U N A M

FES ACATLÁN

ANÁLISIS ESTADISTICO DE LOS FACTORES QUE INFLUYEN EN LA ESPERANZA DE VIDA

MODELO DE REGRESIÓN LINEAL

ALUMNOS:

CHÁVEZ CARLOS MARÍA DE JESÚS 417057808

ESCALANTE LÓPEZ JULIO CÉSAR 314100885

LÓPEZ RODRÍGUEZ ALICIA 314110730

RABELL MAGAÑA SHIRLEY ANAHI 314297699

VENTURA GARCIA IVAN 314254380

Resumen

El análisis de regresión es la técnica estadística de uso más frecuente para investigar y modelar la relación entre variables. Su atractivo y utilidad, generalmente son resultado del proceso conceptualmente lógico de usar una ecuación para expresar la relación entre una variable de interés (la respuesta) y un conjunto de variables relacionadas.

El presente trabajo busca interpretar, analizar y con ello buscar un optimo modelo de regresión lineal, que nos ayude a entender factores que pueden influir en la esperanza de vida.

Especificamente por el tipo de cantidad de variables consideradas, se emplea para la obtención de resultados el método estadístico de regresión lineal múltiple, lo que permite al final presentar un método que con mayor profundidad y sin pérdida de generalidad, nos proporcione mayor información en la relación de la esperanza de vida y los factores que contribuyen a fluctuaciones de la misma.

Palabras clave: Esperanza de vida, correlación de variables, modelo de regresión lineal múltiple.

Índice

1. Marco Teórico **5**
1.1. Del contexto de las variables 5
1.2. Del modelo de estudio 6

2. Variables **9**
2.1. Características naturales 9
2.2. Análisis exploratorio 10

3. Estimación del modelo y validación de hipotesis **13**
3.1. Estimación del modelo 14

4. Validación de supuestos **15**
4.1. Normalidad 15
4.2. Homocedasticidad 16
4.3. Independencia 17
4.4. Forma funcional errónea 18

5. Conclusiones **19**

6. Bibliografía **20**

Índice de cuadros

1. Características generales de las variables de estudio 9
2. Estadísticos principales 12
3. Resultados usando la función lm de R 14

Índice de figuras

1. Matriz de gráficas de dispersión 11
2. Esperanza de vida 11
3. Histograma: Variables vs Densidad (teorica y propuesta) 12
4. Dispersión, densidad y correlación de los datos 13
5. Normal Q-Q 16
6. Residuos vs Ajustados 17
7. Errores vs Indices de observación 18

1. Marco Teórico

1.1. Del contexto de las variables

Al comienzo del siglo XIX la esperanza de vida comenzó a incrementarse significativamente en los países más industrializados, lo que llevó a una desigualdad en la distribución de salud en todo el mundo, a medida que ha avanzado el tiempo ha mejorado la tecnología dando como resultado mejor calidad en la salud y una mayor esperanza de vida. De igual forma en esa época las enfermedades infecciosas sin los medicamentos y medidas de salud pública competentes resultaban en que las personas muerieran a una edad muy temprana, sin embargo esto ha ido cambiando a lo largo del tiempo porque empezó a haber una mejora en estos generando la "transición de la salud", llamada así por epidemiólogos debido a que en esta etapa empezó a aumentar sustancialmente la expectativa de vida (Roser M. et. al., 2019). En cambio este aumento propicia un mayor impacto de las enfermedades crónicas no transmisibles sobre la salud de personas de 60 años y más, debido a que hay un aumento en la proporción de este grupo y son más vulnerables a adquirirlas. (Domínguez A. et. al. 2005).

Otro aspecto importante es la carga negativa asociada a las enfermedades que usualmente se cuantifica mediante tasas de mortalidad, prevalencia o de incidencia, además de los años de vida potencial perdidos que la mortalidad genera en distintas edades, pero con ninguna de estas se puede cuantificar de manera factible dicho efecto sobre la calidad de vida (Domínguez A. et. Al., 2005).

Cuando el gasto público en salud es bajo, en los países de bajos ingresos el déficit se compensa con el gasto privado, aproximadamente en un 85 % a cargo de los usuarios. Esto significa que los pagos se realizan directamente al acceder a los servicios de salud. Dichos pagos no permite mancomunar riesgos, y pueden muy bien conducir a desembolsos ruinosos con grandes probabilidades de sumir a los hogares en la pobreza (Organización Mundial de la Salud, 2009).

Además los recursos externos se están convirtiendo en una fuente muy importante de financiación de la salud en los países de bajos ingresos, esto porque los recursos externos representaron el 17 % del gasto en salud en estos países en 2006, en comparación con el 12 % del gasto sanitario total en el año 2000, y en algunos países de bajos ingresos, dos tercios del total del gasto sanitario se financia mediante recursos externos. En estas situaciones, la previsibilidad de la asistencia es una preocupación importante (Organización Mundial de la Salud, 2009).

