



INVESTIGACIÓN APLICADA
a las Ciencias sociales
en Stata

 Anáhuaconline


COMUNICACIÓN
CIENTÍFICA

Jorge Alberto Pérez Cruz
Rolando Israel Valdez Ramírez
Fortino Vela Peón

Investigación aplicada a las Ciencias Sociales en Stata

Jorge Alberto Pérez Cruz
Rolando Israel Valdez Ramírez
Fortino Vela Peón



Ediciones Comunicación Científica se especializa en la publicación de conocimiento científico de calidad en español e inglés en soporte de libro impreso y digital en las áreas de humanidades, ciencias sociales y ciencias exactas. Guía su criterio de publicación cumpliendo con las prácticas internacionales: dictaminación de pares ciegos externos, autenticación antiplagio, comités y ética editorial, acceso abierto, métricas, campañas de promoción, distribución impresa y digital, transparencia editorial e indexación internacional.

Cada libro de la Colección Ciencia e Investigación es evaluado para su publicación mediante el sistema de dictaminación de pares externos y autenticación antiplagio. Invitamos a ver el proceso de dictaminación transparentado, así como la consulta del libro en Acceso Abierto.



[DOI.ORG/10.52501/cc.155](https://doi.org/10.52501/cc.155)



Investigación aplicada a las Ciencias Sociales en Stata

Jorge Alberto Pérez Cruz
Rolando Israel Valdez Ramírez
Fortino Vela Peón



Pérez Cruz, Jorge Alberto

Investigación aplicada a las ciencias sociales STATA / Jorge Alberto Pérez Cruz, Rolando Israel Valdez Ramírez, Fortino Vela Peón .— Ciudad de México: Comunicación Científica, 2025. (Colección Ciencia e Investigación).

207 páginas : gráficas, ilustraciones ; 23 × 16.5 centímetros.

DOI: 10.52501/cc.155

ISBN: 978-607-9104-78-8

1. Ciencias sociales – Métodos estadísticos – Programas para computadora. 2. Estadística – Programas para computadora. I. Valdez Ramírez, Rolando Israel, coautor. II. Vela Peón, Fortino, coautor. I. Título

LC: HA32 P47

DEWEY: 300.285555 P47

La titularidad de los derechos patrimoniales y morales de esta obra pertenece a las autoras D.R. © Jorge Alberto Pérez Cruz, Rolando Israel Valdez Ramírez y Fortino Vela Peón, 2025. Reservados todos los derechos conforme a la Ley. Su uso se rige por una licencia Creative Commons BY-NC-ND 4.0 Internacional, <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode.es>

Primera edición en Ediciones Comunicación Científica, 2025

Diseño de portada: Francisco Zeledón • Interiores: Guillermo Huerta

Ediciones Comunicación Científica, S. A. de C. V., 2025

Av. Insurgentes Sur 1602, piso 4, suite 400,

Crédito Constructor, Benito Juárez, 03940, Ciudad de México,

Tel.: (52) 55-5696-6541 • Móvil: (52) 55-4516-2170

info@comunicacion-cientifica.com • www.comunicacion-cientifica.com

 comunicacioncientificapublicaciones  @ ComunidadCient2

ISBN 978-607-9104-78-8

DOI: 10.52501/cc.155



Esta obra fue dictaminada mediante el sistema de pares ciegos externos.
El proceso transparentado puede consultarse, así como el libro en acceso abierto,
en <https://doi.org/10.52501/cc.155>

Índice

<i>Introducción</i>	9
1. Análisis de regresión lineal y su metodología	13
2. Análisis sobre las personas que presentan SARS-CoV-2 en México: enero a septiembre 2020.	41
3. Análisis del ingreso a través de cuantiles.	87
4. El modelo de Solow con datos de corte transversal	101
5. La utilidad de los modelos de suavizamiento exponencial en el pronóstico de la curva de infectados por COVID-19 en México. . . .	119
6. Un modelo con variable dependiente trunca en combinación con la técnica de diferencia en diferencias para evaluar una política pública	147
7. Modelo de datos panel para estimar el papel de la educación en la desigualdad por ingresos.	155
8. Modelo de ecuaciones estructurales con variables latentes: análisis de la satisfacción en programas sociales	175
<i>Referencias</i>	203
Sobre los autores.	205

Resumen

Este libro está diseñado para estudiantes, profesionistas, investigadores, y todo aquel interesado en el uso de Stata para el análisis de datos. En los distintos capítulos mostramos los procesos sistemáticos que implican desarrollar e implementar análisis de estadística descriptiva, el uso de gráficos y la estimación de modelos de regresión lineal y logística, a través de aplicaciones en donde se emplearán estructuras de datos como corte transversal, series de tiempo, y de panel, el cual brinda la posibilidad de utilizarse a través de pantallas o comandos, además de que brinda la posibilidad de programar con un lenguaje sencillo. La selección de las aplicaciones responde a los principales retos que enfrentan en la actualidad los profesionistas e investigadores.

Palabras clave

Ciencias sociales, Métodos estadísticos, Programas para computadora, Estadística Programas para computadora.

Introducción

En la actualidad, hemos observado que la cantidad de datos que se generan sobre prácticamente todo lo que nos rodea crece a un ritmo acelerado y gran parte de estos se encuentran disponibles en línea, lo que hace que estén al alcance para todos aquellos que tengan interés por realizar análisis cuantitativo, ya sea desde una perspectiva de la estadística descriptiva o inferencial. Esta forma tan vertiginosa en que se genera la información implica una serie de retos:

- En los estudios o trabajos de investigación que presentan datos estadísticos de encuestas, censos, registros administrativos, cuentas nacionales, entre otros, rápidamente pierden vigencia, por lo que se requiere constante actualización.
- Las bases de datos cada vez son más extensas, debido a que la disponibilidad de datos, a lo largo del tiempo, ha aumentado al igual que el conjunto de variables que se emplean en los análisis; además, en el caso de información a nivel local y regional se ha observado que ha crecido.
- En los modelos de regresión, donde se analiza un fenómeno específico, generalmente son más los factores que intervienen en su explicación; en ese sentido, se ha convertido en una necesidad de incorporar un mayor número de variables para identificar sus impactos.
- Se han diversificado las formas en que se puede representar un fenómeno, por ejemplo, cuando se habla de violencia social, se puede manejar

en términos de delitos denunciados, medidas de percepción, intensidad de la violencia, entre otras mediciones; sin embargo, los métodos de estimación también se han diversificado, de tal forma que existen múltiples métodos para explicar tan solo un fenómeno.

- En la selección de información relevante para ciertos grupos de la población, así como la forma de presentarla para que pueda ser comprendida por la población que muestra interés en el tema. Por ello, las infografías han tomado relevancia por lo simple y entendible que pueden llegar a ser para la mayoría de la población.

Sin duda, existen más retos que afrontar en este escenario tan dinámico y globalizado de la información. La capacidad de adaptabilidad a este entorno demanda al analista cuantitativo el conocimiento en métodos de estimación, el manejo de bases de datos, la construcción de variables, el uso de software especializado y capacidad analítica para la toma de decisiones.

Este libro está diseñado para estudiantes, profesionistas, investigadores, y todo aquel interesado en el uso de Stata para el análisis de datos. En los distintos capítulos mostramos los procesos sistemáticos que implican desarrollar e implementar análisis de estadística descriptiva, el uso de gráficos y la estimación de modelos de regresión lineal y logística, a través de aplicaciones en donde se emplearán estructuras de datos como corte transversal, series de tiempo, y de panel, el cual brinda la posibilidad de utilizarse a través de pantallas o comandos, además de que brinda la posibilidad de programar con un lenguaje sencillo. La selección de las aplicaciones responde a los principales retos que enfrentan en la actualidad los profesionistas e investigadores.

De esta forma, en este libro se tendrá la oportunidad de conocer y aplicar los comandos más importantes empleados para la construcción de indicadores, elaboración de gráficas, estimación y pronóstico de modelos. Es importante mencionar que, con respecto a la parte de la estimación y pronósticos, se desarrolla el análisis considerando la metodología en el contexto de regresión lineal, partiendo de elementos teóricos que sustentan la relación de variables hasta las pruebas de validación de los modelos. Cabe señalar que este libro guía al lector de forma pragmática en el análisis de regresión, absteniéndose del desarrollo matemático, y enfocándose princi-

palmente en cada uno de los procedimientos que requieren realizarse en el análisis estadístico y de regresión a través de los comandos del programa Stata y en la interpretación de cada uno de los resultados obtenidos de dicho procedimiento.

Si bien cada capítulo del presente libro puede estudiarse por separado, también puede entenderse como un todo, ya que se encuentra estructurado yendo de lo más elemental a lo complejo. La lógica detrás de la organización de los capítulos se centra en la estructura de datos que se trata en cada capítulo, siendo los primeros de corte transversal, después con series de tiempo, datos agrupados, y, finalmente, datos de panel.

Por otro lado, los temas que se incluyen en el libro son variados con el objetivo de captar la atención de varias disciplinas de las ciencias sociales, en particular los principales temas que se abordan son de economía, políticas públicas, política social.

Para cada uno de los temas se emplean distintos métodos y técnicas de estimación. Uno de los métodos más recurridos es el de Mínimos Cuadrados Ordinarios (MCO) que se utiliza en el capítulo cuatro con una estructura de corte transversal, el método Máxima Verosimilitud (MV) se implementa en el capítulo seis para estimar un modelo con variable dependiente trunca, y en el capítulo ocho para realizar la estimación de un modelo de ecuaciones estructurales. En el capítulo cinco se muestran distintos métodos de suavización de series de tiempo, mientras que en el capítulo siete, dentro de la metodología de datos panel, se emplea el método de Mínimos Cuadrados Generalizados y Mínimos Cuadrados Generalizados Factibles.

1. Análisis de regresión lineal y su metodología

Representación de la ecuación y forma gráfica del modelo de regresión lineal

El uso de la estadística para el análisis de datos puede partir desde cuadros que presentan a una variable o un conjunto de estas ofreciendo medidas de tendencia central o de dispersión, así como representaciones gráficas, pruebas de hipótesis y la asociación de variables, donde en esta última el interés se centra en establecer el tipo de relación que existe entre ellas, así como los impactos que se generan entre una variable que se identifica como dependiente y otra u otras que se consideran independientes. Existe una diversidad de formas de realizar el análisis entre un par, o más, de variables, sin embargo, cuando se pretende construir un modelo que vincule la relación entre un conjunto de variables, la forma tradicional de realizarlo es a través del análisis de regresión.

