Introducción.

El objetivo de este reporte es evaluar el Producto Interno Bruto (PIB) de distintos países y la influencia que distintas variables tendrían para generar un modelo de regresión lineal que prediga su valor. Antes de iniciar la exposición quisiera dejar en claro 2 puntos que serán transversales a todo el análisis:

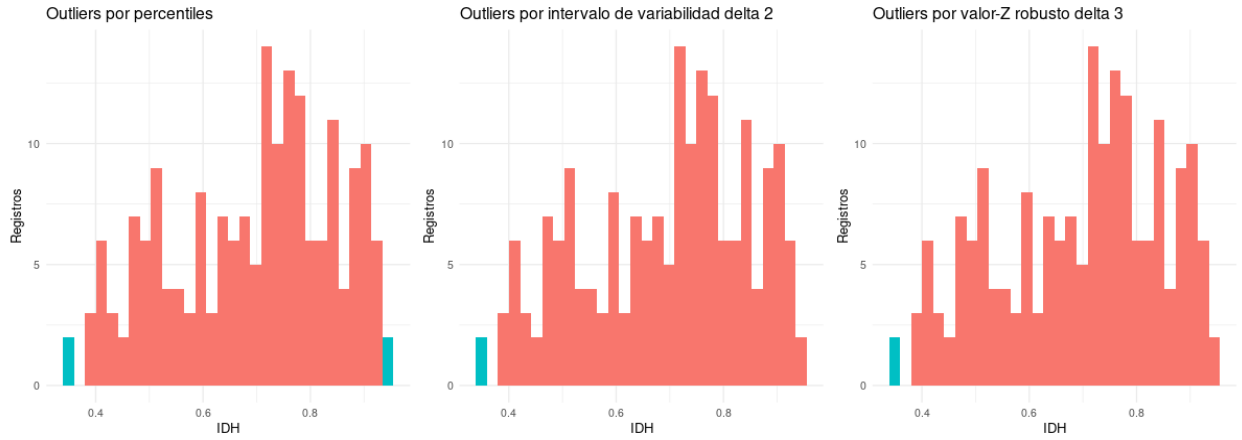
1. Por razones que se explicarán más adelante, la variable PIB se trabajará con escala logarítmica; siempre que aparezca mencionada la variable se deberá asumir así a menos que se haga una salvedad explícita.

2. Dado que los datos empleados en el análisis nos fueron facilitados sin fuente ni contexto (organización que informa, año de recolección, método de recolección, etc) su veracidad no será cuestionada, asumiendo siempre que los datos son correctos.

Presencia de valores *outliers*.

En una primera etapa nos centraremos en la detección de candidatos a *outliers* en 4 variables que me parecieron relevantes para entender el fenómeno del PIB. Mi criterio de selección estuvo basado en la correlación que estas variables presentan con el PIB.

1. Índice de Desarrollo Humano (**IDH**).



Al mirar los 3 histogramas salta a la vista el grupo de registros escindido a la izquierda, el cual es detectado por las 3 técnicas. Para la búsqueda de candidatos a *outliers* por valores que estén bajo el 1% y sobre el 99% se pudo identificar 4 registros: **Níger** y la **República Centroafricana** por ser los valores más bajos (en cian a la izquierda del gráfico), y a **Australia** y **Noruega** por ser los valores más altos (en cian a la derecha del gráfico).

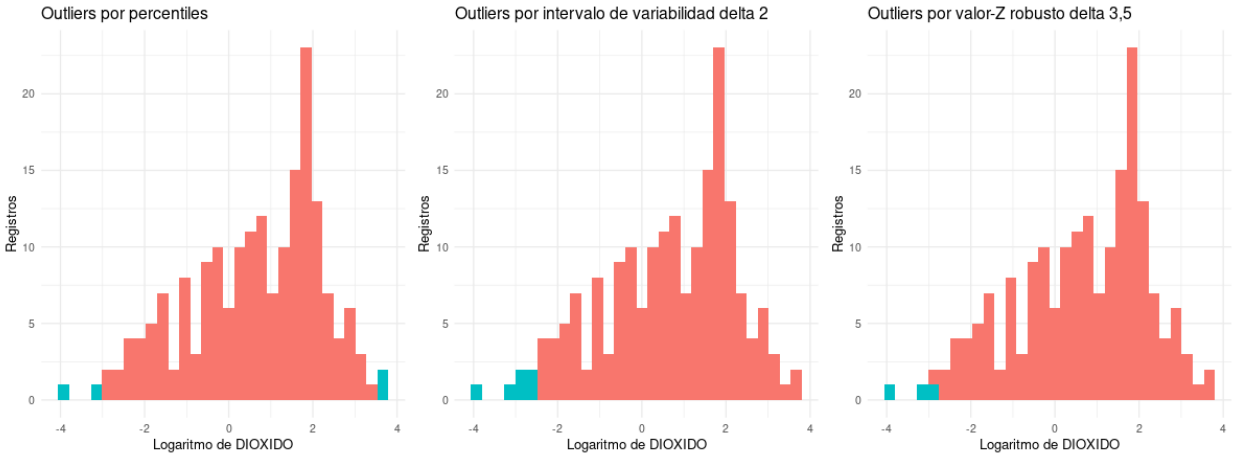
Por otro lado, para los métodos de búsqueda de candidatos a *outliers* por intervalo de variabilidad (cuyo valor de delta se fijó en 2 desviaciones estándar) y por encontrarse fuera del rango fijado por el valor-Z robusto (cuyo valor de delta se fijó

