Trabajo Practico R

por Francisco Di Menna, Valentin Paniagua, Lorenzo Lavieri y Tomas Ripoll Brandoni

Ejercicio 1: Análisis econométrico con datos de Gapminder

Parte 1: Ingreso por persona

En primer lugar hagamos una introducción sobre Gapminder. Gapminder es un dataset el cual contiene información variada (población, expectativa de vida, mortalidad infantil, ingresos, religión predominante, entre otros) de varios países de diferentes regiones del mundo entre los años 1960 y 2010. A lo largo de este trabajo práctico, iremos filtrando la información del dataset según convenga para responder cada una de las consignas propuestas.

Inciso 1:

En este primer inciso nos piden que grafiquemos la variable *income_per_person* para la Argentina. Tras correr la correspondiente línea de código, el gráfico muestra una tendencia creciente de largo plazo, reflejando el aumento general del nivel de vida en el país. Sin embargo, dicha tendencia está acompañada por fuertes oscilaciones, lo que evidencia la alta volatilidad macroeconómica característica de Argentina. Por ejemplo, puede observarse un fuerte crecimiento entre 1990 y 1998, asociado al plan de convertibilidad y la apertura económica del período menemista. Posteriormente, se produce un profundo descenso, vinculado con la crisis financiera y el colapso del régimen cambiario en 2001. Finalmente, durante los años posteriores se aprecia una recuperación significativa del ingreso per cápita.

En resumen, el gráfico refleja un comportamiento procíclico y volátil del ingreso per cápita argentino, con períodos de expansión y crisis bien marcados, lo que resulta consistente con la dinámica económica del país a lo largo de toda su historia.

Inciso 2:

En este segundo inciso se solicita estimar distintos modelos de regresión para explicar la evolución del ingreso per cápita argentino en función del tiempo. Con este fin, se divide la muestra en dos subconjuntos: un conjunto de entrenamiento (train) que incluye todos los años menos los últimos diez, y un conjunto de prueba (test) que contiene esos diez años finales.

A partir del conjunto de entrenamiento se ajustan tres modelos:

- Un modelo lineal simple, que asume una relación lineal y constante entre el año y el ingreso per cápita.

- Un modelo polinómico de grado 2, que permite capturar cierta curvatura en la tendencia, reflejando posibles aceleraciones o desaceleraciones en el crecimiento.

- Un modelo polinómico de grado 10, mucho más flexible, que puede ajustarse casi perfectamente a las variaciones del conjunto de entrenamiento, aunque con un alto riesgo de

sobreajuste.

Los tres modelos se utilizan luego para predecir el ingreso per cápita en el conjunto de prueba (test). Al comparar los resultados, se observa que el modelo lineal reproduce correctamente la tendencia general, pero no logra capturar las fluctuaciones de corto plazo. El modelo de grado 2 mejora el ajuste al permitir una forma ligeramente curvada, representando mejor los períodos de aceleración y desaceleración del ingreso per cápita. Por su parte, el modelo de grado 10 se ajusta muy bien a los datos, pero presenta un comportamiento excesivamente sensible a las variaciones dentro del periodo analizado evidenciando un claro caso de sobreajuste. Este resultado ilustra la clásica disyuntiva entre sesgo y varianza: los modelos más complejos pueden capturar mejor los datos históricos, pero pierden capacidad de generalización.

En conclusión, el análisis muestra que un modelo polinómico de bajo grado (como el de grado 2) logra un equilibrio razonable entre ajuste y estabilidad predictiva, mientras que modelos demasiado flexibles tienden a replicar el ruido de la serie más que su tendencia estructural.

Inciso 3:

En este último inciso de la primera parte se busca analizar la relación entre los ingresos per cápita de distintos países sudamericanos en este caso, Chile, Perú, Uruguay y Brasil con el objetivo de examinar el grado de similitud en sus trayectorias de desarrollo y en sus ciclos económicos.

En primer lugar, se calcula la matriz de correlaciones entre los niveles del ingreso per cápita. Los resultados muestran correlaciones muy elevadas (en torno a 0.85 o superiores) entre todos los países, lo que sugiere que las economías de la región comparten una tendencia de largo plazo común. Esto puede interpretarse como evidencia de un proceso de convergencia económica parcial, donde las mejoras en el ingreso promedio siguen una dirección ascendente y relativamente sincronizada. Factores estructurales comunes como la integración comercial, la especialización en exportaciones de recursos naturales y la exposición a shocks internacionales contribuyen a esta evolución conjunta.

