



Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Estadística Aplicada

**MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN
FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA
EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA**

Ing. Ricardo Alberto Meza Santos

Asesorado por el Mtro. Luis Carlos Bolaños Mendez

Guatemala, mayo de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA



FACULTAD DE INGENIERÍA

**MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN
FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA
EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA**

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

RICARDO ALBERTO MEZA SANTOS

ASESORADO POR EL MTRO. LUIS CARLOS BOLAÑOS MENDEZ

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ESTADÍSTICA APLICADA

GUATEMALA, MAYO DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA



NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

NÓMINA DEL TRIBUNAL EXAMINADOR

DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Mtro. Ing. Edgar Darío Álvarez Cotí
EXAMINADOR	Mtro. Ing. Edwin Adalberto Bracamonte Orozco
EXAMINADOR	Mtro. Ing. William Eduardo Fagiani Cruz
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

**MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN
FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA
EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA**

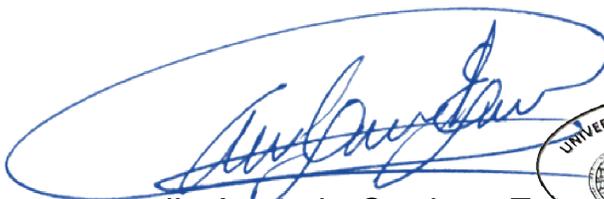
Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha 19 de julio 2021.

Ing. Ricardo Alberto Meza Santos

LNG.DECANATO.OI.326.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA**, presentado por: **Ricardo Alberto Meza Santos**, que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:



Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, mayo de 2022

AACE/gaoc



Guatemala, mayo de 2022

LNG.EEP.OI.326.2022

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

“MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA”

presentado por **Ricardo Alberto Meza Santos** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

“Id y Enseñad a Todos”

Mtro. Ing. Edgar Darío Álvarez Colí
Director

Escuela de Estudios de Postgrado
Facultad de Ingeniería





Guatemala 27 de febrero 2022.

M.A. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

M.A. Ingeniero Álvarez Cotí:

Por este medio informo que he revisado y aprobado el Informe Final del trabajo de graduación titulado **“MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA”** del estudiante **Ricardo Alberto Meza Santos** quien se identifica con número de carné **2531250570101** del programa de Maestría en Estadística Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014*. Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,

MSc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestría en Estadística Aplicada
Escuela de Estudios de Postgrado

Guatemala, 1 agosto de 2021.

M.A. Ing. Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

Estimado M.A. Ing. Álvarez Cotí

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: **“MODELO ESTADÍSTICO DE PREDICCIÓN DE LA DIMENSIÓN EDUCATIVA DEL IDH EN FUNCIÓN DE LA INVERSIÓN PORCENTUAL DEL PIB Y TASA DE COBERTURA EDUCATIVA NETA EN LOS DIFERENTES NIVELES EN GUATEMALA”** del estudiante **Ricardo Alberto Meza Santos** del programa de Maestría en **Estadística Aplicada**, identificada con número de carné: CUI **2531250570101**.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.



Luis Carlos Bolaños Méndez
Ing. electrónico
Col. 7653

Mtro. Ing. Luis Carlos Bolaños Méndez
Colegiado No. 7653
Asesor de Tesis

ACTO QUE DEDICO A:

- Dios** Por ser la fuente de fortaleza en mi vida y otorgarme la bendición de concluir este proceso.
- Mis padres** Juan Francisco Meza y Aura Margarita Santos, por su apoyo y guía a lo largo de toda mi vida.
- Mi hermana** Karen Janeth Meza, por ser parte de mi vida comprometiéndome a ser ejemplo continuo.
- Mi sobrino** Liam Joaquín Zacarias Meza, por ser una brisa de aire fresco y ser un recordatorio constante de lo fácil que es ser feliz.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser mi casa de estudios, y la fuente de formación profesional de que he alcanzado.
Facultad de Ingeniería	Por albergarme en sus pasillos y formarme en sus aulas, regalándome además muchos de mis amigos más entrañables.
Mis amigos de la maestría	Por ser los compañeros de desvelos y fines de semana interminables, que fueron fuente de apoyo para no claudicar, gracias, Julio Monterroso, Erick Marroquín y Andrea Corado.
Mi amigo	Luis Fernando Alvarado, quien además de su amistad, compartió conmigo la aventura de cursar la maestría.
Mi asesor	Luis Carlos Bolaños Méndez por compartir sus conocimientos y brindarme su orientación.
Mi Primo	El ingeniero Luis Manuel Ávila, a quien considero un hermano y amigo, que siempre me impulso para alcanzar esta y muchas metas.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES.....	V
LISTA DE SÍMBOLOS	IX
GLOSARIO	XI
RESUMEN.....	XIII
PLANTEAMIENTO DEL PROBLEMA.....	XV
OBJETIVOS.....	XIX
RESUMEN DEL MARCO METODOLÓGICO	XXI
INTRODUCCIÓN	XXV
1. MARCO REFERENCIAL.....	1
2. MARCO TEÓRICO.....	5
2.1. Modelos de regresión y predicciones	5
2.1.1. Supuestos estadísticos de los datos.....	6
2.1.1.1. Homocedasticidad	6
2.1.1.2. Homogeneidad	7
2.1.1.3. Independencia	7
2.1.1.4. Linealidad	7
2.1.1.5. Normalidad	8
2.1.2. Análisis de correlación.....	10
2.1.3. Análisis de regresión	11
2.1.4. Series temporales.....	14
2.2. índice de Desarrollo Humano	16
2.2.1. Programa de las Naciones Unidas para el Desarrollo	16

2.2.2.	Dimensiones de desarrollo	17
2.2.3.	Dimensión de educación	17
2.2.4.	Producto interno bruto	18
3.	PRESENTACION DE RESULTADOS.....	19
3.1.	Descripción de la población y la muestra	19
3.2.	Evaluación de la calidad de los datos	20
3.3.	Evaluación gráfica de los efectos	23
3.4.	Análisis de varianza de los modelos generales de cada país	25
3.5.	Elección de modelos según el coeficiente de determinación ...	27
3.6.	Elección de modelo bajo criterios de información y parsimonia.....	30
3.7.	Modelo seleccionado para Guatemala	32
3.7.1.	Supuestos estadísticos del modelo electo para Guatemala.....	34
3.7.1.1.	Supuesto de no multicolinealidad.....	34
3.7.1.2.	Supuesto de normalidad de residuos ...	35
3.7.1.3.	Supuesto de homocedasticidad	36
3.7.1.4.	Supuesto de independencia de residuos.....	38
3.8.	Modelo seleccionado para Bolivia	39
3.8.1.	Supuestos estadísticos del modelo electo para Bolivia.....	41
3.8.1.1.	Supuesto de no multicolinealidad.....	41
3.8.1.2.	Supuesto de normalidad de residuos ...	42
3.8.1.3.	Supuesto de homocedasticidad	43
3.8.1.4.	Supuesto de independencia de residuos.....	45

3.9.	Modelo seleccionado para México	46
3.9.1.	Supuestos estadísticos del modelo electo para México	48
3.9.1.1.	Supuesto de no multicolinealidad	48
3.9.1.2.	Supuesto de normalidad de residuos ..	49
3.9.1.3.	Supuesto de homocedasticidad.....	50
3.9.1.4.	Supuesto de independencia de residuos	52
3.10.	Modelo seleccionado para Perú	53
3.10.1.	Supuestos estadísticos del modelo electo para Perú.....	55
3.10.1.1.	Supuesto de no multicolinealidad	55
3.10.1.2.	Supuesto de normalidad de residuos ..	56
3.10.1.3.	Supuesto de homocedasticidad.....	57
3.10.1.4.	Supuesto de independencia de residuos	59
3.11.	Comparación de los modelos propuestos.....	60
3.11.1.	Modelos propuestos para cada país.....	61
3.11.2.	Gráficos de cajas para los modelos propuestos de los países.....	61
4.	DISCUSIÓN DE RESULTADOS	67
4.1.	Resultado del análisis preliminar de la información.....	67
4.2.	Resultado de la evaluación gráfica de los efectos.....	68
4.3.	Resultado del análisis de varianza de los modelos generales de cada país	69
4.4.	Resultados del proceso de modelización	70
4.4.1.	Depuración de los modelos basado en el coeficiente de determinación	70

4.4.2.	Depuración de los modelos basado en los criterios parsimonia y de información AICc y BIC....	71
4.5.	Validación de supuestos de los modelos y selección.....	72
4.5.1.	Resultado del supuesto de multicolinealidad.....	72
4.5.2.	Resultado del supuesto de normalidad	73
4.5.3.	Resultado del supuesto de homocedasticidad de residuos.....	73
4.5.4.	Resultado del supuesto de independencia de residuos.....	74
4.6.	Resultados de comparación entre modelos seleccionados de los distintos países	74
CONCLUSIONES.....		77
RECOMENDACIONES		79
REFERENCIAS.....		81

ÍNDICE DE ILUSTRACIONES

FIGURAS

1.	Descomposición de una serie de tiempo aditiva en: observaciones, tendencia, estacionalidad y aleatoriedad	15
2.	Gráfica de la prueba de normalidad para Guatemala	21
3.	Gráfica de la prueba de normalidad para Bolivia	21
4.	Gráfica de la prueba de normalidad para México.....	22
5.	Gráfica de la prueba de normalidad para Perú	22
6.	Gráfica de efectos sobre la variable dependiente para Guatemala.....	23
7.	Gráfica de efectos sobre la variable dependiente para Bolivia.....	24
8.	Gráfica de efectos sobre la variable dependiente para México.....	24
9.	Gráfica de efectos sobre la variable dependiente para Perú.....	24
10.	Diagrama de Pareto del modelo electo para Guatemala.....	34
11.	Gráfico de prueba de normalidad de residuos para Guatemala.....	36
12.	Gráfico de homocedasticidad de residuos para Guatemala.....	37
13.	Gráfico de residuos versus orden para el modelo de Guatemala	39
14.	Diagrama de Pareto del modelo electo para Bolivia	41
15.	Gráfico de prueba de normalidad de residuos para Bolivia.....	43
16.	Gráfico de homocedasticidad de residuos para Bolivia.....	44
17.	Gráfico de residuos versus orden para el modelo de Bolivia	46
18.	Diagrama de Pareto del modelo electo para México.....	48
19.	Gráfico de prueba de normalidad de residuos para México	50
20.	Gráfico de homocedasticidad de residuos México	51
21.	Gráfico de residuos versus orden para el modelo de México.....	53
22.	Diagrama de Pareto del modelo electo para Perú	55

23.	Gráfico de prueba de normalidad de residuos para Perú	57
24.	Gráfico de homocedasticidad de residuos para Perú	58
25.	Gráfico de residuos versus orden para el modelo de Perú	60
26.	Gráfico de cajas para Guatemala	62
27.	Gráfico de cajas para Bolivia	63
28.	Gráfico de cajas para México	64
29.	Gráfico de cajas para Perú	65

TABLAS

I.	Variables del estudio.....	XXII
II.	Comprobación del supuesto de normalidad para los distintos modelos ..	20
III.	Análisis de varianza para el modelo completo de Guatemala.....	26
IV.	Análisis de varianza para el modelo completo de Bolivia	26
V.	Análisis de varianza para el modelo completo de México.....	27
VI.	Análisis de varianza para el modelo completo de Perú	27
VII.	Resumen de los modelos de regresión evaluados para Guatemala.....	28
VIII.	Resumen de los modelos de regresión evaluados para Bolivia.....	28
IX.	Resumen de los modelos de regresión evaluados para México	29
X.	Resumen de los modelos de regresión evaluados para Perú.....	29
XI.	Mejor modelo para Guatemala.....	31
XII.	Mejor modelo para Bolivia.....	31
XIII.	Mejor modelo para México.....	31
XIV.	Mejor modelo para Perú.....	32
XV.	Análisis de varianza del modelo electo para Guatemala	33
XVI.	Coeficientes del modelo electo para Guatemala.....	33
XVII.	Coeficientes del modelo electo para Guatemala.....	35
XVIII.	Prueba de contraste Breusch-Pagan para Guatemala	37
XIX.	Prueba de contraste de Durbin-Watson para Guatemala	38

XX.	Análisis de varianza del modelo electo para Bolivia.....	40
XXI.	Coeficientes del modelo electo para Bolivia.....	40
XXII.	Coeficientes del modelo electo para Bolivia.....	42
XXIII.	Prueba de contraste Breusch-Pagan para Bolivia.....	44
XXIV.	Prueba de contraste de Durbin-Watson para Bolivia.....	45
XXV.	Análisis de varianza del modelo electo para México.....	47
XXVI.	Coeficientes del modelo electo para México.....	47
XXVII.	Coeficientes del modelo electo para México.....	49
XXVIII.	Prueba de contraste Breusch-Pagan para México.....	51
XXIX.	Prueba de contraste de Durbin-Watson para México.....	52
XXX.	Análisis de varianza del modelo electo para Perú.....	54
XXXI.	Coeficientes del modelo electo para Perú.....	54
XXXII.	Coeficientes del modelo electo para Perú.....	56
XXXIII.	Prueba de contraste Breusch-Pagan para Perú.....	58
XXXIV.	Prueba de contraste de Durbin-Watson para Perú.....	59
XXXV.	Modelos seleccionados para los distintos países.....	61

LISTA DE SÍMBOLOS

Símbolo	Significado
β_i	Coefficiente sub i de los modelos
gl	Grados de libertad
S^2	Varianza muestral
μ	Media poblacional
σ	Varianza poblacional

GLOSARIO

AICc	Criterio de la información Akaike
Aleatoriedad	Carencia de una tendencia o patrones reconocibles en una serie de datos.
BIC	Criterio de la información Bayesiano
Correlación	Relación o correspondencia entre dos o más variables.
Hipótesis nula	Supuesto que se desea validar o invalidar.
Homocedasticidad	Uniformidad de la varianza.
Homogeneidad	Igualdad en el comportamiento de todos los elementos que conforman un conjunto determinado de datos.
IDH	Índice de Desarrollo Humano
Independencia	Falta de relación entre variables, es decir, el comportamiento de una de ellas no se ve afectado por los valores que toma la otra.
Indicador	Valor estadístico de referencia al tamaño de parámetros o atributos.

PIB

Producto Interno Bruto

RESUMEN

El propósito de esta investigación fue proponer un modelo de predicción para la dimensión de educación del índice de desarrollo humano que sea confiable y que refleje las condiciones al momento del estudio del sistema de educación en Guatemala y contrastar dicho modelo como modelos similares de Bolivia, México y Perú.

El objetivo de esta investigación fue construir un modelo con el máximo nivel ajuste, para ello se aplicó análisis de regresión, y poder ofrecer una herramienta que sea capaz de pronosticar a corto y medio plazo la dimensión de educación del índice de desarrollo humano (IDH) en función de las variables propuestas, que se determinó fueron estadísticamente significativas.

La metodología utilizada en la investigación tuvo un enfoque cuantitativo, aplicando un diseño no experimental de tipo descriptivo que consistió en seleccionar el modelo que ofrece el máximo ajuste, se realizó seleccionando aquellos modelos cuyo coeficiente de determinación fueron los mayores, por su alta capacidad de representar el fenómeno, a ello se sumó el análisis de criterios de información Akaike (AICc) y Bayesiano (BIC), aplicados a cada modelo, también se aplicó el principio de parsimonia para que el modelo fuera estadísticamente representativo y también fuera el más simple.

Uno de los resultados más llamativos fue el efecto casi nulo de la inversión destinada a educación en los diferentes países, siendo la razón principal que el presupuesto está destinado en su mayor parte a cubrir gastos puramente

operativos, por un lado, por otra parte, también se explica por reducida variación con respecto al tiempo.

Se concluyó, contrastando los modelos de los diferentes países identificando las diferencias fundamentales entre cada sistema resaltando que la ventaja manifiesta de los demás países con respecto a Guatemala es el nivel de cobertura neta educativa en los diferentes niveles, como pieza principal para elevar la dimensión de educación del índice de desarrollo humano.

La principal recomendación es aumentar el nivel de cobertura de los diferentes niveles de manera efectiva, haciendo accesible la educación para la población por medio de infraestructura y recurso humano a través de la inversión específica en estos rubros, priorizando las áreas con mayor necesidad.

PLANTEAMIENTO DEL PROBLEMA

- Contexto general

El índice de desarrollo humano (IDH) es una forma aceptada, generalmente, de medir la calidad de vida de los habitantes de un país o región, este índice se construye tomando en cuenta tres dimensiones, las cuales son salud, riqueza y educación, siendo esta última, fundamental para alcanzar las primeras dos; esta información es recogida y presentada por el Programa de las Naciones Unidas para el Desarrollo conocido por sus siglas PNUD, y que según informes presentados por este programa actualmente le otorgan a Guatemala un índice que lo sitúa por debajo de la posición 128 de 192 países, ubicando a la nación en un nivel medio mundial, pero que es el más bajo de América Latina y de Centro América, aunque ha mostrado una tendencia a conservar dichas posiciones e índices en los últimos años.

El hecho de que Guatemala se haya estancado en su Índice de Desarrollo Económico es una muestra del estancamiento en materia de las políticas que favorecen el desarrollo de la sociedad como tal, a pesar de que el PNUD existe desde 1965 (con presencia en Guatemala desde 1975) y aporta esta información de carácter objetivo, se evidencia la poca eficiencia de las políticas aplicadas en la búsqueda del desarrollo integral de la sociedad.

- Descripción del problema

Guatemala no ha sido capaz de elevar sus índices de desarrollo humano en los últimos 5 años, y se sospecha que es debido a políticas poco eficientes en

el área de la educación, una de las 3 dimensiones para la construcción del índice, pero también como un elemento indirecto que afecta a las otras dos dimensiones, en el caso de salud donde no se dispone de personal capacitado para prestar estos servicios y en el caso del producto interno bruto (PIB) donde el aporte de su contribución individual es escaso en comparación con otros países donde se dispone de mano de obra especializada debido a su formación.

Se desconoce si existe algún grado significativo de correlación entre la proporción del IDH correspondiente a la dimensión de educación y el porcentaje de inversión, del PIB, que el estado hace al rubro de educación.

Se requiere un modelo que permita pronosticar la dimensión de educación en función de las variables generadoras como el porcentaje del PIB destinado a educación y que sean estadísticamente más significativas, con la finalidad de alcanzar los estándares educativos mínimos mundiales para Guatemala.

- Formulación del problema

Se determinaron las preguntas a responder relacionadas con el problema propuesto, y delimitó a un marco de tiempo y a los países incluidos en el estudio.

- Pregunta central

¿Cuál es el modelo con máximo ajuste que permita predecir el valor de la dimensión de educación del índice de desarrollo humano (IDH) en función de la inversión en educación como un porcentaje del PIB y el nivel de cobertura neta educativa en los diferentes niveles?

- Preguntas auxiliares
 - ¿Cuál es el nivel de correlación entre la dimensión educativa del índice de desarrollo humano (IDH) y la inversión en educación como un porcentaje del PIB y tasa de cobertura educativa neta en los diferentes niveles?
 - ¿Cuál es el modelo estadístico óptimo que describe mejor el comportamiento de la dimensión educativa del índice de desarrollo humano (IDH) cuando hay variaciones del presupuesto de educación en su forma porcentual y en las tasas de cobertura neta educativa de los diferentes niveles?
 - ¿Qué diferencias existen entre los modelos de inversión de educación de los países con valores en la dimensión de educación superiores al nacional y que tienen un contexto similar?

- Delimitación del problema

El problema se analizará con los datos históricos de inversión en educación en Guatemala, y los resultados de la dimensión de educación asociados a ellos de manera anual, en el periodo comprendido del 2004 al 2019, se contemplará la tasa de cobertura educativa neta.