en 3) se identificaron sólo valores en el extremo inferior: **Níger** y la **República Centroafricana** (en cian a la izquierda del gráfico).

	Grupo	Promedio	Desviación_estándar	Correlación_con_PIB
1	IDH original	0.6924415	0.1547061	0.9394354
2	Percentiles 1% al 99%	0.6934875	0.1500047	0.9350782
3	Intervalo de variabilidad delta 2	0.6961323	0.1513451	0.9364372
4	Valor-Z robusto con delta 3	0.6961323	0.1513451	0.9364372

Al revisar el cambio de los estadísticos entre el dato original y los 3 métodos, no parece ser muy significativo pues tanto el promedio como la desviación estándar presentan un cambio en el tercer decimal, por lo cual no parece muy interesante excluir a los candidatos a *outlier* del análisis.

2. Logaritmo de las Emisiones de dióxido de carbono en toneladas per cápita (**DIOXIDO**).



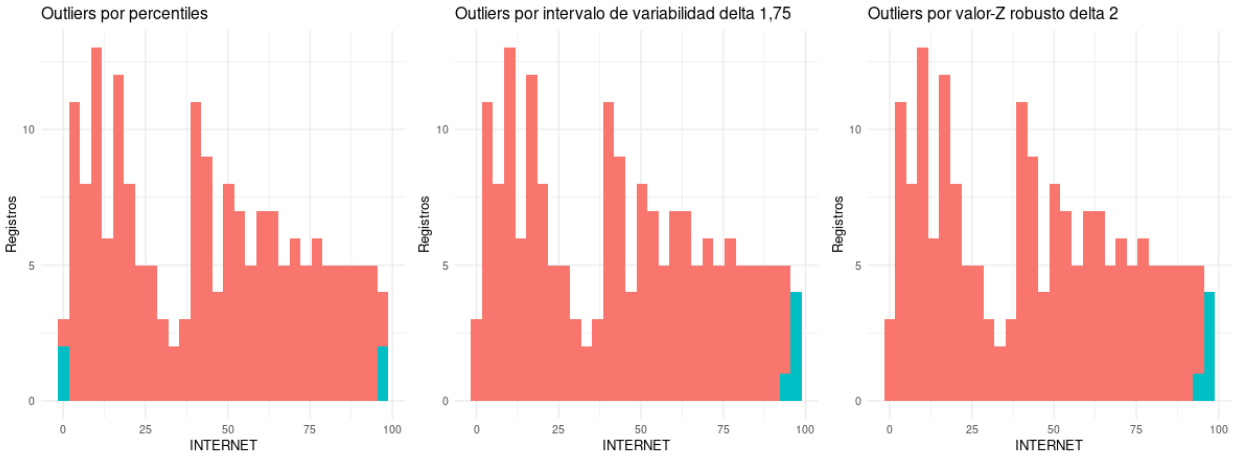
En el caso de las emisiones de dióxido de carbono he decido trabajar con el logaritmo, pues la correlación lineal con el logaritmo del PIB mejoraba significativamente. Con el método de los percentiles 1% y 99% se identificaron 4 candidatos a *outliers*: **Burundí** y **Chad** con menores emisiones (en cian a la izquierda del gráfico), y **Trinidad y Tobago** y **Qatar** por el lado de mayor emisión (en cian a la derecha del gráfico).

Tanto con el método del intervalo de variabilidad (cuyo delta se fijó en 2 desviaciones estándar) como con el método del rango entre valores-Z robustos (con valor delta fijado en 3,5) se obtuvieron sólo registros candidatos a *outliers* hacia el lado de las menores emisiones, dato que se correlaciona con la mayor densidad de datos que se ven hacia valores mayores. 6 países, **Burundí**, **Chad**, **República Democrática Del Congo**, **Rwanda**, la **República Centroafricana** y **Malawi**, fueron identificados con el primer método, mientras que el rango del valor-Z robusto sólo identificó a los 3 primeros países como candidatos.

