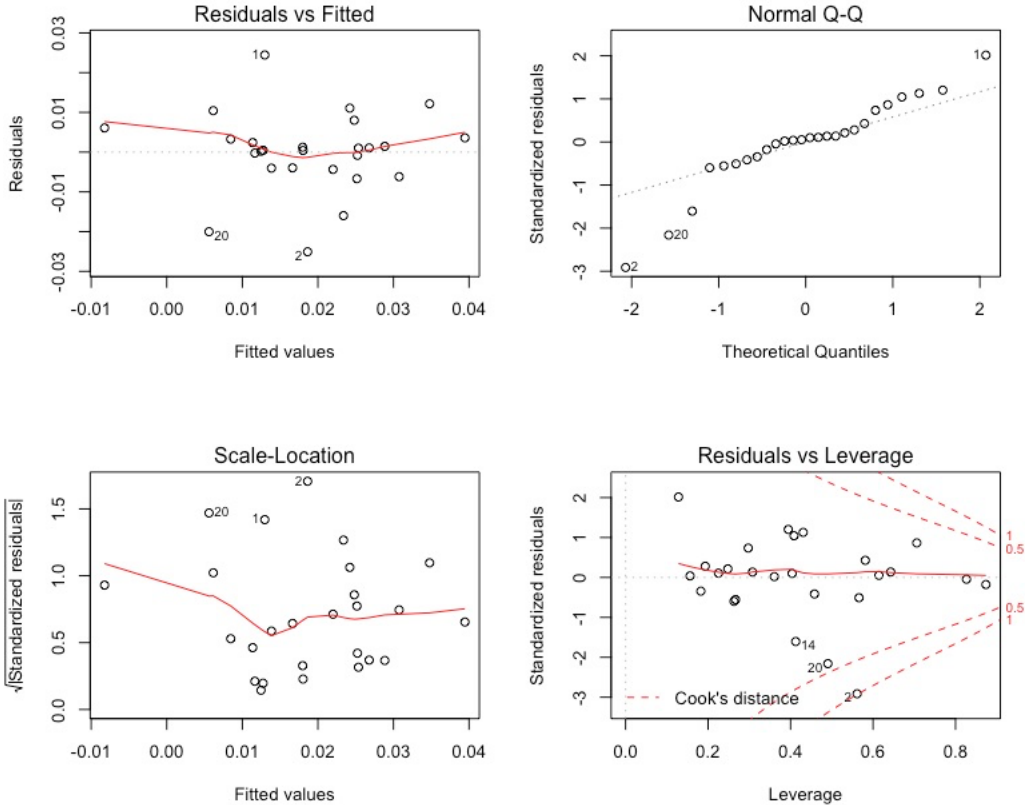
El p-valor asociado a la prueba es superior a 0.05, por lo que podemos concluir que el modelo no tiene problemas de heterocedasticidad.

Por último, y lo que se considera más importante para esta aplicación del modelo es el impacto que están teniendo los valores extremos en la inferencia del modelo. Por esta razón hemos decidido eliminar de la muestra los valores extremos de la variable dependiente (esta acción se podría haber hecho al inicio de la realización de los modelos al igual que la prueba de normalidad). Aunque de esta forma nos permite mostrar cuán sensible es este método a valores extremos.

Para ello hemos utilizado Rosner Test, el cuál ha identificado que las observaciones de Estonia, Luxemburgo y Noruega eran valores extremos para la variable de flujos de entrada de IED (Anexo 5). Una vez eliminadas dichas observaciones a pesar de no ser adecuado eliminar más

del 5% de las observaciones de la muestra, volvemos a observar en el modelo 3 como el modelo sigue siendo sensible a valores extremos, tal y como se puede observar en el análisis de los residuos del modelo.

Ilustración 4. Análisis de residuos del modelo 3



En relación con el análisis de la significación de las variables, los 3 modelos robustos a penas explican las diferencias en la IED. Ya que en ninguno de los casos son variables estadísticamente significativas.

Este modelo nos permite extraer dos conclusiones principales. La primera y más importante es el problema que ocasionan los valores extremos en muestras que no son muy grandes (normalmente en estudios de variables macroeconómicas) ya que hay países finitos. Y esto se ensalzado por el hecho de que la muestra de este estudio contiene países con características económicas muy diversas. Como es el caso de la economía de Estados Unidos en comparación por ejemplo con la de Hungría. Especializaciones económicas diferentes al igual que extensión geográfica y relaciones internacionales distintas, estas diferencias podrían ser estudiadas en estudios posteriores a este trabajo desde un punto de vista económico y político.

Por otra parte, y lo que también resulta sorprendente es que a pesar de que la estimación sufre de valores extremos ninguna de las variables que a priori resultan económicamente relevantes para la explicación de atracción de IED han mostrado significación estadística (Denisia, 2010).

Una de las posibles explicaciones de la no significación de estas variables es que no se está teniendo en cuenta el componente temporal. Estos efectos son estudiados en el siguiente apartado en la introducción de un análisis de datos panel.

4.2. Resultados modelo de regresión lineal múltiple por MCO para datos panel

En relación con los países observados, hay países que ya no forman parte del modelo puesto que tenían muchos valores no observados de la variable dependiente, por lo que se ha optado a retirarlos del estudio. Estos países son Chile, Estonia, Finlandia, Corea, Holanda, Suiza, España y Noruega.