A continuación, se repite el ejercicio pero tomando las tasas de crecimiento interanual del ingreso per cápita, en lugar de sus niveles. Esta transformación permite capturar las fluctuaciones de corto plazo y analizar hasta qué punto las economías se mueven de manera sincronizada en términos de ritmo de crecimiento, más allá de su nivel de desarrollo.

Los resultados muestran que las correlaciones entre las tasas de crecimiento son positivas, pero significativamente menores que las observadas para los niveles. Esto indica que, aunque las economías sudamericanas tienden a crecer en el largo plazo, la intensidad y el momento de ese crecimiento difieren entre países. Por ejemplo, algunos pueden experimentar expansiones más fuertes en determinados períodos (como durante el boom de los commodities entre 2003 y 2011), mientras otros enfrentan desaceleraciones o crisis internas.

En otras palabras, los países comparten shocks externos comunes, como variaciones en los precios internacionales o condiciones financieras globales, pero la respuesta interna depende de sus políticas

macroeconómicas, la estabilidad institucional, la estructura productiva y la dependencia de ciertos sectores (por ejemplo, minería en Perú o soja en Brasil).

Por lo tanto, mientras los niveles de ingreso per cápita muestran un patrón de crecimiento conjunto y sostenido, las tasas de crecimiento revelan una dinámica más heterogénea y menos sincronizada, que refleja la diversidad estructural y las asimetrías en la capacidad de respuesta ante los ciclos económicos internacionales.En conclusión, este análisis evidencia que las economías de Sudamérica comparten una tendencia ascendente común, pero presentan diferencias marcadas en sus trayectorias de corto plazo, lo que sugiere que los procesos de crecimiento regional no son plenamente coordinados ni uniformes.

Parte 2

Inciso 5:

En este inciso se analiza la relación entre la expectativa de vida total y la expectativa de vida de las mujeres en el año 2010, utilizando los datos del dataset *gapminder*.

El gráfico de dispersión elaborado permite visualizar cómo se relacionan ambas variables a nivel global. Cada punto representa un país, ubicado en el eje horizontal la expectativa de vida general y en el eje vertical la expectativa de vida femenina. A simple vista, se observa una fuerte asociación positiva: los países con una mayor expectativa de vida total tienden a tener también una mayor expectativa de vida para las mujeres.

Esta relación se complementa con la línea de tendencia agregada al gráfico mediante el ajuste de una regresión simple. Dicha recta confirma la correlación positiva y significativa entre ambas variables, lo que implica que la expectativa de vida femenina crece sistemáticamente con la expectativa de vida promedio del país.

Desde un punto de vista interpretativo, esta asociación es esperable. En prácticamente todos los países del mundo, las mujeres viven más años que los hombres, debido a diferencias biológicas, sociales y de comportamiento. Por lo tanto, la expectativa de vida femenina actúa como un indicador más sensible del nivel sanitario y del desarrollo humano de cada sociedad.

Además, el gráfico muestra que los países con expectativas de vida más bajas (por debajo de los 60 años) presentan una mayor dispersión, mientras que aquellos con niveles altos de desarrollo (por encima de los 75 años) tienden a concentrarse en torno a la recta de regresión. Esto sugiere que en los países más avanzados las brechas entre sexos se reducen y la mortalidad se estabiliza, mientras que en las regiones con menores niveles de desarrollo existen diferencias más amplias entre la población general y la femenina.

En síntesis, el análisis revela que existe una relación positiva, fuerte y coherente entre ambas variables.

Inciso 6:

En este inciso se profundiza el análisis anterior mediante una regresión lineal simple, donde la variable dependiente es la expectativa de vida total (*life_expectancy*) y la variable explicativa es la expectativa de vida femenina (*life_expectancy_female*).

El modelo estimado permite cuantificar la relación entre ambas variables y verificar su significancia estadística. El resultado del ajuste muestra un coeficiente positivo y altamente significativo con un intervalo de confianza del 1 por ciento (0,01) para la variable *life_expectancy_female*, lo cual confirma la relación creciente entre ambas medidas. En términos económicos, esto implica que los países con mayor longevidad femenina tienden a registrar también una mayor expectativa de vida promedio en toda su población.