	Grupo	Promedio	Desviación_estándar	Correlación_con_PIB
1	Logaritmo de DIOXIDO original	0.6337138	1.578118	0.9140737
2	Percentiles 1% al 99%	0.6448819	1.502518	0.9140962
3	Intervalo de variabilidad delta 2	0.7532314	1.455174	0.9066636
4	Valor-Z robusto con delta 3.5	0.6971610	1.508302	0.9124330

Al comparar las técnicas se observa que el mayor cambio se produce con la exclusión de los candidatos a *outliers* identificados por el intervalo de variabilidad, tanto para promedio como para desviación estándar como par correlación con el logaritmo del PIB, quizás por el hecho de que se eliminaría un mayor número de registros. Sin embargo la correlación no mejora, por lo que no considero significativo para continuar dejar afuera ningún valor.

3. Porcentaje de uso de internet (**INTERNET**).



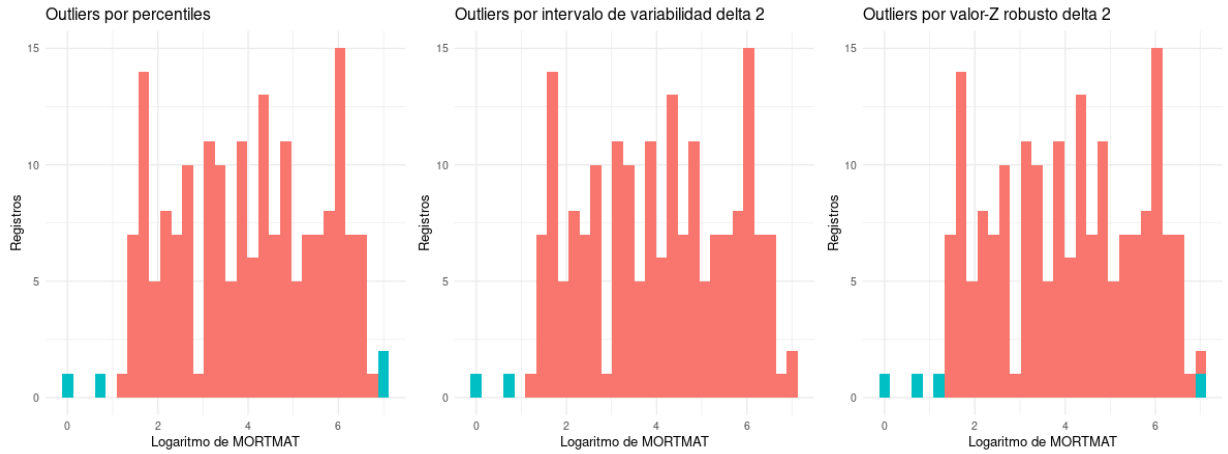
En el caso del Porcentaje de uso de internet, el método de los percentiles 1% y 99%, como es esperable dado el número de registros, vuelve a identificar 4 candidatos a *outliers*: **Eritrea** y **Timor Oriental** como los que tienen menor uso de internet (en cian a la izquierda del gráfico), y **Noruega** e **Islandia** como aquellos candidatos con un mayor uso (en cian a la derecha del gráfico).

En esta ocasión los métodos del intervalo de variabilidad (con delta fijado en 1,75 desviaciones estándar) y del rango entre valores-Z robustos (con delta fijado en 2) entregaron 5 registros sólo por el lado del mayor uso de internet (en cian a la derecha de los respectivos gráficos): **Liechtenstein**, **Andorra**, **Dinamarca**, **Noruega** e **Islandia**.

	Grupo	Promedio	Desviación_estándar	Correlación_con_PIB
1	INTERNET original	44.11293	28.92764	0.8797536
2	Percentiles 1% al 99%	44.00348	28.35472	0.8747101
3	Intervalo de variabilidad delta 1.75	42.68672	27.97981	0.8752167
4	Valor-Z robusto con delta 2	42.68672	27.97981	0.8752167

Comparando los métodos se observa una mayor diferencia con los promedios, desviación estándar y correlaciones obtenidas con el intervalo de variabilidad y el valor-Z robusto, pero que no impresiona ser significativa, por lo que dejaré estos valores dentro del cálculo del modelo.

4. Logaritmo de la Tasa de mortalidad materna por cada 100 nacimientos (**MORTMAT**).



Para el análisis de la Tasa de mortalidad materna también me decanté por el uso de la escala logarítmica, pues también mejoraba la correlación con respecto al logaritmo del PIB. Al mirar el histograma resaltan inmediatamente un par de registros con menor mortalidad materna que parecen buenos candidatos a valores *outliers*. Este par es bien identificado por el método del intervalo de variabilidad con un delta de 2 desviaciones estándar: **Bielorrusia** e **Israel** (en cian a la izquierda del gráfico).

Con los métodos de los percentiles 1% y 99% y del rango entre valores-Z robustos (con delta fijado en 2) se identifican 4 registros: 2 a cada lado de la escala, en el caso del primero, y 3 a la izquierda y 1 a la derecha para el segundo (en color cian en los respectivos gráficos): **Bielorrusia** e **Israel** por el lado de la menor mortalidad materna y **Chad** y **Sierra Leona** con mayor mortalidad materna, con los percentiles; mientras que con el valor-Z robusto tenemos a **Bielorrusia**, **Israel** y **Polonia** por la izquierda y a **Sierra Leona** por la derecha.

