# Evaluación 1 - Minería de Datos

Pablo Herrera Gálvez

2022-07-17

### Contexto.

- Se cuenta con información de 28 potenciales variables explicativas del PIB per capita para 188 países.
- Se realiza un análisis de valores atípicos u *outliers*.
- Posteriormente se plantean posibles modelos predictivos del producto de cada país.
- Por último se estiman tres modelos de regresión lineal, cada uno con sus consideraciones.

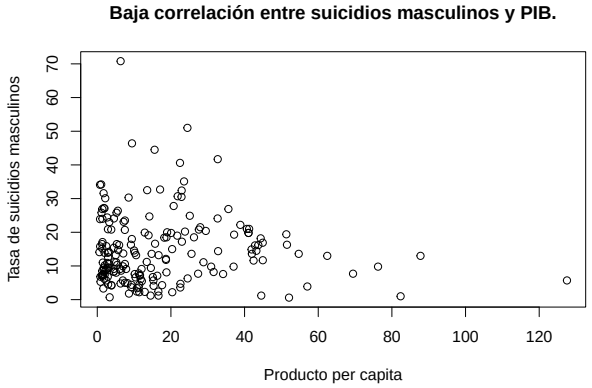
## I. Revisión de valores atípicos u outliers.

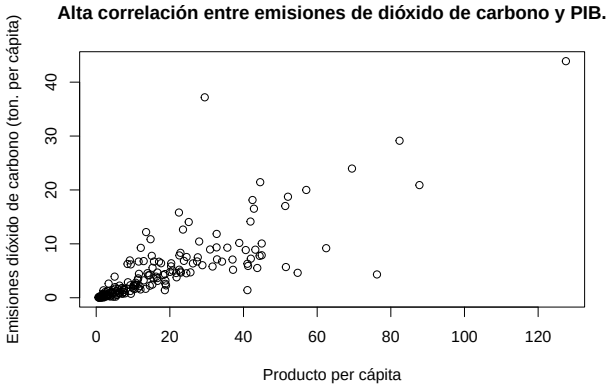
Antes de comenzar con el análisis de valores atípicos se realiza una primera revisión del total de variables contenidas en la base de datos. Se estima el coeficiente de correlación de Pearson **entre el PIB per cápita y las 28 variables numéricas restantes.**

Table 1: Variables con baja correlación con PIB per cápita.

| VARIABLE | CORRELACIÓN |
|---|---|
| AFECTADOS | -0.087 |
| BOSQUE | -0.062 |
| DESASTRE | -0.049 |
| POB | -0.045 |
| SUICIDIOMAS | -0.025 |
| PRISION | 0.071 |

A partir de esto, se identifican 6 variables que presentan un coeficiente de correlación reducido (entre -0.1 y 0.1). Para ilustrar, se grafica un caso de baja correlación y otro de alta correlación positiva.




Posteriornmente, se analizaron gráficamente las 22 variables restantes en busca de valores atípicos. Con esto se escogieron 5 variables que podrían presentar *outliers*. Estas cinco son el *PIB* más 4 potenciales variables explicativas: *IPC, HOMICIDIO, RENOVABLE* y *DIOXIDO*.

Para ilustrar el proceso se incluye la distribución de densidad del 50% superior del PIB y de la tasa de homicidios, lo que permite identificar valores anómalos en el extremo superior de ambas distribuciones. En el caso del producto, Qatar es el país que más se aleja del resto de la distribución, mientras que para el caso de la tasa de homicidios, sobresale Honduras.