El valor del intercepto representa la expectativa de vida total teórica cuando la expectativa de vida femenina fuera cero (valor sin interpretación práctica, pero útil para determinar el punto de partida del modelo). En cambio, el coeficiente de pendiente indica cuántos años aumenta la expectativa de vida total, en promedio, cuando la expectativa de vida femenina aumenta un año. Dado que el coeficiente es cercano a 1, la relación entre ambas variables es casi proporcional.

El R^2 nos indica cuánto explica nuestro modelo a la variable independiente. El mismo siempre se ubica en el intervalo (0,1). En nuestra regresión nos arroja un R^2 de 0.86 que refleja una gran parte de la variabilidad en la expectativa de vida total se explica por la variabilidad en la expectativa de vida femenina. Esto sugiere que el modelo ajusta adecuadamente los datos y que la expectativa de vida femenina es un predictor muy sólido del promedio nacional de esperanza de vida.

En el gráfico correspondiente, la nube de puntos muestra una clara alineación en torno a la recta de ajuste, lo cual refuerza visualmente la fortaleza de la relación lineal. Las desviaciones leves que se observan pueden deberse a diferencias estructurales entre países, como brechas de género, acceso desigual a la salud o distintos niveles de desarrollo económico.

En resumen, la evidencia empírica muestra que la expectativa de vida femenina y la total de la población están estrechamente relacionadas de manera positiva y casi lineal, validando la hipótesis de que los avances en salud y bienestar de las mujeres se reflejan directamente en la mejora del conjunto de la población.

Inciso 7:

En este inciso se aplica un test t de una muestra para analizar si la diferencia entre la expectativa de vida femenina y la expectativa de vida total es significativamente mayor que cero.

El resultado del test arroja un estadístico $t = 7.3957$, con un p-valor = 3.5×10^{-12}, lo cual es extremadamente pequeño. Dado que este p-valor es mucho menor que cualquier nivel de significancia habitual (por ejemplo, 0.05 o incluso 0.01), se rechaza la hipótesis nula con total confianza.

En términos prácticos, esto significa que existe evidencia estadísticamente significativa de que la expectativa de vida de las mujeres es superior a la expectativa de vida total promedio. El intervalo de

confianza del 95% indica además que, en promedio, las mujeres viven al menos 1.46 años más que el conjunto de la población.

Desde el punto de vista económico y social, este resultado es coherente con lo que se observa a nivel mundial: las mujeres suelen vivir más que los hombres debido a diferencias biológicas, factores de comportamiento (menor exposición a riesgos laborales o conductas de riesgo), y mayor utilización del sistema de salud.

En conclusión, el test confirma de manera robusta que la longevidad femenina supera a la del promedio poblacional, consolidando la evidencia observada en los gráficos previos y reforzando la relación positiva entre bienestar, salud y género.

Inciso 8:

En este inciso se extiende el modelo de regresión del inciso 6 con el objetivo de explicar la expectativa de vida total (*life_expectancy*) no solo a partir de la expectativa de vida femenina (*life_expectancy_female*), sino también incorporando el nivel de ingreso per cápita (*income_per_person*) como una segunda variable explicativa.

La estimación del modelo múltiple permite evaluar si, al controlar por el nivel de ingresos, la relación entre la expectativa de vida femenina y la general se mantiene significativa, y en qué medida la riqueza de un país contribuye a explicar las diferencias en longevidad.

Los resultados muestran que ambos coeficientes tanto el de *life_expectancy_female* como el de *income_per_person* son positivos. El coeficiente asociado a *life_expectancy_female* es significativo pero el coeficiente asociado a *income_per_person* es estadísticamente no significativo. Esto implica que, a la hora de explicar *life_expectancy,* utilizar *income_per_person* no nos aporta ninguna información. Una interpretación posible es que el efecto del ingreso sobre la salud y la longevidad ya está reflejado indirectamente en la expectativa de vida femenina, dado que ambos indicadores capturan dimensiones similares del desarrollo humano y del acceso a mejores condiciones de vida. Por otro lado, el R^2 del modelo múltiple es prácticamente igual al del modelo simple del inciso 6, lo que reafirme que la inclusión del ingreso per cápita no mejora la capacidad explicativa del modelo.