En particular, el análisis de regresión lineal (ARL) permite comprender la forma y la intensidad en la que se vinculan un par, o más, de variables, y es a través de la estimación de una ecuación que se puede representar esa relación de la forma simple de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (1)$$

o en su versión múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i \quad (2)$$

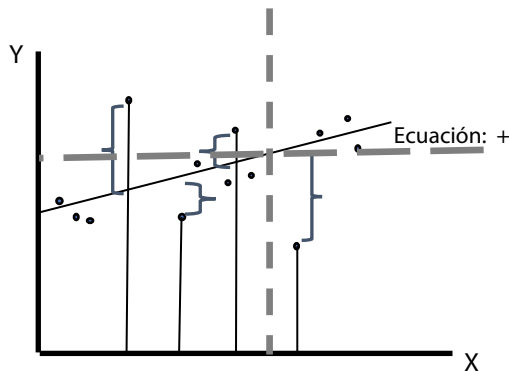
En el caso simple solamente existe una variable independiente (exógena) representada por la variable X , y esta explica a la variable dependiente (endógena) Y ; la variable u representa el término estocástico o de error, y que contempla aquellas variables que, en cierto grado, pueden explicar a la variable dependiente, pero que resulta prácticamente imposible medirlas. Los parámetros representados por las beta son los coeficientes que se estiman por medio del método de Mínimos Cuadrados Ordinarios (MCO) u otros métodos que podrían aplicarse en respuesta al incumplimiento de los supuestos del método de MCO, estos parámetros representan los valores resultados de la relación que guardan las variables en el modelo. El parámetro β_0 representa el intercepto del modelo, y su valor constituye el valor promedio de la variable dependiente, de tal forma que cuando el valor de la variable independiente es igual a cero, el valor de la variable dependiente será igual a β_0 . En lo correspondiente al parámetro β_1 representa el valor que dimensiona la relación entre la variable dependiente e independiente, este valor se conoce como el valor de impacto entre X y Y .

En el caso de la ecuación, en su versión múltiple, el número de variables independientes (exógenas) que explican a la variable dependiente (endógena) se incrementan a un número definido de K variables. La diferencia con la ecuación simple son los diferentes β_k que acompañan a cada una de las variables independientes, en donde cada uno de estos dimensiona la relación de impacto que tiene cada variable X con la variable dependiente Y .

En este sentido, el propósito en el ARL es encontrar la ecuación que permita conocer la relación de las variables determinada por los valores de los coeficientes beta. Los valores de los coeficientes se obtienen empleando el método de MCO. Recibe este nombre debido a que este método de estimación calcula el valor de los coeficientes beta considerando que la distancia que existe entre el valor que estima la ecuación calculada, con los valores observados, sea la mínima posible, el término cuadrático hace referencia a que en su forma lineal la suma de los valores de esa distancia sería igual a cero, es por ello por lo que deben considerarse en forma cuadrática, igual como sucede con el cálculo de varianza. En la gráfica 1.1 se ilustra, de forma general, la estimación de la ecuación entre la variable Y y X , en donde cada uno de los puntos representa los valores correspondientes a esas variables. Las líneas punteadas son las medias de cada una de las variables, de tal

forma que en la intercepción de estas medias cruzará la recta de regresión. Esta recta se obtiene una vez que se han estimado tanto el valor del intercepto β_0 como de la pendiente β_1 (en el gráfico se presentan ambos coeficientes con un acento circunflejo comúnmente denominado como el símbolo de gorro $\hat{}$, que significa que son los valores estimados). Sustituyendo los valores de X en la ecuación de regresión estimada se obtendrán los valores de la variable dependiente que el modelo predice. Graficando los valores estimados de Y y los valores originales de X se obtiene la recta de regresión que se presenta en la gráfica 1.1. Como se observará en este gráfico, la variable de error no aparece en la ecuación estimada, esto se debe a que el cálculo de los coeficientes beta estimados, deber garantizar una recta que atraviese en promedio a la mitad del conjunto de puntos entre Y y X, de tal manera que el desvío que existe entre el valor observado de Y y el valor de estimado para Y (o geométricamente, como se puede observar en la gráfica 1.1, la distancia vertical entre los dos valores de Y antes señalados), es conocido como el error del modelo. Una propiedad algebraica de los errores es que al sumarse para la totalidad del conjunto de valores de la variable dependiente de un total de cero. Esto sucede así, debido a que la distancia que existe entre los valores de Y observada y Y estimada que están por arriba de la recta de regresión es la misma distancia que existe entre esas dos variables por debajo de la recta de regresión, su única diferencia es el signo, lo que al sumarse da un total de cero.

Gráfica 1.1. Recta de regresión y errores del modelo



Fuente: Elaboración propia.

Supuestos del modelo de regresión lineal

Los supuestos que se tienen que considerar en la estimación de las beta bajo el método de MCO son los siguientes:

- El valor medio de los errores del modelo es igual a cero.
- La varianza de los errores condicionados a diferentes valores de las variables independientes es constante.
- No existe autocorrelación entre los errores del modelo.
- Ninguna de las variables independientes se aproxima a una constante.
- Se han incluido de acuerdo con la teoría las variables relevantes en el modelo.
- No existe relación perfecta entre las variables independientes.
- No existe relación entre los errores del modelo y los valores de las variables independientes.
- Los estimadores de los beta son insesgados, es decir, los valores de los coeficientes estimados son iguales a los verdaderos valores poblacionales, siempre y cuando se cumplan los supuestos de muestra aleatoria, parámetros lineales y media condicional cero.
- La varianza de los coeficientes estimados es la más pequeña entre los diversos valores posibles de estos coeficientes, siempre y cuando se cumplan los supuestos del punto anterior, más el de varianza constante.

El cumplimiento de estos supuestos es fundamental para estimar la ecuación que mejor se ajuste a los datos y que sea capaz de predecir el comportamiento de la variable que se pretende explicar. El incumplimiento de uno o algunos de los supuestos podría obligar a emplear otros métodos de estimación distintos al de MCO, que permitan estimar la ecuación de regresión que mejor represente la relación de variables. La decisión de emplear un par, o más, de variables estará determinado por un proceso que incluye desde la revisión de la teoría hasta la validación estadística y el pronóstico o construcción de escenarios a partir de la ecuación estimada.

Metodología para la estimación del modelo de regresión lineal

El punto de partida en el ARL es establecer sobre qué variables se pretende conocer su comportamiento y explicarlo, ya sea que el análisis se realice a través del tiempo o a través de las diferentes unidades de análisis; a esta variable se le denomina variable dependiente o endógena. Por ejemplo, se podría tener el interés por comprender el comportamiento del precio del petróleo, las ventas de un sector económico, la cotización del precio peso-dólar, el valor de las viviendas en una región del país, las variaciones de empleo durante alguna pandemia, los efectos económicos de algún tratado de libre comercio, entre otras variables que podrían ser de interés.

Relación de variables y causalidad

Posterior a la definición de la variable dependiente, sobre la cual se pretende explicar su comportamiento, se recurrirá a la búsqueda de modelos teóricos con el propósito de identificar aquellos factores o variables que intervienen en la explicación de la variable dependiente, o bien, en algunas otras disciplinas como la sociología, ciencia política, demografía, entre otras, se puede explorar sobre estos factores, los cuales se denominarán variables independientes o exógenas. Este aspecto es fundamental debido a que da validez a la relación que se plantea, lo que significa que desde una perspectiva teórica o exploratoria se esboza la relación causal entre la variable dependiente e independiente, así como la forma en que se vincula, es decir, podría existir una relación directa o indirecta. Cuando se omite este paso en los análisis de regresión, y simplemente se define de manera arbitraria la o las variables que podrían explicar la variable dependiente, existe el riesgo de encontrar una relación más en un sentido de casualidad que en un sentido de causalidad.¹

¹ Es necesario tener mucho cuidado sobre la noción de causalidad que se considera. La causalidad que considera el ARL es una causalidad empírica.

Un ejemplo de lo anterior se podría encontrar bajo el siguiente argumento de alguien que considera que el precio de las viviendas en la Ciudad de México está determinado por el número de perros que viven en situación de abandono en el Estado de México. Al recolectar datos de ambas variables y realizarse el ARL se concluye que, desde el punto de vista estadístico, existe una relación entre ambas variables; sin embargo, esta relación no es racional, debido a que no tiene nada que ver el número de perros en situación de calle en el Estado de México y el valor de la vivienda en la Ciudad de México, esta relación estadística existe por casualidad y no por causalidad. Este tipo de relaciones también se conocen como relaciones espurias. De esta forma, estadísticamente se podría encontrar relación de variables que en la vida práctica no tendrían sentido, por lo que resulta prioritario que en el ARL se fundamenten las relaciones de variables a través de hipótesis, razonamientos sustentados en elementos teóricos.

Cabe mencionar que, además de la revisión teórica para identificar las variables que pueden explicar a una determinada variable dependiente, existe la posibilidad de consultar revistas científicas y de difusión de actualidad que aborden el tema que se pretende analizar, con el objeto de tomarlas como referencia en relación con la forma en que se han incorporado variables independientes al modelo, así como la manera en que se miden cada una de las variables que se incluyen en el modelo.

Con la revisión teórica se identifica el conjunto de variables que se emplearán en el análisis de regresión, y en la mayoría de las ocasiones se parte de una ecuación teórica (modelo teórico) que las vincule, la cuales generalmente tendrán que ser transformadas en una versión econométrica. En caso de que no se especifique un modelo teórico, se podrá plantear directamente la relación de variables en forma general siguiendo la estructura que se presenta en las ecuaciones (1) o (2). En este paso se tendrá un modelo teórico especificado en forma de modelo econométrico. En Blanchard, Amighini y Giavazzi (2012) se plantea que el nivel de competencia y regulación en un mercado de bienes y servicios condicionará el nivel de margen de ganancias de las empresas:

Podemos imaginar que el margen depende del grado de competencia existente en el mercado de productos. Cuanto mayor es el grado de

competencia, menor es el margen y viceversa, cuanto menor es el grado de competencia, mayor es el margen [...] Por otro parte, el margen también depende del grado de regulación del mercado de productos. Para verlo, imaginemos un mercado de productos muy regulado con una gran cantidad de barreras comerciales: las barreras comerciales limitarán el número de productos extranjeros que pueden venderse en el mercado y, por tanto, reducirán el grado de competencia existente en el mercado. Por consiguiente, cuanto mayor sea el grado de regulación del mercado de productos, menor será el grado de competencia. Podemos expresarlo formulando el margen como una función positiva de la regulación del mercado de productos (*rmp*)

$$\mu = f(RMP)" \text{ (p. 164)}$$

Se plantea la relación teórica entre la regulación del mercado de productos (*RMP*) y el margen de las ganancias, e incluso se señala puntualmente el tipo de relación que se espera entre ambas variables. Siguiendo la estructura de la ecuación (1), el modelo econométrico podría quedar planteado de la siguiente manera:

$$\mu = \alpha + \beta RMP + u \quad (3)$$

De esta forma se cuenta con un modelo econométrico; sin embargo, podría no ser la forma definitiva del modelo, se tendrán que considerar otros aspectos para determinar la forma en que se incluirán las variables al modelo.