La información puede ser recopilada del Ministerio De Educación (MINEDUC), Instituto Nacional de Estadística (INE), Ministerio de Finanzas Públicas (MINFIN) y del Programa de las Naciones Unidas para el Desarrollo (PNUD).

OBJETIVOS

General

Construir un modelo estadístico con nivel máximo de ajuste, por medio del análisis de regresión, que ofrezca una herramienta para la predicción de la dimensión educativa del índice de desarrollo humano (IDH) en función de la inversión en educación como un porcentaje del PIB y de las tasas de cobertura educativa neta en los diferentes niveles.

Específicos

1. Estimar el nivel de correlación entre la dimensión de educación del índice de desarrollo humano (IDH) en función de la inversión en educación como un porcentaje del PIB y de las tasas de cobertura educativa neta en los diferentes niveles, mediante el análisis de regresión, para aplicarlo como criterio de selección del modelo óptimo.
2. Seleccionar el mejor modelo estadístico, que ayude a predecir con mayor precisión la dimensión de educación, en función de la inversión realizada en forma porcentual del PIB, evaluando los diferentes modelos desarrollados, utilizando los criterios de información Akaike (AICc) y Bayesiano (BIC).
3. Comparar la inversión otorgada a educación en Guatemala con respecto a Bolivia, México y Perú, países con contextos similares al guatemalteco, por medio de modelos estadísticos de regresión que proyecten el comportamiento de la dimensión de educación en cada uno de ellos.

RESUMEN DEL MARCO METODOLÓGICO

- Características del estudio

El enfoque del estudio realizado tiene un enfoque cuantitativo, se determinó un valor porcentual, el cual es la dimensión de educación, una de las tres dimensiones que conforman el índice de desarrollo humano (IDH), que está planteado en función del porcentaje de producto interno bruto (PIB) asignado a dicho rubro y de las tasas de cobertura educativa neta de los diferentes niveles.

El diseño del estudio que se aplicó fue no experimental, pues se buscó construir un modelo de regresión estadístico tomando parte de la información socioeconómica de Guatemala y otros países latinoamericanos de contexto similar, específicamente en el área de educación, se analizó en su estado original sin ninguna manipulación; además es transversal pues se estudiaron datos de manera independiente para cada país involucrado en el estudio. Fue longitudinal, pues se analizó el comportamiento de los mismos países a lo largo del tiempo con el propósito de la construcción de un modelo predictivo, funcional a mediano plazo.

El tipo del estudio fue descriptivo, dado que se infirió el efecto de las variables explicativas: porcentaje de producto interno bruto (PIB) asignado a dicho rubro y las tasas de cobertura educativa neta en los diferentes niveles.

- Unidades de análisis

El objeto de estudio fue la dimensión de educación, del periodo comprendido del año 2004 al 2019 y datos socioeconómicos relacionados

directamente con ella descritos en la Tabla I, la cual se encuentra dividida en subpoblaciones dadas por los espacios geográficos específicos ordenados por países, entre ellos a Guatemala, Bolivia, México y Perú, de la cual se extrajeron los datos anuales del periodo ya mencionado en cada país, que serán estudiadas en su totalidad.

Tabla I. **Variabes del estudio**

Variable	Definición teórica	Definición operativa	Escala
Dimensión Educación (<i>I_EDU</i>)	Nivel de alcance logrado según los estándares educativos, del Programa de las Naciones Unidas para el Desarrollo.	Porcentual, adimensional, rango de 0 a 1	De razón
Porcentaje del PIB invertido en educación (<i>%PIB</i>)	Dato histórico de la inversión porcentual que ha asignado cada gobierno en el rubro de educación en los diferentes países del estudio.	Porcentual, adimensional, rango de 0 a 1	De razón
Tasa de cobertura educativa neta en preprimaria (<i>PRE</i>)	Datos recopilados de Ministerio De Educación y/o institutos de estadística de cada país sobre cantidad relativa de personas en edad nivel preprimaria.	Porcentual, adimensional, rango de 0 a 1	De razón
Tasa de cobertura educativa neta en primaria (<i>PRI</i>)	Datos recopilados de Ministerio De Educación y/o institutos de estadística de cada país sobre cantidad relativa de personas en edad nivel primaria.	Porcentual, adimensional, rango de 0 a 1	De razón
Tasa de cobertura educativa neta en secundaria (<i>SEC</i>)	Datos recopilados de Ministerio De Educación y/o institutos de estadística de cada país sobre cantidad relativa de personas en edad nivel secundaria.	Porcentual, adimensional, rango de 0 a 1	De razón

Fuente: elaboración propia.

- Fases del estudio

El estudio fue estructurado en 6 fases, las cuales se describen a continuación.

- Fase 1: Revisión bibliográfica

El análisis bibliográfico se realizó sobre los libros, revistas, tesis y artículos, consultados para elaborar la investigación. Son documentos con no más de 10 años de antigüedad, especializados en estadística, socioeconomía, y en el caso de las tesis solo de nivel de postgrado.

- Fase 2: Gestión y recolección de la información

Los datos utilizados corresponden al total de la población, que provino del informe anual generado por cada país en el Programa de las Naciones Unidas para el Desarrollo y de instituciones estatales oficiales.

- Fase 3: Análisis preliminar de la información

Los datos fueron analizados sometiéndolos a pruebas de multicolinealidad, aleatoriedad, normalidad, homogeneidad, entre otras, con lo que se determinó su comportamiento y bajo estos criterios se confirmó la viabilidad del tratamiento que se aplicó posteriormente a estos, en la construcción del modelo de regresión.

- Fase 4: Modelización

Se construyeron modelos a partir de los datos, dentro de los modelos que se propusieron, se aplicaron análisis para determinar el nivel de ajuste, de criterios de información y parsimonia, lo que ayudó a construir modelos de máxima verosimilitud, funcionales con datos composicionales para cada país.

- Fase 5: Validación y selección

Se evaluó cada modelo bajo los criterios de las siguientes pruebas de regresión de significancia de coeficientes, coeficientes de determinación ajustado y el cumplimiento de los supuestos de multicolinealidad, homocedasticidad de residuos, normalidad de residuos e independencia de residuos, con base en los resultados obtenidos se seleccionó el mejor modelo estadístico en base al nivel máximo de ajuste, al nivel de correlación.

- Fase 6: Redacción de informe final

El informe final detalló el proceso empleado con el que se dio respuesta a las preguntas que se desprendieron del problema y que se convirtieron en los objetivos del estudio, donde se describió las diferentes herramientas estadísticas que se emplearon, y las conclusiones que se obtuvieron a partir de ellas.

INTRODUCCIÓN

El presente trabajo es una innovación al determinar la dimensión en educación del índice de desarrollo humano en función de las variables de porcentaje de inversión y la cobertura neta educativa en los diferentes niveles. Se desarrolla utilizando técnicas de análisis multivariado determinando un modelo estadísticamente significativo para Guatemala y los diferentes países con los que se les compara.

El problema de investigación consiste en que se desconoce si existe algún grado significativo de correlación entre la proporción del IDH correspondiente a la dimensión de educación y el porcentaje de inversión del PIB entre otros indicadores. Por ello se requiere un modelo que permita pronosticar dicha razón y a la vez sea estadísticamente válido con la finalidad de alcanzar los estándares educativos mínimos mundiales para Guatemala.

En el resultado del estudio se determina un modelo de pronóstico con un nivel máximo de ajuste de la dimensión de educación, con ello se aporta al sistema educativo nacional una herramienta importante para la implementación de políticas de desarrollo, beneficiando a la sociedad guatemalteca en general.

La investigación emplea un enfoque cuantitativo, un diseño no experimental de tipo descriptivo, esquematizada en seis fases, siendo estas: la revisión bibliográfica, gestión y recolección de la información, análisis preliminar de la información, modelización, validación y selección, por último, redacción de informe final, estructurada en los capítulos que se describen a continuación.

El estudio es factible, ya que la información que requiere para realizar el análisis estadístico es de acceso público por medio de fuentes oficiales como el ministerio de educación, ministerio de finanzas públicas e instituto nacional de estadísticas y replicada en bases de datos de organismos internacionales como la CEPAL, UNESCO y UNICEF entre otras.

El capítulo uno presenta el análisis de investigaciones previas realizadas en áreas sociales y de educación, con ello se logró establecer los criterios para la elaboración de la investigación que se está presentando, con esta información fue posible determinar de manera inicial que modelos eran los más viables dada la naturaleza de los datos.

El capítulo dos contiene el marco teórico, con los fundamentos teóricos necesarios para la construcción de los modelos propuestos, estos fundamentos fueron recopilados de la literatura estadística especializada, desarrollada por diferentes autores, con esto se formó una concepción más amplia del tema, y de allí una mayor capacidad crítica de análisis de los resultados.

La investigación presenta los resultados obtenidos en el capítulo tres, donde se muestran los análisis previos, los modelos propuestos, el método de depuración aplicado, el modelo final seleccionado y la comparación entre países, concluyendo en el capítulo cuatro se discuten los resultados, explicando la naturaleza y comportamiento del modelo, aplicando un análisis interno y externo, dando pie a las conclusiones y recomendaciones que se presentan al final.

1. MARCO REFERENCIAL

La medida del desarrollo humano y de las variables que intervienen en su análisis se han reducido a índices o escalas porcentuales que, si bien son resultados continuos se ven influenciados por los efectos de tener cotas superiores e inferiores que tienden a frenar o amortiguar los modelos de regresión al aproximarse a ellos, si no se les da a los datos el tratamiento adecuado, dado que los porcentajes no son exclusivos de las áreas sociales, a continuación encontraremos diferentes áreas donde se plantearon estrategias diferentes en cuanto al manejo de la información que serán tomadas como base en la resolución del problema estadístico.

Elacqua y Martínez (BID, 2018) describen como en las últimas dos décadas en Latinoamérica, ha crecido la inversión en educación de 3.6 % a 5.3 % del PIB, mientras que en Guatemala en el mismo año se destinó el 2.93 % de su PIB para ello, destacando que en los países con mejores resultados se invierten anual mente 8000 US\$ por estudiante en comparación con Latinoamérica que invierte 2000 US\$ por estudiante, y sumado a ello la poco eficiente gestión de los recursos.

Liscano y Ortiz (2018) recomiendan modelos de regresión lineal y mixto cuando se está trabajando con datos composicionales, este tipo de dato se define como multivariado ya que el resultado de la suma es constante, y se utiliza por conveniencia, esta clase de datos se presentan como porcentajes, proporciones o concentraciones, se aplican por lo general en áreas de estudio como geología, arqueología, economía, ciencias políticas y ciencias forenses. El que los datos no puedan ser negativos y la suma sea constante implica que las técnicas

multivariantes habitualmente utilizadas no son adecuadas para su análisis y modelización, en este caso fue utilizado para proyectar la intención de voto en unas elecciones distritales en Colombia.

López-Roldan y Fachelli (2016) los modelos multivariados con datos composicionales habitualmente son empleados en estudios de investigación social, en estos casos es habitual el uso de transformaciones log-cociente, para la aplicación de regresión logística se trata de predecir una variable cualitativa o categórica, con la ventaja, frente al modelo de regresión clásico, de no tener que establecer la serie de condiciones de aplicación que dificultan su utilización y sus posibilidades, en particular, en el contexto de estudios por encuesta, en este caso relaciona el voto con el nivel educativo.

Enke, Graue y Mehdiyev (2011) presentan un sistema de predicción en tres etapas. Etapa 1, se realiza el análisis de regresión múltiple para definir las variables que tienen una fuerte relación con la salida. En la segunda fase se implementa la evolución diferencial con base en el *Fuzzy Clustering* (algoritmo de agrupamiento) tipo 2 para crear un modelo de predicción. En la tercera fase se utiliza una red *Fuzzy* neural (clasificación por peso) tipo-2 para realizar el razonamiento de la predicción. Este modelo a superado a los modelos clásicos de predicción financiera basados en índices.

Montero (2016) cuando la variable dependiente es un porcentaje, se puede estimar un modelo de probabilidad lineal, sobre todo si el modelo sólo tiene valores intermedios. Sin embargo, cuando los porcentajes están muy próximos a los extremos superior o inferior ya no se comportan igual cuando están en mitad de la tabla porque se “frenan”, se “tuercen” acotados los límites superior e inferior. Entonces hay que estimar un modelo de respuesta fraccional, que es similar a un *logit* o *probit*. Hay dos versiones similares *francreg* y *betareg* que se pueden

estimar como *logit*, *probit* *log-log*, entre otros. La única diferencia entre ambas es que la última no se puede ajustarse cuando se tienen 0 y 1 exactos en la variable dependiente.

Trucco (CEPAL, 2014) en el documento educación y desigualdad en América latina de la serie políticas sociales, destaca el papel fundamental de la educación en el desarrollo social, no limitando su importancia únicamente a los años de escolaridad, señalando que la calidad de esta potencializa en el aumento en el IDH en todas sus dimensiones, reafirmando el papel destacado de la educación y justificando su importancia.

Abuín (2007) indica que al momento elegir las variables que se consideren como generadoras, estas deben poseer un sentido numérico, se debe evitar variables que se repitan, deben cumplir con un sentido teórico asociado al problema y que entre la variable dependiente y las generadoras debe darse una relación de proporcionalidad.

Llaugel (2011) sugiere emplear el método de mínimos cuadrados para construir el modelo de regresión múltiple, permite un mejor ajuste entre variables dependiente y regresoras. Utilizar el coeficiente de determinación para la calidad del ajuste que ofrece la ecuación de regresión.

Walpole Myers y Myers, (2012) resaltan que el coeficiente de correlación únicamente mide la fuerza existente en un modelo lineal, mientras el coeficiente de determinación puede ser usado en relaciones no lineales y en relaciones con dos o más variables generadoras. Teniendo, el coeficiente de determinación una mayor utilidad.

“Cualesquiera conclusiones acerca de una relación causa y efecto deben basarse en los conocimientos de los especialistas en la aplicación de que se trate.” (Anderson, Sweeney y Williams, 2001, p. 565). Esto significa que la experiencia del analista y la profundidad de sus conocimientos sobre el tema son más importantes que propiamente el análisis de regresión que se realice.

En general se pueden destacar los siguientes elementos, una forma objetiva aceptada de poder comparar resultados es en forma de índices o porcentajes, indistintamente del tamaño de las poblaciones, si bien los modelos de regresión en consenso pueden ser aplicados, la limitante que da propiamente el resultado de la variable dependiente y de los propios índices utilizados para su predicción sufren un efecto de amortiguamiento o frenado, que determina la necesidad de un tratamiento especial, asociado a los datos composicionales, y modelos de regresión múltiple o multivariado, y en algunos aplicando métodos posteriores a la regresión que le den robustez a las predicciones finales.

2. MARCO TEÓRICO

A continuación, se establece las referencias teóricas estadísticas y sobre el campo de aplicación del estudio.

2.1. Modelos de regresión y predicciones

Llevar a cabo la construcción de un modelo de regresión y predicción confiable comienza con la calidad de los datos, en algunos casos simplemente son tomados y utilizados en la forma en la que se encuentran, y en otros casos requiere un tratamiento o transformación previa, que permita utilizar dicha información de la manera más adecuada, en este sentido la información debe ser sometida a ciertas consideraciones que deben tomarse en cuenta.

Construir un modelo estadístico que permita predecir resultados con un grado de confiabilidad aceptable entre al menos dos variables, una variable objetivo y una o más variables generadoras, requiere de una buena cantidad de herramientas que pueden ayudar con la operatoria pero que no pueden sustituir el análisis de un profesional de la información.

Dentro de los aspectos a considerar, previo a iniciar la construcción del modelo deseado se debe cumplir con un análisis de la información que habrá de ser considerada en un espacio de tiempo determinado, empleando la técnica de análisis de correlación la cual mide el nivel de asociación entre variables. Cuando los resultados del análisis de correlación confirman la viabilidad de los datos, se puede construir el modelo que tenga como fin predecir a futuro los valores en

función de las variables evaluadas previamente, este proceso lleva por nombre análisis de regresión, que medirá el nivel de confiabilidad de nuestro modelo.

2.1.1. Supuestos estadísticos de los datos

La calidad del modelo de regresión está en función de la calidad de la información o datos que se utilicen, en el caso de la construcción de un modelo es imprescindible garantizar la calidad de los datos y por consiguiente del modelo que se construya a partir de ellos, existe una serie de supuestos sobre el comportamiento de los datos y sus residuos, que deben ser cumplidos para este cometido.

La gran mayoría de autores destacan que los supuestos estadísticos que deben ser verificados previo a la construcción de un modelo son los siguiente: homocedasticidad, homogeneidad, independencia, linealidad y normalidad.

2.1.1.1. Homocedasticidad

Según Hair, Anderson, Tatham y Black (1999) los datos presentan homocedasticidad cuando se analiza la dispersión de la varianza de ellos y se puede identificar un comportamiento constante, la uniformidad de la varianza de los datos nos lleva a concluir que existe homocedasticidad en caso contrario heterocedasticidad.

Guàrdia, Freixa, Però, y Turbany (2008) indican que cuando se está seguro la muestra presentan un comportamiento normal, pueden ser aplicada la prueba F de Snedecor para determinar la igualdad de varianzas, que ofrece una gran potencia, pero es muy sensible a las desviaciones en la distribución normal, una segunda alternativa la prueba de *Bartlett*.

2.1.1.2. Homogeneidad

Gómez, Aparicio y Patiño (2010) indican que el supuesto de homogeneidad es observable cuando la media o la varianza de los datos es susceptible a los cambios, muchas veces presentado por un proceso inadecuado en la toma de datos o por la presencia de datos atípicos, sugiriendo para su comprobación la aplicación de la prueba *t-student* para evaluación de la homogeneidad.

2.1.1.3. Independencia

Díaz (2006) hace énfasis en que la independencia de los datos está fuertemente asociada a la aleatoriedad de estos, por lo que sugiere analizar el comportamiento del cambio signos en los residuos en busca de series largas de signos sin cambio ya sea positivos o negativos, de ser así, recomienda la aplicación de una prueba de rachas para determinar la independencia de los datos.

2.1.1.4. Linealidad

Hair et al. (1999) destacan que el supuesto de linealidad tiene implicación directa al realizar un análisis de regresión, por la dependencia que tiene de la correlación dada entre las variables que sean analizadas, a pesar de ello, no es posible representar los efectos no lineales. Si en caso se diera la no linealidad se hace necesario el análisis de todas las relaciones para determinar de qué manera influyen en la correlación. Con todo esto lo que se pretende es determinar si los datos se logran ajustar a una recta de regresión, para que el modelo pueda ser considerado válido.

Lind, D., Marchal, W. y Wathen, S. (2012) recomiendan que se verifique el cumplimiento del supuesto de linealidad, evaluando la correlación entre las variables, para demostrar la existencia de relación entre las mismas, por medio de la prueba de significancia del coeficiente de correlación.

La prueba de significancia compara un valor t calculado versus un valor t teórico donde un valor calculado mayor a un valor t teórico confirma la significancia del coeficiente de correlación y por consiguiente del supuesto de linealidad.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (\text{Ec.1})$$

2.1.1.5. Normalidad

Hair et al. (1999) destaca la importancia de la prueba del supuesto de normalidad, debido a que la mayor parte de las pruebas estadísticas paramétricas que se pueden realizar sobre los datos, son válidas cuando los datos se ajustan a una distribución normal.

Para evaluar el supuesto de normalidad se dispone de diversas pruebas y la adecuada selección de ellas, esto permite identificar cuando la distribución de los datos se ajusta a una distribución normal, como se puede apreciar a continuación.

