Ingeniería Informática – 2022/2023

Análisis de esperanzas de vida

Raúl Aguilar Arroyo y Alberto Penas Diaz

INTRODUCCIÓN Y OBJETIVOS

Se pretende determinar si factores como consumo de alcohol, PIB, mortalidad infantil, el IMC, el nº de hospitales o la escolarización influyen en la esperanza de vida de un país. Asi como intentar localizar los mejores estimadores de esta.

Descripción de los datos:

Numero de observaciones muestrales: 3111
Fuentes: WHO y UNESCO

Variables:

- **Variable principal**
  - life_expectancy
- **Variables secundarias mas relevantes**
  - life_expect = esperanza de vida en años
  - alcohol = consumo de alcohol
  - age5-19obesity = obesidad por cada 100k niños
  - doctors = doctores por cada 100k habitantes
  - gni_capita = PIB per capita ($)
  - che_gdp = % inversión en sanidad del PIB
  - une_infant = mortalidad infantil por cada 100k

Nota: en los análisis se utilizará un nivel de significación del 5%, i.e. α = 0.05.
Nota 2: la variable life expect esta calculada por la WHO y la variable une_life esta calculada por la UNESCO

ESTUDIO DESCRIPTIVO

La figura 1 muestra las medidas resumen de la variable principal "life_expect", haremos una distinción entre países desarrollados y en desarrollo.
Las características mas llamativas de las variables son el amplio rango de valores [36,84], asi como la alta desviación estándar la cual es mucho mas notable en los países en desarrollo y menor en los países desarrollados, lo mismo sucede con todos los cuartiles. También se puede observar que el mínimo pertenece a los países en desarrollo (cuya media se encuentra por debajo de la media general), y el máximo a los desarrollados (cuya media se encuentra por encima de la media general), hecho muy lógico a priori.

	life_expect	developed	developing
Media	69,15	78,71	67,12
Desviación estandar	9,13	3,2	8,68
Coef. Variación	0,13	0,04	0,13
Mín	36,23	70,13	36,23
Máx	84,17	84,17	82,95
Asimetria	-0,72	-0,79	-0,67
Curtosis	-0,26	-0,29	-0,38
Q1	63,2	76,9	61,07
Q2(mediana)	71,6	79,59	69,41
Q3	75,54	81,14	73,82

