UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

GRADO EN ECONOMÍA

TRABAJO DE FIN DE GRADO

TÍTULO: *Aproximación Lineal para calcular el índice ERPE*

AUTOR: *Julia López Regino*

TUTOR: David Casado de Lucas

CURSO ACADÉMICO: *2015/2016*

CONVOCATORIA: *Febrero*

Índice

Resumen

En el Consejo Europeo de 2010 se fijaron y aprobaron los objetivos de la Estrategia Europa 2020, para promover el crecimiento de los países miembro. Elaboraron un marco con 5 objetivos claros donde destaca como uno de los principales la lucha contra la pobreza.

Para medir el impacto de esta lucha definieron un nuevo indicador que mide la población en riesgo de pobreza y exclusión social, ERPE (AROPE en inglés).

El objetivo de este trabajo es tratar de encontrar una primera aproximación, utilizando el modelo de regresión lineal múltiple, que nos permita conocer el ERPE de cada país con los datos macroeconómicos a los que tiene acceso el público general, ya que la fórmula utilizada por la Unión Europea contabiliza los usuarios uno a uno.

Para ello vamos a realizar un estudio transversal y longitudinal con los datos disponibles, para encontrar una fórmula que sea válida para todos los países y a lo largo del tiempo.

INTRODUCCIÓN

Estrategia Europa 2020

En el Consejo Europeo del 17 de junio de 2010, la Unión Europea aprobó la estrategia **Europa 2020**, para promover el crecimiento de los países miembro a lo largo de esta década.

Con esto tratan, no sólo de salir de la crisis, sino de paliar las deficiencias del modelo de crecimiento actual, sentando las bases para un crecimiento inteligente, sostenible e integrador.

En esta estrategia se fijaron 5 objetivos principales, que luego cada país puede adaptar a sus necesidades, y que podemos resumir en:

- **Empleo**: que el 75% de la población de 20 a 64 años tenga acceso al empleo.
- **I+D**: inversión del 3% del PIB de la UE.
- **Cambio climático y sostenibilidad energética**:
    1. Reducir las emisiones de gases de efecto invernadero un 20% (llegando a un 30 si se da la posibilidad)
    2. Que el 20% de la energía provenga de fuentes renovables.
    3. Aumento del 20% de la eficiencia energética.
- **Educación**:
    1. Que la tasa de abandono escolar temprano se sitúe por debajo del 10%
    2. Que como mínimo un 40% de la población entre 30 y 34 años completen los estudios terciarios.
- **Lucha contra la pobreza y la exclusión social**: Reducir como mínimo en 20 millones el número de personas en situación o riesgo de pobreza y exclusión social.

En este trabajo vamos a centrarnos en este último objetivo de lucha contra la pobreza y la exclusión social, estudiando el indicador que la Comisión Europea ha elaborado para la consecución de ese quinto objetivo, ERPE (AROPE en Inglés), que determina el porcentaje de la población que está en riesgo de pobreza y exclusión social.

Indicador ERPE (AROPE)

El indicador ERPE (Personas En Riesgo de Pobreza y Exclusión) amplía al tradicional concepto de riesgo de pobreza ya que no considera sólo los términos monetarios (nivel de renta) sino que utiliza un concepto multidimensional de pobreza y exclusión social utilizando los siguientes subindicadores: tasa de riesgo de pobreza después de transferencias sociales, carencia material severa y hogares con baja intensidad de trabajo (indicador del paro). Las personas se cuentan solo una vez en el caso de estar incluidas en más de un subindicador de los mencionados anteriormente y no se aplica a las personas de 60 años o más.

Los datos de estas variables se publican anualmente, según la encuesta de calidad de vida de cada país, y está estandarizado para todos los países de la Unión Europea, lo que nos permite la comparación entre países.

El objetivo de este trabajo es llegar a una primera aproximación lineal, con un modelo bueno y sencillo, que nos permita usar estas variables para estimar el valor del ERPE, ya que la fórmula que hemos encontrado (y que utiliza Eurostat para este indicador), se basa en el uso de datos microeconómicos que no están disponibles para el público general.

$$\mathrm{AROPE} = \frac{\sum_{\forall i\_EQ\_INC20 < AROPT60\_or\_SEV\_DEP\_or\_WI < 0.2} \mathrm{RB050a}_i}{\sum_i \mathrm{RB050a}_i} \cdot 100$$

Esta fórmula nos indica que el ERPE se calcula como una proporción entre el sumatorio de cada individuo que cumple una de las tres condiciones establecidas y el sumatorio total de personas contempladas.

Vamos a realizar varios ejercicios que nos ayuden a determinar esta relación tanto conjunta (para todos los países) como para países individualizados, viendo sus especificaciones.

Para conseguir nuestro objetivo, vamos a utilizar como herramienta estadística el análisis de regresión lineal. Realizaremos el estudio tanto transversal como longitudinalmente, tratando de ajustar el modelo con la mayor fidelidad posible. Realizaremos también, una serie de ejercicios complementarios con el fin de consolidar nuestros resultados.

Partimos de unos datos fiables, ya que los hemos obtenido de una fuente fiable, como es la Unión Europea, por lo que estudiaremos los posibles problemas clásicos de este modelo (colinealidad, heterocedasticidad, distribución no normal, etc.) que aparecen, pero en un grado tan pequeño que no tiene consecuencias determinantes.

Conjunto de datos

Variables utilizadas (subindicadores y ERPE):

Como ya hemos explicado en la introducción, vamos a ajustar varios modelos con el fin de encontrar una primera aproximación a la función del indicador ERPE (AROPE). Para ello, vamos a utilizar las tres variables que lo componen:

1. **Tasa de riesgo de pobreza**: población situada por debajo del 60% de la mediana de los ingresos (después de transferencias sociales).
2. **Carencia material severa**: Se considera carencia material severa cuando no se tiene acceso a 4 o más items de la siguiente lista:
    - Gatos relacionados con la vivienda (hipoteca o alquiler, recibos, comunidad...) o en compras a plazos en los últimos 12 meses.
    - No puede permitirse una comida de carne, pollo o pescado al menos cada dos días.
    - No puede permitirse mantener la vivienda con una temperatura adecuada durante los meses fríos.
    - No puede permitirse ir de vacaciones al menos una semana al año.
    - No tiene capacidad para afrontar gastos imprevistos (de 650 euros).
    - No puede permitirse un automóvil.
    - No puede permitirse un teléfono.
    - No puede permitirse una televisión a color.
    - No puede permitirse una lavadora.
3. **Hogares con muy baja intensidad de trabajo**: esta variable comprende las personas de 0 a 59 años que viven en hogares en los que sus miembros en edad de trabajar lo hicieron menos del 20% de su potencial de trabajo en el año anterior a la entrevista, ya que este es el periodo de referencia de los ingresos.

    Básicamente, se calculan los meses en los que los miembros en edad de trabajar han estado activos y por otro, el total de meses que podían haber trabajado. Se calcula el ratio entre estos dos datos y se determina si es inferior al 20%.

    IT= $\frac{Meses\ trabajando}{Meses\ que\ podían\ haber\ trabajado}$, si el resultado es < 20% se contabiliza como hogar con muy baja intensidad laboral.

Todos los datos de estas series las obtenemos de Eurostat.

Análisis

El primer modelo que vamos a estimar, se trata de un estudio transversal, en el que vamos a utilizar los últimos datos disponibles de las variables (2014) de todos los países de la unión europea, incluyendo también los diferentes agregados de países que tenemos disponibles. En la siguiente tabla se muestran los valores numéricos de estos datos:

Tabla 1: ERPE y componentes para todos los países en 2014

País	Y	X1	X2	X3
Alemania	20,6	16,7	5,0	10,0
Austria	19,2	14,1	4,0	9,1
Bélgica	21,2	15,5	5,9	14,6
Bulgaria	40,1	21,8	33,1	12,1
Chipre	27,4	14,4	15,3	9,7
Croacia	29,3	19,4	13,9	14,7
Dinamarca	17,8	11,9	3,2	12,1
Eslovaquia	18,4	12,6	9,9	7,1
Eslovenia	20,4	14,5	6,6	8,7
España	29,2	22,2	7,1	17,1
Estonia	26,0	21,8	6,2	7,6
Finlandia	17,3	12,8	2,8	10,0
Francia	18,5	13,3	4,8	9,6
Grecia	36,0	22,1	21,5	17,2
Holanda	16,5	11,6	3,2	10,2
Hungria	31,1	14,6	23,9	12,2
Irlanda	27,4	15,3	8,4	21,0
Islandia	11,2	7,9	1,4	4,9
Italia	28,3	19,4	11,6	12,1
Letonia	32,7	21,2	19,2	9,6
Lituania	27,3	19,1	13,6	8,8
Luxemburgo	19,0	16,4	1,4	6,1
Malta	23,8	15,9	10,2	9,8
Noruega	13,5	10,9	1,2	5,9
Polonia	24,7	17,0	10,4	7,3
Portugal	27,5	19,5	10,6	12,2
Reino Unido	24,1	16,8	7,3	12,2
República Checa	14,8	9,7	6,7	7,6
Rumania	40,2	25,4	26,3	6,4
Suecia	16,9	15,1	0,7	6,4

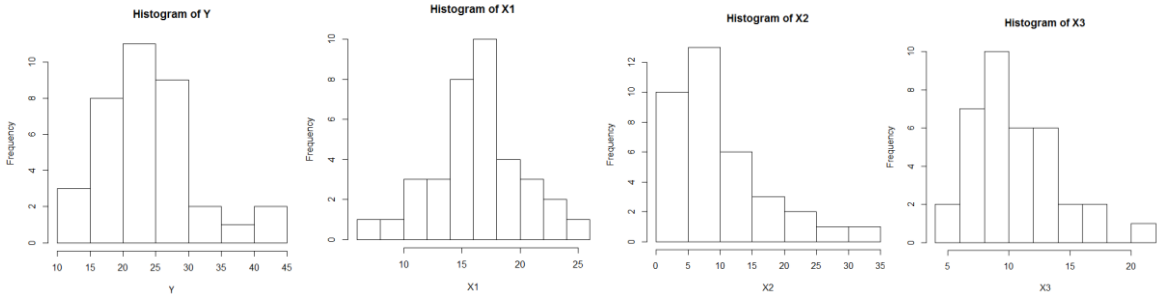
Fuente: elaboración propia a partir de datos de Eurostat

Podíamos utilizar también los datos de ciertos agregados de países (UE(28), zona euro, etc.) Pero hemos decidido excluirlos puesto que son agregados de países ya contemplados como país individual.