Una forma visual de identificar la normalidad en el comportamiento de los datos o de sus residuos es por medio de un gráfico *Q-Q plot*, el cual muestra los datos observados versus los datos esperados, donde se espera que la dispersión de los puntos se encuentre en su gran mayoría alrededor de una recta a 45°

desde el origen, esto nos da indicios de que el supuesto se cumple, aunque vale resaltar que no es una prueba definitiva.

Para la comprobación del supuesto de normalidad, Guisande (2006) sugiere la utilización de la prueba de *Shapiro-Wilks* que por lo general se recomienda para muestras no mayores a 30 datos, donde se calcula un p-valor que debe ser comparado con el nivel de significancia deseado, que en caso de ser mayor se concluye que los datos o sus residuos tiene un comportamiento que se ajusta a una distribución normal, una prueba de normalidad alternativa a *Shapiro-Wilks* es la de *Ryan-Joiner*, ya que ambas se basan regresión y correlación.

$$R_p = \frac{\sum Y_i b_i}{\sqrt{s^2(n-1)\sum b_i^2}} \quad (\text{Ec.2})$$

Donde:

Y_i : Observaciones ordenadas de los datos

b_i : Ponderación normal de los datos ordenados

S^2 : Varianza muestras.

Según Marques (2001) cuando una serie o muestra posee más de 50 datos la prueba más recomendable para la comprobación del supuesto de normalidad es *Kolmogórov-Smirnov* aunque también es posible utilizar *Anderson-Darling* ya que ambas basadas en la distribución empírica. Donde el objetivo es encontrar un p-valor que sea mayor al nivel de significancia que se pretende, para concluir si se cumple con el supuesto.

Kolmogorov-Smirnov

$$D = \max \{D^+, D^-\}$$

Donde:

$$D^+ : \max_i \left\{ \frac{i}{n} - Z_{(i)} \right\}$$

$$D^- : \max_i \left\{ \left(Z_{(i)} - \frac{i-1}{n} \right) \right\}$$

$$Z : F(X_{(i)})$$

$F_{(i)}$: Función de distribución normal

$x_{(i)}$: Estadístico del i^{esimo} orden de una muestra aleatoria $1 \leq i \leq n$

n : Tamaño de la muestra

(Ec.3)

Anderson-Darling

$$A^2 = -N - \frac{1}{N} \sum (2i - 1)(\ln F(Y_i) + \ln(1 - F(Y_{N+1})))$$

$F(Y_i)$: Función de distribución normal acumulada

Y_i : Observaciones ordenas de los datos

(Ec.4)

2.1.2. Análisis de correlación

Lind et al. (2012) recomienda como paso inicial para evaluar la relación entre dos variables la construcción de diagramas de dispersión, con el fin de ofrecer una representación visual del comportamiento de estas, aunque al tratarse de una inspección visual no puede considerarse una prueba definitiva para identificar algún tipo de tendencia, pero sí como un indicio.

Después de elaborado el diagrama de dispersión, y haber identificado algún patrón o tendencia asociado a un modelo particular, se procede, de ser necesario, a realizar las transformaciones que ayuden a linealizar el

comportamiento de los datos, que se puede observar en un nuevo diagrama de dispersión, repitiendo este paso hasta que se esté satisfecho con el resultado.

Cuando los datos ya presentan esta tendencia lineal se puede proceder con un análisis cuantitativo de la relación entre los datos observados, para ello calcularemos el coeficiente de correlación tal y como se indica en la Ec.5.

Navidi (2006), ve el cálculo del coeficiente de correlación como una medida para cuantificar de manera porcentual que tan fuerte es la relación entre las variables, siendo un valor absoluto del resultado próximo a la unidad un indicador de una relación muy fuerte y lo opuesto cuando el resultado está próximo a cero.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right) \quad (\text{Ec.5})$$

Para el cálculo del coeficiente de correlación se utilizarán los valores transformados de haberse dado el caso, con ellos se calculará las medias de cada uno y sus respectivas desviaciones, con estos valores se alimenta la ecuación y calcular el valor r, para su interpretación.

2.1.3. Análisis de regresión

Walpole, Myers, Myers y Ye (2012), señalan que en todo modelo de regresión vamos a encontrar una variable dependiente o de respuesta y al menos una variable que será la variable independiente o regresora que tradicionalmente se representan como “Y” y “X” respectivamente, aunque esto no quiere decir que no se puedan representar bajo otra simbología.

Lind et al. (2012) simplifica la definición a una ecuación formada por al menos dos variables, una de ellas dependiente y el resto independientes, que refleja la relación gráfica y cuantificada entre las variables y que puede ser utilizada para realizar predicciones futuras.

La mayoría de los autores consideran el modelo básico de regresión, el que se presenta a continuación:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (\text{Ec.6})$$

Donde β_0 y β_1 representan el intercepto en el eje “y” y la pendiente respectivamente, por otra parte “ ε_i ” constituye el término aleatorio. Si el valor exacto del término aleatorio fuera conocido se podría conocer el valor de Y; pero, como solo puede estimarse, el modelo de regresión queda simplificado de la siguiente forma:

$$y = \beta_0 + \beta_1 x \quad (\text{Ec.7})$$

Donde β_0 y β_1 se obtiene a partir de las siguientes ecuaciones:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{Ec.8})$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (\text{Ec.9})$$

Webster (2000) señala que el objetivo del análisis de regresión es determinar la ecuación de la recta que mejor se ajuste a los datos reales

obtenidos a partir de una muestra. Para encontrar la ecuación de la recta se utiliza el método matemático conocido como mínimos cuadrados ordinario, en donde el resultado obtenido será los coeficientes β_0 y β_1 de la recta, y que esta se ajuste lo mejor posible a todos los valores de la muestra representativa.

Posada y Noguera (2007) destacan que el criterio de la información de Akaike (AIC) considera cambios en la bondad de ajuste y el número de parámetros existentes en el modelo que al ser evaluados en la ecuación AIC ofrecen un parámetro de la calidad del modelo y cuando se tiene varios modelos diseñados para realizar la misma predicción, el mejor modelo se determinara por el que ofrezca el menor valor.

Existentes dos variantes para el cálculo del criterio de la información Akaike, AIC, utilizado para muestras con más de 50 datos y AICc cuando el conjunto de datos es pequeño.

$$AIC = 2k - 2 \times \ln(L)$$

k: Número de parámetros del modelo.

ln(L): Función de log – verosimilitud.

(Ec.10)

$$AICc = AIC + \frac{2k(k + 1)}{N - k - 1}$$

N: El tamaño de la muestra de datos

(Ec.11)

Posada y Noguera (2007) indican la importancia del calcular el criterio de información Bayesiano (BIC), para los diferentes modelos propuestos como un indicador de la bondad de ajuste de la función de verosimilitud, la cantidad de

datos de la muestra y el número de parámetros con que cuenta el modelo, siendo la mejor opción el modelo con el BIC de menor valor.

$$BIC = -2 \times \ln(L) + \ln(N) \times k$$

k: Número de parámetros del modelo
N: El tamaño de la muestra de datos
ln(L): Función de log – verosimilitud

(Ec.12)

2.1.4. Series temporales

González (2009) señala que cuando se identifica una serie temporal, en primer lugar, existe un orden cronológico, no hay independencia entre las observaciones, y que incluso muchas de ellas manifestarán dependencia entre ellas, algo que debe ser muy tomado en cuenta en el proceso de análisis.

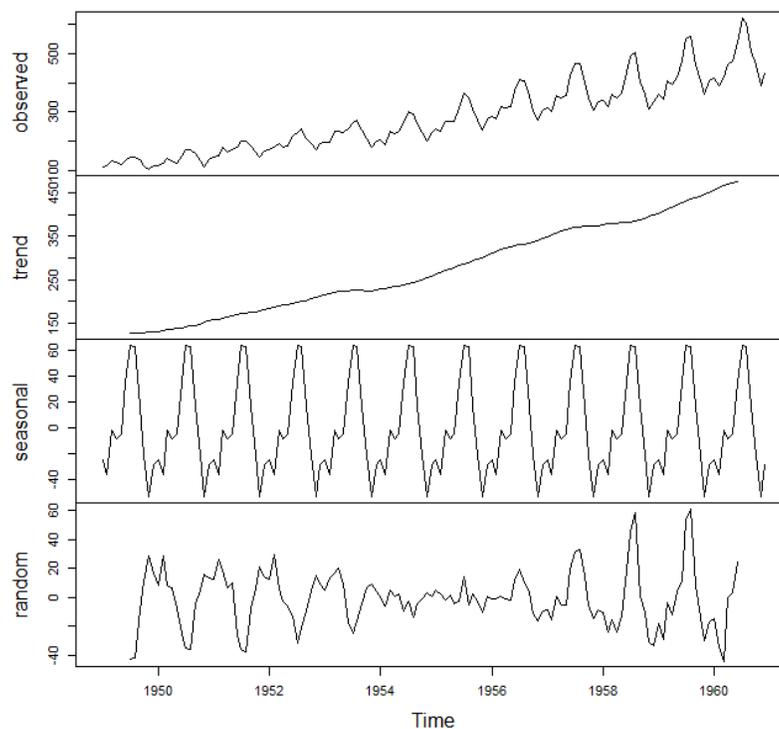
Debido a la naturaleza de las series temporales, los modelos desarrollados a partir de ellas pueden aprovechar la dependencia entre observaciones y a partir de ello proyectar el comportamiento futuro relativamente inmediato. Las técnicas desarrolladas para el análisis de datos o registros influenciados de manera significativa por el tiempo son un área de estudio a la cual se le conoce como análisis de series temporales.

Las series de tiempo pueden aplicarse en la mayoría de los campos de estudio, por ejemplo, en salud, educación, economía, entre otros se puede mencionar el producto interno bruto trimestral, las ventas anuales, los precios de los combustibles, entre otras. En meteorología se pueden encontrar series de lluvia anual, velocidad mensual del viento, temperaturas diarias, por mencionar algunas.

Lind et al. (2012), amplia la definición de series temporales, como un conjunto de registros ordenados de forma cronológica, y la periodicidad de su registro se puede realizar de forma anual, mensual, semanal, diaria o de cualquier manera en la que el tiempo pueda ser medido.

Dadas las condiciones que presentan las series de tiempo es necesario analizar otros aspectos como la estacionalidad, la tendencia, la aleatoriedad y el ajuste.

Figura 1. **Descomposición de una serie de tiempo aditiva en: observaciones, tendencia, estacionalidad y aleatoriedad**



Fuente: Escutia, I. (2019) *Descomposición de series de tiempo*. Consultado el 9 de noviembre de 2020. Recuperado de RPubS by RStudio: <https://rpubs.com/ltzelEscutia/546278>

Antúnez (2011) señala que la estacionalidad se puede entender como un factor que se repite en espacios de tiempo similares y que por lo general no suele

ser superior a 12 meses, sugiriendo aplicar la prueba de Dickey Fuller, para su determinación.

La tendencia, según González (2009), se puede presentar tanto lineal como no lineal, en el comportamiento general de los datos y muestra cómo se comporta la media a lo largo del tiempo.

Por último, Gómez et al. (2010) se refiere al ajuste de los datos con respecto a una función de distribución de probabilidad.

2.2. Índice de Desarrollo Humano

El índice de desarrollo humano (IDH) se creó para hacer hincapié en que la ampliación de las oportunidades de las personas debería ser el criterio más importante para evaluar los resultados en materia de desarrollo. El crecimiento económico es un medio que contribuye a ese proceso, pero no es un objetivo en sí mismo. (PNUD, 2016, pág. 1)

Dicho de forma sencilla el índice de desarrollo humano es un indicador porcentual que refleja que tanto se satisfacen los estándares básicos mundiales en lo referente a la calidad de vida en una sociedad como la guatemalteca.

2.2.1. Programa de las Naciones Unidas para el Desarrollo

El Programa de las Naciones Unidas para el Desarrollo (PNUD) fue establecido en 1965, en el 2010 el índice de desarrollo humano (IDH) sufrió un ajuste en el método utilizado para su cálculo y es el método que actualmente se utiliza, este programa está presente en muchos países del mundo y en particular

en Guatemala desde el año de 1975, para aportar información que oriente al desarrollo del país.

El Programa de las Naciones Unidas para el Desarrollo (PNUD) es la red mundial de la ONU para el desarrollo, que propugna el cambio y hace que los países tengan acceso al conocimiento, a la experiencia y a los recursos necesarios para ayudar a que las personas se labren un futuro mejor. El programa está presente en 177 países y territorios, y colabora con gobiernos y ciudadanos para que den con sus propias soluciones frente a los desafíos que plantea el desarrollo nacional y mundial. De este modo, a medida que desarrollan su capacidad local, los países se benefician del personal del PNUD y de su amplia variedad de asociados para obtener resultados. (ONUSIDA, 2015, página 2)

2.2.2. Dimensiones de desarrollo

Como se menciona en el Informe nacional de desarrollo humano Guatemala, (s.f.) este se descompone en tres dimensiones básicas las cuales son disfrutar de una vida larga y saludable (salud), acceso a educación (educación) y nivel de vida digno (nivel de vida).

2.2.3. Dimensión de educación

Como se indica en el Informe nacional de desarrollo humano Guatemala, (s.f.) la dimensión de educación es un indicador compuesto que toma en cuenta aspectos como la esperanza educativa en niños y la escolaridad obtenida por los adultos mayores de 25 años.

Este indicador mide el porcentaje de alcance en educación obtenido por un país con respecto a un estándar básico a nivel mundial, y esta dimensión aporta el 33 % del peso del índice de desarrollo humano que puede ser alcanzado por una sociedad o país.

2.2.4. Producto interno bruto

“El producto interno bruto (PIB) es un indicador económico que refleja el valor monetario de todos los bienes y servicios finales producidos por un país o región en un determinado periodo de tiempo, normalmente un año. Se utiliza para medir la riqueza que genera un país. También se conoce como producto bruto interno (PBI)” (Sevilla, 2020).

3. PRESENTACION DE RESULTADOS

A continuación, se presentan los resultados obtenidos en función del cumplimiento de los objetivos planteados en y que se detallan a continuación.

Objetivo 1. Estimar el nivel de correlación entre la dimensión de educación del índice de desarrollo humano (IDH) en función de la inversión en educación como un porcentaje del PIB, nivel de cobertura neta educativa en los niveles de preprimaria, primaria y secundaria, mediante el análisis de regresión, para aplicarlo como criterio de selección del modelo óptimo.

3.1. Descripción de la población y la muestra

El método por el cual se calcula la dimensión en educación (I_EDU) es utilizado oficialmente desde el 2010 aunque es posible encontrar algunos datos anteriores a esta fecha, para la construcción de los modelos se utilizaron:

- Guatemala, los datos poblacionales abarcan desde el 2004 al 2019, siendo un total de 16 datos.
- Bolivia, los datos poblacionales desde el 2005 al 2009, 2011 y del 2013 al 2018, siendo un total de 12 datos.
- México, los datos poblacionales desde el 2004 al 2017, siendo un total de 14 datos.
- Perú, los datos poblacionales desde el 2004 hasta el 2019, siendo un total de 16 datos.

3.2. Evaluación de la calidad de los datos

Se comprobó el supuesto de normalidad de los datos de las variables dependientes en cada modelo de regresión propuesto para Guatemala, Bolivia, México y Perú, sobre los datos se aplicó la prueba de normalidad de *Ryan-Joiner* proporcionada por la aplicación Minitab® 19 y que es similar a *Shapiro-Wilk*, recomendadas para muestras pequeñas, por medio de las siguientes hipótesis:

H_0 : Los datos siguen una distribución normal

H_a : Los datos no siguen una distribución normal

Bajo el criterio, de aceptar H_0 si el p-valor es mayor al nivel de significancia α que en este caso se tomó 0.05.

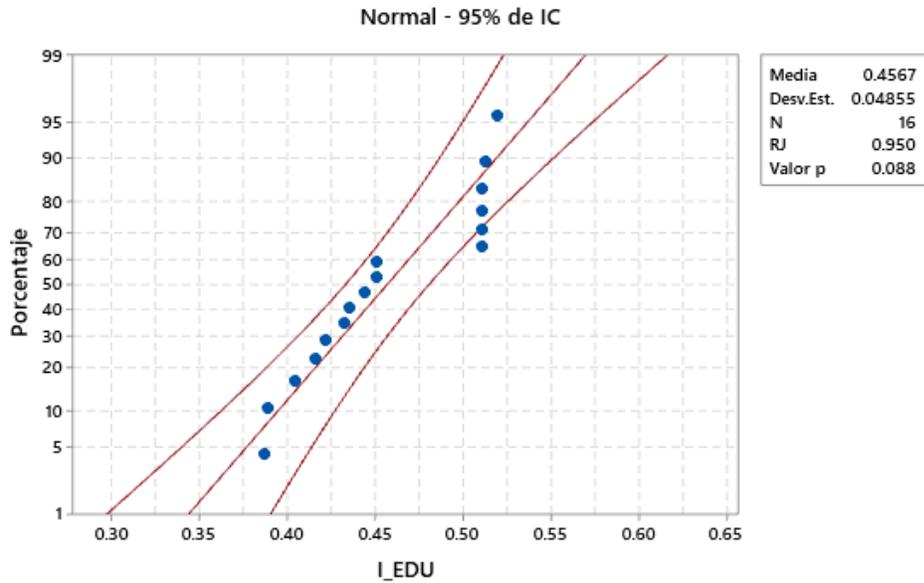
En la Tabla II se muestra el resultado de la evaluación obtenido por la prueba de normalidad de la variable de respuesta en cada uno de los modelos propuesto para cada país, determinando cierto nivel de ajuste con respecto a una distribución normal.

Tabla II. **Comprobación del supuesto de normalidad para los distintos modelos**

País	Variable	R	p-valor	Conclusión
Guatemala	I_EDU	0.950	0.088	No se rechaza H_0
Bolivia	I_EDU	0.970	0.100	No se rechaza H_0
México	I_EDU	0.988	0.100	No se rechaza H_0
Perú	I_EDU	0.961	0.100	No se rechaza H_0

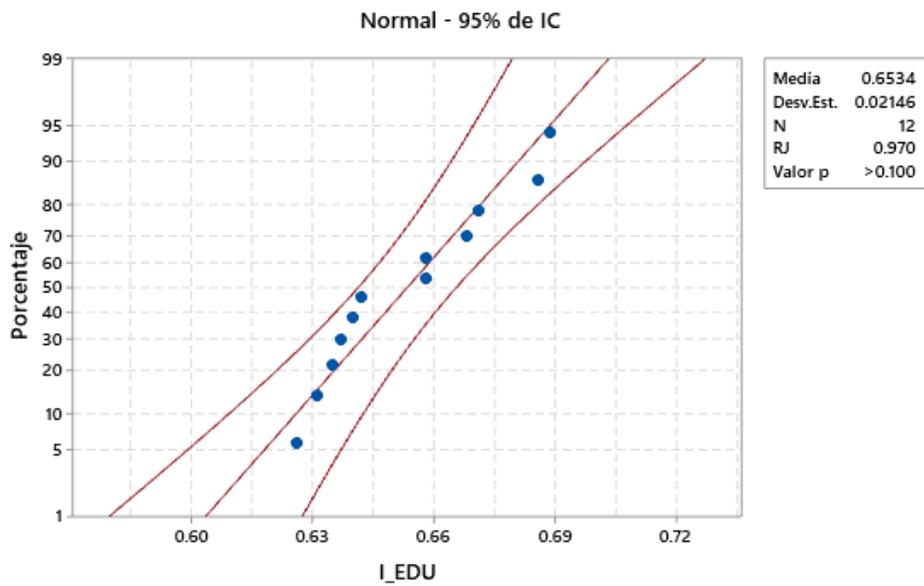
Fuente: elaboración propia.