	Grupo	Promedio	Desviación_estándar	Correlación_con_PIB
1	Logaritmo de MORTMAT original	4.007415	1.631808	-0.8436535
2	Percentiles 1% al 99%	4.015273	1.574418	-0.8494702
3	Intervalo de variabilidad delta 2	4.046779	1.594943	-0.8526013
4	Valor-Z robusto con delta 2	4.046735	1.573656	-0.8524138

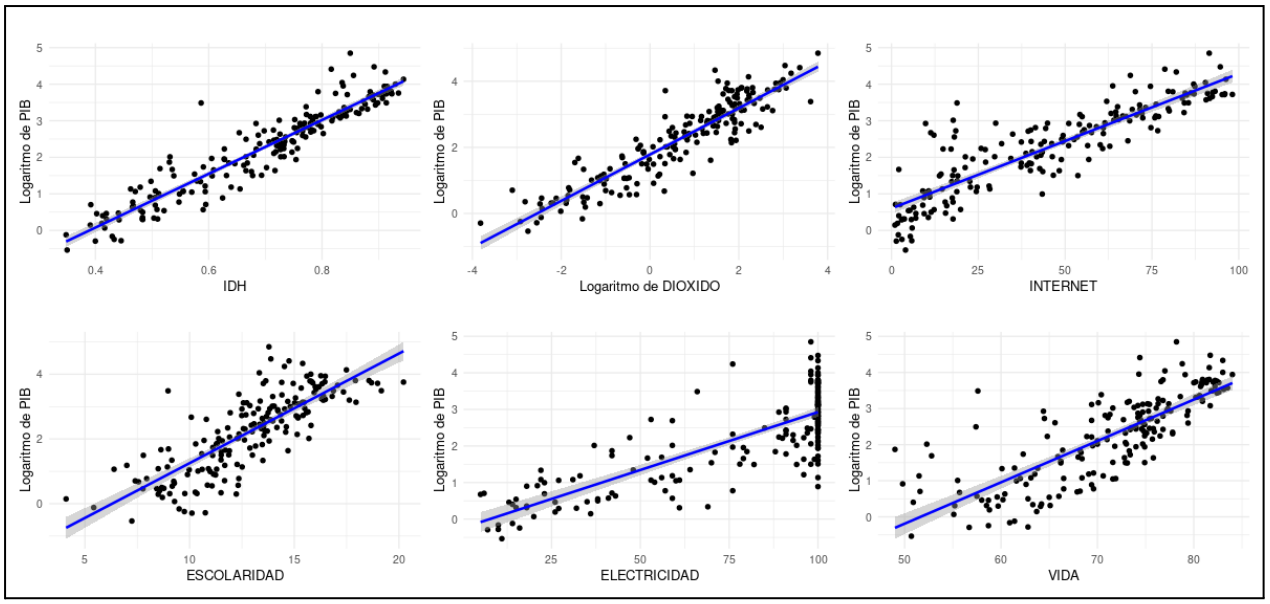
Finalmente, al comparar los métodos se observa un cambio de la media hacia valores sutilmente mayores, acompañado por una leve disminución de la desviación estándar, especialmente con el intervalo de variabilidad lo cual, asociado a una leve mejora de la correlación con el logaritmo del PIB, los hace candidatos a *outliers* a ser considerados al momento de armar el modelo final.

Hipótesis y Análisis preliminar de los datos.

Para el diseño del modelo predictor de valores del PIB estimé que se utilizarían aquellas variables con mayor nivel de correlación, independientemente del signo de éstas. Con esta idea en mente he revisado la correlación entre todas las variables, tanto en su escala original como en su escala logarítmica, contra los valores del PIB en escala logarítmica. Las razones para usar la escala logarítmica en el PIB son:

- En el gráfico de dispersión se observa una mayor tendencia a la agregación alrededor de una recta con la mayoría de las variables.
- Los valores absolutos obtenidos en los cálculos de correlación son mayores con la mayoría de las variables.

Finalmente me he quedado con 11 variables que presentan una correlación absoluta mayor a 0,7, valor que fue definido arbitrariamente como significativo. De estas 11 variables con alta correlación 6 tienen signo positivo (a mayor valor en la variable mayor será el valor en el PIB), y 5 presentan signo negativo (a mayor valor en la variable menos valor en el PIB). Con esto en mente se pueden plantear las siguientes hipótesis:



1. *El Índice de desarrollo humano (**IDH**) estará presente en el modelo con un argumento positivo*: esto pues es la variable que presenta la mayor correlación positiva con el logaritmo del **PIB** (0,9394), además de que tiene el **PIB** per cápita como parte de su cálculo, junto con el Índice de educación e Índice de esperanza de vida.

