

ECONOMETRICS IS THE ORIGINAL DATA SCIENCE.

- JOSHUA ANGRIST

PO: YOU SET ME UP TO FAIL? WHY?

MASTER SHIFU: IF YOU ONLY DO WHAT YOU CAN, YOU'LL NEVER BE MORE THAN YOU ARE NOW.

PO: BUT I DON'T WANT TO BE MORE! I LIKE WHO I AM!

MASTER SHIFU: YOU DON'T EVEN KNOW WHO YOU ARE.

- KUNG FU PANDA 3

NO SÉ SI ES CASUALIDAD QUE YO ME SIENTA ASÍ

SIEMPRE QUE TÚ ESTÁS CERQUITA DE MÍ.

- BENITO ANTONIO MARTÍNEZ OCASIO (*BAD BUNNY*)

Índice general

<i>Antes de comenzar: ¿Qué es la inferencia causal?</i>	15
<i>Los negocios son matemáticas</i>	21
<i>Python para hacer Econometría</i>	27
<i>El modelo de resultados potenciales</i>	41
<i>Experimentos y Pruebas A/B</i>	55
<i>Una guía para entender y hacer modelos de Regresión Lineal</i>	65
<i>Cómo se usan los modelos de series de tiempo para proyectar las ventas en una empresa</i>	99
<i>Efectos Fijos</i>	125
<i>Diferencias en Diferencias</i>	141
<i>La única forma de crear valor es iterando a partir de los datos</i>	155
<i>Cómo hacer Investigación de mercados con inteligencia artificial</i>	165

Agradecimientos 175

Glosario 177

Bibliografía 185

Índice de figuras

1. Broad Street en Londres durante el brote de cólera de 1854. John Snow apunta a la bomba de agua como fuente de la enfermedad. Fuente: imaginado por el autor y creado por DALL-E 3. 15
2. El mapa que John Snow trazó para identificar que el origen del cólera no era el *miasma*, sino el agua contaminada. Ubicar geográficamente los brotes en un mapa es algo que no se había hecho antes y ayudó a identificar con facilidad que la fuente era una bomba específica. Cerrar esa bomba de agua logró salvar algunas vidas, pero el modelo de pensamiento de Snow sigue salvando muchas vidas más hasta la fecha. 17
3. Crecimiento del PIB de México de 1990 a la actualidad. Como milenial, me ha tocado vivir 3 crisis: la crisis del peso mexicano (o efecto tequila, la llamada “gran recesión” y la crisis del Covid-19). Fuente: Elaboración propia con datos del Banco Mundial. 22
4. Florence en su estudio. Fuente: elaborado por el autor con Dall-e 28
5. El Diagrama de la Rosa que Florence Nightingale presentó y que demostró que las muertes venían más por las enfermedades que por la batalla. 28
6. La Lotería es un juego de azar tradicional mexicano. Imagen de Alex Covarrubias, vía Wikipedia (Dominio público) 30
7. En una conferencia de estadística: "Levanta la mano si estás familiarizado con el sesgo de selección... como pueden ver, es un término que la mayoría de las personas conoce...".
Fuente: xkcd 45
8. “Hemos recibido 500 respuestas y encontramos que a las personas les encanta responder a las encuestas”.
Fuente: Jono Hey, Sketchplanations 48
9. Diagrama de dispersión generado a partir del bloque de código de Python. Fuente: Elaboración propia. 66
10. Diagrama de dispersión con línea de regresión. 67

11. Le llamamos “error” a la diferencia entre una observación y la línea de regresión que “predice” dónde debería estar Y_i dado X_i . Fuente: Elaboración propia con Python. 68
12. El objetivo del método de mínimos cuadrados es encontrar la línea que minimice la suma de los errores al cuadrado. Fuente: Elaboración propia con Python. 69
13. Andrei Márkov (izquierda) y Carl Friedrich Gauss (derecha) jugando ajedrez. Fuente: imaginado por mi y hecho con chatGPT. 76
14. Un diagrama de dispersión nos muestra que X y Y son variables que no parecen tener ninguna relación entre sí. 76
15. No es sino hasta que agregamos una tercera variable que podemos revelar la verdadera relación que hay entre las variables. 77
16. Los residuales y los valores de predicción no muestran estar correlacionados. La gráfica muestra una linea de regresión cercana a cero. 81
17. Gráfico que compara residuales vs cada uno de los predictores en un diagrama de dispersión. No se ve un patrón específico. 83
18. El 5 % del p-value es lo mismo que uno en veinte. Con ese parámetro, no es necesario recurrir al p-hacking para obtener resultados “significativos”.
Fuente: xkcd 88
19. Hay una correlación entre el número de premios Nobel que gana un país y su consumo de chocolate. Naturalmente, esa correlación no implica causalidad. Fuente: Leo Prinz, 2020 (actualizado con datos hasta 2024) 91
20. Caminata aleatoria. Cualquier parecido con una serie de tiempo real, es mera coincidencia. 101
21. Diferencia visual entre una serie no-estacionaria y la misma serie con una diferencia aplicada. Nota que la serie se vuelve estacionaria. 104
22. Simulación de un proceso AR(3). 106
23. Índice Global de la Actividad Económica (IGAE). El gráfico lo muestra como un gráfico de líneas. En una serie estacionaria, no se debe observar una tendencia clara. Este indicador a simple vista se puede observar que no es estacionario. Fuente: INEGI. 110
24. Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie original (sin diferenciación). 111
25. Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie diferenciada. Nota que en la serie diferenciada, el orden de rezago a elegir es menor al de la serie original. 112
26. Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie diferenciada. Nota que en la serie diferenciada, el orden de rezago a elegir es menor al de la serie original. 115

27. Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie diferenciada. Nota que en la serie diferenciada, el orden de rezago a elegir es menor al de la serie original. 117
28. Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de nuestra simulación de un proceso ARMA(1,3). 118
29. En una regresión lineal que no distingue los datos entre diferentes medios, no parece haber ningún efecto entre el gasto de publicidad y las ventas. 128
30. Sólo es necesario incluir el medio como variable de control. Hacerlo delata la verdadera relación que hay entre la publicidad y las ventas. 129
31. Cada medio de publicidad tiene un valor promedio de ventas (\bar{Y}_i) y del costo del anuncio (\bar{X}_i). Las líneas en el gráfico muestra donde se cruzan esos valores promedio, con una desviación estándar para determinar el tamaño de la línea. Este es un paso determinante para crear nuestras variables *within*. 131
32. Al absorber los efectos fijos, los datos se conjuntan en un solo punto medio de publicidad y ventas. El resultado es que el coeficiente de una regresión conjunta captura el efecto independiente del medio que se está usando. 133
33. Bajo un modelo neoclásico ortodoxo, el establecer un salario mínimo crea un desequilibrio artificial en el mercado que causa desempleo. Este modelo fue la base de muchas políticas de libre mercado. La evidencia que presentaron Card y Krueger no muestran este efecto en la realidad. Fuente: Elaboración propia. 141
34. El supuesto de tendencias paralelas implica que, de no ser por la aplicación del tratamiento, ambos grupos seguirían la misma tendencia. Eso convierte a la diferencia entre tendencias en el efecto causal. 145
35. Precios promedio de bebidas en la zona de Berkeley. Fuente: Elaboración propia con datos de Silver *et al.* (2014) 148

Índice de cuadros

1. Tabla de resumen de las muertes antes y después del cambio de la toma de agua por la compañía Lambeth. Fuente: Elaboración propia basada en Snow (1859), p. 90. 16
2. Promedios simulados según tipo de escuela 45
3. Datos de ventas y costos de anuncios en Google Ads 126
4. Gasto en publicidad y ventas por año (Google Ads) 132
5. Efecto de la publicidad en las ventas hecha con un modelo de mínimos cuadrados con las variables centradas por grupos. 134
6. Modelo de efectos fijos (PanelOLS) – Variable dependiente: Ventas 135
7. Modelo de efectos fijos (PanelOLS) – Variable dependiente: Ventas 137
8. Resultados del estudio de Card & Krueger (1994) 142

Antes de comenzar: ¿Qué es la inferencia causal?

A mediados del siglo XIX hubo un brote de cólera que cambió el mundo por completo¹.

El cólera es una enfermedad terrible. Hoy en día sabemos con certeza que su origen proviene de una bacteria. Saber esto es muy útil: si nos llegara a infectar esa bacteria, simplemente tendríamos que tomar un antibiótico y asunto resuelto. Nuestra vida no estaría en peligro.

Pero para las personas que vivían en Londres en 1854, saber la verdad era cuestión de vida o muerte.

A lo largo de la historia hubo muchas pandemias de cólera. La primera registrada es la de 1817 a 1824 y desde entonces la Organización Mundial de la Salud (OMS) considera siete pandemias distintas. Aunque las pandemias son eventos separados, algunas se han extendido por décadas, como la que comenzó en 1961 y sigue vigente hasta el día de hoy. El brote de 1854 de Londres sucedió en la calle Broad Street y mató a 616 personas.

En la época, la teoría dominante de la causa del cólera era el miasma

El *miasma* se refería a **aires malos**. La idea era que las enfermedades se transmitían por los aires de una persona enferma. Es una teoría que duró por muchos siglos desde Hipócrates, más de 400 años antes de Cristo hasta que la teoría de los gérmenes la sustituyó².

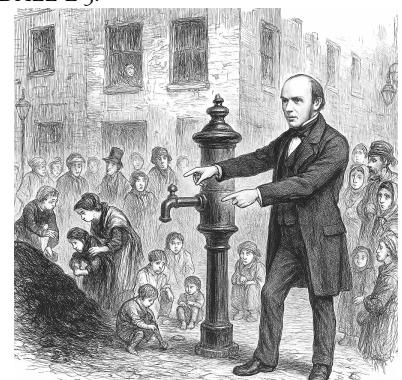
John Snow aún no sabía de la teoría de los gérmenes, pero no le convencía la teoría del *miasma* y decidió ponerla a prueba.

Usó lo que los economistas conocemos como un **experimento natural**³.

En la zona de la epidemia habían dos compañías que tomaban el agua del río Támesis y la distribuían en la ciudad: *Southwark and*

¹ John Snow luego le llamaría “El brote más terrible de cólera que haya ocurrido en este reino”.

Figura 1: Broad Street en Londres durante el brote de cólera de 1854. John Snow apunta a la bomba de agua como fuente de la enfermedad. Fuente: imaginado por el autor y creado por DALL-E 3.



² Fue hasta los 1860s que Louis Pasteur publicó su trabajo

³ Un *experimento natural* ocurre cuando las condiciones del mundo real imitan un experimento aleatorio. Veremos más al respecto más adelante.

Vauxhall y *Lambeth*. El agua que llevaban estaba muy contaminada. **Visiblemente** contaminada. Fue durante el mismo periodo que Lambeth decidió cambiar su toma de agua un poco más río arriba⁴.

La gran aportación de John Snow fue haber tomado nota de las muertes por cólera antes y después del cambio de la toma de agua. El cuadro 1 muestra este registro para las dos empresas. Snow notó que algunos distritos recibían su agua de ambas compañías, pero otros tomaban su agua de una compañía en específico⁵.

Esta era la clave para identificar sin lugar a dudas que la causa de la enfermedad era el agua y no los malos aires.

Observa las muertes por cólera de los distritos donde el suministro de agua es Southwark & Vauxhall. Prácticamente no hay cambio. En cambio, pon atención a las muertes por cólera de los distritos que suministraban su agua de Lambeth: cayeron dramáticamente. Las muertes de los distritos que se suministraban de ambas compañías también cayeron, pero ese dato no sería tan informativo sin los datos por separado.

Suministro de Agua	Muertes en 1849	Muertes en 1854
Lambeth	162	37
Southwark & Vauxhall	2261	2458
Lambeth y Southwark & Vauxhall	3905	2547

La inferencia causal se trata más de sentido común que de la estadística en sí misma

Más de 150 años después, el desafío sigue siendo similar. Hoy en día tenemos mucha más disponibilidad de datos que en 1859, y a pesar de eso, estamos en el punto en el que somos más vulnerables a la desinformación.

Durante la pandemia de Covid-19, la desinformación fue responsable de la muerte de muchas personas. Un estudio demostró que en Estados Unidos, las personas afiliadas al partido Republicano fueron más vulnerables a las noticias falsas sobre los efectos de las vacunas. La consecuencia fue que hubo un **exceso de muertes** 43 % mayor entre republicanos que entre demócratas⁶.

El problema es que estamos tan inundados ahora con información, que lo más crítico es saber qué hacer con ella.

John Snow reunió muchos datos sobre las muertes por cólera. Pero

⁴ El problema era que todos los desechos de la ciudad desembocaban en el río Támesis. Entre esos desechos, naturalmente, estaban las heces de todos los habitantes de una ciudad. El problema llegó a ser tan grande que en el verano de 1858 el olor insopportable obligó a los miembros del parlamento a abandonar el palacio de Buckingham en un episodio que se conoció como *el gran hedor*. Incidentalmente, fue esta la razón que motivó al parlamento a cambiar el sistema de drenaje de la ciudad.

⁵ John Snow. *On the Mode of Communication of Cholera*. John Churchill, London, 1849. URL <https://archive.org/details/b28985266/page/n3/mode/2up>. Accessed: 2025-04-23

Cuadro 1: Tabla de resumen de las muertes antes y después del cambio de la toma de agua por la compañía Lambeth. Fuente: Elaboración propia basada en Snow (1859), p. 90.

⁶ Jacob Wallace, Paul Goldsmith-Pinkham, and Jason L. Schwartz. Excess death rates for republican and democratic registered voters in florida and ohio during the covid-19 pandemic. *JAMA Internal Medicine*, 183(9):916–923, 2023. doi: 10.1001/jamainternmed.2023.1154. URL <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2807617>

su contribución más importante fue darle sentido a esos datos para transformarlos en información. No se limitó a recolectar datos: él entendía la importancia de explicar las causas y efectos y comunicarlos correctamente.

Esa es la esencia de la inferencia causal.



Figura 2: El mapa que John Snow trazó para identificar que el origen del cólera no era el *miasma*, sino el agua contaminada. Ubicar geográficamente los brotes en un mapa es algo que no se había hecho antes y ayudó a identificar con facilidad que la fuente era una bomba específica. Cerrar esa bomba de agua logró salvar algunas vidas, pero el modelo de pensamiento de Snow sigue salvando muchas vidas más hasta la fecha.

Resumen del capítulo

En este primer capítulo viajamos en el tiempo al Londres de 1854 para conocer a John Snow, el que podría ser el primer detective de datos de la historia, y su lucha contra el cólera.

Lo que hicimos fue usar su historia para entender la idea más importante de este libro. Vimos cómo Snow desafió una teoría dominante pero incorrecta (el miasma) no con opiniones, sino con un método brillante. Usó un **experimento natural** —el hecho de que una compañía de agua cambiara su fuente mientras otra no— para crear una comparación justa y demostrar sin lugar a dudas que la verdadera causa de la epidemia era el agua contaminada.

Esto es importante porque la historia de John Snow es la historia de origen de la inferencia causal. Nos enseña la lección fundamental: para encontrar la verdad, no basta con tener datos; necesitas un ‘diseño de comparación’ inteligente. Sin un grupo de control creíble (la compañía que no cambió su agua), la evidencia de Snow habría sido solo una correlación interesante. Esta idea de buscar comparaciones justas para

aislar la causa real es la base de todo lo que haremos en este libro.

¿Cómo te ayuda esto? Este capítulo te instala un nuevo ‘software’ en el cerebro. De ahora en adelante, cuando escuches que “A causó B”, tu primera pregunta no será sobre la estadística, sino sobre el método: “¿Con qué lo compararon?”. Te da el superpoder del escepticismo estructurado. Te enseña a pensar como un detective, buscando siempre el grupo de control o el “experimento natural” escondido en los datos para poder aislar la verdadera causa. Es la base para no caer en la desinformación y para empezar a construir argumentos sólidos.

Conviértete en un detective de datos: ejercicios introductorios

La inferencia causal empieza con la forma en que piensas. Estos ejercicios son para que practiques la lógica de detective que usó John Snow.

1. **El Miasma Moderno (Conceptual):** La teoría del ‘miasma’ era una explicación fácil pero incorrecta. Piensa en un “miasma moderno”: una correlación que la gente comúnmente confunde con una causa. (Ejemplo: “Las empresas que más invierten en publicidad en el Super Bowl son las más exitosas”). ¿Por qué esta afirmación podría ser una correlación y no una causa directa?
2. **El Grupo de Control es tu Ancla (Conceptual):** Imagina que John Snow solo hubiera tenido los datos de la compañía ‘Lambeth’ (la que mejoró su agua). Habría visto que las muertes bajaron y podría haber dicho que esa era la prueba. ¿Por qué su argumento habría sido mucho más débil sin los datos de la compañía ‘Southwark & Vauxhall’? ¿Qué otras explicaciones para la caída de las muertes no podría haber descartado?
3. **Diseñando una Comparación Justa (Conceptual):** Quieres saber si una nueva estrategia de descuentos (‘tratamiento’) realmente aumenta el valor promedio de compra en tu tienda en línea. Usando la lógica de John Snow, ¿cuál sería un buen “grupo de control”? ¿Por qué comparar las ventas de esta semana (con descuentos) con las de la semana pasada (sin descuentos) podría ser una comparación injusta?
4. **Encontrando Experimentos Naturales (Conceptual):** Un [experimento natural](#) ocurre cuando el mundo te “asigna” por accidente un grupo de tratamiento y uno de control. Describe un posible experimento natural que podrías usar para estudiar el efecto de:
 - a) El trabajo remoto en los precios de las rentas de oficinas. (Pista: piensa en una gran empresa que anuncia el teletrabajo permanente, afectando una zona de la ciudad más que otras).
 - b) Una nueva ciclovía en el número de accidentes viales. (Pista: la ciclovía se construye en una avenida principal, pero no en otra avenida paralela muy similar).
5. **Una Imagen Vale Más que Mil Números (Conceptual):** El capítulo dice que Snow no solo recolectó datos, sino que los comunicó correctamente. Además de su famosa tabla, ¿qué otra herramienta (visual) usó para que su argumento fuera tan convincente, y por qué crees que fue tan efectiva?
6. **Sentido Común y Estadística (Reflexión):** La frase clave del capítulo es “La inferencia causal se trata más de sentido común que de la estadística en sí misma”. ¿Significa esto que las matemáticas no son

importantes? ¿O que el ‘diseño’ del estudio es la base sobre la cual se construye cualquier análisis estadístico? Justifica tu respuesta usando el caso de John Snow.

7. **El Universo Contrafactual (Conceptual):** Para una familia que vivía en un distrito servido por ‘Lambeth’, el “resultado contrafactual” es lo que les *hubiera* pasado si ‘Lambeth’ *no hubiera* cambiado su fuente de agua. ¿Cómo usó John Snow los datos de la compañía ‘Southwark & Vauxhall’ como una ventana a ese universo contrafactual?
8. **Aplicando la Lógica al Negocio (Conceptual):** El director de una cadena de cafeterías observa que las sucursales que tienen más de 10 empleados tienen mayores ventas. Propone como política que todas las sucursales contraten hasta tener al menos 10 empleados para aumentar las ventas. Como analista de datos inspirado en John Snow, ¿qué preguntas harías antes de apoyar esa decisión? ¿Qué datos buscarías para determinar si la relación es causal o solo una correlación?

Regalo: Completa el curso educativo por email.

Si este es tu primera vez que te expones a la econometría y la inferencia causal, probablemente te parezca interesante este recurso adicional que he hecho.

Se trata de un curso educativo por correo donde explico

1. **El *mindset* de la inferencia causal.**
2. **Cómo diseñar un proyecto de investigación con econometría.**
3. **Y cómo comunicar los resultados**

Consíguelo gratis en <https://inferenciacausal.com>

O escanea el código QR:



Los negocios son matemáticas

Es mejor haber entendido por qué fallaste que ser ignorante de por qué tuviste éxito

- Robert A. Burgelman

CUANDO SALÍ DE LA UNIVERSIDAD, tenía un solo objetivo: abrir mi propio negocio. Había emprendido de muchas formas durante la carrera: había vendido promociones para una taquería que comenzamos en mi familia (falló) y había aprendido a crear productos que solucionan problemas y sabía vender. Ya estaba listo para las grandes ligas (según yo).

El año era 2009: justo en medio de la gran recesión.

La gran recesión afectó los ingresos de mi generación de una manera brutal⁷. Quienes salimos al mundo laboral ese año nos encontramos un escenario post-apocalíptico sin opciones de trabajo y con negocios cerrando por doquier.

Y fue en los negocios que cerraban donde vi oportunidad (no hagas esto en casa).

Mi café favorito estaba en venta. Era un café en el centro histórico, con una clientela establecida y operando al 100 %. Era la oportunidad perfecta para poner en práctica todo lo que había aprendido los últimos cuatro años en la universidad (o eso creía yo).

El único problema era que no tenía dinero.

Decidí juntar 10 amigos, hacerlos socios y comprar el negocio. Venía junto con una renta mensual de 11 mil pesos y un barista muy hábil, aunque un poco antipático. Hice modificaciones mínimas y abrimos al público ese mismo mes. Jamás olvidaré la sensación de abrir un negocio y comenzar a recibir clientes ese mismo día. Pensé que si seguía a ese mismo ritmo, recuperaría mi inversión en menos

⁷ Jesse Rothstein. The lost generation? labor market outcomes for post great recession entrants. *NBER Papers*, 2020. DOI: 10.3386/w27516. URL <https://www.nber.org/papers/w27516>; and Raymundo M. Campos-Vazquez, Gerardo Esquivel, Parama Ghosh, and Enrique Medina-Cortina. Long-lasting effects of a depressed labor market: Evidence from mexico after the great recession. *Labour Economics*, 81:102332, 2023. DOI: 10.1016/j.labeco.2023.102332. URL <https://doi.org/10.1016/j.labeco.2023.102332>

de un mes y me imaginaba como el próximo Steve Jobs antes de los 30.

Para noviembre de ese mismo año, acabé con una neumonía que casi me mata y el negocio quebrado.

Cometí dos errores grandes con ese negocio

Si no hubiera cometido estos errores, habría tenido un negocio exitoso en lugar del rotundo fracaso que viví.

- **El primer error fue no haberme ensuciado las manos lo suficiente.** Pensaba que mis conocimientos de administración eran suficientes para manejar cualquier situación. Si volviera al pasado, habría dedicado más tiempo a aprender a hacer todo en la operación del negocio. ¡Al menos hubiera aprendido a hacerla de barista!
- **Mi segundo error fue seguir demasiado mi intuición y muy poco a lo que decían los datos.** También fue por arrogancia. Saliendo de la universidad sentía que ya lo sabía todo y que todos los demás estaban equivocados.

Nunca más.

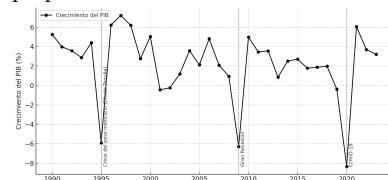
Usar datos en los negocios significa tener la humildad de aceptar que no lo sabemos todo

En retrospectiva, veo que no fue una idea absurda. No estaba loco, simplemente no entendí los números.

Para convencer a 10 amigos a que invirtieran conmigo usé gráficas del ciclo económico. Les expliqué que probablemente estábamos en el punto más bajo y que la economía no tenía más opción que subir. Los datos históricos me respaldaban, pero esta no era una recesión normal, sino una crisis financiera con raíces estructurales profundas..

Si hubiera sido más cuidadoso al revisar los datos, probablemente no habría tomado un riesgo tan alocado.

Figura 3: Crecimiento del PIB de México de 1990 a la actualidad. Como millennial, me ha tocado vivir 3 crisis: la crisis del peso mexicano (o efecto tequila), la llamada “gran recesión” y la crisis del Covid-19). Fuente: Elaboración propia con datos del Banco Mundial.



Este libro está diseñado para un mundo donde la Inteligencia Artificial puede hacer análisis de datos

La primera sección se enfoca en el *mindset* de la inferencia causal.

Hoy en día ya es posible subir una base de datos a ChatGPT y pedirle que limpie y prepare los datos para hacer análisis. Luego le puedes pedir que haga una **regresión lineal** y que te dé su interpretación de los resultados. Finalmente, le puedes pedir que haga las **pruebas de hipótesis** más comunes.

Pero incluso con todo eso, hay algo que la IA todavía no puede hacer: distinguir entre **correlación vs. causalidad**. Eso sólo lo vas a poder hacer tú. Porque eres tú el que tiene una *ventana de contexto* más completa⁸, tienes ojos y oídos y puedes salir al mundo a ver si tus hipótesis tienen sentido o no.

La segunda sección tiene elementos de negocios indispensables para generar crecimiento basado en datos.

El análisis de datos hoy en día es un elemento integral de los negocios. No se trata de un accesorio adicional: los datos son tu negocio. Cada paso que damos en negocios, lo debemos hacer tomando la evidencia como un elemento central.

En la última sección veremos los modelos más avanzados. Porque al inicio estaremos intentando hacer experimentos, pero cuando no es posible hacer uno, nos apoyaremos en los llamados **experimentos naturales** para hacer modelos de **diferencias en diferencias**⁹ que nos entreguen resultados con interpretación similar a la que tendríamos si hicieramos un experimento.

It's the economist way.

¿Qué necesitas saber antes de comenzar?

Lo más importante son tus ánimos de aprender y tu curiosidad.

Procuro a lo largo de este libro mantener las ideas a nivel intuitivo. Sin embargo, hay algunos temas que requieren conocimientos previos:

- En el capítulo de Regresión, cuando paso de la regresión con una variable a múltiples variables, explico el resultado final usando álgebra lineal. Es un segmento que te puedes saltar, pero me pareció indispensable explicarlo de esa manera.

⁸ La ventana de contexto es el número de *tokens* (palabras o frases) que una Inteligencia Artificial usa en su memoria para dar respuestas. Todos los días aumenta más, al grado de que pueden recordar bibliotecas enteras, pero aún así la forma en que los humanos damos contexto y conectamos ideas lo considero más potente.

⁹ Este es el modelo que más me interesa y el que siempre me aseguraba de enseñar en mis clases. En palabras simples, se trata de hacer una comparación de antes y después con un grupo similar que sirva como control para poder identificar los efectos. Es uno de los modelos más poderosos y de mayor crecimiento en la econometría.

- En general, a lo largo del libro, usamos modelos sencillos, pero es necesario tener una intuición sobre lo que representa una **variable aleatoria** y cómo funciona la probabilidad básica.
- Comenzamos el libro con un repaso de programación en Python. Aunque es muy básico, es suficiente para entender el código en los siguientes capítulos o al menos formar una intuición. Sin embargo, te recomiendo hacer más ejercicios para solidificar tu conocimiento.

Resumen del capítulo

En este capítulo te conté una historia de fracaso: la de mi primer negocio, un café que abrí con toda la arrogancia de un recién graduado y que quebró en meses.

Lo que hicimos fue usar esa experiencia para establecer la filosofía de este libro. Identificamos dos errores fatales: uno operativo, por no “ensuciarme las manos”, y uno mucho más profundo y analítico: confiar ciegamente en mi intuición en lugar de escuchar humildemente a los datos. Vimos que este libro está diseñado para un mundo con Inteligencia Artificial, donde tu verdadero valor no será correr el código, sino pensar, usar tu contexto único para diferenciar la causa de la casualidad.

Esto es importante porque el fracaso es el mejor maestro, y la lección fundamental que enseña es la humildad. Este capítulo argumenta que el análisis de datos no es una herramienta para genios que lo saben todo, sino todo lo contrario: es un acto de humildad. Es el reconocimiento de que nuestras coronadas pueden estar equivocadas y que necesitamos evidencia para navegar un mundo complejo. Es la justificación de por qué este libro existe: para darte el ‘mindset’ causal que ninguna IA puede replicar.

¿Cómo te ayuda esto? Este capítulo es el “porqué” de todo lo que aprenderás a continuación. Te prepara mentalmente para el viaje, demostrando que el objetivo no es memorizar fórmulas, sino desarrollar un nuevo instinto. Te da permiso para dudar de tu propia intuición y te muestra un camino para tomar decisiones más sólidas. Entender esto te ayudará a ver cada capítulo no como una lección de estadística, sino como un arma más en tu arsenal para tomar mejores decisiones en los negocios y en la vida.

Lecciones del fracaso: ejercicios de reflexión

La teoría se entiende mejor cuando se conecta con la experiencia. Estos ejercicios son para que reflexiones sobre las lecciones de este capítulo y las apliques a tu propio mundo.

1. **Tus dos errores (Reflexión personal):** En el capítulo confieso dos errores: uno operativo (“no ensuciarse las manos”) y uno analítico (“seguir la intuición sobre los datos”). Piensa en un proyecto —personal, académico o profesional— que no salió como esperabas. ¿Puedes identificar un error de ejecución y un error de juicio o análisis que hayas cometido?
2. **La cita inicial (Conceptual):** La cita de Robert Burgelman dice: “Es mejor haber entendido por qué fallaste que ser ignorante de por qué tuviste éxito”. ¿Cómo se aplica esta frase a la historia del café que

quebró? ¿Por qué es tan fundamental para el método de este libro el entender a fondo las causas de un fracaso?

3. **Intuición vs. Datos (Reflexión personal):** El autor admite que, por arrogancia, usó datos históricos del ciclo económico para justificar una corazonada. Piensa en una decisión importante que hayas tomado basándote fuertemente en tu intuición. En retrospectiva, ¿qué datos concretos podrías haber buscado para validar o refutar esa intuición antes de actuar?
4. **Tu valor en la era de la IA (Conceptual):** El capítulo argumenta que la IA puede correr una regresión, pero no diferenciar causa de correlación porque le falta “contexto”. Da un ejemplo de un conocimiento de contexto que tú tienes sobre tu trabajo, tu industria o tus estudios, que una IA —analizando solo una base de datos— no podría saber.
5. **El mapa del libro (Conceptual):** La estructura del libro es: 1) ‘Mindset’ Causal, 2) Aplicaciones de Negocio, 3) Modelos Avanzados. ¿Por qué crees que es crucial empezar con el ‘mindset’ antes de saltar directamente a las fórmulas y el código de los modelos más complejos?
6. **La humildad de los datos (Reflexión):** Se argumenta que “usar datos en los negocios significa tener la humildad de aceptar que no lo sabemos todo”. ¿Estás de acuerdo? ¿Por qué crees que a muchas personas y empresas les cuesta adoptar esta mentalidad y prefieren confiar en la “experiencia” o la jerarquía?
7. **“It’s the economist way” (Conceptual):** Esta frase se usa para describir el enfoque de usar ‘experimentos naturales’ y modelos como ‘Diferencias en Diferencias’ cuando un experimento real no es posible. Basado en lo que leíste en el capítulo anterior sobre John Snow, ¿qué crees que define “la forma del economista” de abordar un problema causal?
8. **Gestionando tus expectativas (Reflexión personal):** El capítulo lista algunos conocimientos previos recomendados (álgebra lineal, probabilidad, Python). ¿Cuál de estas áreas sientes que es tu punto más fuerte y cuál el más débil? ¿Qué acción podrías tomar esta semana para reforzar el área donde te sientes menos seguro?

Python para hacer Econometría

Todos en este país deben aprender cómo programar una computadora... porque te enseña cómo pensar.

– Steve Jobs

SI NO SABES CON QUÉ LENGUAJE COMENZAR A APRENDER ECONOMETRÍA, tu mejor opción es [Python](#).

En las escuelas enseñan con [Eviews](#) o con [Stata](#). [R](#) también es un lenguaje genial para la estadística: es el que he usado por años. Pero si tuviera que empezar de cero hoy a aprender econometría lo haría con Python.

La razón es que la [inteligencia artificial \(IA\)](#) ha cambiado la forma en la que programamos

- Python es un lenguaje de programación que se puede usar para hacer de todo, no sólo para estadística.
- Esto quiere decir que puedes integrarlo con diferentes soluciones y hacer productos con tus datos¹⁰.
- Estamos en una nueva era donde la inteligencia artificial (IA) es la que se encarga de crear y modificar el código y tú eres el encargado de pensar¹¹.

Hace un par de años programar en Python era una barrera gigante, hoy es relativamente trivial¹².

Lo que importa es entender los modelos de la econometría, cómo funcionan y cómo usarlos. Esto requiere que nos adentremos a la filosofía sobre cómo entendemos causas y efectos y a conocer bien las herramientas a nuestra disposición.

Y nadie lo entendió mejor que una enfermera Rockstar del Siglo XIX.

¹⁰ Por ejemplo, puedes hacer una API completa para comunicarte con estadísticas oficiales y desplegarlos en un *dashboard*.

¹¹ Jingxuan Liu, Chenshuo Xia, Yihan Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023. URL <https://arxiv.org/abs/2305.01210>

¹² En mis clases de investigación de operaciones, les invito a mis alumnos a que creen y modifiquen código de Python con ayuda de IA. Aún tiene su reto pensar, pero la implicación es que se requiere un conocimiento más general de todo lo que puedes hacer con el lenguaje de programación y no pasando horas tecleando. Dicho esto, saber de algoritmos y entender la sintaxis del lenguaje sigue siendo el atajo para ahorrarse muchos dolores de cabeza.

Antes de que existiera Python, teníamos a Florence Nightingale

Florence Nightingale fundó la enfermería moderna, pero salvó aún más vidas gracias a su genialidad estadística.

La conocían como la dama de la linterna. Se le veía por las noches rondando en ayuda de los soldados durante la guerra de Crimea¹³. Se ofreció como voluntaria junto a un equipo de 38 enfermeras para atender a los heridos en combate.

Ahí fue donde hizo la contribución más grande a la estadística, que la hizo famosa.

No es sino hasta que se registran datos de forma meticulosa que los patrones comienzan a emerger.

Cuando llegó con su equipo a la guerra, se dio cuenta de que los malos cuidados médicos cobraban más vidas que las balas del enemigo. Había pocas medicinas, se ponía poca atención a la higiene y las infecciones eran comunes. Nightingale comenzó a hacer registros cuidadosos de todo y lo comunicó al gobierno británico.

Impulsó cambios importantes que **redujeron las muertes** de 42 % a 2 % en el hospital.

Convenció al gobierno británico de estos cambios gracias a gráficas innovadoras como la figura 5.

El gráfico de arriba se llama diagrama de rosa. El área azul representa las muertes por enfermedades infecciosas prevenibles, el área roja son las muertes por heridas en batalla y el área negra son otras causas. El poder de este gráfico es que se vuelve evidente de inmediato lo importante que es la higiene para prevenir muertes en el hospital.

En la época, hacer este tipo de gráficos requería muchas horas de trabajo. Hoy puedes hacerlo en minutos gracias a Python.

El primer paso para hacer econometría con Python

No necesitas instalar nada para empezar a hacer econometría.

Python es un lenguaje de programación general que puedes instalar en tu computadora. Sólo necesitas [descargarlo](#), instalarlo y descargar los módulos apropiados. Hacerlo de esta forma requiere un poco de experiencia y que sepas usar la terminal, entre otras cosas.

¹³ Annie Gracey Swenson. *Medical Women of America: A Short History of the Pioneer Medical Women of America and a Few of Their Colleagues in England*. Monarch Book Company, Chicago, 1910. URL https://archive.org/details/medicalwomenvict00swen_0/page/n6/mode/1up



Figura 4: Florence en su estudio. Fuentte: elaborado por el autor con Dall-e

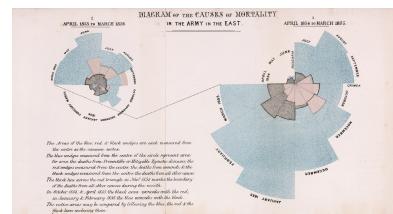


Figura 5: El Diagrama de la Rosa que Florence Nightingale presentó y que demostró que las muertes venían más por las enfermedades que por la batalla.

Pero hay una forma más fácil que no requiere descargas ni conocimiento técnico.

- Entra a [Google Colab](#)¹⁴.
- Si no tienes cuenta en Google, tiene que crearla.
- Da click en Nueva notebook (cuaderno Jupyter) y comienza a trabajar.
- Para ejecutar un bloque de código, sólo necesitas picar el botón con el símbolo de play  del lado izquierdo del bloque de código o usa `ctrl + Enter` (`cmnd + Enter` en Mac)

¹⁴ colab.research.google.com

Los notebook de Google Colab son la mejor forma de hacer ciencia de datos con Python, porque trabajar con datos requiere mucha prueba y error. Repetir pequeños bloques de código una y otra vez sin alterar el resto del programa. Si nunca has usado Python, las notebook te ayudarán a comenzar sin preocuparte por los detalles de instalación¹⁵. Si eres un experto en Python te será evidente los dolores de cabeza que estás evitando al trabajar así¹⁶.

Familiarízate con Python

La mejor forma de aprender cualquier lenguaje de programación es jugar con él.

Observa las siguientes operaciones. Ejecuta y modifica los ejemplos a tu gusto. Intenta predecir lo que vas a obtener de resultado antes de ejecutar.

- Sumas y restas (`2+2`, `8-3`).
- Multiplicaciones y divisiones (`7*2`, `9/3`).
- Potencias y raíces cuadradas (`3**2`, `16**(1/2)`).
- Concatenación de cadenas de texto (`'Hola' + '' + 'mundo'`).¹⁷
- Comparaciones (`5 > 3`, `'hola' == 'adiós'`)

Pero este libro no se trata de aprender a programar.

Si quieres ser un maestro de la econometría, no necesitas reinventar la rueda. Puedes usar el código que hicieron otras personas.

El código de los expertos.

¹⁵ En mis clases usamos google Colab. Antes teníamos que perder muchas clases en la instalación y configuración, Colab tiene todo listo para comenzar.

¹⁶ En particular si estás tratando de trabajar en local. La recomendación para trabajar en local con Python es que cada proyecto tenga su propio ambiente virtual y que cada ambiente cargue sus propios módulos. En si, es una filosofía de trabajo que rompe con la forma en que estamos acostumbrados a trabajar con software como Excel, donde abrir la aplicación implica cargar todo lo que necesitas para trabajar. Este es un punto medio que no requiere que seas un experto en ciencias de la computación, pero tampoco te deja sin la posibilidad de personalizar tu ambiente de trabajo.

¹⁷ Nota que Python interpreta el símbolo de suma de manera muy relajada. Las de cadenas de texto (el texto entre comillas) no se pueden sumar, así que de manera automática toma la decisión de concatenar el texto. Por eso, el código incluye un espacio.

Módulos de Python para hacer econometría

Un módulo es un paquete con funciones que hizo alguien más para solucionar un problema.

A diferencia de los programas estadísticos tradicionales como Stata o Eviews, Python requiere de paquetes especiales para hacer econometría. Cada paquete contiene funciones particulares para lo que deseas. La diferencia es que esos paquetes con funciones se descargan aparte y es necesario llamarlos cuando los quieras usar.

Aquí hay algunos módulos de Python que son útiles para hacer econometría:

- **Statsmodels** : Modelos estadísticos y herramientas para realizar análisis de datos.
- **Pandas** : Estructuras de datos flexibles y eficientes para manipular y analizar datos. Útil para trabajar con datos en formato tabular.
- **NumPy**: Biblioteca para el cálculo numérico en Python. Proporciona funciones y herramientas para trabajar con arreglos numéricos.

En capítulos posteriores veremos ejemplos de **statsmodels** y **numpy**, porque son los módulos que se usan para hacer modelos estadísticos y manipulación avanzada de datos.

Aquí aprenderemos a usar **pandas** para manejar datos de manera visual y tabular.

Un juego de lotería con Python

La Lotería es un juego tradicional mexicano parecido al juego de Bingo.

En este juego, en lugar de números se sacan tarjetas con diferentes personajes u objetos como una campana, la muerte o un borracho.

Al inicio del juego se reparten cartas con pictogramas distribuidos de manera aleatoria. El jugador debe marcar lo que aparece en su carta. Gana el jugador que marca su carta completa.

Algunos conceptos importantes:

- **Mazo**. Es el conjunto de todas las cartas individuales.



Figura 6: La Lotería es un juego de azar tradicional mexicano. Imagen de Alex Covarrubias, vía Wikipedia (Dominio público)

- **Carta.** Cada carta tiene una imagen y su nombre. Por ejemplo: El valiente.
- **Tabla.** Cada jugador tiene una tabla con 16 cartas aleatorias que debe de llenar conforme el gritón las menciona.
- **El gritón.** Es la persona encargada de dar a conocer la siguiente carta a todos los jugadores.

Para comenzar, invocamos los módulos.

```
import pandas as pd
import random
```

Con `pandas` ahora tienes el poder de crear y manipular bases de datos.

Con `pandas` puedes cargar datos desde archivos `csv` o Excel, visualizarlo, quitar filas, cambiar columnas y hacer lo que sea con tus datos. Como nuestro proyecto es una lotería, vamos a necesitar números aleatorios, que son la especialidad del módulo `random`.

Comencemos a incluir las cartas en una lista.

```
cartas = ["La maceta", "El borracho", "La campana", "El catrin", "El violoncello", "La sandia",
          "La chalupa", "El gorrito", "El arpa", "El camaron", "El barril", "La dama", "La bota", "El
          pajaro", "El melon", "El cotorro", "La palma", "El mundo", "El apache", "El pescado", "La
          muerte", "El alacran", "El gallo", "La calavera"]
```

Los conocedores de la lotería se podrán dar cuenta de que me faltó poner algunas cartas.

No es problema. Podemos incluir las cartas que nos faltan más adelante. Usemos la función `append()` para agregar la carta de “El diablito” a nuestro mazo de cartas. También podemos usar `extend()` para agregar más elementos al mazo desde otra lista.

```
# Con append podemos agregar un elemento adicional que nos faltaba
cartas.append("El diablito")
# Con extend podemos agregar los elementos de una lista a otra
cartas.extend(["El valiente", "La corona", "El barril"])
# Con print() mostramos
print(cartas)
```

Con esta lista ya podemos repartir las tablas

Una tabla de lotería tiene 16 cartas: cuatro a lo ancho y cuatro a lo largo. Usa `pandas` para crear un `DataFrame` para la tabla de cada jugador. Este `DataFrame` tendrá 16 cartas únicas del mazo y una columna adicional para marcar las cartas.

```
# Crear un dataframe con una columna de carta
deck_df = pd.DataFrame(cartas, columns=['Carta'])
```

Y el siguiente código crea una tabla con 16 cartas asignadas de manera aleatoria. Esta es la tabla que te reparten al inicio del juego. Incluimos una columna para marcar si la carta ya fue cantada o no.

```
def crear_tabla(deck_df):
    # Usamos sample(16) para tomar 16 cartas aleatorias sin repetición
    # .sample() es un método de pandas, y aquí lo usamos directamente
    # desde el DataFrame deck_df. También podrías importar solo sample
    # desde pandas, pero aquí lo dejamos todo a través del objeto DataFrame.
    tabla = deck_df.sample(16).reset_index(drop=True)
    tabla['Marcada'] = False # Agregar una nueva columna para marcar las cartas
    return tabla

# Ejemplo de la creación de una tabla para un jugador
tabla_jugador = crear_tabla(deck_df)
print("Tabla del Jugador:")
print(tabla_jugador)
```

Como el juego apenas está por comenzar, todas las casillas de la tabla deben de comenzar indicando `False`.

	Carta	Marcada
0	El valiente	<code>False</code>
1	El violoncello	<code>False</code>
2	El alacran	<code>False</code>
3	La palma	<code>False</code>
4	El gallo	<code>False</code>
5	El gorrito	<code>False</code>

La función `crear_tabla` es una función hecha por nosotros, que toma la lista de cartas y utiliza `random.sample` para seleccionar 16 cartas únicas de esa lista. `random.sample` es útil porque automáticamente se asegura de que no haya duplicados en la selección.

La columna “Marcada” aparece como una lista de “Falsos” porque es una de esas variables que sólo toma como valores Falso y Verdadero. Al comenzar, todos son falsos porque el gritón aún no “canta” ninguna de las cartas.

Cantar las cartas

Ahora que tenemos las tablas, necesitamos una forma de “cantar” las cartas y que los jugadores revisen sus tablas.

Podemos hacer esto con otra función muy sencilla. Hemos creando `cantar_carta()`, que selecciona una carta del mazo de forma aleatoria.

```
def cantar_carta(deck_df):
    return deck_df.sample().iloc[0]['Carta']

# Ejemplo de cantar una carta
carta_cantada = cantar_carta(deck_df)
print("Carta Cantada:", carta_cantada)
```

Carta Cantada: La calavera

Este código tiene un error. Normalmente cuando el “gritón” canta las cartas de lotería, ya no las reemplaza en el mazo y no volverán a salir. Para solucionar esto tendríamos que hacer que se elimine el elemento de la carta, pero dejaremos esto como ejercicio al lector.

Esta función selecciona una carta al azar de la lista de cartas. Cada vez que se llama a la función, simula al “gritón” cantando una nueva carta.

Marcando las Cartas

El Data Frame `tabla_jugador` tiene dos columnas. La primera tiene nuestras cartas y la segunda nos ayuda a marcar si el gritón ya dijo nuestra carta. Es nuestra columna de frijolitos¹⁸.

En la siguiente función primero se verifica si la carta cantada está en nuestra tabla. Lo hacemos con la palabra clave `if`, que cambia el elemento de la segunda columna a `True` si la tenemos.

¹⁸ Es común poner un frijol o algún objeto que tengas a la mano en tu tabla, para tener marcadas las cartas que ya te salieron y darte cuenta cuando completes tu lotería.

```
def marcar_carta(tabla, carta_cantada):
    if carta_cantada in tabla['Carta'].values:
        tabla.loc[tabla['Carta'] == carta_cantada, 'Marcada'] = True
        print("¡Carta marcada!")
    else:
        print("Esta carta no está en tu tabla.")

# Ejemplo de uso de la función para marcar la tabla con la tarjeta
marcar_carta(tabla_jugador, carta_cantada)
print(tabla_jugador)
```

En el ejemplo de arriba, Python sacó la carta “El gallo” y la marcó verdadera en nuestra base de datos.

```
¡Carta marcada!.
   Carta  Marcada
0   El valiente  False
1  El violoncello  False
2    El alacran  False
3     La palma  False
4      El gallo  True
5     El gorrito  False
6      La sandia  False
7      La muerte  False
8       El arpa  False
9      La chalupa  False
10     El catrin  False
11     La maceta  False
12     El cotorro  False
13      La dama  False
14      La bota  False
15      El mundo  False
```

De lo contrario, el sistema simplemente nos dirá que la carta no está en nuestra tabla y podemos volver a pedir al gritón que cante otra carta.