Para el estudio de estos análisis se ha optado por la regresión por MCO de efectos fijos, puesto que el tamaño de la muestra es reducido y las observaciones no son aleatorias. Sino que son datos macroeconómicos de países en concreto que pertenecen a la OCDE y de los cuáles se disponen datos. El primer modelo es un modelo de regresión lineal por MCO agrupado. Para el estudio de efectos fijos, se han realizado dos modelos con efectos fijos, el segundo modelo toma variables *dummy* los países y el tercer modelo los diferentes periodos de tiempo. Los resultados de las dos regresiones son los siguientes:

Ilustración 5. Resultado de regresión lineal por MCO datos panel 1

```

Residuals:
    Min       1Q   Median       3Q      Max
-0.77982 -0.02380 -0.00131  0.02500  0.97025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.071e-01  1.757e-01   0.610   0.5427
IED_out2    -6.945e-02  7.117e-02  -0.976   0.3303
Prod        -3.317e-03  2.223e-03  -1.492   0.1372
Infl        2.176e-03  4.085e-03   0.533   0.5949
Salar       4.032e-06  2.511e-06   1.606   0.1098
Ind2        -6.644e-01  2.595e-01  -2.560   0.0112 *
Agri2       -1.256e+00  1.159e+00  -1.083   0.2799
Cons2       -2.325e-01  6.928e-01  -0.336   0.7375
Real_est2   -1.382e-01  5.575e-01  -0.248   0.8045
Imp2        2.335e-01  4.451e-02   5.247  3.8e-07 ***
UE          -9.406e-03  4.571e-02  -0.206   0.8372
desarrollo  5.898e-02  5.676e-02   1.039   0.3000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1201 on 208 degrees of freedom
Multiple R-squared:  0.3012,    Adjusted R-squared:  0.2642
F-statistic: 8.149 on 11 and 208 DF,  p-value: 8.379e-12

```


Tal y como se puede observar para el modelo agrupado hay variables que resultan estadísticamente significativas como es el caso de los impuestos a corporaciones en los que el signo del coeficiente es económicamente intuitivo. A mayor sean los impuestos a negocios menor será la IED, ya que los impuestos suelen desincentivar la IED.

Ilustración 6. Resultado de regresión lineal por MCO datos panel

Residuals:				
Min	1Q	Median	3Q	Max
-0.83526	-0.01514	0.00207	0.02188	0.92836
Coefficients: (2 not defined because of singularities)				
	Estimate	Std. Error	t value	Pr(> t)
IED_out2	-7.548e-02	7.656e-02	-0.986	0.325451
Prod	-1.917e-02	6.518e-03	-2.942	0.003666 **
Infl	9.081e-04	5.121e-03	0.177	0.859440
Salar	4.863e-06	6.340e-06	0.767	0.443982
Ind2	-1.438e+00	1.290e+00	-1.115	0.266248
Agri2	-8.094e-01	3.811e+00	-0.212	0.832041
Cons2	-1.184e+00	1.632e+00	-0.726	0.468979
Real_est2	-1.744e+00	1.738e+00	-1.003	0.316964
Imp2	6.525e-01	1.787e-01	3.652	0.000336 ***
UE	7.183e-01	5.091e-01	1.411	0.159872
desarrollo	7.454e-01	5.337e-01	1.396	0.164185
PaisAustria	5.073e-01	1.847e-01	2.747	0.006591 **
PaisBelgium	4.372e-01	2.039e-01	2.145	0.033241 *
PaisCzech Republic	9.891e-02	9.959e-02	0.993	0.321898
PaisDenmark	4.656e-01	2.057e-01	2.264	0.024707 *
PaisFrance	5.763e-01	2.360e-01	2.442	0.015527 *
PaisGermany	5.799e-01	1.832e-01	3.166	0.001801 **
PaisGreece	1.272e-01	1.421e-01	0.895	0.372070
PaisHungary	-1.021e-01	1.077e-01	-0.948	0.344424
PaisIceland	1.119e+00	6.041e-01	1.853	0.065404 .
PaisItaly	4.745e-01	1.674e-01	2.834	0.005099 **
PaisLatvia	-1.003e-01	1.298e-01	-0.773	0.440671
PaisLithuania	-5.281e-02	1.200e-01	-0.440	0.660358
PaisLuxembourg	3.653e-01	3.756e-01	0.973	0.331906
PaisMexico	NA	NA	NA	NA
PaisNew Zealand	9.923e-01	5.787e-01	1.715	0.087988 .
PaisPoland	2.776e-02	9.738e-02	0.285	0.775878
PaisPortugal	1.390e-01	1.223e-01	1.137	0.257000
PaisSlovak Republic	-3.661e-02	1.279e-01	-0.286	0.775066
PaisSlovenia	NA	NA	NA	NA
PaisUnited States	1.397e+00	6.191e-01	2.256	0.025216 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.1186 on 191 degrees of freedom				
Multiple R-squared: 0.4334, Adjusted R-squared: 0.3474				
F-statistic: 5.038 on 29 and 191 DF, p-value: 2.596e-12				

El segundo modelo recoge la información y las diferencias de IED entre los distintos países. Se puede concluir cómo hay países como Alemania, Italia o Austria son más propensos a tener una mayor IED que otros. Al igual que la variable productividad se vuelve estadísticamente significativa, aunque el signo de su coeficiente no es económicamente intuitivo, ya que a mayor sea la productividad de un país, más incentivos pueden tener los inversores para realizar IED en el país.

Ilustración 7. Resultado de regresión lineal por MCO datos panel 3

```
Residuals:
    Min       1Q   Median       3Q      Max
-0.72405 -0.03190  0.00003  0.02755  0.93576

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
IED_out2    -7.689e-02  7.534e-02  -1.021  0.3087
Prod        -3.576e-03  2.258e-03  -1.584  0.1148
Infl        2.169e-03  5.532e-03   0.392  0.6953
Salar       4.498e-06  2.549e-06   1.765  0.0791 .
Ind2        -6.369e-01  2.623e-01  -2.428  0.0161 *
Agri2       -1.186e+00  1.198e+00  -0.990  0.3235
Cons2       -2.281e-01  7.139e-01  -0.319  0.7497
Real_est2   -8.953e-02  5.645e-01  -0.159  0.8741
Imp2        2.438e-01  4.614e-02  5.284 3.33e-07 ***
UE          -4.696e-03  4.605e-02  -0.102  0.9189
desarrollo  6.399e-02  5.785e-02  1.106  0.2700
Year2007    7.960e-02  1.790e-01  0.445  0.6570
Year2008    7.434e-02  1.807e-01  0.411  0.6812
Year2009    8.509e-02  1.777e-01  0.479  0.6326
Year2010    1.048e-01  1.773e-01  0.591  0.5551
Year2011    1.072e-01  1.788e-01  0.599  0.5495
Year2012    9.288e-02  1.789e-01  0.519  0.6042
Year2013    7.992e-02  1.783e-01  0.448  0.6544
Year2014    7.816e-02  1.792e-01  0.436  0.6633
Year2015    6.996e-02  1.791e-01  0.391  0.6965
Year2016    1.094e-01  1.793e-01  0.610  0.5425
Year2017    1.670e-02  1.799e-01  0.093  0.9261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1205 on 198 degrees of freedom
Multiple R-squared:  0.3938,    Adjusted R-squared:  0.3264
F-statistic: 5.846 on 22 and 198 DF,  p-value: 1.361e-12
```