Por último, para complementar nuestro análisis, incorporamos un gráfico de las regresiones. En los mismos se observa que tanto el modelo de regresión simple como el modelo de regresión múltiple son prácticamente iguales lo que nuevamente confirma nuestra idea, es decir, que incorporar *income_per_person* como variable explicativa no mejora la capacidad predictiva de nuestro modelo.

Inciso 9:

En este último inciso se amplía el modelo de regresión con el fin de analizar los determinantes múltiples de la expectativa de vida total. Para ello, se incluyen como variables explicativas:

- *income_per_person*: el ingreso per cápita, indicador del nivel de desarrollo económico;

- *child_mortality:* la tasa de mortalidad infantil, que refleja la calidad del sistema de salud y las condiciones sanitarias básicas;

- *children_per_woman*: la tasa de fertilidad, que captura aspectos demográficos y sociales relacionados con la estructura poblacional.

El modelo inicial muestra que los coeficientes asociados a *income_per_person* y *child_mortality* son altamente significativos y con los signos esperados:

- El coeficiente de *income_per_person* es positivo, indicando que un mayor nivel de ingreso se asocia con una mayor esperanza de vida.

- El coeficiente de *child_mortality* es negativo, lo que significa que a medida que disminuye la mortalidad infantil, aumenta la esperanza de vida promedio, lo cual es coherente con la evidencia empírica y la teoría del desarrollo.

Por el contrario, la variable *children_per_woman* resulta no significativa, lo que sugiere que, una vez controlados los efectos del ingreso y la mortalidad infantil, la fertilidad no tiene un impacto independiente relevante sobre la expectativa de vida. En consecuencia, se estima un modelo corregido, excluyendo la variable no significativa (*children_per_woman*) e incorporando life_expectancy_male como nuevo predictor, con el fin de capturar la posible influencia del componente masculino en la expectativa de vida total.

Los resultados del modelo corregido confirman y fortalecen las conclusiones anteriores:

- La mortalidad infantil mantiene un efecto negativo fuerte y significativo.

- El ingreso per cápita arroja un valor positivo pero no significativo.

- La expectativa de vida masculina también es significativa y positiva, lo que sugiere que las mejoras en la salud y en la calidad de vida de los hombres contribuyen directamente a elevar la expectativa de vida total del país.

Dado que el resumen de la segunda regresión nos arrojó que, el ingreso per cápita es no significativo, lo quitamos de la regresión y corremos una tercera regresión donde las únicas variables explicativas son *child_mortality* y *expectancy_life_male*.

Por otra parte, para complementar nuestro análisis, observamos y comparamos los R^2 ajustado de cada una de las regresiones. . El primer modelo nos arrojó un valor de 0.79 mientras que el segundo y tercer modelo nos devuelve un valor de 0.89. Por lo tanto, el modelo más apropiado para predecir el comportamiento de la expectativa de vida es el tercero.

Conclusión final:

A lo largo de este trabajo se exploraron distintas dimensiones del desarrollo económico y social utilizando el dataset *gapminder*, que reúne información comparada entre países sobre ingreso, salud y demografía.

En la primera parte, el análisis se centró en la evolución del ingreso per cápita argentino, mostrando una clara tendencia creciente de largo plazo pero con alta volatilidad, reflejo de la recurrencia de crisis económicas y cambios estructurales en el país. El ejercicio de regresión lineal y polinómica permitió observar que los modelos más simples capturan adecuadamente la tendencia general, mientras que los más complejos tienden a sobreajustarse, perdiendo capacidad predictiva. Este resultado ilustra la importancia de equilibrar la complejidad del modelo con su capacidad de generalización.

Posteriormente, el análisis de las correlaciones entre países sudamericanos reveló que las economías de la región presentan trayectorias de ingreso per cápita fuertemente relacionadas en el largo plazo, aunque sus tasas de crecimiento muestran comportamientos más dispares. Esto sugiere la existencia de tendencias estructurales comunes vinculadas a factores externos como los precios de los commodities o la apertura comercial, pero también ciclos económicos propios que responden a las particularidades de cada país.