Inténtalo. Es divertido.

Verificando el Ganador

La función `all()` verifica si todos los valores en la columna “Marcada” tiene valor verdadero. Es una forma rápida de encontrar al ganador.

Puedes seguir cantando y marcando cartas hasta que completes tu

tabla de lotería.

```
def verificar_ganador(tabla):
    return all(tabla['Marcada'])

# Ejemplo de verificar si nuestra tabla es ya ganadora
if verificar_ganador(tabla_jugador):
    print("¡Felicitaciones, has ganado!")
else:
    print("Sigue jugando.")
```

Sigue jugando.

Cuando tu tabla tiene todas las cartas marcadas, has ganado¹⁹.

Preguntas frecuentes en las primeras sesiones de Python

Tengo ya bastante experiencia enseñando a programar por primera vez para reconocer los problemas más comunes al inicio. Regresa a esta lista si te encuentras en problemas, tal vez encuentres tu problema aquí.

- **Me apareció error.** En los lenguajes de programación no hay errores genéricos. Siempre tienes que revisar con detalle.²⁰
- **No puedo cargar el módulo.** Asegúrate que estés utilizando el entorno correcto donde el módulo está instalado. Si estás usando [Google Colab](#), los módulos más comunes como `pandas` y `numpy` ya están preinstalados. Si estás en tu propia máquina, quizás necesites instalar el módulo usando pip, por ejemplo, `pip install pandas`.
- **Mi código no hace lo que espero.** Revisa cada línea cuidadosamente. Asegúrate de entender qué hace cada parte del código. A veces, un pequeño error como una letra mal escrita o una indentación incorrecta puede causar problemas²¹.
- **No entiendo el error que me muestra Python.** Los mensajes de error pueden ser confusos al principio. Lee el mensaje completo. A menudo, la última línea te da una pista sobre lo que está mal. Si no entiendes el mensaje, intenta buscarlo en internet. Es muy probable que alguien más haya tenido el mismo problema.

¹⁹ Mi recomendación es que hagas todo este código a mano en Colab y ejecutes cada uno de los bloques a la vez hasta que entiendas bien qué hace cada uno. Para cuando llegues al final, si es entretenido ejecutar el código una y otra vez hasta que ganas. Intentalo por tu cuenta.

²⁰ Pro tip: Los errores en Python se leen de abajo para arriba. Lo que sale al final es la descripción del error y de ahí vas subiendo y revisando los detalles.

²¹ Python es particularmente especial con la identación (los espacios a la izquierda). La identación sirve para identificar si una parte del código pertenece a una función o a un ciclo `for`. Pon mucha atención a esos detalles.

- **El código se ejecuta, pero no pasa nada.** Verifica que estés llamando a las funciones correctamente y que estés pasando los argumentos correctos. También asegúrate de que cualquier cambio que esperes ver se esté mostrando o guardando adecuadamente²².
- **¿Cómo instalo un módulo de Python?** Generalmente, puedes instalar módulos de Python usando pip. Por ejemplo, para instalar `matplotlib`, usarías `pip install matplotlib` en tu terminal o línea de comandos²³.
- **¿Cómo sé qué módulo usar para una tarea específica?** La experiencia te ayudará a conocer qué módulos son mejores para diferentes tareas. Mientras tanto, busca recomendaciones en línea o en libros de texto sobre Python. La comunidad de Python es muy activa y hay muchos recursos disponibles.
- **¿Cómo puedo mejorar en Python?** La práctica es clave. Trabaja en pequeños proyectos, resuelve problemas y trata de leer y entender el código de otras personas. También, participar en comunidades en línea y foros puede ser muy útil.

Ya sabes usar Python, ahora aprendamos econometría

Lo más importante es practicar.

No hay libro que te dé la suficiente experiencia antes de comenzar a construir tus propios modelos. Necesitas empezar hoy mismo a modelar, a obtener datos y a jugar con ellos. Es a prueba y error que tu mente te hará un experto en econometría.

Es momento de comenzar con las matemáticas.

Resumen del capítulo

Python te da superpoderes de econometrista.

No es una exageración. Lo que a Florence Nightingale le tomó semanas de trabajo meticoloso para tabular, calcular y dibujar a mano, tú lo puedes replicar en minutos con unas pocas líneas de código. Si te pidiera que calcularas los coeficientes de una regresión hace 100 años, tendrías que sudar sobre los números por días; hoy lo haces con un comando en segundos. Es lo más cercano a la magia de verdad que conozco.

Pero como en toda historia de superhéroes, hay un truco: los poderes no vienen solos, hay que aprender a usarlos.

²² Python puede crear una función sin ejecutarla. Tienes que verificar que estás llamando la función y que tenga algo que mostrar. En el código de arriba, usamos regularmente la función `print` para mostrar algo en la consola.

²³ En Colab, este tipo de funciones se llaman usando `!` antes. Por ejemplo, `!pip install matplotlib`. Esto es para indicar que el comando es uno que debe de ejecutar la terminal, y no el notebook. Es muy útil saber esto para instalar módulos más allá de los que tiene Colab de manera estándar.

Y desgraciadamente, ahora mismo, todavía no sabes usar Python.

El camino para dominar un lenguaje de programación requiere horas frente a la computadora. No se aprende leyendo, se aprende haciendo. Te vas a equivocar, vas a ver mensajes de error rojos y frustrantes, y vas a tener que solucionar problemas por tu cuenta. Cada error que arregles, cada función que logres hacer funcionar, es como subir de nivel. Entre más tiempo le dediques, mejorarán tus habilidades para hacer magia con los datos.

Este libro no te va a hacer un experto de la noche a la mañana, pero te dará el mapa y las herramientas. Los siguientes ejercicios son tu primer entrenamiento. No los saltes. Ábrete un notebook en Google Colab y hazlos uno por uno. Empieza a construir tu poder.

Manos a la obra: ejercicios de práctica

Los siguientes ejercicios están diseñados para que los hagas en un notebook de Google Colab. Van de lo más básico a lo más complejo, preparándote para los análisis que haremos en los próximos capítulos.

1. **Variables y Operaciones Básicas:** Crea dos variables, `gasto_tv = 230.1` y `gasto_radio = 37.8`. Crea una tercera variable, `gasto_total`, que sea la suma de las dos anteriores. Imprime el resultado.
2. **Listas en Python:** Crea una lista llamada `medios` que contenga los strings: `'TV'`, `'Radio'` y `'Newspaper'`. Usa la función `append()` para añadir `'Redes Sociales'` a la lista. Imprime la lista final.
3. **Tu primera función:** Escribe una función en Python llamada `a_miles` que reciba un número (ej. las ventas) y lo divida entre 1000. Llama a la función con el número `15300` e imprime el resultado.
4. **Introducción a NumPy:** Importa la librería NumPy como `np`. Crea un array de NumPy a partir de la siguiente lista de ventas: `ventas_lista = [22.1, 10.4, 9.3, 18.5, 12.9]`. Usa las funciones de NumPy para calcular la media (`np.mean()`) y la suma total (`np.sum()`) de las ventas.
5. **Introducción a Pandas:** Importa la librería Pandas como `pd`. Crea un DataFrame a partir del siguiente diccionario: `datos = {'Canal': ['TV', 'Radio', 'Periodico'], 'Inversion': [1000, 500, 200]}`. Imprime el DataFrame.
6. **Cargando datos reales:** Busca en internet el archivo `advertising.csv` que usamos en este libro (está en repositorios como Kaggle o en el GitHub del libro). Sube el archivo a tu entorno de Google Colab y usa pandas para cargarlo en un DataFrame llamado `datos_publicidad`.
7. **Selección de datos:** Del DataFrame `datos_publicidad`, selecciona únicamente la columna `'Sales'` y guárdala en una variable llamada `y`. Luego, selecciona las columnas `'TV'`, `'Radio'` y `'Newspaper'`.

y guárdalas en un nuevo DataFrame llamado `X`.

8. **Visualización simple:** Importa `matplotlib.pyplot` como `plt`. Crea un diagrama de dispersión (`plt.scatter()`) que muestre la relación entre la inversión en `'TV'` (eje x) y las `'Sales'` (eje y). Añade etiquetas a los ejes.
9. **Filtrando datos:** Crea un nuevo DataFrame llamado `inversion_alta_tv` que contenga únicamente las filas de `datos_publicidad` donde el gasto en `'TV'` fue mayor a 250. ¿Cuántas filas tiene este nuevo DataFrame? (Usa `len()` o `.shape`).
10. **Preparando datos para la regresión (Parte 1):** El capítulo de regresión te mostrará que para calcular los coeficientes, necesitamos una matriz `X` que incluya una columna de unos para el intercepto. Usando NumPy, convierte la columna `'TV'` del DataFrame a un array de NumPy. Luego, crea un nuevo array que contenga dos columnas: una de puros unos y la otra con los datos de `'TV'`. (Pista: investiga la función `np.column_stack`).
11. **Operaciones con matrices en NumPy:** Usando la matriz de 2 columnas que creaste en el ejercicio anterior (llamémosla `X_matrix`), calcula $X'X$ (la transpuesta de `X` multiplicada por `X`). Imprime el resultado y sus dimensiones (`.shape`). (Pista: la transpuesta se obtiene con `.T` y la multiplicación de matrices con el operador `@`).
12. **Correlación:** Usando el DataFrame original `datos_publicidad`, calcula la matriz de correlación entre todas las variables. ¿Qué par de variables (sin contar una variable consigo misma) tiene la correlación más alta? (Pista: usa el método `.corr()`).
13. **Aplicando una función a una columna:** Crea una nueva columna en `datos_publicidad` llamada `'Sales_en_miles'` aplicando la función `a_miles` que creaste en el ejercicio 3 a cada elemento de la columna `'Sales'`. (Pista: investiga el método `.apply()`).
14. **Reto - Simulación simple:** Simula 100 lanzamientos de un dado de 6 caras usando `np.random.randint(1, 7, 100)`. Guarda los resultados en un array. Calcula cuántas veces salió el número 6.

Regalo: Repositorio en Github con el código completo

Todo el código yo mismo lo probé. La forma más fácil en que lo puedes comprobar por tu cuenta es abriendo sesión en Google Colab y comenzar a poner el código. Todas las bases de datos las puedes encontrar en el repositorio en línea que está en el siguiente enlace:

<https://github.com/marionomics/econometria>

O escanea el código QR:



El modelo de resultados potenciales

Algebra's like sheet music: the important thing isn't "can you read music". It's "can you hear it".

Can you hear it?

– Niels Bohr en la película Oppenheimer (2023)

*Pensaba que contigo iba a envejecer
En otra vida, en otro mundo podrá ser*

– Bad Bunny

PARA ENTENDER LA ECONOMETRÍA NECESITAS COMPRENDER EL MULTIVERSO.

El multiverso se basa en la interpretación de Hughes Everett (1957) de la mecánica cuántica ²⁴. Pero la interpretación popular que aparece en las películas y cómics es que cada decisión que tomamos genera un universo nuevo donde las cosas podrían ser muy diferentes a lo que conocemos como realidad.

Por ejemplo: hay un universo en el que no te gusta el chocolate²⁵.

Le vamos a llamar a ese universo un **contrafactual**, porque se opone a la realidad. Naturalmente, el hubiera no existe y no podemos obtener datos de los contrafetales. Sólo podemos imaginarlo.

Así que vamos a imaginarlo.

¿Cómo sería tu vida si hubieras entrado en Harvard?

¿Tendrías hoy mejores ingresos?

La relación entre la educación y los ingresos es una pieza clave de la ciencia económica. Hay un mar de teorías respecto a la existencia de un efecto positivo: entre mayor educación mayores ingresos, y hay muchos estudios dedicados a entender cómo funciona esta relación causal y cuál es su intensidad.

²⁴ Paul Busch, Teiko Heinonen, and Pekka Lahti. Heisenberg's uncertainty principle. *Physics Reports*, 452(6):155–176, 2007. DOI: 10.1016/j.physrep.2007.05.006

²⁵ No creo que sea un universo en el que te gustaría estar, pero cada quién.

El trabajo que inició todo es el de Gary Becker, en el que explica que la educación, el entrenamiento y la experiencia son iguales a la inversión que se hace en tecnología para el capital, pero en este caso conforman *capital humano*. Cuando vamos a estudiar, lo hacemos con la esperanza de que esto nos genere beneficios futuros, como ingresos mayores.

Se vería algo así como²⁶

$$w = w_0 e^{rS}. \quad (1)$$

Donde w_0 es el salario inicial sin educación, S serían los años de educación, y r la tasa de retorno de la educación. Es un modelo simple, pero cuenta una historia muy poderosa en una época en la que la econometría disponible no permitía evaluar si esta relación era causal.

De acuerdo a Heckman, Lochner y Todd²⁷, el retorno de la educación no es el mismo para todos. Igual que Beckman, ellos consideran que la educación es una inversión que genera retornos a largo, plazo, pero en su análisis causal, ellos identifican que el efecto es diferente para personas dependiendo de su habilidad individual, su entorno familiar y la calidad de su educación. Los resultados son similares en otros estudios que usan inferencia causal para estudiar el efecto de la educación en los ingresos²⁸

Respecto a la calidad de la educación, surge una duda: ¿qué tan importante es en si la calidad de estudiar en una institución de élite?

Estoy hablando de la diferencia que tiene estudiar en una institución como Harvard, el MIT o Stanford. En Estados Unidos, esas universidades de gran prestigio son muy codiciadas, y ganarse un lugar ahí es un sueño al que muy pocos pueden acceder. Por eso en la película de *Spiderman: No way home*, Peter Parker llega a los extremos de poner en riesgo el multiverso porque no lo aceptaron en el MIT (y porque no se le ocurrió mandar una carta de reconsideración)²⁹.

La pregunta es: ¿Realmente era para tanto? Probablemente Peter Parker podría haber ido a una universidad local. Su título no vendría con el gran nombre de las universidades que aparecen en las películas, pero tienen la misma validez y se habría desarrollado profesionalmente y habría sido feliz con su novia y su amigo³⁰. Al menos eso es lo que Dale y Krueger estiman en un estudio en el que encuentran que, para la mayoría de los estudiantes, asistir a una universidad selectiva no tiene básicamente ningún impacto en sus ingresos.

²⁶ Gary S. Becker. Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5, Part 2):9–49, 1962. DOI: 10.1086/258724. URL <https://www.journals.uchicago.edu/doi/10.1086/258724>

²⁷ James J. Heckman, Lance J. Lochner, and Petra E. Todd. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Econometrics*, 144(1):306–348, 2008

²⁸ David Card. The causal effect of education on earnings. In *Handbook of Labor Economics*, volume 3, pages 1801–1863. Elsevier, 1999; and Gladys Lopez-Acevedo. Mexico: Two decades of the evolution of education and inequality. *World Bank Policy Research Working Paper*, (3919), 2006

²⁹ Spider-man: No way home, 2021

³⁰ En una versión más cuerda de la película, el Dr. Strange simplemente le dice a Parker que lo deje en paz y entren él y sus amigos a una universidad regular, como cualquier otra persona habría hecho. Es una película más aburrida, pero al menos es un universo en el que Peter Parker la pasa bien.

Siendo justos, una de las características distintivas del personaje de Peter Parker es que es de bajos ingresos. En este caso, Dale y Krueger sí identifican que ir a una escuela selectiva tiene efectos positivos³¹. Lo mismo para quienes vienen de grupos minoritarios. Esto es un aspecto clave para entender cómo funciona la causalidad.

La razón por la que las personas de grupos minoritarios y de bajos ingresos si observan una mejora en sus ingresos por entrar a una escuela selectiva es porque la naturaleza de estas escuelas *excluye* de manera natural a aquellos que no tienen los recursos o los contactos para entrar en ellas. Dicho de otra forma: los papás de los alumnos que entran en las universidades más exclusivas ya tienen los ingresos y las conexiones que les aseguran un mejor ingreso, independientemente de si entran a una universidad selectiva o no. Por otra parte, a los chicos de un origen menos privilegiado, entrar a una escuela donde pueden hacer conexiones que les permitirán mejores oportunidades es una gran ventaja.

Son dos tipos de estudiantes **fundamentalmente diferentes**.

Para identificar el efecto, debemos ir más allá de la diferencia de medias

Esto es a lo que yo le llamo el *error de novatos*, cuando se trata de hacer identificación causal.

Supongamos que deseas saber el **efecto** que tiene entrar en una escuela más selectiva en los ingresos. Decides entonces que es buena idea evaluar la diferencia entre los ingresos promedio de las personas que fueron a una escuela selectiva Y_1 y el de los egresados de una no selectiva Y_0 .

$$E[Y_1] - E[Y_0] \quad (2)$$

La E denota la **esperanza** o valor esperado de lo que está entre paréntesis³². Normalmente denotamos con Y el **resultado** o la variable en la que esperamos ver un efecto. En este caso es el ingreso.

El subíndice 1 o 0 nos indica el **grupo** al que pertenece la variable. Un 1 nos da al **grupo de tratamiento** y un 0 al **grupo de control**. Entonces, Y_1 se lee como los ingresos de un individuo del grupo de tratamiento: de alguien que sí asistió a una escuela selectiva; mientras que Y_0 son los ingresos de alguien que no asistió a una escuela selectiva.

³¹ Dale S. Berg and Alan B. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 2002

³² La $E(\cdot)$ se lee como el *valor esperado* o esperanza. Es un concepto más general que simplemente el promedio, pues aplica a más que sólo números.

Tratamiento y **control** son parte del lenguaje en los estudios clínicos³³. Cuando quieras hacer un experimento para saber si una medicina funciona, divides a tus sujetos en dos grupos y a uno le aplicas la medicina y al otro no. Luego mides los resultados. En este caso, la "medicina" sería la asistencia a una escuela selectiva.

Hagamos una simulación de una base de datos que identifica a 180 alumnos. La mitad de ellos entró a una escuela selectiva y la otra mitad no lo hizo. El siguiente bloque de código nos regresa una tabla con estos estudiantes ficticios, un nivel de habilidad aleatorio y un nivel de ingresos definido por la ecuación

$$Y_i = 1000 + 250D_i + 250A_i + \varepsilon_i, \quad (3)$$

donde Y_i indica el ingreso del estudiante i , D_i es una **variable indicadora (dummy)** que indica si el estudiante ingresó a una escuela selectiva, y A_i indica la habilidad del estudiante. La letra griega ε (se lee *epsilon*) indica un término de error que nosotros llamaremos con la creación de un número aleatorio con distribución normal.

Este es el código que genera la tabla:

```
import pandas as pd
import numpy as np
import random

# Creando un diccionario con 'id' como claves y range(179) como valores
alumnos = {'id': list(range(180))}

# Creando el DataFrame
df = pd.DataFrame(alumnos)

# La semilla ayuda a tener el mismo resultado
random.seed(42)
np.random.seed(42)

# La habilidad es aleatoria.
df['habilidad'] = [np.random.normal(0, 1) for _ in range(len(df))]
ruido_habilidad = np.random.normal(0, 0.2, len(df))

# Añadiendo la columna 'selectivas' con una elección aleatoria de 0 o 1 para cada fila
df['selectivas'] = (df['habilidad'] + ruido_habilidad > 0.5).astype(int)

# Añadiendo la columna 'ingresos' con el cálculo especificado
df['ingresos'] = 1000 + df['selectivas'] * 250 + df['habilidad'] * 250 + np.random.normal(0,
                                         280, len(df))

# Mostrando las primeras filas del DataFrame
```

³³ Requiere un poco de imaginación identificar cuál sería el tratamiento en cada caso particular. Hay estudios en los que es muy evidente. Por ejemplo, una subida en el salario mínimo podría verse como un tratamiento, pero se siente un poco raro cuando el tratamiento es el color de piel. Basta recordar que es un recurso útil para identificar causas y efectos.

```
| df.head()
```

Grupo	Habilidad promedio	Ingresos promedio	N
Escuela selectiva	1.018	1616.3	48
Escuela no selectiva	-0.477	1106.4	132

Notas: La variable **habilidad** se generó aleatoriamente y afecta tanto la probabilidad de ingresar a una escuela selectiva como los ingresos.

El [sesgo de selección](#) se observa claramente: quienes ingresaron a una escuela selectiva tienen mayor habilidad en promedio. Comparar directamente los ingresos promedio entre grupos sin controlar esta diferencia produciría un estimador sesgado del efecto del tipo de escuela.

De acuerdo a la forma en que estamos creando los ingresos en esta tabla, estos dependen de asistir a una escuela selectiva. En particular, podemos observar que si alguien entra en una escuela selectiva, automáticamente le estamos incluyendo 250 a sus ingresos. No es lo que encontraron Dale y Krueger, pero sigamos esta simulación para ver a dónde nos lleva. El siguiente código nos muestra la diferencia de promedios de la ecuación (2).

```
| df[df['selectivas'] == 1]['ingresos'].mean() - df[df['selectivas'] == 0]['ingresos'].mean()
```

676.1391841200148

Los datos nos dicen que hay una diferencia de 676 (digamos que son miles de) dólares. Pero esto es diferente que los 250 que nosotros establecimos como la diferencia en la ecuación 3. ¿Por qué nos aparece un efecto distinto?

Sesgo de Selección

Si recuerdas la discusión sobre el artículo de Dale y Krueger, los estudiantes que entran a una escuela selectiva son **fundamentalmente distintos** a los que no.

En particular vimos que los estudiantes que tienen padres con dinero y conexiones resultan tener mayores ingresos que el resto, independientemente del tipo de universidad a la que asisten. En el código aglomeramos todas las características que hacen a alguien más propenso a entrar en una escuela selectiva en la variable de "habilidad". Si la persona imaginaria que generamos en la simulación

Cuadro 2: Promedios simulados según tipo de escuela

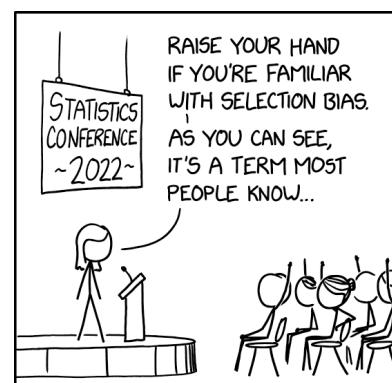


Figura 7: En una conferencia de estadística: "Levanta la mano si estás familiarizado con el sesgo de selección... como pueden ver, es un término que la mayoría de las personas conoce...". Fuente: [xkcd](#)

tiene una habilidad mayor o igual a 0.5, entonces entrará a una escuela selectiva.

Es una forma de emular las circunstancias reales en las que un grupo de personas tiene características particulares que lo hacen más propenso a entrar en el **grupo de tratamiento**. Como puedes notar la habilidad no es una característica **observable**, pero definitivamente hace que la selección no sea aleatoria.

En un experimento, nos interesa que la selección del grupo de tratamiento sea aleatorio. Cuando no se dan las condiciones para que la selección sea aleatoria como en un experimento, se dice que tenemos **sesgo de selección**.

En otras palabras, si vas a una zona rica en Nueva York tendrás más probabilidad de encontrar a un futuro estudiante de Harvard que si vas a una zona pobre en Puerto Rico (el territorio con tasa de pobreza más alta en EE.UU.).

El beneficio de entrar a una escuela más selectiva

La comparación no se debe hacer entre los estudiantes que entraron a escuelas selectivas y los que no. La comparación debe ser a **las mismas personas en universos paralelos**.

Imaginemos que Alicia logró entrar a Harvard, pero Bernardo no. Alicia sabe tres idiomas, tuvo tutorías personalizadas durante la preparatoria y pasaba las tardes en clases extracurriculares. Bernardo trabaja por las tardes para apoyar a su familia, tiene buenas calificaciones, pero no ha tomado tutorías extra. Alicia tiene computadora en casa y buen acceso a internet, Bernardo tiene una computadora descompuesta y no hay internet en su casa.

No podemos comparar a Alicia con Bernardo y pensar que las diferencias en sus ingresos vienen de su inscripción a Harvard.

Lo ideal sería comparar a Alicia en el universo 1, donde si entró a Harvard, con Alicia del universo 0, donde no entró a Harvard. Digamos que Y_i son los ingresos de Alicia. Podemos incluir un subíndice más para indicar el universo.

Y_{1i} indica los ingresos de Alicia en el universo 1 y Y_{0i} en el universo 0.

Finalmente, D_i es una **dummy** que indica si Alicia entró a Harvard ($D_i = 1$) o no ($D_i = 0$).

El ingreso de Alicia entonces es

$$Y_i = \begin{cases} Y_{1i} & \text{si } D_i = 1 \\ Y_{0i} & \text{si } D_i = 0 \end{cases}$$

El ingreso de Alicia en el universo cero (Y_{01}) no lo podemos observar, pues está en otro universo donde ella no entró a Harvard³⁴. Aún así podemos denotar el efecto de haber entrado a esa universidad en sus ingresos como $Y_{1i} - Y_{0i}$, por lo tanto el ingreso de Alicia en función de D_i sería

$$Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$$

Escribir este efecto aún cuando no se puede observar es un ejercicio que nos ayuda a entender el sesgo que se genera al comparar grupos diferentes.

Cuando $D_i = 1$, los ingresos de Alicia son Y_{1i} y cuando $D_i = 0$ sus ingresos son Y_{0i} . Para observar esto no se necesita imaginación. Pero el ingreso promedio de aquellas personas que entraron a una escuela selectiva *si no hubiéran entrado en una*, $E[Y_{0i}|D_i = 1]$ es un **contrafactual**. Es algo que no pasó, pero que podría pasar. Son los ingresos de todas las personas en situación similar a la de Alicia ($D_i = 1$), **si no hubiéran entrado a Harvard**.

³⁴ Que no sea observable no significa que no lo podamos incluir en el modelo. No podemos entrar a un universo paralelo, pero con algunos supuestos y estadística, nos sirve mucho pensar en lo que *pasaría* en escenarios que no existen. A estos se les conocen como **contrafactual**.

Descubriendo el sesgo de selección

La diferencia de medias viene con un sesgo de selección escondido. Para descubrirlo hace falta manipular un poco las ecuaciones.

Nota que

$$E[Y_1] - E[Y_0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \quad (4)$$

La ecuación de arriba nos muestra la diferencia entre los ingresos observados. Es simplemente lo que pasó en la realidad.

Pero lo que realmente necesitamos comparar es

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \quad (5)$$

La ecuación 5 representa el **tratamiento promedio en la unidad tratada**. En inglés se expresa como *Average treatment on the treated* y lo verás con las siglas **ATT**.

La diferencia entre la ecuación 5 y la ecuación 4 es sutil. Nota que $E[Y_{0i}|D_i = 1]$ es un contrafactual³⁵. Son los ingresos que tendrían las personas que entraron a Harvard si no hubieran entrado.

Nota que podemos agregar este contrafactual a la ecuación 4 como suma y como resta para descubrir el *ATT* de la ecuación 5, acompañado de un elemento adicional, al que llamamos sesgo de selección.

³⁵ El subíndice de Y_{0i} es diferente a lo que pasó en realidad, expresado por el $D_i = 1$.

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{ATT}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Sesgo de selección}} \quad (6)$$

Es un truco sucio, ¿lo notaste? Solo incluí el contrafactual como suma y como resta otra vez para no alterar la ecuación. Hacer esto revela que la ecuación inicial incluye el sesgo de selección.

Para deshacernos del sesgo de selección necesitamos que las condiciones de nuestro estudio se asemejen lo más posible a lo que veríamos en un experimento. En un experimento, la selección es aleatoria, lo que hace que $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$, eliminando el sesgo de selección.

¿Por qué funciona? Imagina que estás haciendo un estudio para identificar si existe discriminación en el proceso de contratación de las empresas en Estados Unidos. Hay muchas formas de tratar de hacer ese tipo de estudios, pero como se trata de un tema delicado, hacer una encuesta sería un esfuerzo inútil: si una empresa discrimina, no estaría dispuesta a decírselo a un extraño en una encuesta, por mucho que le asegures que será anónima.

Para solucionar ese problema, Bertrand y Mullainathan³⁶ idearon un *experimento* en el que mandaron currículums falsos a empresas que anunciaban puestos de trabajo. Los currículums eran idénticos entre sí, exceptuando una diferencia clave: algunos tenían nombres que sonaban más “blancos” y otros tenían nombres más “afroamericanos”. Encontraron que las empresas respondían un 50 % más a los postulantes con nombres más blancos que a los nombres afroamericanos.

Este es un tipo de estudio que se conoce como estudios de *auditología*. Aquí podríamos decir que el *tratamiento* D_i es el color de piel y el valor de resultado Y_i es la respuesta binaria de si respondieron o no. Nota que al mandar los currículums de manera aleatoria, se cumple

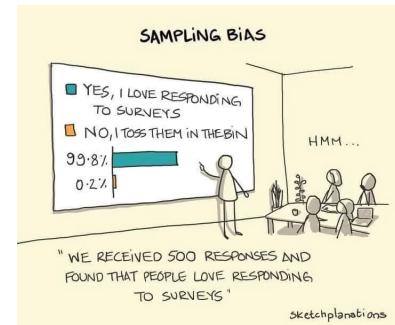


Figura 8: “Hemos recibido 500 respuestas y encontramos que a las personas les encanta responder a las encuestas”.
Fuente: Jono Hey, Sketchplanations

³⁶ Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *NBER*, 2003. DOI: 10.3386/w9873

que $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$. El valor promedio de las respuestas que reciben aquellos que no tienen un color particular de piel es independiente de su color de piel.

Hemos eliminado exitosamente el sesgo de selección.

El experimento aleatorio ideal

Hay una manera de eliminar el sesgo de selección.

Los ensayos controlados aleatorizados (RCTs) son experimentos donde nos aseguramos de eliminar todas las variables que podrían afectar nuestro resultado. Los puedes encontrar en inglés como RCTs: *Randomized controlled trials*. Son el *gold standard* de los estudios científicos.

Los resultados de un **ensayo controlado aleatorizado (RCT)** son considerados causales.

Un ejemplo de experimento en economía es el que hicieron Miguel & Kremer (2003)³⁷ para encontrar los efectos de tratar contra las lombrices a alumnos en las escuelas en Kenia. Al eliminar los parásitos no sólo mejoraron la salud de los alumnos, también aumentaron la participación en clase de las escuelas **y de escuelas cercanas**.

Para hacer un experimento necesitas:

- **Controlar las condiciones del estudio.** En la medida de lo posible, elige hacer tu estudio en grupos que sean comparables y homogéneos.
- **Divide en dos grupos: tratamiento y control.** El grupo de tratamiento es donde esperamos ver el efecto, el **grupo de control** sirve para comparar los resultados.
- **Asigna los grupos aleatoriamente.** Todos los participantes en el estudio deben tener la misma probabilidad de pertenecer a cualquier grupo.
- Sigue y analiza tus resultados. Verifica que todo salió de acuerdo al plan y analiza los datos. Veremos más sobre esto a lo largo del libro.

Cuando un experimento está bien diseñado, no se necesitan modelos estadísticos demasiado complejos.

Desafortunadamente, los experimentos no siempre son posibles de ejecutar. Requieren de planeación, tiempo y recursos que no siempre

³⁷ Edward Miguel and Michael Kremer. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72 (1):159–217, 2003. DOI: 10.1111/j.1468-0262.2004.00481.x

escalan.

Aún así, diseñar el **experimento ideal** para identificar el efecto que buscamos es un paso esencial en el diseño de nuestra estrategia de identificación.

Experimento ideal y experimentos naturales

El experimento ideal requiere que imaginemos cómo seguiríamos los pasos para hacer un experimento si no tuviéramos ninguna limitación de recursos, tiempo o incluso ética.

Por ejemplo, el estudio de Dale & Krueger (2002)³⁸ requeriría que tomáramos a un grupo de estudiantes y mandáramos de manera aleatoria a la mitad a Harvard y registrar los ingresos de los alumnos en ambos grupos al salir.

Pensar en el experimento ideal soluciona la mitad del problema.

Cuando entiendes cuál es el experimento ideal, es más fácil entender la estrategia que debes usar con las limitantes de la vida real. Dale y Krueger no usaron un experimento para encontrar los efectos de entrar a las escuelas selectivas. Lo que hicieron fue una estrategia ingeniosa para que los datos funcionaran como si hubieran hecho un experimento.

Usaron un experimento natural.

Los experimentos naturales son situaciones que encontramos que generan condiciones para analizar un fenómeno. Por ejemplo, Dale & Krueger (2002) aprovecharon a los alumnos que fueron aceptados en escuelas selectivas, pero decidieron no ingresar por motivos externos. Este es un grupo que sí se puede comparar con quienes entraron a las universidades selectivas.

No encontraron diferencias significativas por entrar en una universidad selectiva.

Hay una excepción: los alumnos con menos ventajas en su entorno familiar sí se beneficiaron de asistir a una escuela selectiva. Esto es porque la universidad es una forma de hacer conexiones personales que pueden traer ventajas para toda la vida. Los alumnos con mayor ventaja en su contexto familiar ya tienen acceso a estas conexiones, pero para los menos aventajados, la universidad es una oportunidad significativa.

³⁸ Dale S. Berg and Alan B. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 2002

Diseña tu experimento ideal con Inteligencia Artificial

Diseñar el experimento ideal es el primer paso para encontrar la estrategia de identificación.

Para diseñar el experimento natural necesitamos:

- Encontrar la relación causal que nos interesa.
- Reducir las variables a indicadores para medirlas.

El resto lo podemos hacer con ayuda de Inteligencia Artificial.

Usaremos chatGPT para este ejercicio.

Algunas advertencias sobre chatGPT:

- Los modelos grandes de lenguaje (LLM) como chatGPT funcionan como una predicción de la siguiente frase que viene en el texto. Es como el predictor de texto de tu teléfono, pero más avanzado.
- El resultado que obtengas depende del modelo que estés usando (yo muestro el resultado de chatGPT 4, que es el mejor en el mercado en el momento en que escribo esto).
- La mejor manera de obtener buenos resultados con chatGPT es interactuando con lo que te arroja. Pídele modificaciones o agrega información para que corrija, no te quedes con lo primero que te arroja.

Usa el siguiente comando. Puedes copiar y pegar directamente en chatGPT.

Eres un experto en econometría e inferencia causal.

Diseña el experimento ideal para identificar el efecto sobre las variables que te diré a continuación.

Un experimento ideal es una descripción detallada de un experimento que se podría hacer para obtener efectos causales, sin reparos en los recursos, tiempo o dilemas éticos que pueda causar.

Te voy a describir las variables que deseo estudiar y su relación causal que busco identificar.

¿Estás listo?

A continuación tienes que indicarle las variables que deseas medir. Por ejemplo, yo escribí esto para pedir indicaciones sobre el experimento ideal de las universidades selectivas.

Deseo conocer el efecto de asistir a una universidad selectiva en los ingresos

Lee con atención el resultado. Si consideras que lo que describe es algo viable, deberías intentarlo.

Con más experiencia, es buena idea que intentes diseñar estos experimentos ideales por tu cuenta. No sólo es un gran ejercicio mental, probablemente te dará la respuesta sobre la estrategia que debes usar en tu proyecto.

Resumen del capítulo

Para hacer inferencia causal, primero tuvimos que abrir la puerta al multiverso.

Lo que hicimos en este capítulo fue establecer el lenguaje para hablar sobre la causalidad. La idea clave es que para medir el verdadero efecto de algo (el "tratamiento"), necesitamos comparar el resultado que observamos en nuestro universo con el resultado que *hubiera ocurrido* en un universo paralelo donde no se recibió el tratamiento. A esto le llamamos el **contrafactual**.

Esto es importante porque nos revela el error más común y peligroso en el análisis de datos: el **sesgo de selección**. Comparar a la gente que fue a Harvard con la que no fue, no nos dice el efecto de ir a Harvard. Nos dice que los dos grupos ya eran diferentes *desde el principio*. La diferencia que vemos en sus sueldos es una mezcla del efecto real de la universidad y de esas diferencias preexistentes.

¿Cómo te ayuda esto? El modelo de **resultados potenciales** te da una estructura mental para desarmar cualquier afirmación causal que escuches. De ahora en adelante, cuando veas una comparación entre dos grupos, tu cerebro inmediatamente preguntará: "Ok, pero, ¿estos grupos eran iguales antes de que todo pasara?". Te obliga a pensar en el **experimento ideal**, el experimento perfecto que eliminaría el sesgo de selección. Y pensar en ese experimento imposible es, irónicamente, el primer paso para encontrar una solución inteligente y posible en el mundo real.

Aterrizando el multiverso: ejercicios conceptuales

Estos no son ejercicios de código, son para que los pienses, los discutas o los escribas. El objetivo es que la "música" del álgebra de los resultados potenciales empiece a sonar en tu cabeza.

1. **Definiendo el contrafactual:** Una empresa implementa un nuevo programa de bienestar (clases de yoga gratis) para sus empleados. Quieren saber si el programa reduce el estrés. Para una empleada llamada Laura, que **sí** participó en el programa:

- Define en palabras qué serían Y_{1i} , Y_{0i} y D_i .
- ¿Cuál de los dos resultados potenciales (o "universos") podemos observar para Laura?

- ¿Cómo expresarías el efecto causal individual del programa para ella?
2. **Detectando el sesgo de selección:** Un blog de tecnología publica un artículo que dice: "Los usuarios que activan la autenticación de dos factores (2FA) en sus cuentas sufren un 80% menos de hackeos". ¿Por qué una simple comparación entre los que usan 2FA y los que no, probablemente exagera el efecto real del 2FA? Describe qué tipo de persona es más propensa a activar el 2FA y cómo eso genera un sesgo de selección.
 3. **El ATT en palabras:** En el ejemplo de Harvard, vimos que el Tratamiento Promedio en los Tratados (ATT) es $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$. Explica con tus propias palabras qué significa el término contrafactual $E[Y_{0i}|D_i = 1]$. ¿Por qué es fundamentalmente diferente de $E[Y_{0i}|D_i = 0]$ (los ingresos de los que no fueron a Harvard)?
 4. **Calculando el sesgo:** Imagina los siguientes datos sobre un programa de tutorías para mejorar las calificaciones:
 - Calificación promedio de alumnos **con** tutoría: 9.1
 - Calificación promedio de alumnos **sin** tutoría: 7.5
 - Calificación promedio que los alumnos **con** tutoría *hubieran obtenido si no la hubieran tomado* (el contrafactual): 8.5
 Calcula: a) La diferencia de medias simple, b) El verdadero ATT del programa de tutorías, y c) El sesgo de selección. ¿Qué nos dice el signo del sesgo de selección sobre los alumnos que eligen tomar tutorías?
 5. **El poder de la aleatoriedad:** En un [ensayo controlado aleatorizado \(RCT\)](#) bien diseñado, el sesgo de selección es cero. Usando los datos del ejercicio anterior, si el experimento se hubiera asignado al azar, ¿qué valor esperaríamos que tuviera el contrafactual calificación promedio que los alumnos **con** tutoría hubieran obtenido si no la hubieran tomado? ¿Qué implica esto sobre los dos grupos antes de empezar las tutorías?
 6. **Diseña tu experimento ideal (Fácil):** Quieres saber el efecto causal de poner música clásica en una cafetería sobre el gasto promedio por cliente. Describe paso a paso el [experimento ideal](#) (un RCT) que diseñarías. Define la población, el tratamiento, el grupo de control, cómo harías la asignación aleatoria y qué variable de resultado medirías.
 7. **Diseña tu experimento ideal (Difícil):** Quieres saber el efecto causal de haber crecido con un perro en la infancia sobre los niveles de empatía de una persona en la edad adulta. ¿Por qué es imposible (y antiético) hacer un RCT para esta pregunta?
 8. **Pensando en experimentos naturales:** Para la pregunta del ejercicio anterior (perros y empatía), ¿se te ocurre alguna situación del mundo real que pueda funcionar como un [experimento natural](#)? (Pista: piensa en situaciones que "asignan" un perro a una familia de forma casi aleatoria, por razones ajenas a sus características personales).
 9. **Interpretando la fórmula:** La fórmula de descomposición es: *Diferencia de Medias = ATT + Sesgo de Selección*. Describe un escenario del mundo real donde el ATT sea prácticamente cero, pero observemos una gran diferencia de medias positiva debido a un fuerte sesgo de selección. (Pista: piensa en productos o servicios exclusivos).

10. ¡A usar la IA!: Toma una pregunta causal que te interese personalmente (ej: ¿dormir 8 horas diarias mejora el humor?, ¿leer ficción aumenta la creatividad?, ¿usar bicicleta para ir al trabajo reduce el estrés?). Usa el *prompt* exacto que te di al final del capítulo en ChatGPT u otra IA. Pega la respuesta que te dé y luego escribe un párrafo criticándola: ¿es realmente un “experimento ideal”? ¿Qué limitaciones prácticas u éticas ignoró?

Experimentos y Pruebas A/B

Outsized returns often come from betting against conventional wisdom, and conventional wisdom is usually right.

Given a 10 percent chance of a 100 times payoff, you should take that bet every time. But you're still going to be wrong nine times out of ten...

We all know that if you swing for the fences, you're going to strike out a lot, but you're also going to hit some home runs.

The difference between baseball and business, however, is that baseball has a truncated outcome distribution. When you swing, no matter how well you connect with the ball, the most runs you can get is four. In business, every once in a while, when you step up to the plate, you can score 1,000 runs.

This long-tailed distribution of returns is why it's important to be bold.

Big winners pay for so many experiments.

- Jeff Bezos.

If your experiment needs statistics, you ought to have done a better experiment

- Ernest Rutherford

EN MAYO DE 2024, mi esposa decidió que abriríamos una **sucursal de la tienda** en Mazatlán³⁹.

La ciudad nos queda a 3 horas y media y pensamos que sería una buena oportunidad para expandir el negocio y aprender a operarlo. En una semana nos instalamos en un local, pusimos cámaras y un sistema de inventarios y echamos a andar la aventura.

Ahora sólo queda encontrar clientes que quieran ir a la tienda.

Una de las estrategias que intentamos fue hacer publicidad usando Instagram. El problema: ninguno de los dos tenemos experiencia en *pautar*⁴⁰.

Así que hice lo que si sabía hacer: un experimento.

Dividí el anuncio en tres elementos e hice tres versiones diferentes de cada uno. En total fueron $2^3 = 8$ diferentes anuncios. Cada anuncio tenía una diferente combinación de llamado de atención, de oferta de valor y de llamado a la acción.

³⁹ Un poco de historia. Mi esposa llegó de Rusia en 2016 conmigo, donde nos conocimos. Ella estudió antropología social en San Petersburgo, pero se decidió a convertirse en *lashista*. Comenzó a poner pestanas en nuestro departamento en Durango y de ahí el negocio creció. Hoy es una empresa donde da cursos y provee material para *lashistas*. Conoce más en mariapestanuras.com

⁴⁰ es la palabra *fancy* que usan para decir "poner publicidad"



Lanzamos los anuncios, con un miedo inmenso de que tal vez estamos echando dinero a la hoguera⁴¹.

Cuando se acabó la campaña, encontramos una diferencia enorme entre anuncios. El anuncio más efectivo costaba **40 pesos (MXN) por lead**⁴², a diferencia del menos efectivo que costaba más de 200. El objetivo era encontrar el mensaje correcto⁴³.

Y lo logramos. Por una fracción del presupuesto que habríamos puesto en contratar a un “experto” en mercadotecnia.

Los experimentos son el gold standard experimental de la ciencia

Hoy en día, las plataformas para hacer publicidad en línea tienen incluida la creación de pruebas A/B. Una **prueba A/B** es un experimento en el que se ponen dos versiones diferentes de un anuncio y la plataforma optimiza para encontrar rápido y en automático cuál es la versión más efectiva. Hay otros servicios que hacen algo similar: stripe te permite hacer pruebas A/B en los formularios de pago para que hagas pruebas con detalles del diseño o el *copy* y puedas elegir siempre la más efectiva.

En realidad la oportunidad de hacer experimentos en los negocios es ilimitada.

La mayor barrera de los experimentos en los negocios es el tiempo. El punto de un experimento es aislar todas las variables que podrían afectar los resultados. Si queremos conocer el efecto que tiene una campaña publicitaria en las ventas, tenemos que poner en contexto

⁴¹ La hoguera en este caso se conoce como Meta.

⁴² Prospecto.

⁴³ Hacer esto transforma el problema de encontrar un anuncio efectivo a encontrar aquellos anuncios con costo de adquisición del cliente menor a su *lifetime value*.

la época del año en que se hizo, el medio en el que se ejecuta, la segmentación que se aplicó y hasta si estaba mercurio retrógrado⁴⁴.

Simplemente no hay tiempo de poner a prueba todos y cada uno de los detalles⁴⁵.

¿Por qué se usan experimentos para hacer inferencia causal en los negocios?

Imagina esta situación: hay una discusión en tu compañía sobre el encabezado en tu *landing page*.

Acabas de contratar a un economista en el área de marketing que viene con nuevas ideas y quiere cambiar el encabezado que han tenido en los últimos 5 años. El nuevo encabezado estaría más centrado en el cliente y no es una descripción de la empresa. De acuerdo a él, esto aumentaría la tasa de conversión de la página, pero su jefe no está de acuerdo y quiere dejar el copy actual.

La solución del 99 % de las empresas: organizar reuniones para debatir y encontrar la respuesta.

El problema con este enfoque es que hacer una reunión puede llegar a ser muy costoso y existe el riesgo a que no se llegue a ningún acuerdo. Si llamas a una reunión a 10 empleados y cada uno de ellos está ganando **50 dólares la hora, entonces se trata de una reunión de 500 dólares**.

Lo peor es que lo más probable es que ninguno de los asistentes a la reunión pueda dar con los argumentos suficientes para decidir el uso de algún copy. En el mejor de los casos, se tomará una decisión que no está basada en datos. En el peor de los casos, la decisión será más un reflejo de las dinámicas de poder en la empresa que de la efectividad del *copy*.

¿De qué forma sí podrías encontrar la respuesta correcta, libre de controversias? usando un experimento

Éstas son las razones para usar experimentos en los negocios:

- **Razón #1: los experimentos quitan cualquier fuente de duda razonable en la causalidad entre variables.** Cuando se hace un experimento, puedes afirmar con seguridad que la relación que encuentras es causal. Si haces un análisis de la información que no viene de experimentos, siempre habrá quien niegue tus resultados diciendo que “correlación no implica causalidad”⁴⁶.