Asimismo diversas investigaciones han estudiado la relación entre la esperanza de vida y el ingreso nacional per cápita. Por su parte Preston encontró una relación positiva entre estas variables, sin embargo ante un crecimiento de los ingresos, el aumento de la esperanza de vida en los países pobres es más significativo mientras que en los países ricos su

efecto disminuye. En este influyente estudio, se encontró que la esperanza de vida al nacer de países de todo el mundo estaba relacionada con su ingreso nacional bruto per cápita, una cantidad que se utiliza con frecuencia como indicador de riqueza nacional, nivel de vida y desarrollo económico. Preston demostró que la relación transversal entre la esperanza de vida y el ingreso nacional per cápita entre países se puede describir con precisión mediante la llamada curva de Preston, con rápidos aumentos de la esperanza de vida en los países con ingresos más bajos y aumentos más lentos en los países con ingresos más altos (Shkolnikov et. al, 2019).

Otros autores han establecido la relación que existe entre la esperanza de vida y el gasto en salud. Lago (2013) concluye que existe una asociación positiva entre estas variables, y ante el aumento de los recursos dirigidos al sector salud se presenta una disminución de la tasa de mortalidad materna e infantil y un incremento de la esperanza de vida. En particular, indica que cuando aumenta el gasto total en salud también aumenta la esperanza de vida aunque a una tasa decreciente, lo que que podría deberse a la existencia de rendimientos decrecientes del gasto en términos de los resultados en salud. Es decir, cuando los niveles de ingreso y gasto son bajos, aumentarlos genera un mayor impacto en el estatus de salud poblacional, el cual se vuelve cada vez menos significativo a medida que el gasto aumenta.

Al igual que la relación entre la esperanza de vida y el ingreso nacional per cápita, el efecto marginal de un aumento porcentual del gasto en salud (% PIB) sobre los resultados en salud es diferente entre países según su nivel de desarrollo. Esto es, el impacto es mayor para aquellos países que se encuentran en una etapa temprana del desarrollo; en cambio, tiene un menor efecto para aquellos que han alcanzado altos niveles de desarrollo (Sanmartín, 2019).

1.2. Del modelo de estudio

Para tratar de identificar que factores influyen en la esperanza de vida, se hara un modelo de regresión lineal múltiple, entendiendo la ecuación como:

$$Yi = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_n X_{in} + \varepsilon_i$$

Donde

- α: ordenada al origen.
- β_i: promedio que tiene el incremento en una unidad de la variable predictora X_i sobre la variable dependiente Y, manteniéndose constantes el resto de variables.
- ε: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

Para saber si los parámetros asociados al modelo son significativos, necesitamos una prueba de hipótesis de la forma:

$$H_0 : \beta_j = 0 \qquad vs \qquad H_1 : \beta_j \neq 0$$

Si no rechazamos la hipótesis nula, se dice que la variable X_j no es significativa en el modelo de regresión.

Para saber el desempeño del modelo usamos el coeficiente de determinación múltiple R^2 ajustado, el cual toma en cuenta el número de variables en el modelo:

$$\hat{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right)$$

Donde k es el número de variables.

Dicho contraste utiliza la R^2, la cual representa la calidad de ajuste de una regresión, y mide la variabilidad explicada por el modelo contra la variabilidad total de la muestra. Además el R^2 ajustado, toma en cuenta la cantidad de variables en el modelo como se vió previamente en la fórmula, obteniendo así un mejor indicador de nuestro modelo.

Otra cosa que es importante mencionar es el contraste de la significancia total del modelo, esta prueba de hipótesis esta como sigue:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0 \qquad vs \qquad H_1 : \beta_j \neq 0 \text{ para algún j}$$

En esta prueba de hipótesis se utiliza el estadístico de prueba F, si tenemos un $p - value < \alpha$ se rechaza la hipótesis y el modelo explica la variable respuesta en función de las variables con las que se pretende expresar.

Si no se cumple la significancia del modelo, de alguna de las variables o la calidad de ajuste de este es muy bajo, entonces se puede verificar la forma funcional erronea con un contraste de hipotesis (*Test RESET de Ramsey*) donde las hipotesis son las siguientes:

$$H_0 : \delta_1 = \delta_2 = 0 \qquad vs \qquad H_1 : \delta_1 \neq 0 \wedge \delta_2 \neq 0$$

Dicho contraste sirve para saber si nuestras variables explicativas siguen alguna potencia para poder ajustarse mejor, además se utiliza cuando tenemos sospechas de no linealidad.

Por otro lado, si la muestra $\left\{(Y_i, X_i^t)^t\right\}; i \in \mathbb{N}$ está dada, entonces:

$$\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$$

Para lo cual, se debería de cumplir que la muestra $\{\varepsilon_i\} : i \in \mathbb{N}$ cumple con las siguientes caracteristicas:

- **Provienen de una distribución normal**

En este punto verificaremos que la distribución de los errores (ε) con base en $\hat{\varepsilon}_1$,...,$\hat{\varepsilon}_m$ se distribuyen normales, en particular con media cero y varianza constante.