El último modelo muestra que no hay ningún año que sea significativamente diferente a otro. Esto no económicamente intuitivo puesto que el periodo de la muestra comprende los años de la crisis económica mundial de 2008 y la del EURO en 2013. Aún así esto puede ser debido a un problema de omisión de variables relevantes o a la complejidad en sí de explicar los determinantes de la IED.

Selección del modelo

Para la selección del modelo podemos observar que el modelo escogido es el de efectos fijo. Si hubiésemos computado el modelo de efectos aleatorios, tendría sentido aplicar el contraste de Hausman pero para nuestro tipo de datos no sería adecuado.

También podemos computar una prueba F para seleccionar si es preferible utilizar el modelo con efectos fijos individualizados o agrupado (Anexo 6). Para ello calculamos otro modelo que incluye los dos efectos fijos, y podemos concluir que ya que el p valor no es inferior a 0.05 no se puede incluir que los efectos fijos sean significativos.

4.3. Resultados del modelo gravitacional

El primer aspecto que es importante a resaltar es la dificultad de la realización del modelo gravitacional para los flujos netos de IED. Al igual que pasaba con los modelos anteriores, esta variable por definición puede tomar valores iguales a 0 o negativos, debido a que es la diferencia entre la inversión y la desinversión anual. Esto imposibilita tomar logaritmos de dicha variable y el posterior análisis de las elasticidades.

También es importante destacar que ciertas variables como los impuestos a corporaciones o salarios no son variables propias para incluir en el modelo gravitacional puesto que no implican una relación bilateral entre los países. Por lo que el objetivo de este modelo no es tanto analizar los causantes intrínsecos de la inversión sino las relaciones bilaterales en relación con la IED entre los países de la muestra.

Este modelo tal y como se ha explicado en el apartado de la metodología implica tomar logaritmos de las variables numéricas por lo que el modelo final que se ha llevado a cabo es el siguiente:

$$\log(IED_{in_{ij}}) = \log(dist_{ij}) + \log(PIB_i) + \log(PIB_j)$$

Los resultados tras aplicar el método de MCO para la estimación de los parámetros, es el siguiente:

Ilustración 8. Resultados del Gravity Model

```
Call:
lm_robust(formula = log(IED_in) ~ log(dist) + log(PIB1) + log(PIB2),
          data = DatTFG3, clusters = dist, se_type = "stata")

Standard error type: stata

Coefficients:
              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept) -10.3744    3.9020  -2.659 1.075e-02 -18.2286  -2.5201 46
log(dist)    -0.9984    0.2986  -3.344 1.649e-03  -1.5994  -0.3975 46
log(PIB1)     0.8063    0.1068   7.547 1.388e-09   0.5913   1.0214 46
log(PIB2)     0.9300    0.1982   4.694 2.439e-05   0.5312   1.3289 46

Multiple R-squared:  0.3843 , Adjusted R-squared:  0.3763
F-statistic: 29.7 on 3 and 46 DF, p-value: 7.804e-11
```

Este modelo no incluye efectos fijos determinados aleatorios, aunque es importante saber que existe la posibilidad de incluirlos, al igual que utilizar distintas extensiones del modelo gravitacional para solucionar ciertos problemas de endogeneidad como introducir variables instrumentales en la regresión (Shepherd, 2013).

Cómo podemos observar en la Ilustración 8 no hay variables que sean estadísticamente significativas. Esto puede ser debido a la falta de observaciones en la base de datos utilizadas y que al igual que en los modelos anteriores puede estar afectada por valores extremos y problemas de heterocedasticad. Algo que es muy común en este tipo de modelos.

5. Conclusiones

El método de los MCO es una técnica comúnmente usada para la estimación de los parámetros de regresiones lineales tanto simples como múltiples. Aunque también puede ser usado para la estimación de parámetros en modelos gravitacionales tal y como este estudio ha enseñado en la realización de los tres modelos.

Se puede observar que, aunque la regresión lineal es un método que es usado comúnmente para el análisis de estudios socioeconómicos, las condiciones que implican pueden ser difíciles de conseguir con datos reales. Este trabajo ha permitido analizar el efecto que tiene en las estimaciones el hecho de tener valores extremos en la muestra y como afectaba esto a la inferencia e imposibilita la obtención de un modelo fiable. Esta sensibilidad a dichos valores es mucho mayor en estudios de datos de corte transversal como en estudios de datos panel.

También es importante destacar el uso de efectos fijos en regresiones lineales con datos panel en aquellas muestras que no son escogidas de forma aleatoria, lo que implica que el modelo sea susceptible a tener problemas de heterocedasticidad. En relación con la selección de modelos para analizar fenómenos económicos hemos podido ver como hay una batería amplia de distintos modelos que se pueden aplicar, aunque cada uno tiene sus propias limitaciones. Como el caso de los modelos gravitacionales los cuáles nos permiten estudiar las relaciones bilaterales de la variable sujeta estudio, en nuestro caso la IED, pero por lo contrario nos imposibilita estudiar ciertas variables que son intrínsecas de cada país como es por ejemplo estudiar la productividad. Aunque esto último se pudiese incluir de forma dicotómica, no se podría estudiar el efecto de la productividad en términos de elasticidad.