Figura 1: Medidas resumen

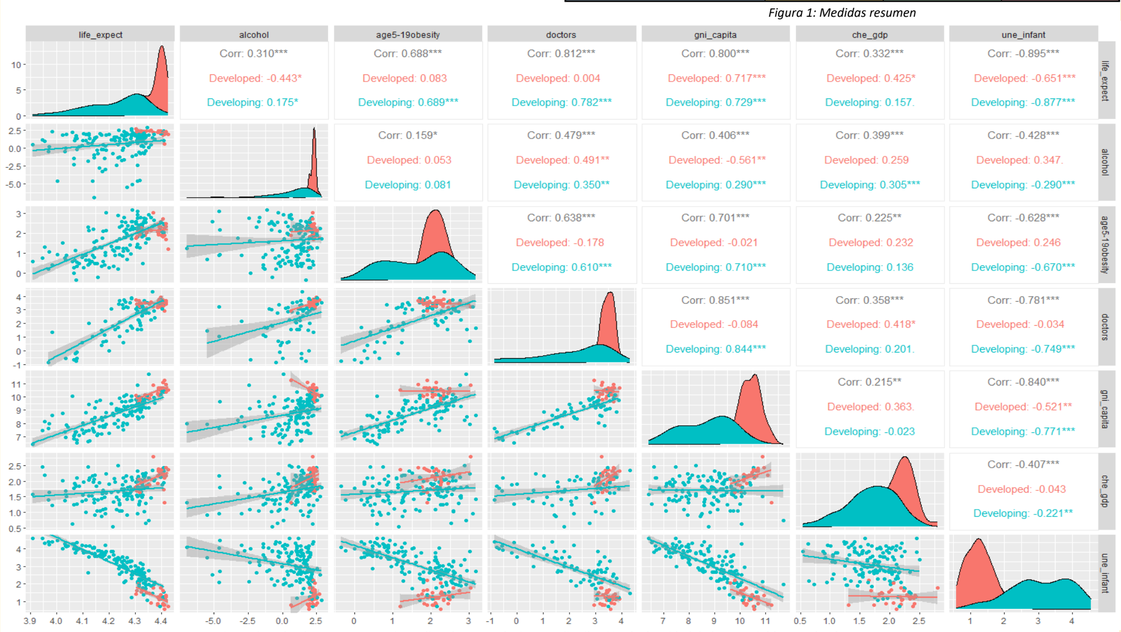

Figura 2: Nube de puntos, regresiones lineales , correlaciones y distribución de densidad de las 7 variables empleadas

La matriz anterior (figura 2) muestra un análisis multivariante de las variables secundarias mas destacables y la variable principal.
Para mejorar la legibilidad se ha aplicado una transformación logarítmica de los datos y se ha seleccionado solo un año (2013). Además se ha coloreado los países desarrollados en rojo y los países en desarrollo en azul.
Son especialmente interesantes las nubes de puntos que genera la esperanza de vida con respecto del resto de variables (columna 1), también es destacable la correlación que guarda con las estas variables (fila 1).
En la diagonal se pueden observar graficas de densidad que añadiendo la distinción por colores ayuda a visualizar como se reparten los datos .
En las nubes de puntos se puede ver como los países desarrollados tienen unos datos mas "compactos" y menos distribuidos que los países en desarrollo que muestran puntos mucho mas repartidos por el grafico.
A priori parece que la mortalidad infantil es un estimador excelente de la esperanza de vida, asi como sorprende la poca relación que guarda con la inversión en sanidad en países desarrollados, también sorprende la aparente ausencia de relación entre la esperanza de vida y el consumo de alcohol (incluso tiene una correlación pequeña pero positiva).
Cabe también mencionar que la división entre desarrollados y en desarrollo es muy relevante ya que muestran tendencias y correlaciones muy distintas entre ellos, por lo cual esta división hace que el estudio sea mas preciso.

AJUSTE DE DISTRIBUCIONES

En el análisis descriptivo hemos observado que contamos con una asimetría negativa importante, por lo que difícilmente los datos seguirán una distribución normal, asi que probaremos varias distribuciones a ver cual se ajusta mas. Por la forma que muestran los histogramas. Parece que las distribuciones que mas se asemejan a la forma de nuestros datos es la normal, log-normal o Weibull. Por lo que hemos probado las tres. En la figura 3 podemos ver que todas muestran un ajuste muy parecido por lo que por simplicidad escogeremos la distribución normal.
Si planteamos la siguiente hipótesis:
$H_0 \rightarrow \mu \neq \hat{x}$
$H_1 \rightarrow \mu = \hat{x}$
Obtenemos un p-value < 2.2e-16 es decir prácticamente 0, por lo que podemos asumir con casi total seguridad que los datos siguen una distribución normal de media $\hat{x} = 71{,}02$.
Por otro lado podemos intentar probar que la desviación muestral es igual $\sigma = \sigma_0$:
$H_0 \rightarrow \sigma^2 \neq \sigma_0^2$
$H_1 \rightarrow \sigma^2 = \sigma_0^2$
Lo cual nos da un r p-value = 0.02013 por lo que podemos rechazar la hipótesis nula y aceptar que $\sigma^2 = \sigma_0^2 = 8$
Por lo que podemos concluir que nuestros datos siguen una distribución normal Z(71.02, 8),