Figura 2. **Gráfica de la prueba de normalidad para Guatemala**



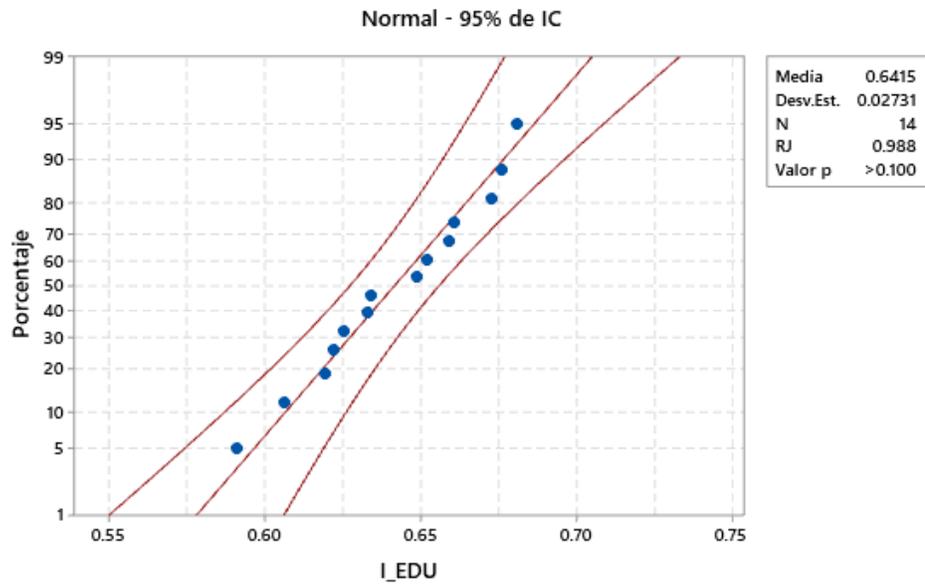
Fuente: elaboración propia.

Figura 3. **Gráfica de la prueba de normalidad para Bolivia**



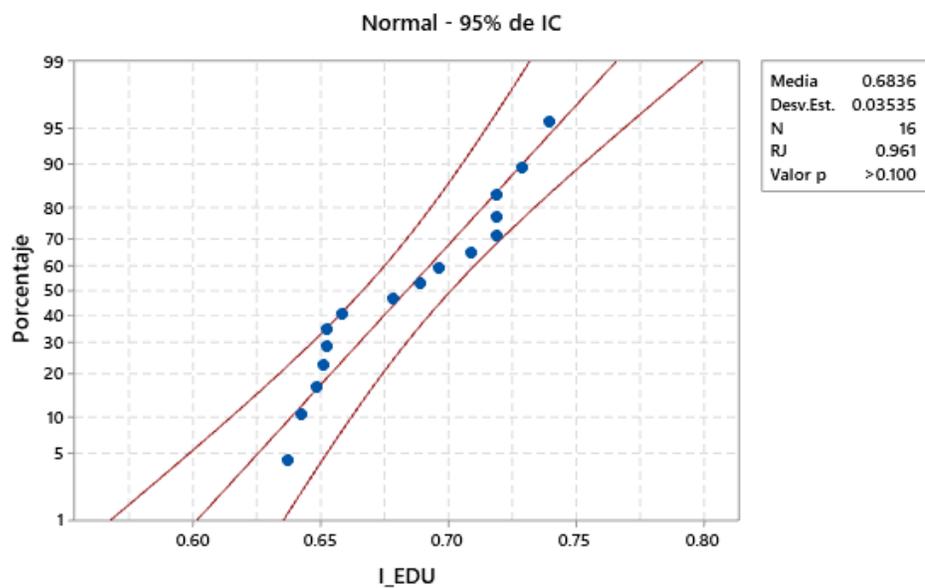
Fuente: elaboración propia.

Figura 4. **Gráfica de la prueba de normalidad para México**



Fuente: elaboración propia.

Figura 5. **Gráfica de la prueba de normalidad para Perú**



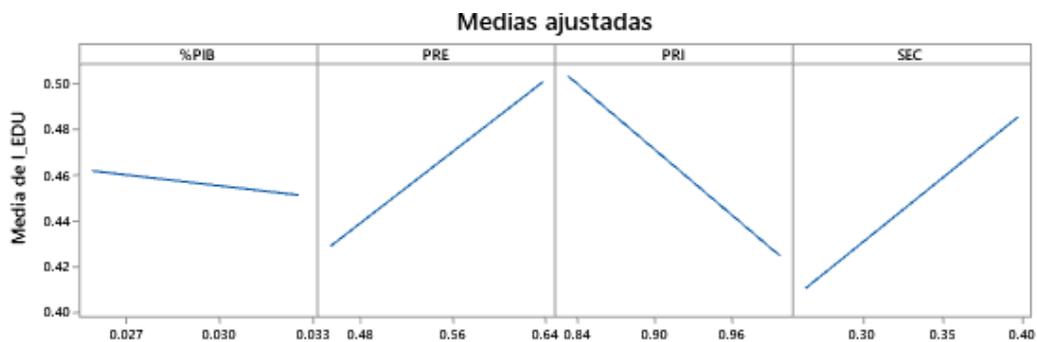
Fuente: elaboración propia.

Los gráficos anteriores, que se observan en las Figuras uno, dos, tres y cuatro Corresponden a los países de Guatemala, Bolivia, México y Perú respectivamente, en las cuales se ilustra el comportamiento Normal de la variable de respuesta I_EDU, sin ningún tipo de transformación respaldando los resultados que se muestran en la Tabla I.

3.3. Evaluación gráfica de los efectos

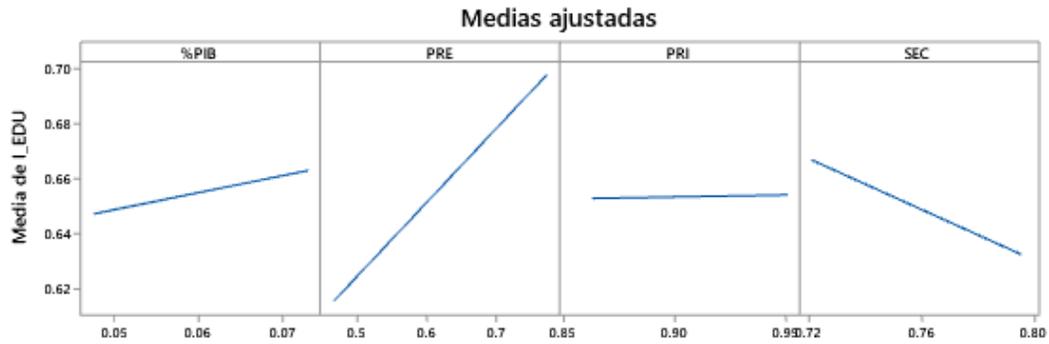
Se evaluó gráficamente los efectos de las variables predictoras propuestas sobre la variable dependiente obteniendo los siguientes resultados:

Figura 6. **Gráfica de efectos sobre la variable dependiente para Guatemala**



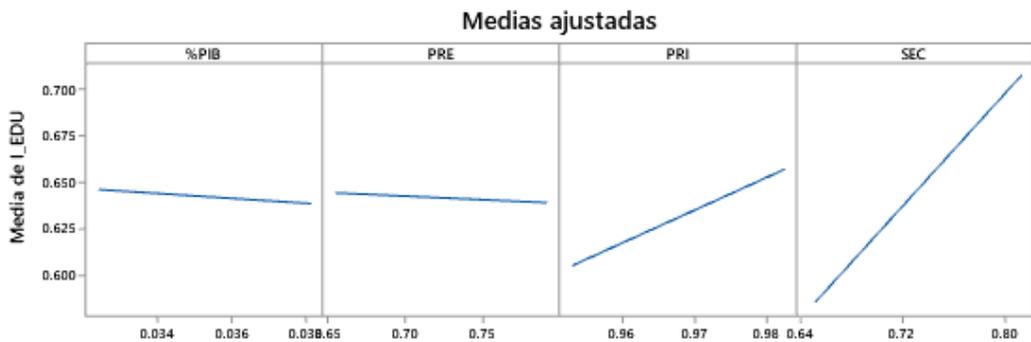
Fuente: elaboración propia.

Figura 7. **Gráfica de efectos sobre la variable dependiente para Bolivia**



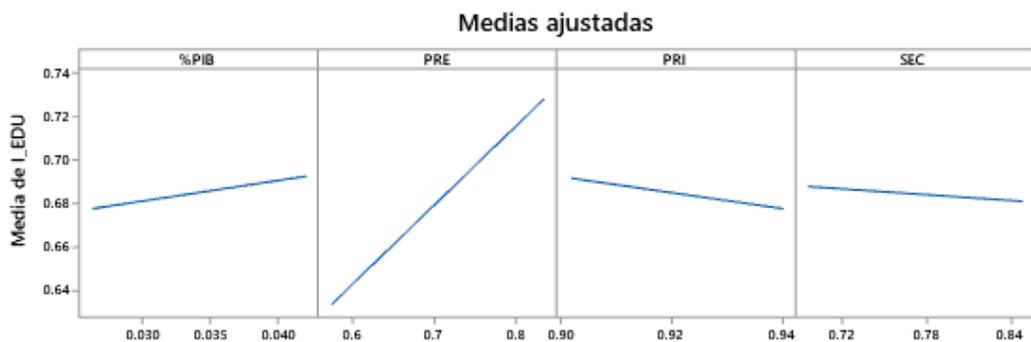
Fuente: elaboración propia.

Figura 8. **Gráfica de efectos sobre la variable dependiente para México**



Fuente: elaboración propia.

Figura 9. **Gráfica de efectos sobre la variable dependiente para Perú**



Fuente: elaboración propia.

El análisis gráfico ofreció indicios de las variables que son mayor fuente de variabilidad y por tanto determinar si deben o no deben ser incluidas en el modelo, en general los efectos con pendientes más pronunciadas son candidatos para formar parte del modelo después de haber cumplido con los supuestos estadísticos.

3.4. Análisis de varianza de los modelos generales de cada país

Se evaluó gráficamente los efectos de las variables predictoras propuestas sobre la variable dependiente obteniendo los siguientes resultados:

Se realizó un análisis de varianza multivariado, incluyendo todas las variables propuestas y a partir de ese modelo general se inició la depuración que llevo al modelo definitivo mediante el contraste de cada una de las variables predictoras bajo la siguiente prueba de hipótesis:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

Bajo el criterio de rechazar H_0 si el p-valor es menor al nivel de significancia α que en este caso se tomó 0.05.

Los términos que no tienen una asociación estadísticamente significativa para sus respectivos modelos se encuentran resaltados con negrilla en las tablas III, IV, V y VI que se presentan a continuación.

Tabla III. **Análisis de varianza para el modelo completo de Guatemala**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	4	0.033525	94.82 %	0.033525	0.008381	50.33	0.000
%PIB	1	0.011317	32.01 %	0.000059	0.000059	0.36	0.562
PRE	1	0.006566	18.57 %	0.002348	0.002348	14.10	0.003
PRI	1	0.014194	40.14 %	0.001675	0.001675	10.06	0.009
SEC	1	0.001448	4.10 %	0.001448	0.001448	8.70	0.013
Error	11	0.001832	5.18 %	0.001832	0.000167		
Total	15	0.035357	100.00 %				

Fuente: elaboración propia.

Tabla IV. **Análisis de varianza para el modelo completo de Bolivia**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	4	0.004646	91.72 %	0.004646	0.001161	19.39	0.001
%PIB	1	0.000010	0.21 %	0.000055	0.000055	0.92	0.370
PRE	1	0.004554	89.91 %	0.000699	0.000699	11.67	0.011
PRI	1	0.000000	0.01 %	0.000001	0.000001	0.01	0.909
SEC	1	0.000081	1.60 %	0.000081	0.000081	1.35	0.283
Error	7	0.000419	8.28 %	0.000419	0.000060		
Total	11	0.005065	100.00 %				

Fuente: elaboración propia.

Tabla V. **Análisis de varianza para el modelo completo de México**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	4	0.009000	92.85 %	0.009000	0.002250	29.20	0.000
%PIB	1	0.000279	2.88 %	0.000025	0.000025	0.33	0.582
PRE	1	0.000336	3.46 %	0.000020	0.000020	0.25	0.627
PRI	1	0.006522	67.28 %	0.000310	0.000310	4.02	0.076
SEC	1	0.001863	19.22 %	0.001863	0.001863	24.18	0.001
Error	9	0.000694	7.15 %	0.000694	0.000077		
Total	13	0.009694	100.00 %				

Fuente: elaboración propia.

Tabla VI. **Análisis de varianza para el modelo completo de Perú**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	4	0.017698	94.41 %	0.017698	0.004424	46.45	0.000
%PIB	1	0.015367	81.97 %	0.000053	0.000053	0.56	0.470
PRE	1	0.002092	11.16 %	0.000362	0.000362	3.80	0.077
PRI	1	0.000236	1.26 %	0.000211	0.000211	2.21	0.165
SEC	1	0.000003	0.02 %	0.000003	0.000003	0.03	0.866
Error	11	0.001048	5.59 %	0.001048	0.000095		
Total	15	0.018746	100.00 %				

Fuente: elaboración propia.

3.5. Elección de modelos según el coeficiente de determinación

El proceso de depuración requirió la construcción varios modelos hasta encontrar los mejores de ellos, en las tablas VII, VIII, IX y X se presenta un resumen de los resultados que se obtuvieron para cada país al utilizar diferentes modelos, indicando el número de variables que incluyó el modelo, el coeficiente de determinación, el coeficiente de determinación ajustado y las variables que

conformaran el modelo, resaltando en negrilla los que presentaron el mejor nivel de correlación.

Tabla VII. **Resumen de los modelos de regresión evaluados para Guatemala**

No.	Vars	R-cuad.	R-cuad. (ajust)	%PIB	PRE	PRI	SEC
1	1	87.3	86.4				X
2	1	45.5	41.6		X		
3	2	90.0	88.5		X	X	
4	2	88.6	86.9		X		X
5	3	94.7	93.3		X	X	X
6	3	90.7	88.4	X	X	X	
7	4	94.8	92.9	X	X	X	X

Fuente: elaboración propia.

Tabla VIII. **Resumen de los modelos de regresión evaluados para Bolivia**

No.	Vars	R-cuad.	R-cuad. (ajust)	%PIB	PRE	PRI	SEC
1	1	89.7	88.6		X		
2	1	72.9	70.2				X
3	2	90.1	87.9	X	X		
4	2	90.0	87.8		X	X	
5	3	91.7	88.6	X	X		X
6	3	90.6	87.1		X	X	X
7	4	91.7	87.0	X	X	X	X

Fuente: elaboración propia.

Tabla IX. **Resumen de los modelos de regresión evaluados para México**

No.	Vars	R-cuad.	R-cuad. (ajust)	%PIB	PRE	PRI	SEC
1	1	88.0	87.0				X
2	1	54.8	51.1			X	
3	2	92.4	91.0			X	X
4	2	89.6	87.7	X			X
5	3	92.6	90.4	X		X	X
6	3	92.6	90.4		X	X	X
7	4	92.8	89.7	X	X	X	X

Fuente: elaboración propia.

Tabla X. **Resumen de los modelos de regresión evaluados para Perú**

No.	Vars	R-cuad.	R-cuad. (ajust)	%PIB	PRE	PRI	SEC
1	1	91.3	90.7		X		
2	1	82.0	80.7	X			
3	2	93.9	93.0		X	X	
4	2	93.1	92.1	X	X		
5	3	94.4	93.0	X	X	X	
6	3	94.1	92.7		X	X	X
7	4	94.4	92.4	X	X	X	X

Fuente: elaboración propia.

Como criterio para determinar los modelos con mejor correlación se eligió en base al coeficiente de determinación ajustado que es mayor en cada país, para su identificación a los modelos propuestos en cada país se les asigno un número, en el caso de Guatemala el mejor modelo propuesto es el No. 5, Para Bolivia el modelo No.1 y el 5, México se representa mejor por medio del modelo No. 3, mientras que en el caso de Perú los mejores modelos propuesto corresponden a los No. 3 y 5.

Objetivo 2. Seleccionar el mejor modelo estadístico, que ayude a predecir con mayor precisión la dimensión de educación, en función de la inversión realizada en forma porcentual del PIB, evaluando los diferentes modelos desarrollados, utilizando los criterios de información Akaike (AICc) y Bayesiano (BIC).

3.6. Elección de modelo bajo criterios de información y parsimonia

Las tablas XI, XII, XIII y XIV muestran como los coeficientes de determinación redujeron la cantidad de modelos para cada país, en el caso de Guatemala y México a un modelo, y en el caso de Bolivia y Perú dos modelos.

Para la selección del mejor modelo asociado a cada país se aplicaron los siguientes análisis:

- Criterios de información de cada modelo AICc y BIC.
- Principio de parsimonia Interpretando *Cp de Mallows*.

Se utilizó el criterio de información por medio de los índices AICc y BIC, los cuales fueron calculados para cada modelo y se compararon entre ellos, la regla de decisión indica que mientras menor sea el resultado de los índices, el modelo será mejor, en los modelos de Guatemala, Bolivia y México los criterios de información dieron como resultado el mismo modelo, mientras que en el caso de Perú sirvió para determinar que el modelo 3 como el mejor en este caso.

El principio de parsimonia establece que cuando se tiene varios modelos con capacidades similares de explicar el mismo fenómeno la mejor opción siempre es la más sencilla, y para ello se calcula el estadístico *Cp de Mallows*, el cual se debe comparar con el número de parámetros que posee el modelo

incluyendo el termino independiente p siendo este el número de variables sin incluir el termino constante, estableciendo que el valor C_p debe ser menor a p más uno, lo cual indicara que es el modelo más simple con el cual se puede explicar a la variable dependiente.

$$C_p < p + 1$$

Tabla XI. **Mejor modelo para Guatemala**

No.	Vars (p)	R-cuad. (ajust)	Cp de Mallows	AICc	BIC	%PIB	PRE	PRI	SEC
5	3	93.3	3.4	-83.285	-85.422		X	X	X

Fuente: elaboración propia.

Tabla XII. **Mejor modelo para Bolivia**

No.	Vars (p)	R-cuad. (ajust)	Cp de Mallows	AICc	BIC	%PIB	PRE	PRI	SEC
1	1	88.6	0.7	-77.441	-78.986		X		
5	3	88.6	3.0	-69.066	-76.642	X	X		X

Fuente: elaboración propia.

Tabla XIII. **Mejor modelo para México**

No.	Vars (p)	R-cuad. (ajust)	Cp de Mallows	AICc	BIC	%PIB	PRE	PRI	SEC
3	2	91.0	1.5	-85.807	-87.695			X	X

Fuente: elaboración propia.

Tabla XIV. **Mejor modelo para Perú**

No.	Vars (p)	R-cuad. (ajust)	Cp de Mallows	AICc	BIC	%PIB	PRE	PRI	SEC
3	2	93.0	1.9	-95.785	-96.331		X	X	
5	3	93.0	3.0	-92.689	-94.827	X	X	X	

Fuente: elaboración propia.

Después de haber aplicado los criterios de información y parsimonia se confirmó en el caso de Guatemala y Bolivia, los modelos seleccionados bajo el criterio de mejor coeficiente de determinación, en cuanto a México y Perú, cada uno con dos posibles modelos igual de buenos según el coeficiente de determinación, los criterios de información y parsimonia facilitaron el tomar la decisión por un modelo en particular.