Igualmente es importante destacar las limitaciones del estudio, las cuáles radican principalmente en la accesibilidad a los datos y la naturaleza de estos. En lo que se refiere a la accesibilidad de estos, hemos podido observar como ciertos países como Holanda no proporcionan los datos a la OCDE de sus flujos de IED, o en el caso de Turquía el cuál no proporciona sus salarios por motivos de confidencialidad. Y en relación con la naturaleza de los datos, estos han imposibilitado en algunos casos introducir el país en el estudio puesto que sus flujos de IED eran negativos o iguales a 0, lo que imposibilita la toma de logaritmos de tales variables. Esto afecta especialmente en la realización del modelo gravitacional ya que la escala logarítmica es esencial para la realización del modelo.

Por último, cabe destacar la dificultad en explicar los factores de la IED de manera estadística, ya que, muchos determinantes que según la teoría económica son decisivos para una mayor atracción de IED, como es el caso de los salarios o los sectores de especialización de la economía, no han mostrado tener una gran influencia estadísticamente hablando en las diferencias de IED entre los distintos países.

6. Bibliografía

Banco Mundial (n.d). *Inversión extranjera directa, entrada neta de capital (balanza de pagos, US\$ a precios actuales*. Último acceso el 24 de junio de 2021. Obtenido de: <https://datos.bancomundial.org/indicador/BX.KLT.DINV.CD.WD>

Denisia, V. (2010). Foreign direct investment theories: An overview of the main FDI theories. *European journal of interdisciplinary studies*, (3).

Economou, F., Hassapis, C., Philippas, N., & Tsionas, M. (2017). Foreign direct investment determinants in OECD and developing countries. *Review of Development Economics*, 21(3), 527-542.

Faraway, J. J. (2004). *Linear models with R*. Chapman and Hall/CRC.

Gil, C. (2018). Regresión Lineal Simple. Rpubs by RStudio. Último acceso el 24 de junio de 2021. Obtenido de https://rpubs.com/Cristina_Gil/Regresion_Lineal_Simple

Head, K., Mayer, T. & Ries, J. (2010), The erosion of colonial trade linkages after independence *Journal of International Economics*, 81(1):1-14

Hill, R. C., Griffiths, W. E., & Lim, G. C. (2018). *Principles of econometrics*. John Wiley & Sons.

Mogrovejo, J. (2005). Factores determinantes de la inversión extranjera directa en algunos países de Latinoamérica. *Revista latinoamericana de desarrollo economico*, (5), 51-82.

Moudatsou, A., & Kyrkilis, D. (2011). FDI and Economic Growth: Causality for the EU and ASEAN. *Journal of Economic Integration*, 554-577.

OCDE.Stat. (n.d). Metadata for OECD countries and selected non-member economies. Último acceso el 24 de junio de 2021. Obtenido de: <https://stats.oecd.org>

OECD. (n.d). *Foreign direct investment statistics. Explanatory notes*. Último acceso el 24 de junio de 2021. Obtenido de: <https://www.oecd.org/daf/inv/FDI-statistics-explanatory-notes.pdf>

OECD. (n.d). Gross Domestic Product (GDP). Último acceso el 24 de junio de 2021. Obtenido de: <https://stats.oecd.org/index.aspx?queryid=60702>

OECD. (nd). FDI Regulatory Restrictiveness Index. . Último acceso el 24 de junio de 2021. Obtenido de: <https://www.oecd.org/investment/fdiindex.htm>

Sánchez, J. T., & González, C. M. (2004). Los determinantes de la inversión extranjera directa en la UE y los PECO. *ICE, Revista de Economía*, (814).

Shepherd, B. (2013). Gravity model of international trade: A user guide.

UNCTAD. (2019) “World Investment Report”. Último acceso el 24 de junio de 2021. Obtenido de: https://unctad.org/en/PublicationsLibrary/wir2019_en.pdf

Wölwer, A. L., Burgard, J. P., Kunst, J., & Vargas, M. (2018). Gravity: Estimation Methods for Gravity Models in R. *Journal of Open Source Software*, 3(31), 1038.

Wooldridge, J. (2015). Introduction to Econometrics, chapter 11 and 12. *Cengage Learning*.

Anexos

Anexo 1. Matriz de correlaciones.

##	DatTFG.IED_in2	DatTFG.IED_out2	DatTFG.Ind2	DatTFG.Agri2
##	DatTFG.IED_in2	1.00000000	0.83374233	0.198197499
##	DatTFG.IED_out2	0.83374233	1.00000000	-0.069212244
##	DatTFG.Ind2	0.19819750	-0.06921224	1.000000000
##	DatTFG.Agri2	0.04449988	-0.13401202	0.038706791
##	DatTFG.Cons2	0.10083280	0.07796618	0.209099785
##	DatTFG.Real_est2	0.01681467	-0.05106605	-0.485114578
##	DatTFG.Exp2	-0.19964250	0.24218956	-0.203199881
##	DatTFG.Imp2	-0.17208322	0.25404913	-0.206228348
##	DatTFG.Prod	-0.10209675	-0.19113285	-0.004296417
##	DatTFG.Infl	0.09546823	0.08140138	0.135861987
##	DatTFG.Reg	-0.06267748	-0.18964494	0.052692767
##	DatTFG.Salar	-0.29565758	-0.10322661	-0.379363470
##	DatTFG.Impu2	-0.36453089	-0.14511781	-0.041199352
##	DatTFG.Cons2	DatTFG.Real_est2	DatTFG.Exp2	DatTFG.Imp2
##	DatTFG.IED_in2	0.10083280	0.01681467	-0.19964250
##	DatTFG.IED_out2	0.07796618	-0.05106605	0.24218956
##	DatTFG.Ind2	0.20909979	-0.48511458	-0.20319988
##	DatTFG.Agri2	0.36795468	0.19444975	-0.33158625
##	DatTFG.Cons2	1.00000000	-0.36920806	-0.05432035
##	DatTFG.Real_est2	-0.36920806	1.00000000	-0.41423308
##	DatTFG.Exp2	-0.05432035	-0.41423308	1.00000000
##	DatTFG.Imp2	-0.03067275	-0.39915894	0.99281051
##	DatTFG.Prod	0.23004285	0.07703474	-0.28001475
##	DatTFG.Infl	0.56134866	-0.16113051	-0.13635429
##	DatTFG.Reg	0.39847804	0.23588744	-0.37414547
##	DatTFG.Salar	-0.07899812	-0.03813137	0.15571117
##	DatTFG.Impu2	0.17143561	-0.04765921	0.29498940
##	DatTFG.Prod	DatTFG.Infl	DatTFG.Reg	DatTFG.Salar
##	DatTFG.IED_in2	-0.102096751	0.09546823	-0.06267748
##	DatTFG.IED_out2	-0.191132852	0.08140138	-0.18964494
##	DatTFG.Ind2	-0.004296417	0.13586199	0.05269277
##	DatTFG.Agri2	-0.067173473	0.59625613	0.29088434
##	DatTFG.Cons2	0.230042854	0.56134866	0.39847804
##	DatTFG.Real_est2	0.077034742	-0.16113051	0.23588744
##	DatTFG.Exp2	-0.280014750	-0.13635429	-0.37414547
##	DatTFG.Imp2	-0.302158801	-0.10153544	-0.38035848
##	DatTFG.Prod	1.000000000	0.18021614	0.45580872
##	DatTFG.Infl	0.180216141	1.00000000	0.13792292
##	DatTFG.Reg	0.455808721	0.13792292	1.00000000
##	DatTFG.Salar	0.210775118	-0.24356663	0.15676385
##	DatTFG.Impu2	-0.057941347	-0.24102636	0.34805536
##	DatTFG.Salar	DatTFG.Impu2	DatTFG.Salar	DatTFG.Impu2
##	DatTFG.Salar	0.08499679	0.26895444	1.00000000
##	DatTFG.Impu2	0.26895444	1.00000000	1.00000000