En la segunda parte, el foco se desplazó hacia los determinantes de la salud y la expectativa de vida, confirmando empíricamente la estrecha relación entre la longevidad femenina y la expectativa de vida total. Los resultados estadísticos y econométricos mostraron de manera consistente que las mujeres viven más que el promedio poblacional, diferencia que es significativa y sistemática a nivel global. Luego incluimos otras variables explicativas como la mortalidad infantil, expectativa de vida de los hombres y llegamos a que la inclusión de nuevas variables mejora la capacidad de predicción de nuestro modelo

En conjunto, el trabajo muestra cómo el uso de herramientas estadísticas y de visualización en R permite vincular los datos empíricos con los fenómenos económicos y sociales que describen la realidad de los países. Los resultados obtenidos destacan la interdependencia entre crecimiento, desarrollo humano y salud pública, reforzando la idea de que el progreso económico sostenible requiere simultáneamente estabilidad macroeconómica y mejora de las condiciones sociales básicas.

Ejercicio Simulación 2: Simulación de ataque en el juego TEG

El código desarrollado reproduce el juego de mesa TEG, donde un atacante y un defensor se enfrentan lanzando dados hasta que uno de los dos no puede continuar. El objetivo del atacante es eliminar todas las fichas del defensor, mientras que el defensor busca resistir el ataque. A partir de estas reglas, se programaron tres funciones principales en R: resultado_ataque, simular_batalla y probabilidad_ataque, que permiten simular desde una sola tirada de dados hasta miles de batallas completas para estimar probabilidades de victoria.

La primera función, resultado_ataque, reproduce una única tirada de dados entre atacante y defensor. Según las reglas del TEG, el atacante puede tirar hasta tres dados, siempre que deje al menos una ficha sin usar, y el defensor puede tirar hasta tres dados según sus fichas disponibles. En el código, los dados se generan aleatoriamente del 1 al 6 y se ordenan de mayor a menor para comparar los valores más altos. Solo se comparan tantos dados como tenga el jugador con menos dados, y en cada enfrentamiento individual, si el dado del atacante es mayor, el defensor pierde una ficha; si es igual o menor, la pérdida es del atacante. De esta forma, la función devuelve un vector con las pérdidas de ambos jugadores, reflejando exactamente la mecánica del TEG, donde los empates favorecen al defensor.

La segunda función, simular_batalla, encadena varias rondas consecutivas de resultado_ataque para simular una batalla completa. El ciclo continúa mientras el atacante tenga más de una ficha (ya que no puede atacar con la última) y el defensor aún tenga fichas. En cada ronda se determina la cantidad de dados de cada jugador en función de sus fichas restantes y se actualiza el número de fichas después de aplicar las pérdidas de esa ronda. Cuando el atacante llega a una sola ficha o el defensor se queda sin unidades, la batalla termina. Si el atacante conserva solo una ficha, el intento de conquista se da por fallido; si el defensor llega a cero, el atacante ha ganado.

Por último, la función probabilidad_ataque utiliza simular_batalla dentro de un bucle para ejecutar mil simulaciones con el fin de estimar la probabilidad de victoria del atacante. En cada simulación, ambos comienzan con cinco fichas y se registra una victoria si el defensor termina con cero fichas. Al final, la función divide las victorias del atacante entre el número total de simulaciones para obtener una probabilidad promedio.

Los resultados obtenidos en las simulaciones muestran que, cuando ambos jugadores comienzan con cinco fichas, la probabilidad de victoria del atacante ronda el 20 %. Este valor es coherente con las reglas y la lógica del juego: aunque el atacante suele tirar más dados, el hecho de que deba dejar una ficha sin atacar y que los empates favorezcan al defensor generan una ventaja estructural para la defensa. En otras palabras, cuando la batalla se inicia en igualdad de condiciones, el atacante solo gana una de cada cinco veces aproximadamente.

Ejercicio de Análisis de datos: US Presidential Elections 2024

En este trabajo se utiliza un conjunto de datos de encuestas electorales de Estados Unidos (polls.csv), que recopila información sobre distintas mediciones de intención de voto realizadas antes de las elecciones presidenciales de 2024. El dataset incluye variables como las fechas de cierre de cada encuesta, el tamaño de la muestra, la metodología utilizada y los porcentajes estimados para los candidatos demócrata y republicano. A partir de estos datos, se busca analizar qué factores explican las diferencias entre los resultados de las encuestas y el resultado electoral real.