⁴⁴ ¡Es broma!

⁴⁵ Como en muchos casos, la excepción a la regla es el caso más notorio: las *startup* de tecnología son famosas por hacer pruebas A/B de millones de pequeños detalles. Desde el tamaño de la letra, hasta 20 diferentes tonos de azul para probar cuál es el que genera más clicks. La razón es que ellos pueden automatizar y tienen volúmenes gigantescos. Para el resto de nosotros, simplemente no vale la pena.

⁴⁶ Y tendrían razón.

- **Razón #2: el resultado de los experimentos se hacen con matemáticas simples y con poca estadística.** Cuando haces un experimento bien, lo único que necesitas es una diferencia de medias y algunas pruebas estadísticas sencillas y fáciles de interpretar. El resultado es muy claro e intuitivo y no necesita matemáticas complejas.
- **Razón #3: los resultados son más entendibles e interpretables.** A diferencia de lo que muchos creen sobre la ciencia, los científicos siempre buscamos la claridad. Con pocas matemáticas y estadística, es más fácil comunicar los resultados de manera creíble.
- **Razón #4: puede ser mas barato de implementar que hacer reuniones para tomar una decisión.** Sobre todo cuando una empresa es pequeña, es más importante basar nuestras decisiones en evidencia. ¿De verdad es importante tener a todos los jefes de área para decidir el copy de tu página? Eso es algo que debes dejar que decidan tus clientes.

¿Qué pasa cuando no se puede hacer un experimento?

La regla es que siempre debemos comenzar con un experimento, o mejor dicho con un [experimento ideal](#).

La realidad es que en la mayoría de los casos, un experimento no es posible, no es ético, o está fuera de nuestro presupuesto. Nada de eso impide que *imaginemos* cuál sería el **experimento ideal** que nos ayudaría a identificar los efectos que buscamos. Angrist y Pischke (2005)⁴⁷ sugieren que hagamos ese ejercicio imaginándonos como investigadores sin restricciones de presupuesto ni comité de ética que tenga que revisar lo que hacemos.

Como si fueras un científico loco de una película de ficción..

La razón #1 para comenzar con un experimento ideal es que si no puedes imaginar una forma de identificar causas y efectos cuando no tienes ninguna restricción, entonces será difícil que logres identificar los efectos en situaciones normales con las restricciones naturales que da la vida. Por otro lado, el ejercicio de imaginar el experimento te ayuda a entender la relación causal de una forma más precisa. Con esto puedes identificar las variables que se involucran y cómo funcionan.

⁴⁷ J. D. Angrist and J. S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009

Aunque no lo creas, hay preguntas que no se pueden resolver ni aún con un experimento.

¿Cuál es el impacto de la experiencia laboral antes de fundar una empresa?

En un estudio con casi tres millones de emprendedores se encontró que la edad en la que los emprendedores tienen más probabilidad de tener éxito en los negocios es de **42 años**⁴⁸. Aparentemente, la experiencia laboral es un factor clave que da las condiciones ideales para hacer que los negocios tengan un mejor desempeño. ¿Podríamos hacer un experimento para comprobarlo?

El problema es que las personas con más experiencia laboral, también tienen más años de vida.

Con más edad viene más experiencia, pero también podría tratarse de una mayor madurez emocional o una red de contactos más grande. ¿Cómo controlamos estas variables en nuestro experimento? Podríamos imaginar que separamos a dos grupos de emprendedores de manera aleatoria y nos aseguramos que el grupo A emprenda en sus 20 y el grupo B emprenda llegando a los 40 y medimos la diferencia. Pero hay miles de características inseparables de la edad que hacen que los dos grupos sean fundamentalmente distintos y por lo tanto, no sean comparables.

Y por lo tanto, nuestro experimento no puede tener resultados que podamos interpretar correctamente.

Es importante saber distinguir cuándo una pregunta es fundamentalmente imposible de contestar. Aún dentro de un experimento hipotético sin límite de recursos. De esta manera no perdemos el tiempo tratando de hacer comparaciones sin sentido con datos de menor calidad que los que vienen de un experimento.

Esto no quiere decir que el estudio de Azoulay (2020) esté mal: simplemente que no podemos inferir que la relación entre estas variables sea causal.

Ejemplo: El efecto de las búsquedas pagadas en las ventas

El modelo de negocios de Google es uno de los más extraños e ingeniosos que te vas a topar.

Cada vez que haces una búsqueda en Google, se lanza una subasta tras bambalinas. Si alguna vez has intentado poner un anuncio en

⁴⁸ Pierre Azoulay, Benjamin F. Jones, J. Daniel Kim, and Javier Miranda. Age and high-growth entrepreneurship. *American Economic Review: Insights*, 2(1): 65–82, 2020. DOI: 10.1257/aeri.20180582

Google, te darás cuenta de que estás haciendo pujas por términos de búsqueda. Si tienes una tienda de productos para hacer escalada, te interesa aparecer en los primeros términos cuando alguien pone “zapatos para escalar” en el buscador.

Tienes dos formas de lograrlo: haciendo contenido relevante en tu página y esperar a que aparezca, o pagando a Google para que te ponga en los primeros lugares.

El contenido puede venir en forma de un blog. De hecho, antes de las redes sociales, esa era la única manera de hacer que el buscador te suba en los lugares de búsqueda de manera orgánica. Puedes escribir sobre cómo usar correctamente el calzado, sobre accesorios para escalar o sobre técnicas de escalada.

Si haces bien tu contenido, apareces en los primeros lugares aunque no estés pagando a Google.

Por eso el negocio de google es tan peculiar. Si eres la marca más relevante de una categoría, pagar porque te ponga en los primeros lugares no parece tener mucho sentido, pues ya debes de aparecer al inicio. Es difícil saber cuando las búsquedas pagadas realmente están surtiendo efecto o son sólo un gasto innecesario.

Se necesita hacer un experimento para averiguarlo.

Blake, et al. (2015)⁴⁹ hicieron una serie de experimentos con diferentes productos publicados en la plataforma eBay. El experimento consistía en “apagar” las búsquedas pagadas para diferentes productos elegidos al azar y comparar las ventas en los grupos de tratamiento y de control. Lo que encontraron fue que los clicks que venían de la publicidad pagada fueron sustituidos casi en su totalidad por clicks que venían de los resultados de búsqueda.

Las búsquedas pagadas no tenían ningún efecto.

Pero como siempre, los detalles son importantes. Los autores se acercaron a eBay con sus resultados para hacer un experimento a mayor escala con casi el 30 % de los productos de la tienda. El estudio lo hicieron con diferentes mercados e identificando a diferentes tipos de usuarios. Encontraron que sí hay un efecto positivo para los clientes que tienen poco tiempo como usuarios en la tienda y realmente necesitan ayuda para encontrar lo que necesitan. También las marcas más pequeñas dentro de la tienda mejoraron sus ventas gracias a la publicidad, a pesar de que las marcas grandes recibían clicks de búsquedas pagadas y orgánicas por igual.

⁴⁹ T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1):155–174, 2015. URL <http://www.jstor.org/stable/43616924>

Prompt: Diseña el experimento ideal

El primer paso para encontrar la estrategia correcta para nuestro estudio es diseñar el experimento ideal.

Un experimento ideal ignora las limitaciones del mundo real. En un experimento ideal no importa el presupuesto ni las leyes de la física. Podemos usar a las personas como los biólogos usan las bacterias en sus cajas de petri, aislando las condiciones para hacer comparaciones completas y apropiadas.

Por ejemplo, para estudiar el efecto de la educación en los ingresos, podríamos imaginar un experimento en el que compramos dos islas y las poblamos con personas con características comparables (misma cultura, mismo idioma, etc.) y a un grupo les proporcionamos más educación que al otro y medimos la diferencia.

Te dejo este prompt que te permitirá pedirle a la **inteligencia artificial (IA)** que te de ideas sobre cómo hacer un experimento que te interesa. Copia y pega esto en la ventana del chat de la inteligencia artificial. Después de su primera respuesta dile tu idea de lo que quieras estudiar⁵⁰:

Te voy a enseñar a hacer un experimento ideal para inferencia causal.
 Un experimento ideal es un diseño en el que se estudian los efectos causales sin limitaciones de presupuesto, de tiempo o de ética. Por ejemplo, para averiguar los efectos que tiene la educación en los ingresos de las personas podríamos ofrecer un incentivo económico a un **grupo de tratamiento** para que terminen sus estudios y revisar sus resultados en el tiempo.
 Tampoco hay limitaciones en las leyes de la física. Si nos interesa conocer el efecto que tienen las instituciones en el desarrollo económico de una nación, podríamos volver en el tiempo y asignar diferentes estructuras gubernamentales a diferentes estados de México para hacer un análisis *ceteris paribus* de sus resultados económicos. No te limites a los ejemplos que estoy dando. Puedes ser creativo. Por ejemplo, para estudiar el efecto de la educación en los ingresos, podríamos imaginar un experimento en el que compramos dos islas y las poblamos con personas

⁵⁰ Dos detalles de los prompts que uso. Primero, son prompts largos y llenos de detalle. En el momento en el que escribo esto, los modelos de lenguaje grandes como chatGPT requieren de contexto para mejorar sus resultados. Con el tiempo han mejorado. Mi pronóstico es que en el futuro, estos modelos usarán las conversaciones previas que has tenido con el modelo para refinar los resultados y adaptarse a tu lenguaje, lo que dará la ilusión de tener mejores resultados sin necesidad de contexto. El segundo punto es el uso de ejemplos, que dan más claridad sobre el tipo de resultado que quieres obtener. Estas dos son las variables que puedes modificar en el *prompt* para mejorarlo.

con características comparables (misma cultura, mismo idioma, etc.) y a un grupo les proporcionamos más educación que al otro y medimos la diferencia. Se trata de experimentos hipotéticos que nos permiten identificar la estrategia de identificación causal. Te daré un tema que me interesa estudiar y tu me darás una idea de experimentos ideales. ¿Estás listo?

Pasa la información de lo que quieras averiguar. Trata de darle tantos detalles como puedas para que tenga bien el contexto de lo que deseas. Aquí te dejo como ejemplo el texto de arriba sobre Google. El resultado de darle esto la mayoría de las veces es un experimento muy bien diseñado que se asemeja bastante a lo que hicieron Blake, et al. (2015).

El modelo de negocios de Google es uno de los más extraños e ingeniosos que te vas a topar.

Cada vez que haces una búsqueda en Google, se lanza una subasta tras bambalinas. Si alguna vez has intentado poner un anuncio en Google, te darás cuenta de que estás haciendo pujas por términos de búsqueda. Si tienes una tienda de productos para hacer escalada, te interesa aparecer en los primeros términos cuando alguien pone "calzado de escalada" en el buscador.

Tienes dos formas de lograrlo: haciendo contenido relevante y pagando a Google para que te ponga en los primeros lugares.

El contenido puede venir en forma de un blog. De hecho antes de las redes sociales esa era la única manera de hacer que el buscador te suba en los lugares de búsqueda de manera orgánica. Puedes escribir sobre cómo usar correctamente el calzado, sobre accesorios para escalar o sobre técnicas de escalada.

Si haces bien tu contenido, apareces en los primeros lugares aunque no estés pagando a Google.

Por eso el negocio de google es tan peculiar. Si eres la marca más relevante de una categoría, pagar porque te ponga en los primeros lugares no parece tener mucho sentido, pues ya debes de aparecer al inicio. Es difícil saber las búsquedas pagadas realmente están surtiendo efecto o son sólo un gasto innecesario.

Se necesita hacer un experimento para averiguarlo.

Al final, van a pasar una de dos cosas:

1. Vas a encontrar una forma de implementar el experimento.
2. O vas a diseñar tus modelos para que los resultados se asemejen lo más posible a los experimentos.

En cualquiera de esos casos, el diseño de tu estudio será mucho mejor de lo que habría sido si únicamente hubieras salido al mundo

a diseñar encuestas sin pensar en el diseño de tu estudio. Hacer esto mejorará la calidad de los datos que obtienes.

Y cuando se trata de datos, la calidad es órdenes de magnitud más importante que la cantidad.

Resumen del capítulo

En este capítulo vimos que los experimentos no son solo para científicos en laboratorios. Son una herramienta brutalmente efectiva para tomar mejores decisiones en los negocios. Son el **gold standard experimental** para encontrar la verdad.

Lo que hicimos fue bajar la idea del “experimento” de su pedestal académico y ponerla a trabajar en el mundo real, desde una campaña de anuncios en Instagram hasta el debate sobre el encabezado de tu página web. Un experimento, o su primo cercano la **prueba A/B**, es simplemente una forma estructurada de preguntarles a tus clientes qué es lo que prefieren, en lugar de discutirlo en una junta.

Esto es importante porque un experimento bien diseñado es la única forma de callar para siempre el molesto argumento de “correlación no implica causalidad”. Cuando aleatorizas, eliminás el **sesgo de selección** y aíslas el efecto verdadero. El resultado deja de ser una opinión y se convierte en evidencia. Como dice Jeff Bezos, los grandes ganadores pagan por muchísimos experimentos.

¿Cómo te ayuda esto? De ahora en adelante, tu respuesta por defecto ante una discusión de negocio debería ser: “¿Podemos probarlo?”. Este capítulo te da el marco para convertir debates costosos en experimentos baratos. Te enseña a pensar en el **experimento ideal** no como un ejercicio teórico, sino como el primer paso práctico para diseñar un estudio que te dé respuestas claras. Te da el poder de retar la “sabiduría convencional” con datos, no con opiniones.

Afilando la navaja experimental: ejercicios prácticos

Es hora de pensar como un experimentador. Estos ejercicios te ayudarán a aplicar los conceptos del capítulo a situaciones del mundo real.

1. **De la junta al experimento:** En el capítulo se habla del debate sobre cambiar el encabezado de una *landing page*. Diseña una prueba A/B simple para resolver esta discusión. Define: a) El grupo de control (A), b) El grupo de tratamiento (B), c) La métrica clave que usarías para decidir un ganador (la variable de resultado), y d) ¿Por qué es crucial mostrar las dos versiones al mismo tiempo a usuarios aleatorios?

2. **Interpretando resultados:** Imagina que corriste el experimento anterior y obtuviste estos datos:

- Encabezado A (Control): Se mostró a 5,000 visitantes y 400 hicieron clic en el botón “Comprar”.
- Encabezado B (Tratamiento): Se mostró a 5,000 visitantes y 450 hicieron clic en el botón “Comprar”.

Calcula la tasa de conversión para cada versión y el “lift” (la mejora porcentual) que generó el encabezado B. ¿Parece un cambio que valga la pena implementar?

3. **Identificando un experimento fallido:** Una tienda de ropa quiere saber si poner música en vivo los fines de semana aumenta las ventas. El primer fin de semana del mes no ponen música y registran las ventas. El último fin de semana del mes, que es quincena y día de pago para la mayoría de la gente, contratan a un guitarrista y registran las ventas. Las ventas del último fin de semana son 50 % más altas. El dueño concluye que la música fue un éxito rotundo. ¿Por qué esta conclusión es, probablemente, incorrecta? ¿Qué supuesto fundamental de los experimentos se está violando?
4. **Una pregunta in-experimentable:** Eres el gerente de producto de una aplicación de citas. Quieres saber el efecto causal de que una persona ponga “busco algo serio” en su perfil sobre la probabilidad de que consiga una pareja estable en un año. ¿Por qué sería casi imposible diseñar un ensayo controlado aleatorizado (RCT) limpio para responder esta pregunta? (Pista: Piensa en el sesgo de selección y en variables inseparables de la decisión).
5. **Diseñando el experimento ideal:** Una app de *delivery* de comida quiere saber si ofrecer envío gratis en el primer pedido realmente convierte a los usuarios nuevos en clientes recurrentes. Describe el experimento ideal para medir este efecto. ¿Cómo seleccionarías a los participantes y cómo los asignarías a los grupos de tratamiento y control? ¿Qué medirías y por cuánto tiempo?
6. **Buscando la letra pequeña (Efectos Heterogéneos):** El estudio de eBay sobre los anuncios pagados encontró que, aunque el efecto general era casi nulo, sí funcionaban para usuarios nuevos. Imagina que tu experimento de envío gratis (del ejercicio anterior) muestra un efecto general muy pequeño. ¿Qué subgrupos de usuarios investigarías por separado para ver si el envío gratis es muy efectivo para algún nicho en particular?
7. **¿Vale la pena experimentar?:** Una *startup* debate sobre el color de su botón de compra. El diseñador prefiere azul y el CEO prefiere verde. Una junta para discutirlo de 1 hora con 5 personas clave cuesta \$400 en salarios. Una herramienta de software para hacer la prueba A/B cuesta \$50. Explica por qué pagar los \$50 es casi siempre una mejor inversión, incluso si el resultado de la prueba es que ambos colores funcionan igual.
8. **La ética del A/B testing:** Una plataforma de videojuegos quiere saber si aumentar la dificultad de un nivel de forma inesperada causa que los jugadores pasen más tiempo en el juego (por frustración y repetición) o que lo abandonen. Describen un experimento donde a un grupo aleatorio de jugadores se les sube la dificultad sin avisar. ¿Por qué este experimento, aunque metodológicamente podría ser correcto, es éticamente cuestionable?
9. **Poniendo a prueba a la IA:** Usa el *prompt* para diseñar el experimento ideal que viene en este capítulo. Pídele a ChatGPT que diseñe el experimento para resolver el problema de la tienda de ropa con música en vivo (ejercicio 3). ¿La solución que te da la IA corrige los errores que identificaste en el diseño original del dueño de la tienda? Explica por qué sí o por qué no.

Una guía para entender y hacer modelos de Regresión Lineal

I know I'm stereotypical barbie and therefore don't form conjectures concerning causality of adjacent unfolding events. But some things have been happening that might be related

- Barbie (2023)

LA REGRESIÓN ES LA BASE DE LOS MODELOS DE INFERENCIA CAUSAL.

Hay una razón por la que todos los libros de econometría y de ciencia de datos lo cubren. Se trata del modelo por el que debes comenzar **antes de explorar** modelos más complejos.

No hay nada de malo en usar redes neuronales o modelos de random forest en tu proyecto⁵¹, pero si usas **regresión lineal**, tus modelos tendrán los siguientes beneficios.

- Mayor parsimonia. Entre más simple es el modelo, menos problemas te va a causar.
- Serán más fáciles de interpretar. Si necesitas comunicar tus resultados a un jefe o un cliente, necesitas poder decir claramente lo que tus datos significan y las limitaciones.
- Pruebas de robustez. Cuando un modelo pasa pruebas y demuestra que es robusto, podrás tener más confianza de usarlo en tus predicciones.

No es magia. Hay teoremas muy sólidos que ayudan a que entendamos lo que funciona y cuándo funciona.

Este capítulo se trata de sentar esas bases sólidas.

⁵¹ Actualmente está de moda hacer transición de economista a Data Scientist. Una de las razones principales es que alguien que estudia econometría ya lleva un buen avance en los modelos *supervisados* (todos los que vemos en este libro son de ese tipo). Los modelos *no supervisados* llaman mucho la atención por ser los que están detrás de la Inteligencia Artificial, pero siempre la regresión es el punto de inicio también en los libros de Data Science.

El modelo de mínimos cuadrados ordinarios con dos variables

Comencemos con el modelo básico. Tienes una variable X y deseas conocer el efecto que tiene sobre Y. La variable X podría ser el gasto en una campaña publicitaria por Televisión, mientras que Y son las ventas de nuestro producto.

Si tienes suficientes combinaciones de las dos variables, puedes plantear un modelo sobre su comportamiento. Usaremos la base de datos de publicidad, disponible libremente en [kaggle.com](https://www.kaggle.com/datasets/abhishek960/advertising). El siguiente código carga la base de datos directamente del repositorio y muestra un diagrama de dispersión entre los gastos en publicidad por TV y las ventas en millones de unidades.

```
import pandas as pd
import matplotlib.pyplot as plt

# Cargar los datos
data = pd.read_csv('advertising.csv')

# Crear un diagrama de dispersión
plt.figure(figsize=(10, 6))
plt.scatter(data['TV'], data['Sales'], alpha=0.5)
plt.title('Diagrama de Dispersión de Gastos en TV vs Ventas')
plt.xlabel('Gastos en TV ($)')
plt.ylabel('Ventas (Miles de unidades)')
plt.grid(True)
plt.show()
```

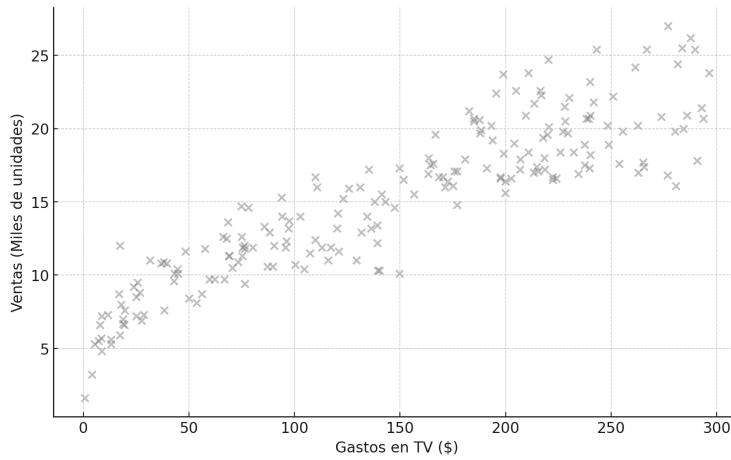


Figura 9: Diagrama de dispersión generado a partir del bloque de código de Python. Fuente: Elaboración propia.

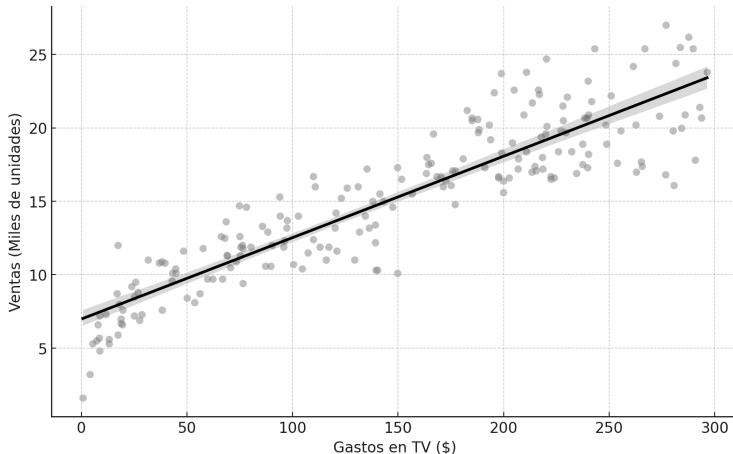
Este es un ejemplo muy claro donde la regresión lineal es el modelo ideal para nosotros: los puntos siguen un patrón muy claro visualmente.

Lo que nos dice la regresión lineal es que existe una línea que se ajusta a los datos. No necesitamos que el ajuste sea perfecto⁵². Es normal pensar que hay muchos factores que afectan las ventas además del gasto publicitario, desde el clima hasta el día del mes pueden generar variaciones. Todos reaccionamos diferente a la publicidad.

Así se ve nuestra línea de regresión.

```
# En esta ocasión usaremos seaborn porque nos ayudará a añadir la recta de ajuste
→ automáticamente
import seaborn as sns

# Crear un diagrama de dispersión con una línea de regresión
plt.figure(figsize=(10, 6))
sns.regplot(x='TV', y='Sales', data=data, scatter_kws={'alpha':0.5})
plt.title('Diagrama de Dispersión de Gastos en TV vs Ventas con Línea de Regresión')
plt.xlabel('Gastos en TV ($)')
plt.ylabel('Ventas (Miles de unidades)')
plt.grid(True)
```



⁵² Si tuviéramos un ajuste perfecto no necesitaríamos de modelos estadísticos.

Figura 10: Diagrama de dispersión con línea de regresión.

Este tipo de línea se genera con un modelo lineal, donde cada punto es producto de una función de tipo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

El punto i se ubica en la coordenada (X_i, Y_i) . El término ε_i es el **error**, la diferencia entre el punto y la línea. Los términos β_0 y β_1 (*se lee beta-cero y beta-uno*) son los parámetros de una función lineal. Usamos letras griegas por convención, y el subíndice cero y uno son una forma práctica de preparar nuestro modelo en caso de que tengamos que usar más parámetros.

El siguiente es un **diagrama de dispersión**.

Nota que en algunos puntos, la línea de regresión “se equivoca” hacia arriba y en otros puntos hacia abajo. Cada punto que compone la línea de regresión es una predicción del valor de Y_i dado X_i , donde ε_i es la diferencia, a la que llamamos el **residual**.

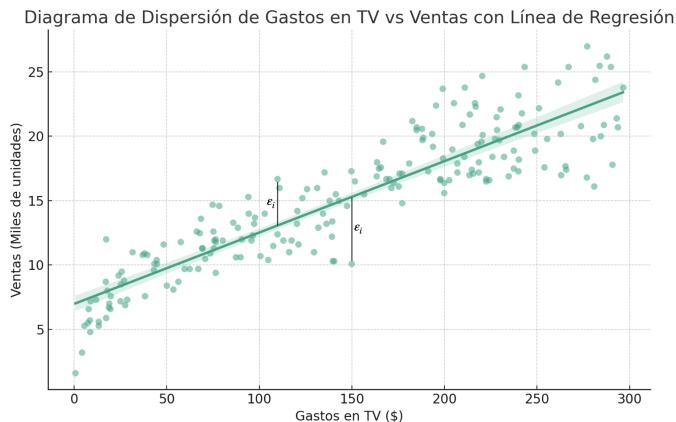


Figura 11: Le llamamos “error” a la diferencia entre una observación y la línea de regresión que “predice” dónde debería estar Y_i dado X_i . Fuente: Elaboración propia con Python.

El modelo lineal tiene la ventaja de que sólo con dos parámetros podemos definir toda la línea.

Si $\beta_0 = 6,97$ y $\beta_1 = 0,0554$, entonces un valor de $X_i = \$150$ en gasto de publicidad por TV implica ventas por 15,29. De hecho, puedes crear una calculadora sencilla en Python para que te muestre el valor de las ventas que corresponde a cualquier gasto en TV⁵³.

```
# Crear la función
def sales(tv):
    b0 = 6.97
    b1 = 0.0555
    return b0 + b1 * tv

# Comprobar el resultado con 150
sales(150)
```

⁵³ No es común ni necesario hacer este tipo de calculadoras. Si deseas obtener el valor de las predicciones en un modelo llamado `model`, basta llamar la función de predicción usando, por ejemplo, `model.predict()`.

15.294999999999998

Podríamos hacer una predicción de Y para cada punto X. Si tu regresión es correcta y la muestra es buena, puedes usar la función para valores de X que no están en tu base de datos. Por ejemplo, este modelo predice que un gasto de publicidad en TV de \$450 traerá ventas por 31.945 miles de unidades.

El método de Mínimos cuadrados ordinarios

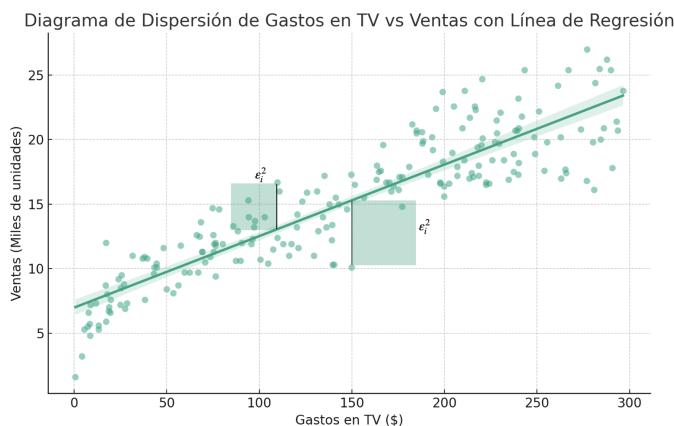
Ya conoces el modelo de regresión lineal, ahora te presento el mejor modelo para resolverlo⁵⁴.

El método de **mínimos cuadrados ordinarios (OLS)** es el método más popular para resolver el modelo de regresión lineal. Se prefiere porque es simple y muy eficiente.

Bajo ciertas condiciones, OLS se considera el mejor estimador lineal insesgado. El acrónimo en inglés es BLUE (Best Linear Unbiased Estimator):

- Best (Mejor): Significa que tiene la menor varianza de las estimaciones.
- Linear (Lineal): El estimador es una función lineal de los valores observados.
- Unbiased (Insesgados): El estimador le *atina* al verdadero valor del parámetro **en promedio**.
- Estimator (Estimador): Es la regla o fórmula que indica cómo estimar los parámetros del modelo.

OLS es una de muchas técnicas que se pueden utilizar para resolver el modelo. Tiene el objetivo de encontrar los valores de β_0 y β_1 que minimizan la suma de los errores al cuadrado. La siguiente imagen muestra cómo se extiende el área de los errores al cuadrado.



La imagen solo muestra el cuadrado de dos puntos. Si pudiéramos mover con libertad los valores de β_0 y β_1 , podríamos ver cómo esos cuadros se hacen más grandes y más chicos, de acuerdo a la distancia con los puntos.

⁵⁴ Hay muchas formas de solucionar el modelo además de la que veremos aquí. El modelo OLS generalmente se selecciona porque tiene propiedades deseables, como lo veremos más adelante.

Figura 12: El objetivo del método de mínimos cuadrados es encontrar la línea que minimice la suma de los errores al cuadrado. Fuente: Elaboración propia con Python.

¿Cómo encontramos los valores de β_0 y β_1 que hacen mínima la suma de los residuales al cuadrado?

Obteniendo los estimadores de OLS

Pasemos la ecuación a notación vectorial, de esta forma nuestra solución aplicará para modelos con más de 2 parámetros, denotando el número de parámetros con k .⁵⁵

- Sea \mathbf{Y} el vector de observaciones de tamaño $n \times 1$ de la variable dependiente (las ventas, en nuestro ejemplo).
- Sea \mathbf{X} una matriz de tamaño $n \times k$ con las observaciones de k variables independientes con n observaciones cada una. Como por lo general nuestro modelo contiene un término constante, incluimos una columna de unos.
- Sea $\boldsymbol{\beta}$ un vector de tamaño $k + 1 \times 1$. Es el vector de los parámetros que deseamos estimar.
- Sea $\boldsymbol{\varepsilon}$ un vector de tamaño $n \times 1$. Es el vector de errores.

Nuestro modelo se vería entonces de la siguiente manera:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n \times k} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k+1} \end{bmatrix}_{k+1 \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1} \quad (7)$$

El modelo de arriba se puede representar de forma compacta como $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Es una forma de representar los datos poblacionales. Sin embargo, lo más común es que tengamos a nuestra disposición los datos de una **muestra**.

Lo que en la práctica significa es que probablemente nunca podremos observar los datos que componen al vector $\boldsymbol{\beta}$, pero si podemos trabajar con una estimación obtenida a través de una muestra⁵⁶.

- Sea $\hat{\boldsymbol{\beta}}$ el vector de estimaciones de los parámetros de la población, bajo el supuesto de que $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.
- Sea \mathbf{e} el vector de residuales. Nuestro objetivo en el método de OLS es minimizar $\sum e_i^2$.

La suma de los residuales al cuadrado (RSS = Residual Sum of Squares) la expresamos en notación vectorial como $\mathbf{e}'\mathbf{e}$ ⁵⁷.

⁵⁵ Esta sección muestra una solución general de la regresión por mínimos cuadrados. La versión *sin covariables* con sólo una variable independiente x y una dependiente y nos lleva a coeficientes de estimación $\hat{\beta}_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$ y $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Este resultado viene de resolver un problema de minimización que se obtiene de manera sencilla con cálculo, pero el tamaño de los polinomios que aparecen con dimensiones superiores hacen que valga la pena mejor no intentar resolverlos por ese método y recurrir al álgebra lineal.

⁵⁶ La forma de la obtención de una muestra está fuera del enfoque de este libro. Normalmente si estás trabajando con datos oficiales el trabajo de muestreo ya fue hecho, pero si tú estás obteniendo datos por tu cuenta propia, si tendrás que asegurarte de que esté bien hecha y cumpla con algunos supuestos.

⁵⁷ En el apéndice explico por qué.

Podemos expresar la RSS como

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}\end{aligned}$$

Si usamos que $\mathbf{y}'\mathbf{X}\hat{\beta} = (\mathbf{y}'\mathbf{X}\hat{\beta})' = \hat{\beta}'\mathbf{X}'\mathbf{y}$, entonces nuestra RSS se verá así

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - 2\hat{\beta}'\mathbf{X}'\mathbf{y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$$

Al igual que haríamos con la versión de dos dimensiones, requerimos obtener las condiciones de primer orden de la ecuación para encontrar el mínimo. Esto lo hacemos con la primera derivada con respecto a $\hat{\beta}$. El truco está en igualar esta derivada a cero.

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

Al despejar esta ecuación podemos obtener los valores de $\hat{\beta}$ que minimizan el valor de los residuales.

Para comprobar que se trata de un mínimo, notamos que la segunda derivada ($2\mathbf{X}'\mathbf{X}$) es una matriz positiva definida (análoga en álgebra lineal a los números positivos)⁵⁸.

De la ecuación anterior podemos obtener las llamadas “ecuaciones normales”.

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'\mathbf{y}$$

Nota que la matriz $\mathbf{X}'\mathbf{X}$ siempre será cuadrada y simétrica con tamaño $k \times k$. Si la inversa de esta matriz existe⁵⁹, la podemos aplicar a ambos lados de la ecuación para despejar $\hat{\beta}$:

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

y por lo tanto

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Esto es lo que tu computadora calcula cuando le pides que haga una regresión lineal con tus datos. Nota que no es necesario tener ningún supuesto en este punto, pues tus estimaciones sólo dependen de tu matriz de datos observados.

Hagamos un ejercicio en Python con los datos de publicidad. El siguiente código presenta los datos en forma de matrices y vectores y los aplica para calcular el vector de $\hat{\beta}$ ⁶⁰:

```
import numpy as np
import pandas as pd
```

⁵⁸ Incluí algunas notas en el apéndice que te podrán ayudar a entender esto mejor.

⁵⁹ ver el apéndice.

⁶⁰ El módulo de `numpy` tiene funciones para hacer operaciones de álgebra lineal que nos ayudarán a comprobar el resultado que acabamos de obtener para encontrar los estimadores de mínimos cuadrados. Revisa la documentación para las funciones de álgebra lineal en <https://numpy.org>

```
# Cargar el conjunto de datos
ruta_archivo = 'advertising.csv'
datos = pd.read_csv(ruta_archivo)

# Seleccionar solo las columnas TV y Sales
tv = datos['TV'].values
ventas = datos['Sales'].values

# Añadir una columna de unos para el término de intercepto
X = np.column_stack((np.ones(tv.shape[0])), tv)

# Aplicar la fórmula de regresión lineal
# Calcular (X'X)^-1
XX_inv = np.linalg.inv(X.T @ X)

# Calcular (X'X)^-1 X'y
beta_hat = XX_inv @ X.T @ ventas

# Coeficientes: Intercepto y Pendiente
intercepto, pendiente = beta_hat

# Mostrar los resultados
print("Intercepto:", intercepto)
print("Pendiente para TV:", pendiente)
```

```
Intercepto: 6.974821488229908
Pendiente para TV: 0.05546477046955883
```

Lo mejor de este resultado es que es relativamente fácil hacerlo escalar para k variables. Queda como ejercicio para el lector modificar el código anterior. Incluye la publicidad por radio y periódicos a la matriz \mathbf{X} y vuelve a calcular los coeficientes.

Propiedades de los estimadores de Mínimos Cuadrados

La propiedad principal de los estimadores es que minimizan la suma de los residuales al cuadrado. Pero hay más propiedades que podemos deducir con ligeras modificaciones. Por ejemplo, podemos sustituir el valor de $\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e}$ dentro de la ecuación normal para obtener:

$$\begin{aligned}(X'X)\hat{\beta} &= X'(X\hat{\beta} + \epsilon) \\ (X'X)\hat{\beta} &= (X'X)\hat{\beta} + X'\epsilon \\ X'\epsilon &= 0\end{aligned}$$

Podemos deducir a partir de este resultado algunas propiedades:

- **Los valores observados de X no están correlacionados con los residuales.**

Que $X'\epsilon = 0$ implica que cada columna de la matriz X tiene correlación muestral de cero con los residuales. Esto sigue siendo verdad aún cuando nuestra regresión incluye una constante.

El siguiente código muestra en Python las predicciones, los residuales y el cálculo de la correlación entre ambos. Nota que las predicciones se calculan con el producto de la matriz X con $\hat{\beta}$, esto es:

$$\hat{Y} = X\hat{\beta} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \hat{\beta}_0 + \hat{\beta}_1 x_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{bmatrix}$$

Los residuales son simplemente $Y - \hat{Y}$.

```
# Calcular los valores predichos y los residuales
predicciones = X @ beta_hat
residuales = ventas - predicciones

# Calcular la correlación entre los valores observados de TV y los residuales
correlacion = np.corrcoef(tv, residuales)[0, 1]

# Mostrar la correlación
print("Correlación entre los valores observados de TV y los residuales:", correlacion)
```

Correlación entre los valores observados de TV y los residuales: 7.864091211939043e-16

Nota que el coeficiente de correlación es prácticamente cero. Esto comprueba la propiedad⁶¹.

- **La suma de los residuales es igual a cero.**

Usemos el cálculo que hicimos de los residuales en la propiedad anterior.

⁶¹ Un truco para sacar el mayor provecho a este libro: no copies y pegues este código. Cópialo de manera consciente en tu editor. Ejecuta los bloques uno a uno. Si algo te causa dudas, hazlo por partes.

```
# Calcular la suma de los residuales
suma_residuales = np.sum(residuales)

# Mostrar la suma de los residuales
print("Suma de los residuales:", suma_residuales)
```

Suma de los residuales: -2.632560835991171e-12

La suma de los residuales en el modelo de regresión lineal es aproximadamente cero. El modelo “ajusta” los datos promediando los errores en ambas direcciones.

- **La media muestral de los residuales es cero.**

Nuevamente, podemos usar los residuales que calculamos antes para obtener este valor promedio⁶².

```
# Calcular la media de los residuales
media_residuales = np.mean(residuales)

# Mostrar la media de los residuales
print("Media de los residuales:", media_residuales)
```

Media de los residuales: -1.3162804179955856e-14

- **El hiperplano de la regresión pasa a través de las medias de los valores observados (\bar{X} y \bar{Y})**

No te intimides por la palabra “hiperplano”. En una regresión con una sola variable, nos referimos a la línea de regresión. En más dimensiones es el equivalente de esta línea⁶³.

Este paso requiere el cálculo de variables adicionales. En primer lugar, calculamos los valores promedio de las ventas y del gasto en campañas de televisión.

Luego calculamos la predicción de ventas promedio, que debería ser igual al promedio que calculamos a partir de los datos.

⁶² Cuando un número viene con una e y un número negativo pequeño es un número muy cercano a cero.

⁶³ Por ejemplo, en tres dimensiones, se vería como una hoja de papel extendida.

```
# Calcular los promedios de TV y Ventas
promedio_tv = np.mean(tv)
promedio_ventas = np.mean(ventas)

# Calcular el valor predicho de Ventas cuando TV es igual a su promedio
ventas_predichas_en_promedio_tv = beta_hat[0] + beta_hat[1] * promedio_tv
```

```
# Mostrar los resultados
print("Promedio de TV:", promedio_tv)
print("Promedio de Ventas:", promedio_ventas)
print("Ventas predichas cuando TV es igual a su promedio:", ventas_predichas_en_promedio_tv)
```

Promedio de TV: 147.0425
 Promedio de Ventas: 15.130500000000001
 Ventas predichas cuando TV es igual a su promedio: 15.1305000000000012

- Los valores de predicción de Y no están correlacionados con los residuales.

Este es un cálculo sencillo hecho con los datos que calculamos al inicio. Debemos de obtener como resultado cero.

```
# Calcular la correlación entre los valores predichos y los residuales
correlacion_predichos_residuales = np.corrcoef(predicciones, residuales)[0, 1]

# Mostrar la correlación
print("Correlación entre los valores predichos de Ventas y los residuales:",
      → correlacion_predichos_residuales)
```

Correlación entre los valores predichos de Ventas y los residuales: 7.818042265014361e-16

- La media de las predicciones de Y para la muestra será igual que la media de los Y observados.

El modelo de regresión lineal se ajusta para minimizar la suma de los cuadrados de los residuales.

Esto resulta en una distribución equilibrada de los residuales alrededor de la línea de regresión.

```
# Calcular la media de los valores predichos
media_predicciones = np.mean(predicciones)

# Mostrar la media de los valores predichos y la media de los valores observados
print("Media de los valores predichos:", media_predicciones)
print("Media de los valores observados (Ventas):", promedio_ventas)
```

Media de los valores predichos: 15.1305000000000014
 Media de los valores observados (Ventas): 15.130500000000001

Listo. Hemos comprobado con nuestros datos que las propiedades de la regresión lineal.

Ahora toca poner atención a los supuestos que hacen que un modelo de mínimos cuadrados tenga sentido.

El teorema de Gauss-Márkov y sus supuestos

I'm BLUE, da-ba-dee-da-ba-day

- Eiffel 65 feat. Gabry Ponte

El teorema de Gauss-Márkov⁶⁴ establece que si tu modelo de regresión lineal satisface cinco supuestos básicos, entonces la regresión por mínimos cuadrados producirá **estimaciones insesgadas** con la varianza más pequeña de **todos** los estimadores lineales posibles.

En otras palabras, el modelo será **BLUE**.

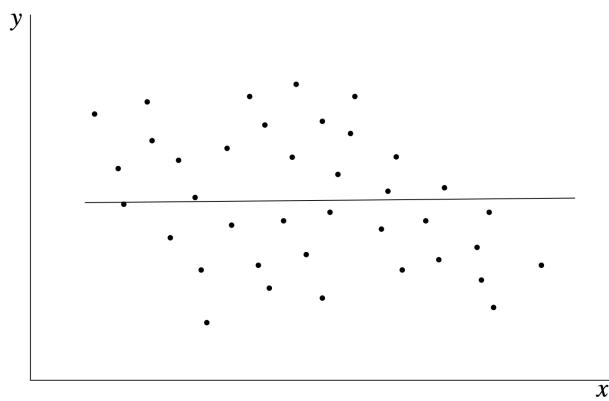
Estos son los supuestos del teorema de Gauss-Márkov

- **Supuesto # 1: Los parámetros deben ser lineales.**

Si lo pensamos detenidamente, se trata de un supuesto fuerte.

Si hiciéramos un diagrama de dispersión, deberíamos ver algo parecido a lo que mostró el diagrama de dispersión del gasto en TV contra las ventas.

Pero a veces nos encontramos con conjuntos de datos que se ven como la imagen 14, en la que no se ve una relación clara entre variables.



⁶⁴ Nota que le puse un acento a Márkov. En ruso, el acento va en la a. De otra forma, el default de los hispanohablantes sería decirle Markov, que suena cómicamente a "zanahoria" en ruso.

Figura 13: Andrei Márkov (izquierda) y Carl Friedrich Gauss (derecha) jugando ajedrez. Fuente: imaginado por mi y hecho con chatGPT.



Figura 14: Un diagrama de dispersión nos muestra que X y Y son variables que no parecen tener ninguna relación entre sí.

Aquí la relación entre las variables no parece ser tan lineal. La línea de regresión parece no estar muy cómoda ahí.

Sin embargo, no debemos dejarnos engañar. La regresión lineal la podemos hacer con múltiples dimensiones (variables). En ocasiones, las variables adicionales de nuestro modelo hacen que la linealidad tenga sentido.

En la imagen del ejemplo, una simple clasificación de las variables revela el patrón oculto.

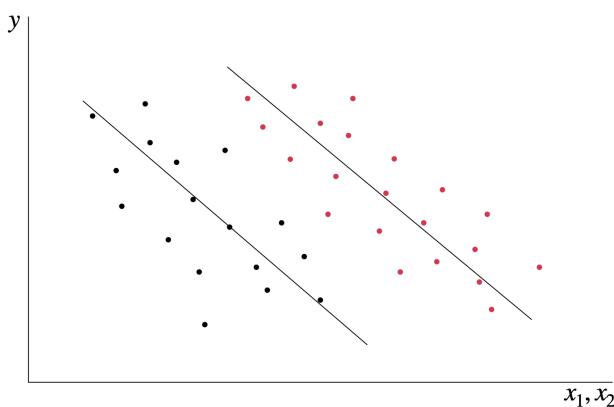


Figura 15: No es sino hasta que agregamos una tercera variable que podemos revelar la verdadera relación que hay entre las variables.

Son los mismos puntos del diagrama de dispersión, pero separarlos por color revela dos relaciones lineales diferentes.

- **Supuesto #2: Los datos deben ser tomados de un muestreo aleatorio de la población.**

Este es un supuesto básico: nuestra muestra debe ser aleatoria.

Si no lo hacemos de esta manera, nos estamos arriesgando a encontrar sesgos en nuestras bases de datos. ¿Cómo nos aseguramos de que nuestra muestra es aleatoria? Si los datos los estamos recabando nosotros, tenemos que tomar las precauciones al momento de diseñar nuestra muestra de que no estamos generando ningún sesgo por la forma en la que estamos recabando los datos.

Por ejemplo: los datos que se obtienen por teléfono generalmente son considerados de menor calidad que las encuestas realizadas en los hogares. ¿Por qué? porque las personas con teléfono podrían tener diferencias clave con las personas que no cuentan con él.

Hacer un muestreo apropiado es mucho más que sólo entrar a un recurso en línea a sacar el número del tamaño de muestra

(Algunos recursos para aprender sobre muestreo son los libros de Pérez López⁶⁵ y el de Wasserman⁶⁶)

- **Supuesto #3: No hay colinealidad: los regresores no están correlacionados perfectamente entre sí.**

En nuestro modelo de álgebra lineal, esto se determina cuando \mathbf{X} es una matriz $n \times k$ de rango completo⁶⁷.