3.7. **Modelo seleccionado para Guatemala**

El modelo de pronóstico propuesto para Guatemala incluye como variables predictoras por su nivel de significancia estadística a preprimaria, primaria y secundaria, por consiguiente, los coeficientes asociados a dichas variables comparten el mismo nivel de significancia además de agregar un término constante.

Para su validación se aplicó la siguiente prueba de contraste tanto a los términos como a los coeficientes y el termino constante, como se puede observar en las tablas XV y XVI, correspondientes al análisis de varianza y análisis de coeficientes respectivamente.

La prueba de contraste aplicado a las variables predictoras, coeficientes y termino constante, bajo la siguiente prueba de hipótesis:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

Bajo el criterio de rechazar H_0 sí el p-valor es menor al nivel de significancia α que en este caso se tomó 0.05.

Tabla XV. **Análisis de varianza del modelo electo para Guatemala**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
<i>Regresión</i>	3	0.033466	94.65 %	0.033466	0.011155	70.79	0.000
<i>PRE</i>	1	0.016092	45.51 %	0.002438	0.002438	15.47	0.002
<i>PRI</i>	1	0.015737	44.51 %	0.002124	0.002124	13.48	0.003
<i>SEC</i>	1	0.001637	4.63 %	0.001637	0.001637	10.39	0.007
<i>Error</i>	12	0.001891	5.35 %	0.001891	0.000158		
<i>Total</i>	15	0.035357	100.00 %				

Fuente: elaboración propia.

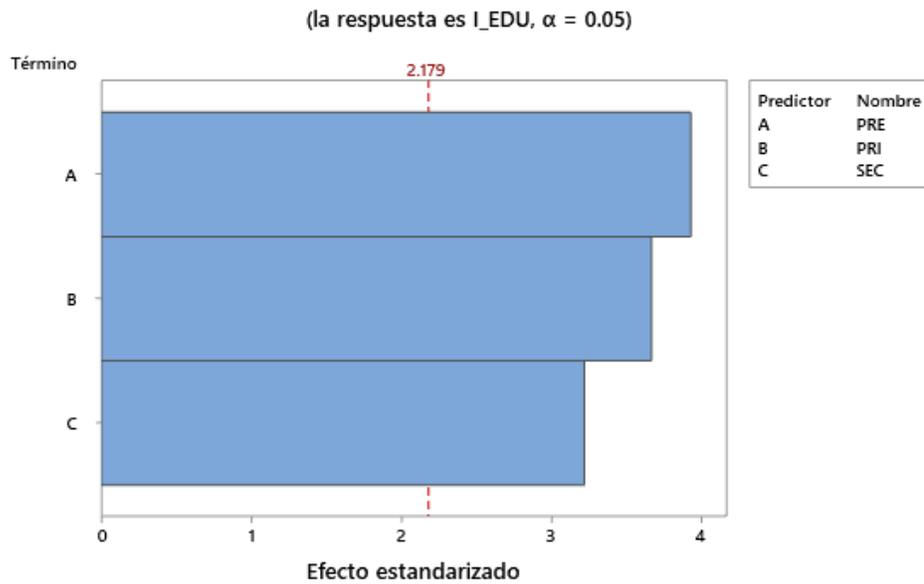
Tabla XVI. **Coefficientes del modelo electo para Guatemala**

Término	Coef	IC de 95 %	Valor T	Valor p
<i>Constante</i>	0.547	(0.257, 0.837)	4.11	0.001
<i>PRE</i>	0.395	(0.176, 0.614)	3.93	0.002
<i>PRI</i>	-0.506	(-0.806, -0.206)	-3.67	0.003
<i>SEC</i>	0.497	(0.161, 0.833)	3.22	0.007

Fuente: elaboración propia.

Algo destacable dentro de los resultados que aporta el análisis de varianza que se muestra en la tabla XV es la contribución de los términos o variables predictoras, que también es posible apreciar de forma gráfica por medio de un diagrama de *Pareto* con los efectos estandarizados, ilustrados en la figura 10.

Figura 10. **Diagrama de Pareto del modelo electo para Guatemala**



Fuente: elaboración propia.

3.7.1. **Supuestos estadísticos del modelo electo para Guatemala**

A continuación, se presentan los resultados y el análisis de los supuestos estadísticos que debe cumplir el modelo propuesto para Guatemala que lo validan.

3.7.1.1. **Supuesto de no multicolinealidad**

Para confirmar este supuesto se evaluó el factor de inflación de la varianza (FIV) el cual puede presentar no correlación si su valor es igual a 1, una correlación moderada si se encuentra entre 1 y 5, y en último caso si supera a 5 se concluye un nivel de correlación alto. Al observar los valores de los factores de inflación de la varianza de cada uno de los términos en la tabla XVII se nota

una correlación moderada en cada uno de ellos, lo que permite concluir que no hay multicolinealidad.

Tabla XVII. **Coefficientes del modelo electo para Guatemala**

Término	FIV
<i>PRE</i>	3.87
<i>PRI</i>	2.83
<i>SEC</i>	4.25

Fuente: elaboración propia.

3.7.1.2. Supuesto de normalidad de residuos

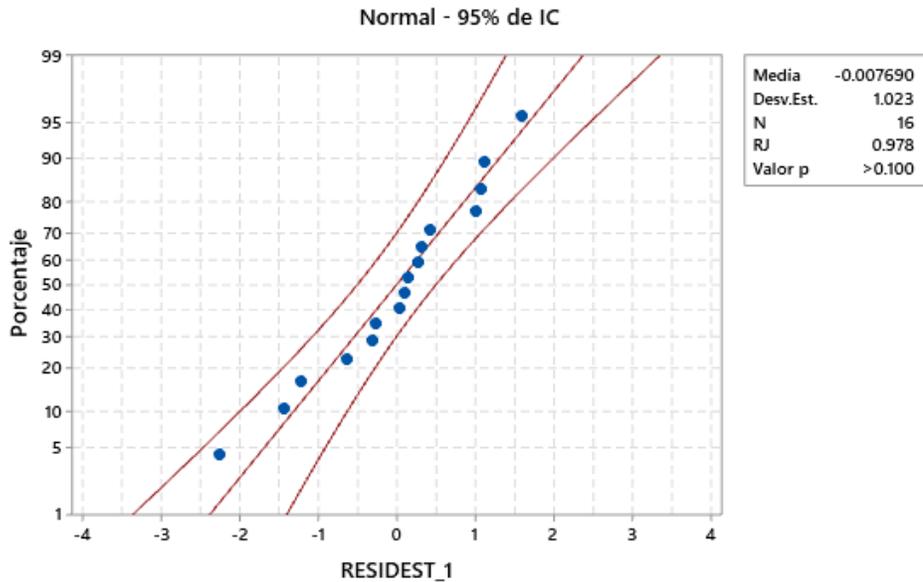
Se aplicó la prueba *Ryan-Joiner* para determinar si los residuos del modelo propuesto presentan un comportamiento normal, en la figura 9 se puede observar cómo los residuos del modelo se ajustan al comportamiento esperado quedando dentro del intervalo de confianza, al realizar el contraste del valor p mayor a 0.100 también se concluye que los residuos tienen un comportamiento que se ajusta a una distribución normal y por consiguiente supuesto de normalidad deseado.

H₀: Los datos siguen una distribución normal

H_a: Los datos no siguen una distribución normal

Bajo el criterio, de aceptar H_0 sí el p-valor es mayor al nivel de significancia α que en este caso se tomó 0.05.

Figura 11. **Gráfico de prueba de normalidad de residuos para Guatemala**



Fuente: elaboración propia.

3.7.1.3. **Supuesto de homocedasticidad**

La homocedasticidad se define como la distribución uniforme de los residuos respecto al valor ajustado, para determinar si el supuesto de homocedasticidad se cumple, se aplicó la prueba de contraste de *Breusch-Pagan* (Tabla XVIII) la cual define la hipótesis nula y alterna como:

H_0 : Los residuos son homocedásticos

H_a : Los resituos son heterocedástios

Bajo el criterio de rechazar H_0 si el valor-p es menor al nivel de significancia α que en este caso corresponde a 0.05, en el caso del modelo seleccionado para Guatemala se concluye que existe homocedasticidad.

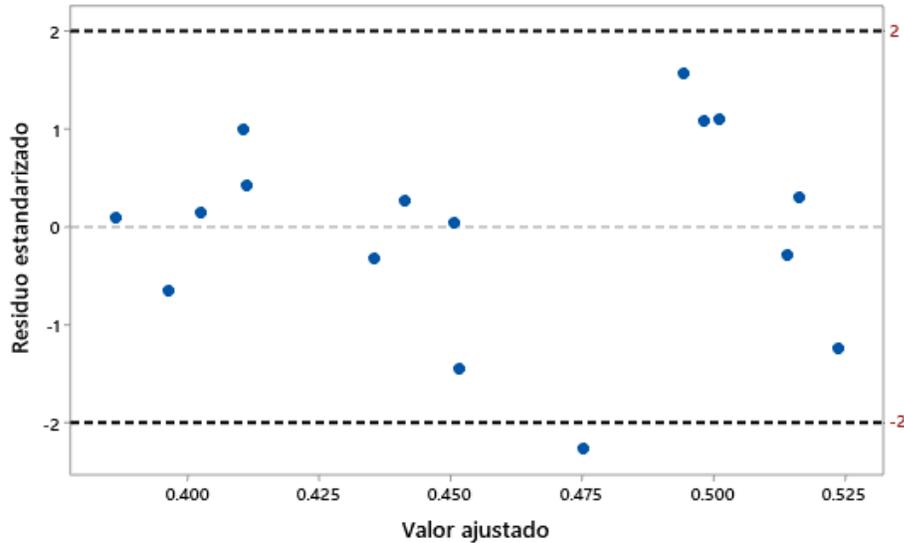
Tabla XVIII. Prueba de contraste Breusch-Pagan para Guatemala

LM (Valor observado)	LM (Valor crítico)	GL	valor-p (bilateral)	alfa
3.520	7.815	3	0.318	0.05

Fuente: elaboración propia.

El cumplimiento del supuesto de homocedasticidad también se evaluó gráficamente, como se puede apreciar en la figura 12 donde la gran mayoría de residuos se encuentran dentro del intervalo de confianza, concluyendo de igual manera que existe homocedasticidad.

Figura 12. Gráfico de homocedasticidad de residuos para Guatemala



Fuente: elaboración propia.

3.7.1.4. Supuesto de independencia de residuos

Para determinar el supuesto de independencia se aplicó un análisis gráfico y a ello se sumó la prueba de contrastes de *Durbin-Watson* para que la conclusión no pasara solamente por una apreciación, en la figura 13 se observa los residuos donde se buscó de identificar una tendencia o comportamiento con algún patrón determinado, que aparentemente no existe, y se procedió a realizar la prueba de contraste con el número de términos del modelo incluido el termino constante y el estadístico de prueba DW como se muestra en la tabla XIX, done el contraste esta dado por:

H_0 : No hay evidencia de correlación

H_a : Hay evidencia de correlación

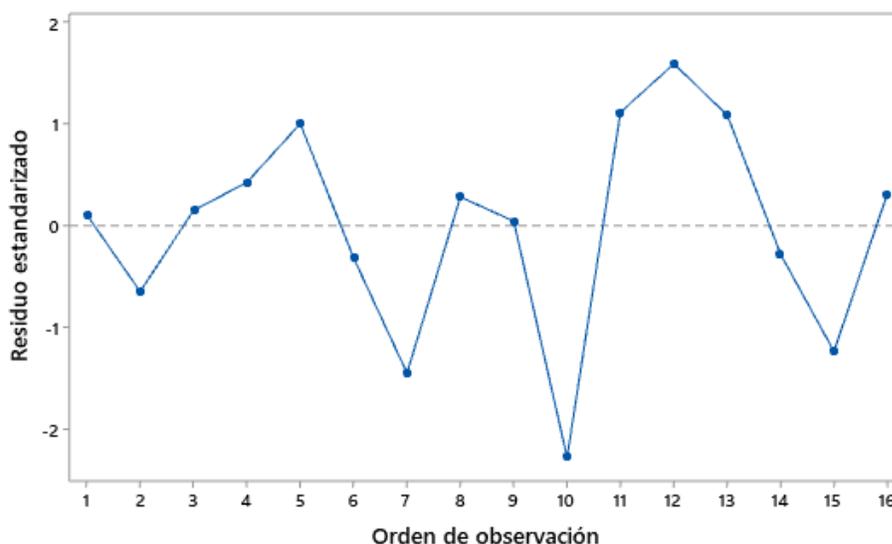
Bajo el criterio de aceptación de H_0 esta dado cuando el estadístico de prueba está dentro del intervalo $D_U < DW < 4 - D_U$. Concluyendo que la independencia existe.

Tabla XIX. Prueba de contraste de Durbin-Watson para Guatemala

n	Parámetros	DW	D_U	$4 - D_U$	alfa
16	4	1.83356	1.722773	2.27727	0.05

Fuente: elaboración propia.

Figura 13. **Gráfico de residuos versus orden para el modelo de Guatemala**



Fuente: elaboración propia.

3.8. Modelo seleccionado para Bolivia

El modelo de pronóstico propuesto para Bolivia incluye como variables predictoras por su nivel de significancia estadística a preprimaria, primaria y secundaria, por consiguiente, los coeficientes asociados a dichas variables comparten el mismo nivel de significancia además de agregar un término constante.

Para su validación se aplicó la siguiente prueba de contraste tanto a los términos como a los coeficientes y el termino constante, como se puede observar en las tablas XX y XXI, correspondientes al análisis de varianza y análisis de coeficientes respectivamente.

La prueba de contraste aplicado a las variables predictoras, coeficientes y termino constante, bajo la siguiente prueba de hipótesis:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

Bajo el criterio de rechazar H_0 sí el p-valor es menor al nivel de significancia α que en este caso se tomó 0.05.

Tabla XX. **Análisis de varianza del modelo electo para Bolivia**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	0.004542	89.68 %	0.004542	0.004542	86.88	0.000
PRE	1	0.004542	89.68 %	0.004542	0.004542	86.88	0.000
Error	10	0.000523	10.32 %	0.000523	0.000052		
Total	11	0.005065	100.00 %				

Fuente: elaboración propia.

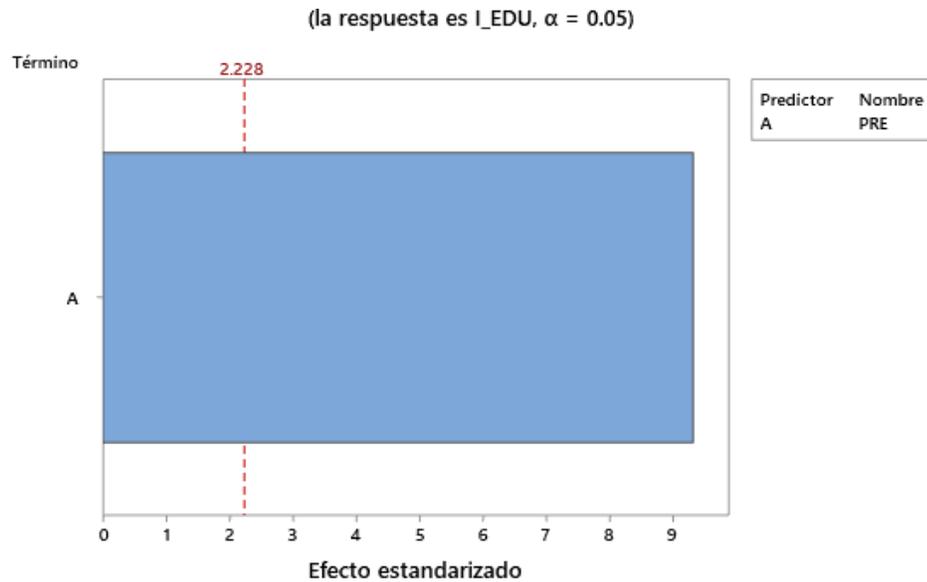
Tabla XXI. **Coefficientes del modelo electo para Bolivia**

Término	Coef	IC de 95 %	Valor T	Valor p
Constante	0.5440	(0.517, 0.570)	45.62	0.000
PRE	0.1802	(0.137, 0.223)	9.32	0.000

Fuente: elaboración propia.

Algo destacable dentro de los resultados que aporta el análisis de varianza que se muestra en la tabla XX es la contribución de los términos o variables predictoras, que también es posible apreciar de forma gráfica por medio de un diagrama de Pareto con los efectos estandarizados, ilustrados en la figura 10.

Figura 14. **Diagrama de Pareto del modelo electo para Bolivia**



Fuente: elaboración propia.

3.8.1. **Supuestos estadísticos del modelo electo para Bolivia**

A continuación, se presentan los resultados y el análisis de los supuestos estadísticos que debe cumplir el modelo propuesto para Bolivia que lo validan.

3.8.1.1. **Supuesto de no multicolinealidad**

Para confirmar este supuesto se evaluó el factor de inflación de la varianza (FIV) el cual puede presentar no correlación si su valor es igual a 1, una correlación moderada si se encuentra entre 1 y 5, y en último caso si supera a 5 se concluye un nivel de correlación alto. Al observar los valores de los factores de inflación de la varianza de cada uno de los términos en la tabla XXII se nota una correlación moderada en cada uno de ellos, lo que permite concluir que no hay multicolinealidad.

Tabla XXII. **Coeficientes del modelo electo para Bolivia**

Término	FIV
PRE	1.00

Fuente: elaboración propia.

3.8.1.2. Supuesto de normalidad de residuos

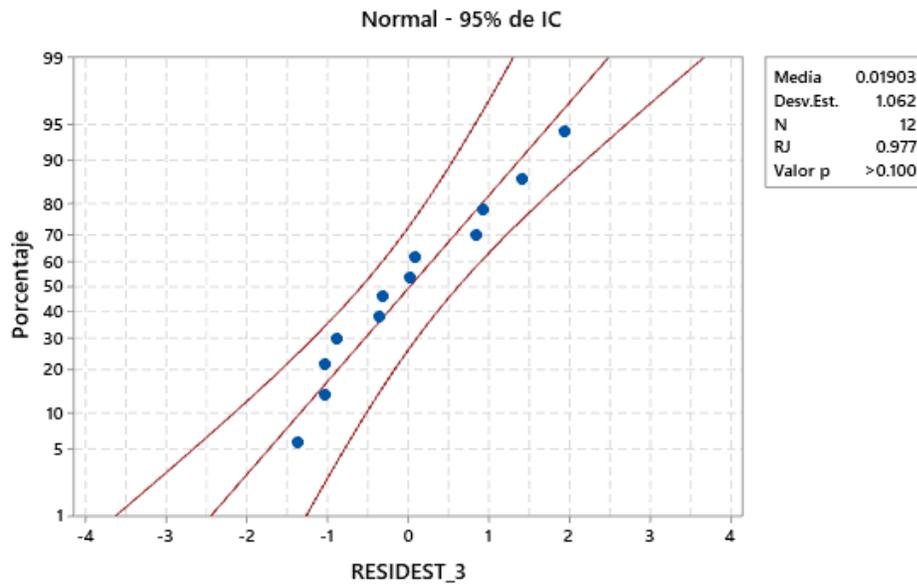
Se aplicó la prueba *Ryan-Joiner* para determinar si los residuos del modelo propuesto presentan un comportamiento normal, en la figura 15 se puede observar cómo los residuos del modelo se ajustan al comportamiento esperado quedando dentro del intervalo de confianza, al realizar el contraste del valor p mayor a 0.100 también se concluye que los residuos tienen un comportamiento que se ajusta a una distribución normal y por consiguiente supuesto de normalidad deseado.