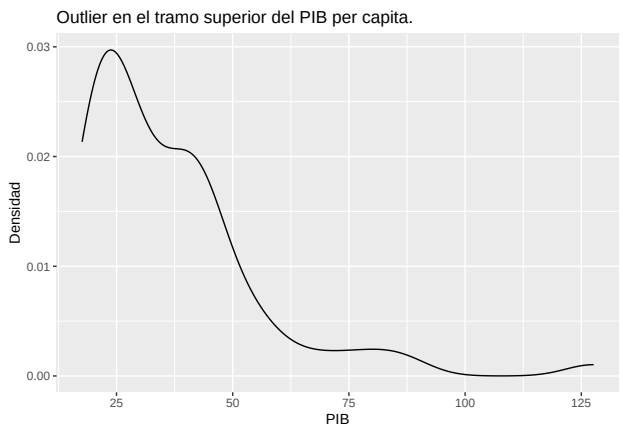




Table 2: Catar lidera a los países con mayor producto per capita.

| PAIS | PIB |
|---|---|
| Qatar | 127.543 |
| Luxembourg | 87.772 |
| Kuwait | 82.369 |
| Singapore | 76.240 |
| Brunei | 69.450 |

Table 3: Honduras sobresale entre los países con mayor tasa de homicidio.

| | PAIS | HOMICIDIO |
|---|---|---|
| 137 | Honduras | 90.4 |
| 66 | Venezuela | 53.7 |
| 109 | Belize | 44.7 |
| 112 | El Salvador | 41.2 |
| 116 | Guatemala | 39.9 |

A las 5 variables mencionadas se les aplican tres métodos de detección de valores atípicos:

- **Método 1. Percentiles.** Como regla, se determinarán como *outliers* todos aquellos valores que formen parte del primer y último percentil de la distribución de cada una de las cinco variables.
- **Método 2. Intervalos de variabilidad.** Dicho intervalo se centra en la media de la variable y sus límites vendrán definidos por un parámetro delta y su desviación estándar. Para el ejercicio se define que el deltra tendrá valor 3.

$$\bar{X} \pm \delta \cdot s = \bar{X} \pm 3 \cdot s$$

- **Método 3. Valor-z robusto.** Analiza la desviación de los datos con respecto a la media. Tiene la particularidad de que el resultado no depende de la cantidad de datos, como ocurre en el Método 1, ni tampoco de estadísticos que se ven afectados por los valores extremos, como ocurre con la media y la desviación estándar del Método 2.

Considerando como outlier sólo aquellas observaciones que son identificadas usando los tres métodos, *RENOVABLE* presenta un outlier y el resto de las variables presentan dos observaciones atípicas cada una.

Estos resultados serán considerados en el modelo final del apartado III. Se realizarán estimaciones con todas las observaciones y luego, excluyendo los ocho países que presentan al menos un outlier en las variables estudiadas.

# II. Planteamiento de hipótesis preliminares.

Del total de 28 variables, se descartaron 6. Con esto, se mantienen 22 variables que podrían ser relevantes para la estimación del PIB. Estas se organizan de acuerdo a las siguientes 5 categorías: **económicas, sociales, salud, tecnología y medio ambiente.**

- **Variables económicas:** incluyen información de desigualdad, inflación, e información de turismo internacional.
  - GINI
  - IPC
  - FAO
  - TURISMO

- **Variables sociales:** incluyen índice de desarrollo humano, violencia, indicadores de género, educación y población inmigrante.
  - IDH
  - HOMICIDIO
  - VIOLENCIA
  - GENERO
  - PARLAMENTO
  - DESERCION
  - ESCOLARIDAD
  - INMIGRANTES

- **Variables de salud:** incluyen tasa de suicidio femenino, tasas de mortalidad infantil y maternal, y expectativa de vida al nacer.
  - SUICIDIOFEM
  - MORTINF
  - MORTMAT
  - VIDA

- **Variables de tecnologia y desarrollo:** incluyen tasa de electrificación, uso de internet y telefonía.
  - ELECTRICIDAD
  - INTERNET
  - CELULAR

- **Variables de medio ambiente:** describen uso de recursos renovables y fósiles además de emisiones de dióxido de carbono.
  - RENOVABLE
  - FOSIL
  - DIOXIDO

**Discusión sobre las variables a incluir en el modelo:**

En primer lugar, se define excluir el Índice de Desarrollo Humano (IDH) debido a que su información ya estaría contenida en otras variables. El IDH concentra información sobre salud, educación y riqueza. Los primeros dos items ya existen en otras variables explicativas, mientras que la riqueza se mide en términos del PIB per cápita (variable explicada).