Esto lo contrastamos a través de una prueba de bondad de ajuste, tales como: Kolmogorov-Smirnov, Cramér-Von Mises y Anderson-Darling, donde se tiene interés sobre las hipótesis $H_0 : F = F_0$ vs $F \not\equiv F_0$, las cuales se basan en una comparación de la función de distribución empírica, denotada por F_n, y la función de distribución planteada en la hipótesis nula (Javier Santibáñez,2018), que en este caso será la dristribución normal, para las cuales usaremos un α del 0.05. SI *p-value* $>\alpha$ NO rechazamos nuestra hipótesis nula H_0, por lo cual no rechazariamos normalidad.

- **Varianza constante (supuesto de homocedasticidad)***

En este punto además de usar gráficas (donde se observa si hay patrones entre $\hat{\varepsilon} - Y$), se hará uso del *Test de Breusch-Pagan*, donde la hipótesis nula es:

$$H_0 : \sigma^2 = \sigma^2 \text{ vs } H_a : \sigma^2 \not\equiv \sigma^2 \text{para algún i} \not\equiv j$$

donde:

1) Se estima el modelo y se calculan los errores estimados ($\tilde{\varepsilon}$).

2) Trabajar con los errores cuadráticos normalizados con el EMV de σ^2_ε, es decir:

$$\hat{\varepsilon_{STi}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon}$$

3) Ajustar el modelo:

$$\hat{\varepsilon_{ST}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

4) Rechaza la hipótesis nula si x_i es significativa para algún i

Por lo tanto, si el valor p asociado a una prueba de heterocedasticidad cae por debajo de un cierto umbral (por ejemplo, 0.05), llegaríamos a la conclusión de que los datos son significativamente heterocedásticos. Para corregir la varianza no constante del modelo se utiliza la tecnica de Mínimos Cuadrados Generalizados (GLS).

- **Correlación nula (supuesto de independencia)****

Llegado a este punto, se plantea el contraste de hipótesis de Durbin-Watson por lo que se propone el siguiente modelo

$\hat{\varepsilon}_i = (\hat{\varepsilon}_i + u; u\ N(0, \sigma_u^2))$

donde $H_0 : \rho = 0$ vs $H_a : \rho \not\equiv 0$

El estadístico de prueba es:

$$d_{DW} \approx 2(1 - \hat{p})$$

- Si $\hat{\rho} = 1$ entonces $d_{DW} = 0$ (correlación lineal directa)
- Si $\hat{\rho} = -1$ entonces $d_{DW} = 4$ (correlación lineal inversa)
- Si $\hat{\rho} = 0$ entonces $d_{DW} = 2$ (correlación nula)

Además bajo el supuesto que $\hat{\varepsilon} \sim N$ entonces si $\rho = 0$ indicará correlación nula siendo asì que los errrores son independientes.
En caso de rechazar la hipótesis $\rho = 0$, entonces corregimos el modelo trabajando con un esquema de modelos autoregresivos.

* Si se rechaza H_0, tenemos heterocedasticidad.
**Si se rechaza H_0, tenemos autocorrelación.

2. Variables

2.1. Características naturales

Para el presente documento se trabajaron con datos del año 2014 con las siguientes variables, las cuales vienen explicadas en el Cuadro 1.

Nombre	Descripción
Esperanza de vida	Esperanza de vida en años
Mortalidad Adulta	Tasas de mortalidad de adultos de ambos sexos (probabilidad de morir entre 15 y 60 años por 1000 habitantes)
Gasto total	Gasto del gobierno general en salud como porcentaje del gasto público total (%)
VIH / SIDA	Muertes por cada 1000 nacidos vivos VIH / SIDA (0-4 años)
Composición de ingresos de los recursos	Índice de desarrollo humano en términos de composición de ingresos de los recursos (índice que varía de 0 a 1)

Cuadro 1: Características generales de las variables de estudio

Además se trabajaron con datos de 131 países de un total de 183, esto porque existían valores faltantes en los registros y podría afectar a nuestro modelo. Se cuenta con información de países desarrollados como los que se encuentran en desarrollo al año en que se obtuvieron los datos.

Cabe mencionar que se los datos se obtuvieron de Kaggle de la siguiente liga: *https://www.kaggle.com/kumarajarshi/life-expectancy-who*, y de dicha información se realizó una muestra para poder manipular mejor los datos, pues se contaban con datos de los años 2000-2015, y trabajar con variables indexadas al tiempo es tema de series de tiempo.

Es importante señalar que los nombres de las variables en los gráficos y tablas mostrados a continuación se encuentran en inglés.

2.2. Análisis exploratorio

Es importante conocer cómo se comportan nuestros datos, para ello en la Figura 1 se muestra una matriz de gráficas donde se observa la dispersión que tienen contra cada una de las variables.