2. *El logaritmo de las emisiones de dióxido de carbono (**DIOXIDO**) estará presente en el modelo con un argumento positivo*: esto pues tiene una fuerte correlación positiva con el logaritmo del **PIB** (0,9141). Además, es de esperar que las sociedades industrializadas y con mayor cantidad de fábricas tengan más ingresos, junto con el

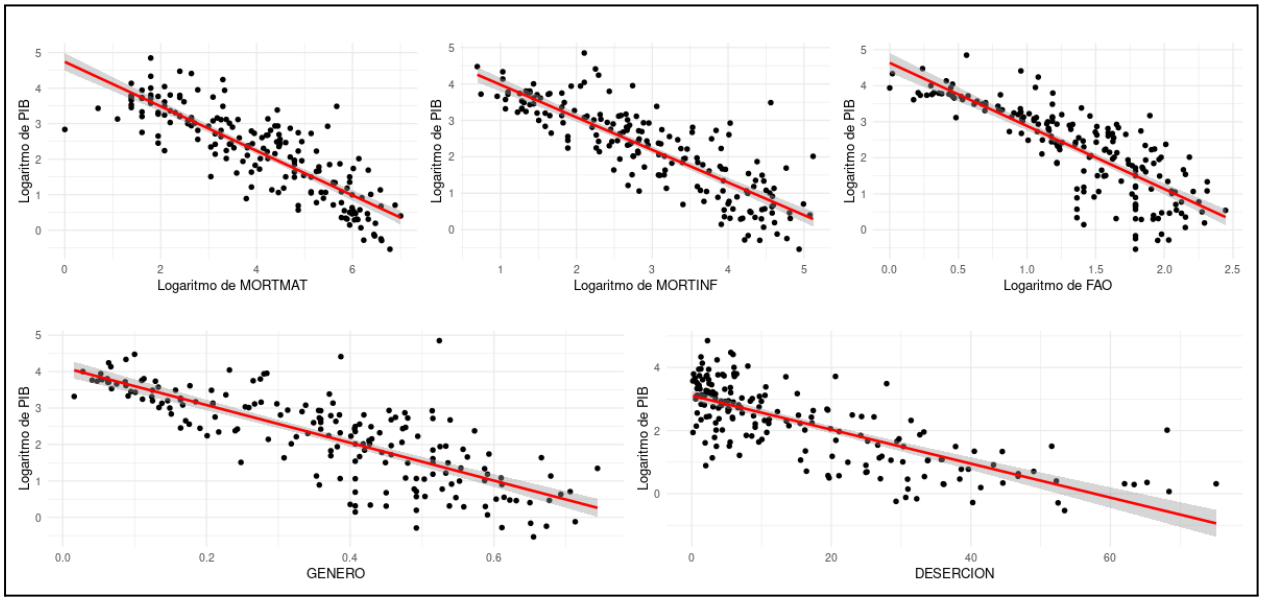
hecho de que la quema de combustibles fósiles sigue siendo parte importante de la generación de electricidad en muchos países.

3. *El porcentaje de uso de internet (**INTERNET**) estará presente en el modelo con un argumento positivo*: esto pues tiene una alta correlación positiva con el logaritmo del **PIB** (0,8798), junto al hecho de que internet es fuente de emprendimientos y negocios que pueden mover millones en la economía, a la vez que se convierte en la principal fuente de ocio importante para las sociedades con mayor nivel de desarrollo.

4. *La cantidad de años de escolaridad (**ESCOLARIDAD**) estará presente en el modelo con un argumento positivo*: esto pues presenta una alta correlación positiva con el logaritmo del PIB (0,8019), además de que las poblaciones con mayor escolaridad presentan una mayor cantidad de mano de obra calificada, asociando una mayor generación de riqueza.

5. *La tasa de electrificación (**ELECTRICIDAD**) estará presente en el modelo con un argumento positivo*: esto pues muestra una significativa correlación positiva con el logaritmo del PIB (0,7998), además de que una sociedad a mayores ingresos se espera que presente un mayor uso de electricidad como fuente de energía.

6. *La expectativa de años de vida (**VIDA**) estará presente en el modelo con un argumento positivo*: esto pues posee una significativa correlación positiva con el logaritmo del PIB (0,7930), junto al hecho de que, al tener mayor capital, las sociedades tienden a cuidar más de las personas, con lo que se espera que alcancen su potencial de longevidad.



7. *El logaritmo de la mortalidad materna (**MORTMAT**) estará presente en el modelo con un argumento negativo*: esto pues presenta la mayor de las correlaciones negativas con el logaritmo del PIB (-0,8437), además de tener sentido que las sociedades que cuenten con mayores recursos pueden destinar más esfuerzos para mantener bien

los niveles de salud de la población, disminuyendo las muertes relacionadas al embarazo.

8. *El logaritmo de la mortalidad infantil (**MORTINF**) estará presente en el modelo con un argumento negativo*: esto pues presenta una de las mayores correlaciones negativas con el logaritmo del PIB (-0,8402), junto con el que, al igual que con **MORTMAT**, tiene sentido que las sociedades que cuenten con mayores recursos pueden destinar más esfuerzos para mantener bien los niveles de salud de la población, disminuyendo la muerte entre sus niños.