Usemos Python para verificar si esto es verdad en los datos de publicidad.

```
import numpy as np
import pandas as pd

# Cargar el conjunto de datos
ruta_archivo = 'advertising.csv'
datos = pd.read_csv(ruta_archivo)

# Seleccionar las columnas de interés: TV, Radio,
# → Newspaper y Sales
tv = datos['TV'].values
radio = datos['Radio'].values
newspaper = datos['Newspaper'].values
ventas = datos['Sales'].values

# Añadir una columna de unos para el término de
# → intercepto
X = np.column_stack((np.ones(tv.shape[0]), tv, radio,
                     newspaper))

# Calcular el rango de la matriz X para verificar la
# → multicolinealidad
rango_X = np.linalg.matrix_rank(X)

# Mostrar el rango de la matriz X
print("Rango de la matriz X:", rango_X)

# Verificar si la matriz X es de rango completo
num_columnas = X.shape[1]
es_rango_completo = rango_X == num_columnas
print("¿Es la matriz X de rango completo (sin
      → multicolinealidad)?", es_rango_completo)
```

Y este es el resultado al ejecutar el código:

```
Rango de la matriz X: 4
¿Es la matriz X de rango completo (sin multicolinealidad)?
→ True
```

⁶⁵ César Pérez López. *Muestreo estadístico: conceptos y problemas resueltos*. Pearson Education S.A., Madrid, 2005

⁶⁶ Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, corrected 2nd printing edition, 2004. ISBN 978-0387402727

⁶⁷ Si \mathbf{X} no tiene rango completo, no se puede calcular la matriz $(\mathbf{X}^\top \mathbf{X})^{-1}$, que es necesaria para obtener los estimadores de mínimos cuadrados. El rango de una matriz es el número de filas o columnas linealmente independientes. Una matriz de rango completo es una en la que ninguna columna de \mathbf{X} se puede escribir como la combinación lineal de otras.

En el código anterior comprobamos que la matriz tiene rango completo y por lo tanto, no presenta colinealidad.

Esta no es la forma tradicional de verificar este supuesto. La forma tradicional es revisar las correlaciones entre las variables. El siguiente bloque de código genera una matriz de correlación.

```
# Calcular la matriz de correlación para las variables
→ predictoras matriz_correlacion = datos[['TV',
→ 'Radio', 'Newspaper']].corr()

# Mostrar la matriz de correlación
print("Matriz de correlación:\n", matriz_correlacion)
```

	TV	Radio	Newspaper
TV	1.000000	0.054809	0.056648
Radio	0.054809	1.000000	0.354104
Newspaper	0.056648	0.354104	1.000000

La forma de usar la matriz de correlación es con una inspección visual.

Aparte de la diagonal de 1s, podemos ver que la correlación más alta es entre el gasto en periódico y el de radio, con un 35 %. Si encontráramos que dos o más variables tienen un coeficiente de correlación demasiado cercano a 1, entonces podríamos sospechar autocorrelación⁶⁸.

En ese caso tendríamos que usar una técnica adicional.

Calcularemos el factor de inflación de la varianza (VIF, por sus siglas en inglés). Esta técnica sirve en los casos en los que inspeccionar la matriz de correlación no da un resultado determinante.

La **multicolinealidad** se puede ocultar: una variable podría ser una combinación lineal de múltiples columnas. Esto es algo que no se vería en la matriz de correlación, pero que si se puede mostrar con este indicador:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (8)$$

donde R_j^2 se hace con la regresión de la variable x_j con todos los demás predictores. El siguiente código hace la prueba en nuestra base de datos:

⁶⁸ Esta técnica no es muy objetiva por si misma. Por lo general se establecen parámetros como “arriba de 0.8” que se consideran demasiado altos. Entre 0.7 y 0.8 se consideran en el borde y abajo de 0.7 se piensa que no debería haber problema.

```

from statsmodels.stats.outliers_influence import
    variance_inflation_factor

# Función para calcular el VIF para cada variable
def calcular_vif(X):
    vif = pd.DataFrame()
    vif["variables"] = X.columns
    vif["VIF"] = [variance_inflation_factor(X.values, i)
        for i in range(X.shape[1])]
    return vif

# Calcular VIF para las variables predictoras
vif_df = calcular_vif(datos[['TV', 'Radio',
    'Newspaper']])

# Mostrar el VIF para cada variable
print("VIF para cada variable:\n", vif_df)

```

Que genera el siguiente resultado

```

VIF para cada variable:
variables      VIF
0          TV  2.486772
1       Radio  3.285462
2  Newspaper  3.055245

```

Todos los factores están por debajo de 5, por lo que **no hay problemas de multicolinealidad** en nuestros datos⁶⁹.

En ocasiones, la matriz de correlaciones será suficiente para encontrar la presencia (o ausencia) de multicolinealidad, pero el VIF es un método que incluye una regla de oro mas fácil de interpretar.

- **Supuesto #4: Exogeneidad. los regresores no están correlacionados con el término de error.**

También se le conoce como el supuesto de media condicional cero, y es probablemente el supuesto más crítico para la inferencia causal.

Se expresa así:

$$E(\epsilon|\mathbf{X}) = 0$$

En palabras, no hay observación dentro de las variables independientes que contengan información sobre el valor esperado del error.

⁶⁹ En el apéndice te he dejado una guía para interpretar el VIF

La manera mas práctica de comprobar esta propiedad es con un gráfico de las predicciones con los residuales. Una inspección visual suele ser suficiente para identificar si existe (o no) un patrón en los datos.

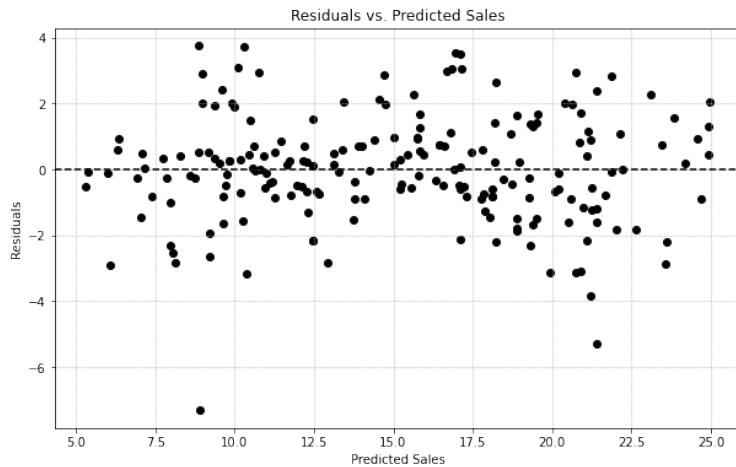


Figura 16: Los residuales y los valores de predicción no muestran estar correlacionados. La gráfica muestra una linea de regresión cercana a cero.

```

import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Cargamos la base de datos
file_path = 'advertising.csv' # Reemplazamos la ruta
data = pd.read_csv(file_path)

# Definir la variable independiente (X) y la dependiente (y)
X = data[['TV', 'Radio', 'Newspaper']]
y = data['Sales']

# Agregar una constante al modelo (para el intercepto)
X = sm.add_constant(X)

# Ajustar un modelo de regresión lineal
model = sm.OLS(y, X).fit()
# Obtener las predicciones y residuales del modelo
predictions = model.predict(X)
residuals = model.resid

# Hagamos un gráfico de los residuales vs. los valores de predicción
plt.figure(figsize=(10, 6))
plt.scatter(predictions, residuals, color='black')
plt.axhline(y=0, color='black', linestyle='--')
plt.xlabel('Predicted Sales')
plt.ylabel('Residuals')
plt.title('Residuales vs. Predicción de ventas')

```

```
| plt.grid(True, which='both', linestyle='--', linewidth=0.5)
| plt.show()
```

Este gráfico muestra que no hay un patrón definido.

Hay otros métodos para comprobarlo, como pruebas de correlación o la prueba de Durbin-Watson, pero por lo general la inspección debería ser suficiente para estos casos.

- **Supuesto #5: Homoscedasticidad. la varianza del error es constante para todos los valores de los regresores.**

El gráfico anterior es útil también para la comprobación de la homoscedasticidad. Podríamos hacer una inspección visual que compruebe que esa varianza es constante.

Pero aquí haremos un gráfico adicional con el siguiente código:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm

# Cargar la base de datos
data = pd.read_csv('advertising.csv')

# Ejecutar la regresión lineal
X = data[['TV', 'Radio', 'Newspaper']] # Variables independientes
y = data['Sales'] # Variable dependiente

# Agregar constante al modelo
X = sm.add_constant(X)

# Ajustar el modelo de regresión
model = sm.OLS(y, X).fit()

# Obtener los residuales y los valores ajustados
residuals = model.resid
fitted = model.fittedvalues

# Gráfico de los residuales vs valores ajustados
plt.figure(figsize=(10, 6))
sns.residplot(x=fitted, y=residuals, color='black', lowess=True)
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residuals vs. Fitted Values')
plt.axhline(y=0, color='black', linestyle='--')
plt.show()
```

Este código crea una línea horizontal en el cero, y la línea gruesa

muestra la relación entre los residuales y los valores ajustados. De nuevo, el gráfico no muestra una tendencia clara (esto es bueno!). En general, una inspección visual del diagrama de dispersión nos ayuda a entender lo que está pasando.

Es difícil capturar de manera visual si hay homoscedasticidad o no. La figura 17 compara residuales y predicciones, pero no podemos dar nuestro veredicto sólo con ver la imagen.⁷⁰ El siguiente código genera el diagrama de dispersión de nuestros datos.

⁷⁰ El problema con la inspección visual es que me hace parecer como alguien que está leyendo hojas de té.

```
# Correcting the Residuals vs Predictors plots
fig, axes = plt.subplots(1, 3, figsize=(18, 6))

# Plotting for each predictor
for i, col in enumerate(['TV', 'Radio', 'Newspaper']):
    sns.scatterplot(x=data[col], y=residuals, color='black', ax=axes[i])
    axes[i].set_title(f'Residuals vs {col}')
    axes[i].set_xlabel(col)
    axes[i].set_ylabel('Residuals')
    axes[i].axhline(y=0, color='black', linestyle='--')

plt.tight_layout()
plt.show()
```

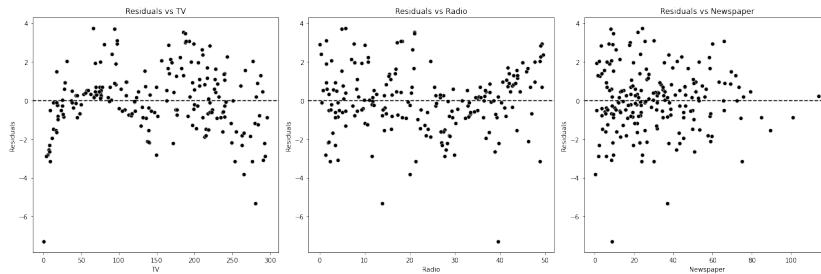


Figura 17: Gráfico que compara residuales vs cada uno de los predictores en un diagrama de dispersión. No se ve un patrón específico.

Decir “una diferencia significativa” es subjetivo. La diferencia que unos ven pequeña otros la ven muy grande. Por eso, lo más seguro es usar una prueba estadística. El siguiente código usa la prueba Breusch-Pagan del módulo `statsmodels` para comprobar nuevamente que no hay **heteroscedasticidad**. La prueba de Breusch-Pagan se obtiene con la regresión de los residuos al cuadrado contra las variables independientes usando una ecuación auxiliar con la forma $\hat{\epsilon}^2 = \gamma_0 + \gamma_1 x + v$, donde $\hat{\epsilon}$ denota los residuales⁷¹, y v es el error de una regresión entre x y $\hat{\epsilon}$. La hipótesis nula es que los errores del modelo tienen varianza constante.

⁷¹ Nota que el error ϵ es diferente que el residual $\hat{\epsilon}$. Los errores vienen del proceso de generación de la información, mientras que los residuales es lo que queda después de haber estimado el modelo.

```
from statsmodels.stats.diagnostic import het_breushpagan

# Aplicar una prueba de Breusch-Pagan
bp_test = het_breushpagan(residuals, model.model.exog)

# Extraer los resultados
bp_test_statistic, bp_test_pvalue = bp_test[:2]

bp_test_statistic, bp_test_pvalue
```

(3.9785268214219682, 0.26379220043199536)

Como el p-value es mayor a 0.05, no podemos rechazar la hipótesis nula de homoscedasticidad. En otras palabras, no hay evidencia de heteroscedasticidad en los residuales del modelo.

Cómo interpretar el reporte de regresión: La guía del economista principiante para que acepten su primer artículo

Hay dos usos para un modelo de regresión: predicción o inferencia.

En la inferencia, estamos tratando de saber **por qué** una variable se comporta de cierta manera. Estos son los métodos más tradicionales y que se acercan más a lo que hacemos en inferencia causal. En predicción, estamos intentando construir un modelo que reconstruya un resultado con información dada⁷².

La **inferencia causal** es diferente a estas dos. Por ejemplo, si aumentamos el precio de nuestro producto un 10 % y observamos que la demanda cae, queremos saber cuánta de esa caída en la demanda se atribuye a los precios y cuánto viene de otros factores externos.

En la sección pasada pasamos por todas las pruebas de hipótesis **antes de ver los resultados de la regresión**. Lo hicimos de esta manera porque una vez cumplimos con los supuestos, podemos enfocarnos en las estimaciones de los parámetros en el modelo.

Hagamos entonces la regresión lineal de un modelo por mínimos cuadrados:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (9)$$

Donde y son las ventas y x_1, x_2 y x_3 representan los diferentes

⁷² Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer, 1 edition, 2013. ISBN 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7. URL <https://www.statlearning.com/>

medios de publicidad. Este modelo ya tiene suficientes dimensiones (variables) para que no sea posible mostrarlo en un gráfico de dispersión. Pero el principio es exactamente el mismo y nosotros no le tenemos miedo a las dimensiones superiores.

Usa este código en Python para hacer tu primera regresión por mínimos cuadrados. Verás un reporte de regresión como el siguiente.

```
import statsmodels.api as sm

# Definir las variables independientes (X) y la variable dependiente (y)
X = data[['TV', 'Radio', 'Newspaper']] # Variables independientes
y = data['Sales'] # Variable dependiente

# Añadir una constante al modelo (intercepción)
X = sm.add_constant(X)

# Realizar la regresión OLS
model = sm.OLS(y, X).fit()

# Mostrar el informe completo de la regresión
model.summary()
```

OLS Regression Results						
Dep. Variable:	Sales	R-squared:	0.903			
Model:	OLS	Adj. R-squared:	0.901			
Method:	Least Squares	F-statistic:	605.4			
Date:	Wed, 07 Feb 2024	Prob (F-statistic):	8.13e-99			
Time:	10:42:10	Log-Likelihood:	-383.34			
No. Observations:	200	AIC:	774.7			
Df Residuals:	196	BIC:	787.9			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	4.6251	0.308	15.041	0.000	4.019	5.232
TV	0.0544	0.001	39.592	0.000	0.052	0.057
Radio	0.1070	0.008	12.604	0.000	0.090	0.124
Newspaper	0.0003	0.006	0.058	0.954	-0.011	0.012
Omnibus:	16.081				2.251	
Prob(Omnibus):	0.000				27.655	
Skew:	-0.431				9.88e-07	
Kurtosis:	4.605				454	

Vamos a interpretar este resultado parte por parte.

La primera sección es un reporte general de cómo fue la regresión, el número de observaciones y el tipo de modelo. También hay algunos estadísticos a los que debemos de poner atención:

- **R cuadrada y R cuadrada ajustada.** Indican el ajuste de los datos a la linea de regresión. El ajuste indica que tan cercanos están los datos a la línea de regresión. Ten cuidado con este indicador y ponlo en contexto, pues en ocasiones el ajuste no significa por sí mismo que sea un mejor modelo. Y al contrario, un mal ajuste no significa necesariamente que sea un modelo que debamos descartar.

Una R^2 de 0.903 significa que un 90.3 % de la variación en las ventas se explica por el modelo. Nada mal.

El R^2 es un número que va del 0 al 1. Números cercanos al 0 significan que no hay ajuste y números cercanos al 1 indican mucho ajuste. El R^2 ajustado toma en consideración el número de predicciones en el modelo. Es una visión más precisa.

- **El estadístico F y Prob(estadístico F)** El estadístico F prueba la hipótesis nula de que todos los coeficientes de regresión son igual a cero. Un estadístico F grande (605.4) indica que esta hipótesis nula es falsa.

No existe una regla clara sobre cuando el valor de F sea inequívocamente grande. Depende mucho del modelo. Por su parte, la probabilidad `Prob(F-statistic)` es un número muy pequeño, cercano a cero. Como podrás intuir, si es una probabilidad, debe estar entre cero y uno. Indica la probabilidad de observar un valor del estadístico de F tan extremo (o más) que el que observamos, asumiendo que la hipótesis nula sea verdadera. Es decir: si todos los coeficientes fueran cero, ¿podría F hacer esto?

- **Grados de libertad.** `Df` y `Df residual` quieren decir “degrees of freedom” o grados de libertad. Se refiere al número de observaciones menos el número de parámetros estimados. En este modelo no es algo que nos pueda causar problema, porque sólo usamos 3 parámetros, pero en modelos más complejos puede ayudarnos a identificar problemas de sobreajuste.
- **AIC y BIC.** Significan respectivamente: “criterio de información de akaike” y “criterio de información bayesiano”. Estos son criterios que se usan para la selección de modelos, no para comprobar su significancia. Usaremos estos cuando tengamos que hacer una comparación entre modelos. Incluí una explicación en el apéndice para explicar esto a más detalle.

La segunda sección del reporte muestra los coeficientes de la regresión y

Esta es la sección del reporte de regresión a la que debes poner

atención para interpretar los resultados y determinar si son significativos.

Veamos cada elemento paso a paso en un modelo lineal sencillo que determina el valor de las ventas (y) en función del gasto en publicidad en medios como TV, Radio o Periódicos (antes de las redes sociales).

Columna #1: coeficientes. Esta es la que determina el valor de tus betas en el modelo de regresión. $\text{const} = 4.6251$ significa que beta cero es igual a 4.62.

En un modelo lineal significa que si el gasto en publicidad fuera cero, aún tendríamos ventas de cuatro mil seiscientas unidades aproximadamente.

El resto de los coeficientes indican la contribución marginal que tiene cada medio a las ventas. De aquí podemos ver que la TV es la que contribuye más a las ventas. Cada 20 dólares gastados en publicidad en TV contribuye a aproximadamente 1.1 (mil) unidades adicionales vendidas.

Columna #2: Errores estándar. Es el ruido de nuestros datos en el modelo. Entre más grande sea el **error estándar**, menos significativo sera el modelo.

Para determinar si un error estándar es grande o pequeño, es necesario compararlo con los coeficientes. Un coeficiente de 0.0544 hace que un error estándar de 0.001 sea pequeño en comparación. Pero si el coeficiente fuera también de 0.001, entonces esos datos muy seguramente no serán significativos.

Si esta comparación aún te parece subjetiva, para eso está la t en la siguiente columna.

Columna #3: Estadístico t. Algunas veces lo verás como t de Student. Es una razón entre la *señal* y el *ruido*. La señal en una regresión es el coeficiente, y el ruido es el error estándar⁷³.

Un tamaño de muestra más grande hace que la señal sea más poderosa. El estadístico t es un número positivo o negativo. Entre más grande sea su valor absoluto, más probable es que los resultados sean significativos.

Esto lo verificamos con el *p-value* en la siguiente columna.

Columna #4: p-value. Es una medida de probabilidad que se obtiene a partir del estadístico t. Indica la probabilidad de obtener un

⁷³ Student era el seudónimo que usaba William S. Gosset cuando trabajaba en la cervecería Guinness. La prueba t permitió a la cervecería hacer crecer la producción sin perder la calidad del producto.

resultado al menos tan extremo como el que observamos, bajo el supuesto de que la hipótesis nula ($\beta_i = 0$) sea verdadera.

Hay una convención de que un p-value menor a 0.05 implica que los resultados son significativos.

Mi consejo es que lo consideres, lo apliques, pero no te cases con esta idea. Después de todo, no hay razón científica que diga que a partir de 0.05 el resultado es significativo por completo.

Cuando adquieres más experiencia en estadística, tomas en contexto el p-value con los intervalos de confianza.

Columnas #5 y #6: Intervalos de confianza. Son el rango en el que se encuentra el verdadero valor del parámetro beta.

Recuerda que los coeficientes son estimaciones que obtenemos a partir de una muestra. El [intervalo de confianza](#) te muestra un rango.

Nota que la columna #5 muestra un número más bajo que el coeficiente y la columna #6 uno más alto. Entre más amplio sea el intervalo, más incertidumbre hay respecto al valor del coeficiente.

Al contrario, un intervalo más angosto significa más certidumbre.

En otras palabras, tenemos un 95 % de certidumbre de que el efecto de los anuncios por TV tienen un efecto en las ventas que va de 0.052 a 0.057 (miles de unidades/dólar).

¿Qué pasa si mi regresión no cumple con los supuestos?

Mi ejemplo de datos se ve muy bonito. Todo funcionó muy bien⁷⁴.

Pero tú y yo sabemos que eso no es lo que pasará cuando lo intentes hacerlo por tu cuenta y con tus propios datos. ¡El problema no eres tú! yo mismo no apostaría a que mi próxima regresión saldrá sin problemas. Son solo gajes del oficio.

La verdad es que es algo muy común. La estrategia que debes tomar depende mucho del problema al que te estás enfrentando.

Resolver este tipo de problemas es algo que podrás aprender a hacer con la práctica. Lo más importante es que conozcas la teoría que vimos en este capítulo a profundidad y que desarrolles una correcta intuición de lo que está pasando al momento de hacer este tipo de regresiones. El resto lo puede hacer la computadora.

Estas son algunas acciones que puedes tomar si tu regresión no cumple con los supuestos que revisamos con anterioridad.

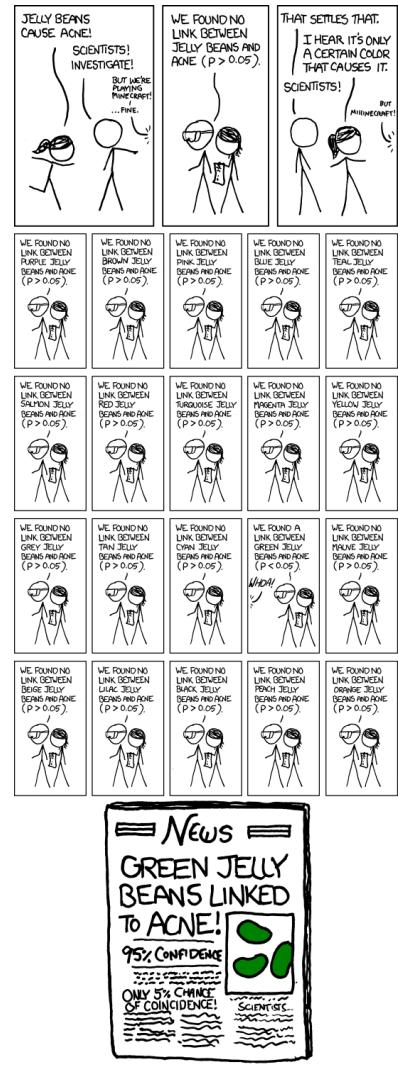


Figura 18: El 5 % del p-value es lo mismo que uno en veinte. Con ese parámetro, no es necesario recurrir al p-hacking para obtener resultados "significativos".

Fuente: [xkcd](https://xkcd.com/217/)

⁷⁴ En los libros nos enfocamos en los datos que funcionan bien, pero cuando tomamos datos de la realidad es cuando surgen los miles de problemas y errores, sin nadie a quién acudir para resolverlos. Hace no mucho, tenía que resolverlo buscando algún problema similar en StackOverflow, pero ahora toda esa información se quedó en los modelos de lenguaje grandes como chatGPT. Vale la pena aprender a pedir ayuda a la IA sobre nuestros problemas con datos.

- Revisa la especificación del modelo.
- Transforma los datos.
- Usa técnicas robustas.
- Incorpora variables adicionales o interacciones.
- Considera métodos no paramétricos u otro tipo de técnicas.

Finalmente, la calidad de los datos es mucho más importante que los modelos.

El modelo de regresión es relativamente simple en comparación a modelos como árboles de regresión o redes neuronales, pero en muchos casos es preferible precisamente por su simplicidad y facilidad de interpretación. Pero lo que más hace la diferencia en tu trabajo no es la complejidad del modelo: es la calidad de los datos.

Si los datos están recabados de una muestra bien diseñada, sin sesgos, con preguntas bien planteadas y congruentes con nuestros objetivos, hasta una diferencia de medias bien hecha es mejor que una red neuronal hecha con malos datos.

Apéndice: Algunas preguntas que te pudieron haber quedado, explicadas con más detalle

¿Por qué $\mathbf{e}'\mathbf{e}$ es la suma de residuos al cuadrado?

En primer lugar, no se debe confundir con $\mathbf{e}\mathbf{e}'$, que es la matriz de varianza-covarianza de los [residuales](#). Si ponemos $\mathbf{e}'\mathbf{e}$ como vectores, se ve claramente que el resultado es una suma de residuales al cuadrado.

$$\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}_{1 \times n} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n \times 1} = \left[e_1 \times e_1 + e_2 \times e_2 + \cdots + e_n \times e_n \right]_{1 \times 1}$$

Guía breve de diferenciación con matrices

Sean a y b vectores de tamaño $k \times 1$.

$$\frac{\partial a'b}{\partial b} = \frac{\partial b'a}{\partial b} = a$$

Sea A una matriz simétrica.

$$\frac{\partial b' A b}{\partial b} = 2Ab = 2b' A$$

Por lo tanto,

$$\frac{\partial 2\beta' \mathbf{X}' \mathbf{y}}{\partial b} = \frac{\partial 2\beta' (\mathbf{X}' \mathbf{y})}{\partial b} = 2\mathbf{X}' \mathbf{y}$$

y

$$\frac{\partial \beta' \mathbf{X}' \mathbf{X} \beta}{\partial b} = \frac{\partial \beta' A \beta}{\partial b} = 2A\beta = 2\mathbf{X}' \mathbf{X} \beta$$

donde $\mathbf{X}' \mathbf{X}$ es una matriz de $k \times k$.

¿Cómo que la inversa de $\mathbf{X}' \mathbf{X}$ podría no existir?

La inversa de una matriz $\mathbf{X}' \mathbf{X}$ (donde \mathbf{X}' es la transposición de la matriz \mathbf{X} y $\mathbf{X}' \mathbf{X}$ es el producto matricial de \mathbf{X}' con \mathbf{X}) podría no existir si la matriz no es invertible.

Aquí hay algunas razones por las cuales $\mathbf{X}' \mathbf{X}$ podría no ser invertible:

1. **Columnas linealmente dependientes:** Si la matriz \mathbf{X} tiene columnas que son combinaciones lineales de otras columnas (es decir, multicolinealidad perfecta), entonces $\mathbf{X}' \mathbf{X}$ no será de rango completo y por lo tanto no tendrá una inversa.
2. **Insuficientes observaciones:** Si hay menos observaciones que variables (es decir, la matriz \mathbf{X} tiene más columnas que filas), entonces $\mathbf{X}' \mathbf{X}$ será de rango deficiente y no invertible.
3. **Datos duplicados o insuficientemente variados:** Si las filas de \mathbf{X} son todas iguales o hay una falta de variabilidad suficiente en los datos, esto también puede conducir a una matriz $\mathbf{X}' \mathbf{X}$ que no sea invertible.

Para asegurar la invertibilidad de $\mathbf{X}' \mathbf{X}$ en un análisis de regresión, a menudo se requiere que la matriz \mathbf{X} tenga rango completo, lo que significa que todas las columnas de \mathbf{X} deben ser linealmente independientes y debe haber un número suficiente de observaciones no duplicadas.

¿Cómo funciona el coeficiente de correlación?

El coeficiente de correlación lineal es un número que va de -1 a 1⁷⁵ y ayuda a entender qué tan relacionada está una variable con la otra.

Aquí algunas reglas generales.

⁷⁵ Sólo en el caso de distribuciones elípticas, como lo es la normal multivariada. Véase

Paul Embrechts, Alexander J. McNeil, and Daniel Straumann. Correlation and dependence in risk management: Properties and pitfalls. In M. A. H. Dempster, editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, 2002. doi: 10.1017/CBO9780511615337.008

- El signo indica la dirección de la correlación. Un coeficiente positivo indica que cuando una variable aumenta, la otra también lo hace. Lo contrario pasa con una correlación negativa. Por ejemplo, podríamos encontrar una correlación positiva entre el calor y las ventas de helado. Por el contrario, podríamos encontrar una correlación negativa entre la cantidad de ejercicio y el nivel de estrés: a mayor ejercicio, menor estrés.
- El valor absoluto del coeficiente indica la fuerza de la relación. Un valor cercano a 1 (ya sea positivo o negativo) indica una relación fuerte, mientras que un valor cercano a 0 indica una relación débil o inexistente.
- Una correlación de 0 indica que no hay una relación lineal entre las variables. Sin embargo, esto no significa que no haya ningún tipo de relación; podría haber una relación no lineal que este coeficiente no detecta.
- Es importante recordar que la correlación no implica causalidad. Dos variables pueden estar correlacionadas sin que una cause a la otra. Por ejemplo, puede haber una correlación entre el consumo de chocolate y el número de premios Nobel por país, pero eso no significa que comer chocolate cause ganar premios Nobel⁷⁶.

¿Cómo funciona el VIF?

El **factor de inflación de la varianza (VIF)** evalúa cuánto se incrementa la varianza de un coeficiente de regresión debido a la multicolinealidad. Se calcula para cada variable predictora y se basa en el nivel en el que esa variable predictoría está correlacionada con las otras variables predictoras en el modelo.

El VIF de una variable se calcula de la siguiente manera:

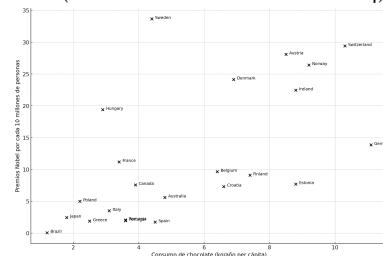
1. Se realiza una regresión lineal donde la variable en cuestión es tratada como la variable dependiente y todas las demás variables predictoras como las independientes.
2. Se calcula el coeficiente de determinación (R^2) de esta regresión.
3. El VIF se calcula como $VIF = \frac{1}{1-R^2}$.

Interpretación del VIF

- Un VIF de 1 indica que no hay correlación entre la variable predictora en cuestión y las demás.

⁷⁶ Aloys Leo Prinz. Chocolate consumption and noble laureates. *Social Sciences & Humanities Open*, 2(1): 100082, 2020. ISSN 2590-2911. DOI: <https://doi.org/10.1016/j.ssho.2020.100082>. URL <https://www.sciencedirect.com/science/article/pii/S2590291120300711>

Figura 19: Hay una correlación entre el número de premios Nobel que gana un país y su consumo de chocolate. Naturalmente, esa correlación no implica causalidad. Fuente: Leo Prinz, 2020 (actualizado con datos hasta 2024)



- Un VIF entre 1 y 5 sugiere una correlación moderada, pero generalmente no es lo suficientemente severa como para requerir atención.
- Un VIF mayor a 5 puede indicar una correlación problemática y podría necesitar atención, dependiendo del contexto y del nivel de precisión necesario en el análisis.
- Un VIF mayor a 10 es comúnmente considerado un indicador claro de multicolinealidad severa.

Importancia del VIF

El **factor de inflación de la varianza (VIF)** es una herramienta útil para detectar multicolinealidad en los modelos de regresión lineal. Al identificar las variables con VIFs altos, los analistas pueden considerar eliminar estas variables, combinarlas con otras, o utilizar técnicas estadísticas para manejar la multicolinealidad y así mejorar la calidad y la interpretación del modelo de regresión.

¿Por qué es importante identificar la multicolinealidad?

La multicolinealidad en los modelos de regresión lineal es problemática por varias razones:

1. **Estimaciones Inestables de los Coeficientes:** Cuando las variables predictoras están altamente correlacionadas, pequeñas variaciones en los datos pueden llevar a grandes cambios en los coeficientes de las variables. Esto hace que los coeficientes sean poco fiables y difíciles de interpretar.
2. **Confianza Reducida en la Significancia de las Variables:** La multicolinealidad puede inflar las varianzas de los coeficientes de las variables predictoras. Esto significa que incluso si una variable es importante en la predicción de la variable dependiente, es posible que no aparezca como significativa en la regresión debido a la alta varianza de su coeficiente.
3. **Interpretaciones Difíciles:** Cuando las variables predictoras están altamente correlacionadas, se vuelve complicado discernir el efecto individual de cada variable sobre la variable dependiente. Esto es porque los efectos de las variables correlacionadas se superponen y se confunden entre sí.

4. **Modelos Sobreajustados:** La multicolinealidad puede llevar a modelos sobreajustados, especialmente si hay un número excesivo de variables predictoras correlacionadas. Un modelo sobreajustado funciona bien con los datos de entrenamiento pero tiende a tener un rendimiento pobre con nuevos datos no vistos.
5. **Dificultad en la Selección de Modelos:** En la presencia de multicolinealidad, es difícil determinar cuál variable debe ser incluida o excluida del modelo. Los criterios de selección de modelos, como el criterio de información Akaike (AIC) o el criterio de información bayesiano (BIC), pueden verse afectados por la multicolinealidad.

Por estas razones, es importante detectar y abordar la multicolinealidad en la fase de análisis de datos para asegurar que el modelo de regresión sea confiable, interpretable y útil para la toma de decisiones.

El supuesto de media condicional cero

Este es el supuesto más crítico de la regresión lineal.

Se le llama el supuesto de **media condicional cero**. Establece que los regresores no deben estar correlacionados con el término de error.

En otras palabras: no hay **endogeneidad**⁷⁷.

En la práctica, implica es que no debe existir ningún patrón en los residuales. No se debe ver que generen patrones lineales o cuadráticos de ningún tipo.

¿Cómo se soluciona la endogeneidad en caso de existir?

Imaginemos que al graficar los residuales vs las predicciones encontramos una relación lineal. Eso implica que en los residuales hay escondida una variable.

Si conocemos lo suficiente sobre nuestras variables podemos encontrar la variable (o una buena proxy) que nos ayude a explicar mejor el comportamiento de nuestra variable de interés.

El truco es entonces:

1. Regresar a la teoría y encontrar la variable que falta.
2. Incluir la variable o una proxy apropiada al modelo de regresión.
3. Volver a hacer las pruebas.

Si en la nueva prueba ya no hay Endogeneidad, se ha solucionado el problema y podemos usar los resultados.

⁷⁷ La endogeneidad es un problema muy profundo que en ocasiones no se puede diagnosticar usando simples pruebas. Todos sabemos que correlación no implica causalidad, pero ¿podría haber causalidad cuando no hay correlación? Imagina un banco central que sube las tasas de interés para evitar los aumentos en la inflación. Si la variable de inflación permanece inmóvil ante los movimientos de la tasa de interés es difícil establecer una relación entre esas variables. Si tuviéramos el contrafactual de la inflación que hubiéramos observado sin los aumentos de la tasa de interés, la relación entre las variables sería clara.

Un poco extra sobre Gauss

Gauss es uno de los matemáticos más famosos con justa razón. Se le conoce como “el principio de las matemáticas”, por sus grandes contribuciones al álgebra, al análisis, la astronomía y la física.

Hay historias increíbles sobre Gauss. Se dice que a los tres años ya le corregía las matemáticas a su papá y que logró descifrar la fecha exacta de su nacimiento años después de que su madre lo olvidó.

Pero la historia más conocida sobre la infancia de Gauss es la de aquella vez que un maestro les dejó la agobiante tarea de **sumar todos los números del 1 al 100**⁷⁸.

La intención del maestro era mantener quietos a los niños por media hora. Gauss llegó casi al instante con la respuesta.

Para llegar al cálculo notó que sumar $100 + 1$ daba el mismo resultado que sumar $99 + 2 : 101$. Este mismo resultado se generaba en todos los 50 pares que se forman en la suma. Por lo tanto el resultado era $101 \times 50 = 5050$.

Es un resultado brillante que además se puede generalizar para cualquier número. La suma consecutiva de los números de 1 a n por lo tanto sería:

$$\sum_{i=1}^n i = \frac{n(n+1)}{2}$$

Pruebas de hipótesis

Las pruebas de hipótesis son un componente fundamental en la estadística y la investigación científica. Aquí está una explicación detallada de su concepto y uso:

¿Qué son las Pruebas de Hipótesis?

Las pruebas de hipótesis son procedimientos estadísticos que se utilizan para determinar si hay suficiente evidencia en una muestra de datos para inferir que una condición particular es verdadera para toda la población. Estas pruebas se basan en dos hipótesis: la hipótesis nula (H_0) y la hipótesis alternativa (H_1 o H_a). La hipótesis nula generalmente representa una afirmación de no efecto o de estado normal, mientras que la hipótesis alternativa representa lo que el investigador busca probar.

Por qué son importantes las Pruebas de Hipótesis

⁷⁸ Brian Hayes. Gauss's day of reckoning. *American Scientist*, 94(3):200–204, 2006. DOI: 10.1511/2006.59.200. URL <https://www.americanscientist.org/article/gausss-day-of-reckoning>

1. **Validación de Resultados:** Permiten validar si un resultado observado en los datos es debido a una variación aleatoria o a un efecto real.
2. **Control de Errores:** Las pruebas de hipótesis controlan las probabilidades de cometer errores de tipo I (falsos positivos) y tipo II (falsos negativos).

Uso en el Modelo de Regresión Lineal por Mínimos Cuadrados

En un modelo de regresión lineal, las pruebas de hipótesis se utilizan para probar supuestos clave:

1. **Linealidad:** La relación entre las variables independientes y la variable dependiente es lineal. Esto se puede probar visualmente o mediante pruebas estadísticas.
2. **Independencia de los Residuos:** Los residuos (diferencias entre los valores observados y los predichos) deben ser independientes. Esto a menudo se verifica con la prueba de Durbin-Watson.
3. **Homocedasticidad:** Los residuos deben tener varianzas constantes. Esto se puede verificar con pruebas como la de Breusch-Pagan.
4. **Normalidad de los Residuos:** En muchos casos, se asume que los residuos siguen una distribución normal, especialmente importante para pequeñas muestras. Se pueden usar pruebas como la de Shapiro-Wilk para verificar esto.
5. **Ausencia de Multicolinealidad:** Se debe asegurar que las variables independientes no estén altamente correlacionadas entre sí. Esto se puede probar con el factor de inflación de la varianza (VIF).

Resumen del capítulo

En este capítulo construimos, pieza por pieza, la herramienta más importante de la econometría: la [regresión lineal](#).

Lo que hicimos fue empezar con una idea simple: trazar una línea recta que resuma la relación entre dos variables, como el gasto en publicidad y las ventas. Después, nos metimos a la sala de máquinas para ver cómo funciona por dentro. Vimos que el método de [mínimos cuadrados ordinarios \(OLS\)](#) encuentra la mejor línea posible minimizando los errores al cuadrado, y hasta nos atrevimos a deducir la fórmula matemática usando álgebra de matrices. Lo más importante: aprendimos los cinco supuestos del Teorema de Gauss-Márkov, que son las reglas del juego que garantizan que nuestras estimaciones sean las mejores posibles (o sea, [BLUE](#)). Finalmente, aprendimos a descifrar la tabla de resultados que nos da Python, para convertirla de un jeroglífico arcano a una historia clara sobre nuestros datos.

Esto es importante porque la regresión lineal es el lenguaje universal del análisis de datos. Es el punto de partida de casi todos los modelos más complejos que existen. Entender sus fundamentos —en especial los supuestos— es lo que diferencia a alguien que simplemente “corre modelos” de alguien que realmente *entiende* lo que está haciendo. Es tu detector de mentiras integrado. Sin él, es fácil engañarse a uno mismo y a los demás con resultados que parecen correctos pero que están fundamentalmente rotos.

¿Cómo te ayuda esto? Ahora tienes un flujo de trabajo completo. Puedes tomar un conjunto de datos, plantear una hipótesis, correr un modelo de regresión, y lo más crucial, diagnosticar si puedes confiar en sus resultados.

Ya puedes interpretar un coeficiente y decir con confianza: “manteniendo todo lo demás constante (el famoso *ceteris paribus* en la vida real), un aumento de X unidades en esta variable se asocia con un cambio de β unidades en el resultado”.

Este capítulo te da las herramientas para empezar a tener conversaciones serias y basadas en evidencia, y para hacer preguntas inteligentes cuando otros te presenten sus análisis.

Poniendo a prueba la regresión: manos al código y a la mente

Es hora de aplicar lo aprendido. Estos ejercicios combinan la interpretación conceptual con la práctica de escribir código. ¡Abre tu notebook de Colab!

1. **Interpretando coeficientes:** Viendo la tabla final de resultados de `statsmodels` en el capítulo:

- Explica en una sola frase, como si hablaras con un colega de marketing, qué significa el coeficiente de `Radio` (0.1070).
- El coeficiente de `Newspaper` es casi cero y su *p-value* es altísimo (0.954). ¿Qué conclusión práctica sacarías sobre invertir en publicidad en periódicos, basándote *únicamente* en este modelo?
- El intercepto (`const`) es 4.6251. ¿Qué significaría este número en el mundo real del problema de la publicidad? ¿Tiene sentido práctico?

2. **OLS desde las entrañas (Código):** El capítulo te retó a expandir el cálculo manual de $\hat{\beta}$ para incluir todas las variables. Acepta el reto: usando solo `numpy`, construye la matriz X (con una columna de unos, 'TV', 'Radio' y 'Newspaper') y el vector y ('Sales'). Calcula el vector de coeficientes $\hat{\beta}$ usando la fórmula $(X'X)^{-1}X'y$. Verifica que los coeficientes que obtuviste son idénticos a los de la tabla de `statsmodels`.

3. **Creando multicolinealidad (Conceptual y Código):** Imagina que en el dataset de publicidad, además de la columna `Newspaper` (gasto en dólares), creas una nueva columna llamada `Newspaper_cents` que es simplemente el gasto en periódicos multiplicado por 100. ¿Qué crees que pasaría con el supuesto de "No multicolinealidad" si intentas correr una regresión con ambas variables? ¿Por qué la matriz $(X'X)$ no tendría inversa?

4. **El diagnóstico visual:** Uno de los gráficos más útiles para un primer diagnóstico es el de los residuales contra los valores predichos.
 - ¿Qué dos supuestos de Gauss-Markov puedes empezar a evaluar con este gráfico?
 - ¿Cómo se vería un gráfico "saludable"(que cumple los supuestos)? ¿Y cómo se vería uno "enfermo"(que los viola)?
 - Genera este gráfico para el modelo final del capítulo.
5. **Corriendo un modelo alternativo (Código):** Carga el dataset `advertising.csv`. Corre una nueva regresión para predecir las ventas (`Sales`) pero esta vez usando *únicamente* `Radio` y `Newspaper` como variables independientes. Muestra la tabla de resultados completa de `statsmodels`.
6. **Comparando modelos:** Observa la tabla de resultados que generaste en el ejercicio anterior y compárala con la del modelo original del capítulo (que incluía 'TV').
 - ¿Qué modelo es mejor para explicar las ventas? Fíjate en la `Adj. R-squared`.
 - ¿Qué pasó con el coeficiente y el p-value de `Newspaper`? ¿Cambió tu interpretación sobre su efectividad? ¿Por qué crees que pasó esto? (Pista: Piensa en la correlación entre las variables).
7. **El supuesto más importante (Conceptual):** En el capítulo insisto en que la exogeneidad ($E(\epsilon|\mathbf{X}) = 0$) es "el supuesto más crítico para la inferencia causal". Conecta esta idea con lo que aprendiste en el capítulo de Resultados Potenciales. ¿Qué problema fundamental causa una variable omitida que está correlacionada tanto con tu X como con tu Y ?
8. **Leyendo los intervalos de confianza:** En la tabla de resultados, el intervalo de confianza del 95 % para `TV` es [0.052, 0.057]. Explica qué significa este rango como si se lo estuvieras presentando a tu jefe, quien no sabe de estadística pero sí de negocios. ¿Por qué es más informativo que solo darle el coeficiente puntual de 0.0544?
9. **Un VIF problemático (Código):** En el DataFrame de publicidad, crea una nueva variable llamada `Radio_y_Diario` que sea la suma de `Radio` y `Newspaper`. Ahora, calcula el VIF para un modelo que intenta predecir las ventas usando `Radio`, `Newspaper` y tu nueva variable `Radio_y_Diario`. ¿Qué le pasó a los VIFs? Explica por qué.
10. **Reto - ¿Relación no lineal?:**

Usando el dataset de publicidad, crea una nueva variable que sea `TV_cuadrado = data['TV']**2`.

Corre una nueva regresión para predecir las ventas usando `TV` y `TV_cuadrado`. Observa los coeficientes y sus p-values. ¿Qué te sugiere esto sobre la relación entre el gasto en TV y las ventas? ¿Es estrictamente lineal?

Cómo se usan los modelos de series de tiempo para proyectar las ventas en una empresa

People assume that time is a strict progression of cause to effect, but actually, from a non-linear, non-subjective viewpoint, it's more like a big ball of wibbly-wobbly, timey-wimey... stuff

– The Doctor

It's tough to make predictions, especially about the future.

– Yogi Berra

Introducción

Si lo que buscas es una herramienta para predecir el comportamiento de una variable en el tiempo, esto es lo más cercano que encontrarás.

La frase “series de tiempo” significa dos cosas:

1. Modelos que analizan la naturaleza de fluctuaciones temporales de una variable (o varias).
2. Registros regulares de datos de una variable en el tiempo (p. ej. registros mensuales del Producto Interno Bruto).

El supuesto clave de las series de tiempo es que podemos extraer información sobre el comportamiento de nuestros datos sólo con el registro de su comportamiento en el tiempo. Más aún, los errores que *no son parte de nuestros datos* también tienen patrones que se pueden capturar y aprovechar para hacer inferencia e incluso predicciones.

Para qué se usan las series de tiempo

Los modelos de **serie de tiempo** son muy populares en economía, pero también te las puedes encontrar en la biología, la física e incluso

para detectar brotes de Dengue⁷⁹.

Pero uno de los usos más populares de las series de tiempo son los negocios.

Los usos en negocios de las series de tiempo incluyen:

- Proyección de ventas.
- Predicción de demanda.
- Finanzas.
- Energía.
- Mercados financieros.
- Optimización de inventarios.

Lo que tienen en común estos usos es que conocer los valores en el futuro es crítico.

En este capítulo veremos los modelos más relevantes para el análisis de series de tiempo. Veremos también cómo comprobar su capacidad de predecir causalidad. Finalmente, veremos aplicaciones específicas de los negocios.

Un ejemplo de una Serie de Tiempo

Definimos una serie de tiempo como una secuencia $\{X_t\}$ de observaciones de una variable aleatoria en el tiempo.

Las observaciones X_t tienen a t como subíndice, que indica el momento en el tiempo de la observación. Esto nos permite crear modelos sobre el comportamiento de la variable en el tiempo. Un ejemplo de una serie de tiempo es una [caminata aleatoria](#):

$$X_t = X_{t-1} + \varepsilon_t,$$

donde ε_t es un [ruido blanco](#).

El siguiente código nos muestra cómo hacer una simulación de una caminata aleatoria en Python y un gráfico para mostrarlo⁸⁰.