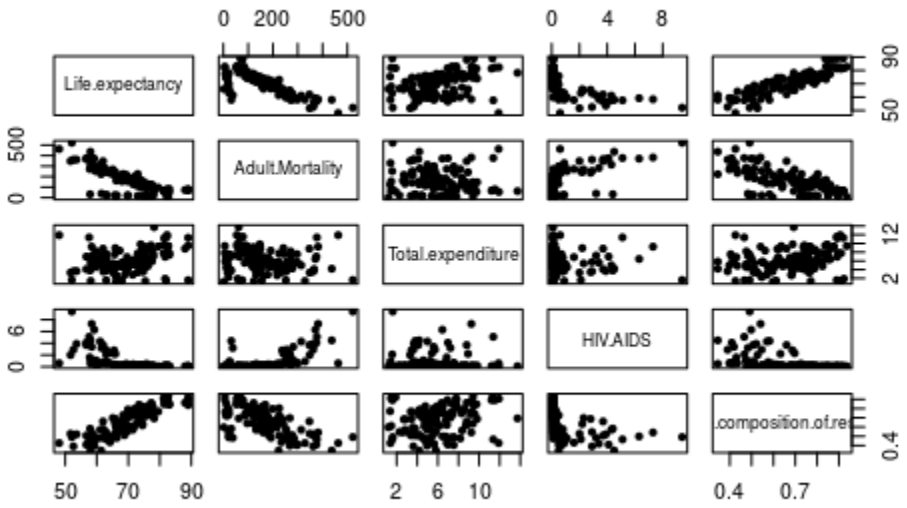

Figura 1: Matriz de gráficas de dispersión

Además es importante ver cómo se distribuyen los datos de nuestra variable principal (Esperanza de vida), para ello se realizó el histograma de la misma, donde podemos observar que tiene su valor mínimo debajo de 50 años y su valor máximo en 90 años, lo que nos indica que la esperanza de vida de la población en 2014 estaba entre los 50-90 años, de igual forma podemos ver que la media está entre los 67 y 72 años aproximadamente.

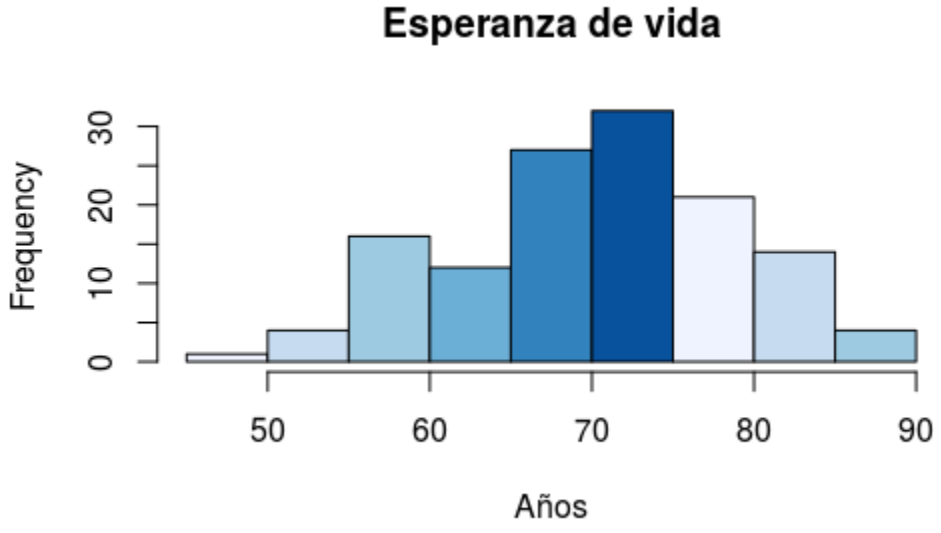

Figura 2: Esperanza de vida

Del mismo modo, para conocer cómo se comportan nuestras variables ya no de forma visual sino estadística, calculamos los estadísticos principales de estas los cuales podemos observar en la siguiente tabla (Cuadro 2).

	Min.	1st QU.	Median	Mean	3rd Qu.	Max
Life. expectancy	48.1	64.65	72.0	70.51	75.8	89.0
Adult. Mortality	2.0	74.50	144.00	160.37	225.00	522.00
Total.expenditure	1.21	4.48	5.82	6.10	7.63	13.73
HIV.AIDS	0.10	0.10	0.10	0.80	0.50	9.40
Income.composition. of.resources	0.34	0.54	0.69	0.66	0.779	0.936

Cuadro 2: Estadísticos principales

En la tabla anterior podemos observar que los diversos países destinan en promedio 6.10 % del gasto total en gastos de salud, y el máximo porcentaje destinado es de 13.73 % en contraparte el mínimo de 1.2 %. La composición de ingresos fue de 0.345 a 0.936 con un promedio de 0.66, es decir, el índice de desarrollo humano en términos de la composición de recursos estuvo en promedio en 66 %.

Por otra parte, las muertes de niños de 0-4 años por VIH/SIDA, fue de 0.100 a 9.4 por cada 1000, lo cual es relativamente bajo. Mientras las tasas de mortalidad en adultos (15-60 años por cada 1000 habitantes) estuvo en promedio en 160.37.

Así mismo, se graficaron los histogramas de las variables comparándolas con la densidad de la distribución normal (línea roja) y con su propia densidad (línea azul), lo cual se observa en la Figura 3.