Si estudiamos independientemente estas variables, obtenemos unas medias de:

Y: 24.01333	**X1**: 16.29667	**X2**: 9.846667	**X3**: 10.41

Mirando el histograma detectamos algunos posibles problemas para garantizar la normalidad de la variable Y (las X_j se consideran no aleatorias).



Fuente: elaboración propia

En un segundo momento, vamos a ampliar el análisis de este modelo realizando un estudio longitudinal en el que vamos a introducir los datos de España desde 2005 hasta 2014, y repitiendo este caso con algunos países más con el fin de encontrar unos coeficientes que se mantengan constantes a lo largo del tiempo, es decir, para poder asegurar que la función utilizada no ha variado. La presentación de los datos es la siguiente:

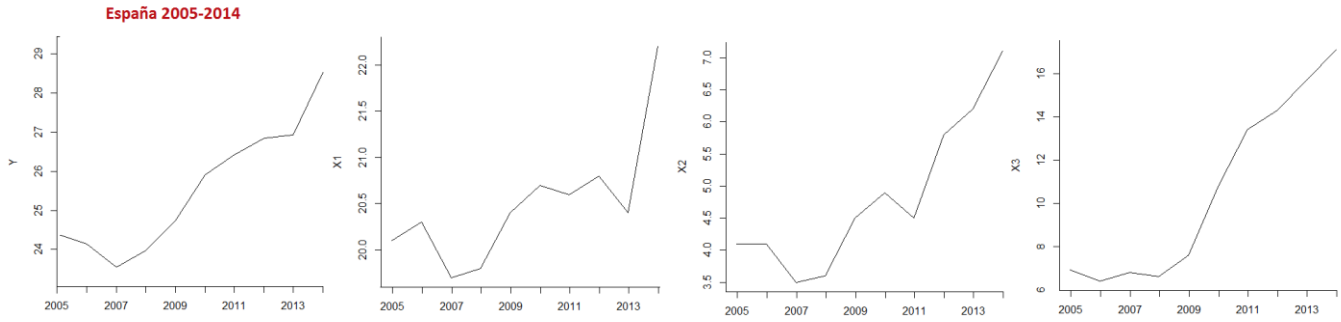
Tabla 2. Datos de España de 2005 a 2014

España	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	24,3	24	23.3	23,8	24,7	26,1	26,7	27,2	27,3	29,2
X1	20,1	20,3	19,7	19,8	20,4	20,7	20,6	20,8	20,4	22,2
X2	4,1	4,1	3,5	3,6	4,5	4,9	4,5	5,8	6,2	7,1
X3	6,9	6,4	6,8	6,6	7,6	10,8	13,4	14,3	15,7	17,1

Fuente: Elaboración propia a partir de los datos de Eurostat

El resto de tablas estudiadas para el análisis se pueden consultar en el apéndice 1.

Como son series temporales, es conveniente dibujar su gráfico para observar la tendencia de cada una de las variables:



Mirando estos gráficos observamos que las variables han evolucionado de forma similar, sobre todo X1 y X2, pero al tratarse de un solo país, es conveniente pensar que esto puede ser una coincidencia, debida a las características propias.

Una vez desarrollados y ajustados, vamos a realizar una serie de ejercicios que nos ayuden a cerciorarnos de estos ajustes, teniendo en cuenta algunas relaciones "curiosas" entre los países.

MODELO DE REGRESIÓN LINEAL

Caracterización del modelo (¿Qué es?)

Para la realización del análisis estadístico vamos a utilizar el modelo de regresión lineal múltiple.

La idea de la que partimos es analizar una función que en términos matemáticos podemos expresar como Y= f (X) o Y= f (X1, X2,…Xn)

El modelo de regresión lineal simple se caracteriza por la expresión:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad [1]$$

Este modelo estima la relación lineal de la variable exógena, X, de la cual tenemos unos valores predeterminados, con la variable endógena, Y, añadiendo un término de error aleatorio, ϵ, que recoge el efecto de otros factores que no conseguimos identificar o son desconocidos, pero que influyen en la Y.

El modelo de regresión múltiple se desarrolla de la misma manera, pero tiene en cuenta más de una variable, que determina a Y. Se caracteriza como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \quad [2]$$

Resumiendo, la variabilidad total de la variable endógena, Y, podemos descomponerla en dos partes:

- La parte que tratamos de explicar con las variables exógenas utilizadas en el modelo, $X_1, X_2, \ldots, X_n$
- La parte que no podemos explicar, ϵ.

Hipótesis

[H1] Sobre Y

[H1-1] Causalidad: La variable dependiente no influye en las variables independientes. Si sucede esto, tendríamos que considerar otros modelos. El significado de las variables es el principal indicador de las relaciones de causalidad.

En nuestro caso esto queda garantizado por la propia definición de las variables.

[H1-2] Aleatoriedad: Dado que al menos se utiliza un término de error aleatorio en el modelo, la variable explicada es también aleatoria independientemente del carácter de las variables explicativas.

[H1-3] Normalidad: Cuando el error es normalmente distribuido y las variables explicativas son deterministas, es fácil deducir (utilizando los resultados para suma de

variables normales) cuál debería ser la distribución teórica de la variable explicada. Cuando las variables explicativas son estocásticas, puede deducirse fácilmente si son normalmente distribuidas, pero no si siguen otras distribuciones.

[H2] Sobre X_j

[H2-1] Determinismo (no aleatoriedad): Suponemos que las variables explicativas no son aleatorias o que son la realización de variables aleatorias (como consideran Newbold y otros), lo que facilita la demostración de los resultados.

[H1-2] Unicidad (no colinealidad): Un grave problema que puede surgir al ajustar un modelo de este tipo es la colinealidad, es decir, una alta correlación entre las variables explicativas. Este problema es fácil de identificar si estudiamos la relación entre las variables explicativas. En nuestro caso, suponemos que no va a suponer un problema por la definición de las variables, pero puede aparecer en los datos.

[H3] Sobre la relación funcional (coeficientes β_j)

[H3-1] Linealidad en los parámetros: Suponemos que los parámetros son lineales. Este es el motivo por el que se habla de regresión lineal, ya que son los parámetros las incógnitas al ajustar el modelo.
Una regresión no lineal sería, por ejemplo, $Y = \beta_0 + \beta_1^2 X_1 + \beta_2^3 X_2 + \epsilon$

[H3-2] Homogeneidad estructural: Suponemos también que los coeficientes son parámetros, es decir, que son independientes de los valores de las variables explicativas. Esto implica que la fórmula que estimemos será válida para todo el rango de estudio ya que sólo existe una fórmula y no varias. Esta hipótesis debemos evaluarla en la fase descriptiva de los datos o comparando varios modelos lineales.

[H3-3] Linealidad en las variables: este modelo supone dependencias lineales entre las variables. Este modelo tiene una gran utilidad porque con frecuencia podemos transformar las variables para seguir aplicando el modelo lineal. Eso sí, debemos tener en cuenta las transformaciones aplicadas a la hora de interpretar los modelos, por ejemplo en $Y = \beta_0 + \beta_1 log(\mathrm{X}) + \epsilon$ no debe olvidarse que β_1 es el coeficiente de $log(\mathrm{X})$, no de X.

[H4] Sobre el término de error (ϵ)

[H4-1] Aleatoriedad: Sólo debemos tener en cuenta los errores aleatorios, ya que suponemos que los errores no aleatorios que puedan presentarse se corrigen antes de aplicar el modelo.

[H4-2] E[ϵ_i]=0: También por lo dicho en el apartado anterior, cualquier error de media no nula se supone ya corregido antes de aplicar el modelo.

[H4-3] Homocedasticidad o varianza uniforme, E[ϵ_i^2] $= \sigma^2$ (esto también puede ocurrir con las variables independientes)**:** La varianza, σ^2, debe ser constante, si esto se viola, hablamos de heterocedasticidad, y suele ser fácil de identificar al representar los datos mediante contrastes de hipótesis más formales. También puede estar causada por la presencia de unos pocos datos atípicos entre otros motivos. Una opción es intentar transformar la variable dependiente. En nuestro caso, observamos que aumenta la dispersión conforme aumenta X3, por lo que vamos a probar a transformar esta variable.

[H4-4] Normalidad: Los términos de error siguen una distribución normal, sea directamente o gracias al teorema central del límite. Es fácil evaluar el incumplimiento de esta hipótesis, por ejemplo a partir de la asimetría, el exceso de curtosis o contrastes de bondad de ajuste específicos para el caso de la normalidad.