Búsqueda y selección de datos

Posterior al planteamiento general del modelo econométrico, se realizará la búsqueda de fuentes de información que proporcionen los datos que servirán de insumos para poder estimar la relación que existe entre el conjunto de variables. Estos datos pueden provenir de bases de datos oficiales o de organismos no gubernamentales, los cuales pueden ser locales, nacionales o

internacionales. Además, se pueden obtener datos de encuestas, informes o registros administrativos, siempre y cuando el conjunto de variables que se emplean coincida ya sea de manera temporal o a través de las unidades de análisis. En el caso de México, las principales fuentes de información son el Instituto Nacional de Estadística y Geografía (INEGI), el Banco de México (Banxico), la Secretaría de Hacienda y Crédito Público (SHCP), el Consejo Nacional de Evaluación de la Política de Desarrollo Social (Coneval), el Consejo Nacional de Población (Conapo), el Instituto Nacional de las Mujeres (Inmujeres) y registros administrativos del sector público. A continuación, se comparten direcciones web donde se puede tener acceso a diversas bases de datos nacionales e internacionales:

[http://www.inegi.org.mx/sistemas/bie/;](http://www.inegi.org.mx/sistemas/bie/)
[https://www.banxico.org.mx/SieInternet/;](https://www.banxico.org.mx/SieInternet/)
<https://stats.oecd.org/index.aspx;>
http://www.hacienda.gob.mx/POLITICAFINANCIERA/FINANZASPUBLICAS/Estadisticas_Oportunas_Finanzas_Publicas/Paginas/unica2.aspx;
<http://datos.bancomundial.org/pais/mexico;>
[https://www.ipums.org/;](https://www.ipums.org/)
http://estadisticas.cepal.org/cepalstat/WEB_CEPALSTAT/Portada.asp;
<http://www.coneval.org.mx/Paginas/principal.aspx;>
[https://datamexico.org/;](https://datamexico.org/)
<http://datos.imss.gob.mx/group/asegurados;>
<https://www.investing.com;>
<https://www.census.gov;>
<https://globalfinancialdata.com/global-macro-data.>

Análisis gráfico de variables

Definidos el conjunto de datos que se emplearán para representar cada una de las variables, se procede al análisis gráfico, en donde particularmente se podrían realizar dos formas gráficas básicas, sin que ello limite que otras formas gráficas puedan desarrollarse, pero al menos es importante que se

muestren las variables en forma de gráfica de dispersión con el objeto de establecer cuatro características básicas:

1. Observar *a priori* el grado de dispersión de los datos, con lo cual será posible definir gráficamente el grado de asociación de estos.
2. Establecer la forma geométrica que se construye a través de los datos.
3. Detectar e identificar puntos aberrantes o atípicos.
4. Definir el tipo de relación directa o inversa que existe entre cada par de variables.

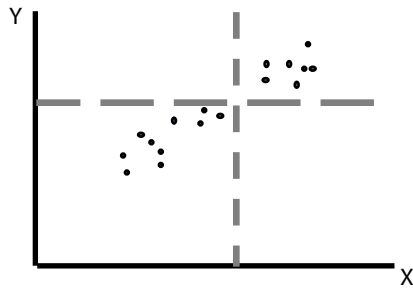
Las gráficas de dispersión se realizan considerando cada variable independiente (en el eje X) contra la variable dependiente (eje Y). Este tipo de gráfico resulta de gran apoyo debido a que proporciona información referente al tipo de relación y el grado de asociación que existe entre cada conjunto de variables que se considera incorporar en el modelo. En la gráfica 1.2 se presenta la relación entre el conjunto de datos de X y de Y, empleando como referencia los ejes cartesianos que se forman con la media de ambas variables. Aquí es posible identificar la distribución del conjunto de datos y establecer el tipo de relación entre estas variables de acuerdo con dicha distribución. Se sabe que cuando la mayoría de los puntos se ubican en los cuadrantes I y III, que se forman con las medias de X y Y, como se aprecia en la gráfica 1.2(a), se establece que la relación entre las variables es positiva o directa. Por el contrario, cuando la mayoría de los puntos se ubican en el cuadrante II y IV se establece que la relación es negativa o inversa como se aprecia en la gráfica 1.2(b). Cuando la distribución de los datos se presenta de manera uniforme entre los cuatro cuadrantes, se establece que no existe relación entre las variables, tal como se aprecia en la gráfica 1.2(c). El grado de asociación lineal de las variables dependerá de que tan cercanos a un comportamiento lineal estén los datos. En este sentido, entre más cercanos estén, más fuerte será la asociación y viceversa.

Existen algunas otras formas geométricas no lineales que se pueden formar entre el conjunto de datos de X y Y que son visibles a través de este tipo de gráficos, de esas formas se partirá para determinar la forma en que se debe introducir la o las variables independientes en el modelo de regre-

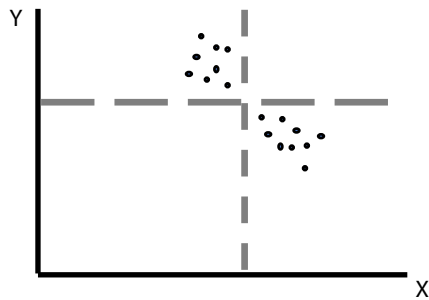
sión, considerando formas polinomiales, exponenciales, logarítmicas, entre otras más.

Gráfica 1.2. Gráfica de dispersión

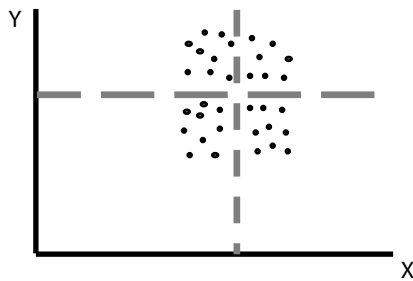
(a) Relación directa entre las variables



(b) Relación inversa entre las variables



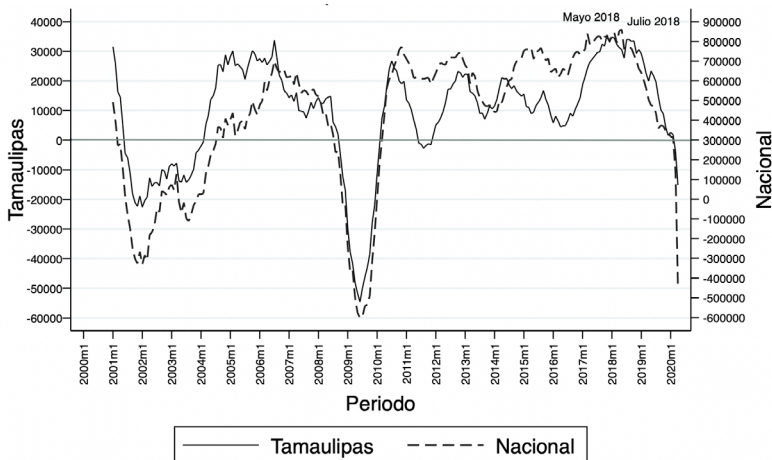
(c) Sin relación entre las variables



Fuente: Elaboración propia

La segunda gráfica básica dependerá del tipo de datos, este se sugiere que se realice para cada una de las variables (graficadas preferentemente en el eje Y), y podría ser un gráfico temporal o por unidades de análisis (graficado en el eje X), este tendrá como propósito analizar la distribución y su trayectoria. Cuando se incluya más de una variable en el gráfico, es importante tomar en consideración que la medición o escala de cada una de las variables coincida, de lo contrario, se recomienda emplear dos ejes de Y. En la figura 1.3 se presenta un ejemplo donde aparece el total de trabajadores que cuenta mensualmente con seguro social desde el año 2000 a 2020 tanto en el estado Tamaulipas, que es representado en el eje de Y del lado izquierdo del gráfico, y a nivel nacional, que presenta sus datos en el eje de Y del lado derecho. En este gráfico existe la posibilidad de realizar un análisis comparativo de dos series de tiempo, con diferentes unidades de medición, visualizando a simple vista a las series incluidas. Como se observa, la diferencia de trabajadores durante la mayoría de los periodos que se analizan se encuentra prácticamente sincronizada en sus movimientos, lo que representa que la evolución de trabajadores asegurados en Tamaulipas responde a los movimientos que se observan a nivel nacional.

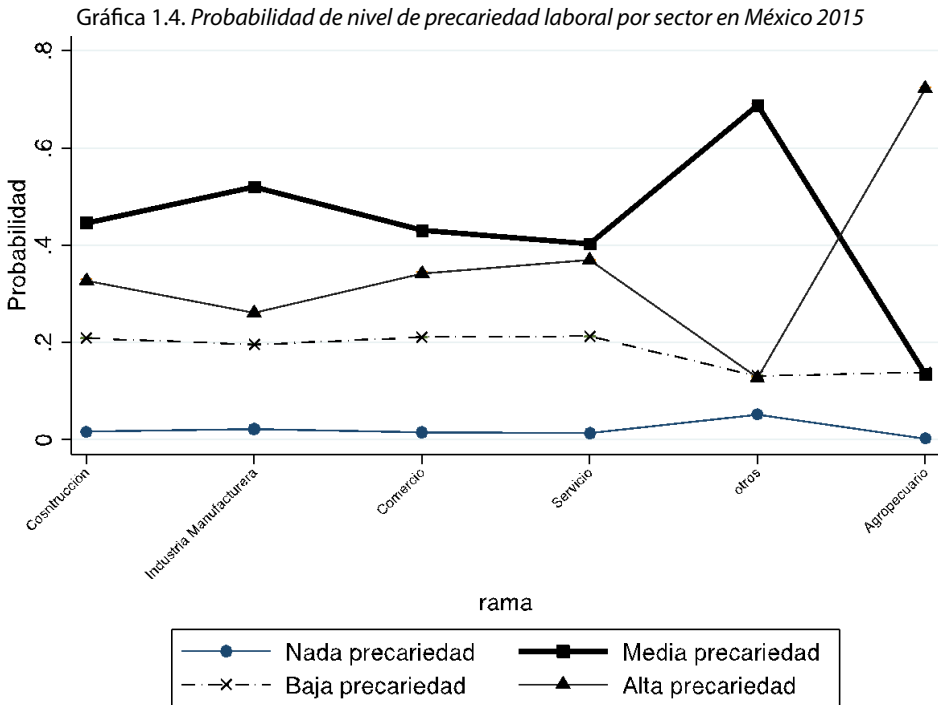
Gráfica 1.3. Diferencia anual del total de trabajadores asegurados en Tamaulipas-Nacional



Fuente: Elaboración propia con cifras del IMSS

Fuente: Elaboración propia con cifras del IMSS.

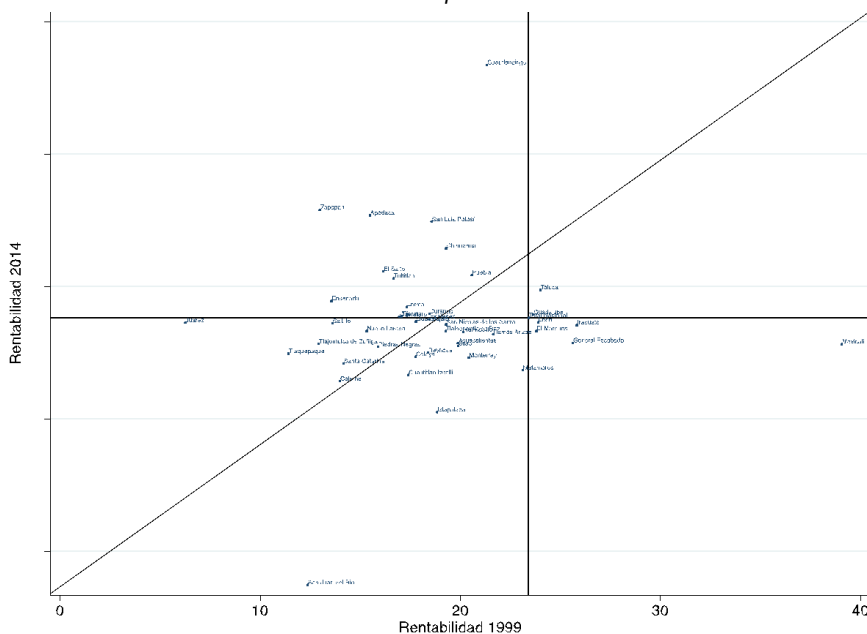
También podríamos representar los datos de una variable categórica en forma de probabilidades para diferentes unidades de análisis. En la gráfica 1.4 se representa la precariedad laboral en cuatro categorías en los principales sectores económicos de México para el periodo 2015. Cada una de las líneas representa diferentes niveles de precariedad, de tal forma que la suma de probabilidades de cada uno de los sectores en cada uno de los niveles de precariedad es igual a la unidad, en este sentido, este gráfico permite comparar la intensidad de la precariedad laboral a través de diferentes sectores económicos. El sector con mayor probabilidad de alta precariedad es el agropecuario con un valor de aproximadamente 0.71, el segundo es el sector de servicios.