H₀: Los datos siguen una distribución normal

H_a: Los datos no siguen una distribución normal

Bajo el criterio, de aceptar H₀ sí el p-valor es mayor al nivel de significancia α que en este caso se tomó 0.05, aceptando H₀ y concluyendo que los residuos tienen un comportamiento normal.

Figura 15. **Gráfico de prueba de normalidad de residuos para Bolivia**



Fuente: elaboración propia.

3.8.1.3. **Supuesto de homocedasticidad**

La homocedasticidad se define como la distribución uniforme de los residuos respecto al valor ajustado, para determinar si el supuesto de homocedasticidad se cumple, se aplicó la prueba de contraste de *Breusch-Pagan* (Tabla XXIII) la cual define la hipótesis nula y alterna como:

H_0 : Los residuos son homocedásticos

H_a : Los resituos son heterocedásticos

Bajo el criterio de rechazar H_0 si el valor-p es menor al nivel de significancia α que en este caso corresponde a 0.05, en el caso del modelo seleccionado para Bolivia se concluye que existe homocedasticidad.

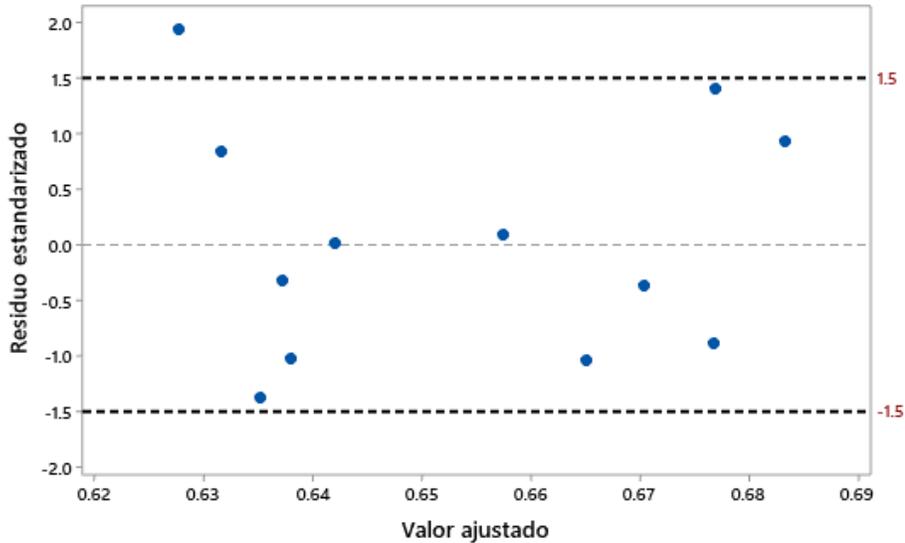
Tabla XXIII. Prueba de contraste Breusch-Pagan para Bolivia

LM (Valor observado)	LM (Valor crítico)	GL	valor-p (bilateral)	alfa
0.707	3.841	1	0.401	0.05

Fuente: elaboración propia.

El cumplimiento del supuesto de homocedasticidad también se evaluó gráficamente, como se puede apreciar en la figura 16 donde la gran mayoría de residuos se encuentran dentro del intervalo de confianza, concluyendo de igual manera que existe homocedasticidad.

Figura 16. Gráfico de homocedasticidad de residuos para Bolivia



Fuente: elaboración propia.

3.8.1.4. Supuesto de independencia de residuos

Para determinar el supuesto de independencia se aplicó un análisis gráfico y a ello se sumó la prueba de contrastes de Durbin-Watson para que la conclusión no pasara solamente por una apreciación, en la figura 13 se observa los residuos donde se buscó de identificar una tendencia o comportamiento con algún patrón determinado, que aparentemente no existe, y se procedió a realizar la prueba de contraste con el número de términos del modelo incluido el termino constante y el estadístico de prueba DW como se muestra en la tabla XXIV, done el contrate esta dado por

H_0 : No hay evidencia de correlación

H_a : Hay evidencia de correlación

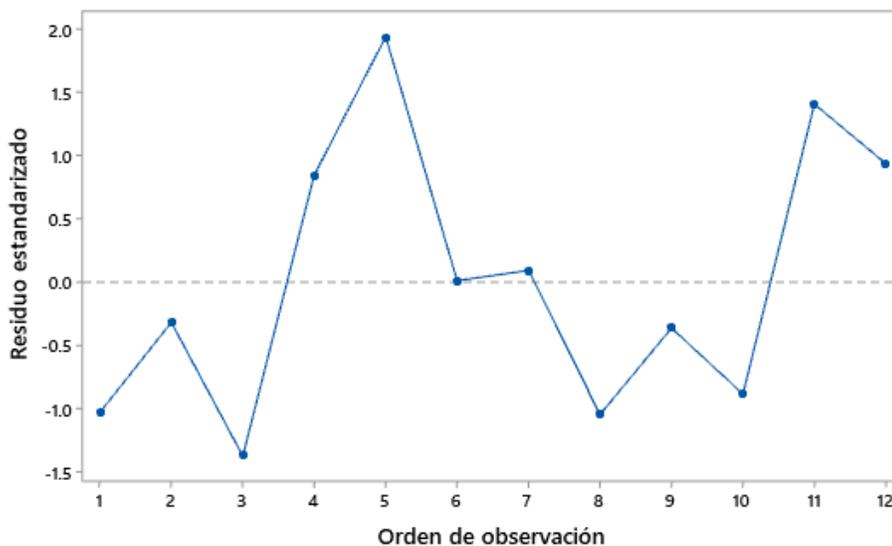
Bajo el criterio de aceptación de H_0 esta dado cuando el estadístico de prueba está dentro del intervalo $D_U < DW < 4 - D_U$. Concluyendo que la independencia existe.

Tabla XXIV. Prueba de contraste de Durbin-Watson para Bolivia

n	Parámetros	DW	D_U	$4 - D_U$	alfa
12	2	1.54051	1.33137	2.66863	0.05

Fuente: elaboración propia.

Figura 17. **Gráfico de residuos versus orden para el modelo de Bolivia**



Fuente: elaboración propia.

3.9. **Modelo seleccionado para México**

El modelo de pronóstico propuesto para México incluye como variables predictoras por su nivel de significancia estadística a preprimaria, primaria y secundaria, por consiguiente, los coeficientes asociados a dichas variables comparten el mismo nivel de significancia además de agregar un término constante.

Para su validación se aplicó la siguiente prueba de contraste tanto a los términos como a los coeficientes y el termino constante, como se puede observar en las tablas XXV y XXVI, correspondientes al análisis de varianza y análisis de coeficientes respectivamente.

La prueba de contraste aplicado a las variables predictoras, coeficientes y termino constante, bajo la siguiente prueba de hipótesis:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

Bajo el criterio de rechazar H_0 sí el p-valor es menor al nivel de significancia α que en este caso se tomó 0.05.

Tabla XXV. **Análisis de varianza del modelo electo para México**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	2	0.008959	92.43 %	0.008959	0.004480	67.12	0.000
PRI	1	0.005314	54.82 %	0.000426	0.000426	6.39	0.028
SEC	1	0.003645	37.61 %	0.003645	0.003645	54.62	0.000
Error	11	0.000734	7.57 %	0.000734	0.000067		
Total	13	0.009694	100.00 %				

Fuente: elaboración propia.

Tabla XXVI. **Coefficientes del modelo electo para México**

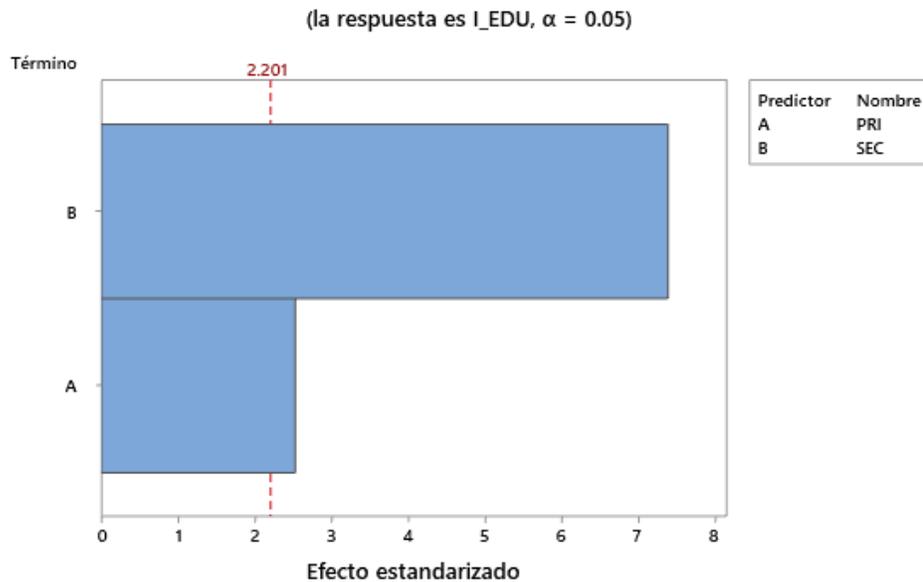
Término	Coef	IC de 95 %	Valor T	Valor p
Constante	-1.082	(-2.289, 0.126)	-1.97	0.074
PRI	1.270	(0.164, 2.375)	2.53	0.028
SEC	0.6708	(0.471, 0.871)	7.39	0.000

Fuente: elaboración propia.

Algo destacable dentro de los resultados que aporta el análisis de varianza que se muestra en la tabla XXV es la contribución de los términos o variables

predictoras, que también es posible apreciar de forma gráfica por medio de un diagrama de Pareto con los efectos estandarizados, ilustrados en la figura 18.

Figura 18. **Diagrama de Pareto del modelo electo para México**



Fuente: elaboración propia.

3.9.1. **Supuestos estadísticos del modelo electo para México**

A continuación, se presentan los resultados y el análisis de los supuestos estadísticos que debe cumplir el modelo propuesto para México que lo validan.

3.9.1.1. **Supuesto de no multicolinealidad**

Para confirmar este supuesto se evaluó el factor de inflación de la varianza (FIV) el cual puede presentar no correlación si su valor es igual a 1, una correlación moderada si se encuentra entre 1 y 5, y en último caso si supera a 5 se concluye un nivel de correlación alto. Al observar los valores de los factores

de inflación de la varianza de cada uno de los términos en la tabla XXVII se nota una correlación moderada en cada uno de ellos, lo que permite concluir que no hay multicolinealidad.

Tabla XXVII. **Coefficientes del modelo electo para México**

Término	FIV
PRI	4.84
SEC	4.84

Fuente: elaboración propia.

3.9.1.2. Supuesto de normalidad de residuos

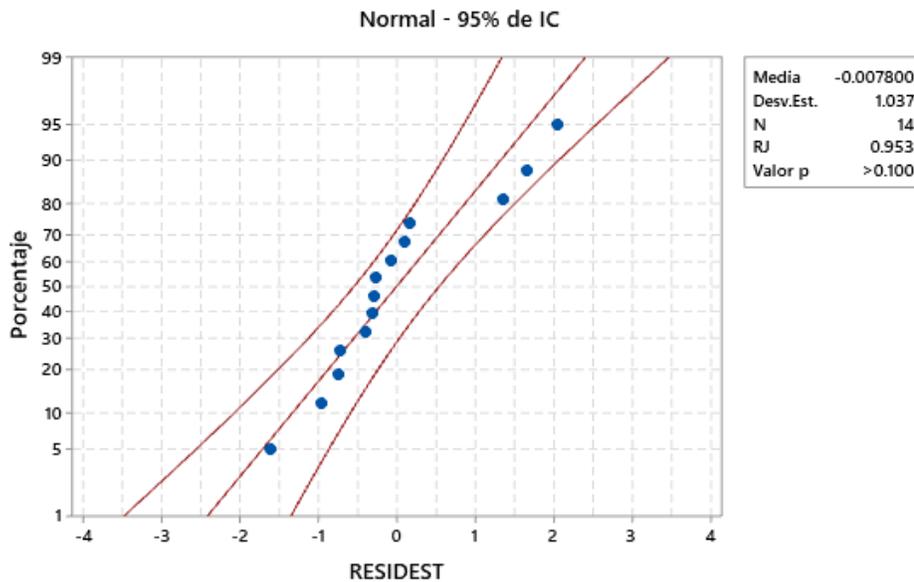
Se aplicó la prueba *Ryan-Joiner* para determinar si los residuos del modelo propuesto presentan un comportamiento normal, en la figura 15 se puede observar cómo los residuos del modelo se ajustan al comportamiento esperado quedando dentro del intervalo de confianza, al realizar el contraste del valor p mayor a 0.100 también se concluye que los residuos tienen un comportamiento que se ajusta a una distribución normal y por consiguiente supuesto de normalidad deseado.

H₀: Los datos siguen una distribución normal

H_a: Los datos no siguen una distribución normal

Bajo el criterio, de aceptar H₀ sí el p-valor es mayor al nivel de significancia α que en este caso se tomó 0.05, aceptando H₀ y concluyendo que los residuos tienen un comportamiento normal.

Figura 19. **Gráfico de prueba de normalidad de residuos para México**



Fuente: elaboración propia.

3.9.1.3. **Supuesto de homocedasticidad**

La homocedasticidad se define como la distribución uniforme de los residuos respecto al valor ajustado, para determinar si el supuesto de homocedasticidad se cumple, se aplicó la prueba de contraste de Breusch-Pagan (Tabla XXVIII) la cual define la hipótesis nula y alterna como:

H_0 : Los residuos son homocedásticos

H_a : Los resituos son heterocedástios

Bajo el criterio de rechazar H_0 si el valor-p es menor al nivel de significancia α que en este caso corresponde a 0.05, en el caso del modelo seleccionado para México se concluye que existe homocedasticidad.

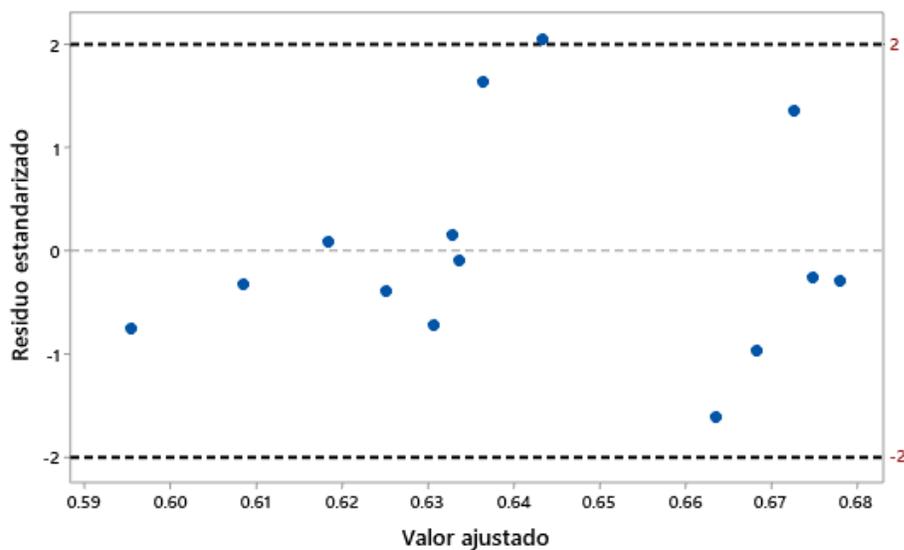
Tabla XXVIII. **Prueba de contraste Breusch-Pagan para México**

LM (Valor observado)	LM (Valor crítico)	GL	valor-p (bilateral)	alfa
1.987	5.991	2	0.370	0.05

Fuente: elaboración propia.

El cumplimiento del supuesto de homocedasticidad también se evaluó gráficamente, como se puede apreciar en la figura 20 donde la gran mayoría de residuos se encuentran dentro del intervalo de confianza, concluyendo de igual manera que existe homocedasticidad.

Figura 20. **Gráfico de homocedasticidad de residuos México**



Fuente: elaboración propia.

3.9.1.4. Supuesto de independencia de residuos

Para determinar el supuesto de independencia se aplicó un análisis gráfico y a ello se sumó la prueba de contrastes de Durbin-Watson para que la conclusión no pasara solamente por una apreciación, en la figura 21 se observa los residuos donde se buscó de identificar una tendencia o comportamiento con algún patrón determinado, que aparentemente no existe, y se procedió a realizar la prueba de contraste con el número de términos del modelo incluido el termino constante y el estadístico de prueba DW como se muestra en la tabla XXIX, done el contraste esta dado por

H_0 : No hay evidencia de correlación

H_a : Hay evidencia de correlación

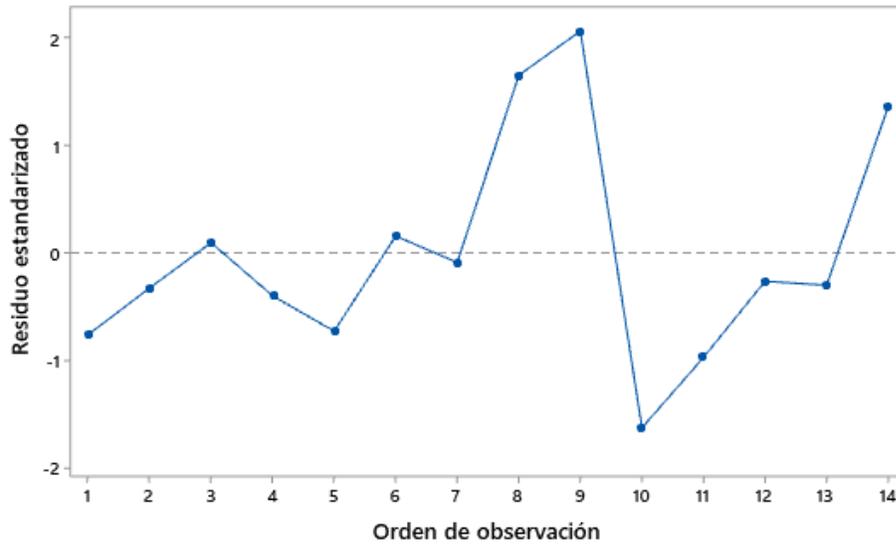
Bajo el criterio de aceptación de H_0 esta dado cuando el estadístico de prueba está dentro del intervalo $D_U < DW < 4 - D_U$. Concluyendo que la independencia existe.

Tabla XXIX. Prueba de contraste de Durbin-Watson para México

n	Parámetros	DW	D_U	$4 - D_U$	alfa
14	3	1.58149	1.55066	2.44934	0.05

Fuente: elaboración propia.

Figura 21. **Gráfico de residuos versus orden para el modelo de México**



Fuente: elaboración propia.

3.10. Modelo seleccionado para Perú

El modelo de pronóstico propuesto para Perú incluye como variables predictoras por su nivel de significancia estadística a preprimaria, primaria y secundaria, por consiguiente, los coeficientes asociados a dichas variables comparten el mismo nivel de significancia además de agregar un término constante.

Para su validación se aplicó la siguiente prueba de contraste tanto a los términos como a los coeficientes y el termino constante, como se puede observar en las tablas XXX y XXXI, correspondientes al análisis de varianza y análisis de coeficientes respectivamente.

La prueba de contraste aplicado a las variables predictoras, coeficientes y termino constante, bajo la siguiente prueba de hipótesis:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

Bajo el criterio de rechazar H_0 sí el p-valor es menor al nivel de significancia α que en este caso se tomó 0.05.

Tabla XXX. **Análisis de varianza del modelo electo para Perú**

Fuente	GL	SC Sec.	Contribución	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	2	0.017608	93.93 %	0.017608	0.008804	100.64	0.000
PRE	1	0.017117	91.31 %	0.017230	0.017230	196.96	0.000
PRI	1	0.000491	2.62 %	0.000491	0.000491	5.61	0.034
Error	13	0.001137	6.07 %	0.001137	0.000087		
Total	15	0.018746	100.00 %				

Fuente: elaboración propia.