[H4-5] Autoincorrelación: los términos de error aleatorios $\boldsymbol{\epsilon_i}$ no están correlacionados entre sí, es decir, no se ven afectados por los distintos valores que puede tomar si las variables toman otros valores. Esto se puede comprobar gráficamente o aplicando contrastes de hipótesis. En nuestro caso, no podría aparecer en los datos por países (no han sido ordenados por ninguna de las variables), pero podría aparecer al considerar los datos temporales.

[H5] Sobre la muestra de datos $(\boldsymbol{Y^{(i)}}, \boldsymbol{X_1^{(i)}}, \boldsymbol{X_2^{(i)}}, \dots, \boldsymbol{X_k^{(i)}})$

[H5-1] Tamaño: En la muestra debe haber al menos $(k+2)$ datos, siendo n el número de variables incluidas, ya que en otro caso no se pueden estimar todos los parámetros.
Además, la lógica nos dice que es más fácil ajustar un modelo (o que el modelo ajustado es mejor) cuantos más datos útiles tengamos. En nuestro caso, podemos considerar sin problema modelos con tres o cuatro parámetros incluso para el caso en el que menos datos tenemos.

Ajuste

Para ajustar este modelo se utiliza, en condiciones normales, el método de mínimos cuadrados ordinarios.

El ajuste del modelo por el método de los mínimos cuadrados ordinarios la realizamos con la idea de que los errores o residuos sean lo más pequeños posibles, ya que como su propio nombre indica, se basa en minimizar la suma de los cuadrados de los residuos o suma residual. Estos residuos son la distancia que existe entre cada valor de la variable que explica el modelo con el valor que proporciona la parte funcional del modelo (la expresión sin incluir el término de error).

Es decir, tratamos de seleccionar los valores de los coeficientes β_j que resuelvan el problema:

$$\min \psi\,(\beta_0\,, \beta_1, \dots, \beta_k) = min \sum_{i=1}^{n} [Y^{(i)} - \hat{Y}^{(i)}]^2 \quad [3]$$

Si nos lo imaginamos representado gráficamente, es más intuitivo. Se trata de seleccionar de entre todas las posibles rectas que cruzan la nube de puntos (en el modelo de regresión lineal múltiple pensamos en planos e hiperplanos), cuál es aquella que minimiza la suma de las distancias de cada punto a su proyección vertical sobre la recta.

Diagnosis (Medida de bondad del ajuste)

Una vez creado y ajustado nuestro modelo, tenemos que validar este ajuste. Para esto, se suele considerar la información de varios resultados, que vamos a resumir en lo siguiente:

- **Ajuste global**: debemos tener en cuenta varios datos que arroja nuestro modelo:
    - **ERR**: error estándar de regresión. Este error tiene las mismas unidades que la variable dependiente Y(i), por lo que se utiliza para comparar dos regresiones sobre la misma Y. El modelo es mejor cuanto menor es esta medida.
    - Coeficiente de determinación, $\boldsymbol{R^2}$. Nos fijamos en este dato si queremos comparar modelos donde las variables y datos son distintos. Si hay varias variables explicativas utilizamos el R^2 ajustado. El modelo es mejor cuanto más cercano a 1 es esta medida, ya que nos indica cual es el porcentaje de la variable endógena explicado por las variables exógenas.

- **p-valor** (significatividad global): queremos que esté dato sea lo más próximo a 0 posible, ayudándonos a rechazar la hipótesis nula de que todos los coeficientes de las variables son nulos a la vez, es decir, rechazamos que el modelo es inútil para explicar la variable dependiente.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$$

$$H_1: \exists \beta_{j0} \neq 0$$

- **Coeficientes**: respecto a la información de las estimaciones de los coeficientes, es importante observar para cada uno la precisión y, después, la significatividad.

  - Coeficiente de correlación parcial: la idea parte de analizar una posible relación entre dos variables que pueda generar colinealidad.
    Para estudiar las relaciones entre dos variables, $X_i \; y \; X_j$, podemos hallar la correlación entre ellas, pero también podemos utilizar las regresiones cruzadas, $X_i = \beta_0 + \beta_1 X_j + \varepsilon$ y $X_j = \beta_0 + \beta_1 X_i + \varepsilon$ o las regresiones parciales cruzadas $\tilde{X}_i = \beta_0 + \beta_1 \tilde{X}_j + \varepsilon$ y $\tilde{X}_j = \beta_0 + \beta_1 \tilde{X}_i + \varepsilon$ que estudian la relación entre estas variables sin el efecto de las demás. También podemos utilizar $Y = \beta_0 + \beta_1 X_i + \varepsilon$ si queremos estudiar el efecto total de la variable X_i sobre Y, o $\tilde{Y} = \beta_0 + \beta_1 \tilde{X}_i + \varepsilon$ para estudiar el efecto parcial.

  - También es importante el **p-valor** de cada parámetro individual, que nos muestra con que probabilidad podemos rechazar la hipótesis $H_0: \beta_j = 0$. Queremos en este caso, al igual que en la medida global, que este valor sea lo más cercano a 0 posible, ya que si esto se cumple, nos indica que el coeficiente es significativamente distinto de 0, y por lo tanto, es útil para explicar la variable dependiente. Si no podemos rechazar esta hipótesis, las conclusiones sobre los coeficientes no son fiables, ya que no podemos fiarnos de las estimaciones ni los contrastes.

- **Residuos:** los residuos son el elemento más informativo sobre la bondad del ajuste de un modelo. Aportan información importante tanto sobre la información de Y que no ha sido explicada por la parte funcional del modelo (expresión menos el

término de error) como del posible incumplimiento de alguna de las hipótesis. El gráfico de residuos frente a valores estimados permite detectar:

- Muestra si hay relación lineal entre los residuos y los valores estimados, lo que sucede cuando no se ha eliminado el efecto de X sobre Y.
- No linealidad
- Heterocedasticidad
- Datos atípicos, autocorrelación, estacionalidad, etc.

Entre las características que pueden tenerse en cuenta están: la media, la mediana, el coeficiente de asimetría, el exceso de curtosis, el histograma, distintos gráficos de residuos, contrastes de aleatoriedad, contrastes de normalidad y contrastes de autocorrelación.

Aplicaciones del modelo

Este modelo permite estimar y sustraer la posible tendencia lineal de unos datos. Es muy útil ya que en la realidad se dan este tipo de relaciones con frecuencia, y si no, es posible aproximar esta relación ya que permite transformar las variables para encontrar una relación lineal entre estos nuevos valores.

Un primer uso que podemos dar a este modelo se basa en sustituir los valores de las variables independientes y considerar el valor ajustado una corrección del error del valor anterior de la variable dependiente, ya que los coeficientes que proporciona nos indican cuánto varía la variable dependiente, cuando una única variable independiente varía en una unidad mientras las otras permanecen constantes. A esta situación se le denomina *ceteris paribus*, y aunque tiene sentido matemático, no lo suele tener en la práctica ya que es muy difícil que las variables explicativas sean 100% independientes, y por lo tanto pueda variar una sin que el resto se vean afectadas.

También nos permite identificar aquellas variables independientes que debemos incluir en el modelo por resultar más explicativas, descartando aquellas variables que no aporten información nueva al modelo, y detectar aquellas interacciones entre variables dependientes que pueden afectar al resultado del modelo.

En algunas ocasiones, cuando obtenemos un modelo bien ajustado, lo podemos utilizar para predecir un resultado de la variable dependiente a partir de las variables independientes seleccionadas.

También existen una serie de aplicaciones más avanzadas, como los intervalos de confianza o los contrastes de hipótesis, que no vamos a entrar a detallar ya que se pueden consultar en libros avanzados sobre este modelo.

Análisis de datos (Modelización)

Una vez situados en el tema en cuestión, procedemos al análisis de los datos, es decir, ajustamos nuestros modelos.

Vamos a partir en los dos análisis de las 3 variables explicadas a lo largo del trabajo. Con lo que nuestro modelo inicial es:

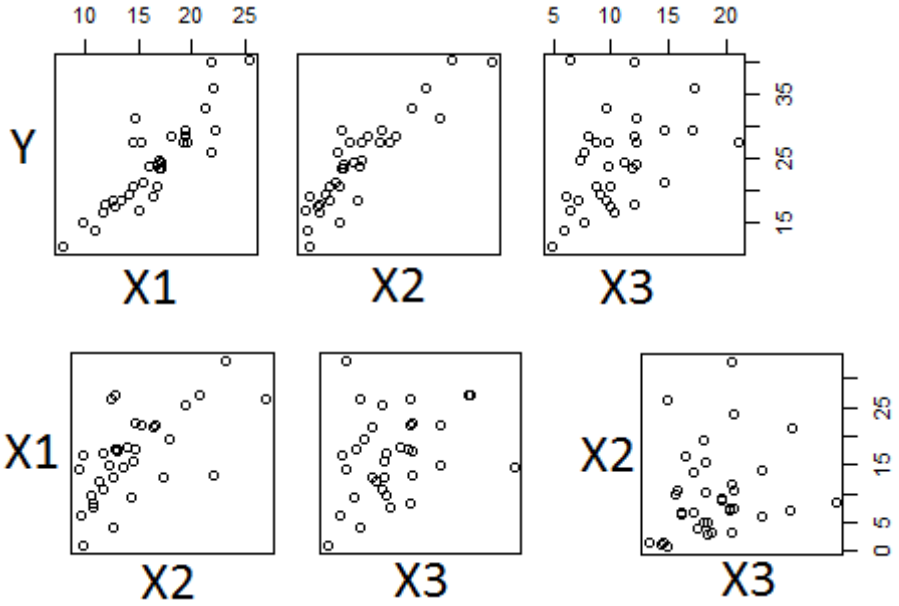
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad [4]$$

Dónde:

- Y= ERPE
- X1= riesgo de pobreza
- X2= carencia material severa
- X3= hogares con baja intensidad laboral

Comenzamos el análisis con un **modelo transversal**, cuyo objetivo es encontrar una aproximación lineal a la formula (función) que podríamos utilizar para calcular el ERPE, partiendo de los datos de las 3 variables utilizadas para todos los países de la Unión Europea (se contemplan 30 países) en el año 2014, último dato disponible.