Fuente: Pérez Cruz, J. A., y Ceballos Álvarez, G. I. (2022). Dimensionando la precariedad laboral en México de 2005 a 2015, a través del Modelo Logístico Ordinal Generalizado. *Nósis. Revista De Ciencias Sociales*, 28(55), 109-135. <https://doi.org/10.20983/noesis.2019.1.6>

Otra representación gráfica de amplio uso en los análisis de datos, es aquella con indicadores de diferentes unidades de análisis que se contrastan en dos periodos de tiempo, tal y como se presenta en la gráfica 1.5. Aquí se presenta el valor de la rentabilidad de 43 municipios en México en dos periodos, en el eje de las abscisas se encuentra el indicador en el año inicial 1999, y el eje de la ordenada se ubica el indicador de 2014. La recta de 45 grados representa un referente respecto a la rentabilidad para cada uno de los municipios, en donde si el punto se localiza por encima de esta recta significa que la rentabilidad para el municipio es mejor en 2014 que en 1999, como es el caso del municipio de Zapopan, Apodaca, San Luis Potosí, Chihuahua, entre otros. Cuando el punto se ubica por debajo de la línea de 45 grados, la rentabilidad en 2014 para el municipio es menor a la observada en 1999, como se observa en los casos de los municipios de Iztapalapa, Matamoros, Reynosa, Monterrey, entre otros.

Gráfica 1.5. Rentabilidad de los municipios más industrializados 1999-2014



Fuente: Elaboración propia con datos de los censos económicos publicados por el INEGI.

Otras gráficas que pueden ser de utilidad son los histogramas que brindan una orientación en relación con el tipo de distribución de probabilidad de cada una de las variables que se incorporen en los modelos. Estos gráficos son un referente para establecer la forma de distribución de las observaciones de las series, así como conocer la presencia de normalidad en las mismas. Dependiendo de los tipos de datos que se empleen, podrían realizarse gráficos específicos para inspeccionar en los datos su estructura, dispersión, tendencia, estabilidad, cambios estructurales, ciclo, estacionalidad, estacionariedad, entre otros. Cada uno de estos gráficos ofrece información relevante que permite conocer la estructura de los datos a través de las diferentes unidades de análisis y a través del tiempo.

Una vez realizada la inspección gráfica es importante realizar un análisis estadístico de los datos de cada una de las variables en donde, de manera elemental, se parte de la media, mediana, moda, varianza, distribución percentil, coeficiente de correlación, valores máximo y mínimo, coeficientes de simetría y curtosis, entre otras medidas estadísticas que contribuyan puntualmente a identificar la composición de la variable, distribución de probabilidad, trayectoria de la variable, asociación entre ellas y nivel de dispersión. Resulta práctico y conveniente como complemento del análisis gráfico.

Esta primera revisión teórica, gráfica y de medidas estadísticas, permite identificar la forma en que podrían vincularse cada una de las variables, de tal manera que, a nivel de variable, podría ser más útil representar cada una de las variables en forma polinomial o logarítmica, en diferencia, en forma de crecimiento, o cualquier otra forma de medición que permita identificar la relación entre las variables consideradas en el modelo de regresión. En ocasiones, las teorías especifican la forma de medición de la variable dependiente e independiente, tal es el caso de la teoría sobre el modelo neoclásico de crecimiento de Solow, el modelo de consumo keynesiano, los modelos de rentabilidad a la escolaridad, entre otros. En otras ocasiones es necesario explorar la forma de vinculación de las variables, pero siempre considerando elementos o hipótesis con referentes teóricos.

Transformación de variables

También el análisis gráfico y estadístico puede contribuir a decidir realizar en alguna transformación de las variables para mejorar el ajuste de los modelos de regresión. En este punto, se especifica el modelo de regresión tomando como referencia la ecuación (1) o (2), incorporando las variables de acuerdo con la medición de cada una de ellas que mejor se ajuste, desde la perspectiva teórica, gráfica y estadística. Las variables se podrían transformar en participaciones, el interés puede centrarse en determinar la proporción que representa cada una de las actividades que se desarrolla en un municipio, estado o país con respecto al total de las actividades económicas de cada una de esas actividades, esta relación se presenta en el cuadro 1.1. En este esquema de medición de la variable es posible identificar las actividades industriales que mayor presencia tienen en cada una de las unidades territoriales con respecto al total de actividades que se desarrollan en cada una de ellas. En el cuadro 1.1, este cálculo está representado por el cociente entre los valores que se interceptan en la diagonal que aparece sombreada en este cuadro, es decir, el valor de la columna de cada ámbito geográfico, entre la fila correspondiente a ese mismo ámbito geográfico.

Cuadro 1.1. *Participaciones de las actividades industriales respecto al resto de las actividades en diferentes ámbitos geográficos*

		<i>Actividad industrial (numerador)</i>		
		<i>Municipal</i>	<i>Estatad</i>	<i>Nacional</i>
Resto de actividades (denominador)	Municipal			
	Estatad			
	Nacional			

Fuente: Elaboración propia.

Otra alternativa es calcular el dato en participaciones de forma cruzada entre diferentes ámbitos geográficos, como se señala en la parte sombreada del cuadro 1.2. El objetivo es establecer la participación que representa una actividad industrial en el municipio con respecto al total de esa actividad

industrial en el estado o en el país. El análisis también se puede realizar a nivel de estado en comparación con el nivel nacional. Existen indicadores, como el coeficiente de localización que resulta de una combinación de cálculo de la forma de los cuadros 1.1 y 1.2, dependerá de los objetivos del análisis para determinar lo que resultaría más útil.

Cuadro 1.2. *Participaciones de las actividades industriales respecto a esa misma actividad en un ámbito geográfico superior.*

		Actividad industrial (numerador)		
		Municipal	Estatad	Nacional
Actividad industrial (denominador)	Municipal			
	Estatad			
	Nacional			

Fuente: Elaboración propia.

Los datos también se podrían representar en forma de tasas de crecimiento, con el propósito de identificar el comportamiento de una variable a lo largo de un determinado periodo de tiempo, para lo cual se podría emplear la fórmula de tasa de crecimiento para dos periodos:

$$\Delta Y_t = \left(\frac{Y_t - Y_{t-1}}{Y_{t-1}} \right) * 100 \quad (4)$$

En donde Y_t es el valor de la variable en el periodo t , mientras que Y_{t-1} es el valor de la variable en el periodo anterior. Otra forma de expresar la tasa de crecimiento es a través de la siguiente expresión:

$$\Delta Y_t = \left[\left(\frac{Y_t}{Y_{t-1}} \right)^{\frac{1}{r}} - 1 \right] * 100 \quad (5)$$

En donde r son los periodos que existen entre el lapso que se calcula la tasa de crecimiento, esta se recomienda cuando el cálculo se realiza en un lapso mayor a dos periodos, con este procedimiento el resultado es más preciso que el previo. Por ejemplo, cuando se calculan la tasa de crecimiento de desempleo entre el mes de enero y diciembre de 2019, de forma pro-

medio mensual, el resultado será más preciso con la fórmula (5) que la (4). En caso de que únicamente interese conocer el resultado de la tasa de crecimiento acumulado de ese periodo entonces la fórmula (4) es más pertinente que la (5).

El analista deberá explorar distintas formas de medición de las variables, revisando en libros o publicaciones de revistas para identificar la diversidad de formas en que podrían considerarse la medición de una variable que requiere ser estimada o que se incorpora en el modelo en forma de variable explicativa o independiente. Incluso, se podrían construir variables empleando simultáneamente los datos en logaritmos, transformándolos en participaciones y calculando sus tasas de crecimiento. De esta manera, la forma de plantear la variable dependerá de los propósitos del análisis, pero también teniendo en consideración la relación teórica que se pretenda probar.

Posteriormente a la definición del modelo, y a la medición de las variables, se estima el modelo de regresión a través del método de regresión que mejor se ajuste a los supuestos del modelo, a la forma de medición de las variables tanto dependientes como independientes y a la distribución de probabilidad de estas. En este apartado es importante señalar que el procedimiento de estimación dependerá desde la forma en que se mida la variable dependiente; podríamos encontrar modelos con variables dependientes binarias, ordinales, cuantitativas, variables latentes, entre otras. También dependerá de la forma de distribución de las variables, lo que permitirá establecer estimaciones con distribuciones normales, poisson, binomiales, distribución χ^2 , entre otras. En el caso donde existan problemas de exogeneidad en las variables, autocorrelación serial o espacial, sesgo de especificación u otra violación de supuestos básicos del modelo de regresión, será necesario determinar un procedimiento de estimación de regresión que permita establecer un mejor ajuste entre el conjunto de variables analizadas.

Los resultados de la estimación tendrán que validarse a través de estadísticos que permitan establecer, en principio, que la construcción del modelo es aceptable; que a nivel individual cada uno de los coeficientes que se incluyó es significativo; que la capacidad de explicación que tienen las variables independientes sobre la dependiente es adecuada; y que los resultados del modelo no violen los supuestos básicos del modelo de regresión, de lo contrario se tendrán que emplear otros métodos de estimación.

Una parte de la validación de la estimación del modelo de regresión conlleva realizar prueba de hipótesis en conjunto o a cada uno de los coeficientes β_k estimados, con el propósito de validar su inclusión en el modelo de regresión o con el propósito de identificar si la relación y la intensidad de esta entre la variable dependiente e independientes se cumple tal y como se establece desde la perspectiva teórica. Por ejemplo, si se desea probar los rendimientos crecientes, constantes o decrecientes en el modelo neoclásico de Solow, se deberá realizar prueba de hipótesis conjunta de los coeficientes estimados del modelo. Las pruebas de hipótesis podrían variar dependiendo de la técnica de estimación que se emplee, probando en cada una de estas los propósitos o intereses que pudieran surgir en cada uno de los análisis que se realicen.