```
import numpy as np
import matplotlib.pyplot as plt

# Fijar la semilla del generador de números aleatorios para
# → reproducibilidad
np.random.seed(42)
```

⁷⁹ Derek A. T. Cummings, Rafael A. Irizarry, Norden E. Huang, Timothy P. Endy, Ananda Nisalak, Kumnuan Ungchusak, and Donald S. Burke. Travelling waves in the occurrence of dengue haemorrhagic fever in thailand. *Nature*, 427(6972):344–347, 2004. DOI: 10.1038/nature02225. URL <https://pubmed.ncbi.nlm.nih.gov/14737166/>; and Norden Huang and Nii O. Attoh-Okine, editors. *The Hilbert-Huang Transform in Engineering*. Taylor & Francis Group, USA, 2005. ISBN 978-0-8493-3422-1

⁸⁰ “Plantamos” una semilla para que sean los mismos números aleatorios y obtengas el mismo resultado que yo exactamente.

```

# Parámetros
N = 100 # Número de pasos en el tiempo
sigma = 1 # Desviación estándar del ruido
X_0 = 0 # Valor inicial

# Inicializar el arreglo para X_t
X_t = np.zeros(N)
X_t[0] = X_0

# Generar el proceso: es una secuencia de números
# aleatorios
for t in range(1, N):
    epsilon = np.random.normal(0, sigma) # Generar el
    # ruido
    X_t[t] = X_t[t-1] + epsilon # Actualizar el valor de
    # X_t

# Graficar el proceso
plt.figure(figsize=(10, 6))
plt.plot(X_t, label='$X_t$')
plt.xlabel('Paso del tiempo')
plt.ylabel('Valor')
plt.title('Simulación de $X_t = X_{t-1} + \varepsilon_t$')
plt.legend()
plt.grid(True)
plt.show()

```

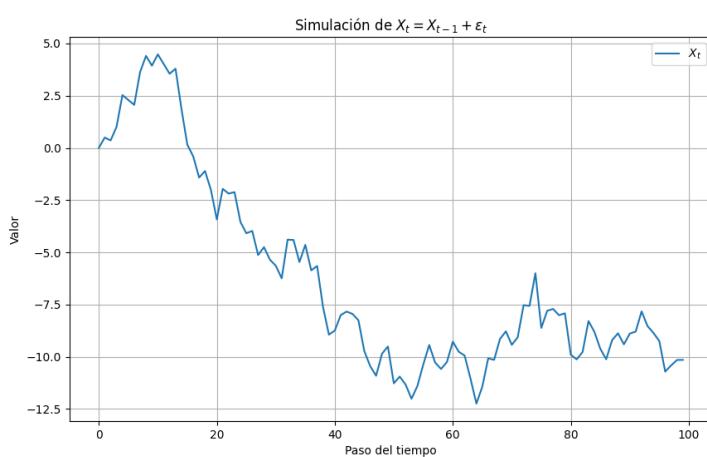


Figura 20: Caminata aleatoria. Cualquier parecido con una serie de tiempo real, es mera coincidencia.

Veamos ahora una de las propiedades más importantes de las series de tiempo: [estacionariedad](#).

Estacionariedad

En un sentido intuitivo, una serie de tiempo es estacionaria cuando las propiedades estadísticas del **proceso que genera la serie** no cambian en el tiempo.

Hay mucha filosofía detrás de esa definición. Para empezar, tenemos el supuesto de que existe un proceso que genera el comportamiento de la serie de tiempo. Luego hay que notar que esta definición no implica la ausencia de cambios en los valores de la serie, sólo que la forma en que los datos cambian permanece constante.

Por eso es un concepto clave: sin estacionariedad, los modelos no funcionan.

En el apéndice de este capítulo te dejo las definiciones formales de estacionariedad y sus ideas clave. En resumen:

- Podemos detectar si un modelo es estacionario usando la prueba Dickey-Fuller: si el p-value de nuestra prueba es inferior a 0.05, entonces nuestra serie es estacionaria⁸¹.
- Si nuestra serie no es estacionaria, podemos diferenciarla. La serie se diferencia restando a cada elemento su rezago $X_t - X_{t-1}$. Una **diferencia de la serie** es suficiente para recobrar estacionariedad en una serie de tiempo.

El siguiente código hace una simulación de la serie de tiempo $y_t = \beta_0 + \beta_1 t + \phi y_{t-1} + \epsilon_t$, así como su diferencia. Se muestran los gráficos correspondientes y los resultados de la prueba Dickey-Fuller, hecha con la paquetería `statsmodels`.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import adfuller

# Parámetros del modelo
np.random.seed(42) # Para reproducibilidad
n = 100 # Número de observaciones
beta0 = 0.5 # Intercepto
beta1 = 0.01 # Tendencia
phi = 0.8 # Coeficiente autoregresivo
epsilon = np.random.normal(0, 1, n) # Términos de error

# Generar la serie de tiempo
t = np.arange(n) # Tiempo
y = np.empty(n)
```

⁸¹ En realidad, la prueba Dickey-Fuller está diseñada para detectar la presencia de **raíces unitarias**, que es la causa fundamental de la no-estacionariedad **estocástica** (aleatoria). La tendencia lineal o los cambios estructurales también son casos de no-estacionariedad, pero esta prueba no los detecta.

```

y[0] = beta0 + betal + epsilon[0] # Inicializar la primera observación

for i in range(1, n):
    y[i] = beta0 + betal * t[i] + phi * y[i-1] + epsilon[i]

# Diferenciar la serie para conseguir estacionariedad
y_diff = np.diff(y)

# Prueba de Dickey-Fuller para la serie original y diferenciada
adf_result_original = adfuller(y)
adf_result_diff = adfuller(y_diff)

# Crear los gráficos
fig, axs = plt.subplots(1, 2, figsize=(14, 6))

# Serie original
axs[0].plot(t, y, label='Serie Original')
axs[0].set_title('Serie de Tiempo No Estacionaria', fontsize=14)
axs[0].set_xlabel('Tiempo', fontsize=12)
axs[0].set_ylabel('Valor', fontsize=12)
axs[0].legend()

# Serie diferenciada
axs[1].plot(t[1:], y_diff, label='Serie Diferenciada', color='orange')
axs[1].set_title('Serie de Tiempo Diferenciada', fontsize=14)
axs[1].set_xlabel('Tiempo', fontsize=12)
axs[1].set_ylabel('Valor', fontsize=12)
axs[1].legend()

plt.tight_layout()
plt.show()

(adf_result_original[0], adf_result_original[1], adf_result_diff[0], adf_result_diff[1])

```

(-2.297082991922703,
 0.1729104907141269,
 -6.61625846070549,
 6.19646154282778e-09)

Algunos puntos relevantes.

- Normalmente cuando observas una serie con una tendencia como la que se muestra en el primer modelo, podemos esperar que la serie no sea estacionaria.
- Por lo general una diferencia debería ser suficiente para lograr estacionariedad. Es posible hacer dos o más diferencias al modelo, pero hay que ser cautelosos, pues podrías hacer la serie extremadamente volátil e introducir patrones artificiales en los datos.

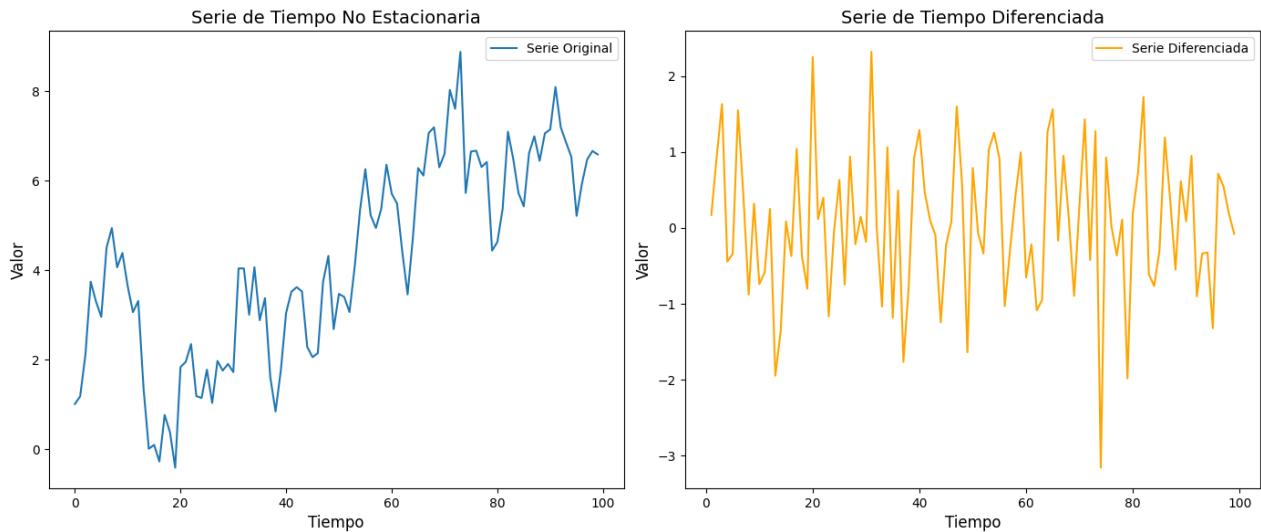


Figura 21: Diferencia visual entre una serie no-estacionaria y la misma serie con una diferencia aplicada. Nota que la serie se vuelve estacionaria.

- Si con una o dos diferencias no logras generar estacionariedad en la serie de tiempo, vale la pena revisar si no existen otros aspectos como cambios estructurales o estacionalidad en la serie. Estos aspectos se ven más adelante con modelos específicos.

El modelo ARIMA es un ejemplo de un modelo que funciona cuando la serie es estacionaria. Es también un modelo popular por su flexibilidad de uso y lo poderoso de sus resultados.

Veamos más a fondo.

El modelo ARIMA

El modelo ARIMA es como tener una bola de cristal que proyecta el futuro con base en el comportamiento en el pasado y nada más.

Una idea muy intuitiva es que cuando vemos un gráfico de líneas, podemos simplemente seguir dibujando a donde pensamos que seguirá la tendencia. Después de todo, tiene más sentido pensar que la línea seguirá una misma tendencia a que el siguiente número será totalmente aleatorio. Esa es la idea básica detrás del modelo ARIMA.

El modelo ARIMA asume una relación **lineal** entre las variables y sus rezagos.

Observa el siguiente modelo:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t \quad (10)$$

Nota que es básicamente un modelo de **regresión lineal**, con la única diferencia de que la variable independiente y la dependiente son la misma, pero en diferentes observaciones en el tiempo. La historia que este modelo cuenta es que el valor de X en el periodo t depende de su valor en el periodo $t - 1$, con una ligera variación aleatoria⁸².

Este es un modelo AR(1).

Modelos AR(p)

El predictor más lógico del valor de una variable en el tiempo son sus rezagos.

Dicho de otra manera: la mejor manera de saber cuáles son las ventas del próximo año es observando las ventas de este año. Es un modelo sencillo, pero muy poderoso. El modelo AR(p) asume que el valor de y_t depende **linealmente** de los primeros p rezagos de la serie.

Observa la siguiente simulación de un modelo AR(3).

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Establecer la semilla para reproducibilidad
np.random.seed(42)

# Número de observaciones
n = 1000

# Coeficientes del proceso AR(3)
phi = np.array([0.5, -0.2, 0.1])

# Término de ruido
sigma = 1
epsilon = np.random.normal(loc=0, scale=sigma, size=n)

# Inicializando la serie temporal
y = np.zeros(n)

# Generando el proceso AR(3)
for t in range(3, n):
    y[t] = phi[0] * y[t-1] + phi[1] * y[t-2] + phi[2] * y[t-3] + epsilon[t]

# Creando un índice de series temporales
dates = pd.date_range(start='2024-01-01', periods=n)
```

⁸² Naturalmente, hay muchas excepciones. Los *shocks* económicos y las situaciones extraordinarias suelen ser difíciles de predecir con este tipo de modelos. Y tiene sentido: si alguien pudiera predecir cuándo es el siguiente *shock* en el mercado financiero, ese alguien podría hacer mucho dinero con ese conocimiento. Pero el mismo comportamiento que genera una oportunidad de arbitraje, tiende a eliminar la posibilidad de explotarla.

```
# Convirtiendo a una serie de pandas para graficar
y_series = pd.Series(y, index=dates)

# Graficando
plt.figure(figsize=(14, 6))
plt.plot(y_series)
plt.title('Proceso Simulado AR(3)')
plt.xlabel('Tiempo')
plt.ylabel('Valor')
plt.show()
```

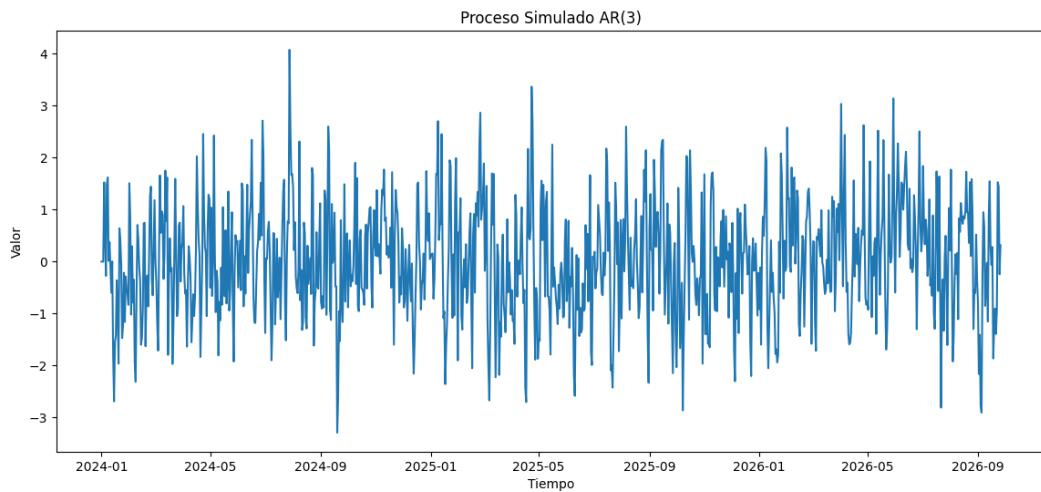


Figura 22: Simulación de un proceso AR(3).

Estamos haciendo trampa.

Estoy creando una simulación de un modelo donde

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \epsilon_t$$

La ventaja es que sabemos que $\phi_1 = 0,5$, $\phi_2 = -0,2$ y $\phi_3 = 0,1$, y podemos hacer pruebas con Python para aprender a hacer modelos AR.

Estos son los pasos:

Paso #1: Verifica si tu serie de tiempo es estacionaria. La estacionariedad es lo que permite que los modelos sean consistentes. Este es el código para hacer la comprobación haciendo una prueba de Dickey-Fuller⁸³:

⁸³ David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431, 1979

```

from statsmodels.tsa.stattools import adfuller

# Realizando el test ADF
adf_result = adfuller(y_series)

# Extraemos cada componente del resultado
test_statistic = adf_result[0]
p_value = adf_result[1]
used_lag = adf_result[2]
n_obs = adf_result[3]
critical_values = adf_result[4]
ic_best = adf_result[5]

# Mostrar los resultados de forma legible
print("===== Prueba de Dickey-Fuller Aumentada =====")
print(f"Estadístico de prueba: {test_statistic:.4f}")
print(f"Valor p: {p_value:.4e}")
print(f"Número de retardos usados: {used_lag}")
print(f"Número de observaciones: {n_obs}")
print("Valores críticos:")
for key, value in critical_values.items():
    print(f" {key}: {value:.4f}")
print(f"Información de criterio (AIC/BIC): {ic_best:.2f}")

```

```

===== Prueba de Dickey-Fuller Aumentada =====
Estadístico de prueba: -15.3109
Valor p: 4.1877e-28
Número de retardos usados: 2
Número de observaciones: 997
Valores críticos:
 1%: -3.4369
 5%: -2.8644
 10%: -2.5683
Información de criterio (AIC/BIC): 2739.69

```

Aprendamos a interpretar esta prueba:

- La prueba se llama Dickey-Fuller Aumentada (ADF). Se usa para identificar la presencia de una **raíz unitaria**. Identificar estacionariedad directamente requeriría que supiéramos el proceso que genera la serie de tiempo. En nuestro caso lo sabemos porque nosotros lo generamos, pero en la vida real eso es justo lo que queremos estimar⁸⁴.
- Una serie de tiempo tiene raíces unitarias si se puede representar por el proceso donde las raíces de la ecuación característica son iguales a uno (o están en el círculo unitario de un espacio complejo).

⁸⁴ Este módulo no genera por su cuenta un reporte organizado y entendible, como el que vimos en la regresión. Pero eso no es un problema. Simplemente agregamos algunas funciones `print` para mostrar los resultados en pantalla.

- Especificación del modelo. La prueba ADF se basa en la estimación del modelo siguiente:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \cdots + \delta_{p-1} \Delta Y_t - p + 1 + \varepsilon_t$$

donde $\Delta Y_t = Y_t - Y_{t-1}$. Es un modelo lineal, igual a los que hemos visto desde el capítulo sobre regresión. Este modelo sirve para identificar si existen o no raíces unitarias. Cada $\delta_i \Delta Y_{t-i}$ representa rezagos de las primeras diferencias en la serie. Estamos también incluyendo una derivada del tiempo βt y el término clave para determinar las raíces unitarias: γY_{t-1} .

- La hipótesis nula de la prueba ADF es que la serie de tiempo tiene raíz unitaria ($\gamma = 0$), lo que significa que no es estacionaria. La hipótesis alternativa es que la serie de tiempo no es estacionaria ($\gamma < 0$).

En otras palabras, como el p-value es muy pequeño ($p < 0,01$), podemos interpretar que nuestra serie no es estacionaria.

Paso #2: Si la serie no es estacionaria, considera transformarla para hacerla estacionaria.

Hay transformaciones que resultan naturales a una serie de tiempo. Por ejemplo, el indicador general de la actividad Económica es evidentemente No-estacionario.

El siguiente código extrae directamente los datos del segundo trimestre desde INEGI y los transforma en un data-frame que se puede usar para evaluarse como serie de tiempo.

```
# Importar las librerías necesarias
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# -----
# Cargar datos desde INEGI
# -----
url = 'https://www.inegi.org.mx/contenidos/programas/igae/2018/tabcularados/ori/IGAE_2.xlsx'
igae_data = pd.read_excel(url, skiprows=5)

# -----
# Limpiar datos
# -----
# Eliminar la primera fila que contiene encabezados no útiles
igae_data_cleaned = igae_data.drop(igae_data.index[0])
igae_data_cleaned.reset_index(drop=True, inplace=True)
```

```

# Obtener la fila correspondiente a "Total" (ya es la primera después de limpiar), ignorando la
# → primera columna
total_series = igae_data_cleaned.iloc[0, 1:].transpose()

# Convertir a valores numéricos y eliminar NaNs
total_series = pd.to_numeric(total_series, errors='coerce').dropna()

# -----
# Crear índice de tiempo
# -----
# Crear fechas mensuales a partir de enero de 1993
dates = pd.date_range(start='1993-01-01', periods=len(total_series), freq='MS')

# Crear DataFrame con índice de fechas
df = pd.DataFrame({'Valor IGAE': total_series.values}, index=dates)

# Convertir explícitamente a tipo float
df['Valor IGAE'] = df['Valor IGAE'].astype(float)

# Confirmar dimensiones (opcional)
print(len(df), len(total_series), len(dates))

```

Sabrás que este código se ejecutó correctamente porque te mostrará las primeras filas de tu base de datos de la IGAE.

Usa este código para mostrarlo en un gráfica.

```

# -----
# Graficar serie original
# -----
plt.figure(figsize=(10, 6))
plt.plot(df.index, df['Valor IGAE'])
plt.title('Serie de Tiempo IGAE')
plt.xlabel('Fecha')
plt.ylabel('Valor IGAE')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

```

Con algo de experiencia vas a aprender a notar cuando una serie no es estacionaria (como es este caso) sólo al ver el gráfico. Pero siempre debes de corroborar tus sospechas haciendo una prueba estadística.

Paso #3: Identifica el orden del rezago p . La siguiente pregunta que tenemos es ¿cuántos rezagos debo usar en mi modelo?

Para resolver este problema usamos la función de autocorrela-

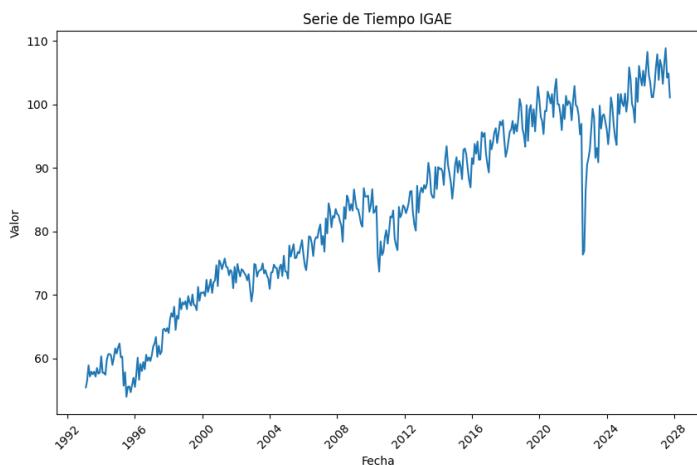


Figura 23: Índice Global de la Actividad Económica (IGAE). El gráfico lo muestra como un gráfico de líneas. En una serie estacionaria, no se debe observar una tendencia clara. Este indicador a simple vista se puede observar que no es estacionario. Fuente: INEGI.

ción (AFC) y la función de autocorrelación parcial (PACF). Este es el gráfico que generan⁸⁵.

Primero hagamos el gráfico de la serie original, para visualizar la diferencia cuando la serie es estacionaria.

```
# -----
# Paso 3: ACF y PACF de la serie original
# -----
y_series = df['Valor IGAE']

fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(12, 8))
plot_acf(y_series, ax=ax1, lags=40)
ax1.set_title('Función de Autocorrelación (ACF) - Serie Original')
plot_pacf(y_series, ax=ax2, lags=40)
ax2.set_title('Función de Autocorrelación Parcial (PACF) - Serie Original')
plt.tight_layout()
plt.show()
```

El truco para identificar el orden del rezago en un modelo AR(p) es mirar el PACF. El punto en el que la autocorrelación cae por debajo del nivel de significancia (no está cubierto por el color en el gráfico), es el nivel de rezago que debemos usar.

En este caso el gráfico nos muestra una caída después del tercer rezago, que es lo que esperamos con el modelo que hemos diseñado.

⁸⁵ Existe una función de `auto_arima` en el módulo `pmdarima`, que encuentra de manera automática el [orden de rezago](#) de manera automática. Dado que es una tarea muy orientada a reglas estructuradas, generalmente está bien hacerlo de esta manera. Lo que estamos viendo en esta sección está más orientado a que tengas un entendimiento general de lo que implica el rezago, por eso lo hacemos con este método visual.

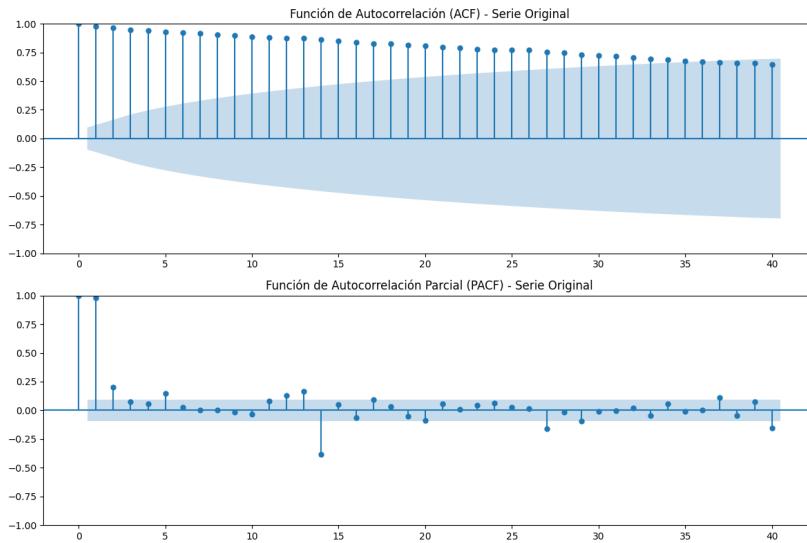


Figura 24: Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie original (sin diferenciación).

```
# -----
# Paso 6: Serie diferenciada (opcional para verificar
#         estacionariedad)
# -----
df_diff = df.diff().dropna()
y_diff = df_diff['Valor IGAE']

fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(12, 8))
plot_acf(y_diff, ax=ax1, lags=40)
ax1.set_title('ACF - Serie Diferenciada')
plot_pacf(y_diff, ax=ax2, lags=40)
ax2.set_title('PACF - Serie Diferenciada')
plt.tight_layout()
plt.show()
```

Paso #4: Ejecutar el modelo y mostrar los resultados

Finalmente estamos listos para ejecutar nuestro modelo AR(3). Usa este código:

```
from statsmodels.tsa.ar_model import AutoReg

# Ajustando un modelo AR(3)
model = AutoReg(y_series, lags=3)
model_fitted = model.fit()

# Mostrando los resultados
model_results = model_fitted.summary()

model_results
```

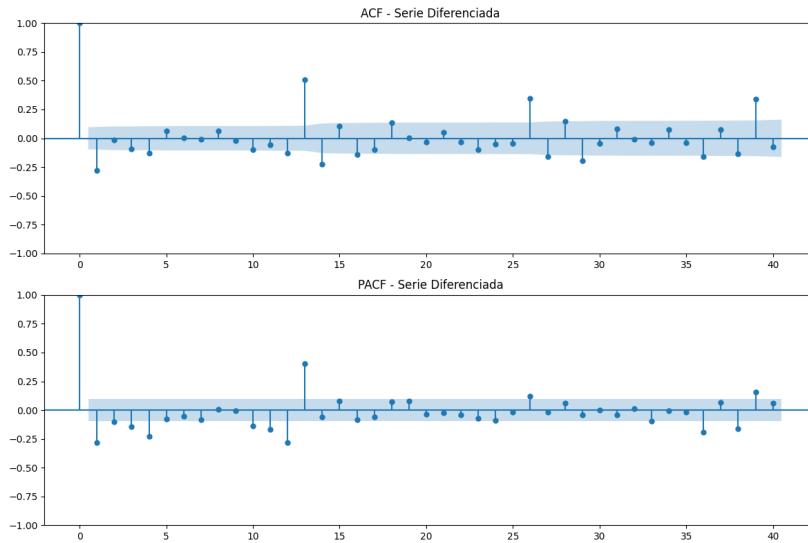


Figura 25: Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie diferenciada. Nota que en la serie diferenciada, el orden de rezago a elegir es menor al de la serie original.

Esta es la tabla con los resultados de la regresión de un [modelo autorregresivo AR\(\$p\$ \)](#) con $p = 3$ en estimación por máxima verosimilitud.

	Coeficiente	Error estándar	IC 95 %
Intercepto	0.019	0.031	[-0.042, 0.080]
y_{t-1} (retardo 1)	0.488***	0.032	[0.426, 0.550]
y_{t-2} (retardo 2)	-0.184***	0.035	[-0.252, -0.116]
y_{t-3} (retardo 3)	0.085**	0.032	[0.023, 0.147]
N (observaciones)	1000		
Log-verosimilitud	-1394.23		
AIC	2798.46		
BIC	2822.98		
Desv. estándar (innovaciones)	0.980		

Notas: Estimaciones mediante máxima verosimilitud condicional.

Niveles de significancia: * $p < 0,1$, ** $p < 0,05$, *** $p < 0,01$

El modelo es estable, ya que todas las raíces del polinomio autoregresivo están fuera del círculo unitario:

Raíces: $2,10, 2,37 \pm 2,37i$ (módulos > 1).

Lo que podemos observar en los resultados de este modelo:

- Los coeficientes son efectivamente cercanos a los que hicimos en la simulación, con un valor significativo ($p < 0.05$)
- La sección donde dice Real e Imaginario nos ayudan a confirmar que nuestro modelo es estable (i.e. todas las unidades de la ecuación característica están fuera del círculo unitario).

Paso #5: Comparar AIC y BIC para encontrar el mejor modelo

El criterio de Información de Akaike (**AIC**) y el Criterio de información Bayesiano (**BIC**) son medidas que nos ayudan a identificar el “mejor modelo” en términos de la información que podemos obtener.

La idea es esta: cuando incluimos más variables a nuestro modelo, podemos tener mayor precisión, pero nos arriesgamos a un sobreajuste. Veamos que sucede si comparamos AR(3) con los modelos AR(2) y AR(4).

La regla de oro es el que el modelo con AIC y BIC más bajo, gana.

Veamos una comparativa con modelos AR(2) y AR(4)

```
# Ajustando modelos AR(2) y AR(4) y comparando AIC y BIC
model_ar2 = AutoReg(y_series, lags=2).fit()
model_ar4 = AutoReg(y_series, lags=4).fit()

# Extrayendo AIC y BIC para cada modelo
aic_bic_comparison = pd.DataFrame({
    'Model': ['AR(2)', 'AR(3)', 'AR(4)'],
    'AIC': [model_ar2.aic, model_fitted.aic, model_ar4.aic],
    'BIC': [model_ar2.bic, model_fitted.bic, model_ar4.bic]
})

aic_bic_comparison
```

	Model	AIC	BIC
0	AR(2)	2805.485454	2825.108467
1	AR(3)	2798.455225	2822.978979
2	AR(4)	2794.998432	2824.420916

Curioso.

El modelo AR(4) tiene un nivel menor de AIC, con un BIC ligeramente mayor. Eso quiere decir que, si no supiéramos el modelo real, bien podríamos considerar el modelo AR(4) como viable.

Generalmente, el AIC se enfoca más en la calidad del ajuste, mientras que el BIC añade una penalización más fuerte por la cantidad de parámetros, favoreciendo modelos más simples.

Paso #6: Generar proyecciones

Finalmente, podemos crear proyecciones a partir de nuestros modelos. En el siguiente código, dividimos la base de datos en datos de entrenamiento (`train`) y de prueba (`test`). La idea es comprobar que nuestras proyecciones ayudan a predecir el valor que estamos buscando.

```

from sklearn.metrics import mean_absolute_error

# Dividiendo los datos en entrenamiento y prueba
train_data = y_series[:int(0.9 * len(y_series))]
test_data = y_series[int(0.9 * len(y_series)):]

# Ajustando el modelo AR(3) al conjunto de entrenamiento
model_train = AutoReg(train_data, lags=3).fit()

# Haciendo predicciones en el conjunto de prueba
predictions = model_train.predict(start=len(train_data), end=len(train_data) + len(test_data) - 1, dynamic=False)

# Calculando el Error Absoluto Medio (MAE)
mae = mean_absolute_error(test_data, predictions)

mae

```

Un gráfico nos puede ayudar a tener mayor claridad. Nota que la proyección en este caso es sólo una línea horizontal en el valor medio de los datos. Lo que esto nos indica es que el modelo predice el siguiente valor de nuestro indicador igual al último dato.

En otras palabras, con la información que tenemos y esta especificación del modelo, es tan probable que suba a que baje⁸⁶.

```

# Creando un gráfico para comparar los valores reales y las
# → predicciones
plt.figure(figsize=(14, 7))
plt.plot(train_data, label='Training Data')
plt.plot(test_data, label='Actual Value')
plt.plot(predictions, label='Forecast', linestyle='--')
plt.title('AR(3) Forecast vs Actuals')
plt.xlabel('Date')
plt.ylabel('Value')
plt.legend()
plt.show()

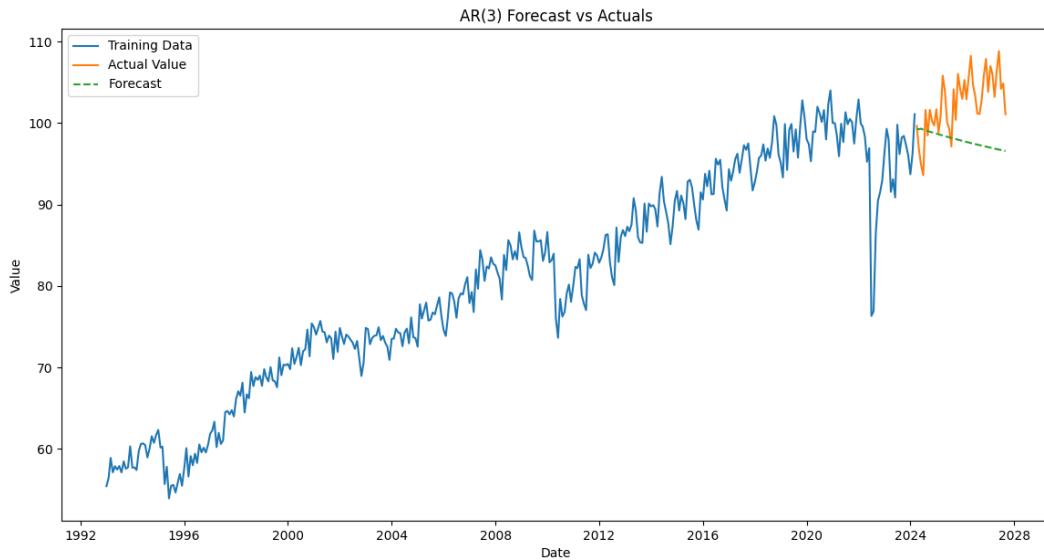
```

⁸⁶ Estoy consciente de que esto puede parecer poco alentador. Pero lo que estamos viendo son sólo las bases, y esta es la versión más sencilla. Lo interesante de este modelo es que se puede incluir mucha más información y refinrar el modelo, pero los principios de aplicación seguirán siendo los mismos. Por eso vale mucho la pena pasar tiempo trabajando lo básico.

Medias Móviles: La MA de ARIMA

El procedimiento que vimos antes se puede aplicar igual con un modelo más complejo.

En ocasiones, nos enfrentamos a dos problemas, que a primera vista parecen no estar relacionados entre sí:



- Los rezagos de la variable no son suficientes para explicar su comportamiento.
- Los rezagos de los errores muestran una tendencia.

Lo curioso es que ambos problemas los podemos solucionar incluyendo dichos rezagos al modelo. Considera el siguiente modelos ARMA(p, q), que contiene un rezago ($p = 1$) y tres rezagos del error ($q = 3$).

Figura 26: Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de la serie diferenciada. Nota que en la serie diferenciada, el orden de rezago a elegir es menor al de la serie original.

```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA

# Estableciendo la semilla para reproducibilidad
np.random.seed(42)

# Número de observaciones
n = 1000

# Generando un término de ruido
epsilon = np.random.normal(loc=0, scale=1, size=n)

# Inicializando la serie temporal
y = np.zeros(n)

# Parámetros para el proceso ARMA(1,3)

```

```

phi = 0.5 # Coeficiente AR
theta = [0.1, -0.2, 0.3] # Coeficientes MA

# Generando el proceso ARMA(1,3)
for t in range(4, n):
    y[t] = phi * y[t-1] + epsilon[t] + theta[0] * epsilon[t-1] + theta[1] * epsilon[t-2] +
           theta[2] * epsilon[t-3]

# Creando un índice de series temporales
dates = pd.date_range(start='2024-01-01', periods=n)

# Convirtiendo a una serie de pandas para graficar y modelar
y_series_arma = pd.Series(y, index=dates)

# Ajustando un modelo ARMA(1,3)
arma_model = ARIMA(y_series_arma, order=(1, 0, 3))
arma_result = arma_model.fit()

# Creando predicciones con el modelo ARMA(1,3) para los últimos 100 puntos de datos
arma_predictions = arma_result.predict(start=n-100, end=n-1)

# Creando un gráfico para comparar los valores reales y las predicciones de ARMA(1,3)
plt.figure(figsize=(14, 7))
plt.plot(y_series_arma[n-100:], label='Actual Values')
plt.plot(arma_predictions, label='ARMA(1,3) Predictions', linestyle='--')
plt.title('ARMA(1,3) Simulation')
plt.xlabel('Date')
plt.ylabel('Value')
plt.legend()
plt.show()

arma_result.summary()

```

Este modelo se ve más interesante.

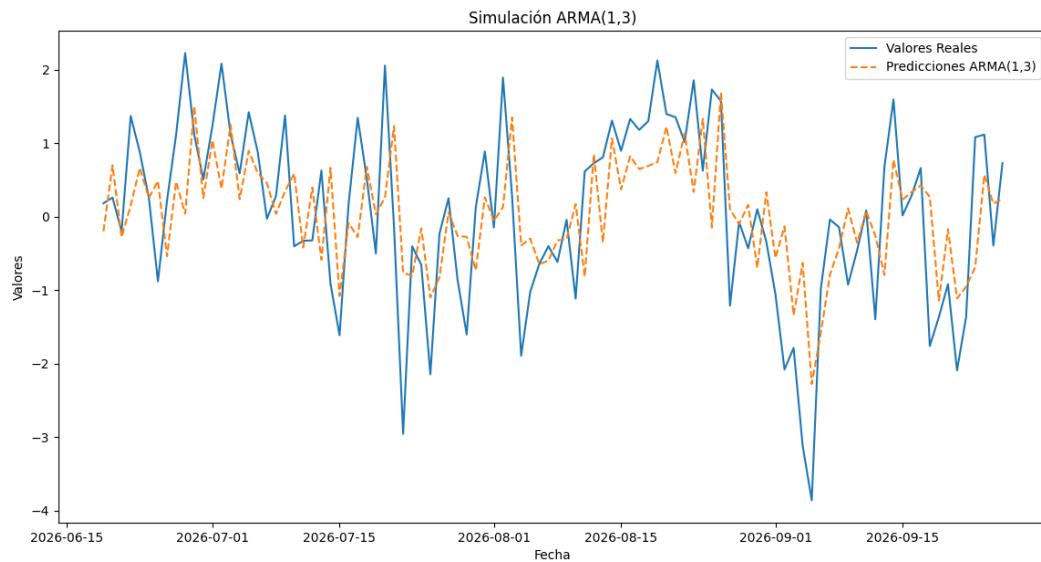
De manera formal, escribimos el modelo como

$$y_t = \sum_i^p \phi_i y_{t-i} + \sum_j^q \theta_j \varepsilon_{t-j} + \varepsilon_t, \quad (11)$$

donde p y q son el nivel de rezago para la parte autorregresiva (AR) y las medias móviles (MA), respectivamente⁸⁷. Nota que el **modelo de medias móviles MA(q)** usa los rezagos del error como posibles predictores de la variable de interés.

Los pasos que seguimos en la sección anterior aplican también aquí. De igual manera, si nos enfrentamos a un **modelo ARMA(p, q)**, debemos verificar que sea estacionario, identificar el orden del proceso y hacer nuestras proyecciones.

⁸⁷ En este caso, $p = 1$ y $q = 1$, que transforman la ecuación (11) en $y_t = \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t$



Una diferencia clave son las funciones **ACF** y **PACF**. Para detectar el orden $MA(q)$ revisamos dónde la función ACF corta de manera abrupta.

Veamos nuevamente el gráfico

```
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Plotting ACF and PACF for the ARMA(1,3) model
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(12, 8))