Lo primero que hacemos al iniciar el análisis es representar gráficamente la relación de todas con todas, y obtenemos los siguientes resultados:



Mirando estos gráficos, vemos como positivo que parece que se da una relación lineal en el comportamiento de la Y respecto a las 3 variables (especialmente con X1 y X2), pero también observamos que entre la X1 Y X2 puede haber cierta información redundante (calculando su correlación vemos que es 0,62), lo que puede generar un problema de colinealidad. No creemos que este problema venga generado por la definición de las variables, ya que han sido propuestas por la Unión Europa, pero si pueden generarlo los datos.

La variable X3, sin embargo, es la que en principio menos relación tiene con el resto, puede indicar que es la "menos importante" para este modelo, además observando el gráfico vemos que puede presentar heterocedasticidad ya que aumenta la dispersión conforme nos alejamos del origen

Realizando un primer ajuste del modelo obtenemos los siguientes resultados:

- Según el ajuste global:

```
Residual standard error: 0.9284 on 26 degrees of freedom
Multiple R-squared:  0.9861,    Adjusted R-squared:  0.9844
F-statistic: 612.8 on 3 and 26 DF,  p-value: < 2.2e-16
```

  Como era previsible, el R^2es muy elevado, el ajustado supera el 0,98, lo que indica que estas variables explican aproximadamente un 98% del ERPE.

  Además, el p-valor es prácticamente 0, lo que nos ayuda a rechazar la hipótesis de que las tres variables conjuntamente tienen un efecto nulo sobre Y.

  Esto nos ayuda a concluir que las tres variables conjuntamente son un buen grupo para ajustar el valor del ERPE.

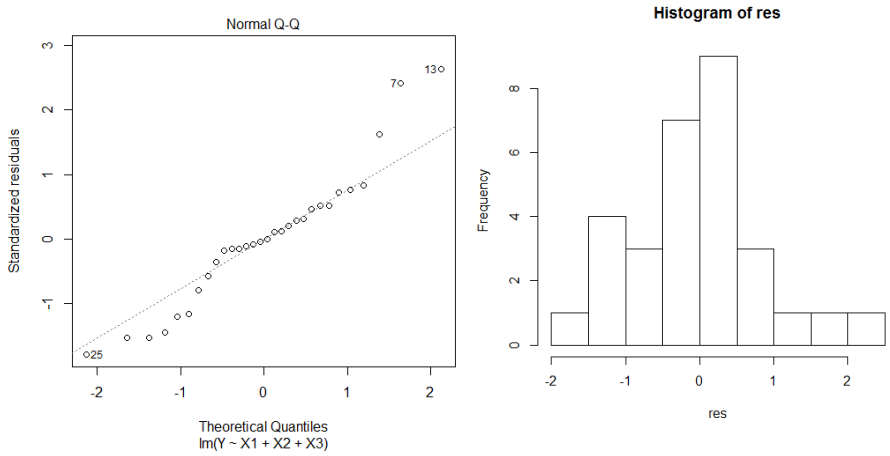
- Ajuste para cada coeficiente:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.20017    0.80091   2.747   0.0108 *
X1           0.83043    0.05424  15.310 1.60e-14 ***
X2           0.52347    0.02765  18.931  < 2e-16 ***
X3           0.30024    0.04875   6.159 1.64e-06 ***
```

  Si observamos estos coeficientes, vemos que todos son significativamente distintos de cero, lo que corrobora nuestra idea de que las tres variables son necesarias, y además consideradamente explicativas.

  Pensamos también que el término independiente β_0 no es determinante ya que sería el valor asignado a un país cuyos tres componentes son 0, lo que no parece posible (a la vista de los datos).

- Análisis de los residuos:



Realizamos la correlación entre los residuos y la variable explicada para ver cuánta información contienen y obtenemos que esta es aproximadamente un 10%

En este momento queremos ver la normalidad de los residuos. Observando el histograma, vemos que los residuos se distribuyen de forma similar a la normal, pero en el gráfico Q-Q, vemos como hay varios puntos que se alejan de este ajuste.

Si nos fijamos en los datos:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.5776  -0.4636  -0.0180   0.4507   2.3202
```

Podemos concluir que se aleja un poco de la distribución de la normal (0,1) ya que la media es aprox. 0,37 y la desviación típica está por encima de 2,5. Esto puede deberse también a la gran disparidad entre los países contemplados en el estudio.

Realizamos un contraste para comprobar la normalidad, shapiro.test.

Este test establece como hipótesis nula que una población está distribuida normalmente, por lo tanto, nos interesa que el p-valor que arroje sea lo más alto posible, puesto que nos ayudara a rechazar la hipótesis de que este modelo no está distribuido normalmente.

```
        Shapiro-Wilk normality test

data:  res
W = 0.95648, p-value = 0.2512
```

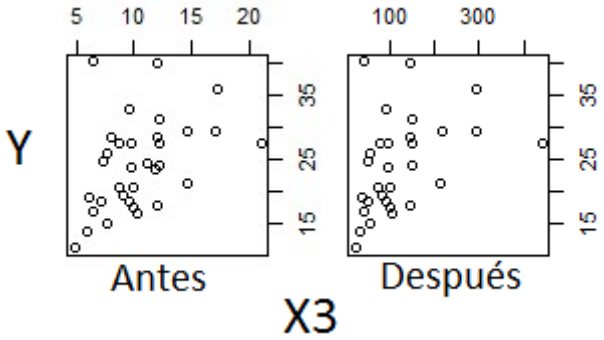
Cómo se puede observar, según este contraste nuestra población está distribuida normalmente, pero no nos quedamos completamente seguros, por lo que decidimos realizar una transformación de variables con el fin de ajustar lo más verazmente posible el modelo.

Cómo la variable que al parecer nos crea problemas es la X3, vamos a probar con su transformación.
La primera idea que nos viene a la mente es transformarla en logaritmo (X3=>log(X3))

El lector que quiera contrastarlo puede realizar los cálculos utilizando el código que puede encontrar en el Apéndice 2.

Los resultados al ajustar el modelo con esta transformación nos indican que no es necesaria, es decir, no mejoran en nada nuestro modelo.

La segunda transformación que probamos es el cuadrado. Esto tiene sentido puesto que el problema localizado es la heterocedasticidad, y quizás, esta transformación puede mejorar el modelo.



Mirando la representación gráfica de la relación, observamos que los puntos se alejan menos de una recta de regresión imaginaria.

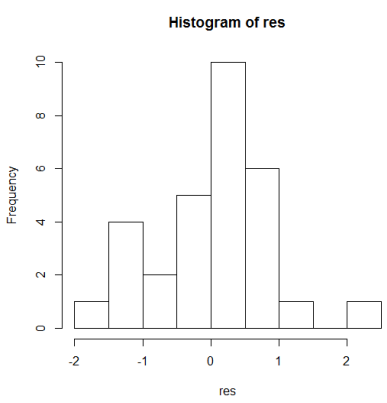
Volvemos a realizar el análisis para este modelo, siguiendo los mismos pasos que en el primero.
El ajuste global del modelo es bueno, ya que tanto el elevado R^2 como el p-valor corroboran esta idea.
En este caso, todos los coeficientes son significativamente distintos de 0, incluido el termino de corrección (β_0) que se hace más "significativamente distinto de cero" (pero esto se debe a la condensación producida con la transformación en X3).

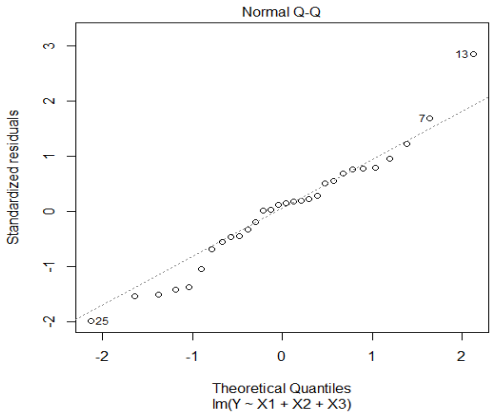
```
Residuals:
    Min      1Q  Median      3Q     Max
-1.6932 -0.4588  0.1113  0.5603  2.4152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.763844   0.716014   5.257 1.71e-05 ***
X1          0.830731   0.052145  15.931 6.24e-15 ***
X2          0.529609   0.026574  19.930  < 2e-16 ***
X3          0.012280   0.001873   6.557 5.94e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8937 on 26 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9856
F-statistic:   662 on 3 and 26 DF,  p-value: < 2.2e-16
```

También realizamos el estudio de los residuos, que es en lo que nos basamos finalmente para seleccionar el mejor modelo.





Vemos como los gráficos se ajustan en mayor medida a la distribución normal, y la recta ajusta mejor. Por otro lado, obtenemos que es un mejor ajuste ya que disminuye la asimetría (en el primer modelo era cercana a 0.45 frente al 0.27 en este caso) y aumenta el exceso de curtosis (de 0.26 a 0.51) aunque sigue estando cerca del valor de una normal.

Realizando el contraste de Shapiro, obtenemos que en este caso el p-valor es 0.3417, lo que nos ayuda a decidir que este modelo es algo mejor ajuste que el anterior.