Para ello, el informe se estructura en torno a tres preguntas principales:

1. ¿Cómo influye el tamaño de la muestra en la precisión de las encuestas?
   Se analiza si un mayor número de casos reduce el error de estimación, considerando tanto modelos lineales como cuadráticos.
2. ¿Cómo afecta la recencia de la encuesta a su precisión?
   Se evalúa si las encuestas más cercanas a la fecha de la elección tienden a ser más exactas.
3. ¿Qué papel cumple la metodología de recolección de datos en la calidad de las predicciones?
   Se comparan los distintos métodos (teléfono, panel online, texto, combinaciones híbridas, etc.) mediante análisis de promedios, varianzas y regresión.

El objetivo general del trabajo es identificar qué características de las encuestas contribuyen a una mayor exactitud en la estimación del resultado electoral, proporcionando evidencia empírica sobre los determinantes de la precisión en el contexto de las elecciones estadounidenses.

Pregunta 1.

El análisis del tamaño muestral muestra que la relación entre la cantidad de encuestados y el error de las predicciones no es lineal. En lugar de una mejora constante al aumentar la muestra, los resultados indican que el efecto del tamaño muestral presenta rendimientos decrecientes: las encuestas más pequeñas tienden a ser imprecisas por limitaciones estadísticas, pero a medida que la muestra crece, el error disminuye rápidamente hasta estabilizarse. A partir de cierto punto, ampliar aún más la muestra no reduce de forma significativa el error, e incluso puede generar leves aumentos.

El uso del logaritmo del tamaño muestral fue clave para capturar esta dinámica, ya que permite modelar mejor las relaciones donde los efectos se reducen progresivamente y los incrementos relativos son más relevantes que los absolutos. Además, la regresión cuadrática sobre el logaritmo del tamaño ofrece un ajuste superior al modelo lineal: presenta un R^2 ajustado más alto, menor error residual y un comportamiento más realista, evidenciando una relación de tipo “U” que refleja la complejidad de los procesos de muestreo y de error en encuestas electorales. En síntesis, el modelo cuadrático describe con mayor precisión la verdadera naturaleza de la relación entre el tamaño muestral y la precisión: el tamaño importa, pero su impacto positivo disminuye conforme crece la muestra.

Pregunta 2.

La correlación entre los días hasta la elección y el error absoluto es de 0.44, lo que indica una relación positiva moderada: las encuestas más lejanas al día de la elección tienden a ser menos precisas. El modelo lineal confirma esta tendencia: el coeficiente positivo muestra que por cada día adicional antes de la elección, el error promedio aumenta. Con un R^2 ajustado de 0.19, el modelo explica una parte relevante de la variación del error, demostrando que la "recency" tiene un efecto claro sobre la precisión. En conclusión, las encuestas más cercanas al día de votación ofrecen estimaciones más exactas que aquellas realizadas con mucha anticipación.

Pregunta 3.

El análisis de la variable metodología revela diferencias claras y estadísticamente significativas en la precisión de las encuestas según el método de recolección de datos. A partir de los promedios obtenidos con aggregate, se observa que las encuestas que emplean métodos mixtos (como Online Panel/Text-to-Web o IVR/Online Panel/Text-to-Web/Email) presentan errores absolutos considerablemente menores mientras que las metodologías basadas únicamente en paneles online o llamadas telefónicas superan los 3 puntos de error promedio. Además, la desviación estándar muestra que los métodos mixtos no solo son más precisos, sino también más consistentes en sus resultados.

Al profundizar con la regresión lineal, se observa que varias metodologías, especialmente aquellas que integran componentes digitales o textuales, tienen coeficientes negativos y significativos, lo que indica un menor error en comparación con el método base (Online Panel). En cambio, las metodologías tradicionales muestran errores mayores o sin diferencias significativas.

En conclusión, los resultados demuestran que la elección del método de recolección influye directamente en la precisión de las encuestas. Las estrategias más modernas e híbridas —que combinan distintos canales de contacto digital— logran una representación más fiel del electorado y reducen el margen de error, mientras que las encuestas puramente telefónicas o de paneles online presentan mayores sesgos y variabilidad en sus estimaciones.