# ACF
plot_acf(y_series_arma, ax=ax1, lags=40)
ax1.set_title('Autocorrelation Function (ACF) for ARMA(1,3)')

# PACF
plot_pacf(y_series_arma, ax=ax2, lags=40)
ax2.set_title('Partial Autocorrelation Function (PACF) for ARMA(1,3)')

plt.tight_layout()
plt.show()
```

Lo que podemos observar:

En el gráfico ACF parece haber un rechazo significativo en el primer retardo y luego una disminución exponencial o una disminución sinusoidal amortiguada en los retardos subsiguientes. Esto es típico de un proceso AR(1) o un proceso ARMA con un componente AR(1).

Figura 27: Función de Autocorrelación (ACF) y Función de Autocorrelación Parcial (PACF) de la serie diferenciada. Nota que en la serie diferenciada, el orden de rezago a elegir es menor al de la serie original.

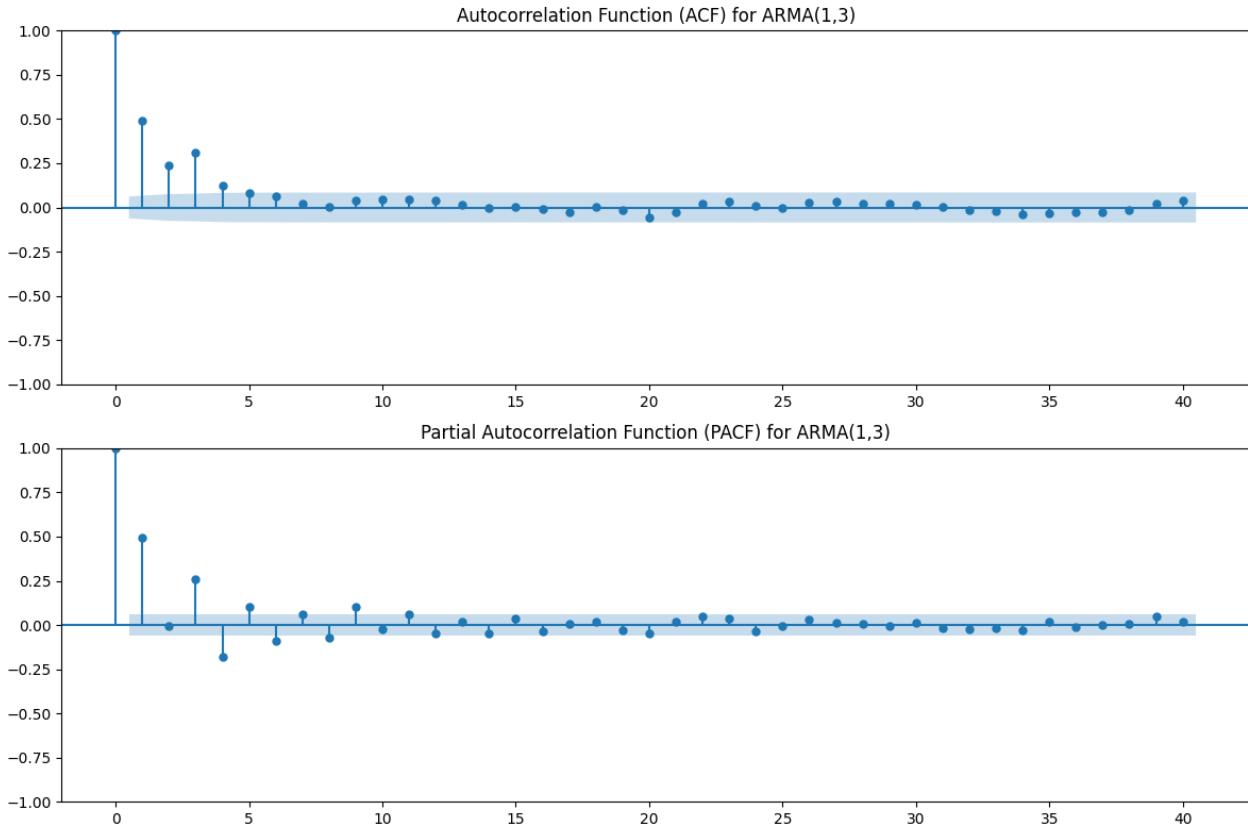


Figura 28: Función de Autocorrelación (AFC) y Función de Autocorrelación Parcial (PAFC) de nuestra simulación de un proceso ARMA(1,3).

El hecho de que el gráfico ACF no se corte después de un número determinado de retardos sugiere la presencia de una componente AR.

El gráfico PACF muestra un rechazo significativo en el primer retardo y rechazos significativos en los retardos tercero y cuarto, lo cual es consistente con un proceso MA(3), ya que el PACF de un proceso MA(q) generalmente muestra rechazos significativos para los primeros retardos y luego se corta.

En la práctica la interpretación del ACF y PACF puede ser más arte que ciencia. Se basa en la identificación de “cortes” y rechazos significativos, por lo que una interpretación plausible de estos gráficos sería que estamos mirando un proceso con una componente autoregresiva de orden 1 (AR(1)) y una componente de media móvil de orden 3 (MA(3)).

¿Y la I del ARIMA?

Finalmente, la I en la palabra ARIMA nos indica la *integración del modelo*, que quiere decir el número de diferencias que requieren nuestros datos para ser estacionarias. Por ejemplo, un modelo ARMA(2,1) con una diferencia, sería un **modelo ARIMA(p, d, q)** con $p = 2$, $d = 1$ y $q = 1$.

Como vimos, que los datos sean estacionarios es muy importante, pero es un problema que se soluciona de una forma relativamente fácil con una diferencia.

La primera diferencia se suele representar como:

$$\Delta X_t = X_t - X_{t-1}$$

Pero podemos ir aún más lejos, si es necesario. Podemos aplicar un segundo orden de integración, también conocido como una segunda diferencia.

$$\begin{aligned}\Delta^2 X_t &= \Delta(\Delta X_t) = X_t - X_{t-1} - (X_{t-1} - X_{t-2}) \\ &= X_t - 2X_{t-1} + X_{t-2}\end{aligned}$$

Es decir, le aplicamos una diferencia a la serie ya diferenciada. Y podemos ir aún más allá con una tercera diferencia.

$$\begin{aligned}\Delta^3 X_t &= \Delta(\Delta^2 X_t) \\ &= (X_t - 2X_{t-1} + X_{t-2}) - (X_{t-1} - 2X_{t-2} + X_{t-3}) \\ &= X_t - 3X_{t-1} + 3X_{t-2} - X_{t-3}\end{aligned}$$

Generalmente, si en dos o tres diferencias el problema no se arregla, tienes problemas más serios con tu serie de tiempo que entender, y una diferencia más no los va a arreglar. Es mejor regresar a las bases.

Apéndice

Aquí van los temas que requieren una explicación adicional, pero que hubieran roto el ritmo de la explicación en el capítulo.

Proceso Estocástico

La palabra **estocástico** significa lo mismo que aleatorio.

La diferencia principal es que *estocástico* viene del griego, mientras que *aleatorio* viene del latín. Para definir correctamente lo que es un proceso estocástico necesitamos algunas definiciones adicionales que vienen de la teoría de la probabilidad:

Un **espacio probabilístico** es una tripla $(\Omega, \mathcal{F}, \mathbb{P})$ donde:

- Ω es un conjunto no vacío conocido como el **espacio muestral**.
- \mathcal{F} es una σ -álgebra (se lee *sigma*-álgebra) de subconjuntos de Ω . Es decir, es una familia de subconjuntos cerrados con respecto a la unión contable y complementaria con respecto a Ω .
- \mathbb{P} es una medida de probabilidad definida para todos los miembros de \mathcal{F} .

Ahora definamos las variables aleatorias

Una **variable aleatoria** es una función $x : \Omega \rightarrow \mathbb{R}$ tal que la imagen inversa de cualquier intervalo $(-\infty, a]$ pertenece a \mathcal{F} , es decir, es una función medible.

Con estas definiciones, podemos llegar a nuestra definición de proceso estocástico:

Un **proceso estocástico real** es una familia de variables aleatorias reales $\mathbf{X} = \{x_t(\omega) \mid t \in T\}$ definidas en el mismo espacio probabilístico $(\Omega, \mathcal{F}, \mathbb{P})$. El conjunto T se le llama el **espacio índice** de procesos. Si $T \subset \mathbb{Z}$, entonces es un proceso estocástico discreto. Si T es un intervalo de \mathbb{R} , entonces es un proceso estocástico continuo.

Ruido blanco

Un ruido blanco es un proceso estocástico que no tiene correlación serial con media cero y varianza constante finita.

De manera formal, un proceso $\{w_t\}$ es un **ruido blanco** si:

- Su primer momento es siempre cero, es decir $E[w_t] = 0$.
- Su segundo momento es finito. Es decir, $[E(w_t - \mu)^2] < \infty$.
- El momento cruzado $E[w_s w_t]$ es cero, para $s \neq t$. Es decir, $cov(w_s, w_t) = 0$.

Raíces unitarias

Consideremos un proceso autorregresivo de orden p :

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t$$

Donde ε_t representa un ruido blanco. Podemos reescribir el mismo proceso como

$$(1 - a_1 L - \cdots - a_p L^p) y_t = a_0 + \varepsilon_t$$

donde L^i es el operador de rezago. La parte entre paréntesis de la izquierda se conoce como la ecuación característica de la serie de tiempo. Consideremos la raíz de esta ecuación:

$$m^p - m^{p-1} a_1 - \cdots - a_p = 0$$

Si la raíz de la ecuación es $m = 1$, entonces el proceso estocástico tiene **raíz unitaria**. Esto se suele comprobar con la [prueba Dickey–Fuller \(ADF\)](#).

Resumen del capítulo

En este capítulo aprendimos el arte de predecir el futuro mirando únicamente el pasado. Eso, en esencia, es el análisis de series de tiempo.

Lo que hicimos fue desglosar la herramienta principal para este trabajo, el [modelo ARIMA\(\$p, d, q\$ \)](#). Vimos que este modelo es como una navaja suiza con tres funciones: la parte Autoregresiva (AR), que asume que el futuro se parece al pasado reciente de la variable; la parte de Medias Móviles (MA), que aprende de los errores de predicción pasados; y la parte Integrada (I), que es el truco de diferenciar los datos para hacerlos estables. Descubrimos que la clave para que todo funcione es la [estacionariedad](#) y aprendimos el flujo de trabajo completo: diagnosticarla con la prueba Dickey–Fuller, corregirla con diferencias, y usar los gráficos ACF y PACF como mapas para elegir la estructura correcta de nuestro modelo.

Esto es importante porque el tiempo es una dimensión crucial en casi cualquier problema económico o de negocios. Saber construir un pronóstico de ventas, demanda o inventario no es magia, es una habilidad técnica muy valiosa. Entender este proceso te protege de cometer el error garrafal de aplicar modelos a datos “inestables” (no estacionarios), lo que produce proyecciones sin sentido. Es la base para cualquier análisis de datos que se desarrolle a lo largo del tiempo.

¿Cómo te ayuda esto? Ahora tienes una metodología paso a paso para tomar una secuencia de datos y construir un pronóstico coherente. Puedes responder preguntas como: “basado en el comportamiento histórico, ¿dónde esperamos que estén nuestras ventas el próximo trimestre?”. Sabes cómo verificar si tus datos son aptos para ser modelados y cómo “curarlos” si no lo son. Puedes construir, comparar (usando AIC y BIC) y generar proyecciones con uno de los modelos de pronóstico más utilizados y respetados en la industria.

Viajando en el tiempo: ejercicios de proyección

La teoría es una cosa, pero con las series de tiempo, la intuición se desarrolla con la práctica. Es hora de aplicar el flujo de trabajo que aprendiste.

1. **Estacionariedad a ojo de buen cubero (Conceptual):** Busca en Google Imágenes los gráficos de las siguientes tres series de tiempo: a) el precio diario de Bitcoin (BTC), b) el número de turistas mensuales que llegan a Cancún, y c) la tasa de desempleo trimestral de tu país. Sin hacer ninguna prueba, solo por inspección visual, ¿cuáles parecen estacionarias y cuáles no? Justifica brevemente por qué. (Pista: busca tendencias claras o patrones que se repiten).
2. **El poder de la diferencia (Código):** Usa el DataFrame del IGAE que construimos en el capítulo.
 - Aplica una primera diferencia a la serie usando `.diff()`.
 - Genera un gráfico de la serie ya diferenciada. ¿Luce más “estable” o “plana” que la original?
 - Corre la prueba de Dickey-Fuller Aumentada (ADF) sobre esta nueva serie diferenciada. ¿Cuál es el p-value? ¿Puedes rechazar ahora la hipótesis nula de que hay una raíz unitaria?
3. **Descifrando el oráculo (Conceptual):** Estás analizando los gráficos ACF y PACF de una serie ya estacionaria para decidir el orden de tu modelo.
 - **Caso A:** El gráfico ACF se corta abruptamente después del segundo rezago (lag 2), mientras que el PACF muestra un decaimiento lento. ¿Qué orden (p,q) de modelo ARMA te sugiere esto?
 - **Caso B:** El gráfico PACF se corta abruptamente después del primer rezago (lag 1), mientras que el ACF decae lentamente. ¿Qué orden (p,q) de modelo ARMA te sugiere esto?
4. **Simulando el pasado (Código):** En el capítulo simulamos un proceso AR(3). Ahora te toca a ti. Simula un proceso **AR(2)** con coeficientes $\phi_1 = 0,6$ y $\phi_2 = 0,2$. Genera 500 observaciones. Después, usa el modelo `AutoReg` de `statsmodels` con `lags=2` sobre tu serie simulada. ¿Los coeficientes que estima el modelo se parecen a los que usaste para crearla?
5. **La batalla de los modelos (Código):** Usando la serie del IGAE ya diferenciada (del ejercicio 2), corre tres modelos para ella: un AR(1), un AR(2) y un AR(3). Imprime una tabla o un resumen comparando sus valores de AIC y BIC. Según estos criterios, ¿cuál de los tres modelos parece ser el mejor?
6. **Escuchando el silencio (Conceptual):** Explica con tus propias palabras qué es un proceso de “ruido blanco”. Si los residuales (los errores de predicción) de tu modelo de serie de tiempo se comportan como un ruido blanco, ¿es una buena o una mala señal para tu modelo? ¿Por qué?
7. **Lanzando la bola de cristal (Código):** Toma el “mejor” modelo que elegiste en el ejercicio 5 para la serie del IGAE.
 - Divide los datos: usa el 90 % para entrenar y reserva el último 10 % para probar.
 - Ajusta tu modelo AR(p) solo con los datos de entrenamiento.
 - Genera un pronóstico para el periodo que cubren los datos de prueba.

- Crea un gráfico que muestre los datos de entrenamiento, los datos reales de prueba, y tu pronóstico. ¿Qué tan bien le atinó tu modelo?
8. **El significado de la “I” (Conceptual):** Un colega te dice que está usando un modelo ARIMA(1,2,1) para proyectar el inventario. ¿Qué significa el número 2 en la posición de la “d” (integración)? ¿Qué tuvo que hacerle a sus datos originales? ¿Es algo común?
9. **Pasado vs. Errores Pasados (Conceptual):** ¿Cuál es la diferencia fundamental en la “historia” que cuenta la parte Autoregresiva (AR) y la que cuenta la parte de Medias Móviles (MA) de un modelo ARIMA?
10. **Reto - Tu propio pronóstico (Código):** Busca en Kaggle o en otra fuente un dataset de series de tiempo que te interese (ej: “avocado prices”, “monthly sunspots”, “shampoo sales”). Carga los datos y realiza el flujo de trabajo completo del capítulo:
- Grafica la serie.
 - Prueba la estacionariedad y diferencia si es necesario (anota el orden ‘d’).
 - Usa los gráficos ACF y PACF para proponer un orden ‘p’ y ‘q’.
 - Ajusta tu modelo ARIMA(p,d,q) y muestra el resumen.
 - Genera un pronóstico para los siguientes 6 períodos.

Efectos Fijos

*I've learned a lot.
And one of the things I've learned is life is really unpredictable.
And people can make forecasts, and they can make predictions.
But those predictions and forecasts may not come true
if there's an unforseeable factor involved*
– Taylor Swift en el iHeartRadio Music Awards en el día de π de 2019.

La regresión lineal es un modelo muy importante porque permite establecer controles en nuestros datos. El problema es que depende de un supuesto clave: inconfundibilidad condicional⁸⁸:

$$(Y_0, Y_1) \perp T | X$$

En otras palabras, requiere que todas las variables de confusión sean conocidas y medidas de tal manera que podamos incluirlas en el modelo y hacer que el grupo de tratamiento se comporte como si hubiera sido fruto de una asignación aleatoria⁸⁹. Pero, a pesar de que no siempre tenemos el lujo de que nuestras variables sean observables, siempre podemos agruparlos con características en común.

Ese es el problema que resuelven los modelos de [datos en panel](#).

⁸⁸ Las variables del lado izquierdo son los contrafactuals, la T es el tratamiento y la X son las covariables que usamos como control.

⁸⁹ En otras palabras, para cada observación, el resultado observado se puede expresar como $Y = T \cdot Y^1 + (1 - T) \cdot Y^0$. Esto implica que la Y y la T son dependientes porque el valor promedio de $T \times Y^1$ no equivale al promedio de $(1 - T) \cdot Y^0$

Cómo se ven los datos en panel

Imagina que estamos estudiando el efecto que hay entre el gasto en publicidad y los ingresos que nos genera.

Para ser mas claros, estamos haciendo una campaña para incrementar las ventas de una e-commerce por tres canales de venta. La primera es por anuncios de Google, el segundo con anuncios en Meta (que incluye Facebook e Instagram) y el tercero es por mail marketing.

Comencemos cargando nuestra base de datos en panel.

```
import pandas as pd
import statsmodels.formula.api as sm

df = pd.read_csv("sales-panel.csv")
df.head()
```

Esto es el encabezado de los datos que nos genera el código de arriba.

	Medio	Año	Costo de publicidad	Ventas
0	Google Ads	2020	1.25	3.4
1	Google Ads	2021	2.00	10.0
2	Google Ads	2022	6.00	13.5
3	Google Ads	2023	5.00	8.0
4	Google Ads	2024	6.00	11.0

Algunas observaciones sobre los datos:

- Estamos agrupando los datos por el medio en el que se hace la publicidad (Anuncios de Google, Anuncios de Facebook y una campaña de email marketing). Eso quiere decir que la base de datos va a repetir cada uno de esos medios en la primera columna⁹⁰.
- El supuesto clave en estos datos es que los clientes que obtenemos a partir de medios diferentes son distintos entre sí⁹¹.
- Es un ejemplo simple, pero en realidad un análisis de panel como este podría ser muy útil para analizar campañas distintas que corren en paralelo.

Hagamos un diagrama de dispersión para analizar los datos.

Lo que deseamos conocer es el efecto que hay entre el costo de la publicidad y las ventas de esa campaña particular.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

# Cargar los datos (ajusta la ruta si es necesario)
df = pd.read_csv("sales-panel.csv")

# Asignar un marcador distinto a cada medio
markers = {
    'Google Ads': 'o',      # Círculo
    'Facebook Ads': 's',    # Cuadrado
    'Email Marketing': 'D'  # Rombo
}
```

Cuadro 3: Datos de ventas y costos de anuncios en Google Ads

⁹⁰ Esta forma de presentar los datos es menos visual que si extendiéramos cada uno de los medios en una columna cada uno (*wide format*), pero hace más fácil aplicar los modelos que veremos en adelante.

⁹¹ Este supuesto tiene sentido, simplemente porque se interactúa diferente en diferentes medios. El medio entonces agrupa muchas características no observables de los grupos. Por ejemplo, podríamos imaginar que los clientes de mail marketing ya confían más en nuestro contenido, mientras que los que vienen de publicidad de Meta podrían desconfiar aún en nosotros. La confianza no se puede medir, pero queda implícita en el medio.

```

}

# Estilo blanco con cuadrícula
sns.set(style="whitegrid")

# Crear el gráfico
plt.figure(figsize=(10, 6))

for medio in df['Medio'].unique():
    subset = df[df['Medio'] == medio]
    plt.scatter(
        subset['Ad cost'],
        subset['Sales'],
        s=100,
        label=medio,
        color='black',
        marker=markers[medio]
    )

# Calcular y graficar la línea de regresión general
X = df['Ad cost'].values
Y = df['Sales'].values
model = np.polyfit(X, Y, 1)
predicted = np.polyval(model, X)
plt.plot(X, predicted, color='gray', linewidth=2)

# Personalización del gráfico
plt.title('Diagrama de dispersión del costo de publicidad
           vs ventas')
plt.xlabel('Costos de publicidad')
plt.ylabel('Ventas')
plt.legend()
plt.grid(True)
plt.show()

```

Le puse formas distintas para que notes a simple vista: una regresión simple no es lo que deseamos hacer.

Si hiciéramos una regresión lineal simple tendríamos que nuestro gasto en publicidad no está aumentando las ventas. Al contrario, ¡Las está haciendo caer! Pero al separarlos por medio nos podemos dar cuenta de que no es así: cada una de las campañas de manera individual tiene un efecto positivo claro en las ventas.

Esto se ve más claro si trazamos una línea de regresión para cada uno de los medios en nuestro diagrama de dispersión.

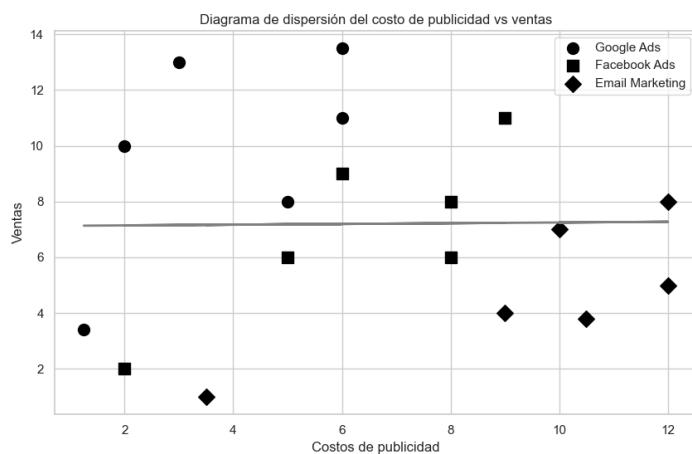


Figura 29: En una regresión lineal que no distingue los datos entre diferentes medios, no parece haber ningún efecto entre el gasto de publicidad y las ventas.

```
# Identificar los medios únicos para asegurarnos de asignar un marcador único para cada uno
unique_media = df['Medio'].unique()

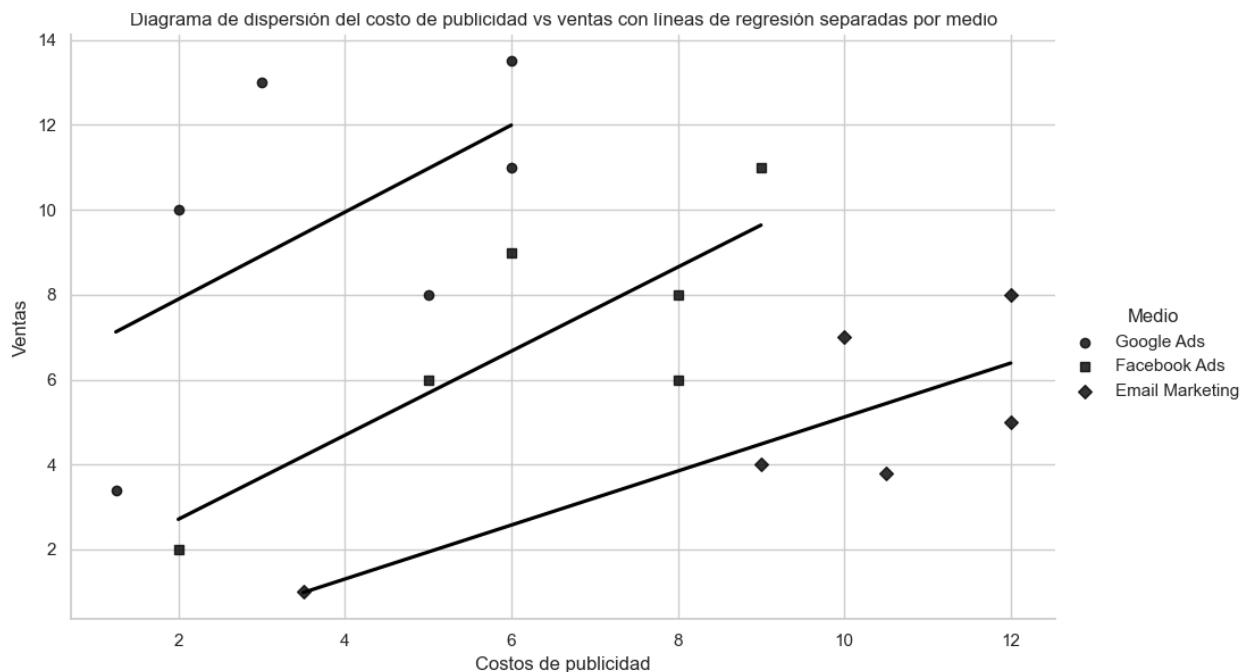
# Crear un diagrama de dispersión con diferentes marcadores para cada medio y líneas de
# → regresión separadas
sns.lmplot(
    x='Ad cost',
    y='Sales',
    data=df,
    hue='Medio',
    markers=['o', 's', 'D', '^'][len(unique_media)],
    palette = ['black'] * len(df['Medio'].unique()),
    height=6, aspect=1.6, ci=None)

# Configurar el título y las etiquetas del gráfico
plt.title('Diagrama de dispersión del costo de publicidad vs ventas con líneas de regresión
# → separadas por medio')
plt.xlabel('Costos de publicidad')
plt.ylabel('Ventas')
plt.show()
```

Los *efectos fijos* son sólo una aglomeración de variables de control en una sola

La clave de los efectos fijos es que:

- Podemos incluir el efecto del tiempo en nuestro modelo, pero el tiempo en sí mismo no es una variable.
- Lo que importa es que estamos capturando todas las características intrínsecas de nuestro medio y estamos asumiendo que son “fijas”.



El modelo de efectos fijos se define en términos generales como

$$Y_{it} = \beta X_{it} + \gamma U_i + \epsilon_{it}$$

donde Y_{it} es el resultado que tiene el individuo i en el tiempo t , que puede medirse en meses, años, trimestres o lo que sea que tenga sentido. Nuevamente, X_{it} es el vector de variables para el individuo i en el tiempo t ⁹².

Nota que ahora incluimos una variable U_i que no tiene subíndice de tiempo t .

Esta representa el conjunto de inobservables del individuo i ⁹³. Este elemento no tiene subíndice t porque asumimos que estos inobservables no tienen variación en el tiempo. Por ejemplo, en una campaña de anuncios de google podemos asumir que el algoritmo que subasta un término de búsqueda es el mismo para todas las observaciones que hacemos.

Variación dentro del individuo (within)

En teoría, los efectos fijos funcionan igual que si usáramos una variable dummy para cada uno de los individuos.

El problema es que no es raro que nuestro panel se componga de más de 3 variables como en el ejemplo. Imaginemos que estamos tra-

Figura 30: Sólo es necesario incluir el medio como variable de control. Hacerlo delata la verdadera relación que hay entre la publicidad y las ventas.

⁹² Es muy común aprender sobre datos en panel con zonas geográficas. Por ejemplo, con registros de diferentes países en el tiempo.

⁹³ La U es porque en inglés se dice *unobservables*.

tando de hacer un panel para una campaña gigantesca ultrasegmentada de contenido con facebook ads. Una campaña así funcionaría haciendo un anuncio para cada pieza de contenido que hacemos, pautando (haciendo publicidad) y dejando que el algoritmo de meta encuentre a los consumidores ideales de ese contenido. Lo que obtenemos es una campaña para cada pieza de contenido que corre en paralelo. Esto puede hacer que el tamaño de nuestra base de datos aumente muy rápido⁹⁴.

Por eso haremos un pequeño truco que nos permitirá obtener el mismo resultado que si usáramos variables *dummy*, pero con un conjunto más manejable de datos.

El primer paso es obtener el valor de la media condicional de publicidad y ventas en cada uno de los medios. Nuestro objetivo será identificar el punto en el que se cruzan los puntos medios de cada grupo para posteriormente juntarlos en un mismo medio.

Así se ven de manera visual.

```
# Calculando el promedio y la desviación estándar del costo de anuncios y las ventas para cada
# medio
media_stats = df.groupby('Medio').agg({'Ad cost': ['mean', 'std'], 'Sales': ['mean',
# 'std']}).reset_index()

# Creando el gráfico
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Ad cost', y='Sales', data=df, hue='Medio', style='Medio', markers=[ 'o', 's',
# 'D'][len(unique_media)])

# Añadiendo las líneas para representar el promedio ± una desviación estándar
for _, row in media_stats.iterrows():
    medio = row['Medio']
    ad_cost_mean = row[('Ad cost', 'mean')]
    ad_cost_std = row[('Ad cost', 'std')]
    sales_mean = row[('Sales', 'mean')]
    sales_std = row[('Sales', 'std')]

    # Dibujando líneas para el costo de anuncios
    plt.plot([ad_cost_mean - ad_cost_std, ad_cost_mean + ad_cost_std], [sales_mean, sales_mean],
             color='black', linestyle='--', linewidth=1)

    # Dibujando líneas para las ventas
    plt.plot([ad_cost_mean, ad_cost_mean], [sales_mean - sales_std, sales_mean + sales_std],
             color='black', linestyle='--', linewidth=1)

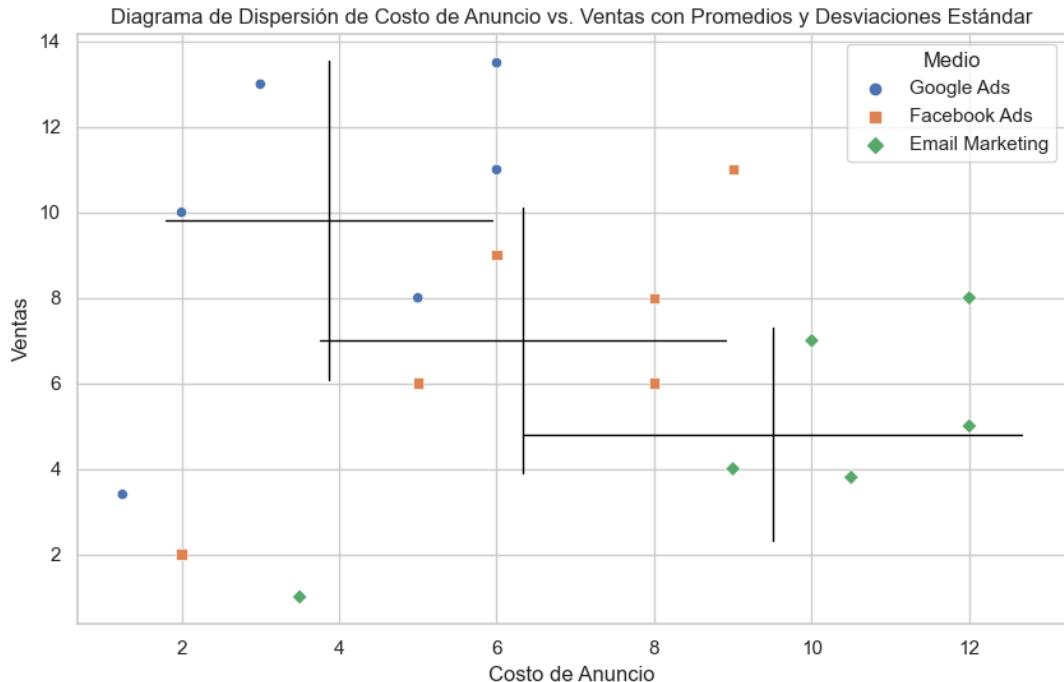
plt.title('Diagrama de Dispersión de Costo de Anuncio vs. Ventas con Promedios y Desviaciones
# Estándar')
plt.xlabel('Costo de Anuncio')
```

⁹⁴ Si de verdad usáramos *dummies* para cada uno de los individuos, serían $n - 1$ variables adicionales las que tendríamos que incluir en nuestra regresión. Si estás usando R como software para hacer tu trabajo y si incluyes la variable de clasificación como un objeto de tipo `factor`, el software lo transforma en *dummies* de forma automática y hace la regresión con la función `lm()`.

```

plt.ylabel('Ventas')
plt.legend(title='Medio')
plt.grid(True)
plt.show()

```



El siguiente paso es restar estas medias de nuestros individuos.

$$\begin{aligned}\ddot{Y}_{it} &= Y_{it} - \bar{Y}_i \\ \ddot{X}_{it} &= X_{it} - \bar{X}_i\end{aligned}$$

Visualmente lo que esto logra es como si “empalmaramos” las cruces que se formaron en el gráfico anterior.

Ahora sólo tenemos que hacer una regresión de \ddot{Y}_{it} contra \ddot{X}_{it} , o bien

$$\begin{aligned}(Y_{it} - \bar{Y}_i) &= \beta(X_{it} - \bar{X}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \\ \ddot{Y}_{it} &= \beta \ddot{X}_{it} + \ddot{\epsilon}_{it}\end{aligned}$$

Estas son las nuevas variables de nuestra regresión de efectos fijos.

Nota que el término de elementos inobservables desaparece. Esto es porque $U_i = \bar{U}_i$, por su propia definición. Es una operación que elimina todos los términos constantes en el tiempo.

Figura 31: Cada medio de publicidad tiene un valor promedio de ventas (\bar{Y}_i) y del costo del anuncio (\bar{X}_i). Las líneas en el gráfico muestran donde se cruzan esos valores promedio, con una desviación estándar para determinar el tamaño de la línea. Este es un paso determinante para crear nuestras variables *within*.

En concreto, la regresión se hace sobre este conjunto de datos.

```
df['Within Ad cost'] = df.groupby('Medio')['Ad cost'].transform(lambda x: x - x.mean())
df['Sales Within'] = df.groupby('Medio')['Sales'].transform(lambda x: x - x.mean())
df.head()
```

Año	Gasto publicitario	Ventas	Gasto interno	Ventas internas
2020	1.25	3.4	-2.625	-6.42
2021	2.00	10.0	-1.875	0.18
2022	6.00	13.5	2.125	3.68
2023	5.00	8.0	1.125	-1.82
2024	6.00	11.0	2.125	1.18

Cuadro 4: Gasto en publicidad y ventas por año (Google Ads)

Notas: El “Gasto interno” y “Ventas internas” corresponden a las variables centradas dentro del medio ([transformación within](#)). Todos los valores corresponden a la plataforma Google Ads.

Y visualmente, podemos observar que colocamos juntos los centros que calculamos con anterioridad. Nuestro modelo es una regresión lineal con los datos modificados de esta manera:

```
# Calculando las estadísticas necesarias para los cruces en el gráfico
media_within_stats = df.groupby('Medio').agg({'Within Ad cost': ['mean', 'std'], 'Sales Within':
    ['mean', 'std']}).reset_index()

# Creando el gráfico con las variables "Within"
plt.figure(figsize=(10, 6))

# Usando scatterplot para los puntos con color, pero sin añadir la línea de regresión aquí
sns.scatterplot(x='Within Ad cost', y='Sales Within', data=df, hue='Medio', style='Medio',
    markers=['o', 's', 'D'][len(df['Medio'].unique())])

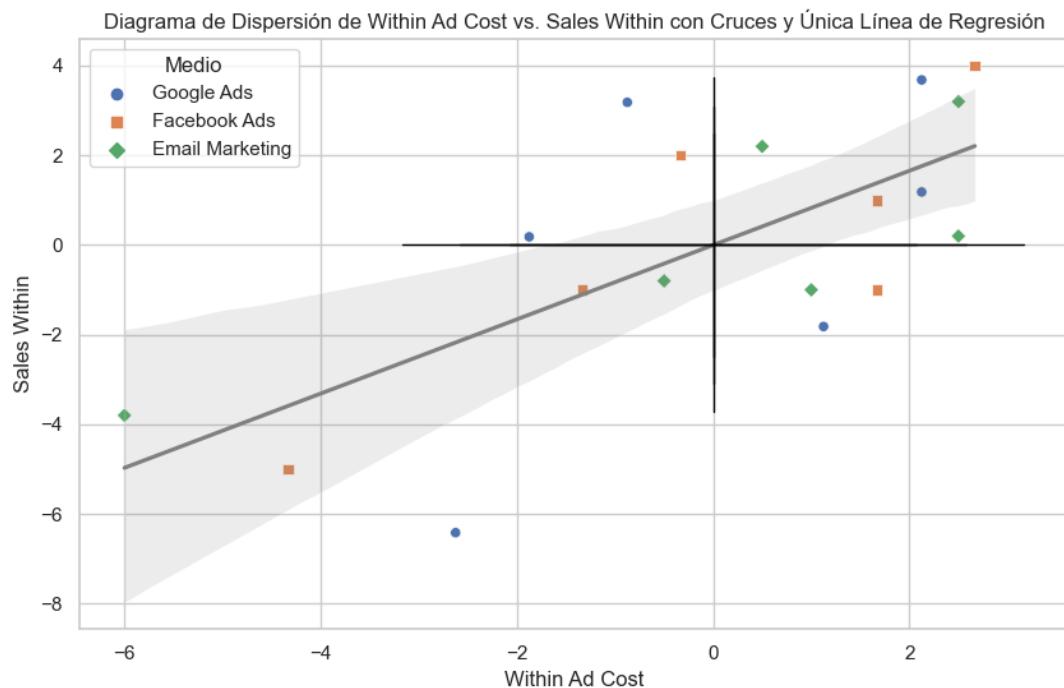
# Añadiendo la línea de regresión con regplot
sns.regplot(x='Within Ad cost', y='Sales Within', data=df, scatter=False, color='gray')

# Añadiendo las cruces que representan la media ± una desviación estándar para las variables
# within
for _, row in media_within_stats.iterrows():
    medio = row['Medio']
    within_ad_cost_mean = row[['Within Ad cost', 'mean']]
    within_ad_cost_std = row[['Within Ad cost', 'std']]
    sales_within_mean = row[['Sales Within', 'mean']]
    sales_within_std = row[['Sales Within', 'std']]

    # Dibujando líneas para Within Ad cost
    plt.plot([within_ad_cost_mean - within_ad_cost_std, within_ad_cost_mean +
        within_ad_cost_std], [sales_within_mean, sales_within_mean],
        color='black', linestyle='-', linewidth=1)
```

```
# Dibujando líneas para Sales Within
plt.plot([within_ad_cost_mean, within_ad_cost_mean], [sales_within_mean - sales_within_std,
→ sales_within_mean + sales_within_std],
color='black', linestyle='-', linewidth=1)

plt.title('Diagrama de Dispersión de Within Ad Cost vs. Sales Within con Cruces y Única Línea de
→ Regresión')
plt.xlabel('Within Ad Cost')
plt.ylabel('Sales Within')
plt.legend(title='Medio')
plt.grid(True)
plt.show()
```



Nota cómo ahora las medias están en cero en los dos ejes.

Ahora todos los datos están en un punto comparable. A esto se le llama “absorber” los efectos fijos.

Naturalmente, ahora podemos aplicar una regresión lineal simple a nuestros datos.

Figura 32: Al absorber los efectos fijos, los datos se conjuntan en un solo punto medio de publicidad y ventas. El resultado es que el coeficiente de una regresión conjunta captura el efecto independiente del medio que se está usando.

```

import statsmodels.api as sm
import statsmodels.formula.api as smf

# Define la fórmula para el modelo de efectos fijos con las
# variables centradas respecto a su grupo
formula = 'Q("Sales Within") ~ Q("Within Ad cost")'

# Ajusta el modelo de efectos fijos
model = smf.ols(formula, data=df).fit()

# Muestra el resumen de los resultados
model.summary()

```

Y listo. En la siguiente tabla mostramos los resultados de la regresión.

	Coeficientes	Error estándar	IC 95 %
Costo de publicidad	0.828	0.214	[0.374, 1.282]
Intercepto	0.000	0.517	[-1.095, 1.095]
R^2	0.483		
R^2 ajustada	0.451		
N (observaciones)	18		

La regresión sobre nuestros datos centrados es una regresión lineal simple. Podemos observar que la relación entre la publicidad y las ventas es positiva al ver el coeficiente. También es una relación significativa, de acuerdo al estadístico t y al *p-value*.

Este es un modelo sencillo con sólo dos variables. Por eso podemos aplicar el truco de centrar las variables en grupos directamente y usar una regresión sencilla de mínimos cuadrados. Sin embargo, en modelos más complejos, tendrás que usar paquetería especializada para el manejo de datos en panel.

Este es el código para hacer la regresión de efectos fijos con la función `PanelOLS`, del módulo `linearmodels`.

```

from linearmodels.panel import PanelOLS

# Preparing the data: Setting 'Medio' and 'Año' as index
df_panel = df.set_index(['Medio', 'Año'])

# Specifying and fitting the model with entity effects (fixed effects)
panel_model = PanelOLS.from_formula('Sales ~ Q("Ad cost") + EntityEffects', data=df_panel)

# Fitting the model

```

Cuadro 5: Efecto de la publicidad en las ventas hecha con un modelo de mínimos cuadrados con las variables centradas por grupos.

```

panel_results = panel_model.fit()

# Displaying the results
panel_results.summary

```

La siguiente tabla condensa la información que nos genera el reporte de nuestra regresión por panel.

	Coeficiente	Error estándar	IC 95 %
Costo publicitario (x_{it})	0.828***	0.229	[0.337, 1.319]
R^2 (dentro)	0.483		
R^2 (entre)	0.667		
R^2 (global)	0.643		
N (observaciones)	18		
Número de entidades	3		
Número de periodos	6		
Efectos incluidos	Entidad (efectos fijos)		
Estadístico F (modelo)	13.09 (p = 0.0028)		
prueba de poolabilidad	F(2,14) = 13.45, p = 0.0006		

Cuadro 6: Modelo de efectos fijos
(PanelOLS) – Variable dependiente:
Ventas

Notas: Estimaciones obtenidas mediante mínimos cuadrados para datos de panel con efectos fijos.

Niveles de significancia: * $p < 0,1$, ** $p < 0,05$, *** $p < 0,01$

¿Qué significa todo esto? Aquí algunos puntos a los que poner atención:

- El modelo muestra un valor de R^2 *within* (dentro), de 0.483. Lo podemos interpretar como que un 48 % de la variación de las ventas se explica por el gasto en publicidad.
- La prueba de agrupabilidad rechaza la hipótesis nula de que se pueda agrupar a todos los medios en un intercepto común. Con esto validamos el uso de un modelo de efectos fijos por entidad, a diferencia de un modelo agrupado (que sería la regresión simple de todos los datos).
- El modelo estima un resultado positivo y significativo. Un aumento en una unidad del gasto en publicidad se asocia con un aumento de 0.828 unidades en ventas. Este resultado ya controla diferencias constantes entre medios.

Nada mal, ¿cierto?.

El efecto fijo de dos vías

Digamos que queremos hacer lo mismo no sólo para los individuos, sino también para el tiempo.

El resultado sería un modelo como este:

$$Y_{it} = U_i + U_t + \beta X_{it} + \epsilon_{it}$$

Lo que este modelo nos da es una estimación que permite comparar la variación entre individuos al mismo tiempo que *entre años*.

Por ejemplo, nota que en los datos anteriores, las ventas del año 2020 son relativamente más bajas que las demás. Es un recordatorio de la época de pandemia, en la que cerraron todos los negocios y las ventas de muchas cosas bajaron, a menos de que vendas cubrebocas. En el caso de las ventas por Facebook, el nivel de ventas que muestran los datos es muy bajo para lo que estamos acostumbrados a vender por ese medio, pero no es un nivel de ventas tan fuera de lo normal para el marketing por e-mail.

Usa este código para hacer una regresión de panel por dos vías.

```
from linearmodels.panel import PanelOLS

# Efectos fijos de dos vías
model = PanelOLS.from_formula('Sales ~ Q("Ad cost") +
    EntityEffects + TimeEffects', data=df_panel)
results = model.fit()
print(results.summary)
```

Nuevamente, veamos un resumen de los resultados en el cuadro 7.

Nota que Python permite hacer este tipo de modelos de una manera muy sencilla. Sólo es necesario incluir en la regresión los `EntityEffects` al mismo tiempo que los `TimeEffects`.

En esta tabla los resultados no parecen ser tan alentadores. El gasto en publicidad no tiene un efecto significativo con las ventas cuando controlamos por medio y por el tiempo a la vez.

Los **efectos fijos de dos vías** son un tema muy interesante, pues capturan las características completas del año, a la vez que segmentan los efectos de los grupos en los que están clasificados los datos. Por eso este tipo de modelos son la base del modelo de **diferencias en diferencias**, que permiten diferenciar el efecto en el tiempo que tienen los grupos de “tratamiento” contra los de “control”.

	Coefficiente	Error estándar	IC 95 %	Cuadro 7: Modelo de efectos fijos (PanelOLS) – Variable dependiente: Ventas
Costo publicitario (x_{it})	-0.089	0.322	[-0.817, 0.639]	
R^2 (dentro)	-0.109			
R^2 (entre)	-0.142			
R^2 (global)	-0.138			
N (observaciones)	18			
Número de entidades	3			
Número de periodos	6			
Efectos incluidos	Entidad y tiempo (efectos fijos)			
Estadístico F (modelo)	0.08 ($p = 0.789$)			
Prueba de agrupabilidad	F(7,9) = 14.19, $p = 0.0003$			

Notas: Estimaciones obtenidas mediante mínimos cuadrados para datos de panel con efectos fijos por entidad y por periodo.

Niveles de significancia: * $p < 0,1$, ** $p < 0,05$, *** $p < 0,01$

Lo que en la práctica quiere decir es que los modelos de diferencias en diferencias son modelos de efectos fijos de dos vías, pero no todos los efectos fijos de dos vías pueden generar un diseño de diferencias en diferencias.

¿Cuál es la diferencia? La forma en la que seleccionamos los grupos para que sean de tratamiento y de control.

En un modelo de diferencias en diferencias, lo que queremos es asemejar el diseño del estudio a un experimento natural.

Eso lo veremos en el siguiente capítulo.

Resumen del capítulo

En este capítulo nos enfrentamos a uno de los mayores enemigos de la inferencia causal: las variables importantes que no podemos ver ni medir.

Lo que hicimos fue aprender a usar una nueva estructura de datos, los [datos en panel](#), que nos permiten seguir a los mismos individuos (empresas, países, o en nuestro caso, medios de publicidad) a lo largo del tiempo. Vimos cómo una regresión simple puede mentirnos descaradamente, mostrándonos una relación negativa donde en realidad había una positiva. La solución fue aplicar la magia de los [efectos fijos](#), un método que controla por todas esas características “inobservables” y constantes de cada individuo, como la “calidad” o la “confianza” inherente a cada medio de publicidad. Lo hicimos de dos formas: a mano, con la transformación *within*, y luego con las herramientas profesionales de Python. Finalmente, subimos el nivel con los [efectos fijos de dos vías](#) para controlar no solo por las diferencias entre individuos, sino también por los shocks que afectan a todos por igual en un año determinado.

Esto es importante porque la mayoría de los problemas del mundo real están plagados de heterogeneidad.

neidad inobservable. La “cultura” de una empresa, la “habilidad” de un vendedor, o la “calidad” de una escuela son factores decisivos que casi nunca tenemos en una columna de nuestros datos. Ignorarlos lleva a conclusiones erróneas. El modelo de efectos fijos es tu primera gran herramienta para aislar un efecto causal en datos que no vienen de un experimento, permitiéndote hacer comparaciones mucho más justas.

¿Cómo te ayuda esto? Ahora tienes un método para analizar datos de individuos a lo largo del tiempo. Cuando sospeches que hay diferencias fundamentales y persistentes entre los grupos que comparas (ej. distintas sucursales de una tienda, distintos países, distintos productos), tu primer instinto será usar un modelo de efectos fijos. Te permite responder a la pregunta: “Dentro de cada sucursal, ¿qué efecto tuvo cambiar X en el resultado Y, una vez que eliminamos las diferencias preexistentes entre la sucursal buena y la mala?”. Es una técnica increíblemente poderosa y es la base para entender el modelo de Diferencias en Diferencias que veremos a continuación.

Controlando el caos: ejercicios de panel y efectos fijos

Es hora de aplicar esta poderosa técnica. Estos ejercicios te ayudarán a solidificar la intuición detrás de los efectos fijos y a practicar su implementación.

1. **La trampa de la regresión simple (Conceptual):** El capítulo mostró cómo una regresión simple (o “agrupada”) daba un resultado engañoso. Describe otro escenario de negocios donde podría pasar lo mismo. Por ejemplo, si analizaras la relación entre las horas trabajadas por los vendedores y las ventas totales, ¿qué “efecto fijo” a nivel de vendedor podría confundir los resultados si lo ignoras?
2. **La magia de la transformación *within* (Conceptual):** El truco de restar la media de cada individuo (“de-meaning”) elimina la variable inobservable U_i . ¿Por qué funciona esto? ¿Qué característica fundamental debe tener esa variable inobservable para que el truco de la resta la haga desaparecer?
3. **Calculando el “Within” a mano (Código):** Usando el DataFrame `sales-panel.csv`, calcula manualmente las variables “within” para el medio `Facebook Ads`. Es decir, calcula la media de `Ad cost` y `Sales` solo para `Facebook Ads`, y luego resta esas medias de los valores originales para obtener `Within Ad cost` y `Sales Within` para ese medio.
4. **Requisitos para un modelo de panel (Conceptual):** Un colega te entrega un archivo de Excel y te pregunta si puede usar un modelo de efectos fijos. ¿Qué dos características clave debe tener esa tabla de datos para que tu respuesta sea “sí”?
5. **Efectos fijos por entidad (Código):** Carga el dataset `sales-panel.csv`. Usando la librería `linearmodels`, corre un modelo de efectos fijos por entidad (`EntityEffects`) para predecir `Sales` a partir de `Ad cost`. Muestra la tabla de resultados. Confirma que el coeficiente que obtienes para `Ad cost` es el mismo (0.828) que el que se obtuvo con la regresión sobre los datos transformados manualmente en el capítulo.
6. **Efectos fijos por tiempo (Código y Conceptual):** Ahora, usando `linearmodels`, corre un modelo de efectos fijos *solo* por tiempo (`TimeEffects`), sin efectos de entidad. ¿Qué coeficiente obtienes

para `Ad cost`? ¿Es significativo? ¿Qué te dice este resultado sobre la importancia de controlar por las características de cada *medio* en lugar de controlar solo por el *año*?

7. **Interpretando el R-cuadrado “Within” (Conceptual):** El reporte de `PanelOLS` con efectos de entidad te da un R^2 *within* de 0.483. Explica en una frase qué significa este número. ¿A qué porcentaje de la varianza se refiere?
8. **El poder de los dos efectos (Código y Conceptual):** Corre el modelo de efectos fijos de dos vías (`EntityEffects + TimeEffects`) como se muestra en el capítulo. El resultado para `Ad cost` ahora no es significativo (*p-value* = 0.789). Propón una historia de negocio que explique por qué podría pasar esto. ¿Qué pudo haber ocurrido en ciertos años que, al controlarlo, “absorbió” el efecto que antes veíamos en la publicidad?
9. **La prueba de ‘Poolability’ (Conceptual):** En el resultado del modelo con solo efectos de entidad, la prueba de “Poolability” tiene un *p-value* muy bajo (0.0006). La hipótesis nula de esta prueba es que los efectos fijos de todos los medios son iguales entre sí. Dado el *p-value*, ¿rechazas o no rechazas la hipótesis nula? ¿Qué te dice esto sobre si fue buena idea usar efectos fijos en lugar de una regresión simple?
10. **Reto - Estructurando datos de panel (Conceptual y Código):** Imagina que tienes un archivo CSV con esta estructura (formato “ancho” o *wide*):

Tienda	Ventas_2022	GastoPubli_2022	Ventas_2023	GastoPubli_2023
A	100	10	120	12
B	200	15	210	16

Los modelos de panel necesitan un formato “largo” (*long*). Describe cómo se vería esta tabla en formato largo. (Pista: investiga la función `pd.melt` de pandas).

Diferencias en Diferencias

Let me finish talking with my husband. He needs to know how good my life could have been

- Evelyn en Everything, Everywhere, All at Once

¿Qué pasa cuando no podemos hacer un experimento para identificar causas y efectos?

Si hay una lección que quiero que recuerdes sobre este libro es que en nuestra mente siempre debe de estar el experimento como forma ideal de identificar causalidad. Y cuando el experimento no sea posible de hacer, o sea muy caro, entonces recurrimos a los datos.

En otras palabras, buscamos un **experimento natural**.