9. *El logaritmo del índice de precios de la FAO (**FAO**) estará presente en el modelo con un argumento negativo*: esto pues presenta una de las mayores correlaciones negativas con el logaritmo del PIB (-0,8077), además de hacer sentido que un país con mayores ingresos destine relativamente menos recursos en alimentar a su población con alimentos básicos e invierta en otro tipo de bienes de consumo.

10. *El índice de desigualdad de género (**GENERO**) estará presente en el modelo con un argumento negativo*: esto pues presenta una correlación negativa significativa con el logaritmo del PIB (-0,7833) junto al hecho de que con menos desigualdad de género más probabilidad de que se generen ingresos por incorporar a la fuerza de trabajo a casi la mitad de la población que, de otra forma, estaría ocupada con labores domésticas.

11. *La tasa de deserción en educación primaria (**DESERCION**) estará presente en el modelo con un argumento negativo*: esto pues presenta una correlación negativa significativa con el logaritmo del PIB (-0,7438) acompañado al sentido que hace el que a menor deserción es esperable una mayor presencia de mano de obra calificada con un mayor impacto en los ingresos, a modo inverso con **EDUCACION**.

	IDH	DIOXIDO_log	INTERNET	ESCOLARIDAD	ELECTRICIDAD	VIDA	MORTMAT_log	MORTINF_log	FAO_log	GENERO	DESERCION
IDH	1.0000000	0.8760986	0.9077456	0.9124004	0.8496588	0.8997976	-0.9087893	-0.9286382	-0.8056136	-0.8330626	-0.7952043
DIOXIDO log	0.8760986	1.0000000	0.8093840	0.7529580	0.8045748	0.7195762	-0.7973560	-0.7723550	-0.7107002	-0.7079937	-0.7644801
INTERNET	0.9077456	0.8093840	1.0000000	0.8094097	0.7289044	0.8073674	-0.8581904	-0.8891018	-0.8331220	-0.8026999	-0.6923412
ESCOLARIDAD	0.9124004	0.7529580	0.8094097	1.0000000	0.7352297	0.7866698	-0.8172378	-0.8461822	-0.7111642	-0.7395079	-0.6773373
ELECTRICIDAD	0.8496588	0.8045748	0.7289044	0.7352297	1.0000000	0.8045639	-0.7794173	-0.7708260	-0.5911597	-0.6565330	-0.7985142
VIDA	0.8997976	0.7195762	0.8073674	0.7866698	0.8045639	1.0000000	-0.8632219	-0.9211800	-0.7028268	-0.7605494	-0.7249872
MORTMAT log	-0.9087893	-0.7973560	-0.8581904	-0.8172378	-0.7794173	-0.8632219	1.0000000	0.9296592	0.7474619	0.8727626	0.7265503
MORTINF log	-0.9286382	-0.7723550	-0.8891018	-0.8461822	-0.7708260	-0.9211800	0.9296592	1.0000000	0.7831427	0.8556239	0.7209420
FAO_log	-0.8056136	-0.7107002	-0.8331220	-0.7111642	-0.5911597	-0.7028268	0.7474619	0.7831427	1.0000000	0.7660375	0.5845942
GENERO	-0.8330626	-0.7079937	-0.8026999	-0.7395079	-0.6565330	-0.7605494	0.8727626	0.8556239	0.7660375	1.0000000	0.6106668
DESERCION	-0.7952043	-0.7644801	-0.6923412	-0.6773373	-0.7985142	-0.7249872	0.7265503	0.7209420	0.5845942	0.6106668	1.0000000

Finalmente, sirva como apoyo a todas las hipótesis planteadas el que, al revisar las correlaciones entre las 11 variables mencionadas podemos observar que existe concordancia entre variables con el mismo signo del argumento: aquellas que tienen correlación positiva con el logaritmo del PIB también presentan correlación positiva entre ellas (en la tabla de arriba, el cuadrante superior izquierdo); aquellas con correlación negativa con el logaritmo del PIB tienen entre ellas correlación positiva (cuadrante inferior derecho). En los otros escenarios la correlación es negativa.

Propuesta de modelo final.

Para la construcción del modelo final quise utilizar 3 enfoques distintos:

- Modelamiento con todas las variables, pero sin transformación de escala.
- Modelamiento sólo con las variables identificadas en el punto anterior.
- Modelamiento con todas las variables con logaritmo del PIB y logaritmo de otras variables, cuando corresponda.

Modelamiento con todas las variables sin transformación

En cuanto al modelo utilizando las variables sin transformación en escala logarítmica, el valor de $\mathbf{R^2}$ obtenido fue de 88,8%, con un $\mathbf{R^2}$ ajustado de 86,83% (el cual mejora hasta el 87,38% si es que se consideran sólo las variables que contribuyen al modelo con un 90% de intervalo de confianza) y una correlación de 0.9424, valores que son bastante buenos, pero aún así inferiores a los obtenidos con las variables transformadas, por lo que esta vía fue descartada.