Finalmente, una vez definido el modelo estimado en donde se han incorporado las variables independientes más relevantes que explican a la variable dependiente, y posterior al haber atendido la violación de supuestos del modelo regresión, es posible calcular pronósticos a partir de asignar valores a las variables independientes. Es recomendable que se acompañe el pronóstico de la variable dependiente con intervalos de confianza con el propósito de establecer los límites en que podrían variar dichos valores. Por otro lado, los resultados del pronóstico son generados reconociendo que la relación entre las variables del modelo permanece constante a lo largo del tiempo o las unidades de análisis, en circunstancias donde se observen factores que modifiquen la trayectoria de la relación de variables, existe una alta probabilidad de cometer errores en el pronóstico.

Clasificación de variables

En el ARL se pueden emplear variables que contienen información numérica, a las cuales se les denomina variables cuantitativas; este conjunto de información podría representarse como datos discretos o datos continuos. En el caso de los datos discretos, son valores numéricos enteros, algunos ejemplos de este tipo de información son accidentes vehiculares, números de personas que asisten a una función de cine, empresas que pertenecen a un sector económico, números de hijos, entre otros casos. Con relación a los

datos continuos, son todos aquellos valores que se presentan en la recta numérica, es decir, son los datos que se representan en forma decimal; los ejemplos más representativos son: altura de las personas, desempleo, precio de bien o servicio, oferta monetaria, entre otros.

Las variables que se emplean en el ARL también pueden constituirse a partir de información no numérica, por lo que sus datos representan atributos o diferenciación entre las unidades de análisis o a lo largo de tiempo; estas variables se conocen como variables cualitativas. De acuerdo con los datos que la componen, la variable cualitativa podría ser nominal u ordinal. En el primer caso, las nominales, son aquellas que contienen datos que corresponden a atributos o categorías de cada una de las unidades de análisis o de tiempo, sin que su valor presente un orden jerárquico. Ejemplos de ellas son la religión, la zona geográfica, los periodos presidenciales, las marcas de vehículos, entre otros. Por el contrario, los datos que conforman una variable cualitativa ordinal reflejan atributos o categorías en donde sus valores muestran un orden jerárquico, tal es el caso de la satisfacción del cliente, la clase social, la intensidad de violencia, los grados de precariedad, el nivel educativo, entre otros.

Tipos de datos

En los modelos de regresión lineal, además de clasificarse de acuerdo con el tipo de variable, también se pueden catalogar considerando al conjunto de datos que sirven de insumo para dimensionar el comportamiento entre las variables. De hecho, comúnmente, el tipo de datos que se utilice se asocia con la posibilidad de cierta violación de los supuestos estadísticos cuando se estiman estos modelos. En este sentido, los datos se pueden representar como datos de corte transversal, series de tiempo o como panel de datos.

En el caso de los datos de corte de transversal, estos corresponden a observaciones captadas en un momento de tiempo para distintas unidades como pueden ser personas, empresas, municipios, estados, países, marca de vehículos, actividades productivas, beneficiarios de programas públicos, entre otros. Un ejemplo de datos de corte transversal lo encontramos en el modelo donde se pretenden analizar los niveles de precariedad (np) de los

trabajadores en México para el año 2020, a través de variables como prestaciones sociales (ps), niveles educativos (ne) y tamaño de la empresa (te). En este caso, las unidades de análisis de cada una de las variables son las personas y para cada una de estas solo existe un dato, debido a que se analiza la información únicamente para 2020, entonces el número de elementos que contendrá la base de datos dependerá del número de personas que se encueste. Las variables que contienen datos de corte transversal en los modelos se denotan con el subíndice i , de tal forma que el modelo que se acaba de ejemplificar quedaría especificado de la siguiente manera:

$$np_i = \beta_0 + \beta_1 ps_i + \beta_2 ne_i + \beta_k te_i + u_i \quad (3)$$

En lo correspondiente a los datos de series de tiempo, son aquellos que se generan a lo largo de un intervalo temporal para solo una unidad de análisis. Representando el ejemplo anterior en forma de series de tiempo, se consideraría analizar la proporción de la población que presenta niveles altos de precariedad (nap) en México de forma anual durante el periodo de 1980 a 2020, siendo la proporción de la población que cuenta con prestaciones sociales (pps), la proporción de la población con nivel educativo profesional (nep) y la proporción de microempresas en México (pme) las variables que la explican. En este ejemplo, la unidad de análisis para cada una de las variables es el país y no las personas, por lo tanto, se cumple que solo existe una unidad de análisis a través de la cual se genera información de forma anual desde 1980 a 2020, lo que nos da un total de 41 datos para cada una de las variables. Para identificar variables con datos de series de tiempo, se asigna a cada una de estas el subíndice t , representando el modelo de la siguiente manera:

$$nap_t = \beta_0 + \beta_1 pps_t + \beta_2 nep_t + \beta_k pme_t + u_t \quad (4)$$

Cabe observar que a los datos de corte transversal hoy en día se les suele denotar como microdatos, mientras que los de serie de tiempo se presentan más comúnmente a nivel de agregación mayor.

Ahora bien, los datos de panel son aquellos que combinan a la dimensión transversal con la dimensión temporal de manera simultánea. Esto es,

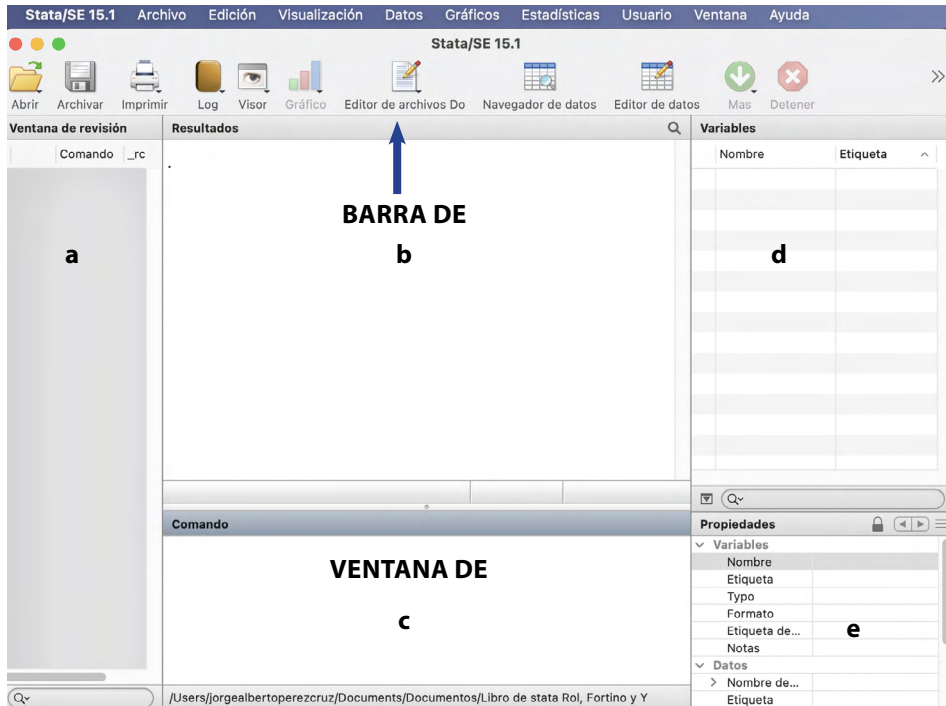
son datos que se presentan para diferentes unidades de análisis a lo largo de un periodo de tiempo. Siguiendo con el ejemplo anterior, se podría tener el interés de analizar la proporción de la población que presenta niveles altos de precariedad (*nap*) en los 43 municipios más grandes de México de forma anual durante el periodo de 2000 a 2020, donde esta será explicada por la proporción de la población que cuenta con prestaciones sociales (*pps*), la proporción de la población con nivel educativo profesional (*nep*) y la proporción de microempresas en México (*pme*) durante este periodo y para estos municipios. De tal manera que multiplicando el número de unidades de análisis (43 municipios) por el total de periodos (21 años) se tendría un total de 903 datos. Para identificar un modelo de datos de panel, a las variables del modelo se le asignan los subíndices *it*, por lo que el modelo quedaría representado como sigue:

$$nap_{it} = \beta_0 + \beta_1pps_{it} + \beta_2nep_{it} + \beta_kpme_{it} + u_{it} \quad (5)$$

Software Stata

El programa Stata es un software con el que se pueden realizar análisis estadísticos descriptivos e inferenciales a partir de bases de datos como microdatos, obtenidos por registros administrativos oficiales o bien, recolectados a través de alguna plataforma de información de la web o incluso los que podrían provenir de la aplicación de una encuesta a una población objetivo. Es un software que puede gestionarse a través de las diferentes pestañas o a través de sintaxis escrito en la venta de comandos: Para esta último es necesario conocer el o los comandos, así como sus posibles opciones. Uno de los propósitos de este libro consiste precisamente en presentar la sintaxis básica que se utiliza para calcular estadísticos, graficar, estimar modelos, los cuales se introducirán en la barra del comando. En la gráfica 1.1, se presenta la pantalla principal del software Stata versión 15.1, en la parte superior aparecen las pestañas con todas las funciones del *software*. Los principales íconos de abrir, guardar, imprimir, entre otros se ubican debajo de estas pestañas. La pantalla principal del software tradicionalmente cuenta con cinco ventanas o secciones:

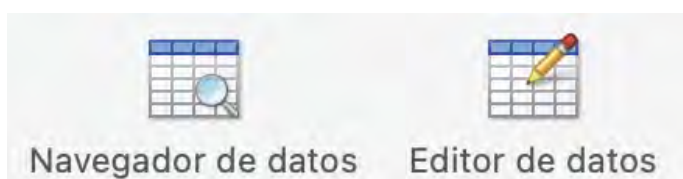
Figura 1.1. Pantalla de inicio de Stata versión 15.1



- (a) Ventana de revisión: se ubican del lado izquierdo de la figura 1.1; en esta parte aparecen el historial de las sintaxis que se han realizado desde que se abrió el programa, cuando aparezcan en rojo significa que existió algún error en la sintaxis.
- (b) Resultados: Se encuentra en el centro de la pantalla del programa, en esta sección se presentan todos los resultados estadísticos e instrucciones que se le proporcionen al programa ya sea a través de las pestañas principales o por medio de los comandos, con excepción de los gráficos, los cuales aparecerán en pantallas adicionales.
- (c) Comando: Se ubica en la parte inferior del centro de la pantalla del programa, es aquí donde se introducen las sintaxis para solicitar estadísticos descriptivos, estimaciones y gráficos.
- (d) Variables: sección ubicada en la parte central derecha del programa donde se enlistan todas las variables contenidas en la base de datos.

- (e) Propiedades: Se ubica en la parte inferior derecha, en esta parte del programa se muestran las características de cada una de las variables, así como las correspondientes a la base de datos.

Los archivos de bases de datos en formato Stata se clasifican con la extensión **dta**, aunque el programa permite importar datos provenientes de otros paquetes estadísticos almacenados con otras extensiones. Para poder acceder a la base de datos se puede hacer a través de los siguientes dos íconos que aparecen en la pantalla principal.