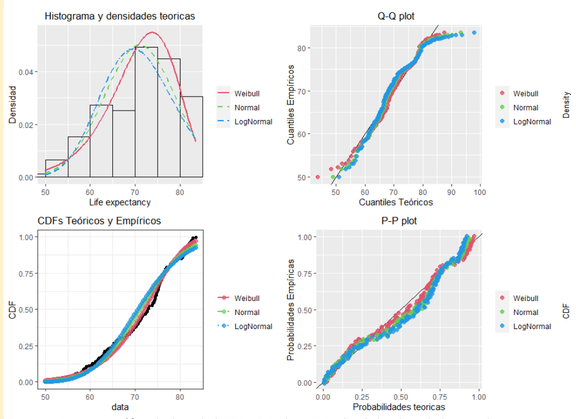

Figura 3: gráfico de densidad, CDFs, P-P plot y Q-Q plot de las densidades a probar

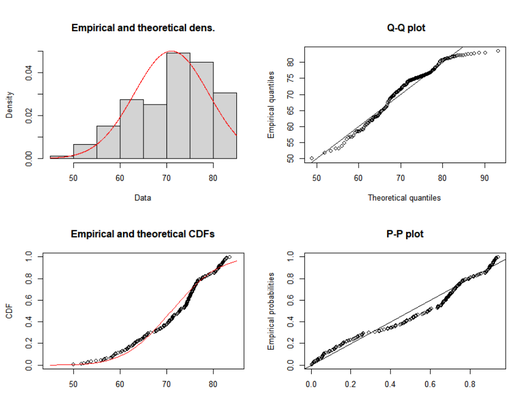

Figura 4: gráfico de densidad, CDFs, P-P plot y Q-Q plot de la distribución normal

INTERVALOS DE CONFIANZA Y CONTRASTE DE HIPÓTESIS

Queremos construir los intervalos de confianza para la media de la esperanza de vida los países según nuestros datos. Según la media y la desviación típica de la esperanza de vida, sacamos los histogramas correspondientes y comprobamos (véase apartado anterior) que los datos siguen una distribución parecida a una normal. También se podría realizar una prueba de bondad de ajuste para ver si nuestros datos se ajustan a una distribución normal, pero es redundante.

	Media	Desviación típica
Esperanza de vida	$\hat{\mu} = 69{,}15$ $IC_{95\%} = (68.829, 69.4708)$	$\hat{\sigma} = 9{,}13$ $IC_{95\%} = (8.6926, 9.6013)$
E.V. países desarrollados	$\hat{\mu} = 78.71$ $IC_{95\%} = (78.4411, 78.9789)$	$\hat{\sigma} = 3.20$ $IC_{95\%} = (2.8510, 3.6175)$
E.V. países en desarrollo	$\hat{\mu} = 67.12$ $IC_{95\%} = (66.78422, 67.45578)$	$\hat{\sigma} = 8.68$ $IC_{95\%} = (8.2239, 9.1752)$

Figura 8: Intervalos de confianza

Los intervalos salen muy estrechos debido al gran número de datos que se han utilizado para su estimación, lo que quiere decir que, con un 95% de confianza, estando en un país desarrollado cualquiera, la esperanza de vida media cae en el intervalo A y, estando en un país en vías de desarrollo, en el intervalo B.
Viendo la media

A pesar de que nuestros datos cuentan con un número significativamente mayor de información sobre países en vías de desarrollo, los intervalos de confianza obtenidos son suficientes como para confirmar que, el principal factor que define la media de la esperanza de vida es el de si un país se considera desarrollado o en desarrollo.

REGRESIÓN MÚLTIPLE

Primero, queremos saber qué variables secundarias son las más significativas a la hora de predecir nuestra variable principal (life_expectancy), las cuales nosotros consideramos:

•alcohol
•doctors
•che_gdp
•une_infant
Para ello utilizaremos los resultados de una regresión múltiple utilizando todas las variables secundarias.
Los resultados se muestran en la figura 5:

	Estimación	Error estándar	Estadístico T	P-valor
Alcohol	-0.2588	0.022	-11.729	0.00000
che_gdp	0.3538	0.034	10.283	0.00000
doctors	0.059	0.0076	7.835	0.00000
une_infant	-0.3085	0.0043	-70.393	0.00000