Modelamiento con variables seleccionadas incluyendo transformaciones

Por otro lado, al trabajar sólo con las variables identificadas en el punto anterior como hipotéticamente relevantes, se aplicó transformación de escala tanto al **PIB** como a **DIOXIDO**, **MORTMAT**, **MORTINF** y **FAO**. Con estos cambios se logró inmediatamente una mejora en los valores: $\mathbf{R^2}$ de 94,27%, $\mathbf{R^2}$ ajustado de 93,92% y una correlación de 0,979. El mejor valor de $\mathbf{R^2}$ ajustado se obtuvo al seleccionar sólo las 6 variables que ayudaban a explicar el modelo con un 95% de intervalo de confianza: 93,98% con **IDH**, **ESCOLARIDAD**, **VIDA**, **DESERCION**, el logaritmo de **DIOXIDO** y el logaritmo de **FAO**. Sin embargo estos valores aún pueden ser mejorados si iniciamos el modelo sin una selección previa de las variables a utilizar.

Modelamiento con todas las variables incluyendo transformaciones

El enfoque que resultó más fructífero fue el de iniciar la búsqueda de con todas las variables, transformando a escala logarítmica tanto el **PIB** como aquellas variables que presentaron mejor correlación y mejor comportamiento lineal en los gráficos de dispersión. Antes que todo, quisiera señalar cómo se utilizarían las variables:

- 11 variables que se trabajaron en su escala normal y que se esperaba que tuvieran argumentos con signo positivo: **IDH**, **INTERNET**, **ESCOLARIDAD**, **ELECTRICIDAD**, **VIDA**, **CELULAR**, **INMIGRANTES**, **FOSIL**, **TURISMO**, **PARLAMENTO** y **BOSQUE**.

- 8 variables que se trabajaron en su escala normal y que se esperaba que tuvieran argumentos con signo negativo: **GENERO**, **DESERCION**, **RENOVABLE**, **VIOLENCIA**, **SUICIDIOFEM**, **HOMICIDIO**, **AFECTADOS** y **DESASTRE**.

- 2 variables que se trabajaron en escala logarítmica y que se esperaba que tuvieran argumentos con signo positivo: **DIOXIDO** y **PRISION**.

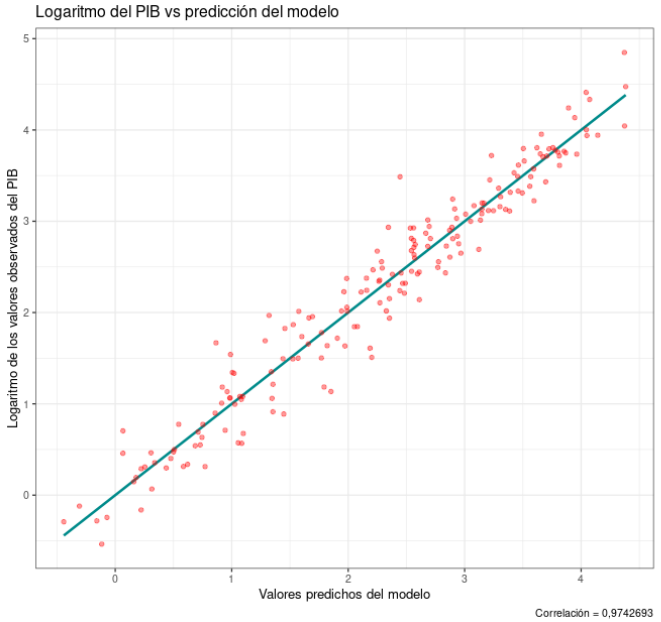
- 7 variables que se trabajaron en escala logarítmica y que se esperaba que tuvieran argumentos con signo negativo: **MORTMAT**, **MORTINF**, **FAO**, **GINI**, **IPC**, **SUICIDIOMAS** y **POB**.