Anexo 2. Shapiro Test para IED_in2

Shapiro-Wilk normality test

data: DatTFG\$IED_in2
 W = 0.41472, p-value = 1.117e-09

Anexo 3. Durbin-Watson modelo 1

```
lag Autocorrelation D-W Statistic p-value
1 -0.07985515 2.155215 0.738
Alternative hypothesis: rho != 0
> |
```

Anexo 4. Breusch Pagan

```
studentized Breusch-Pagan test

data: modelo2
BP = 9.6923, df = 10, p-value = 0.4679
```

Anexo 5. Rosner Test

```
Number of Outliers Detected: 3

  i   Mean.i   SD.i   Value Obs.Num   R.i+1 lambda.i+1 Outlier
1 0 0.042278829 0.19220413 0.9505454     6 4.725531 2.892705 TRUE
2 1 0.009840738 0.08164454 -0.3765540    17 4.732646 2.876209 TRUE
3 2 0.024151655 0.03109964 0.1628727    27 4.460537 2.858923 TRUE
```

Anexo 6. F-test Modelo agrupado vs. Efectos fijos

```
F test for individual effects

data: IED_in2 ~ IED_out2 + Prod + Infl + Salar + Ind2 + Agri2 + Cons2 + ...
F = 1.3257, df1 = 17, df2 = 191, p-value = 0.1799
alternative hypothesis: significant effects
```

Anexo 7. Código R

```
#Libraries
library(readxl)
library(ggplot2)
library(knitr)
library(kableExtra)
library(plotly)
library(car)
library(caTools)
library(VGAM)
library(AER)
library(truncreg)
library(censReg)
library(GGally)
library(texreg)
library(cowplot)
library(sampleSelection)
library(texreg)
```

```

library(ROCR)

#Lectura de la base de datos
library(readxl)
DatTFG <- read_excel("~/Desktop/Analisis TFG/DatTFG.xlsx",
  col_types = c("text", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric",
  "numeric", "numeric", "numeric","numeric", "numeric", "numeric", "numeric", "numeric"))
#View(DatTFG)

#Valoración de NAs
apply(is.na(DatTFG),2,which)#Para observar la columna en la que se situa el valor missing

apply(is.na(DatTFG),2,mean) #Para ver el porcentaje de missings que tiene cada variable (columna)

###Decidimos sustituir los valores missing para Estonia (tupla 6) y Turquía (tupla 29), por la media

#Cambiamos el valor de la IED de inflow
DatTFG$IED_in[DatTFG$Pais == "Estonia"] <- mean(DatTFG$IED_in, na.rm=T)

#Cambiamos el valor de la IED de outflow
DatTFG$IED_out[DatTFG$Pais == "Estonia"] <- mean(DatTFG$IED_out, na.rm=T)

#Corregimos error de variable Salar por el valor de la media en caso de que haya algun valor negativo
DatTFG[is.na(DatTFG$Salar),"Salar"]<- mean(DatTFG$Salar, na.rm=T)

apply(is.na(DatTFG),2,which)

#Eliminamos la observación de holanda, en realidad la siguiente funcion solo escogera las tuplas en
las que se disponga de toda la informacion, por l que eliminamos las observaciones de Holanda y
Estonia

DatTFG<- DatTFG[complete.cases(DatTFG),]
#View(DatTFG)

#Transformacion a variables en terminos de GDP
IED_in2<- DatTFG$IED_in/DatTFG$PIB #flujo de IED de entrada en terminos de PIB
IED_out2<- DatTFG$IED_out/DatTFG$PIB #flujo de IED de salida en terminos de PIB
Ind2<- DatTFG$Ind/DatTFG$PIB #proporcion del sector de la industria en la economía
Agri2<- DatTFG$Agri/DatTFG$PIB #proporcion del sector de la agricultura en la economía
Cons2<- DatTFG$Cons/DatTFG$PIB #proporcion del sector de la construcción en la economía
Real_est2<- DatTFG$Real_est/DatTFG$PIB #proporcion del sector de los bienes raíz en la economía
Exp2<- DatTFG$Exp/DatTFG$PIB #proporcion de las exportaciones en la economía
Imp2<- DatTFG$Imp/DatTFG$PIB #proporcion de las importaciones en la economía
Impu2<- DatTFG$Impu/DatTFG$PIB #proporcion de los impuestos a corporaciones en la economía
DatTFG<-cbind(DatTFG,IED_in2,IED_out2,Ind2,Agri2, Cons2, Real_est2, Exp2,Imp2, Impu2)

library (knitr) #para poder generar tablas con formato html

#Tabla comparativa valores relativos y absolutos
absolutos_relativos <- data.frame(Variables = c("IED_in","IED_out", "Ind", "Agri","Cons",
"Real_est",
"Exp", "Imp", "Impu"),