En segundo lugar se identifican algunos pares de variables que resultan complementarios además de presentar una fuerte correlación entre ellos. Por esto se determina mantener una

sola en cada caso. Un ejemplo de esta situación es la mortalidad de niños(as) menores de 5 años y la mortalidad maternal en partos. Ante esto, se mantiene el indicador de mortalidad maternal por presentar un mayor coeficiente de variación. El segundo caso en que esto ocurre es con el porcentaje de uso de energías renovables y combustibles fósiles. Para este caso se mantiene el indicador de energía renovable por su mayor coeficiente de variación.

A modo de hipótesis, se espera que indicadores de mayor inflación estén asociados a una menor actividad económica, así como también se espera lo mismo ante indicadores más pobres en materia de igualdad de género, violencia y calidad de educación y salud. Por último, es razonable pensar que a mayores niveles de tecnología y desarrollo se observe un mayor producto.

# III. Propuesta y estimación de un modelo de regresión lineal.

Ante las definiciones del apartado anterior, se tomarán en consideración las siguientes variables explicativas:

- **Variables económicas:**

  GINI, IPC, FAO y TURISMO.

- **Variables sociales:**

  HOMICIDIO, VIOLENCIA, GENERO, PARLAMENTO, DESERCION, ESCOLARIDAD e INMIGRANTES.

- **Variables de salud:**

  SUICIDIOFEM, MORTMAT y VIDA.

- **Variables de tecnologia y desarrollo:**

  ELECTRICIDAD, INTERNET y CELULAR.

- **Variables de medio ambiente:**

  RENOVABLE y DIOXIDO.

Con esto se realizarán tres estimaciones en total:

- **Estimación 1:** Incluirá las 19 variables para realizar una estimación lineal del PIB per cápita.
- **Estimación 2:** Se modifica la primera estimación aplicando logaritmo al PIB per capita.
- **Estimación 3:** Se replica la estimación 2 pero excluyendo los países con outliers identificados en el apartado I.

## Estimación 1.

Estimación modelo lineal del PIB per capita.

Este modelo arroja un R cuadrado de **0.863**. Además la correlación entre el PIB y el PIB predicho por el Modelo 1 es de **0.929**.

Table 4: Estimación Modelo 1.

| Variable | Coeficiente | Significancia |
|---|---|---|
| (Intercept) | -26.871051 | *** |
| CELULAR | 0.011984 | |
| DESERCION | 0.040720 | |
| DIOXIDO | 1.332103 | |
| ELECTRICIDAD | -0.033129 | *** |
| ESCOLARIDAD | 0.275258 | |
| FAO | -1.384547 | *** |
| GENERO | -4.391556 | *** |
| GINI | -0.043348 | *** |
| HOMICIDIO | -0.047526 | *** |
| INMIGRANTES | 0.401158 | |
| INTERNET | 0.063740 | |
| IPC | -0.011725 | *** |
| MORTMAT | 0.010295 | |
| PARLAMENTO | 0.068042 | |
| RENOVABLE | 0.072733 | |
| SUICIDIOFEM | -0.037522 | *** |
| TURISMO | 0.050207 | |
| VIDA | 0.523044 | |
| VIOLENCIA | -0.086837 | *** |

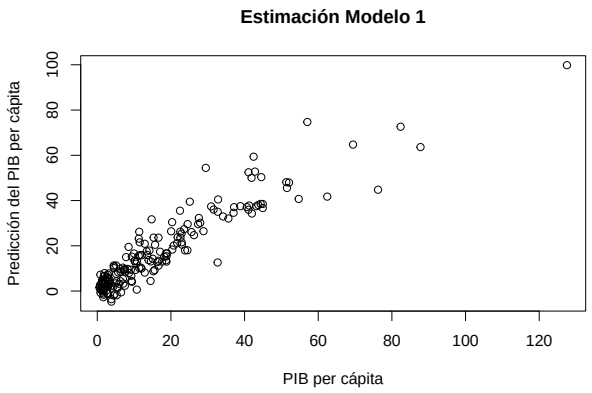
## Estimación 2.