Figura 5: Estimaciones variables secundarias

Los p-valores de todas las variables son ínfimos, lo que significa que todas las variables empleadas para el modelo de regresión múltiple son altamente significativas y proporcionan gran linealidad a la hora de estimar nuestra variable principal. Esto, en términos prácticos, quiere decir que, tanto el consumo alcohol, como la obesidad por cada 100.000 niños, los doctores por cada 100.000 habitantes, el PIB per cápita y la mortalidad infantil son variables muy influyentes en la esperanza de vida por país. Nótese también que no hemos hecho distinción entre paises desarrollados y en vías de desarrollo, pues la variabilidad entre ambos modelos sería ínfima, teniendo en cuenta la dependencia de la esperanza de vida con todas las variables estudiadas.
La linealidad de la regresión, junto con la tabla anterior nos dice además que el alcohol y la mortalidad infantil, de entre las variables escogidas, son claramente lo que más contribuyen a la reducción en la esperanza de vida.

Por último, el modelo de regresión múltiple nos quedaría de la siguiente manera:

Esperanza de vida = 76.81 - 0.2588alcohol + 0.3538che_gdp + 0.059doctors - 0.3085une_infant

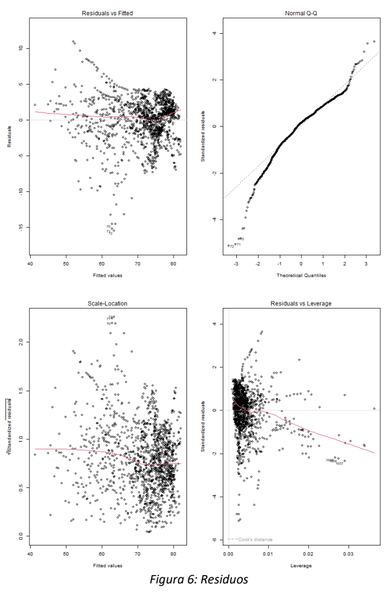

Figura 6: Residuos

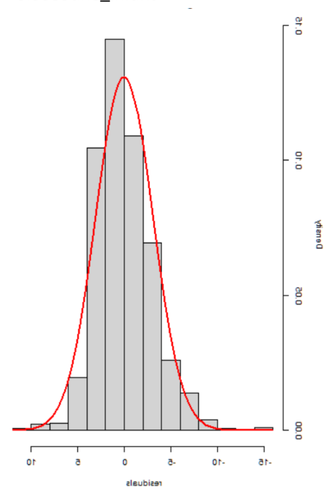

Figura 7: Normalidad de los residuos

Para finalizar, evaluaremos la normalidad de los residuos, realizando el histograma que se muestra en la figura 7. En esta se muestra que los residuos tienen una pequeña asimetría pero que aproximadamente pueden considerarse normales, lo que valida el modelo de regresión elegido.

CONCLUSIONES

Para hacer que este texto ocupe más espacio, se pueden agregar más detalles y ejemplos que respalden las afirmaciones mencionadas. Por ejemplo:

En concreto, se ha podido observar que en los países desarrollados, la esperanza de vida supera los 78 años en promedio, mientras que en los países en desarrollo se ubica por debajo de los 68 años.
En cuanto a la dispersión en la esperanza de vida, se ha podido constatar que en los países desarrollados existe una variabilidad menor en la esperanza de vida de sus ciudadanos, mientras que en los países en desarrollo se observa una mayor variabilidad en esta medida.
En cuanto a las variables inversión en sanidad y doctores por cada 100k ciudadanos, se ha podido observar que en los países donde se registran valores altos en estas variables, mientras que a su ves tienen valores bajos de alcohol y mortalidad infantil, observan una mayor esperanza de vida. Es decir las variables inversión en sanidad, doctores, alcohol y mortalidad infantil son muy influyentes a la hora de estimar la esperanza de vida.