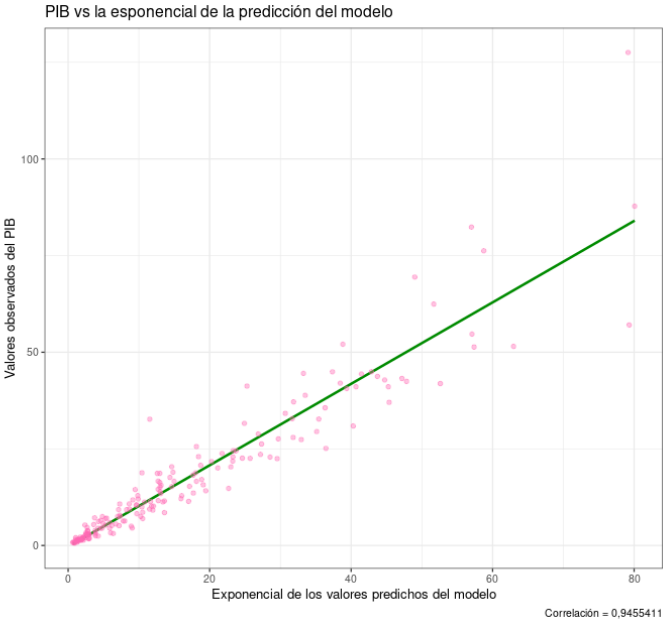
Al realizar la primera estimación del modelo con las variables anteriormente mencionadas se obtuvo valores aún mejores que los anteriormente informados: $\mathbf{R^2}$ de 95,3%, $\mathbf{R^2}$ ajustado de 94,47% y correlación de 0,9762. Sin embargo este candidato a modelo presentaba muchas variables cuyo aporte no parecía significativo, por lo que se decidió empezar a descartar variables.

Luego de múltiples intentos se escogió el candidato a modelo que se obtiene al descartar aquellas variables que para un intervalo de confianza del 95% no estuviesen realmente aportando, según el test-t de la primera estimación. Es así como quedaron sólo 8 variables: **IDH**, **ESCOLARIDAD**, **VIDA**, **CELULAR**, **INMIGRANTES**, **DESERCION**, logaritmo de **DIOXIDO** y logaritmo de **FAO**. expresados en la siguiente ecuación de regresión lineal:

$$\hat{Y} = -0{,}238 + 8{,}532 \times \boldsymbol{PIB} - 0{,}129 \times \boldsymbol{ESCOLARIDAD} - 0{,}028 \times \boldsymbol{VIDA} + 0{,}003 \times \boldsymbol{CELULAR} + 0{,}006 \times \boldsymbol{DESERCION} + 0{,}199 \times \log(\boldsymbol{DIOXIDO}) - 0{,}239 \times \log(\boldsymbol{FAO})$$

Con este modelo se obtiene un $\mathbf{R^2}$ de 94,92%, pero con un aumento del $\mathbf{R^2}$ ajustado, logrando este modelo explicar el 94,69% de la variabilidad del logaritmo del **PIB**, con una correlación de 0,9743 al comparar el predicho del modelo con el logaritmo del **PIB**. Además, al realizar el test-t con estas variables se obtiene que todas ellas aportan al modelo de manera significativa con un intervalo de confianza mayor al 99%. Llama la atención que el intercepto sí se encuentre dentro del intervalo de error (valor-t -0,959), lo cual en ningún caso significa que se pueda prescindir de él para la construcción de este modelo.





Finalmente, al revisar nuestras hipótesis planteadas vemos que sólo se ha podido verificar 3 de ellas:

- Las hipótesis 1 (*El Índice de desarrollo humano (**IDH**) estará presente en el modelo con un argumento positivo*), 2 (*El logaritmo de las emisiones de dióxido de carbono (**DIOXIDO**) estará presente en el modelo con un argumento positivo*) y 9 (*El logaritmo del índice de precios de la FAO (**FAO**) estará presente en el modelo con un argumento negativo*) se verifican según lo planteado.

- Las variables mencionadas en las hipótesis 4 (*La cantidad de años de escolaridad (**ESCOLARIDAD**) estará presente en el modelo con un argumento positivo*), 6 (*La expectativa de años de vida (**VIDA**) estará presente en el modelo con un argumento positivo*) y 11 (*La tasa de deserción en educación primaria (**DESERCION**) estará presente en el modelo con un argumento negativo*), si bien son incluidas en el modelo, llamativamente presentan argumentos con signos opuestos al esperado. Si bien uno se ve tentado de sacarlas para intentar mejorar el modelo, el hecho es que no mejoran los valores, aunque no los empeoran de forma tan relevante, quedando los valores de $\mathbf{R^2}$ ajustado entre 93% y 92%. Para las 2 primeras, **ESCOLARIDAD** y **VIDA**, podría deberse a un efecto modulador sobre **IDH**, recordando que ambas están de alguna forma incluidas en el cálculo de éste. En cuanto a **DESERCION**, quizás sólo se comporta de forma opuesta a **ESCOLARIDAD**.

- Las variables de las otras 5 hipótesis (**INTERNET**, **ELECTRICIDAD**, logaritmo de **MORTMAT**, logaritmo de **MORTINF** y **GENERO**) no entran en el modelo final porque no logran significancia al lado de las otras variables.

- Llamativo es el caso de **CELULAR** pues presentaba una correlación más baja al compararla con el logaritmo del **PIB** (0,6571), si bien de signo positivo, como finalmente presenta el argumento en el modelo. Ya sea como herramienta de trabajo o como importante fuente de ocio y estatus, es llamativo como el teléfono celular termina siendo un indicador importante para estimar la cantidad de riqueza de un país.

Evaluación 1
Curso Minería de datos
Néstor Patricio Rojas Ríos