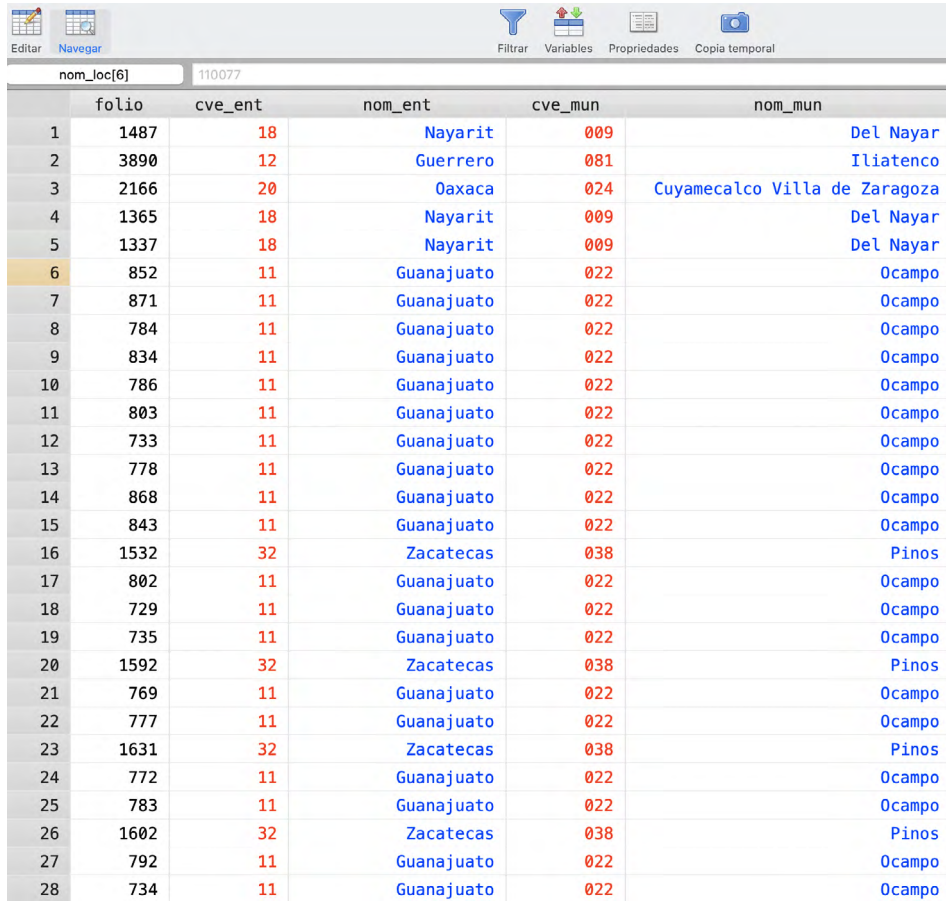


Con ambos es posible examinar los datos. La diferencia radica en que en el navegador de datos solo se puede revisar los datos sin modificarlos, mientras que en el Editor de datos se puede revisar y modificar la información contenida en cada una de las variables. Al examinar la base de datos se podrían clasificar como datos numéricos, texto y etiquetas, en el caso de los datos numéricos aparecen de color negro, las letras en color rojo representan datos no numéricos, y cuando los datos se representan en forma de etiqueta aparecen en color azul. En la figura 1.2 aparece la base de datos de los censos económicos 2014, en donde se representan los diferentes tipos de datos que se pueden emplear en Stata. En el caso de la variable **cve_ent**, aunque sus datos representan valores numéricos, el color que se le asignado a sus valores son rojos, esto significa que para el programa Stata estos valores son no numéricos y representan categorías de la variable. Los datos de la variable **folio** están representados en color negro por lo que Stata lo considera numérico. Para la variable **nom_ent** sus datos están en color azul, esto significa que en Stata se han creado etiquetas que se han asignado a esta variable.

Para los datos numéricos el tipo de almacenamiento dependerá de la extensión de número de campos que contenga cada una de las cifras que se registran en la variable, siendo *byte* el más pequeño y *double*, el de mayor

extensión; de esta forma Stata asignará automáticamente el formato de la variable dependiendo de la extensión de cada uno de los datos que contenga.

Figura 1.2. Base de datos en Stata



	folio	cve_ent	nom_ent	cve_mun	nom_mun
1	1487	18	Nayarit	009	Del Nayar
2	3890	12	Guerrero	081	Iliatenco
3	2166	20	Oaxaca	024	Cuyamecalco Villa de Zaragoza
4	1365	18	Nayarit	009	Del Nayar
5	1337	18	Nayarit	009	Del Nayar
6	852	11	Guanajuato	022	Ocampo
7	871	11	Guanajuato	022	Ocampo
8	784	11	Guanajuato	022	Ocampo
9	834	11	Guanajuato	022	Ocampo
10	786	11	Guanajuato	022	Ocampo
11	803	11	Guanajuato	022	Ocampo
12	733	11	Guanajuato	022	Ocampo
13	778	11	Guanajuato	022	Ocampo
14	868	11	Guanajuato	022	Ocampo
15	843	11	Guanajuato	022	Ocampo
16	1532	32	Zacatecas	038	Pinos
17	802	11	Guanajuato	022	Ocampo
18	729	11	Guanajuato	022	Ocampo
19	735	11	Guanajuato	022	Ocampo
20	1592	32	Zacatecas	038	Pinos
21	769	11	Guanajuato	022	Ocampo
22	777	11	Guanajuato	022	Ocampo
23	1631	32	Zacatecas	038	Pinos
24	772	11	Guanajuato	022	Ocampo
25	783	11	Guanajuato	022	Ocampo
26	1602	32	Zacatecas	038	Pinos
27	792	11	Guanajuato	022	Ocampo
28	734	11	Guanajuato	022	Ocampo

A continuación, en el cuadro 1.3 se señalan los comandos básicos que se emplean en los análisis estadísticos de los datos, los cuales se introducen en la sección de comandos de la figura 1.1.

Además de los comandos que se incluyen en la versión del *software*, existe la posibilidad de emplear algunos otros más específicos que han sido desarrollados por la comunidad de usuarios de Stata y se ponen a disposición de los usuarios para que sean aprobados por los administradores del

software. Estos comandos pueden ser revisados a través de la siguiente sintaxis que se coloca en la ventana de comandos:

ssc describe (aquí se coloca la inicial del nombre del comando sobre el cual se tiene interés)

Se despliega una lista de comando que inician con la letra que se le asignó, en caso de que se desee instalar el comando se tendrá que escribir en la barra de comando la siguiente sintaxis:

ssc install (aquí colocamos el nombre completo del comando que se desea instalar)

Para explorar las diferentes opciones que presenta cada uno de los comandos, se puede recurrir a los documentos de ayuda que acompaña a cada uno de estos comandos, a través de teclear en la sección de comando en el programa:

help (nombre del comando)

Este procedimiento es muy útil, debido a que se puede identificar operaciones específicas en el manejo del comando y obtener información estadística más detallada. Además, en las ayudas se podrán encontrar ejemplos en la forma que podría utilizarse el comando.

También puede utilizarse el comando

findit (aquí colocamos el nombre completo del comando que se desea buscar para posteriormente instalar)

En la web del *software* Stata en la dirección electrónica <https://www.stata.com>, se puede tener acceso a diferentes portales web, manuales, videos, revistas, entre otros formatos más, donde se pueden obtener bases de datos, ejercicios resueltos, aplicación de comandos y soporte del *software*.

Cuadro 1.3. Principales comandos empleados en Stata

Comando	Concepto	Ejemplo
describe	Muestra las características de la base de datos y del conjunto de variables que la conforma.	describe (nombre de variables) – si se deja vacío se desplegarán los datos de la base de datos
list	Muestra los valores que conforman cada una de las variables que se encuentran contenidas en la base de datos.	list (nombre de variables) – si se deja vacío se desplegarán los datos de todas las variables
summarize	Muestra los principales indicadores estadísticos de las variables, tales como el número de observaciones, media, desviación estándar, valor mínimo y máximo.	sum edad – si al final se le agrega una coma y la opción detail se obtiene información estadística adicional de las variables, tal como, simetría, kurtosis, distribución percentil, entre otros.
codebook	Proporciona información sobre el rango de valores, valores únicos, datos perdidos y tabula las etiquetas y sus respectivos valores.	codebook edad
tabulate	Presenta una tabla de los valores de las variables en términos de su tabla de frecuencias; puede emplearse con una o dos variables dando origen a tablas de contingencia.	tabulate edad; tab edad sexo
generate	Crea variables a través de fórmulas específicas o por medio de la interacción de variables.	gen lingreso=log(ingreso); gen tasareal=tasainterresnominal-inflacion.
encode	Transforma la categoría de una variable de texto en valores numéricos.	encode mes; gen(nummes) – a cada mes (representado con el nombre del mes) le asigna un valor
destring	Convierte una variable de texto a una variable numérica.	destring cve_mun, replace – la opción replace sobrescribe la misma variable
clear	Elimina la base de datos que se encuentra en uso en el programa.	Clear
set more off	Evita que en presencia de resultados extensos se haga pausa para mostrar resultados parciales en la sección de resultados de la imagen 1.	set more off

sort	Ordena de menor a mayor el conjunto de una variable.	sort edad
correlation	Establece el grado de asociación entre un conjunto de variables.	corr exportaciones pib
count	Contabiliza las observaciones de la base de datos.	Count
rename	Renombra una variable en la base de datos.	rename edades eda
recode	Recodifica los valores de una variable.	recode edad 1/15=1
drop	Borra variables o un conjunto de datos específicos de una variable.	
twoway		
regress	Estima modelos de regresión simple o múltiple	reg salario educ exp.
predict	Pronostica los valores residuales y de la variable pronosticada.	predict error, r, predict yest, xb
log using	Permite que se guarde cada una de las sintaxis que se emplearon, así como los resultados que se muestren en la sección de resultados.	log using modelcrec.log
log close	Deja de guardar las sintaxis y resultados que aparecen en la sección de resultados.	log close

Fuente: Elaboración propia.

2. Análisis sobre las personas que presentan SARS-CoV-2 en México: enero a septiembre 2020

La pandemia por SARS-CoV-2 ha sido uno de los retos más importantes de salud en el mundo y para México en las últimas décadas. Los altos índices de contagio de la población por este virus han provocado una demanda creciente de los sistemas de salud públicos y privados, así como la demanda de medicamentos y, de manera muy particular, han generado efectos importantes en el ámbito económico y social debido a las medidas restrictivas que se han impuesto sobre aquellas actividades que son consideradas económicas no indispensables.

En este capítulo se pretende mostrar, desde una perspectiva únicamente estadística, el análisis de datos —haciendo hincapié en el uso de gráficas— a través del programa de Stata. La aplicación considerada consiste en analizar la evolución de los contagios en México, así como las características demográficas de la población que se ha contagiado, durante el periodo correspondiente del 21 de enero al 13 de septiembre de 2020. Para lograrlo se emplea la base de datos que es publicada en el portal del Gobierno de México (<https://coronavirus.gob.mx/datos/#DownZCSV>). Estas bases de datos están acompañadas por los correspondientes códigos para poder identificar a las distintas categorías y valores de las variables ahí contenidas).