```

Máximo_absoluto= c (DatTFG[which.max(DatTFG\$IED_in),1],
 DatTFG[which.max(DatTFG\$IED_out),1],
 DatTFG[which.max(DatTFG\$Ind),1],
 DatTFG[which.max(DatTFG\$Agri),1],
 DatTFG[which.max(DatTFG\$Cons),1],
 DatTFG[which.max(DatTFG\$Real_est),1],
 DatTFG[which.max(DatTFG\$Exp),1],
 DatTFG[which.max(DatTFG\$Imp),1],
 DatTFG[which.max(DatTFG\$Impu),1]),

V1 = c(max(DatTFG\$IED_in),max(DatTFG\$IED_out),
 max(DatTFG\$Ind),max(DatTFG\$Agri),
 max(DatTFG\$Cons),max(DatTFG\$Real_est),
 max(DatTFG\$Exp),max(DatTFG\$Imp),
 max(DatTFG\$Impu)),

Máximo_relativo = c (DatTFG[which.max(DatTFG\$IED_in2),1],
 DatTFG[which.max(DatTFG\$IED_out2),1],
 DatTFG[which.max(DatTFG\$Ind2),1],
 DatTFG[which.max(DatTFG\$Agri2),1],
 DatTFG[which.max(DatTFG\$Cons2),1],
 DatTFG[which.max(DatTFG\$Real_est2),1],
 DatTFG[which.max(DatTFG\$Exp2),1],
 DatTFG[which.max(DatTFG\$Imp2),1],
 DatTFG[which.max(DatTFG\$Impu2),1]),

V2 = c(max(DatTFG\$IED_in2),max(DatTFG\$IED_out2),
 max(DatTFG\$Ind2),max(DatTFG\$Agri2),
 max(DatTFG\$Cons2),max(DatTFG\$Real_est2),
 max(DatTFG\$Exp2),max(DatTFG\$Imp2),
 max(DatTFG\$Impu2)),

Mínimo_absoluto= c (DatTFG[which.min(DatTFG\$IED_in),1],
 DatTFG[which.min(DatTFG\$IED_out),1],
 DatTFG[which.min(DatTFG\$Ind),1],
 DatTFG[which.min(DatTFG\$Agri),1],
 DatTFG[which.min(DatTFG\$Cons),1],
 DatTFG[which.min(DatTFG\$Real_est),1],
 DatTFG[which.min(DatTFG\$Exp),1],
 DatTFG[which.min(DatTFG\$Imp),1],
 DatTFG[which.min(DatTFG\$Impu),1]),

V3 = c(min(DatTFG\$IED_in),min(DatTFG\$IED_out),
 min(DatTFG\$Ind),min(DatTFG\$Agri),
 min(DatTFG\$Cons),min(DatTFG\$Real_est),
 min(DatTFG\$Exp),min(DatTFG\$Imp),
 min(DatTFG\$Impu)),

Mínimo_relativo = c (DatTFG[which.min(DatTFG\$IED_in2),1],
 DatTFG[which.min(DatTFG\$IED_out2),1],
 DatTFG[which.min(DatTFG\$Ind2),1],
 DatTFG[which.min(DatTFG\$Agri2),1],
 DatTFG[which.min(DatTFG\$Cons2),1],
 DatTFG[which.min(DatTFG\$Real_est2),1],
 DatTFG[which.min(DatTFG\$Exp2),1]),

```

DatTFG[which.min(DatTFG$Imp2),1],
DatTFG[which.min(DatTFG$Impu2),1]),

V4 = c(min(DatTFG$IED_in2),min(DatTFG$IED_out2),
      min(DatTFG$Ind2),min(DatTFG$Agri2),
      min(DatTFG$Cons2),min(DatTFG$Real_est2),
      min(DatTFG$Exp2),min(DatTFG$Imp2),
      min(DatTFG$Impu2)))

kable(absolutos_relativos, align="c",
      booktabs=TRUE) %>%
kable_styling(position = "center") %>%
row_spec(0, bold=TRUE, background = "#d6d6d6") %>%
column_spec(1, bold=TRUE)

#Tabla estadísticas principales para variables numéricas

library(e1071) #para calcular estadísticas como kurtosis

#Tabla estadísticas
estadisticos <- data.frame(Variables = c("IED_in", "IED_out", "Industria", "Agricultura",
"Construcción", "Bienes raíces",
      "Exportaciones", "Importaciones", "Productividad", "Inflación", "Índice de
restricción",
      "Salario", "Impuesto"),

      Mínimo = c (min(DatTFG$IED_in2), min(DatTFG$IED_out2), min(DatTFG$Ind2),
min(DatTFG$Agri2),
      min(DatTFG$Cons2), min(DatTFG$Real_est2), min(DatTFG$Exp2),
min(DatTFG$Imp2),
      min(DatTFG$Prod), min(DatTFG$Infl), min(DatTFG$Reg),
min(DatTFG$Salar),
      min(DatTFG$Impu2)),

      Media = c (mean(DatTFG$IED_in2), mean(DatTFG$IED_out2),
mean(DatTFG$Ind2), mean(DatTFG$Agri2),
      mean(DatTFG$Cons2), mean(DatTFG$Real_est2), mean(DatTFG$Exp2),
mean(DatTFG$Imp2),
      mean(DatTFG$Prod), mean(DatTFG$Infl), mean(DatTFG$Reg),
mean(DatTFG$Salar),
      mean(DatTFG$Impu2)),

      Mediana = c (median(DatTFG$IED_in2), median(DatTFG$IED_out2),
median(DatTFG$Ind2),
      median(DatTFG$Agri2), median(DatTFG$Cons2),
median(DatTFG$Real_est2),
      median(DatTFG$Exp2), median(DatTFG$Imp2), median(DatTFG$Prod),
median(DatTFG$Infl),
      median(DatTFG$Reg), median(DatTFG$Salar), median(DatTFG$Impu2)),

      Máximo = c (max(DatTFG$IED_in2), max(DatTFG$IED_out2),
max(DatTFG$Ind2), max(DatTFG$Agri2),
      max(DatTFG$Cons2), max(DatTFG$Real_est2), max(DatTFG$Exp2),
max(DatTFG$Imp2),