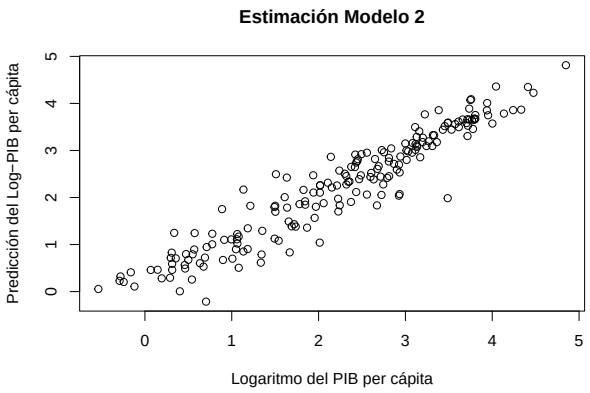
Estimación modelo lineal del logaritmo del PIB per capita.

Este segundo modelo arroja un R cuadrado de **0.904**. Además la correlación entre el PIB y el PIB predicho por el Modelo 2 es de **0.951**.

Table 5: Estimación Modelo 2.

| Variable | Coeficiente | Significancia |
|---|---|---|
| (Intercept) | 1.316697 | |
| CELULAR | 0.002600 | |
| DESERCION | -0.002910 | *** |
| DIOXIDO | 0.035059 | |
| ELECTRICIDAD | 0.005953 | |
| ESCOLARIDAD | 0.039964 | |
| FAO | -0.059302 | *** |
| GENERO | -0.781748 | *** |
| GINI | 0.001730 | |
| HOMICIDIO | 0.003221 | |
| INMIGRANTES | 0.007057 | |
| INTERNET | 0.008995 | |
| IPC | 0.000006 | ** |
| MORTMAT | -0.000467 | *** |
| PARLAMENTO | 0.002037 | |
| RENOVABLE | -0.000983 | *** |
| SUICIDIOFEM | 0.001448 | |
| TURISMO | 0.002116 | |
| VIDA | -0.002437 | *** |
| VIOLENCIA | -0.006525 | *** |



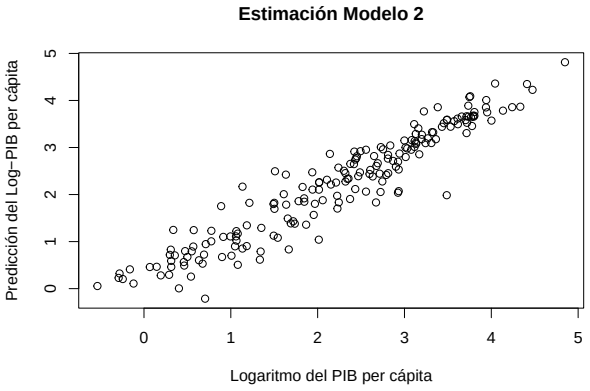
## Estimación 3.

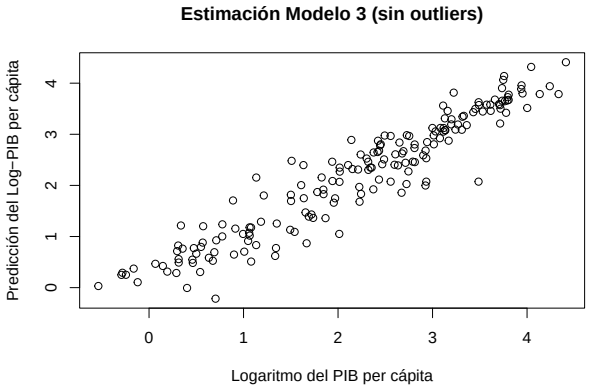
Estimación modelo lineal del logaritmo del PIB per capita excluyendo los 8 países que presentan outliers. Los 8 países excluídos en esta estimación son: Qatar, Luxemburgo, Belarus, Sudan, Venezuela, Honduras, Paraguay y Trinidad y Tobago.

El tercer modelo arroja un R cuadrado de **0.901**. Además la correlación entre el logaritmo del PIB y el logaritmo del PIB predicho por el Modelo 3 es de **0.949**.