Esta base de datos se encuentra en formato csv, por lo que se requerirá en principio importarlos al programa Stata a través de la siguiente sintaxis

```
import delimited ".../200913COVID19MEXICO.csv"
```

Dada la naturaleza de los datos, es necesario emplear el comando **import delimited** para poder importar la base de datos, posteriormente se recomienda que pueda guardarse en el formato del programa que es la extensión *dta*. En la sintaxis (1.1) se observan tres puntos suspensivos antes del nombre del archivo que se descargó del portal mencionado previamente, estos puntos representan la ubicación del archivo que hemos descargado, y es por ello por lo que le solicitamos que lo importe de la ubicación donde guardamos el archivo con extensión *csv*.

La base de datos se acompaña de un catálogo donde se asocian los valores de cada una de las variables con las etiquetas que se señalan en el mismo. Es muy importante referenciar a este catálogo debido a que nos permite comprender y dar lecturas a los resultados estadísticos y gráficos, por lo que se recomienda etiquetar cada una de las variables de la base de datos de acuerdo con las etiquetas del catálogo. A lo largo de este capítulo se irá especificando la forma de hacerlo con las variables que se vayan empleando.

Por otro lado, se hace mención que los datos que se emplean en el análisis son exclusivamente de la población contagiada por SARS-CoV-2 de acuerdo con el periodo especificado previamente. Por tal motivo, de la base de datos se excluye a la población que no presenta contagio, para lo cual se utilizó la siguiente expresión:

drop if resultado!=1

Aquí se señala que, de acuerdo con el catálogo, la variable denominada *resultado*, capta la población que ha sido diagnosticada como positivo al virus SARS-CoV-2. En este caso el valor de uno en esta variable representa a la población que ha dado positivo al virus, por lo tanto, se excluirán a todas aquellas personas que tengan un valor distinto al uno, tal y como se señala en esta expresión.

Una vez clasificada a la población contagiada por el virus SARS-CoV-2 en México durante el periodo de estudio, lo primero que se realiza es conocer los detalles sobre la base de datos que se estará analizando a lo largo de este capítulo. Para identificarlos, se emplea el comando **describe**, en la ventana de comandos y se obtienen los resultados que se presentan en el cuadro 2.1. En este cuadro se observa que el total de contagiados por el virus SARS-

CoV-2 al 13 de septiembre de 2020 fue de 668 381 personas. La base de datos cuenta con un total de 35 variables, en la cual aparecen con el nombre *vars*; de igual forma se presenta el tamaño del archivo que contiene esta base de datos (*size*). Adicionalmente, los resultados de este comando permiten identificar el tipo de dato (*storage format*), el formato de presentación de los datos (*display format*), las etiquetas de los valores de las variables (*value label*) y la etiqueta de la variable (*variable label*).

describe

Cuadro 2.1. Resultados del comando **describe** en Stata

<i>Contains</i>	<i>Data</i>			
obs:	668381			
vars:	35			
size:	114293151			
Variable name	Storage type	Display format	Value label	Variable label
fecha_actua- li~n	str10	%10s		FECHA_ACTUALIZACION
id_registro	str6	%9s		ID_REGISTRO
origen	byte	%8.0g		ORIGEN
sector	byte	%8.0g		SECTOR
entidad_um	byte	%8.0g		ENTIDAD_UM
sexo	byte	%8.0g		SEXO
entidad_nac	byte	%8.0g		ENTIDAD_NAC
entidad_res	byte	%8.0g		ENTIDAD_RES
municipio_res	Int	%8.0g		MUNICIPIO_RES
tipo_paciente	byte	%8.0g		TIPO_PACIENTE
fecha_ingreso	str10	%10s		FECHA_INGRESO
fecha_sintomas	str10	%10s		FECHA_SINTOMAS
fecha_def	str10	%10s		FECHA_DEF
intubado	byte	%8.0g		INTUBADO
neumonia	byte	%8.0g		NEUMONIA
edad	Int	%8.0g		EDAD

<i>Contains</i>	<i>Data</i>		
nacionalidad	byte	%8.0g	NACIONALIDAD
embarazo	byte	%8.0g	EMBARAZO
habla_lengua_~g	byte	%8.0g	HABLA_LENGUA_INDIG
diabetes	byte	%8.0g	DIABETES
epoc	byte	%8.0g	EPOC
asma	byte	%8.0g	ASMA
inmusupr	byte	%8.0g	INMUSUPR
hipertension	byte	%8.0g	HIPERTENSION
otra_com	byte	%8.0g	OTRA_COM
cardiovascular	byte	%8.0g	CARDIOVASCULAR
obesidad	byte	%8.0g	OBESIDAD
renal_cronica	byte	%8.0g	RENAL_CRONICA
tabaquismo	byte	%8.0g	TABAQUISMO
otro_caso	byte	%8.0g	OTRO_CASO
resultado	byte	%8.0g	RESULTADO
migrante	byte	%8.0g	MIGRANTE
pais_nacional~d	str57	%57s	PAIS_NACIONALIDAD
pais_origen	str38	%38s	PAIS_ORIGEN
uci	byte	%8.0g	UCI

Fuente: Elaboración propia.

Los resultados del cuadro 2.1 únicamente permiten identificar la composición de la base de datos. Por otro lado, para identificar los primeros datos estadísticos puntuales, el comando **summarize** resulta útil, debido a que este comando proporciona información relacionada con el número de observaciones, la media, la desviación estándar, mínimo y máximo. Al colocar el comando **summarize** en la barra de comando del programa Stata, sin incluir ninguna variable, se asume que se está solicitando información estadística de todas las variables, que en este caso de acuerdo con el cuadro 2.1, existe un total de 35 variables, la información se presenta en el cuadro 2.2. En este último cuadro aparecen celdas vacías, esto se debe a que sobre las variables que no son numéricas no se pueden obtener estadísticos. Del

cuadro 2.1 se observa que son aquellas variables donde sus tipos de datos son cadenas de valores alfanuméricos o de texto (*strings*).

summarize

Cuadro 2.2. Datos estadísticos de medidas centrales y en desviación

<i>Variable</i>	<i>Obs</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Min</i>	<i>Max</i>
fecha_actu~n	0				
id_registro	0				
Origen	668381	1.631771	0.4823242	1	2
Sector	668381	8.818859	3.765292	1	13
entidad_um	668381	16.26995	8.39075	1	32
Sexo	668381	1.52159	0.499534	1	2
entidad_nac	668381	16.98182	9.489879	1	99
entidad_res	668381	16.48385	8.304719	1	32
municipio_~s	668381	39.72954	53.04093	1	999
tipo_pacie~e	668381	1.245942	0.4306447	1	2
fecha_ingr~o	0				
fecha_sint~s	0				
fecha_def	0				
intubado	668381	73.61254	40.97776	1	99
neumonia	668381	1.812287	0.5160008	1	99
Edad	668381	44.79368	16.64612	0	118
nacionalidad	668381	1.0032	0.0564804	1	2
embarazo	668381	51.87989	47.45143	1	98
habla_leng~g	668381	5.001881	16.8265	1	99
diabetes	668381	2.137932	5.323005	1	98
epoc	668381	2.244082	4.980176	1	98

<i>Variable</i>	<i>Obs</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Min</i>	<i>Max</i>
asma	668 381	2.231471	4.970846	1	98
inmusupr	668 381	2.267789	5.167484	1	98
hipertension	668 381	2.079133	5.146193	1	98
otra_com	668 381	2.43037	6.593922	1	98
cardiovasc~r	668 381	2.249765	5.084537	1	98
obesidad	668 381	2.084513	5.078815	1	98
renal_cron~a	668 381	2.245387	5.03424	1	98
tabaquismo	668 381	2.208604	5.19474	1	98
otro_caso	668 381	19.74595	38.06623	1	99
resultado	668 381	1	0	1	1
migrante	668 381	98.74068	5.015484	1	99
pais_nacio~d	0				
pais_origen	0				
uci	668 381	73.63738	40.93556	1	99

Fuente: Elaboración propia.

En el cuadro 2.2, se obtiene información sobre la población contagiada por el virus SARS-CoV-2 en México durante el periodo de análisis. Pero, antes de revisarlas, es importante hacer mención que cuando se analizan variables donde los datos que las constituyen son valores categóricos (conocidas como variables cualitativas), es decir, valores con los que se codifican los estados o los municipios o aquellos que permiten identificar a una mujer o a un hombre o variables binarias sobre respuestas positivas o negativas, en ninguno de estos casos tiene sentido obtener estadísticos. En este caso, la variable de edad es cuantitativa, la edad promedio de los contagiados por el virus SARS-CoV-2 en México es de 44.8 años con una desviación estándar de 16.6 años, se pueden identificar infantes con cero años y adultos mayores con una edad máxima de 118 años.

Los aspectos que se tienen que tomar en consideración cuando se analiza una base de datos son los valores que representan categorías de respues-

tas como *No Sé* o *No Respondió* que generalmente se codifican con valores entre 98 y 99, respectivamente. Por ejemplo, la base de datos podría contener una variable de número de hijos muertos al nacer, donde algunas personas pudieron haber contestado 1, 2, 3, ... o más, pero también alguien pudo haber decidido no responder, por lo tanto se le asignará el valor categórico de 99; si no se considera esta situación, se podría cometer el error de incluirlo en el análisis estadístico, lo que sobredimensionaría el valor medio y de dispersión de la variable debido a que el programa lo considerará como un valor de 99 niños muertos al nacer. Por lo anterior, siempre resulta útil emplear los comandos de **describe** y **summarize** conjuntos con el catálogo que acompaña la base de datos para poder identificar estos detalles.

El comando **codebook** explora la composición de las variables en la base de datos. Al igual que el comando **summarize**, el comando **codebook** muestra información estadística de la variable, tal como el tipo de dato, rango de valores de la variable, únicos valores de la variable, celdas sin datos, tabulación muy limitada de los datos, media y desviación estándar. El despliegado de la información dependerá del tipo de variable que se solicite. En los cuadros 2.3 y 2.4 se aplicó el comando para dos variables: sexo y edad, respectivamente, en el caso de la primera es categórica y la segunda, numérica. Este comando permite identificar el total de valores perdidos, la cantidad de datos únicos y una tabulación previa de los datos de la variable. De los resultados del cuadro 2.3 se encuentra que la variable sexo es numérica tipo byte, el número representa una categoría, que de acuerdo con el catálogo 1 corresponde a las mujeres, 2 a los hombres y 99 no respondió. Por esta razón, el rango es entre 1 y 2, lo que significa que todas las personas contagiadas contestaron esta pregunta, incluso, el valor único de 2 confirma que solamente la variable analizada se compone de dos valores a lo largo de las 668 381 personas contagiadas. Además, no se encuentra ninguna celda vacía. De acuerdo con la tabulación, existe un total de 319 760 mujeres contagiadas y 348 621 hombres durante el periodo de registro de la base de datos.

codebook sexo

Cuadro 2.3. Estadísticos del comando **codebook** para la variable *sexo*

<i>Sexo</i>	<i>Sexo</i>
type: numeric (byte)	
range: [1,2]	units: 1
unique values: 2	missing .: 0/668381
tabulation: Freq. Value	
319760 1	
348621 2	

Fuente: Elaboración propia.