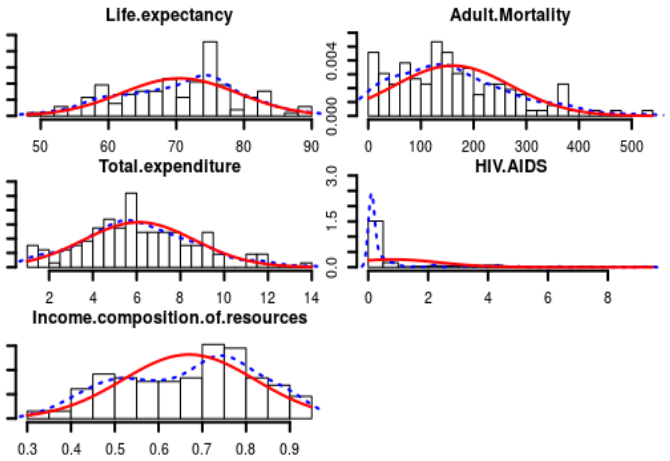

Figura 3: Histograma: Variables vs Densidad (teorica y propuesta)

Como se puede observar no se ajustan tan bien a una distribución normal, y sus densidades tienen dos picos o son a una cola.

Otra cosa que es relevante para nuestro estudio es la correlación que existe entre cada una de las variables, y obtuvimos que las variables más correlacionadas entre sí son "Composición de ingresos de los recursos" - "Esperanza de vida" con una correlación de 89.2 % y una correlación negativa entre "Mortalidad adulta"-"Esperanza de vida" de 77 %. Además las menos correlacionadas fueron "VIH/SIDA"-"Total de gasto" con -9.6 % y "Total de gasto"-"Mortalidad adulta" con -14 %.

En la Figura 4 podemos observar la dispersión, densidad y correlación de los datos.

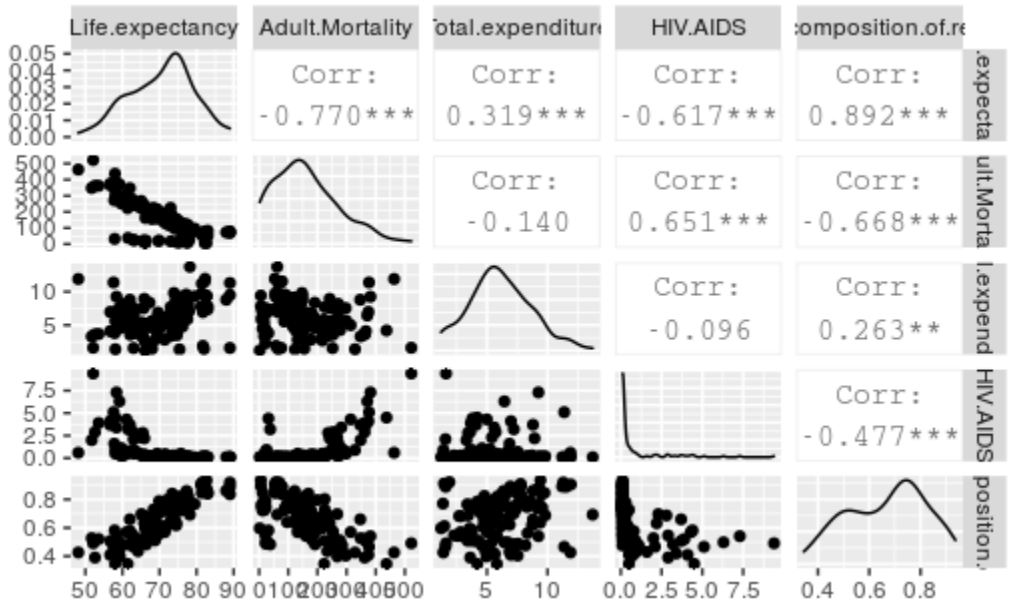

Figura 4: Dispersión, densidad y correlación de los datos

3. Estimación del modelo y validación de hipotesis

En la mayoría de las investigaciones –sin importar el campo del conocimiento en las que se desarrollen– en las cuales se realicen mediciones, observaciones o experimentos de donde se obtengan datos de diferentes variables; es fundamental determinar algún tipo de relación de dependencia entre las variables con el fin de hacer predicciones o pronósticos de eventos futuros de acuerdo con el comportamiento de ellas (Cardona, D. et. al, 2013).

3.1. Estimación del modelo

Con la información descrita anteriormente respecto al análisis exploratorio de los datos y sus características probabilísticas se plantea el siguiente modelo de regresión lineal múltiple para predecir la esperanza de vida en años.

$$\hat{y} = 47{,}614113 - 0{,}017949x_1 + 0{,}355162x_2 - 0{,}844894x_3 + 36{,}285086x_4 + \varepsilon$$

Donde:
x_1: Tasa de mortalidad de adultos de ambos sexos
x_2: Gasto del gobierno en salud como porcentaje del gasto público total
x_3: Muertes por cada mil nacidos vivos VIH/SIDA (0-4 años)
x_4: Índice de desarrollo humano en términos de composición de ingresos de los recursos

El resultado anterior nos permite constatar las conclusiones obtenidas por previas investigaciones que establecen la existencia de una relación lineal de la esperanza de vida con estas variables.

En la Tabla 3 se presentan los resultados obtenidos con la función lm del programa R.