```

```

max(DatTFG$Prod), max(DatTFG$Infl), max(DatTFG$Reg),
max(DatTFG$Salar),
max(DatTFG$Impu2)),

DT = c (sd(DatTFG$IED_in2), sd(DatTFG$IED_out2), sd(DatTFG$Ind2),
sd(DatTFG$Agri2),
sd(DatTFG$Cons2), sd(DatTFG$Real_est2), sd(DatTFG$Exp2),
sd(DatTFG$Imp2),
sd(DatTFG$Prod), sd(DatTFG$Infl), sd(DatTFG$Reg), sd(DatTFG$Salar),
sd(DatTFG$Impu2)),

```

```

Kurtosis = c (kurtosis(DatTFG$IED_in2), kurtosis(DatTFG$IED_out2),
kurtosis(DatTFG$Ind2),
kurtosis(DatTFG$Agri2), kurtosis(DatTFG$Cons2),
kurtosis(DatTFG$Real_est2),
kurtosis(DatTFG$Exp2), kurtosis(DatTFG$Imp2),kurtosis(DatTFG$Prod),
kurtosis(DatTFG$Infl), kurtosis(DatTFG$Reg),
kurtosis(DatTFG$Salar),kurtosis(DatTFG$Impu2)))

```

```

kable(estadisticos, digits=2, align="c",
booktabs=TRUE) %>%
kable_styling(position = "center") %>%
row_spec(0, bold=TRUE, background = "#d6d6d6") %>%
column_spec(1, bold=TRUE)

```

#Boxplots de las variables numéricas

```

Conf3x4 = matrix(c(1:12), nrow=4, byrow=TRUE)
layout(Conf3x4)
layout.show(12)
boxplot(DatTFG$IED_in2, main='Entrada IED')
boxplot(DatTFG$IED_out2, main='Salida IED')
boxplot(DatTFG$Ind2, main='Industria')
boxplot(DatTFG$Agri2, main='Agrario')
boxplot(DatTFG$Cons2, main='Construcción')
boxplot(DatTFG$Real_est2, main='Bienes raíces')
boxplot(DatTFG$Exp2, main='Exportaciones')
boxplot(DatTFG$Imp2, main='Importaciones')
boxplot(DatTFG$Impu2, main='Impuestos')
boxplot(DatTFG$Salar, main='Salarios')
boxplot(DatTFG$Infl, main='Inflación')
boxplot(DatTFG$Prod, main='Prod')

```

#Histogramas

```

Conf3x4 = matrix(c(1:12), nrow=4, byrow=TRUE)
layout(Conf3x4)
layout.show(12)
hist(DatTFG$IED_in2, main='Entrada IED')
hist(DatTFG$IED_out2, main='Salida IED')
hist(DatTFG$Ind2, main='Industria')
hist(DatTFG$Agri2, main='Agrario')
hist(DatTFG$Cons2, main='Construcción')
hist(DatTFG$Real_est2, main='Bienes raíces')
hist(DatTFG$Exp2, main='Exportaciones')
hist(DatTFG$Imp2, main='Importaciones')

```

```

hist(DatTFG$Impu2, main='Impuestos')
hist(DatTFG$Salar, main='Salarios')
hist(DatTFG$Infl, main='Inflación')
hist(DatTFG$Prod, main='Prod')

#Grafico de correlaciones
datacorr <- data.frame(DatTFG$IED_in2,DatTFG$IED_out2, DatTFG$Ind2, DatTFG$Agri2,
DatTFG$Cons2, DatTFG$Real_est2, DatTFG$Exp2, DatTFG$Imp2, DatTFG$Prod, DatTFG$Infl,
DatTFG$Reg, DatTFG$Salar, DatTFG$Impu2 ) #creamos un data frame con las variables numericas
correlacion<-cor(datacorr) #correlation matrix
correlacion

plot_ly( type="heatmap", x=colnames(correlacion), y=rownames(correlacion), z=correlacion,
colors='Blues')

#Tabla de estadisticos variables ctaegoricas

#UE
data_aux<- DatTFG$UE
data_aux<- factor(data_aux, labels=c("No UE","UE"))
estadisticas2<-as.data.frame(summary(data_aux, na.rm=TRUE))

estadisticas2<-cbind(estadisticas2, c(round(estadisticas2[,1]/ sum(estadisticas2[,1]),4)))
colnames(estadisticas2)<-c( "Frecuencia Absoluta", "Frecuencia Relativa")

kable(estadisticas2, digits=2, align="c",
      booktabs=TRUE) %>%
  kable_styling(position = "center") %>%
  row_spec(0, bold=TRUE, background = "#d6d6d6") %>%
  column_spec(1, bold=TRUE)

#En desarrollo

data_aux<- DatTFG$desarrollo
data_aux<- factor(data_aux, labels=c("País desarrollado","País en vias de desarrollo"))
estadisticas2b<-as.data.frame(summary(data_aux, na.rm=TRUE))

estadisticas2b<-cbind(estadisticas2b, c(round(estadisticas2b[,1]/ sum(estadisticas2b[,1]),4)))

colnames(estadisticas2b)<-c( "Frecuencia Absoluta", "Frecuencia Relativa")

kable(estadisticas2b, digits=2, align="c",
      booktabs=TRUE) %>%
  kable_styling(position = "center") %>%
  row_spec(0, bold=TRUE, background = "#d6d6d6") %>%
  column_spec(1, bold=TRUE)

#Gráfico de las variables categoricas
ggplot(DatTFG, aes(x= UE)) +
  geom_bar(fill="#56B4E9") +
  stat_count(geom = "text",
            aes(label = paste(round(..count..)/sum(..count..)*100, "%")),
            position=position_fill(vjust=0.5), colour="black") +
  scale_x_discrete(labels=c("No UE","UE"))