Tabla XXXI. **Coefficientes del modelo electo para Perú**

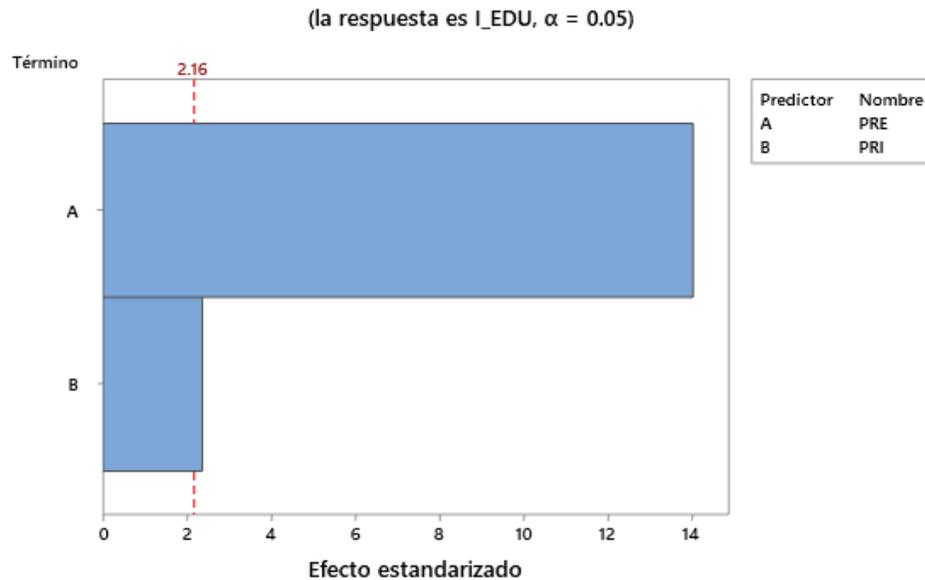
Término	Coef	IC de 95 %	Valor T	Valor p
Constante	0.853	(0.444, 1.262)	4.51	0.001
PRE	0.3900	(0.3300, 0.4500)	14.03	0.000
PRI	-0.484	(-0.925, -0.043)	-2.37	0.034

Fuente: elaboración propia.

Algo destacable dentro de los resultados que aporta el análisis de varianza que se muestra en la tabla XXX es la contribución de los términos o variables

predictoras, que también es posible apreciar de forma gráfica por medio de un diagrama de Pareto con los efectos estandarizados, ilustrados en la figura 22.

Figura 22. **Diagrama de Pareto del modelo electo para Perú**



Fuente: elaboración propia.

3.10.1. **Supuestos estadísticos del modelo electo para Perú**

A continuación, se presentan los resultados y el análisis de los supuestos estadísticos que debe cumplir el modelo propuesto para Perú que lo validan.

3.10.1.1. **Supuesto de no multicolinealidad**

Para confirmar este supuesto se evaluó el factor de inflación de la varianza (FIV) el cual puede presentar no correlación si su valor es igual a 1, una correlación moderada si se encuentra entre 1 y 5, y en el peor de los casos si

supera a 5 se concluye un nivel de correlación alto. Al observar los valores de los factores de inflación de la varianza de cada uno de los términos en la tabla XXXII se nota una correlación moderada en cada uno de ellos, lo que permite concluir que no hay multicolinealidad.

Tabla XXXII. **Coefficientes del modelo electo para Perú**

Término	FIV
PRE	1.00
PRI	1.00

Fuente: elaboración propia.

3.10.1.2. Supuesto de normalidad de residuos

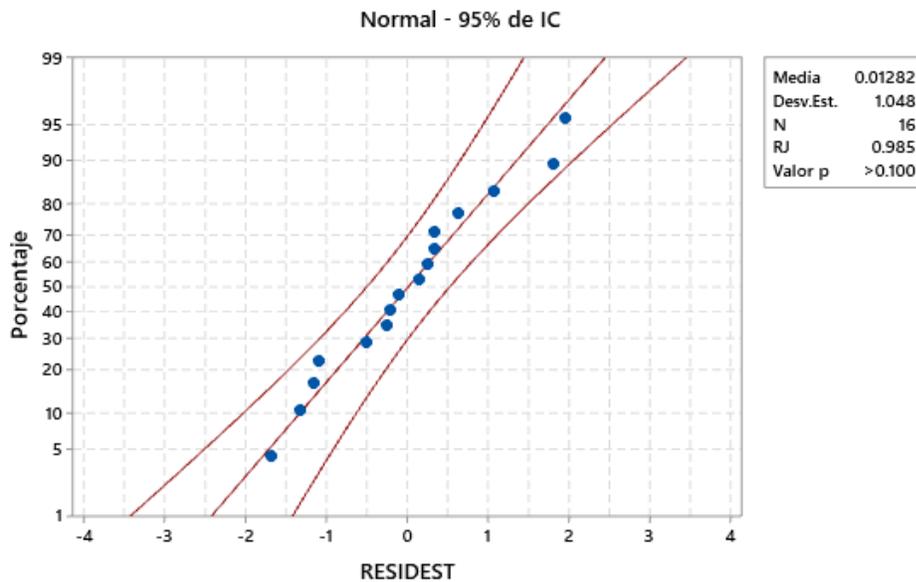
Se aplicó la prueba *Ryan-Joiner* para determinar si los residuos del modelo propuesto presentan un comportamiento normal, en la figura 23 se puede observar cómo los residuos del modelo se ajustan al comportamiento esperado quedando dentro del intervalo de confianza, al realizar el contraste del valor p mayor a 0.100 también se concluye que los residuos tienen un comportamiento que se ajusta a una distribución normal y por consiguiente supuesto de normalidad deseado.

H_0 : *Los datos siguen una distribución normal*

H_a : *Los datos no siguen una distribución normal*

Bajo el criterio, de aceptar H_0 sí el p-valor es mayor al nivel de significancia α que en este caso se tomó 0.05, aceptando H_0 y concluyendo que los residuos tienen un comportamiento normal.

Figura 23. **Gráfico de prueba de normalidad de residuos para Perú**



Fuente: elaboración propia.

3.10.1.3. **Supuesto de homocedasticidad**

La homocedasticidad se define como la distribución uniforme de los residuos respecto al valor ajustado, para determinar si el supuesto de homocedasticidad se cumple, se aplicó la prueba de contraste de *Breusch-Pagan* (Tabla XXXIII) la cual define la hipótesis nula y alterna como:

H_0 : Los residuos son homocedásticos

H_a : Los resituos son heterocedásticos

Bajo el criterio de rechazar H_0 si el valor-p es menor al nivel de significancia α que en este caso corresponde a 0.05, en el caso del modelo seleccionado para Perú se concluye que existe homocedasticidad.

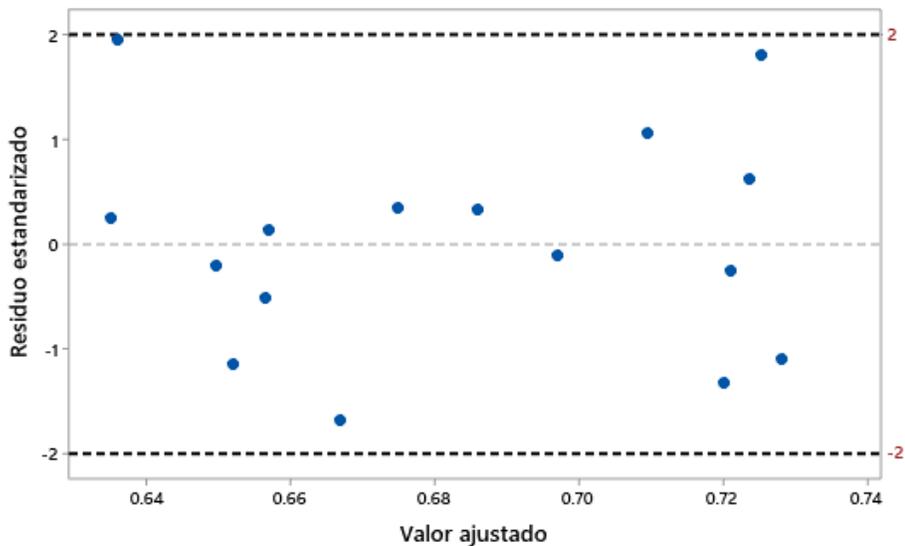
Tabla XXXIII. Prueba de contraste Breusch-Pagan para Perú

LM (Valor observado)	LM (Valor crítico)	GL	valor-p (bilateral)	Alfa
1.987	5.991	2	0.370	0.05

Fuente: elaboración propia.

El cumplimiento del supuesto de homocedasticidad también se evaluó gráficamente, como se puede apreciar en la figura 24 donde la gran mayoría de residuos se encuentran dentro del intervalo de confianza, concluyendo de igual manera que existe homocedasticidad.

Figura 24. Gráfico de homocedasticidad de residuos para Perú



Fuente: elaboración propia.

3.10.1.4. Supuesto de independencia de residuos

Para determinar el supuesto de independencia se aplicó un análisis gráfico y a ello se sumó la prueba de contrastes de *Durbin-Watson* para que la conclusión no pasara solamente por una apreciación, en la figura 25 se observa los residuos donde se buscó de identificar una tendencia o comportamiento con algún patrón determinado, que aparentemente no existe, y se procedió a realizar la prueba de contraste con el número de términos del modelo incluido el termino constante y el estadístico de prueba DW como se muestra en la tabla XXXIV, done el contrade esta dado por:

H_0 : No hay evidencia de correlación

H_a : Hay evidencia de correlación

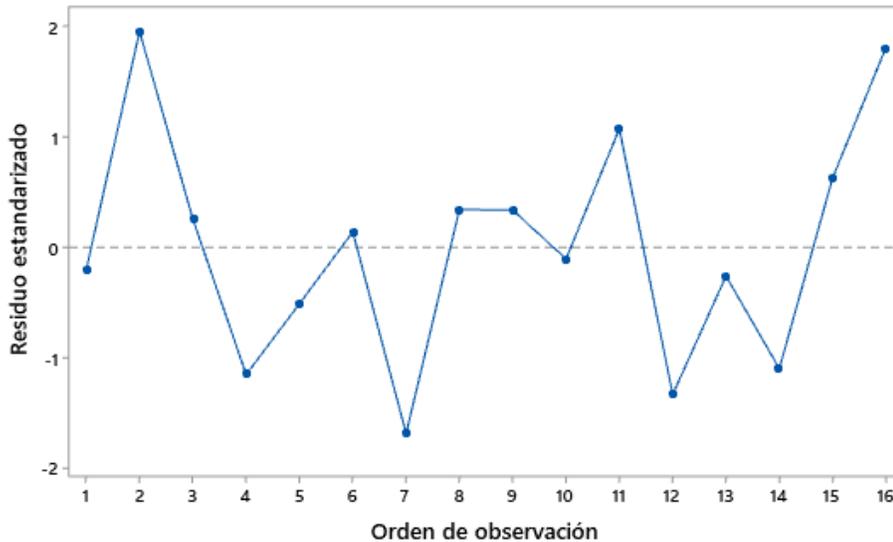
Bajo el criterio de aceptación de H_0 esta dado cuando el estadístico de prueba está dentro del intervalo $D_U < DW < 4 - D_U$. Concluyendo que la independencia existe.

Tabla XXXIV. Prueba de contraste de Durbin-Watson para Perú

n	Parámetros	DW	D_U	$4 - D_U$	alfa
16	3	1.93919	1.5386	2.4616	0.05

Fuente: elaboración propia.

Figura 25. **Gráfico de residuos versus orden para el modelo de Perú**



Fuente: elaboración propia.

Objetivo 3. Comparar la inversión otorgada a educación en Guatemala con respecto a Bolivia, México y Perú, países con contextos similares al guatemalteco, por medio de modelos estadísticos de regresión que proyecten el comportamiento de la dimensión de educación en cada uno de ellos.

3.11. Comparación de los modelos propuestos

Todos los modelos ofrecen un nivel de máximo ajuste y una capacidad de predicción bastante cercana con a los valores observados, pero se puede observar en las ecuaciones las variables de predicción cambian y aunque parecidos responden a características propias los modelos educativos en cada país, la principal similitud es que la variable %PIB no es utilizada en ningún modelo.

3.11.1. Modelos propuestos para cada país

Los modelos que se muestran en la tabla muestran diferencias significativas con respecto a los demás en la tabla XXXV, estas diferencias se pueden apreciar en el valor constante, los coeficientes y las variables incluidas.

Tabla XXXV. **Modelos seleccionados para los distintos países**

País	Modelo Propuesto
Guatemala	$I_{EDU} = 0.547 + 0.395 PRE - 0.506 PRI + 0.497 SEC$
Bolivia	$I_{EDU} = 0.5440 + 0.1802 PRE$
México	$I_{EDU} = -1.082 + 1.270 PRI + 0.6708 SEC$
Perú	$I_{EDU} = 0.853 + 0.3900 PRE - 0.484 PRI$

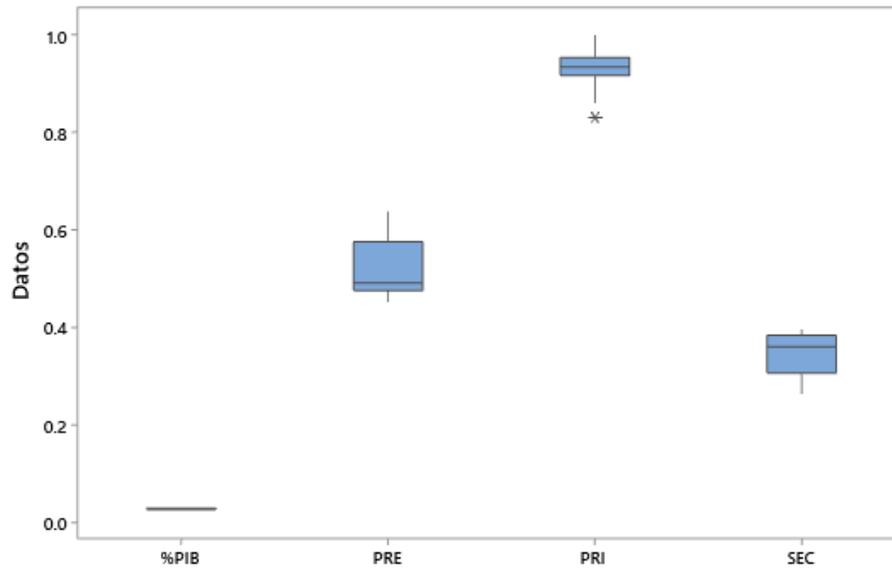
Fuente: elaboración propia.

3.11.2. Gráficos de cajas para los modelos propuestos de los países

Las figuras 26, 27, 28 y 29 representan el comportamiento de cada variable en el modelo, el tamaño de cada caja representa el nivel de cambio sufrido por cada variable en el periodo analizado, siendo las cajas más grandes las que experimentaron esto de forma significativa y por el contrario las cajas más reducidas las que lo experimentaron en menor escala.

En el caso de Guatemala los cambios más grandes se experimentaron en preprimaria y secundaria, y en menor medida primaria, pero en todo caso lo suficiente en cada caso para que cada variable fuera considerada estadísticamente significativa, aunque por debajo del 65 % muy atrás de los niveles de los otros países, mientras que la inversión practicante fue un elemento constante.

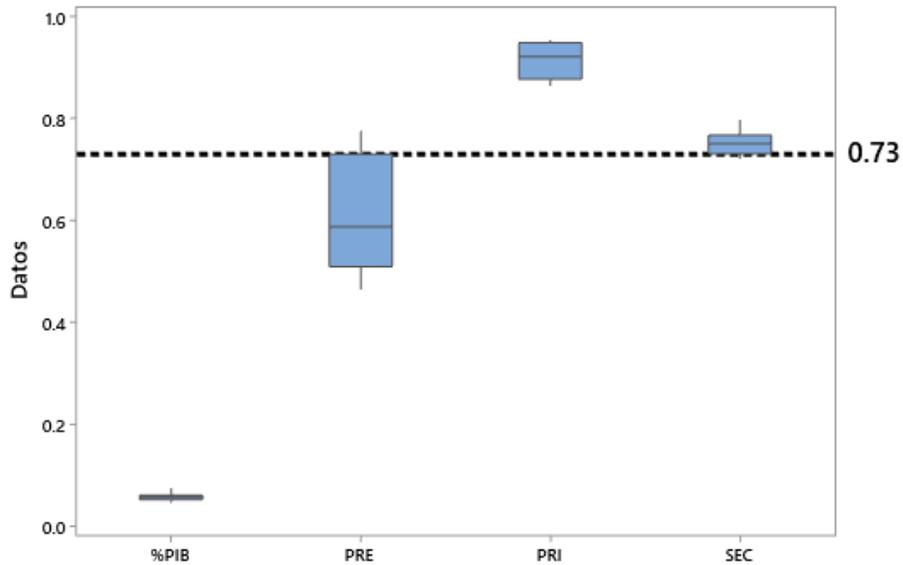
Figura 26. Gráfico de cajas para Guatemala



Fuente: elaboración propia.

En cuanto a Bolivia los cambios más grandes se experimentaron en preprimaria y primaria, y en menor escala en secundaria, todos los niveles experimentaron un aumento, siendo el más significativo en preprimaria, tanto que todo el modelo se pudo explicar por esta variable, mientras que la inversión practicante fue un elemento constante.

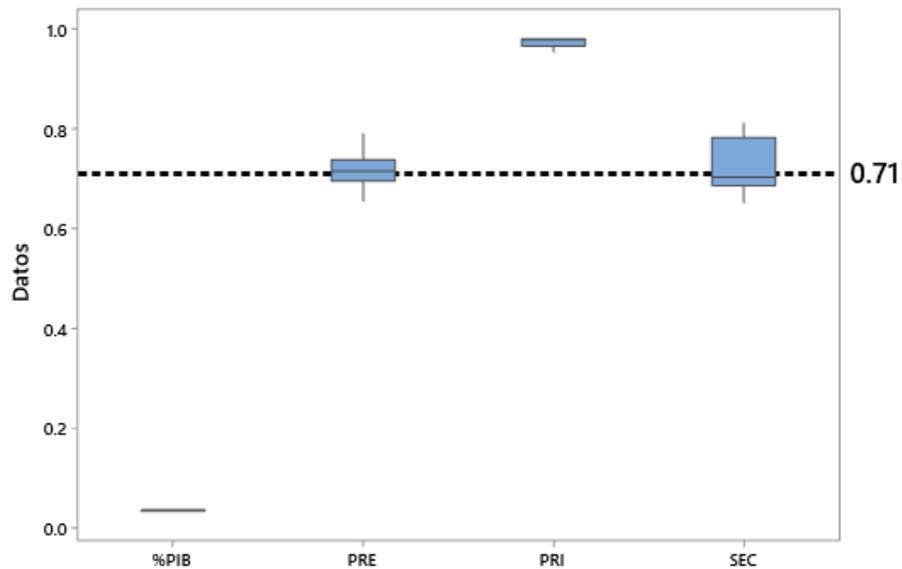
Figura 27. Gráfico de cajas para Bolivia



Fuente: elaboración propia.

Para México los cambios fueron menos notorios en los distintos niveles, los cambios más grandes se experimentaron en secundaria y preprimaria, y en menor escala en primaria, todos los niveles experimentaron un aumento, siendo el más significativo en secundaria, este cambio en comparación con preprimaria fue mayor y al estar al mismo nivel como muestra la línea de referencia del gráfico dio prioridad secundaria y a primaria en el modelo, mientras que la inversión practicante fue un elemento constante.