Una vez determinado esto, probamos si incluir todas las variables que conformarían los términos de segundo orden de una serie de Taylor sería conveniente o no.
Ahora contemplamos las siguientes variables: Y = TodosERPE, X1 = TodosRiesgo, X2 = TodosCarencia, X3 = TodosParo, X4 = X1^2, X5 = X2^2, X6 = X3^2, X7 = X1*X2, X8 = X1*X3, X9 = X2*X3

Llegamos a la conclusión de que nuestro modelo no mejora al tener en cuenta estas variables. De hecho, si vamos eliminando las variables, nos muestra que el mejor modelo es el que hemos elegido con anterioridad.

El lector interesado puede realizar los cálculos utilizados en el Ejercicio 1 del apéndice 2.

Concluimos en este apartado que el modelo que mejor ajusta nuestro caso es:

$$\hat{Y} = 3.76 + 0.83X_1 + 0.52X_2 + 0.01X_3^2 \text{ [5]}$$

Aunque también es aceptable el modelo sin la variable X3 transformada:

$$\hat{Y} = 2.20 + 0.83X_1 + 0.52X_2 + 0.30X_3 \text{ [6]}$$

Llegamos a esta conclusión puesto que con la variable X3 transformada, nuestro modelo cumple las hipótesis teóricas con más certeza, pero tenemos que tener en cuenta que sólo hemos considerado los datos de 2014. En un estudio más amplio sería interesante corroborarlo con datos de otros años.

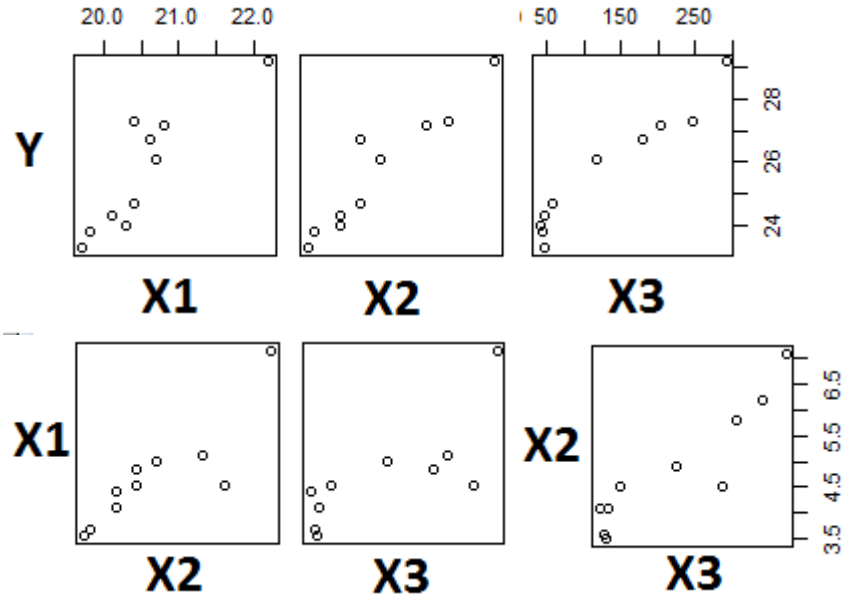
Modelo longitudinal

Ahora vamos a elaborar una tabla con los resultados de una serie de modelos ajustados para diferentes países en un periodo temporal de 10 años, con el fin de establecer una función que sea válida a lo largo del tiempo, es decir, queremos probar si la función que utilizan ha sido la misma a lo largo del tiempo o ha sufrido variaciones. Para esto vamos a probar con los datos de España y algún país más para intentar buscar unos coeficientes que se mantengan constantes a lo largo del tiempo.

España 2005-2014

Vamos a utilizar el modelo con la variable X3 transformada que hemos seleccionado anteriormente.

Comenzamos igualmente elaborando la matriz de correlaciones de todas las variables con todas, en el primer caso, para los datos de España.



Mirando estos gráficos, debemos destacar la relación que se observa entre la Y y la X3 (0,974), vemos que para el caso concreto de España, esta es la variable más influyente. Esto no es del todo sorprendente puesto que sabemos cómo indican muchos expertos que en España el principal problema es el mercado laboral.
Atendiendo a las correlaciones, debemos destacar la fuerte influencia que tienen estas variables en el Y y también entre ellas. No se observan datos atípicos, si acaso el último año difiere un poco, pero sigue estando alineado.

Hay que advertir que en los análisis de regresión múltiple los gráficos por pares de variables no son totalmente fiables (Montgomery y otros).

Comenzamos el análisis del modelo tras el ajuste, fijándonos en los coeficientes globales.

```
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9694
F-statistic: 96.17 on 3 and 6 DF,  p-value: 1.836e-05
```

El ajuste global nos indica que el modelo está bien ajustado y nos pide seguir analizando los resultados.

Si miramos los coeficientes de cada variable, nos indica que hay dos términos que no son significativamente distintos de 0.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.929539   5.758347   1.203  0.27415
X1          0.813501   0.318949   2.551  0.04346 *
X2          0.046046   0.323632   0.142  0.89152
X3          0.014355   0.003294   4.358  0.00478 **
```

Probamos a quitar esas variables, con lo que el modelo que ajustamos queda:

$$Y = \beta_1 X_1 + \beta_3 X_3 + \epsilon \text{ [7]}$$

```
Call:
lm(formula = Y ~ 0 + X1 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56999 -0.22028  0.09934  0.22259  0.34915