```

```

ggplot(DatTFG, aes(x= desarrollo)) +
  geom_bar(fill="#56B4E9") +
  stat_count(geom = "text",
            aes(label = paste(round(..count..)/sum(..count..)*100, "%")),
            position=position_fill(vjust=0.5), colour="black") +

  scale_x_discrete(labels=c("País desarrollado", "País en vías de desarrollo"))

#####
# REGRSIÓN LINEAL CORTE TRANSVERSAL #
#####

#Test normalidad de las distintas variables

shapiro.test(DatTFG$IED_in2)

library(MASS)
modelo1<- rlm (IED_in2 ~ IED_out2+ Prod+ Infl+ Reg+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+
Imp2 + UE + desarrollo, data= DatTFG )
vif(modelo1) #Hemos de quitar las variables de regulacion y UE

modelo2<- rlm (IED_in2 ~ IED_out2+ Prod+ Infl+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+ Imp2 +
desarrollo, data= DatTFG )

#Test multicolinealidad
vif(modelo2)

#Test autocorrelacion
library(lmtest)
library(car)
durbinWatsonTest(modelo1,simulate = TRUE,reps = 1000)

#No se puede concluir con un p-value de 0.97 que haya problemas de autocorrelacion
Conf2x2 = matrix(c(1:4), nrow=2, byrow=TRUE)
layout(Conf2x2)
layout.show(4)

plot(modelo2)

#Test homogeniedad de varianzas
library(lmtest) #Test de Breusch-Pagan
bptest( modelo2 ) #Podemos concluir homogeniedad en los residuos

#En el gráfico se puede detectar presencia de outliers
#Test de Rosner k=3 se sospecha de la existencia de 3 outliers
library(EnvStats)
test <- rosnerTest(DatTFG$IED_in2,
  k = 3)
test

#la observacion 6, 17 y 17 son outliers, los eliminamos de la base de datos
data2<- DatTFG[-c(6, 17, 27),]

```

```

modelo3<- rlm (IED_in2 ~ IED_out2+ Prod+ Infl+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+ Imp2 +
desarrollo, data= data2 )
vif(modelo3)
plot(modelo3)

```

```
## Tabla de resultados modelo 1 modelo 2 modelo 3
```

```

summary(modelo1)
summary(modelo2)
summary(modelo3)

```

```

#####
# REGRSIÓN LINEAL DATOS PANEL EF. F #
#####

```

```

DatTFG2 <- read_excel("~/Desktop/Análisis TFG/DatTFG2.xlsx", col_types = c("text", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric",
"numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric"))

```

```

library(AER)
library(plm)

```

```
DatTFG2
```

```

IED_in2<- DatTFG2$IED_in/DatTFG2$PIB
IED_out2<- DatTFG2$IED_out/DatTFG2$PIB
Ind2<- DatTFG2$Ind/DatTFG2$PIB
Agri2<- DatTFG2$Agri/DatTFG2$PIB
Cons2<- DatTFG2$Cons/DatTFG2$PIB
Real_est2<- DatTFG2$Real_est/DatTFG2$PIB
Exp2<- DatTFG2$Exp/DatTFG2$PIB
Imp2<- DatTFG2$Imp/DatTFG2$PIB
Impu2<- DatTFG2$Impu/DatTFG2$PIB

```

```

DatTFG2<-cbind(DatTFG2,IED_in2,IED_out2,Ind2,Agri2, Cons2, Real_est2, Exp2,Imp2, Impu2)
DatTFG2$pais<- factor(DatTFG2$pais)

```

```

library(car)
scatterplot(IED_in2~ Year|Pais, data=DatTFG2 )

```

```

#Heterogeneidad entre los paises
library(gplots)
plotmeans(IED_in2 ~ Pais, data=DatTFG2 )

```

```

#Heterogeneidad entre el tiempo
plotmeans(IED_in2 ~ Year, data=DatTFG2 )

```

```

DatTFG2$Year<- factor(DatTFG2$Year)
#Modelo lineal
panelmodel1<- lm(IED_in2~ IED_out2+ Prod+ Infl+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+ Imp2
+ UE + desarrollo, data = DatTFG2 )
#Modelos con efectos fijos de estado
panelmodel2<- lm(IED_in2~ IED_out2+ Prod+ Infl+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+ Imp2
+ UE + desarrollo + Pais -1, data = DatTFG2 )

```



```

#Modelo con efectos fijo de tiempo
panelmodel3<- lm(IED_in2~ IED_out2+ Prod+ Infl+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+ Imp2
+ UE + desarrollo + Year -1, data = DatTFG2 )

library (plm)

panelmodel4<- plm(IED_in2~ IED_out2+ Prod+ Infl+ Salar+ Ind2+ Agri2+ Cons2+ Real_est2+ Imp2
+ UE + desarrollo , index=c("Pais", "Year"), model="within", data=DatTFG2)

pFtest(panelmodel4, panelmodel1 )

#####
# GRAVITY MODEL MCO #
#####

DatTFG3 <- read_excel("~/Desktop/Analisis TFG/DatTFG3.xlsx", col_types = c("numeric", "text",
"text", "numeric", "numeric", "numeric", "numeric"))

#View(DatTFG3)

#Eliminar missing data y valores de 0
DatTFG3 <- DatTFG3[complete.cases(DatTFG3[,c("IED_in", "dist", "PIB1", "PIB2" )]),]
DatTFG3 <- DatTFG3[DatTFG3$IED_in != 0, ]
DatTFG3 <- DatTFG3[DatTFG3$dist != 0, ]

#Estadística descriptiva

plot(log(DatTFG3$PIB1 * DatTFG3$PIB2), log(DatTFG3$IED_in), type = "p", xlab = "ln_gdp", ylab
= "ln_IEDin")
abline(lm(log(DatTFG3$IED_in)~log(DatTFG3$PIB2 * DatTFG3$PIB1)))

library(estimatr)
gravity1 <- lm_robust(log(IED_in) ~ log(dist) + log(PIB1) + log(PIB2) , cluster = dist, data =
DatTFG3, se_type = "stata")
summary(gravity1)

```