Table 6: Estimación Modelo 3.

| Variable | Coeficiente | Significancia |
|---|---|---|
| (Intercept) | 1.165840 | |
| CELULAR | 0.002589 | |
| DESERCION | -0.002656 | *** |
| DIOXIDO | 0.042120 | |
| ELECTRICIDAD | 0.006080 | |
| ESCOLARIDAD | 0.043401 | |
| FAO | -0.048948 | *** |
| GENERO | -0.846974 | *** |
| GINI | 0.000606 | |
| HOMICIDIO | 0.006648 | |
| INMIGRANTES | 0.005591 | |
| INTERNET | 0.008496 | |
| IPC | -0.000279 | *** |
| MORTMAT | -0.000398 | *** |
| PARLAMENTO | 0.002029 | |
| RENOVABLE | -0.001288 | *** |
| SUICIDIOFEM | 0.001455 | |
| TURISMO | 0.002013 | |
| VIDA | -0.000751 | *** |
| VIOLENCIA | -0.006379 | *** |



## Interpretación.

- Con el paso del modelo 1 al modelo 2 y la aplicación del logaritmo al PIB per capita, se obtienen resultados lineales. Esta modificación también trae consigo un aumento del R cuadrado ajustado y de la correlación entre la variable dependiente y su valor predicho.
- Luego con el paso del modelo 2 al modelo 3 no existen cambios importantes. Las variables explicativas que resultan significativas son las mismas en ambos casos y tampoco se diferencian el R cuadrado ajustado ni la correlación entre la variable dependiente real y la predicha por el modelo.
- Sobre las variables significativas es posible mencionar lo siguiente:
  - A mayor deserción escolar, se predicen menores valores del PIB per capita.
  - Mayores niveles de inflación están asociados a menores niveles del producto.
  - Menor igualdad de género se asocia a menores niveles del producto.
  - A menor mortalidad de la madre en parto, se observa un mayor PIB.
  - Es curioso también identificar que la estimación arroja una relación negativa entre la esperanza de vida y el PIB.

# Anexos.

## Anexo 1. Descripción de variables.

- PIB: PIB per cápita - miles de millones de USD / habitante.
- POB: Población - millones.
- IDH: Índice de desarrollo humano.
- GINI: Coeficiente de Gini.
- IPC: Índice del precio al consumidor.
- FAO: Índice de precios alimenticios de la FAO.
- GENERO: Índice de desigualdad de género.
- ELECTRICIDAD: Tasa de electrificación.
- ESCOLARIDAD: Años de escolaridad promedio.
- SUICIDIOFEM: Tasa de suicidios femeninos - cada 100.000 personas.
- SUICIDIOMAS: Tasa de suicidios masculinos - cada 100.000 personas.
- BOSQUE: Porcentaje de superficie que corresponde a bosques.
- FOSIL: Porcentaje de uso de combustibles fósiles.
- DIOXIDO: Emisiones de dióxido de carbono - Toneladas per cápita.
- DESASTRE: Población afectada por desastres naturales - miles.
- AFECTADOS: Población sin hogar por desastres naturales - miles.
- HOMICIDIO: Tasa de homicidios - cada 100.000 personas.
- MORTINF: Mortalidad de niños menores de 5 años - miles.
- MORTMAT: Tasa de mortalidad maternal - cada 100 nacimientos.
- TURISMO: Turistas internacionales - millones.
- INTERNET: Porcentaje de uso de internet.
- VIOLENCIA: Porcentaje de víctimas de violencia entre parejas.
- VIDA: Expectativa de vida - años.
- CELULAR: Subscripciones a telefonía celular - cada 100 personas.
- DESERCION: Tasa de deserción escolar primaria.
- PRISION: Tasa de encarcelamiento - 100.000 personas.
- RENOVABLE: Porcentaje de uso de energías renovables.
- PARLAMENTO: Porcentaje de escaños parlamentarios ocupados por mujeres.
- INMIGRANTES: Porcentaje de población que es inmigrante.