Residuals:				
Min.	1st QU.	Median	3rd QU.	Max
-10.3808	-1.6174	-0.0501	1.6143	9.976
Coefficients:				
	Estimate	Std. Error	t value	Pr(|t|)
(Intercept)	47.6141	2.0476	23.253	2e-16***
Adult.Mortality	-0.0179	0.0038	-4.656	8.04e-06***
Total.expenditure	0.3551	0.114	3.187	0.001816**
HIV.AIDS	-0.8449	0.2300	-3.673	0.000353***
Income.composition	36.2850	2.4884	14.581	2e-16***
Residual standard	error:	3.102 on 126	degrees	of freedom
Multiple Rˆ2	0.8741	Adjusted	Rˆ2	0.8701
F-statistic:	218.7 on 4	and 126 DF	p-value:	2.2e-16

Cuadro 3: Resultados usando la función lm de R

Por un lado, se puede observar el p-value de la prueba de hipótesis que contrasta $H_0 = \beta_i = 0$ vs $H_a = \beta_i \neq 0$ asociado a cada uno de los coeficientes. Considerando un nivel de significancia $\alpha = 0{,}05$ en cada caso se rechaza la hipótesis nula y por lo tanto existe suficiente evidencia estadística para concluir que el coeficiente es distinto de cero, es decir su variable asociada es significativa para el modelo.

En este sentido, el parámetro constante y las variables tasa de mortalidad de adultos de ambos sexos, gasto del gobierno en salud como porcentaje del gasto público total, muertes por mil nacidos vivos VIH/SIDA (0-4 años) y el índice de desarrollo humano en términos de composición de ingresos de los recursos son significativas para el modelo lo cual es consistente con la literatura revisada.

Expresando este valor como un porcentaje, el error cuadrático se puede interpretar como el porcentaje de la variación de los valores de la variable dependiente que se puede explicar con la ecuación de regresión.

Esto revela que la ecuación de regresión explica en un 87.41 % los valores observados de la muestra de esperanza de vida del año 2014.

En la mayoría de situaciones prácticas no es común obtener coeficientes de determinación tan altos, pero existen valores aceptables que varían de acuerdo con la rama del conocimiento sobre el que se verse el estudio o investigación.

Por último, el p-value general asociado al estadístico F indica que el modelo en conjunto es significativo, es decir se rechaza la hipótesis nula que considera a todos los coeficientes iguales a cero y se concluye que al menos uno de los coeficientes del modelo es distinto de cero.

4. Validación de supuestos

Con el propósito de verificar que el modelo cumple con los supuestos del modelo de regresión lineal múltiple, a continuación se presenta el análisis realizado sobre los residuos estimados.

4.1. Normalidad

Debe cumplirse que los residuos estimados $\hat{\varepsilon_1}, ..., \hat{\varepsilon_{131}}$ siguen una distribución normal, en particular con media cero y varianza constante.

En la siguiente gráfica de diagnóstico se comparan los cuantiles de los residuos estimados con los cuantiles teóricos de la distribución normal. Se puede observar un buen ajuste de los cuantiles muestrales con algunas desviaciones en las colas, lo cual sugiere que los residuos siguen una distribución normal.

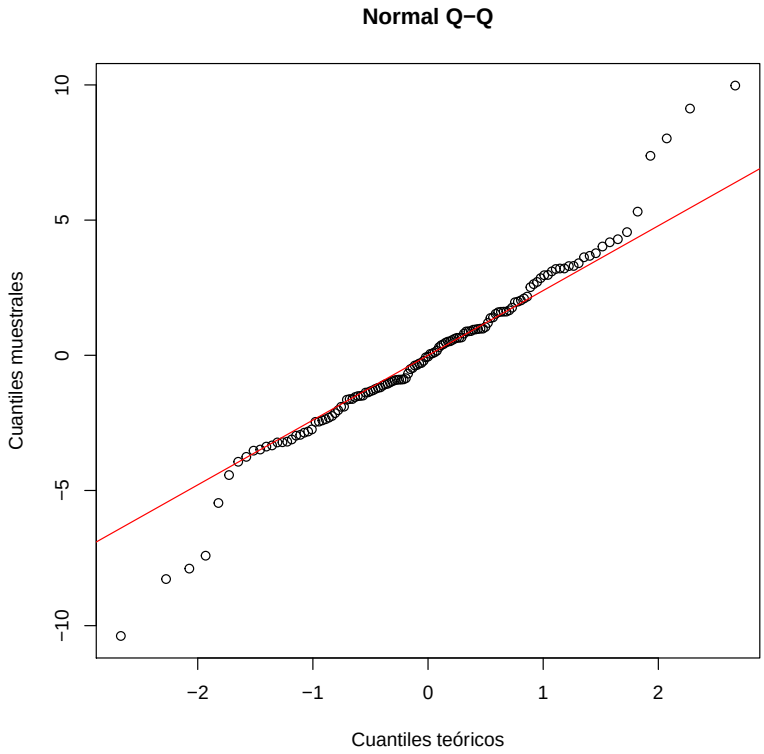

Figura 5: Normal Q-Q

Adicionalmente al realizar las pruebas de bondad de ajuste de Kolmogorov-Smirnov, Cramer-von Mises y Anderson-Darling se tiene que el p-value de cada prueba es mayor al nivel de significancia $\alpha = 0{,}05$ por lo que no se rechaza la hipótesis nula de que los residuos siguen una distribución normal y se concluye que no hay suficiente evidencia estadística para decir que los errores no son una muestra de la distribución normal.