Los economistas llamamos experimento natural a un evento que se parece mucho a un experimento, pero ocurre sin que nadie lo haya planeado. Un ejemplo clásico es el de Card & Krueger⁹⁵, que midieron el efecto de un aumento en el salario mínimo en el empleo. La teoría neoclásica dictaba que el mercado laboral se debía comportar igual a cualquier otro mercado, con curvas de oferta y demanda. Si en lugar de bienes y servicios, el empleado está ofreciendo su trabajo, entonces un aumento en el salario mínimo debería tener como consecuencia una caída en el empleo.

Los datos indican que un aumento en el salario mínimo no tiene efecto alguno en el empleo.

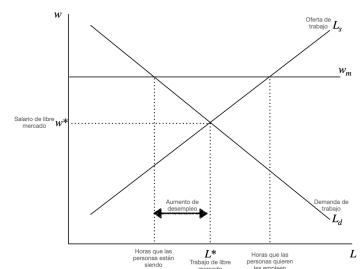
Al enterarse de que habría un aumento en el salario mínimo en Nueva Jersey, los investigadores fueron a los establecimientos de comida rápida a recolectar datos. Registraron el número de empleados, salarios promedio y otros datos en Nueva Jersey y Pensylvania. Hicieron registros en ambos estados antes y después de la implementación de la medida.

Los resultados se vieron así.

La diferencia en diferencia sería:

⁹⁵ David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4):772–793, 1994

Figura 33: Bajo un modelo neoclásico ortodoxo, el establecer un salario mínimo crea un desequilibrio artificial en el mercado que causa desempleo. Este modelo fue la base de muchas políticas de libre mercado. La evidencia que presentaron Card y Krueger no muestran este efecto en la realidad. Fuente: Elaboración propia.



	PA	NJ
Empleo Antes	23.33 (1.35)	20.44 (0.51)
Empleo Después	21.17 (0.94)	21.03 (0.52)
Cambio en el Empleo Medio	-2.16 (1.25)	0.59 (0.54)

Cuadro 8: Resultados del estudio de Card & Krueger (1994)

$$[E(L_{PA}|t=1) - E(L_{PA}|t=2)] - [E(L_{NJ}|t=1) - E(L_{NJ}|t=2)] \quad (12)$$

Tomando los datos de la tabla⁹⁶:

$$(23,33 - 21,17) - (20,44 - 21,03) = 2,76 \quad (13)$$

Es decir, el empleo parece incluso haber aumentado.

Pero con un error estándar de 1.36, no podemos estar seguros de que esos resultados sean causales. O bien, no hay efectos significativos.

Si queremos conocer el efecto real que tuvo el aumento en el salario mínimo, tenemos que tratar a Pennsylvania como un **contrafactual**. Lo que esto quiere decir es que los empleos en ambas ciudades se deberían comportar igual antes del tratamiento y, por lo tanto, Nueva Jersey hubiera tenido el mismo comportamiento que Pennsylvania **si no se hubiera implementado el cambio**⁹⁷.

Cuando tomas la diferencia en el tiempo y entre regiones, lo que te queda es el efecto causal.

El modelo básico de DiD de 2x2

Vamos a hacer una generalización del modelo que te presenté de Card & Krueger.

Estas son las características del modelo:

- **Dos períodos:** $t = 1$ (antes del tratamiento) y $t = 2$ (después del tratamiento).
- **Dos grupos:** $G_i = 2$ (unidades tratadas en el periodo 2), y $G_i = \infty$ (unidades nunca tratadas).
- El modelo puede o no incluir covariables X .
- Hay disponible un número grande de observaciones independientes o clusters.

Nota que usamos la palabra *tratamiento*, como en los estudios clínicos. En el estudio de Card & Krueger, el tratamiento fue la reducción del salario mínimo. En el caso de que se trate de una implementación

⁹⁶ Usamos como notación L_i como el nivel de empleo del estado $i = \{PA, NJ\}$, y t en este caso es el periodo de tiempo del estudio. Lo dividimos en antes de la implementación del aumento del salario mínimo ($t = 1$) y después de la misma ($t = 2$).

⁹⁷ En realidad, el estudio de Card & Krueger está lleno de detalles interesantes que son una especie de manual de diseño de estudios cuasi-experimentales. Por ejemplo, el estudio se enfoca en los restaurantes de comida rápida en ambas zonas, focalizado en ciudades aledañas en ambos estados. De esta manera, no es posible atribuir las diferencias a variables difíciles de medir y de incluir en el estudio, como la cultura, o la educación. El trabajo en los restaurantes de comida rápida en ambos estados es igual, por lo que podemos decir que es un trabajo comparable.

de una política o de una campaña, es fácil imaginar cómo esta puede ser un tratamiento, pero los casos en que observamos un efecto fortuito fuera de nuestras manos (por ejemplo, la imposición de una ley), requiere un poco de imaginación verlo como un tratamiento.

Resultados potenciales en el modelo DiD 2x2

Definimos $Y_t(g)$ como el resultado potencial en el periodo t si las unidades se exponen al tratamiento por primera vez en el periodo g .

El parámetro causal que buscamos es el **Efecto de Tratamiento Promedio en los Tratados** en el periodo $t = 2$ (*Average Treatment Effect among the Treated, ATT*):

$$ATT = \underbrace{E[Y_{t=2}(2) \mid G = 2]}_{\text{Se estima a partir de los datos}} - \underbrace{E[Y_{t=2}(\infty) \mid G = 2]}_{\text{Componente contrafactual}}$$

Para validar nuestro modelo de **DiD** (DiD) consideraremos cuatro supuestos:

Supuesto #1: No interferencia y valores únicos de tratamiento

En inglés, este supuesto se conoce como **SUTVA** (*Stable Unit Treatment Value Assumption*). Bajo este supuesto de efectos causales, los resultados potenciales de una observación dada responden únicamente a su propio estatus de tratamiento y son invariantes a la asignación de tratamiento en otras unidades.

Esto implica que los resultados observados en el tiempo t se realizan como:

$$Y_{i,t} = \sum_{g \in \mathcal{G}} \mathbb{1}\{G_i = g\} \cdot Y_{i,t}(g)$$

Es decir, para unidades que son tratadas en el periodo $t = 2$, observamos $Y_{i,t}(2)$; y para aquellas que no han recibido tratamiento para $t = 2$, observamos $Y_{i,t}(\infty)$.

Supuesto #2: No anticipación

Para todas las unidades i , se cumple que $Y_{i,t}(g) = Y_{i,t}(\infty)$ para todos los grupos en los periodos previos al tratamiento.

Esto significa que las unidades tratadas no cambian su comportamiento antes de que el tratamiento comience, en anticipación a lo que ocurrirá.

Este supuesto no se suele comprobar mediante pruebas estadísticas, sino observando el contexto del **experimento natural**. En muchos estudios, se argumenta que el supuesto se cumple porque la implementación fue rápida y sin previo aviso.

Un ejemplo es el estudio de Card & Krueger⁹⁸, donde el aumento del salario mínimo en Nueva Jersey se implementó poco después de su anuncio.

Otro ejemplo aún más drástico también es de David Card. En septiembre de 1980, Fidel Castro anunció que cualquier ciudadano cubano que lo deseara podía abordar un bote en el puerto de Mariel para emigrar a Estados Unidos. Como resultado, cerca de 125,000 migrantes llegaron a Miami en un periodo muy corto, aumentando la fuerza laboral de la ciudad en un 7% (y en 20% entre los trabajadores cubanos)⁹⁹.

Este estudio, al igual que el del salario mínimo, es controversial porque contradice los modelos neoclásicos que comparan el mercado laboral con un mercado de bienes, con curvas de oferta y demanda. La sorpresa fue que no se observó un aumento del desempleo ni una caída de los salarios en Miami.

Una parte clave de la fuerza del argumento es que el experimento natural cumplía con el **supuesto de no anticipación**. Si Castro hubiese hecho el anuncio con 1 o 2 años de anticipación, el mercado podría haberse ajustado

⁹⁸ David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4):772–793, 1994

⁹⁹ David Card. The impact of the mariel boatlift on the miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257, 1990

Supuesto #3: superposición fuerte

Significa que para cada posible valor del tratamiento, existe una probabilidad positiva de recibirla dentro de cada grupo de covariables.

Formalmente, para algún $\epsilon > 0$, $\mathbb{P}[G = 1 | X] < 1 - \epsilon$ casi seguramente.

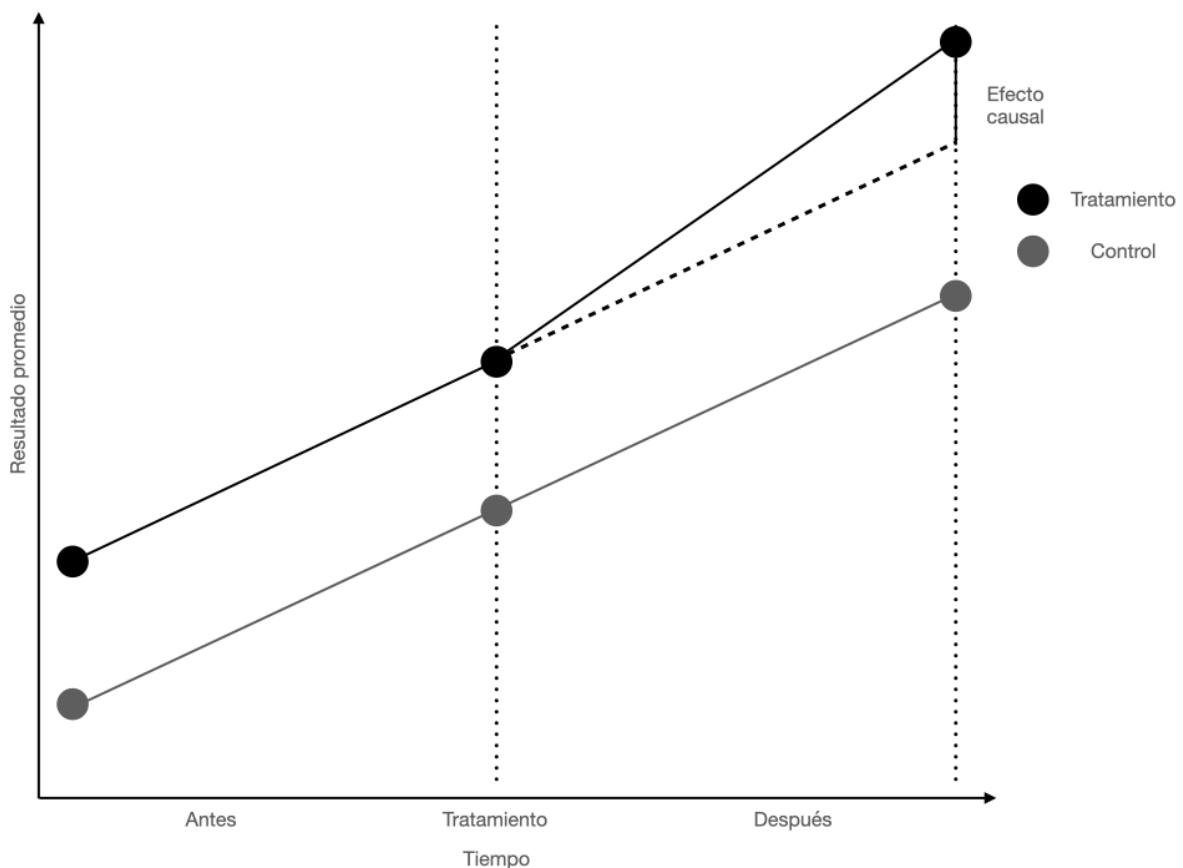
De manera intuitiva: si me dices X , no puedo decir si esa unidad es tratada con 100% de confianza.

Supuesto #4: tendencias paralelas condicionadas

El elemento más importante en el modelo de Diferencias en Diferencias es el **contrafactual**.

En el estudio del salario mínimo, los autores fueron muy cuidadosos al seleccionar las ciudades que iban a comparar. Nueva Jersey y Pennsylvania son estados vecinos. La ciudad de Nueva Jersey y de Philadelphia están separadas únicamente por un puente. Esto permite que podamos asumir con mayor tranquilidad que los efectos que estamos observando no se puedan adjudicar a la cultura o al clima.

De manera visual, se vería algo así:



Continuemos con el ejemplo de una campaña de mercadotecnia.

Para conocer el efecto verdadero de una campaña, necesitamos que exista un **contrafactual**. En otras palabras, debe haber algo con qué comparar la tendencia de las ventas. Si observamos un alza en

Figura 34: El supuesto de **tendencias paralelas** implica que, de no ser por la aplicación del tratamiento, ambos grupos seguirían la misma tendencia. Eso convierte a la diferencia entre tendencias en el efecto causal.

las ventas, podría ser parte natural de un ciclo, o podría ser parte de un boom general en la economía.

Incluir un grupo de control nos da certeza de que ese incremento se debe a la campaña y no a otros factores¹⁰⁰.

De manera formal:

$$E[Y_{t=2}(\infty)|G = 2, X] - E[Y_{t=1}(\infty)|G = 2, X] = \\ E[Y_{t=2}(\infty)|G = \infty, X] - E[Y_{t=1}(\infty)|G = \infty, X] \quad (14)$$

casi seguramente.

Dicho de otra manera, en la ausencia de tratamiento, en cada estrato de covariable, la evolución promedio del resultado Y entre las unidades tratadas en el periodo 2¹⁰¹ es la misma que la evolución promedio del resultado Y entre las unidades que permanecieron sin tratamiento¹⁰².

En otras palabras, las unidades **no tratadas** las tratamos como ese universo paralelo en el que no se hizo el tratamiento.

¹⁰⁰ Por ejemplo, una empresa que vende llantas lanza cada año una campaña para anunciar su marca el día del padre y ve que sus ventas crecen. Como sabemos, correlación no implica causalidad: sus ventas habrían aumentado lo mismo sin la campaña, simplemente por el efecto de la fecha y por el reconocimiento que ya tiene la marca. El problema es que hacer un experimento para comprobar esto requiere que en algún año no se haga campaña en alguna sucursal. ¿Vale la pena? Probablemente. Pero mientras llegan los resultados lo que la gerencia va a ver es que el de Marketing no hizo su trabajo.

¹⁰¹ Representada en el lado izquierdo de la ecuación

¹⁰² El lado derecho de la ecuación. El $G = \infty$ es nuestro indicador de las unidades que permanecieron sin tratamiento.

Identificación del modelo 2x2 sin covariables

Sin la existencia de anticipación y con tendencias paralelas, podemos demostrar que el estimador

$$\tau_{ATT} = \underbrace{(E[Y_{i,t=2}|G_i = 2] - E[Y_{i,t=1}|G_i = 2])}_{\text{Cambio en el grupo tratado}} - \\ \underbrace{(E[Y_{i,t=2}|G_i = \infty] - E[Y_{i,t=1}|G_i = \infty])}_{\text{Cambio en el grupo de comparación}} \quad (15)$$

es una “diferencia en diferencias” de medias poblacionales.

¿Cómo obtenemos este estimador? Usando efectos fijos de dos vías, por supuesto.

Hagamos un ejemplo.

En el año 2014, el gobierno de Berkeley, California implementó un impuesto a las bebidas azucaradas. Como todo economista sabe, este tipo de impuestos “al pecado” tienen intenciones que van más

allá de lo recaudatorio: sirven para modificar los incentivos de los consumidores¹⁰³.

En México también se implementó un impuesto similar en las mismas fechas. De acuerdo a Colchero, Molina & Guerrero-López¹⁰⁴, el impuesto se reflejó en una reducción de 6.3 % en las compras de bebidas azucaradas y un incremento de 16.2 % en la compra de agua embotellada en hogares de ingreso bajo y medio. Afortunadamente para el país, esta fue una implementación a nivel federal. La única desventaja de esto es que no permite que hagamos una comparación con un grupo de control para saber si ese efecto que causamos fue causal o se puede deber a otros factores exógenos.

Aquí es donde la experiencia de Berkeley, California resulta en un mejor experimento natural ideal.

Un grupo de investigadores recolectó los datos de puntos de venta en las tiendas en el área de Berkeley y zonas aledañas para revisar los efectos del impuesto en los precios de las bebidas azucaradas¹⁰⁵. El objetivo de este tipo de impuesto es, en primer lugar, hacer crecer los precios del producto, que a su vez deben de causar una reducción del consumo. Hay quienes asumen que los productos azucarados como los refrescos son tan adictivos que un alza en los precios no causaría un gran efecto en la demanda, pero los trabajos en México y en Berkeley han mostrado una diferencia significativa.

Comencemos con una gráfica:

```
import pandas as pd
import matplotlib.pyplot as plt

# Carga el conjunto de datos
file_path = "public_use_weighted_prices2.csv" # Asegura que el archivo esté en el directorio
→ correcto
df = pd.read_csv(file_path)

# Agrupa por año, mes, ubicación e impuesto, calculando el precio promedio
df_grouped = df.groupby(['year', 'month', 'location', 'tax'])['price'].mean().reset_index()

# Convierte año y mes a un formato de fecha para el eje x
df_grouped['fecha'] = pd.to_datetime(df_grouped[['year',
→ 'month']].astype(int).astype(str).agg('-'.join, axis=1), format='%Y-%m')

# Crea la gráfica de líneas
plt.figure(figsize=(12, 6))

# Define diferentes estilos de línea para cada categoría
line_styles = ['-·', '--·', '-·.', ':·']
```

¹⁰³ A. C. Pigou. *The Economics of Welfare*. Macmillan and Co., London, 1920. URL https://archive.org/details/dli_bengal.10689.4260.

¹⁰⁴ M. Arantxa Colchero, Mariana Molina, and Carlos M. Guerrero-López. After mexico implemented a tax, purchases of sugar-sweetened beverages decreased and water increased: Difference by place of residence, household composition, and income level. *Journal of Nutrition*, 147(8):1552–1557, 2017. doi: 10.3945/jn.117.251892

¹⁰⁵ Lynn D. Silver, Shu Wen Ng, Suzanne Ryan-Ibarra, Lindsey Smith Taillie, Marta Induni, Donna R. Miles, Jennifer M. Poti, and Barry M. Popkin. Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in berkeley, california, us: A before-and-after study. *PLOS Medicine*, 14(4):e1002283, 2017. doi: 10.1371/journal.pmed.1002283

```

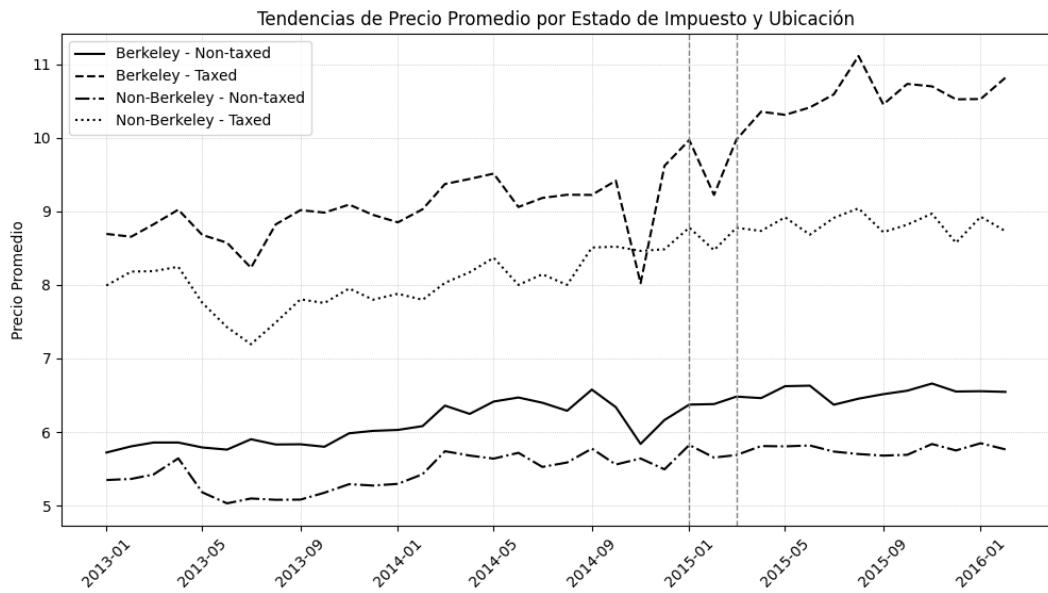
for idx, ((ubicacion, impuesto), grupo) in enumerate(df_grouped.groupby(['location', 'tax'])):
    plt.plot(grupo['fecha'], grupo['price'], label=f"{ubicacion} - {impuesto}",
              linestyle=line_styles[idx % len(line_styles)], color='black')

# Agrega líneas verticales para enero y marzo de 2015
plt.axvline(pd.to_datetime('2015-01-01'), color='gray', linestyle='--', linewidth=1)
plt.axvline(pd.to_datetime('2015-03-01'), color='gray', linestyle='--', linewidth=1)

# Configura la gráfica
plt.xlabel('Tiempo')
plt.ylabel('Precio Promedio')
plt.title('Tendencias de Precio Promedio por Estado de Impuesto y Ubicación')
plt.legend()
plt.xticks(rotation=45)
plt.grid(True, linestyle=':', linewidth=0.5)

# Muestra la gráfica
plt.show()

```



La gráfica muestra cuatro casos diferentes en dos dimensiones: productos con o sin impuestos y ventas realizadas en Berkeley y fuera de la zona de Berkeley. Las líneas verticales muestran el periodo de transición de la implementación del impuesto. La gráfica está mostrando los precios promedio de cada uno de los casos en el tiempo: las dos líneas de la parte alta son productos que son sujetos a recibir el impuesto (refrescos, jugos, té, bebidas energéticas) y las líneas de abajo son productos que nunca fueron sujetos al impuesto.

Figura 35: Precios promedio de bebidas en la zona de Berkeley. Fuente: Elaboración propia con datos de Silver *et al.* (2014)

De manera visual, la diferencia en diferencias es la diferencia que tiene la separación de las líneas de arriba con las de abajo: nota que después de la aplicación del impuesto las líneas de arriba se separan más. Esto es evidencia de que el aumento de precio de las bebidas azucaradas en Berkeley (la línea más alta) se debe al impuesto y no a factores exógenos.

Ahora hagamos la estimación de estas diferencias en diferencias usando un modelo sencillo de efectos fijos de dos vías:

$$P_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 (D_i \times T_t) + \gamma_i + \delta_t + \varepsilon_{it}$$

En este modelo, P_{it} representa el precio. Tenemos dos variables *dummy*: Una que indica si se trata de un producto al que se le aplica el impuesto (D_i) y otra que indica si estamos en un momento anterior o posterior a la implementación del impuesto.

Los coeficientes γ_i y δ_t son los efectos fijos que estamos aplicando a nuestra regresión de dos vías¹⁰⁶.

Resultado de la regresión de efectos fijos de dos vías (TWFE)

El siguiente código hace la regresión de efectos fijos de dos vías:

```
# Importamos las librerías necesarias
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Cargamos el dataset
file_path = "public_use_weighted_prices2.csv"
df = pd.read_csv(file_path)

# Agrupamos por año, mes, ubicación e impuesto, calculando el precio promedio
df_grouped = df.groupby(['year', 'month', 'location', 'tax'])['price'].mean().reset_index()

# Convertimos 'year' y 'month' en un formato de fecha para el análisis
df_grouped['fecha'] = pd.to_datetime(df_grouped[['year',
    ↵ 'month']].astype(int).astype(str).agg('-'.join, axis=1), format='%Y-%m')

# Definimos la fecha de implementación del impuesto (enero de 2015)
tax_implementation_date = pd.to_datetime("2015-01-01")

# Creamos las variables necesarias para el modelo
df_grouped['PostTax'] = (df_grouped['fecha'] >= tax_implementation_date).astype(int) # 1
    ↵ después del impuesto, 0 antes
```

¹⁰⁶ La parte $D_i \times T_t$ se conoce como término **de interacción**. Es una *dummy* que toma el valor de 1 cuando la unidad de tratamiento D_i y la dummy de tiempo T_t están en de manera conjunta en el tratamiento.

```

df_grouped['Taxed'] = (df_grouped['tax'] != "Non-taxed").astype(int) # 1 si está gravado, 0 si
                                                               ↵ no

# Estimamos el modelo TWFE con efectos fijos por ubicación y por tiempo
modelo = smf.ols("price ~ Taxed * PostTax + C(location) + C(fecha)", data=df_grouped).fit()

# Mostramos los resultados de la regresión
print(modelo.summary())

```

La siguiente tabla muestra un resumen de los resultados:

	Coeficiente	Error estándar	IC 95 %
Zona con impuesto (Taxed)	2.753***	0.059	[2.636, 2.871]
Periodo posterior (PostTax)	0.535***	0.148	[0.241, 0.828]
Interacción: Taxed × PostTax	0.709***	0.098	[0.515, 0.903]
Intercepto	6.035***	0.151	[5.736, 6.333]
R^2	0.978		
R^2 ajustado	0.970		
N (observaciones)	152		
Estadístico F	121.0 (p <0.001)		
Efectos fijos incluidos	Ubicación y tiempo (mensual)		

Notas: Modelo OLS con efectos fijos de dos vías.

La variable de interacción captura el efecto causal del impuesto sobre los precios en Berkeley.

Niveles de significancia: * $p < 0,1$, ** $p < 0,05$, *** $p < 0,01$

El coeficiente β_3 muestra nuestro estimador de diferencias en diferencias (el [término de interacción](#)). Un coeficiente de $\beta_3 = 0,709$ indica que, en promedio, las bebidas azucaradas vieron un incremento adicional de 0.71 en comparación con las bebidas no-azucaradas.

Y es un resultado estadísticamente significativo.

Podemos hacer un modelo idéntico en el que comparamos el efecto dentro y fuera del área de Berkeley. Este sería el modelo:

$$P_{it} = \beta_0 + \beta_1 Berkeley_i + \beta_2 T_t + \beta_3 (D_i \times T_t) + \gamma_i + \delta_t + \varepsilon_{it}$$

Este es el resultado.

```

# Importamos las librerías necesarias
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf

```

```

# Cargamos el dataset
file_path = "public_use_weighted_prices2.csv"
df = pd.read_csv(file_path)

# Agrupamos por año, mes, ubicación e impuesto, calculando el precio promedio
df_grouped = df.groupby(['year', 'month', 'location', 'tax'])['price'].mean().reset_index()

# Convertimos 'year' y 'month' en un formato de fecha para el análisis
df_grouped['fecha'] = pd.to_datetime(df_grouped[['year',
    ↵ 'month']].astype(int).astype(str).agg('-'.join, axis=1), format='%Y-%m')

# Definimos la fecha de implementación del impuesto (enero de 2015)
tax_implementation_date = pd.to_datetime("2015-01-01")

# Creamos las variables de tratamiento:
df_grouped['T_t'] = (df_grouped['fecha'] >= tax_implementation_date).astype(int) # 1 si es
    ↵ después del impuesto, 0 antes
df_grouped['Berkeley'] = (df_grouped['location'] == "Berkeley").astype(int) # 1 si la ubicación
    ↵ es Berkeley, 0 en otros lugares

# Creamos la variable de interacción entre Berkeley y el tiempo después del impuesto
df_grouped['Berkeley_T_t'] = df_grouped['Berkeley'] * df_grouped['T_t']

# Estimamos el modelo TWFE con efectos fijos por ubicación y por tiempo
formula = "price ~ Berkeley + T_t + Berkeley_T_t + C(location) + C(fecha)"
twfe_berkeley_model = smf.ols(formula, data=df_grouped).fit()

# Mostramos los resultados de la regresión
print(twfe_berkeley_model.summary())

```

	Coeficiente	Error estándar	IC 95 %
Zona Berkeley	2.715***	0.349	[2.023, 3.408]
Periodo alternativo (T_t)	0.688	0.908	[-1.111, 2.487]
Interacción: Berkeley $\times T_t$	0.381	0.601	[-0.811, 1.572]
Intercepto	4.626***	0.596	[3.445, 5.807]
R^2	0.149		
R^2 ajustado	-0.148		
N (observaciones)	152		
Estadístico F	0.50 (p = 0.992)		
Efectos fijos incluidos	Mensuales		

Notas: Modelo OLS con efectos fijos mensuales.

La interacción simulada no presenta efectos significativos, lo cual refuerza la validez del análisis principal.

Niveles de significancia: * $p < 0,1$, ** $p < 0,05$, *** $p < 0,01$

Aquí los resultados no tienen ese efecto significativo del modelo anterior. ¡Pero eso es algo bueno!

Esto parece indicar que el efecto viene más por el producto con impuesto y no tanto por la diferencia que hay dentro y fuera de Berkeley. Esto es esperado si la especificación original estaba capturando un efecto causal real: cuando reemplazamos nuestra variable de tratamiento por una arbitraria, el efecto desaparece.

Resumen del capítulo

En este capítulo aprendimos a cazar [experimento naturals](#), esos momentos afortunados en los que el mundo real nos regala un grupo de tratamiento y un grupo de control sin que nosotros tengamos que intervenir. La herramienta para analizar estos regalos es una de las más elegantes de la econometría: el modelo de [DiD](#).

Lo que hicimos fue entender su lógica simple pero poderosa. Tomamos un grupo afectado por un evento (el “tratamiento”) y uno que no lo fue (el “control”), y medimos un resultado antes y después. La primera diferencia calcula el cambio en el tiempo para cada grupo. La segunda diferencia —la diferencia entre estos dos cambios— es nuestro efecto causal. Este truco mágico nos permite descontar cualquier tendencia que hubiera ocurrido de todas formas, aislando el verdadero impacto del evento. Vimos que este método, que parece una simple resta de promedios, es en realidad un modelo de [efectos fijos de dos vías](#) donde el coeficiente clave es el del término de interacción.

Esto es importante porque nos acerca un paso más a establecer causalidad con datos del mundo real, sin necesidad de un [ensayo controlado aleatorizado \(RCT\)](#). El supuesto clave que lo sostiene todo, el de [tendencias paralelas](#), nos obliga a ser rigurosos y a justificar por qué nuestro grupo de control es un buen “universo paralelo” del grupo de tratamiento. Dominar el modelo DiD te permite evaluar el impacto de políticas, eventos y decisiones de negocio con un nivel de credibilidad muy alto.

¿Cómo te ayuda esto? Ahora, cuando leas en las noticias que una nueva ley o un evento afectó a una región pero no a otra, tu cerebro inmediatamente pensará: “¡esto es un problema de Diferencias en Diferencias!”. Tienes una receta clara para estimar su impacto: encuentra un grupo de control creíble, dibuja sus trayectorias para verificar visualmente que eran paralelas antes del evento, y luego corre una regresión con una interacción. Esta es la herramienta que te permitirá medir el verdadero efecto de una campaña de marketing en una ciudad piloto, el impacto de una nueva regulación, o cualquier evento que divida al mundo en “tratados” y “no tratados”.

Buscando experimentos naturales: ejercicios de DiD

La mejor forma de entender el modelo DiD es aplicándolo. Estos ejercicios te ayudarán a identificar su estructura, a pensar en sus supuestos y a implementarlo en la práctica.

1. **Identificando los componentes de DiD (Conceptual):** Para el estudio clásico de Card y Krueger sobre el salario mínimo:

- ¿Cuál es el grupo de tratamiento y cuál el de control?

- ¿Cuál es el periodo “antes” y “después”?
 - ¿Cuál es la variable de resultado que se está midiendo?
 - Usando los números de la tabla, verifica el cálculo del estimador DiD de 2.76.
2. **El supuesto sagrado (Conceptual):** Mirando el gráfico de los precios de las bebidas en Berkeley, enfócate en el periodo *antes* de las líneas verticales que marcan la implementación del impuesto. ¿Dirías que el supuesto de tendencias paralelas se cumple a simple vista entre las bebidas que recibirán el impuesto y las que no? ¿Por qué sí o por qué no?
3. **Construyendo la interacción (Código):** Usando el DataFrame `df_grouped` del capítulo (el de los precios de las bebidas), crea manualmente la variable de interacción `Taxed_x_PostTax` multiplicando la columna `Taxed` por la columna `PostTax`. Luego, corre una regresión simple con `smf.ols` usando la fórmula `price ~ Taxed + PostTax + Taxed_x_PostTax`. Confirma que el coeficiente de tu variable de interacción es idéntico al que se obtiene con la sintaxis `Taxed * PostTax`.
4. **¿Un buen DiD? (Conceptual):** Una empresa implementa un programa de 4 días laborales en su oficina de Guadalajara (tratamiento) para ver si aumenta la satisfacción, mientras que su oficina de Monterrey (control) sigue con 5 días. Miden la satisfacción en ambas oficinas en marzo (antes) y en mayo (después). ¿Por qué este podría ser un mal diseño de DiD? ¿Qué supuesto fundamental es difícil de justificar entre dos equipos y ciudades tan diferentes?
5. **Interpretando la regresión de DiD (Conceptual):** En el modelo de regresión $Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 (D_i \times T_t) + \epsilon_{it}$:
- ¿Qué representa el coeficiente β_1 ? (La diferencia promedio en Y entre los grupos *antes* del tratamiento).
 - ¿Qué representa el coeficiente β_2 ? (El cambio promedio en Y para el grupo de *control* a lo largo del tiempo).
 - ¿Por qué β_3 es nuestro efecto causal de interés?
6. **Violando la no anticipación (Conceptual):** El gobierno anuncia hoy que dentro de dos años, se prohibirá la venta de plásticos de un solo uso. Un investigador quiere medir el impacto de esta ley en las ventas de los productores de plástico usando un diseño DiD. ¿Por qué el supuesto de “no anticipación” está claramente violado aquí? ¿Cómo podrían las empresas cambiar su comportamiento *antes* de que la ley entre en vigor?
7. **Tu propio análisis DiD (Código):** Usando los datos de Berkeley, realiza un análisis ligeramente distinto. Esta vez, el “tratamiento” será estar físicamente en Berkeley.
- Crea la variable `PostTax` (1 si es Ene-2015 o después).
 - Crea la variable `Berkeley` (1 si `location` es “Berkeley”).

- Corre una regresión de `price` en función de `Berkeley`, `PostTax`, y la interacción `Berkeley * PostTax`, incluyendo efectos fijos de tiempo (`+ C(fecha)`).
 - Interpreta el coeficiente del término de interacción. ¿Qué te dice sobre el cambio de precios en Berkeley después de la fecha del impuesto, en comparación con las otras áreas?
8. **La prueba de placebo (Conceptual):** Una forma de ganar confianza en un resultado de DiD es hacer una “prueba de placebo”. Imagina que en el estudio del impuesto a las bebidas, finges que el impuesto se implementó un año *antes*, en Enero de 2014, cuando en realidad no pasó nada. Corres tu análisis DiD usando esta fecha falsa. Si tu modelo original es robusto, ¿qué resultado esperarías para el coeficiente de interacción en esta prueba de placebo? ¿Un valor grande y significativo, o uno cercano a cero y no significativo? ¿Por qué?
9. **DiD y Efectos Fijos (Conceptual):** Explica con tus propias palabras la relación entre un modelo de Diferencias en Diferencias y un modelo de Efectos Fijos de Dos Vías. ¿Son lo mismo? ¿O uno es un caso especial del otro?
10. **Reto - Encuentra tu propio experimento natural:** Piensa en un evento reciente (la apertura de una nueva línea de metro en tu ciudad, la legalización de un producto en un estado pero no en el vecino, el lanzamiento de una nueva función en una app que solo un grupo de usuarios recibió al principio). Describe cómo usarías un diseño de DiD para estudiar su impacto. Define claramente tu grupo de tratamiento, tu grupo de control, el periodo antes/después y la variable de resultado que te interesaría medir.

La única forma de crear valor es iterando a partir de los datos

Failure isn't a necessary evil. In fact, it isn't evil at all. It is a necessary consequence of doing something new.

– Ed Catmull

En 2011, durante mi primer semestre en la maestría en Economía, me decidí inscribir a un evento: Startup Weekend.

El evento empezaba un viernes por la tarde. El reto era crear un proyecto de negocio durante el fin de semana y presentarlo en Domingo. Era la primera vez que hacía algo así junto a personas tan talentosas: habían programadores, diseñadores y con perfil de negocios, cada persona más extraordinaria que la siguiente.

En la pared colgaba una lona que decía “No talk, all action”.

Me quedé enganchado con esa idea. La filosofía detrás de estas cuatro palabras era no pasar demasiado tiempo planeando y comenzar a ejecutar lo antes posible. La mejor forma de saber si una idea funciona es poniéndola a prueba.

Las metodologías de negocios basadas en experimentos

2008 fue un año muy intenso en la historia de la humanidad.

Lo recuerdo bien porque fue el año en el que salí de la universidad. En mis últimos semestres tomaba una clase donde el profesor nos explicaba en tiempo real lo que estaba pasando en los mercados financieros y cómo el mundo caía en una inminente recesión. Todo eso mientras me preparaba para salir de mi licenciatura.

Salí a un mundo muy diferente al que existía mientras estudié la licenciatura.

Ese mismo año, Eric Ries publicó el libro “The *Lean Startup*”. Era una filosofía diferente a la creación de negocios, que permitía a las empresas adaptarse en tiempos cambiantes, no importa su tamaño. Su adopción fue inmediata en el mundo de las startups de tecnología.

¿En qué consiste esta metodología?

La idea central es **fracasa rápido, fracasa barato**. En otras palabras, comienza tu negocio asumiendo que en tu primera versión vas a fracasar. El producto no es lo que el cliente buscaba, te equivocaste en el canal para promocionarlo, el nicho de mercado no era en mejor. Todas estas cosas pueden y van a pasar. Tu trabajo es identificarlas al menor costo posible y cambiar rápidamente para encontrar el mejor ajuste entre producto y mercado.

Y nadie mejor para enseñarnos cómo se hace que un monje inglés del siglo XVIII.

Bayes y las bolas de Billar

Thomas Bayes era un ministro presbiteriano, mejor conocido por el teorema que lleva su nombre.

En 1763, Bayes publicó un ensayo donde explicaba una forma diferente de hacer estadística. Su impacto fue tan grande que hay una clara distinción entre la estadística frecuentista de la Bayesiana. Mientras la estadística frecuentista interpreta la probabilidad como la frecuencia a largo plazo de un evento, la bayesiana la entiende como una medida de la creencia sobre un evento.

Imaginemos un juego de billar.

Yo le pego a la bola blanca mientras tú no estás viendo. La bola se mueve por toda la mesa hasta que se detiene en algún punto aleatorio. Mido la distancia de la orilla de la mesa a donde cayó la pelota, registro lo que medí y guardo la bola blanca.

La posición de la bola está oculta para ti, pero te voy a dar algunas pistas para que la adivines.

Ahora lancemos bolas rojas. Cada bola que lancemos va a caer en un lugar aleatorio de la mesa. Yo sólo te voy a decir si cayó a la derecha o a la izquierda de la bola que lanzamos al inicio.

¿Podrías decirme en dónde cayó la pelota?

Empieza con un conocimiento previo, pero actualízalo sobre la marcha

Nuestra hipótesis es que existe un modelo de negocios que se ajusta de manera ideal a un mercado existente.

El problema es que no hay forma de saberlo antes de hacer la prueba. No hay libro, clase o mentoría que te dé el secreto para crear un negocio innovador que genere ganancias millonarias. Si fuera conocimiento público, alguien más ya lo habría hecho y la oportunidad desaparecería.

Se llama la **hipótesis de los mercados eficientes**.

Cuando el panorama es incierto, lo mejor que podemos hacer es experimentar. Esto implica tomar los pasos del método científico y aplicarlos en todos los aspectos de nuestro modelo de negocio. Todo está sujeto a modificación, si los datos nos dice que no está funcionando.

El objetivo es lograr esto al menor costo posible.

MVP: Mínimo Producto Viable

Fracasa rápido, fracasa barato.

Un día estás platicando con tus amigos y decides aliarte con Dani para abrir una heladería. Platican por horas sobre los sabores que van a ofrecer, en dónde se van a establecer y el nombre que le van a poner. Incluso diseñaron un logo en una servilleta.

Sólo hay un problema: no tienen dinero para hacer todo eso.

La mayoría de las personas piensa que, si su idea es muy buena, podrían ir con un inversionista, mostrarle el plan de negocios y levantar el capital que necesitan. Ahora sólo queda implementar el plan y sentarse en un camastro en la playa a ver cómo la cuenta de banco se incrementa. Listo, eres millonario.

Excepto que eso nunca ha sucedido en la historia.

La realidad es que no hay plan de negocios que sobreviva a la realidad. No hay forma de tener todo listo. No importa si eres el mejor estratega de negocios del mundo, en algún momento tienes que hacer pruebas y fallar. No hay forma de evitarlo.

El truco es fallar de forma inteligente.

En nuestra heladería, nuestro sueño podría ser un local comercial en el centro comercial más grande de la ciudad. ¿Cuál sería la versión

más ligera de este modelo de negocio? Depende de lo que deseamos comprobar.

Si deseas comprobar si hay demanda por los sabores, ¿de verdad necesitamos ese local?

Un mínimo producto viable es la versión mínima con la que puedes comprobar tu hipótesis. Por ejemplo, puedes comenzar vendiendo los helados en línea. De esta manera puedes observar cuáles son los sabores que les gustan y cuáles no.

Solo hay un detalle importante: para saber lo que funciona y lo que no, hay que usar los principios de la **inferencia causal**.

Métricas y experimentos

Todo en tu modelo de negocios es sujeto a ser parte de un experimento.

En una startup, el lanzamiento del producto se hace de manera continua. En cada versión de tu modelo de negocio hay pequeñas variaciones que hacer. Cada variación implica una comparación de las métricas relevantes.

Estos son los elementos de tu modelo de negocio de los que puedes hacer variaciones relevantes, en términos generales.

- El segmento de mercado al que te diriges.
- La localización.
- Los canales por los que te publicitas.
- La estrategia de precios.
- Todas las características del producto o servicio.

Todos estos aspectos pueden afectar tus ventas, la operación de tu empresa y tu rentabilidad.

Cuidado con las métricas de vanidad

En un mundo lleno de datos, es fácil perdernos entre lo que es fácil y lo que vale la pena medir.

- Es fácil medir los followers, likes y reposts en nuestra página de redes sociales.

- Es fácil dar seguimiento al número de suscriptores de nuestro newsletter o al tráfico de la página.

Pero esas son métricas que no nos dicen mucho sobre el rendimiento real de nuestro negocio. A estas métricas les llamamos **métrica de vanidad**. Las *métricas de vanidad* se sienten bien, pero seguirlas no ayuda a hacer mejoras sustanciales en el negocio.

Por el contrario, deseas que tus métricas sean *accionables*.

Algunos ejemplos de Métricas accionables:

- Tasas de conversión
- Tasas de opt-in donde los usuarios pongan su correo en una landing page.
- Costo de adquisición de clientes (**CAC**).

No siempre es fácil seguir estas métricas.

En ocasiones es necesario hacer fórmulas complejas para darles seguimiento. En otros casos requiere de hacer registros cuidadosos que toman tiempo y recursos adicionales. Pero estas son las métricas que vale la pena seguir.

Si no estás siguiendo estas métricas, estás dejando dinero en la mesa.

Realiza cambios basados en la evidencia

Hasta el momento hemos:

- Lanzado un producto mínimo viable.
- Definiste tus métricas, evitando las métricas de vanidad.

Es momento de divertirnos.

El juego se llama “usa el método científico para definir tu negocio”. La idea es hacer modificaciones continuas a tu modelo de negocio de acuerdo a lo que la evidencia te dice.

Si la evidencia dice que los helados en cono son mejores que en vaso, elige helados en cono. Si la evidencia dice que los sabores tradicionales se venden más que los exóticos, pero los exóticos se venden más caro, puedes elegir enfocar tus sabores de acuerdo al cliente al que quieras atender.

No importa lo que digan los libros, tu tía o los “expertos en mercadotecnia”. Lo único que importa es lo que digan los datos.

Cómo hacer una empresa basada en datos

Esta metodología no sólo sirve para startups o empresas de tecnología, también se puede implementar en empresas en marcha.

Lo importante es que se sigan los siguientes principios:

- Lanzamiento continuo: En ningún momento estaremos en la versión definitiva de la empresa. Siempre hay algo que modificar.
- Fracasa rápido, fracasa barato: Si quieras probar una hipótesis, hazlo de la manera más mínima posible y con iteraciones cortas.

Así se vería en nuestro emprendimiento de helados.

- **Paso #1: Identifica tu métrica de interés.** Por ejemplo, imagina que quieras aumentar las ventas de helados en tu empresa en invierno. ¿Es posible?
- **Paso #2: Formula tu hipótesis.** ¿Qué variable crees que es clave para hacer crecer las ventas? Sabemos que es un mito que comer helado en invierno causa resfriados. Si viviéramos en un mundo de *homo economicus*, a nadie le importaría comer helado en invierno, pero si tendrían problemas con hacerlo en un lugar cerrado, porque circulan los virus (espero que hayamos aprendido algo con la pandemia!). ¿Qué pasaría si vendiéramos helados con probióticos durante el invierno?
- **Paso #3: Identifica el mínimo viable.** Lo peor que podrías hacer es crear un nueva línea de productos y hacer un gran lanzamiento. El mínimo viable en este caso sería tomar una muestra de tus clientes y ofrecerles el nuevo producto para estudiar su aceptación (o no).
- **Paso #4: Analiza tus métricas.** Digamos que para probar nuestra hipótesis, vamos a un evento de navidad y ponemos un stand donde vendemos el producto que estamos probando. Para hacer nuestro estudio más robusto, podemos poner otro stand con nuestra oferta regular. De esta manera podremos comprobar la diferencia en las ventas.
- **Paso #5: Toma acción.** Si las ventas en el stand con el producto nuevo son más altas que en el stand regular, significa que realmente hay interés en el producto novedoso.

Los libros tienen un problema: estas líneas las estás leyendo de forma lineal.

Da la impresión de que lo que te estoy diciendo es lineal, pero en realidad es cíclico. Vas a tomar acción y formular hipótesis de forma continua. Si haces esto sin parar, estás reduciendo riesgos de negocio y mejorando tus métricas de negocio.

Te lo aseguro.

Resumen del capítulo

En este capítulo, cambiamos el enfoque de los modelos estadísticos a la filosofía que los pone en acción: una metodología para construir y mejorar un negocio iterando a partir de los datos.