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X1 1.171491   0.009067  129.20 1.44e-14 ***
X3 0.012869   0.001188   10.84 4.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3289 on 8 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9998
F-statistic: 3.058e+04 on 2 and 8 DF,  p-value: 2.926e-16
```

Para el caso concreto de España, el efecto de la carencia material a la hora de determinar el riesgo de pobreza y la exclusión social, no es significativamente distinto de 0.

Si analizamos los residuos, el histograma nos indica que debemos rechazar la normalidad de los residuos, sin embargo, nuestro ajuste supera el contraste de normalidad de Shapiro, por lo que decidimos quedarnos con este modelo.

Tras hacer este estudio longitudinal, puede ser interesante tener en cuenta la propia tendencia temporal, debida a la evolución de la economía, que puede alterar la relación entre las variables. Para ello probamos a introducir el efecto del tiempo, t, como variable explicativa, para ver si ha tomado papel en la evolución de nuestras variables. Al realizar el ajuste teniendo en cuenta esta variable, obtenemos que el efecto de esta variable no es significativamente distinto de 0.

El lector puede comprobarlo con el código del Ejercicio 2 del apéndice 2.

También se puede considerar un modelo entre variables desprovistas de tendencia (Novales).

Realizamos el mismo análisis para el resto de países que vamos a considerar.
En el apéndice 2 se puede encontrar el código utilizado para los ajustes de cada país, ya que por motivos de espacio aquí mostramos sólo la tabla que hemos construido a partir de los resultados.

Para cada país hay unos coeficientes significativos y otros que no podemos asegurar que sean distintos de 0, por lo tanto, vamos a suponer que su valor es 0 y tener solo en cuenta los valores significativos.

Hemos realizado el ajuste de los modelos para cada país y para todos sus residuos han superado el contraste de normalidad de Shapiro.

El fin de este ejercicio es corroborar que si utilizamos una media de los coeficientes obtenidos de cada modelo ajustado por país individual obtenemos unos coeficientes similares a los del modelo general ajustado.

Estos son los datos obtenidos:

Tabla 3. Coeficientes obtenidos al ajustar los modelos con los datos de 2005 a 2014 de cada país incluido individualmente

	β_0	β_1	β_2	β_3
España	0	1,17	0	0,01
Portugal	0	0,97	0,83	0
Italia	0	1,16	0,50	0
Grecia	19,16	0	0,79	0
Alemania	0	0,95	0	0,04
Bélgica	14,70	0	1,12	0
MEDIA	5,64	0,71	0,54	0,01

Fuente: Elaboración propia

Recordamos que nuestro modelo quedo ajustado de la siguiente manera:

$$\hat{Y} = 3.76 + 0.83X_1 + 0.52X_2 + 0.01X_3^2$$

Mirando los coeficientes que hemos obtenido en este segundo estudio, observamos que el término independiente difiere un poco (muy poco ya que la escala considerada es de 0 a 100), pero esto no es preocupante ya que el vector (X_1, X_2, X_3) no suele tomar valores cercanos a (0,0,0)

El resto de coeficientes nos indican que este modelo es una buena aproximación (si acaso, el efecto de X1 es el que se ve más distorsionado) para estimar el ERPE de cada país si conocemos los datos de las variables.

Pensamos que esto se debe a que hemos considerado 6 países, no a que haya habido cambios en la definición o determinación del índice ERPE.

Debemos indicar que hemos construido una tabla similar con el primer modelo ajustado (sin la transformación de la variable X3) pero los resultados difieren en mayor medida, y alguno de los ajustes de los países no superan el contraste de normalidad de sus residuos.

Una vez conseguido el objetivo principal del trabajo hemos realizado una serie de ejercicios que pueden ayudarnos a concluir el estudio asegurándonos de los resultados.

El primero de estos ejercicios es un estudio de panel, para **evaluar la evolución proporcional del índice ERPE durante esos años.**

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 Y_{t-1} + \epsilon \text{ [8]}$$

Primero consideramos los datos de todos los países, y obtenemos unos valores que se encuentran en el límite para determinar si es influyente o no, por lo que no está claro que pueda aportar información nueva.
Realizamos un ajuste incluyendo sólo esta nueva variable: $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon$
Con este ejercicio determinamos que sí es influyente en este modelo, ya que rechazamos la hipótesis de que el efecto sea 0, y los residuos superan los contrastes de normalidad. Determinamos que en general el valor del año anterior es un buen predictor, lo que muestra que el índice ERPE evoluciona de forma aproximadamente lineal.
Probamos ahora para el caso concreto de España, y en un modelo completo, en el que incluimos las 4 variables, determinamos que este valor anterior no es influyente, ya que la nueva variable incluida no es significativamente distinta de 0.
Sin embargo, si probamos a ajustar un modelo sólo con el año anterior, sí es influyente, lo que nos indica que Y_{t-1}es un buen predictor del ERPE, pero no aporta información nueva si se incluyen (X_{1t}, X_{2t}, X_{3t})

El lector que quiera comprobarlo puede realizar los cálculos con el código del Modelo 3 del apéndice 2.

La conclusión que sacamos de todo esto es que si tenemos disponibles los datos de las tres variables (pobreza, carencia material y paro) podemos predecir mejor los resultados, pero si no disponemos de ellas, y sólo conocemos los valores del indicador ERPE a lo largo del tiempo, podemos ajustar también un buen modelo para predecir el ERPE del año en cuestión.

La siguiente pregunta que nos hacemos es si existen **diferencias según el país**.

Para contestar a esto realizamos un estudio de panel con 3 países, España, Italia y Francia. En un primer momento vamos a crear una variable cualitativa, llamada país, que asigne automáticamente 1 a España, 2 a Italia y 3 a Francia:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_4 + \epsilon \text{ [9]}$$

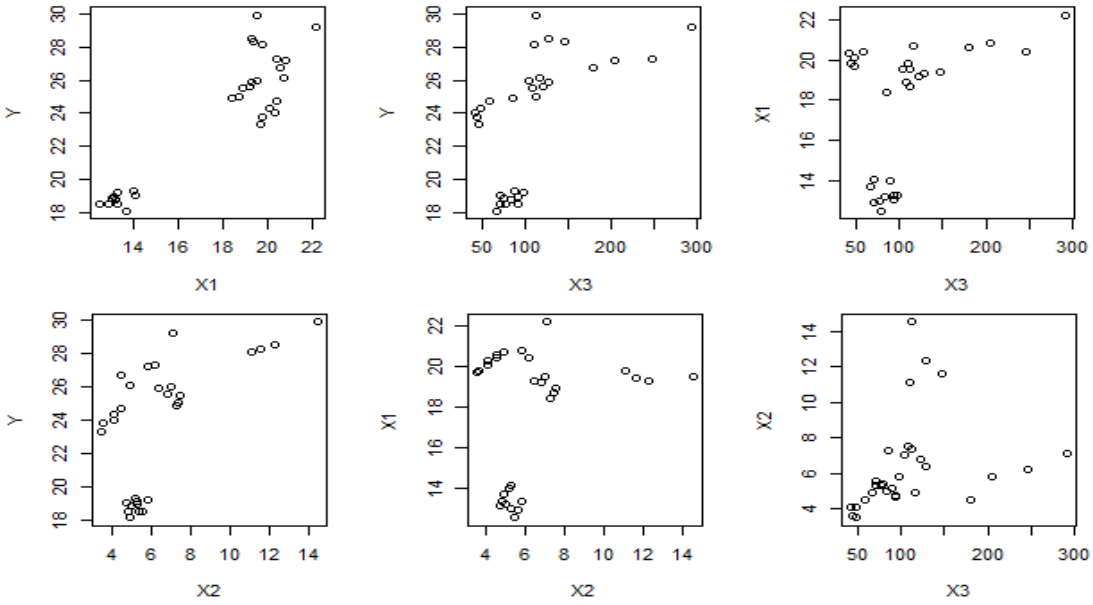
Los resultados nos indican que esta variable no es influyente, con lo que no podemos probar diferencias significativas entre los tres países.

Probamos ahora a crear dos variables indicadoras, X4 y X5, que nos ayuden a responder a nuestra pregunta. Estas variables creadas de forma binaria nos van a permitir observar de otra forma si existe alguna relación entre alguno de los países.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_4 + \beta_5 X_5 + \epsilon \text{ [10]}$$

Tras el primer ajuste, nos sale que la variable X5 no es significativa para nuestro modelo, por lo que probamos el modelo sin incluirla (pero sí X4). En este caso, obtenemos que esta última variable creada es significativa al 1%, lo que nos hace pensar que no hay diferencias grandes entre España e Italia, pero sí con Francia. Esto no es muy sorprendente, ya que partimos de la base de que son dos países mediterráneos, con características más similares entre ellos que con los países continentales.

Por último realizamos un estudio de panel con el fin de obtener **patrones de evolución similares** entre el indicador ERPE en varios países (España, Alemania, Francia e Italia)



Observando los gráficos y la matriz de correlaciones no podemos determinar ningún tipo de relación entre estas variables, por lo tanto, no se demuestra que la evolución del ERPE haya sido paralela en estos paises

Probamos a relacionar los países entre sí, para llegar a una conclusión más profunda. Empezamos relacionando España e Italia ya que en modelos anteriores hemos visto que existe cierta relación.

$$Y_t^{(ES)} = \beta_0 + \beta_1 Y_t^{(IT)} + \varepsilon \text{ [11]}$$

Con este ajuste, corroboramos la idea, ya que el modelo nos arroja unos resultados favorables.

El término independiente no es determinante en esta relación, pero conociendo el ERPE de Italia (en caso de no tener acceso a los datos de España) podríamos aproximar el ERPE de España y este se vería explicado en más de un 99% (si nos fijamos en el R2 ajustado) y rechazando la hipótesis de que no es explicativo con total seguridad (p-valor prácticamente 0). Además este ajuste supera el contraste de normalidad de Shapiro.
Hay que señalar que esto no garantiza que haya sido o que será así en otros años.

En el resto de combinaciones no obtenemos ningún resultado significativamente distinto de 0, por lo que podemos concluir que existe una evolución proporcional entre España e Italia, pero no entre las demás combinaciones. Esto puede deberse a la selección de países, es probable, que si probamos alguna relación entre países con características similares, por ejemplo, España-Portugal o Francia-Bélgica también observemos cierta relación, pero esto dejamos que el lector lo investigue si quiere ahondar en el tema.

Conclusiones

Presentamos este trabajo con el fin de facilitar el cálculo del índice ERPE para cualquier individuo, ya que es difícil acceder a los datos microeconómicos de la encuesta de calidad de vida a partir de los que se elaboran estas variables y este índice.

Hemos quedado satisfechos con la función obtenida, ya que hemos demostrado que es una buena aproximación lineal del tema que nos atañe.

Las funciones determinadas han sido:

$$\hat{Y} = 2.20 + 0.83X_1 + 0.52X_2 + 0.30X_3$$

$$\hat{Y} = 3{,}76 + 0{,}83X_1 + 0{,}52X_2 + 0{,}01X_3^2$$

Esta función la interpretamos como:

- Si X_1 aumenta una unidad mientras el resto de variables permanecen constantes, el índice ERPE aumenta 0,83.
- Si X_2 aumenta una unidad mientras el resto de variables permanecen constantes, el índice ERPE aumenta 0,52.
- Si X_3 aumenta una unidad mientras el resto de variables permanecen constantes, el ERPE aumenta en el primer caso 0,30, y si tomamos la variable transformada 0,01

También hemos visto algunas de las características intrínsecas de cada país, por ejemplo, para España, la variable más significativa es la que mide la intensidad laboral, lo que concuerda con la idea de que en España el principal problema viene por el mercado laboral. Sin embargo, para Bélgica el riesgo de pobreza y exclusión social viene determinado (en gran medida) por la carencia material, es decir, por aquella parte de la población que no tiene acceso a algunos de los puntos incluidos en la variable X2.

Podemos concluir que el tiempo no es una variable determinante, lo que nos indica que el ERPE no evoluciona simplemente por el paso del tiempo, es decir, por la evolución natural de la economía (inflación, cambio de moneda, modificaciones de la definición de pobreza, etc.)

Hemos buscado también si se ha dado una evolución paralela de este indicador en los países con el fin de encontrar un patrón de evolución proporcional, pero los resultados no nos muestran ninguna relación clara.

Sólo hemos podido demostrar esta relación entre España e Italia, lo que tiene sentido puesto que son dos países mediterráneos y comparten algunas similitudes.

Quizás ampliando este estudio, contemplando más países, se puede determinar un patrón que nos permita clasificar a los países en diferentes grupos, pero esto es sólo una hipótesis que se podría estudiar en una extensión de este trabajo.

Referencias

[1] Casado, D. *Apuntes de Métodos de Regresión*

[2] Datos: Eurostat, http://ec.europa.eu/eurostat/data/database

[3] Estrategia Europa 2020: http://ec.europa.eu/europe2020/index_es.htm

[4] Montgomery, D.C., E.A. Peck and G.G. Vining (2001, 3rd ed). Introduction to Linear Regression Analysis. John Wiley & Sons, Inc.

[5] Newbold, Paul, L. Carlson, William, M. Thorne, Betty. *Estadística para administración y economía*, 8ª Edición. Pearson.

[6] Novales Cinca, A. *Análisis de Regresión. Apuntes de Econometría superior*. https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf

[7] Novales Cinca, A. *Estadística y Econometría*. McGraw-Hill.

Apéndice 1. Tablas de los datos utilizados en el modelo 2 (Estudio longitudinal)

Portugal	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	26,1	25,0	25,0	26,0	24,9	25,3	24,4	25,3	27,5	27,5
X1	19,4	18,5	18,1	18,5	17,9	17,9	18,0	17,9	18,7	19,5
X2	9,3	9,1	9,6	9,7	9,1	9,0	8,3	8,6	10,9	10,6
X3	6,0	6,6	7,2	6,3	7,0	8,6	8,3	10,1	12,2	12,2

Italia	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	25,6	25,9	26,0	25,5	24,9	25,0	28,1	29,9	28,5	28,3
X1	19,2	19,3	19,5	18,9	18,4	18,7	19,8	19,5	19,3	19,4
X2	6,8	6,4	7,0	7,5	7,3	7,4	11,1	14,5	12,3	11,6
X3	11,0	11,3	10,2	10,4	9,2	10,6	10,5	10,6	11,3	12,1

Grecia	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	29,4	29,3	28,3	28,1	27,6	27,7	31,0	34,6	35,7	36,0
X1	19,6	20,5	20,3	20,1	19,7	20,1	21,4	23,1	23,1	22,1
X2	12,8	11,5	11,5	11,2	11,0	11,6	15,2	19,5	20,3	21,5
X3	7,6	8,1	8,1	7,5	6,6	7,6	12,0	14,2	18,2	17,2

Alemania	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	18,4	20,2	20,6	20,1	20,0	19,7	19,9	19,6	20,3	20,6
X1	12,2	12,5	15,2	15,2	15,5	15,6	15,8	16,1	16,1	16,7
X2	4,6	5,1	4,8	5,5	5,4	4,5	5,3	4,9	5,4	5,0
X3	12,0	13,6	11,5	11,7	10,9	11,2	11,2	9,9	9,9	10,0

Francia	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	18,9	18,8	19,0	18,5	18,5	19,2	19,3	19,1	18,1	18,5
X1	13,0	13,2	13,1	12,5	12,9	13,3	14,0	14,1	13,7	13,3
X2	5,3	5,0	4,7	5,4	5,6	5,8	5,2	5,3	4,9	4,8
X3	8,7	9,1	9,6	8,8	8,4	9,9	9,4	8,4	8,1	9,6

Bélgica	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Y	22,6	21,5	21,6	20,8	20,2	20,8	21,0	21,6	20,8	21,2
X1	14,8	14,7	15,2	14,7	14,6	14,6	15,3	15,3	15,1	15,5
X2	6,5	6,4	5,7	5,6	5,2	5,9	5,7	6,3	5,1	5,9
X3	15,1	14,3	13,8	11,7	12,3	12,7	13,8	13,9	14,0	14,6

Apéndice 2. Código utilizado para el análisis de los modelos

El programa utilizado para realizar el análisis de los datos en el trabajo es R-project. A continuación está disponible todo el código utilizado a lo largo del trabajo, con el fin de facilitar al lector la repetición de nuestro análisis

Lo primero que hacemos es instalar algunos paquetes adicionales:

```
# Instalar paquetes (los que se pueda)
install.packages('e1071')
install.packages('tseries')
install.packages('randtests')
install.packages('lmtest')
install.packages('car')
# Cargar paquetes (al inicio de cada sesión de R)
library(e1071)
library(tseries)
library(randtests)
library(lmtest)
library(car)
```

Modelo 1. Datos de muchos países en 2014

```
TodosERPE = c(21.2, 40.1, 14.8, 17.8, 20.6, 26.0, 27.4, 36.0, 29.2, 18.5, 29.3, 28.3, 27.4, 32.7,
27.3, 19.0, 31.1, 23.8, 16.5, 19.2, 24.7, 27.5, 40.2, 20.4, 18.4, 17.3, 16.9, 24.1, 11.2, 13.5)
TodosRiesgo = c(15.5, 21.8, 9.7, 11.9, 16.7, 21.8, 15.3, 22.1, 22.2, 13.3, 19.4, 19.4, 14.4, 21.2,
19.1, 16.4, 14.6, 15.9, 11.6, 14.1, 17.0, 19.5, 25.4, 14.5, 12.6, 12.8, 15.1, 16.8, 7.9, 10.9)
TodosCarencia = c(5.9, 33.1, 6.7, 3.2, 5.0, 6.2, 8.4, 21.5, 7.1, 4.8, 13.9, 11.6, 15.3, 19.2, 13.6,
1.4, 23.9, 10.2, 3.2, 4.0, 10.4, 10.6, 26.3, 6.6, 9.9, 2.8, 0.7, 7.3, 1.4, 1.2)
TodosParo = c(14.6, 12.1, 7.6, 12.1, 10.0, 7.6, 21.0, 17.2, 17.1, 9.6, 14.7, 12.1, 9.7, 9.6, 8.8, 6.1,
12.2, 9.8, 10.2, 9.1, 7.3, 12.2, 6.4, 8.7, 7.1, 10.0, 6.4, 12.2, 4.9, 5.9)

Y = TodosERPE
X1 = TodosRiesgo
X2 = TodosCarencia
X3 = TodosParo

# matriz de correlaciones
MATRIX = cbind(Y, X1, X2, X3); pairs(MATRIX); cor(MATRIX)
#Ajustamos el modelo
linReg = lm(Y ~ X1 + X2 + X3)
summary(linReg)
# Observamos los residuos para validar el ajuste
plot(linReg)
res = linReg$residuals
hist(res)
shapiro.test(res)
skewness(res);  kurtosis(res)  # Asimetría y exceso de curtosis.

#Transformación de variables:
```

- Logaritmo: renombramos X3 = log(TodosParo) Realizamos el mismo proceso anterior
- Cuadrado: renombramos X3 = TodosParo^2. Realizamos el mismo proceso anterior

Ejercicio 1: Incluyendo los términos de segundo orden de la serie de Taylor

```
Y = TodosERPE
X1 = TodosRiesgo
X2 = TodosCarencia
X3 = TodosParo
X4 = X1^2
X5 = X2^2
X6 = X3^2
X7 = X1*X2
X8 = X1*X3
X9 = X2*X3

MATRIX = cbind(Y, X1, X2, X3, X4, X5, X6, X7, X8, X9); pairs(MATRIX); cor(MATRIX)
linReg = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9)
summary(linReg)
res = linReg$residuals
shapiro.test(res)
#Se van quitando las variables que no son significativamente distintas de 0.
```

Modelo 2. Modelo longitudinal para países individualmente de 2005 a 2014

```
## [España, 2005 a 2014] ##

EsERPE = c(24.3, 24.0, 23.3, 23.8, 24.7, 26.1, 26.7, 27.2, 27.3, 29.2)
EsRiesgo = c(20.1, 20.3, 19.7, 19.8, 20.4, 20.7, 20.6, 20.8, 20.4, 22.2)
EsCarencia = c(4.1, 4.1, 3.5, 3.6, 4.5, 4.9, 4.5, 5.8, 6.2, 7.1)
EsParo = c(6.9, 6.4, 6.8, 6.6, 7.6, 10.8, 13.4, 14.3, 15.7, 17.1)
Y = EsERPE
X1 = EsRiesgo
X2 = EsCarencia
X3 = EsParo^2

MATRIX = cbind(Y, X1, X2, X3); pairs(MATRIX); cor(MATRIX)

linReg = lm(Y ~ X1 + X2 + X3)
summary(linReg)
plot(linReg)
res = linReg$residuals
hist(res)
shapiro.test(res)

linReg = lm(Y ~ 0 + X1 + X3)
summary(linReg)
plot(linReg)
res = linReg$residuals
hist(res)
shapiro.test(res)
```

Ejercicio 2. ¿Quitar el efecto del tiempo?

```
t = 1:10
linReg = lm(Y ~ 0 + X1 + X3 + t)
summary(linReg)
res = linReg$residuals
shapiro.test(res)
```

[Italia, 2005 a 2014]

```
ItERPE = c(25.6, 25.9, 26.0, 25.5, 24.9, 25.0, 28.1, 29.9, 28.5, 28.3)
ItRiesgo = c(19.2, 19.3, 19.5, 18.9, 18.4, 18.7, 19.8, 19.5, 19.3, 19.4)
ItCarencia = c(6.8, 6.4, 7.0, 7.5, 7.3, 7.4, 11.1, 14.5, 12.3, 11.6)
ItParo = c(11.0, 11.3, 10.2, 10.4, 9.2, 10.6, 10.5, 10.6, 11.3, 12.1)
Y = ItERPE
X1 = ItRiesgo
X2 = ItCarencia
X3 = ItParo^2
```

[Portugal, 2005 a 2014]

```
PoERPE = c(26.1, 25.0, 25.0, 26.0, 24.9, 25.3, 24.4, 25.3, 27.5, 27.5)
PoRiesgo = c(19.4, 18.5, 18.1, 18.5, 17.9, 17.9, 18.0, 17.9, 18.7, 19.5)
PoCarencia = c(9.3, 9.1, 9.6, 9.7, 9.1, 9.0, 8.3, 8.6, 10.9, 10.6)
PoParo = c(6.0, 6.6, 7.2, 6.3, 7.0, 8.6, 8.3, 10.1, 12.2, 12.2)
Y = PoERPE
X1 = PoRiesgo
X2 = PoCarencia
X3 = PoParo
```

[Grecia, 2005 a 2014]

```
GrERPE = c(29.4, 29.3, 28.3, 28.1, 27.6, 27.7, 31.0, 34.6, 35.7, 36.0)
GrRiesgo = c(19.6, 20.5, 20.3, 20.1, 19.7, 20.1, 21.4, 23.1, 23.1, 22.1)
GrCarencia = c(12.8, 11.5, 11.5, 11.2, 11.0, 11.6, 15.2, 19.5, 20.3, 21.5)
GrParo = c(7.6, 8.1, 8.1, 7.5, 6.6, 7.6, 12.0, 14.2, 18.2, 17.2)
Y = GrERPE
X1 = GrRiesgo
X2 = GrCarencia
X3 = GrParo^2
```

[Alemania, 2005 a 2014]