En la siguiente tabla se presentan los p-value obtenidos de los contrastes de hipótesis realizados:

Prueba de bondad de ajuste	p-value
Kolmogorov-Smirnov	0.677
Cramer-von Mises	0.4923
Anderson-Darlin	0.3583

4.2. Homocedasticidad

Además se tiene que verificar que la varianza de los errores es constante.

La gráfica de los residuos estimados vs los valores ajustados (Figura 6) permite analizar este supuesto. En esta gráfica se puede apreciar que no existen patrones sino que los

residuos se distribuyen aleatoriamente a lo largo de los valores ajustados, lo que sugiere que se cumple la varianza constante de los errores (homocedasticidad); sin embargo **es necesario realizar una prueba de hipótesis**.

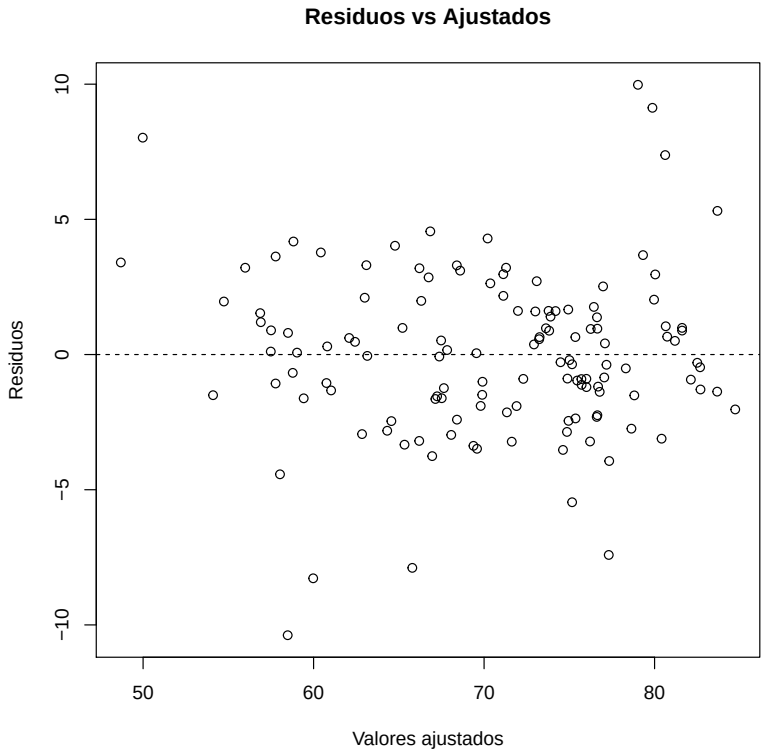

Figura 6: Residuos vs Ajustados

El p-value del test de Breusch-Pagan es 0.7817, dado que es superior al nivel de significancia $\alpha = 0{,}05$ no se rechaza la hipótesis nula de la homocedasticidad de los residuos y se concluye que no hay suficiente evidencia estadística para decir que la varianza de los residuos es no constante.

4.3. Independencia

A continuación se presenta la gráfica de los errores vs los índices de las observaciones, no se observa algún patrón por lo que podría considerarse que no existen indicios de una estructura de dependencia en los errores.

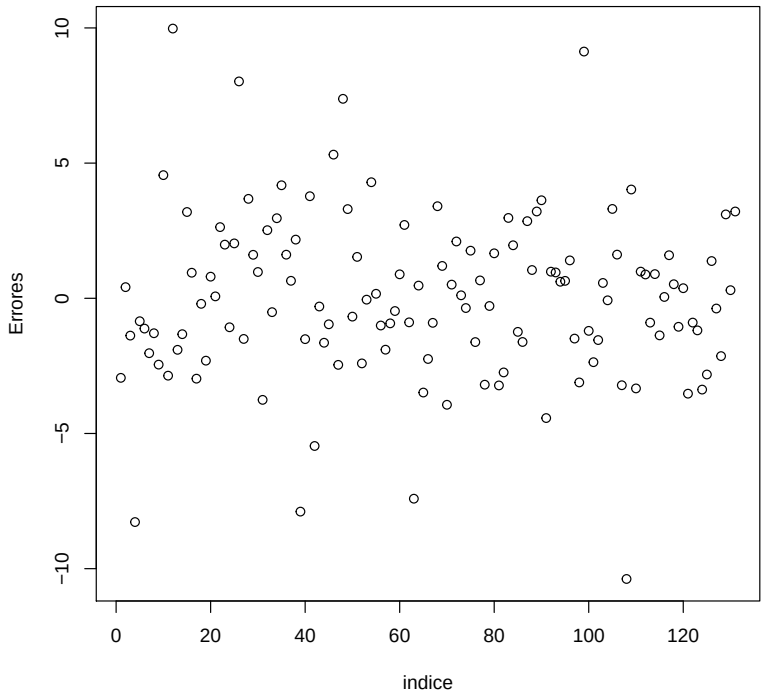

Figura 7: Errores vs Indices de observación

Al realizar la prueba de Durbin-Watson se tiene un p-value de 0.2954 el cual es mayor al nivel de significancia $\alpha = 0{,}05$ por lo que no se rechaza la hipótesis nula que considera que los errores no están autocorrelacionados y por lo tanto no hay suficiente evidencia estadística para concluir que los errores están autocorrelacionados.