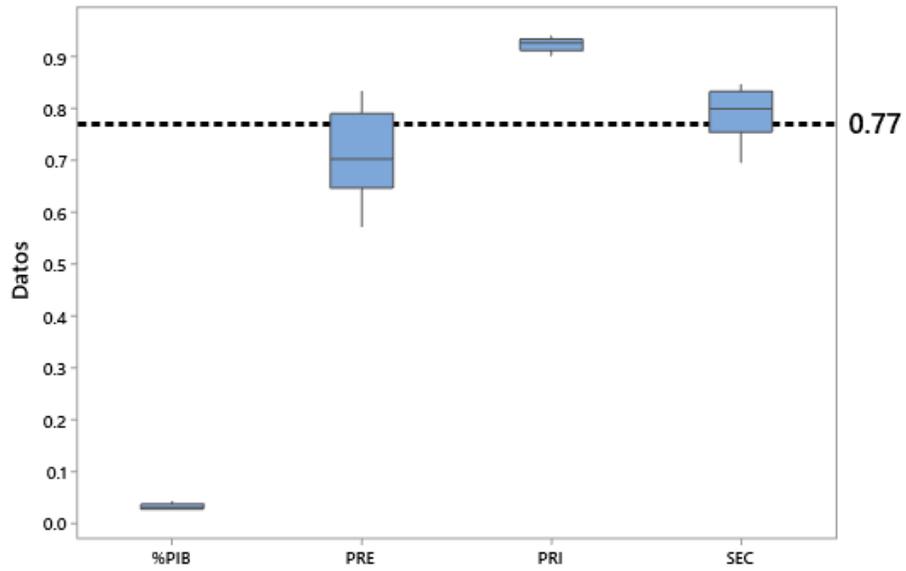
Figura 28. Gráfico de cajas para México



Fuente: elaboración propia.

Por último Perú, experimento sus mayores cambios de cobertura educativa en Preprimaria y secundaria, y bastante menor en Primaria, nuevamente la línea de referencia muestra la coincidencia de dos variables en un nivel siendo estas preprimaria y secundaria, donde el cambio significativamente mayor produce mayor representatividad del moldeo por lo que secundaria fue eliminada del modelo, en cuanto a primaria a pesar de su crecimiento en mucha menor escala fue suficiente para ser considerada una variable estadísticamente significativa, y nuevamente la inversión no formó parte del modelo.

Figura 29. Gráfico de cajas para Perú



Fuente: elaboración propia.

4. DISCUSIÓN DE RESULTADOS

Una vez que se contó con la base referencial y técnica para la elaboración de la información, el trabajo de investigación se estructuró en tres etapas, la primera de ellas consistió en el análisis de los datos recabados, para determinar la viabilidad de construir un modelo con base en ellos.

La segunda etapa fue la modelización, en la cual se generaron múltiples propuestas de modelos para cada país, a estos modelos se les aplicaron diferentes pruebas y con base en ellas se presentó un modelo concreto.

4.1. Resultado del análisis preliminar de la información

Se analizaron los datos históricos de Guatemala, Bolivia, México y Perú, de la dimensión de educación del índice de desarrollo humano, se aplicó una estandarización de los datos de cada modelo y a estos resultados se les aplicó la prueba de normalidad de *Ryan-Joiner* que en concepto es similar a la de *Shapiro-Wilks*, recomendada cuando las muestras o en este caso las poblaciones son pequeñas, en la sección 3.2 del presente trabajo se aprecia el resultado de la prueba de contrastes y los gráficos de normalidad también conocidos como *Q-Q plots*, por los que se confirmó el comportamiento normal de las poblaciones para cada país dada esta importante condición inicial con un 95 % de nivel de confianza.

4.2. Resultado de la evaluación gráfica de los efectos

En la sección 3.3 se presentó un apartado donde se evaluaba el efecto de las variables predictoras sobre la variable independiente, entendiéndose que a mayor pendiente o inclinación en la gráfica el efecto es mayor en el modelo, mientras que cuanto más cercana a una línea horizontal, el efecto es mucho menor, gracias a este criterio se puede observar las variables más significativas para cada modelo.

En el caso de Guatemala el análisis gráfico presento como candidatas casi seguras a conformar el modelo las variables PRE, PRI y SEC, mientras aparentemente se queda descartada la variable %PIB.

Para Bolivia las variables candidatas más probables fueron PRE y PRI, y por el contrario SEC y %PIB quedan excluidas.

En cuanto a México las variables con mayor probabilidad de formar parte del modelo fueron PRI y SEC, quedando fuera del mismo %PIB y PRE.

Por último, Perú sometido al mismo análisis presento únicamente la variable PRI como la más probable descartando al resto.

Es importante señalar que el método gráfico no fue determinante para la construcción del modelo, pero si brindo un punto de partida para la realización del análisis global de los modelos de los distintos países.

4.3. Resultado del análisis de varianza de los modelos generales de cada país

En el presente trabajo se denominó modelos generales a los modelos que incluyen todas las variables predictoras propuestas, este modelo fue el punto de partida para la creación de los diferentes modelos que fueron propuestos para cada país.

Sobre el análisis de varianza, ANOVA como también se le conoce presenta una ventaja sobre el método gráfico, permite cuantificar la contribución al modelo por parte de cada variable y también su nivel de significancia estadística por medio del valor-p cuando este rechaza o acepta la hipótesis alterna según la prueba de contraste aplicada.

En el caso de Guatemala se confirmó que PRE, PRI y SEC eran estadísticamente significativas y %PIB no lo era por tanto es candidata para formar parte del modelo.

Para Bolivia el ANOVA descarto SEC que había sido considerada originalmente después del análisis gráfico, aunque no fue descartada del todo por el margen tan estrecho entre ser estadísticamente significativa o no serlo, dejando el modelo únicamente con la variable PRE.

En el caso de México a las variables excluidas se añadió SEC aunque bajo reserva dado la poca diferencia existente para ser considerada estadísticamente significativa.

Por último, Perú experimento lo mismo con la variable PRI, ubicada justo en la frontera entre ser significativa o no serlo.

4.4. Resultados del proceso de modelización

Se presentaron los siete mejores modelos para cada país, y como criterio de selección inicial se utilizó el coeficiente de determinación ajustado para determinar los mejores modelos en cada caso, esto permitió simplificar las opciones como se puede apreciar en la sección 4.4.1. descrita a continuación.

En el caso de Guatemala y México, fue más que suficiente con el coeficiente de determinación, pero en el caso de Bolivia y Perú se hizo necesario un segundo método de análisis para garantizar la elección de la mejor opción posible, usando para ello los criterios de información AICc y BIC junto con el principio de parsimonia por medio del estadístico *Cp de Mallows* como se detalla en la sección 4.1.2. y que llevaron a la obtención de los modelos con el máximo ajuste.

4.4.1. Depuración de los modelos basado en el coeficiente de determinación

Para Guatemala se logró determinar un único modelo para la predicción de I_EDU en función de las variables PRE, PRI y SEC, coincidiendo con las estimaciones previas del análisis gráfico y del análisis de varianza con un coeficiente de determinación del 93.3 %.

Para Bolivia se lograron determinar dos modelos para la predicción de I_EDU, el primero en función de %PIB, PRE y SEC, y un segundo modelo en función de la variable PRE, el primer modelo coincidió con la evaluación realizada a partir de la gráfica mientras que el segundo apegado al análisis de varianza y en ambos casos con idéntico coeficiente de determinación de 88.6 %.

En el caso de México, tal como sucedió con Guatemala el coeficiente de determinación que alcanzo un 91 % permitió identificar un solo modelo como el mejor para la predicción de I_EDU en función de PRE y PRI.

En cuanto a Perú se lograron determinar dos modelos como sucedió con Bolivia para la predicción de I_EDU, el primero en función de %PIB, PRE y PRI, y un segundo modelo en función de las variables PRE y PRI, el primer modelo coincidió con la evaluación realizada a partir de la gráfica mientras que el segundo apegado al análisis de varianza y en ambos casos con idéntico coeficiente de determinación de 93 %.

4.4.2. Depuración de los modelos basado en los criterios parsimonia y de información AICc y BIC

En el caso de Guatemala y México se confirmó que los modelos resultantes del análisis del coeficiente de determinación fueron congruentes con los resultados obtenidos por medio de los criterios de parsimonia y de información AICc y BIC, donde de manera general los resultados más bajos bajo ambos criterios son las mejores opciones.

Con respecto a Bolivia y Perú ambos países con dos modelos igual de buenos bajo el criterio de mejor coeficiente de determinación, la aplicación de los criterios de parsimonia y de la información AICc y BIC permitió inclinar la balanza en el caso de Bolivia por el modelo de predicción de I_EDU en función de la variable PRE, mientras que en el caso de Perú el modelo de predicción de la variable I_EDU se definió en función de las variables PRE y PRI.

4.5. Validación de supuestos de los modelos y selección

Los modelos propuestos para Guatemala, Bolivia, México y Perú en este punto ofrecen el máximo ajuste con base en los planteamientos sobre los que se construyeron los modelos, para ello se aplicaron las pruebas de multicolinealidad, Normalidad de residuos, homocedasticidad de residuos e independencia de residuos.

4.5.1. Resultado del supuesto de multicolinealidad

El supuesto de multicolinealidad fue evaluado por medio del estadístico conocido como factor de inflación de varianza (FIV) cuyo valor es calculado para cada variable predictora, en cada uno de los modelos propuestos para cada país como resultado del proceso de modelización que se realizó, y cuyo valor para considerar que se cumplió el supuesto no debe exceder de 5 según el software empleado para el análisis, aunque varios autores llevan el valor máximo hasta 10.

En el caso de Guatemala las variables predictoras incluidas en el modelo con máximo ajuste fueron PRE, PRI y SEC, que tuvieron valores FIV de 3.84, 2.83 y 4.25 respectivamente, en ninguna de ellas el valor supero el límite teórico de 5 por lo que se concluyó que el supuesto se cumplió.

Con respecto a Bolivia la variable predictora en el modelo con máximo ajuste fue PRE, con un valor FIV de 1, este valor no supero el límite teórico de 5 por lo que se concluyó que el supuesto se cumplió.

Para México las variables predictoras incluidas en el modelo con máximo ajuste fueron PRI y SEC, que tuvieron valores FIV de 4.84 en ambas, observando

que en ninguna de ellas el valor supero el límite teórico de 5 por lo que se concluyó que el supuesto se cumplió.

En cuanto a Perú las variables predictoras incluidas en el modelo con máximo ajuste fueron PRE y PRI, que tuvieron valores FIV de 1 en ambas, observando que en ninguna de ellas el valor supero el límite teórico de 5 por lo que se concluyó que el supuesto se cumplió.

4.5.2. Resultado del supuesto de normalidad

El supuesto de normalidad fue evaluado para cada modelo de dos maneras, con el gráfico de estandarización de residuos donde estos permanecen dentro del intervalo de confianza y por medio de pruebas de contraste del valor p como resultado de la prueba de normalidad de *Ryan-Joiner* dado lo reducida de las poblaciones como fue presentando en el capítulo anterior.

Para los modelos seleccionados en cada país se alcanzó la misma conclusión por medio del método gráfico y el de contrastes, los residuos cumplen con el supuesto de normalidad necesario para dar por valido cada uno de dichos modelos con un nivel de confianza del 95 %.

4.5.3. Resultado del supuesto de homocedasticidad de residuos

El supuesto de homocedasticidad fue evaluado para cada modelo de dos maneras, con el gráfico de residuos versus ajustes ubicando los residuos dentro de los intervalos de confianza y por medio de pruebas de contraste del valor p como resultado de la prueba de homocedasticidad de *Breusch-Pagan* presentando en el capítulo anterior.

Para los modelos seleccionados en cada país se alcanzó la misma conclusión por medio del método gráfico y el de contrastes, los residuos cumplen con el supuesto de homocedasticidad necesario para dar por válido cada uno de dichos modelos con un nivel de confianza del 95 % como es habitual.

4.5.4. Resultado del supuesto de independencia de residuos

El supuesto de independencia de residuos fue evaluado para cada modelo de dos maneras, con el gráfico de residuos versus orden sin identificar patrones o tendencias y por medio de pruebas de contraste del valor p como resultado de la prueba de independencia de *Durbin-Watson* presentando en el capítulo anterior.

Para los modelos seleccionados en cada país se alcanzó la misma conclusión por medio del método gráfico y el de contrastes, los residuos cumplen con el supuesto de independencia necesario para dar por válido cada uno de dichos modelos con un nivel de confianza del 95 %.

4.6. Resultados de comparación entre modelos seleccionados de los distintos países

Los modelos responden a diferentes contextos por lo que es congruente que las similitudes entre ellos sean escasas, sin embargo, las diferencias manifiestas ayudan a entender el porqué del rezago en Guatemala, todos los modelos experimentaron crecimiento y sostenimiento de los niveles de cobertura alcanzados, siendo la diferencia principal el nivel de atraso en esas áreas.

La inversión en educación representada por %PIB es poco significativa, no por su falta de relación en el modelo sino porque está implícita en el aumento de cobertura.

CONCLUSIONES

1. Se estimó el nivel de correlación por medio de su relación directa con el coeficiente de determinación, de cada uno de los modelos propuestos, determinando el mejor ajuste al contrastarlos, entre los diferentes modelos generados para los países analizados, en el caso de Guatemala y México identificando los modelos que al final fueron los seleccionados, en cuanto a Bolivia y Perú se logró reducir el número de modelos a dos por país, que fueron analizados bajo otros criterios para alcanzar el máximo ajuste en cada caso particular.
2. Se seleccionaron los modelos con el máximo nivel de ajuste aplicando los criterios de parsimonia, información de Akaike (AICc) y Bayesiano (BIC), en el caso de Guatemala y México confirmando la elección hecha bajo el criterio del coeficiente de determinación y en el caso de Bolivia y Perú donde permitió definir un único modelo dentro de las opciones que el coeficiente de determinación ya había reducido logrando con ello.
3. La inversión en educación está implícita en el aumento en la tasa de cobertura neta de los diferentes niveles, por lo que un aumento del presupuesto en educación puede tener relación directa con el aumento de cobertura que se extienda a un aumento en la dimensión de educación y de igual manera en el índice de desarrollo humano.
4. Se construyó un modelo con nivel máximo de ajuste para cada uno de los países: Guatemala, Bolivia, México y Perú con valores de 93.3 %, 88.6 %, 88.6 %, 88.6 %.

91 % y 93 % respectivamente, el denominador común fue la no inclusión de la variable %PIB por el bajo efecto presentado en cada nivel.

RECOMENDACIONES

1. El modelo proporcionado a partir de este estudio permite predecir a corto y mediano plazo la dimensión de educación de educación con un excelente nivel de ajuste como lo deja ver el coeficiente de determinación, y podrá extender su uso en la medida en la que los niveles de cobertura neta educativa no cambien demasiado, pero también es una invitación a estar pendiente de cambios sutiles en los índices que lo conforman para, poder actualizar el modelo en la medida en la que el tiempo transcurra, y de suma utilidad por instituciones que fomenten el desarrollo por medio de la educación.
2. Utilizar los modelos propuestos en esta investigación representan una alternativa confiable y rápida para la predicción de la dimensión de educación del índice de desarrollo humano por su alto nivel de ajuste, en cada caso superior al 90 %.
3. Aplicar otros criterios para identificar ente una variedad de modelos el que ofrezca el mejor ajuste es una práctica por demás recomendada, pues en la medida en la que el modelo se vuelva obsoleto, puede generar nuevos modelos con el mismo nivel de ajuste y sea necesario discriminar ente ellos el mejor utilizando para ello los criterios de parsimonia y de información como el AICc y el BIC, que se debe implementar por las instituciones o entidades que estén dispuestas a emplearlo.
4. La comparación de modelos o benchmarking de modelos realizada entre los modelos generados para cada país incluido en el estudio, permite

identificar áreas de mejora con respecto a la realidad nacional generando posibles líneas de acción en la implementación de políticas de desarrollo educativo.

REFERENCIAS

1. Anderson, D., Sweeney, D., Williams, T. (2012). *Estadística para Negocios y Economía*. México: Cengage Learning.
2. Barros, V., Gallegos, D. y Pavón, C., (2018). Muestreo para el levantamiento de datos acerca de la enseñanza de física experimental en Guayaquil. *Revista Lasallista de Investigación* 15 (2), 223-231. Recuperado de <http://repository.lasallista.edu.co:8080/ojs/index.php/rldi/article/view/1859/210210317>
3. Enke, D., Grauer, M. y Mehdiyev, N. (2011). Stock Market Prediction with Multiple Regression, Fuzzy Type-2 Clustering and Neural Networks. *Procedia Computer Science* 6 (2011), 201-206. Recuperado de <https://doi.org/10.1016/j.procs.2011.08.038>
4. Escutia, I. (9 de 11 de 2019). *Descomposición de series de tiempo*. Recuperado de RPubs by RStudio. <https://rpubs.com/ltzelEscutia/546278>
5. Frías-Navarro, D. (2011). *Técnica estadística y diseño de investigación*. España: Palmero ediciones.
6. Guisande, C. (2006). *Tratamiento de datos*. España: Universidad de Vigo.

7. Gutiérrez, H., y De la Vara, R., (2012) *Análisis y diseño de experimentos*. México: McGraw-Hill.
8. Lind, D., Marchal, W. y Wathen, S. (2012). *Estadística Aplicada a los Negocios y Economía* México: McGraw-Hill Interamericana.
9. Montgomey, D., Hines, W. (1996). *Probabilidad y estadística para ingeniería y administración*. México: Compañía Editorial Continental.
10. Morduchowicz, A. (2006) *Los indicadores educativos y las dimensiones que los integran*. Buenos Aires, Argentina: IIPE – UNESCO.
11. Navidi, W. (2006). *Estadística para ingenieros*. México: McGraw-Hill Interamericana.
12. Pawlowsky-Glahn, V. & Buccianti, A. (2011), *Compositional data analysis: Theory and applications*, United Kingdom: John Wiley & Sons.
13. Programa de las Naciones Unidas para el Desarrollo. (2016) *Informe nacional de desarrollo humano*. Guatemala: PNUD.
Recuperado de <http://desarrollohumano.org.gt/desarrollo-humano/calculo-de-idh/>
14. Posada, S., y Rosero, R. (2007). Comparación de modelos matemáticos: una aplicación en la evaluación de alimentos para animales. *Revista Colombiana de Ciencias Pecuarias* 20 (2), 141-148.

Recuperado de <https://www.redalyc.org/articulo.oa?id=295023034006>

15. ONUSIDA. (2015). *Programa de las Naciones Unidas para el Desarrollo*. Recuperado de https://www.unaids.org/sites/default/files/media_asset/PNUD_es.pdf
16. Torrents, D., Fachelli, S. (2015). El efecto del origen social con el paso del tiempo: la inserción laboral de los graduados universitarios españoles durante la democracia. *Revista Complutense de Educación* 26 (2) 331-349. Recuperado de <https://revistas.ucm.es/index.php/RCED/article/view/43070/45512>
17. Sevilla, A. (2021) *Producto interior bruto (PIB)*. Recuperado de Economipedia <https://economipedia.com/definiciones/producto-interior-bruto-pib.html>.
18. Triola, M. (2018). *Estadística*. México: Pearson Educación.
19. Walpole, R., Myers, R., Myers, S., Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias*. México: Pearson Educación.
20. Webster, A. (2000). *Estadística aplicada a los negocios y la economía*. Colombia: McGraw-Hill.
21. Wooldridge, J. (2010). *Introducción a la econometría: un enfoque moderno*. México: Thomson Learning.