```
AlERPE = c(18.4,20.2,20.6,20.1,20.0,19.7,19.9,19.6,20.3,20.6)
AlRiesgo = c(12.2,12.5,15.2,15.2,15.5,15.6,15.8,16.1,16.1,16.7)
AlCarencia = c(4.6,5.1,4.8,5.5,5.4,4.5,5.3,4.9,5.4,5.0)
AlParo = c(12.0,13.6,11.5,11.7,10.9,11.2,11.2,9.9,9.9,10.0)
Y = AlERPE
X1 = AlRiesgo
X2 = AlCarencia
X3 = AlParo^2
```

[Bélgica 2005-2014]

```
BeERPE = c(22.6,21.5,21.6,20.8,20.2,20.8,21.0,21.6,20.8,21.2)
BeRiesgo = c(14.8,14.7,15.2,14.7,14.6,14.6,15.3,15.3,15.1,15.5)
BeCarencia = c(6.5,6.4,5.7,5.6,5.2,5.9,5.7,6.3,5.1,5.9)
BeParo = c(15.1,14.3,13.8,11.7,12.3,12.7,13.8,13.9,14.0,14.6)
Y = BeERPE
X1 = BeRiesgo
X2 = BeCarencia
X3 = BeParo^2
```

Modelo 3. Estudio de panel para evaluar el carácter predictivo del ERPE de años anteriores

```
## [Muchos países, del 2005 al 2014] ##

TodosERPE = c(21.2, 40.1, 14.8, 17.8, 20.6, 26.0, 27.4, 36.0, 29.2, 18.5, 29.3, 28.3, 27.4, 32.7,
27.3, 19.0, 31.1, 23.8, 16.5, 19.2, 24.7, 27.5, 40.2, 20.4, 18.4, 17.3, 16.9, 24.1, 11.2, 13.5)
TodosRiesgo = c(15.5, 21.8, 9.7, 11.9, 16.7, 21.8, 15.3, 22.1, 22.2, 13.3, 19.4, 19.4, 14.4, 21.2,
19.1, 16.4, 14.6, 15.9, 11.6, 14.1, 17.0, 19.5, 25.4, 14.5, 12.6, 12.8, 15.1, 16.8, 7.9, 10.9)
TodosCarencia = c(5.9, 33.1, 6.7, 3.2, 5.0, 6.2, 8.4, 21.5, 7.1, 4.8, 13.9, 11.6, 15.3, 19.2, 13.6,
1.4, 23.9, 10.2, 3.2, 4.0, 10.4, 10.6, 26.3, 6.6, 9.9, 2.8, 0.7, 7.3, 1.4, 1.2)
TodosParo = c(14.6, 12.1, 7.6, 12.1, 10.0, 7.6, 21.0, 17.2, 17.1, 9.6, 14.7, 12.1, 9.7, 9.6, 8.8, 6.1,
12.2, 9.8, 10.2, 9.1, 7.3, 12.2, 6.4, 8.7, 7.1, 10.0, 6.4, 12.2, 4.9, 5.9)
TodosERPE2013 =
c(20.8,48.0,14.6,18.3,20.3,23.5,29.5,35.7,27.3,18.1,29.9,28.5,27.8,35.1,30.8,19.0,
34.8,24.0,15.9,18.8,25.8,27.5,40.4,20.4,19.8,16.0,16.4,24.8,13.0,14.1)
Y = TodosERPE
X1 = TodosRiesgo
X2 = TodosCarencia
X3 = TodosParo^2
X4 = TodosERPE2013

MATRIX = cbind(Y, X1, X2, X3, X4); pairs(MATRIX); cor(MATRIX)

# ¿Añade información el ERPE anterior?
linReg = lm(Y ~ X1 + X2 + X3 + X4)
summary(linReg)
res = linReg$residuals
shapiro.test(res)

# ¿Añadir información nueva a X4?
linReg = lm(Y ~ X4)
summary(linReg)
res = linReg$residuals
shapiro.test(res)

## [España, 2014] ##

# Hemos quitado el dato de 2005 de las cuatro primeras variables. De la última, hemos quitado
el último dato
EsERPE = c(24.0, 23.3, 23.8, 24.7, 26.1, 26.7, 27.2, 27.3, 29.2)
EsRiesgo = c(20.3, 19.7, 19.8, 20.4, 20.7, 20.6, 20.8, 20.4, 22.2)
EsCarencia = c(4.1, 3.5, 3.6, 4.5, 4.9, 4.5, 5.8, 6.2, 7.1)
EsParo = c(6.4, 6.8, 6.6, 7.6, 10.8, 13.4, 14.3, 15.7, 17.1)
anteriorEsERPE = c(24.3, 24.0, 23.3, 23.8, 24.7, 26.1, 26.7, 27.2, 27.3)
```

```
Y = EsERPE
X1 = EsRiesgo
X2 = EsCarencia
X3 = EsParo^2
X4 = anteriorEsERPE
#realizamos el mismo ajuste que el anterior
```

Modelo 4. Diferencia entre países

```
## [Tres países, del 2005 al 2014] ##
# España
EsERPE = c(24.3, 24.0, 23.3, 23.8, 24.7, 26.1, 26.7, 27.2, 27.3, 29.2)
EsRiesgo = c(20.1, 20.3, 19.7, 19.8, 20.4, 20.7, 20.6, 20.8, 20.4, 22.2)
EsCarencia = c(4.1, 4.1, 3.5, 3.6, 4.5, 4.9, 4.5, 5.8, 6.2, 7.1)
EsParo = c(6.9, 6.4, 6.8, 6.6, 7.6, 10.8, 13.4, 14.3, 15.7, 17.1)

# Italia
ItERPE = c(25.6, 25.9, 26.0, 25.5, 24.9, 25.0, 28.1, 29.9, 28.5, 28.3)
ItRiesgo = c(19.2, 19.3, 19.5, 18.9, 18.4, 18.7, 19.8, 19.5, 19.3, 19.4)
ItCarencia = c(6.8, 6.4, 7.0, 7.5, 7.3, 7.4, 11.1, 14.5, 12.3, 11.6)
ItParo = c(11.0, 11.3, 10.2, 10.4, 9.2, 10.6, 10.5, 10.6, 11.3, 12.1)

# Francia
FrERPE = c(18.9, 18.8, 19.0, 18.5, 18.5, 19.2, 19.3, 19.1, 18.1, 18.5)
FrRiesgo = c(13.0, 13.2, 13.1, 12.5, 12.9, 13.3, 14.0, 14.1, 13.7, 13.3)
FrCarencia = c(5.3, 5.0, 4.7, 5.4, 5.6, 5.8, 5.2, 5.3, 4.9, 4.8)
FrParo = c(8.7, 9.1, 9.6, 8.8, 8.4, 9.9, 9.4, 8.4, 8.1, 9.6)

# Tres países
TresERPE = c(EsERPE, ItERPE, FrERPE)
TresRiesgo = c(EsRiesgo, ItRiesgo, FrRiesgo)
TresCarencia = c(EsCarencia, ItCarencia, FrCarencia)
TresParo = c(EsParo, ItParo, FrParo)
Y = TresERPE
X1 = TresRiesgo
X2 = TresCarencia
X3 = TresParo^2

# Método 1 de identificar al país: Variable cualitativa 'país'
X4 = c(rep(1,10), rep(2,10), rep(3,10))
MATRIX = cbind(Y, X1, X2, X3, X4); pairs(MATRIX); cor(MATRIX)
linReg = lm(Y ~ X1 + X2 + X3 + X4)
summary(linReg)

# Método 2 de identificar al país: Variables indicadoras X4 y X5
X4 = c(rep(0,10), rep(0,10), rep(1,10))
X5 = c(rep(0,10), rep(1,10), rep(0,10))

linReg = lm(Y ~ X1 + X2 + X3 + X4 + X5)
summary(linReg)

linReg = lm(Y ~ X1 + X2 + X3 + X4)
summary(linReg)
res = linReg$residuals
shapiro.test(res)
```

Modelo 5. Evolución paralela entre países

```
# [Varios países, 2005-2014] #
# España
EsERPE = c(24.3, 24.0, 23.3, 23.8, 24.7, 26.1, 26.7, 27.2, 27.3, 29.2)
# Alemania
AlERPE = c(18.4, 20.2, 20.6, 20.1, 20.0, 19.7, 19.9, 19.6, 20.3, 20.6)
# Francia
FrERPE = c(18.9, 18.8, 19.0, 18.5, 18.5, 19.2, 19.3, 19.1, 18.1, 18.5)
# Italia
ItERPE = c(25.6, 25.9, 26.0, 25.5, 24.9, 25.0, 28.1, 29.9, 28.5, 28.3)

MATRIX = cbind(EsERPE, UeERPE, AlERPE, FrERPE, ItERPE); pairs(MATRIX);
cor(MATRIX)

# Relacionar España con Italia
        1º linReg = lm(EsERPE ~ ItERPE)
           summary(linReg)
        2º linReg = lm(EsERPE ~ 0 + ItERPE)
           summary(linReg)
           res = linReg$residuals
           shapiro.test(res)

# Relacionar España con Francia
linReg = lm(EsERPE ~ FrERPE)
summary(linReg)
res = linReg$residuals
shapiro.test(res)

# Relacionar España con Alemania
linReg = lm(EsERPE ~ AlERPE)
summary(linReg)
res = linReg$residuals
shapiro.test(res)

#De igual manera se generan las relaciones entre los demás países.
```