De esta manera, como previamente se verificó que se cumple la normalidad de los errores se concluye que no podemos rechazar la independencia en los mismos. Cumpliendo el supuesto de independencia del modelo de regresión

4.4. Forma funcional errónea

Al ajustar el modelo lineal, la prueba de Ramsey arrojaba p-values demasiado pequeños lo cual implicaba que las variables explicativas necesitaban ser de grado mayor a 1, llegando hasta un grado 6 lo cual es un grado demasiado elevado. Al elevar las variables al cuadrado y al cubo el nuevo modelo arrojaba una R^2 de 91.6 % con lo cual se puede argumentar que el modelo se ajusta de mejor manera al original por una diferencia del 4 %. Este nuevo modelo cumple todas las pruebas descritas anteriormente a excepción la prueba de Breusch-Pagan lo cual implicaba que la varianza de los errores dejaba de ser constante,

por lo que se optó al tener un grado demasiado alto para las variables explicativas en la prueba de Ramsey así como una mejor calidad de ajuste (91.6 % que no se alejaba mucho del modelo pasado del 87.41 %), se decidió optar por el modelo original el cual cumple todos los supuestos de una regresión lineal.

5. Conclusiones

Por lo tanto, después de realizar las pruebas de bondad de ajuste de Anderson-Darling, Kolmogorov-Smirnov y Cramer-von Mises para verificar normalidad en los errores, la prueba de Breusch-Pagan para validar homocedasticidad de los errores y la prueba Durbin-Watson para verificar correlación nula en los residuos, se concluye que el modelo que explica a la variable esperanza de vida en años en términos de las covariables tasa de mortalidad de adultos de ambos sexos, gasto del gobierno en salud como porcentaje del gasto público total, muertes por mil nacidos vivos VIH/SIDA (0-4 años) e índice de desarrollo humano en términos de composición de ingresos de los recursos resulta un modelo adecuado ya que explica la variabilidad de la esperanza de vida en años en un 87.41 % (R^2) lo cual nos dice que es un modelo que se ajusta de manera adecuada a los datos de la muestra y valida los supuestos del modelo de regresión lineal múltiple.

Con nuevos datos al año anterior (2020) se podría contrastar si las variables explicativas han cambiado o inclusive si han incrementado para poder tener una mejor estimación de la esperanza de vida, de igual forma la calidad de ajuste del modelo no es mala pero sí podría mejorar.

6. Bibliografía

Referencias

[1] Barahona-Urbina, P. (2011). Factores determinantes de la esperanza de vida en Chile. Anales de la Facultad de Medicina, 72(4), 255-259.*http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1025-55832011000400006&lng=es&tlng=es*

[2] Cardona Madariaga, D. F., Rivera Lozano, M., Hernán Cárdenas, E., González Rodríguez, J. L. (2013, 1 octubre). Vista de Aplicación de la regresión lineal en un problema de pobreza | Interacción. Revistas Universidad Libre. *https://revistas.unilibre.edu.co/index.php/interaccion/article/view/2315/1767*

[3] Domínguez Alonso, E., Seuc, A. H. (2005). Esperanza de vida ajustada por algunas enfermedades crónicas no transmisibles. Revista Cubana de Higiene y Epidemiología, 43(2), 0-0.

[4] Lago, F. P., Geri, M., Moscoso, N. S., Monterubbianesi, P. D. Gasto total en salud y resultados. Universidad de Costa Rica, Ciencias Económicas, 31(2), 101-116

[5] Montero Granados. R (2016): Modelos de regresión lineal múltiple. Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España.

[6] Organización Mundial de la Salud (OMS). (2009). Organización Mundial de la Salud 2009. *http://www.who.int/whosis/whostat/2009/es/index.html*

[7] Revista Interacción Vol. 12. Octubre 2012-2013. págs. 73-84 Universidad Libre.Facultad de Ciencias de la Educación

[8] Roser, M., Ortiz-Ospina, E., Ritchie, H. (2013). Life expectancy. Our World in Data.

[9] Sanmartín-Durango, D., Henao-Bedoya, M. A., Valencia-Estupiñan, Y. T., Restrepo-Zea, J. H. (2019). Eficiencia del gasto en salud en la OCDE y ALC: un análisis envolvente de datos. Lecturas De Economía, (91), 41-78. *https://doi.org/10.17533/udea.le.n91a02*

[10] Shkolnikov, V. M., et. al. (2019). Patterns in the relationship between life expectancy and gross domestic product in Russia in 2005-15: a cross-sectional analysis. The Lancet. Public health, 4(4), 181–188. *https://doi.org/10.1016/S2468-2667(19)30036-2*

[11] Walpole, R. Myers, R. (1999). Probabilidad y estadística para ingenieros. (6a ed.). México: Prentice Hall.