Lo que vimos fue la idea central de *Lean Startup* y su mantra: “fracasa rápido, fracasa barato”. Usamos la analogía de Thomas Bayes y las bolas de billar para entender el proceso: comenzamos con una creencia (nuestra idea de negocio) y usamos experimentos (lanzar un **MVP**, medir, aprender) para actualizar continuamente esa creencia y acercarnos a la verdad de lo que el mercado realmente quiere. Aprendimos a distinguir entre las métricas que solo inflan el ego (**métrica de vanidad**) y las que de verdad nos guían para tomar decisiones (accionables), como el **CAC**.

Esto es importante porque un plan de negocios no es más que una hipótesis esperando ser refutada. Las mejores herramientas de econometría son inútiles si no tienes un sistema para hacer las preguntas correctas y aprender de las respuestas de forma sistemática. Esta metodología te enseña que el fracaso no es el fin, sino una fuente valiosa de datos. Es la mentalidad que conecta la rigurosidad de la inferencia causal con la velocidad y agilidad que demandan los negocios modernos.

¿Cómo te ayuda esto? Este capítulo te da un mapa para aplicar todo lo que has aprendido. Tienes un ciclo claro —hipótesis, MVP, medir, actuar, repetir— que puedes usar para cualquier proyecto, ya sea lanzar una empresa desde cero, un nuevo producto, o simplemente mejorar una campaña de marketing. Te enseña a no enamorarte de tu plan, sino del proceso de aprendizaje. Ahora puedes proponer no solo un análisis, sino un ‘experimento de bajo costo’ para resolver una duda de negocio, y sabes qué tipo de métricas debes observar para que la decisión sea la correcta.

Iterando hacia el éxito: ejercicios de estrategia

La teoría es fácil, pero la verdadera lección de este capítulo está en la acción. Estos ejercicios son para que pienses como un estratega que usa los datos como su principal consejero.

1. **De la Gran Idea al MVP (Conceptual):** Tu sueño es abrir una cafetería de especialidad con un ambiente único, granos de café exóticos y un espacio de coworking. Es un proyecto grande y caro. Describe un **MVP** para esta idea. ¿Cuál es la versión más simple y barata que podrías lanzar en un mes para probar la hipótesis más importante: “la gente de mi barrio está dispuesta a pagar más por un café de alta calidad”?

2. **Vanidad vs. Acción (Conceptual):** Para tu proyecto de cafetería, clasifica las siguientes métricas como de Vanidad o Accionables y justifica por qué:
 - a) El número de seguidores en la cuenta de Instagram de la cafetería.
 - b) El número de clientes que regresan por segunda vez en una semana.
 - c) El costo de los anuncios en redes sociales dividido entre el número de clientes nuevos que esos anuncios generaron ([CAC](#)).
 - d) El número de ‘likes’ en la foto de un latte art.
3. **El Ciclo de Iteración (Conceptual):** Tu MVP de la cafetería (ejercicio 1) fue un carrito de café por las mañanas. Los datos muestran que vendes mucho café, pero casi nadie compra tus granos exóticos de \$30 la bolsa; la mayoría pide el café de la casa. Ahora estás en el paso de “Tomar acción”. ¿Cuál sería la siguiente ‘iteración’ o ‘experimento’ lógico para seguir aprendiendo sobre tu mercado?
4. **La Analogía Bayesiana (Conceptual):** Explica la metodología de ‘Lean Startup’ usando la analogía de las bolas de billar de Thomas Bayes.
 - ¿Qué representa la “bola blanca” cuya posición es desconocida?
 - ¿Qué representan las “bolas rojas” que lanzas a la mesa?
 - ¿Qué es la “información” que obtienes de cada bola roja?
 - ¿Cómo “actualizas tu creencia” sobre la posición de la bola blanca con esa información?
5. **Diseñando un Experimento de Producto (Conceptual):** Tienes una panadería y quieres probar si una nueva línea de pan de masa madre sin gluten tendría éxito. En lugar de reformar toda tu panadería (el plan a largo plazo), diseña un [MVP](#). ¿Cómo podrías probar esta hipótesis en un solo fin de semana y con una inversión mínima? ¿Qué métrica accionable medirías para decidir si la idea tiene futuro?
6. **“No talk, all action” (Reto Personal):** Piensa en un pequeño proyecto o idea que hayas tenido y dejado en el tintero. Describe el MVP más simple que podrías lanzar en una semana para dejar de planear y empezar a actuar y obtener datos reales.
7. **Pivatar o Perseverar (Conceptual):** Tu MVP del pan sin gluten (ejercicio 5) revela algo curioso: casi nadie compró el pan, pero docenas de personas te preguntaron si podías venderles la masa madre activa para hacer pan en casa. En la jerga de ‘Lean Startup’, esta es una señal para “pivotar”. Explica qué significa y cuál sería la nueva hipótesis principal de tu negocio.
8. **Midiendo el CAC (Cálculo Simple):** Para promover tu nuevo servicio de masa madre, inviertes \$100 en un anuncio en un blog de cocina. Ese anuncio genera 40 clics a tu página. De esos 40 visitantes, 10 te dejan su correo electrónico. De esos 10 correos, 2 te hacen una compra de \$15 cada uno. ¿Cuál es tu Costo de Adquisición de Cliente ([CAC](#))? ¿Fue rentable esta campaña?
9. **Los Mercados Eficientes y la Experimentación (Conceptual):** El capítulo menciona la [hipótesis de los mercados eficientes](#). Explica con tus propias palabras por qué esta idea económica refuerza la necesidad de experimentar en lugar de solo buscar una “fórmula secreta” para un negocio.

10. **El Fracaso es Información (Conceptual):** La cita de Ed Catmull dice que el fracaso “es una consecuencia necesaria de hacer algo nuevo”. ¿Cómo se conecta esta idea con el proceso de iteración? ¿Por qué un experimento que “falla” (es decir, que prueba que tu hipótesis era incorrecta) es increíblemente valioso?

Cómo hacer Investigación de mercados con inteligencia artificial

What you call love was invented by guys like me to sell nylons

– Don Draper

Todos los negocios necesitan hacer investigación de mercado.

Imagínate que te subes a un avión y el piloto te dice que los monitores de vuelo están descompuestos. No hay comunicación con las torres de control, no hay forma de saber cómo será el clima en el trayecto y no hay forma de saber si vas en la ruta correcta y llegarás al aeropuerto correcto (o a algún aeropuerto siquiera). ¿Te subes?

¡Por supuesto que no!

La investigación de mercado es la forma en la que un negocio obtiene información sobre su entorno. Puede ser de manera formal (con encuestas, entrevistas, **focus-group** y observación de las métricas) o informal (la observación simple). Todas las empresas la tienen, porque todas reciben información de su entorno.

Son nuestros monitores de vuelo.

Cuando una marca de pasteles instantáneos decidió quitarle ingredientes a su producto y ¡vendieron más que nunca

A veces no hacer nada es hacerlo todo

- Teniente Harina

En los años 50s toda la economía en Estados Unidos estaba creciendo.

Cuando la economía está creciendo, tener un producto medianamente innovador era garantía de que te harías millonario. A pesar de

eso, General Mills no lograba levantar las ventas de su mezcla para hacer pastel en casa. Era sin duda un producto innovador

Lo que hicieron a continuación, convirtió a las mezclas en un éxito y revolucionó el mercado para siempre.

La necesidad era real: las amas de casa estaban hambrientas de soluciones que les ahorren tiempo de las tareas del hogar. Así que los ejecutivos decidieron contratar a Ernest Dichter para hacer una **investigación de mercado**. Dichter fue el pionero de los **grupos de enfoque**, que juntan en un sólo cuarto a un grupo de amas de casa para encontrar las verdaderas causas del problema.

Tras entrevistar a más de cien amas de casa, Dichter encontró que la razón por la que no compraban era **culpa**. ¡Ellas querían participar en la elaboración del pastel! Un producto “demasiado fácil” de preparar las hacía ver cómo flojas y no cómo amas de casa de verdad (recuerda que eran los años 50).

La solución: quitar ingredientes a la mezcla y dejar que las amas de casa aporten más al proceso.

Ahora las amas de casa debían incluir huevo y leche a la mezcla para que quedara completa. Para la empresa esto significaba menos costos, pero más importante es que las ventas por fin crecieron. La mezcla era ahora más valiosa, porque permitía hacer un pastel, ser parte del proceso y terminarlo con facilidad.

Así de importante es tener un buen conocimiento del mercado.

Pero ¿Un estudio de mercado es sólo hacer preguntas?

Si, pero no es fácil hacer las preguntas correctas.

Para hacer las preguntas correctas necesitas entender el problema que estás solucionando, al cliente para quien estás haciendo tu solución y cómo estás contribuyendo a la solución. El problema y tu solución son importantes porque ayudan a identificar lo que debes cambiar para hacer tu oferta más valiosa, pero esas son cosas que están bajo tu control. Lo que no está control es tu cliente y sus decisiones.

Hay cosas de tu cliente que puedes aprender observando, pero otras se las vas a tener que preguntar directamente.

- ¿Por qué te eligió en lugar de la competencia?
- ¿Cómo usa tu producto / servicio?

- ¿Percibe algún riesgo de que tu producto no cumpla su promesa?

En realidad el cielo es el límite.

Puedes preguntarle lo que se te ocurra a tus clientes. El problema es que la mayoría de nosotros no sabemos diseñar las preguntas correctas y los clientes no tienen ningún incentivo para decírnos la verdad. Hace un par de años, si no tenías el don de diseñar encuestas o no tenías a tu disposición a un buen sociólogo, mala suerte.

Pero hoy podemos apoyarnos de la IA para crear una encuesta increíble.

Usemos a chatGPT para diseñar una encuesta en el estilo del Mom Test

En 2013, Rob Fitzpatrick publicó “The Mom Test”.

En este libro, explicaba cómo cuando diseñamos una pregunta para una investigación de mercado, solemos mezclar objetivos y acabamos queriendo publicitar nuestro producto o servicio. El resultado es que hacemos las preguntas “amañadas” para que nos den una respuesta positiva. Pero eso no es útil para nosotros: queremos que si alguien odia nuestro producto, nos lo diga con todo lujo de detalle.

En otras palabras, quieres que incluso tu mamá te diga si hay algo en tu producto que necesitas cambiar.

El libro está lleno de ejemplos, geniales. Si los estudias, puedes transformar las preguntas en tu estudio de mercado para que cumplan con el objetivo y mejores la calidad de tus respuestas.

O puedes decirle a chatGPT que haga esa modificación por ti.

Estos son mis pasos:

- Define lo que quieras conocer de tus clientes.
- Redacta o elige las preguntas que quieras hacerles.
- Transformalas al estilo del mom test.

Hagamos un ejemplo.

Define lo que quieres conocer de tus clientes

Digamos que tienes una e-commerce de ropa y quieres identificar puntos de fricción en el proceso de compra. En este caso hay miles

de elementos que puedes probar por tu cuenta haciendo una [prueba A/B](#). Pero también hay aspectos que no puedes observar y que tienes que únicamente puedes saber si se los preguntas directamente al cliente.

Tres ejemplos de esto serían:

- **La calidad percibida del producto.** ¿Por qué no la puedes medir con los datos de interacción? Como es una tienda en línea, hay detalles que las fotos no pueden mostrar. Algunos clientes lo podrían poner en las reseñas, pero muchos probablemente sólo dejen de comprar. Adios cliente y adiós datos.
- **Experiencia de soporte al cliente.** Esto incluye aspectos como la facilidad del menú, la rapidez y utilidad de las respuestas y la satisfacción general.
- **La experiencia del checkout.** Hay aspectos como qué tan fácil o difícil fue usar el carrito de compra, si les gustaría incluir opciones de pago o la claridad en costos de envío.

Puse ejemplos de cosas que no podrías medir usando pruebas A/B, pero tal vez también deseas investigar aspectos que se pueden medir en pruebas A/B usando una encuesta.

La razón principal de esto es que el tiempo es limitado y sólo puedes hacer pruebas limitadas a la vez. Muchos expertos recomiendan sólo usar una prueba a la vez, pero los modelos de panel y de diferencias en diferencias en este libro te pueden ayudar a hacer más. Aún con esto, tal vez no quieras usar tu tiempo y recursos de las pruebas A/B en hacer estudios triviales.

Redacta o elige las preguntas que quieres hacer

Comencemos definiendo lo que queremos conocer.

Usa el siguiente prompt en chatGPT o el mejor modelo de inteligencia artificial al que tengas acceso.

Ayúdame a hacer las preguntas para un estudio de mercado.
Giro: e-commerce de ropa. Mercado meta: Gen-Z de Mexico y latinoamerica. Propuesta unica de valor: Te damos puntos por tu ropa vieja que te dan descuentos en la nueva. Estamos luchando para combatir el fast fashion.

Te daré información sobre lo que quiero conocer de ellos y me darás 15 opciones de preguntas tocando los temas siguientes:

- La calidad percibida del producto. - Experiencia de soporte al cliente. Incluye aspectos como la facilidad del menú, la rapidez y utilidad de las respuestas y la satisfacción general. - La experiencia del checkout.
- Hay aspectos como qué tan fácil o difícil fue usar el carrito de compra, si les gustaría incluir opciones de pago o la claridad en costos de envío.

Normalmente deseamos que la extensión de nuestro estudio sea suficientemente larga para que nos de información valiosa, pero no tan larga para que la contesten hasta acabarla (con sinceridad). Por eso le pedimos 15 preguntas, pero seleccionamos sólo las mejores. También toma en cuenta que la IA es muy buena para inventar, pero no todo lo que te de serán preguntas de calidad.

Modifica el prompt anterior con los datos de tu negocio o emprendimiento y sigamos.

El mom test, explicado

Te voy a ahorrar leer un libro entero en este texto de 3 minutos.

El mom test es una técnica donde las preguntas se diseñan para que no te contesten lo que piensan que quieras oír. Digamos que queremos preguntar cómo califican la calidad de la ropa en nuestra e-commerce. Hay muchas razones por las que alguien que conteste el cuestionario no quiere contestarte con la verdad¹⁰⁷.

- Tal vez sienten que decir algo negativo sería muy confrontativo.
- Tal vez realmente quieren agradarte.
- O simplemente les estás haciendo una pregunta en la que tienen que pensar demasiado su respuesta y te acaban contestando lo que sea.

En todos estos casos, el culpable no es el que responde, eres tú.

Para entender mejor el mom test, observa estas preguntas. Son ejemplos de lo que no se debe hacer. ¿Puedes identificar por qué son malas preguntas?

- ¿Crees que {X} es buena idea de negocio?

¹⁰⁷ Rob Fitzpatrick. *The Mom Test: How to Talk to Customers and Learn If Your Business Is a Good Idea When Everyone Is Lying to You*. CreateSpace Independent Publishing Platform, 2013. ISBN 978-1492180746

- ¿Comprarías {X}?
- ¿A qué precio te parece justo comprar {X}?

Estas preguntas están mal planteadas porque alguien que no estaría dispuesto a comprar tu producto te puede fácilmente decir que sí, es buena idea. Tú te vas feliz con la impresión de que tienes la idea más genial de negocio y al final creas algo que a nadie le interesa. *Sad to be you.*

En lugar de eso, puedes preguntar

- ¿Cuáles son los mayores problemas que enfrentas cuando haces {actividad relacionada con X}?
- Cuéntame sobre la última vez que necesitaste algo similar a {X}. ¿Qué solución utilizaste?
- ¿Qué has pagado anteriormente por productos o servicios similares a {X}? ¿Cómo decidiste que valía la pena ese precio.

La razón por la que estas preguntas son mejores es que están diseñadas para sacarle la verdad al encuestado.

Las preguntas del *mom test* se enfocan en experiencias pasadas y decisiones reales. Hacer esto refleja mejor cómo se comportaría el usuario en una situación específica. Además, son preguntas que ayudan a entender el contexto y las motivaciones del usuario.

Ahora transformemos nuestras preguntas.

Transformando nuestro estudio al estilo del mom test

Para mí, el mom test es una metodología fácil de aprender, pero difícil de implementar.

Me toma mucho tiempo y trabajo hacer esas modificaciones. Afortunadamente los modelos de lenguaje grandes tienen ya cargada en su entrenamiento la información relacionada al mom test, porque hay miles de posts de blogs sobre el tema (y probablemente el libro mismo ahí venga cargado). Lo que antes nos tomaba horas, lo podemos hacer en minutos.

Puedes hacerle un prompt sencillo:

```
| redacta las preguntas en el estilo del mom test
```

Listo.

Tu encuesta ahora genera preguntas más interesantes. Son preguntas mejor diseñadas, que te generarán respuestas de mayor calidad. Puedes modificar más tu instrumento, pidiéndole aspectos específicos, como hacerlo en una [escala de Likert](#)¹⁰⁸.

¹⁰⁸ Son de esas preguntas donde puntuas del 1 al 5 o del 1 al 7 qué tan de acuerdo estás con lo que se expresa.

Cuándo no usar datos cuantitativos

Hemos llegado al final de este libro.

Y como cierre, te quiero invitar a que **no uses lo que aprendiste aquí**. Al menos no si no tienes una buena razón de hacerlo.

En ocasiones, aprender este tipo de técnicas es emocionante y queremos aplicarlo cuanto antes y usarlo en todos los problemas que se nos presentan. Pero eso es como si yo te presentara un martillo y tú quisieras usarlo en todo tipo de problemas. Incluso en aquellos que claramente necesitan un desarmador.

Parece que te estoy diciendo algo obvio, pero lo he visto en más de una ocasión.

No todos los estudios de mercado necesitan diseñar una entrevista y aplicarla a un diseño muestral estadísticamente significativo. Algunas cosas simplemente requieren que salgas al mundo y le preguntes a las personas sobre tu producto. La técnica que te enseñé arriba te puede ayudar a diseñar preguntas de tipo [Mom Test](#), pero no hay nada que sustituya salir al mundo real y preguntarle a las personas reales cómo se sienten.

La tecnología sigue avanzando, y las técnicas que podemos usar de inferencia causal seguirán ampliándose.

Con el avance de la IA, tendremos más implementaciones interesantes. Procesar la información de video, de audio, transcripciones y emociones abrirán la puerta a mucha investigación innovadora. Aprender a usar python y la IA nunca ha sido tan importante.

Ya diste el primer paso. Y por eso, te felicito.

Resumen del capítulo

En este capítulo final, cerramos el círculo. Vimos que, a pesar de todo el poder de los modelos que hemos aprendido, a veces la respuesta más profunda no está en los números, sino en una buena conversación.

Lo que hicimos fue explorar la investigación de mercados como el sistema de navegación de un negocio. La historia de la mezcla para pastel de General Mills nos enseñó una lección inolvidable: a veces, el problema no es el producto, sino una emoción humana (la culpa) que ninguna métrica cuantitativa puede detectar. Aprendimos la filosofía del *Mom Test* para diseñar preguntas que nos den la verdad cruda en lugar de cumplidos inútiles, enfocándonos en experiencias pasadas en vez de opiniones hipotéticas. Y lo más importante, descubrimos cómo usar la Inteligencia Artificial como un asistente experto para diseñar encuestas y entrevistas de alta calidad en minutos.

Esto es importante porque un negocio es, en esencia, una relación con sus clientes. Si solo miras los datos cuantitativos, solo estás escuchando la mitad de la historia. Entender las motivaciones, frustraciones y el contexto de tus clientes —el ‘insight’ cualitativo— es lo que te permite crear valor verdadero. Este capítulo es el recordatorio crucial de que la inferencia causal no se trata solo de modelos, sino de una búsqueda honesta de la verdad, y a veces, el camino más corto es simplemente preguntar de la manera correcta.

¿Cómo te ayuda esto? Ahora tienes una herramienta para investigar el “porqué” detrás del “qué” que te muestran tus datos. Cuando tus métricas te señalen un problema (por ejemplo, “la gente abandona el carrito de compra aquí”), tienes un método para descubrir la causa raíz hablando con tus clientes. Puedes usar la IA para generar rápidamente una guía de entrevista que te dará ‘insights’ mucho más profundos y honestos. Esta es la pieza que conecta tu rigor analítico con la empatía humana, dándote una visión completa para tomar mejores decisiones y, al final, construir algo que a la gente realmente le importe.

La conversación final: ejercicios de investigación y cierre

Has llegado al final. Estos últimos ejercicios son para solidificar el arte de hacer buenas preguntas y para reflexionar sobre el camino recorrido.

1. **Identificando malas preguntas (Conceptual):** Un amigo está desarrollando una app para meditar y le pregunta a sus potenciales usuarios: “¿Crees que una app que te ayude a reducir el estrés con meditaciones guiadas es una buena idea?”. Usando los principios del *Mom Test*, explica por qué esta pregunta, aunque bien intencionada, es inútil para validar el negocio.
2. **Reformulando al estilo ‘Mom Test’ (Conceptual):** Reescribe la pregunta del ejercicio anterior. Formula tres preguntas alternativas al estilo ‘Mom Test’ que le darían a tu amigo información mucho más valiosa. (Pista: enfócate en el problema y en el comportamiento pasado. Por ejemplo: “Cuéntame sobre la última vez que te sentiste estresado, ¿qué hiciste al respecto?”).
3. **Tu socio IA (Práctico):** Elige un producto o servicio que uses con frecuencia. Imagina que tu objetivo es descubrir “qué es lo que más frustra a los usuarios avanzados de este producto”. Sigue el flujo de trabajo del capítulo:
 - Escribe un ‘prompt’ para ChatGPT (o similar) pidiéndole 10 preguntas iniciales para una encuesta sobre este tema.
 - Escribe un segundo ‘prompt’ pidiéndole que transforme esas 10 preguntas al estilo del ‘Mom Test’.
 - Comparte la mejor pregunta “antes” y su versión “después” del ‘Mom Test’.

4. **La lección del pastel (Conceptual):** La historia de la mezcla para pastel de General Mills es un caso clásico de investigación de mercados. ¿Cuál era el ‘insight’ cualitativo clave que ninguna prueba A/B o análisis de ventas les estaba mostrando?
5. **Cuantitativo y Cualitativo (Conceptual):** Estás analizando los datos de una plataforma de cursos en línea y observas un dato cuantitativo: “El 80 % de los usuarios que compran un curso nunca lo terminan”. Ahora, formula tres preguntas cualitativas (estilo ‘Mom Test’) que le harías a esos usuarios para entender el “porqué” detrás de ese dato.
6. **“No hacer nada es hacerlo todo” (Conceptual):** El capítulo cita al Teniente Harina. ¿Cómo se aplica esta frase a la solución que encontró General Mills? ¿En qué sentido “quitarle” algo al producto fue la mejor decisión que pudieron tomar?
7. **Saber cuándo guardar el martillo (Reflexión):** El capítulo termina con la advertencia de no usar estas técnicas para todo. Describe un problema de negocio donde correr un modelo de efectos fijos sería la herramienta ‘incorrecta’ o excesiva, y donde una simple conversación con cinco clientes sería infinitamente más útil.
8. **Tu propio ‘Mom Test’ (Práctico/Reto):** Piensa en un proyecto real en el que estés trabajando (en tu empleo, estudios o un proyecto personal). Define una hipótesis clave que tengas sobre tus “clientes” o “usuarios”. Escribe 3 preguntas al estilo ‘Mom Test’ que podrías hacerle a alguien mañana mismo para empezar a validar esa hipótesis.
9. **Conectando los puntos (Reflexión):** ¿Cómo se conecta la filosofía de ‘iteración’ del capítulo anterior con la ‘investigación de mercados’ de este capítulo? ¿En qué parte del ciclo “hipótesis, MVP, medir, actuar” pondrías las entrevistas a clientes?
10. **Tu siguiente paso (Reflexión Personal):** Has llegado al final. ¡Felicidades! Este libro fue el primer paso. ¿Cuál es el concepto o la habilidad más importante que te llevas? Y más importante aún, ¿cuál será el ‘siguiente paso’ que darás para seguir practicando y aprendiendo sobre inferencia causal, Python e IA?

Agradecimientos

No basé mi carrera en tener hits.

Tengo hits porque yo senté las base'

- ROSALÍA

No se puede escribir un libro si no es rodeado de una red de apoyo enorme.

Quiero dedicar esta página a darle su espacio a tantos a quienes agradezco y que han sido clave para la realización de esta obra.

En primer lugar están mi esposa María y mis hijos Román y Natalia, por su paciencia y por su apoyo. Al iniciar este proyecto, me pude robar el mejor espacio de la casa para hacerlo mi estudio y tuve la dicha que poder compartir momentos valiosos en familia en paralelo a la escritura de este texto. Masha en particular ha sido fuente de inspiración. Cada vez que por fin entiendo algo más sobre cómo funcionan y cómo hacer crecer los negocios, encuentro ejemplos de Masha ya aplicándolos.

Agradezco a los directivos de mi facultad. Desde el momento en el que comencé este proyecto a que se imprimió hubo un cambio de cuerpo directivo, pero en todo momento he sentido todo el respaldo y el apoyo. En su momento el Dr. José Ramón Duarte Carranza me dio sin ninguna duda ni miramientos el permiso para tomar mi sabático y embarcarme en este aventura, y ahora recibo con mucho aprecio el respaldo del Dr. Jesús Sotelo Asef para su publicación.

Tengo en la FECA uno de los trabajos más gratificantes que podría pedir, y agradezco de todo corazón a todos mis amigos y compañeros que me han acompañado en la Facultad. En particular a mis compañeros y amigos del cuerpo académico: César, Paco, Nacho, Julieta y Brenda, que me atrevo a mencionar ya sólo por nombre por la amistad que se ha forjado en los años de trabajo conjunto. Más allá de las paredes de la facultad, agradezco a la editorial UJED, con especial énfasis en Manuel: todos sabemos la labor titánica que haces por la

editorial y por la facultad. Gracias también a los revisores anónimos, cuyas observaciones mejoraron sustancialmente este trabajo. Y un agradecimiento especial a Alina, que puso todo su corazón para que la portada quedara fantástica.

Pero uno de los más grandes agradecimientos es a mis alumnos de cuarto semestre de la carrera de Licenciado en Economía y Negocios Internacionales. Por años, los alumnos de ese grupo han sido mis conejillos de indias y me han permitido probar conceptos de una forma y otra. Este libro nació a partir de la clase de Aplicación de Principios Económicos, que desapareció del plan de estudios, pero que yo usé como una especie de pre-econometría.

Glosario

prueba A/B Experimento controlado en el que se comparan dos variantes (A y B) que difieren en un solo elemento; la asignación aleatoria de usuarios permite estimar el efecto causal de la variante sobre un KPI (p.ej., tasa de conversión). [44, 146](#)

ACF Autocorrelation Function: gráfico de las autocorrelaciones muestrales en distintos rezagos. [103](#)

AIC Akaike Information Criterion: medida de selección de modelos que penaliza la complejidad; menor AIC \Rightarrow mejor ajuste/parcimonia. [99](#)

ATT Average Treatment on the Treated (tratamiento promedio en los tratados): $E[Y_1 - Y_0 \mid D = 1]$, el efecto causal medio para las unidades que efectivamente recibieron el tratamiento. [37, 125](#)

BIC Bayesian Information Criterion: versión más restrictiva que el AIC para comparar modelos. [99](#)

BLUE Siglas de Best Linear Unbiased Estimator; afirma que, bajo los supuestos de Gauss-Márkov, OLS es el estimador lineal insesgado con menor varianza. [64](#)

CAC Customer Acquisition Cost. Costo total de captar un nuevo cliente, calculado dividiendo la inversión en marketing y ventas entre el número de clientes adquiridos en un periodo. [139](#)

caminata aleatoria Proceso $X_t = X_{t-1} + \varepsilon_t$ donde ε_t es ruido blanco; ejemplo clásico de serie no estacionaria. [86](#)

contrafactual Resultado potencial que *no* se observa en la realidad; describe lo que habría ocurrido bajo un estado alternativo del mundo. [31, 37](#)

correlación vs. causalidad Distinción fundamental entre asociaciones estadísticas (correlación) y relaciones de causa–efecto (causalidad);

confundirlas conduce a inferencias erróneas. [19](#)

DataFrame Estructura tabular de pandas con índice y columnas etiquetadas; análoga a una hoja de cálculo pero optimizada para operaciones vectorizadas. [26](#)

datos en panel Conjunto de observaciones $\{Y_{it}\}$ que combina dimensión transversal (i) y temporal (t) para los mismos individuos o unidades. [109](#)

prueba Dickey–Fuller (ADF) Contrastación estadística para la hipótesis de raíz unitaria; rechazar la hipótesis nula implica estacionariedad. [107](#)

diferencia de la serie Transformación $\Delta X_t = X_t - X_{t-1}$ (o de orden superior) usada para inducir estacionariedad. [88](#)

diferencias en diferencias Estrategia cuasi-experimental que identifica un efecto causal comparando la variación temporal de un grupo tratado con la de un grupo de control. [125](#)

diferencias en diferencias Diseño cuasi-experimental que estima un efecto causal al comparar la evolución temporal de un grupo tratado y uno de control, bajo el supuesto de tendencias paralelas. [19](#)

efectos fijos Método de estimación que controla inobservables constantes dentro de cada unidad (o periodo) añadiendo una constante específica; se basa en la variación *within*. [112](#)

efectos fijos de dos vías Modelo que introduce efectos fijos *por entidad* (U_i) y *por tiempo* (U_t) simultáneamente. [120](#)

endogeneidad Situación en la que un regresor está correlacionado con el término de error, violando el supuesto de exogeneidad y sesgando las estimaciones. [81](#)

error estándar Estimación de la desviación típica de la distribución muestral de un coeficiente; mide la precisión de la estimación. [75](#)

estacionariedad Propiedad de un proceso estocástico cuyas características estadísticas (media, varianza y autocovarianza) permanecen constantes en el tiempo. [87](#)

estadístico F Contrasta la hipótesis nula de que todos los coeficientes (excepto la constante) son cero contra la alternativa de que al menos uno es distinto de cero. [74](#)

Eviews Software comercial orientado a análisis econométrico y pronóstico; integra interfaz gráfica para estimar modelos, ejecutar scripts y generar reportes de series de tiempo. [21](#)

exceso de muertes Diferencia entre el número observado de fallecimientos en un periodo y el número esperado según patrones históricos; indicador para evaluar el efecto causal de crisis sanitarias, desastres o políticas públicas. [14](#)

experimento ideal Diseño hipotético sin restricciones prácticas ni éticas que ayuda a clarificar el mecanismo causal y la estrategia de identificación cuando un ensayo real no es factible. [40, 46](#)

experimento natural Situación del mundo real en la que eventos, políticas o cambios institucionales generan una asignación *casi aleatoria* de un tratamiento, lo que permite estimar efectos causales sin un ensayo controlado tradicional. [13, 126](#)

grupo de enfoque Técnica cualitativa que reúne a varias personas bajo la guía de un moderador para explorar actitudes, opiniones y motivaciones acerca de un producto o idea. [143](#)

teorema de Gauss–Márkov Resultado que demuestra que, si se cumplen cinco supuestos (linealidad, muestreo aleatorio, ausencia de colinealidad perfecta, exogeneidad y homoscedasticidad), OLS es BLUE. [64](#)

gold standard experimental Expresión que alude al ensayo controlado aleatorio como referencia máxima para identificar efectos causales, pues elimina sesgos mediante aleatorización y control. [44](#)

Google Colab Servicio gratuito en la nube que ofrece cuadernos Jupyter con GPU/CPU preconfiguradas y librerías de ciencia de datos; permite ejecutar código Python sin instalar nada en local. [23, 29](#)

grupo de control Conjunto de unidades que no recibe el tratamiento en un experimento o quasi-experimento; permite aproximar el resultado contrafactual y aislar el efecto causal. [39](#)

grupo de tratamiento Conjunto de unidades que recibe la intervención o variante bajo estudio; se compara con el grupo de control para calcular el efecto causal. [49](#)

grupo de tratamiento Conjunto de unidades que reciben la intervención o condición cuyo efecto causal se desea estimar. [36](#)

heteroscedasticidad Caso opuesto a la homoscedasticidad; la varianza del error cambia con el nivel de los regresores, violando un supuesto clave de OLS. [71](#)

homoscedasticidad Propiedad según la cual la varianza de los errores es constante para todos los valores de los regresores. [70](#)

inferencia causal Disciplina que estudia cómo identificar y cuantificar relaciones de causa–efecto a partir de datos observacionales o experimentales, combinando teoría estadística y supuestos sobre el proceso generador de datos. [15](#)

inteligencia artificial (IA) Campo de la informática que desarrolla sistemas capaces de realizar tareas que normalmente requieren inteligencia humana; en ciencia de datos incluye aprendizaje automático y modelos generativos que hoy asisten la programación y el análisis econométrico. [21, 49](#)

término de interacción Producto de dos variables (p. ej. $D_i \times T_t$) que captura un efecto conjunto; en DiD su coeficiente es el estimador causal. [132](#)

intervalo de confianza Rango que, con una probabilidad pre-especificada (p. ej. 95 %), contiene el verdadero valor del parámetro poblacional. [76](#)

investigación de mercado Proceso sistemático de obtención, organización y análisis de información relevante sobre consumidores, competidores y entorno, para apoyar la toma de decisiones de negocio. [144](#)

Lean Startup Metodología de creación de negocios de Eric Ries basada en ciclos rápidos “construir–medir–aprender”, experimentación continua y aprendizaje validado para reducir riesgo e inversión. [136](#)

escala de Likert Formato de pregunta cerrado en el que el encuestado indica su nivel de acuerdo con una afirmación en una gradación ordinal (p. ej. 1 – 5 o 1 – 7). [149](#)

media condicional cero Supuesto clave que exige que el valor esperado del error sea cero condicional a los regresores ($E[\varepsilon|X] = 0$); garantiza exogeneidad. [81](#)

hipótesis de los mercados eficientes Proposición de que toda la información relevante se incorpora instantáneamente a los precios de

mercado; de ser cierta, las oportunidades de ganancia “fáciles” se eliminan de forma inmediata. [137](#)

mínimos cuadrados ordinarios (OLS) Método que estima los coeficientes de la regresión lineal minimizando la suma de los residuos al cuadrado. [57](#)

modelo autorregresivo AR(p) Modelo lineal que explica Y_t a partir de sus p rezagos: $Y_t = \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t$. [98](#)

modelo ARIMA(p, d, q) Extensión del ARMA a series no estacionarias mediante d diferencias (Integración). [105](#)

modelo ARMA(p, q) Combina componentes AR(p) y MA(q) en un solo esquema: $Y_t = \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$. [102](#)

modelo de medias móviles MA(q) Modelo que expresa Y_t como combinación lineal de los rezagos del error: $Y_t = \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t$. [102](#)

Mom Test Metodología de Rob Fitzpatrick para formular preguntas que obliguen al entrevistado (incluso tu madre) a dar retroalimentación honesta basada en hechos y no en halagos. [149](#)

multicolinealidad Situación en la que dos o más regresores están altamente correlacionados, provocando estimaciones inestables y errores estándar inflados. [67](#)

MVP Sigla de *mínimo producto viable*. Versión más sencilla de un producto o servicio que permite poner a prueba la propuesta de valor con el mínimo tiempo y coste posibles, obteniendo retroalimentación temprana del mercado. [137](#)

supuesto de no anticipación Los resultados potenciales previos al tratamiento no se ven afectados por la futura asignación de éste. [126](#)

notebook (cuaderno Jupyter) Documento interactivo que combina texto, código ejecutable y resultados; facilita el análisis econométrico iterativo y reproducible. [23](#)

NumPy Paquete fundamental para computación numérica en Python; ofrece arrays multidimensionales y rutinas de álgebra lineal que subyacen a la mayoría de los modelos econométricos. [24](#)

orden de rezago Número de retardos (p en AR, q en MA) incorporados al modelo. [96](#)

PACF Partial Autocorrelation Function: muestra las autocorrelaciones

condicionales; útil para fijar el orden p de un modelo AR. [103](#)

pandas Biblioteca de Python que provee estructuras de datos como *Series* y *DataFrame*; pilar para la limpieza, transformación y explotación de datos tabulares en econometría. [24](#)

prueba de poolabilidad Contrastación (F-test) para decidir si un modelo agrupado sin efectos fijos es apropiado frente a un modelo con efectos fijos. [119](#)

pruebas de hipótesis Procedimientos estadísticos que contrastan afirmaciones sobre parámetros mediante un estadístico y una regla de decisión; en inferencia causal se usan para evaluar la significancia de los efectos estimados. [19](#)

p-value Probabilidad de obtener un estadístico tan extremo como el observado si la hipótesis nula fuera cierta; valores pequeños sugieren rechazar H_0 . [75](#)

Python Lenguaje de programación de propósito general, ampliamente usado en ciencia de datos y econometría gracias a su sintaxis sencilla y a un ecosistema extenso de bibliotecas estadísticas. [21](#)

R Lenguaje de programación y entorno especializado en estadística y visualización de datos; base de muchos paquetes económicos (p. ej. *plm*, *fixest*) y punto de referencia para análisis reproducible. [21](#)

raíz unitaria Raíz del polinomio autoregresivo igual a 1; su presencia indica no-estacionariedad. [93](#)

ensayo controlado aleatorizado (RCT) Diseño experimental en el que las unidades se asignan al tratamiento o al control mediante un procedimiento aleatorio, eliminando el sesgo de selección y permitiendo estimar efectos causales de manera creíble. [39](#)

R^2 Medida de ajuste que indica la proporción de la variación de la variable dependiente explicada por el modelo; varía entre 0 y 1. [79](#)

regresión lineal Modelo estadístico que relaciona una variable dependiente con una o más variables independientes mediante una función lineal; punto de partida clásico para estimar efectos causales cuando se cumplen sus supuestos. [19](#), [53](#), [91](#)

residuales Diferencias entre los valores observados y los predichos por el modelo; se emplean para diagnosticar los supuestos de la regresión. [77](#)

resultados potenciales Marco teórico que asocia a cada unidad dos posibles resultados: el que observaría si recibe el tratamiento (Y_1) y el que observaría si no lo recibe (Y_0). Sólo uno se observa en la práctica. [31, 33, 35, 37, 39, 41](#)

ruido blanco Secuencia de innovaciones $\{\varepsilon_t\}$ con media cero, varianza constante finita y nula correlación serial. [86](#)

serie de tiempo Secuencia ordenada $\{X_t\}$ de observaciones de una variable aleatoria a lo largo del tiempo. [85](#)

sesgo de selección Distorsión que surge cuando la asignación al tratamiento no es aleatoria y está correlacionada con los resultados potenciales, de modo que la comparación directa entre grupos produce estimaciones sesgadas. [35](#)

Stata Paquete estadístico propietario muy popular en econometría aplicada; destaca por su sintaxis compacta y amplia colección de comandos para modelos lineales, paneles y series de tiempo. [21](#)

statsmodels Biblioteca de Python orientada a econometría y estadística; incluye regresiones lineales, modelos de series de tiempo y pruebas de hipótesis. [24](#)

SUTVA Stable Unit Treatment Value Assumption: no-interferencia entre unidades y único valor de tratamiento. [125](#)

tendencias paralelas Supuesto clave de DiD: en ausencia de tratamiento, la evolución promedio del resultado sería igual en ambos grupos. [127](#)

métrica de vanidad Indicador que “luce” bien (p. ej. followers, likes) pero no orienta la toma de decisiones porque no refleja directamente la salud o el crecimiento económico del negocio. [139](#)

variable aleatoria Función que asigna valores numéricos a los resultados de un experimento aleatorio; concepto base para describir distribuciones y estimadores en econometría causal. [20](#)

variable indicadora (dummy) Variable binaria que toma valor 1 cuando la unidad pertenece a un grupo (p. ej. tratamiento) y 0 en caso contrario; útil para codificar categorías en modelos econométricos. [34](#)

factor de inflación de la varianza (VIF) Índice que cuantifica cuánto aumenta la varianza del coeficiente de un regresor debido a la mul-

ticolinealidad; valores superiores a 5–10 indican posible problema.
[79](#), [80](#)

transformación within Centrado de cada variable respecto a su media por individuo: $\ddot{X}_{it} = X_{it} - \bar{X}_i$; elimina el término constante U_i de los efectos fijos. [116](#)

Bibliografía

Spider-man: No way home, 2021.

J. D. Angrist and J. S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Pierre Azoulay, Benjamin F. Jones, J. Daniel Kim, and Javier Miranda. Age and high-growth entrepreneurship. *American Economic Review: Insights*, 2(1):65–82, 2020. DOI: [10.1257/aeri.20180582](https://doi.org/10.1257/aeri.20180582).

Gary S. Becker. Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5, Part 2):9–49, 1962. DOI: [10.1086/258724](https://doi.org/10.1086/258724). URL <https://www.journals.uchicago.edu/doi/10.1086/258724>.

Dale S. Berg and Alan B. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 2002.

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *NBER*, 2003. DOI: [10.3386/w9873](https://doi.org/10.3386/w9873).

T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1):155–174, 2015. URL <http://www.jstor.org/stable/43616924>.

Paul Busch, Teiko Heinonen, and Pekka Lahti. Heisenberg's uncertainty principle. *Physics Reports*, 452(6):155–176, 2007. DOI: [10.1016/j.physrep.2007.05.006](https://doi.org/10.1016/j.physrep.2007.05.006).

Raymundo M. Campos-Vazquez, Gerardo Esquivel, Parama Ghosh, and Enrique Medina-Cortina. Long-lasting effects of a depressed labor market: Evidence from mexico after the great recession. *Labour Economics*, 81:102332, 2023. DOI: [10.1016/j.labeco.2023.102332](https://doi.org/10.1016/j.labeco.2023.102332). URL <https://doi.org/10.1016/j.labeco.2023.102332>.

David Card. The impact of the mariel boatlift on the miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257, 1990.

David Card. The causal effect of education on earnings. In *Handbook of Labor Economics*, volume 3, pages 1801–1863. Elsevier, 1999.

David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4):772–793, 1994.

M. Arantxa Colchero, Mariana Molina, and Carlos M. Guerrero-López. After mexico implemented a tax, purchases of sugar-sweetened beverages decreased and water increased: Difference by place of residence, household composition, and income level. *Journal of Nutrition*, 147(8):1552–1557, 2017. DOI: 10.3945/jn.117.251892.

Derek A. T. Cummings, Rafael A. Irizarry, Norden E. Huang, Timothy P. Endy, Ananda Nisalak, Kumnuan Ungchusak, and Donald S. Burke. Travelling waves in the occurrence of dengue haemorrhagic fever in thailand. *Nature*, 427(6972):344–347, 2004. DOI: 10.1038/nature02225. URL <https://pubmed.ncbi.nlm.nih.gov/14737166/>.

David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431, 1979.

Paul Embrechts, Alexander J. McNeil, and Daniel Straumann. Correlation and dependence in risk management: Properties and pitfalls. In M. A. H. Dempster, editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511615337.008.

Rob Fitzpatrick. *The Mom Test: How to Talk to Customers and Learn If Your Business Is a Good Idea When Everyone Is Lying to You*. CreateSpace Independent Publishing Platform, 2013. ISBN 978-1492180746.

Brian Hayes. Gauss's day of reckoning. *American Scientist*, 94(3): 200–204, 2006. DOI: 10.1511/2006.59.200. URL <https://www.americanscientist.org/article/gausss-day-of-reckoning>.

James J. Heckman, Lance J. Lochner, and Petra E. Todd. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Econometrics*, 144(1):306–348, 2008.

Norden Huang and Nii O. Attoh-Okine, editors. *The Hilbert-Huang Transform in Engineering*. Taylor & Francis Group, USA, 2005. ISBN 978-0-8493-3422-1.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in*

- R. Springer Texts in Statistics. Springer, 1 edition, 2013. ISBN 978-1-4614-7137-0. DOI: [10.1007/978-1-4614-7138-7](https://doi.org/10.1007/978-1-4614-7138-7). URL <https://www.statlearning.com/>.
- Jingxuan Liu, Chenshuo Xia, Yihan Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023. URL <https://arxiv.org/abs/2305.01210>.
- Gladys Lopez-Acevedo. Mexico: Two decades of the evolution of education and inequality. *World Bank Policy Research Working Paper*, (3919), 2006.
- Edward Miguel and Michael Kremer. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217, 2003. DOI: [10.1111/j.1468-0262.2004.00481.x](https://doi.org/10.1111/j.1468-0262.2004.00481.x).
- A. C. Pigou. *The Economics of Welfare*. Macmillan and Co., London, 1920. URL <https://archive.org/details/dli.bengal.10689.4260>.
- Aloys Leo Prinz. Chocolate consumption and noble laureates. *Social Sciences & Humanities Open*, 2(1):100082, 2020. ISSN 2590-2911. DOI: <https://doi.org/10.1016/j.ssh.2020.100082>. URL <https://www.sciencedirect.com/science/article/pii/S2590291120300711>.
- César Pérez López. *Muestreo estadístico: conceptos y problemas resueltos*. Pearson Education S.A., Madrid, 2005.
- Jesse Rothstein. The lost generation? labor market outcomes for post great recession entrants. *NBER Papers*, 2020. DOI: [10.3386/w27516](https://doi.org/10.3386/w27516). URL <https://www.nber.org/papers/w27516>.
- Lynn D. Silver, Shu Wen Ng, Suzanne Ryan-Ibarra, Lindsey Smith Taillie, Marta Induni, Donna R. Miles, Jennifer M. Poti, and Barry M. Popkin. Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in berkeley, california, us: A before-and-after study. *PLOS Medicine*, 14(4):e1002283, 2017. DOI: [10.1371/journal.pmed.1002283](https://doi.org/10.1371/journal.pmed.1002283).
- John Snow. *On the Mode of Communication of Cholera*. John Churchill, London, 1849. URL <https://archive.org/details/b28985266/page/n3/mode/2up>. Accessed: 2025-04-23.
- Annie Gracey Swenson. *Medical Women of America: A Short History of the Pioneer Medical Women of America and a Few of Their Colleagues in England*. Monarch Book Company, Chicago, 1910. URL <https://www.archive.org/details/medicalwomendo00swenrich>.

//archive.org/details/medicalwomenvict00swen_0/page/n6/
mode/1up.

Jacob Wallace, Paul Goldsmith-Pinkham, and Jason L. Schwartz. Excess death rates for republican and democratic registered voters in florida and ohio during the covid-19 pandemic. *JAMA Internal Medicine*, 183(9):916–923, 2023. DOI: 10.1001/jamainternmed.2023.1154. URL <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2807617>.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, corrected 2nd printing edition, 2004. ISBN 978-0387402727.