Con respecto a los datos de la edad que se presentan en el cuadro 2.4, se observa que la variable es numérica, el rango de edad de los contagiados es entre 0 y 118 años, existe un total de 114 valores únicos en la variable edad y no se cuenta con ninguna celda vacía. Los datos de la media y desviación estándar ya se conocían con el comando **summarize**, pero no la distribución percentil, este indicador que se ordena de menor a mayor permite identificar cómo se distribuyó el conjunto de datos de la edad, en donde se observa que 10 % de los contagiados tiene una edad de 25 años, la mediana o la mitad de las observaciones cuenta con edades de 44 años, mientras que 75 % de los pacientes cuenta con 56 años edad.

codebook edad

Cuadro 2.4. Estadísticos del comando **codebook** para la variable *edad*

<i>Edad</i>	<i>Edad</i>
type: numeric (int)	
range: [0,118]	units: 1
unique values: 114	missing .: 0/668381
mean: 44.7937	
std. dev: 16.6461	
percentiles: 10%	25% 50% 75% 90%
25	32 44 56 68

Fuente: Elaboración propia.

Por otra parte, para desplegar un cuadro de información con datos ordenados de forma ascendente o descendente, con base en alguna variable, lo hacemos mediante la combinación de dos comandos: **sort** y **list**. Un ejemplo de lo anterior se presenta en el cuadro 2.5, en donde se muestran 10 datos, los cuales han sido ordenados por la fecha de síntomas. Adicionalmente, agregamos características como sexo, edad y tipo de paciente. Si bien el comando **sort** ordena a todas las observaciones, en el comando **list** indicamos que sólo nos muestre los primeros 10 casos, es por ello por lo que agregamos en la sintaxis la expresión **in** y posteriormente el rango deseado, en este caso 1/10 significa “del 1 al 10”. Si quisiéramos los primeros 15 casos, tendríamos que especificar 1/15. De este resultado se puede identificar que de las 10 personas infectadas con el virus, seis eran hombres y solamente tres se hospitalizaron (véase el catálogo de la base de datos para clasificar la categoría de los valores de cada una de las variables).

```
sort fecha_sintomas
list sexo edad tipo_paciente fecha_sintomas in 1/10
```

Cuadro 2.5. *Desplegado de los datos de las variables*

	<i>sexo</i>	<i>edad</i>	<i>tipo_p~e</i>	<i>fecha_si~s</i>
1	2	36	1	13/01/20
2	1	27	1	29/01/20
3	1	43	2	06/02/20
4	2	45	1	19/02/20
5	2	72	2	21/02/20
6	2	36	2	22/02/20
7	2	41	1	22/02/20
8	2	59	1	23/02/20
9	1	51	1	24/02/20
10	1	19	1	25/02/20

Fuente: Elaboración propia.

Ahora bien, para realizar un seguimiento de los casos totales por día, es decir, la suma de las personas infectadas en un día, emplearemos la variable

de *fecha de síntomas* con el propósito de agrupar el total de infectados diarios que percibieron los síntomas del virus, de tal forma que de aquí en adelante los datos que se presentan corresponderán al total de pacientes infectados con SARS-CoV-2 de acuerdo con la fecha que sintieron los síntomas.

Para poder crear una variable que integre a estos pacientes, se emplea el comando **egen** que permite crear variables en función de otras variables, así como también mediante operaciones de suma y producto. En este sentido, lo que se desea es sumar el total de pacientes infectados por día, así que se emplea la opción **sum()** con la variable resultado. Es importante señalar que la variable *resultado* contiene valores de 1 debido a que así se clasificaron las personas que resultaron con prueba positiva ante el virus, en caso de que la variable se clasificara con un valor dos, tendríamos que haber creado una variable que cuantificará a cada uno de los pacientes con el valor de uno, de tal manera que al sumarlos coincida con el número de infectados con el virus por día. Finalmente, es importante agruparlos por día, así que la condición **by** y el orden **sort**, permite clasificarlas y sumarlas por días, la sintaxis empleada es como sigue:

by fecha_sintomas, sort: egen personascovid=sum(resultado)

El nombre de la variable que se creó, y que contiene el número de casos, es *personascovid* (el nombre asignado es una decisión del analista), en caso de que un día hubiera existido un total de 80 contagiados, a cada paciente en ese día se le asignará el valor de 80. Lo anterior es importante debido a que ya se cuenta con el número de casos de contagio por día, pero ese valor se encuentra asignado a cada uno de los pacientes, entonces al momento de graficarlos se deberá considerar que solamente se requiere un dato por día, pero si se diera el caso de que un día específico existieron 400 contagiados se tendrán a 400 pacientes con un valor de la variable *personascovid* de 400. Para resolver esta situación, se utilizará la media de cada uno de los valores de la variable, de tal forma que el valor obtenido será el valor que se repite a través de los pacientes que coincidieron en fecha de síntomas del virus.

Teniendo en cuenta este detalle, resulta más práctico y fácil de comprender la información de personas contagiadas por el virus SARS-CoV-2 a tra-

vés de la presentación de una gráfica de barras. De esta forma, se parte de emplear el comando **graph bar** y la forma en que se considerará la variable; en este caso la variable es *personascovid* y, como se mencionó previamente, se utilizará el valor medio de la variable. Posteriormente, en la parte de opciones que aparecerá después de la coma, se especificará sobre las variables que se pretende graficar, en este caso la fecha de síntomas es nuestra variable de orden en el eje de las abscisas, y en el eje de las ordenadas se establecerán las personas contagiadas. Al agrupar las observaciones por día es evidente que el tamaño de la muestra disminuye. Ahora la cantidad de filas que tendrá nuestra base de datos corresponderá con la cantidad de fechas para las que se tenga información. En nuestro ejemplo, tenemos datos de 209 días, así que la base de datos se compone de 209 observaciones; el eje de las abscisas se notará saturado en el gráfico, es por ello por lo que se asigna de forma vertical y un tamaño de letra específico a través de la sintaxis:

```
label(angle(vertical) labsize(half_tiny))
```

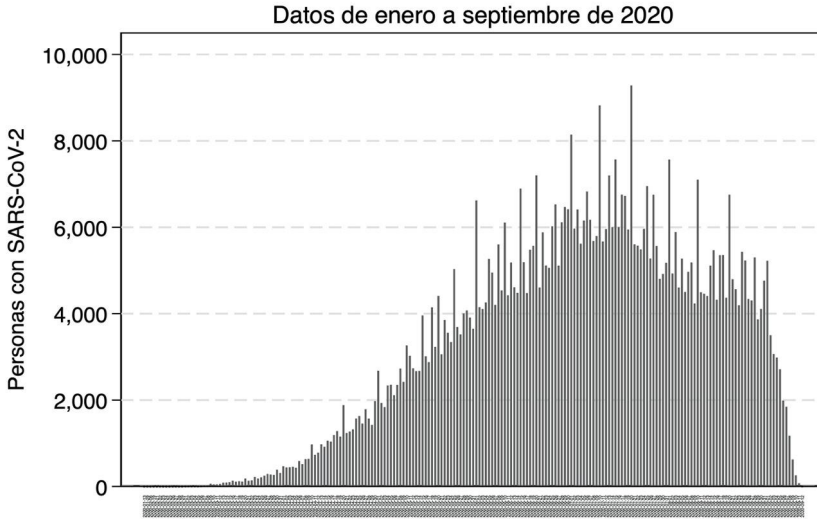
Además, se puntualiza el título del eje de las ordenadas, título del gráfico, subtítulo y la notas. El tipo de gráfico es **s1mono** y se ha nombrado “Mexico”. La sintaxis para construir el gráfico en STATA es la siguiente:

```
graph bar (mean) personascovid, over(fecha_sintomas, label(
angle(vertical)labsize(half_tiny)))ytitle(PersonasconSARS-CoV-2)ylabel(
angle(horizontal) glpattern(dash)) ymtick(, angle(vertical)) tit
le(Personas con SARS-CoV-2 en México) subtitle(Datos de enero
a septiembre de 2020) note(Fuente:Elaboración propia con datos
de la Secretaría de Salud) scheme(s1mono) name(Mexico, repla
ce)
```

El resultado aparece en la gráfica 2.1, de forma tal que se ordena el número de pacientes contagiados con SARS-CoV-2, desde enero a septiembre de 2020, en México. En el gráfico observamos que en julio de 2020 se alcanzó el mayor número de personas contagiadas por el virus, también se observa que el 20 de julio se registraron 9 282 personas contagiadas; es a partir de ese día que se muestra un descenso diario en el número de

contagiados. En el mes de agosto se aprecia una disminución en el número de contagios, pero durante este mes se mantuvo relativamente constante alrededor de los 4 000 contagios diarios. Para el mes de septiembre la tasa de contagio disminuyó a menos de 2 000 por día.

Gráfica 2.1. Gráfica de barras sobre el contagio diario de personas por el virus SARS-CoV-2



Fuente: Elaboración propia con datos de la Secretaría de Salud.

En la figura 2.1 señalamos el número de contagios en México desde enero a septiembre, identificándose un constante crecimiento desde abril hasta julio que alcanzó su máximo nivel, después de ese periodo el número de contagios ha disminuido. Sin embargo, este comportamiento que se presenta corresponde al total de la población contagiada, pero es necesario identificar si el comportamiento por sexo fue distinto. Dentro de la base de datos se cuenta con esa variable donde, de acuerdo con el catálogo de la base de datos, el código 1 corresponde a Mujer, el 2 a Hombre, mientras que 9 significa que no lo especificaron. Para poder asignar etiquetas que permitan una mejor lectura de los resultados de la variable sexo, primero debemos crear la etiqueta y después asignarla a la variable correspondiente. En este caso, creamos la etiqueta con el comando **label define**, después asignamos

el nombre con el que la identificaremos, seguido de los valores que se asocian a cada uno de los números que se consideren, de la siguiente manera:

label define Sexo 1 “Mujer” 2 “Hombre”

El nombre de la etiqueta es *sexo* en donde el 1 corresponde a las mujeres y el 2 a los hombres. Dado que en todos los casos se especificó el sexo de la persona, no se asignó etiqueta para los no especificados. Una vez definida la etiqueta *sexo*, se asigna a la variable *sexo* a través del comando **label values**, tal y como se señala a continuación:

label values sexo Sexo

Pero antes de realizar la gráfica del número de contagios diarios por sexo es fundamental tener en cuenta que la unidad de análisis de la base de datos son pacientes, por lo que es necesario que, para construir indicadores sobre el número de pacientes, primero se agrupe puntualmente a la población que se pretende analizar. Si el propósito es conocer el número de personas contagiadas por día y sexo, entonces tendremos que clasificarlas de esa forma, agrupando a los pacientes por día de contagio y sexo.

Entonces, para analizar la distribución de pacientes contagiados por sexo se crea la variable empleando el comando **egen**, debido a que se construirá una variable a partir de una condición, y se asigna el nombre de la variable. En este caso, se clasificaron los casos por fecha de síntomas y sexo a través de la opción de **group**, tal y como se señala a continuación:

egen grupofechasexo=group(fecha_sintomas sexo)

Retomando la variable *resultados*, que contiene valores de uno que representa cada uno de los pacientes infectados, se contabilizan los casos de pacientes infectados por sexo, empleando la variable que se acaba de crear denominada *grupofechasexo*, y se procede a la suma de los casos que coincidan con cada uno de los grupos, la sintaxis para el cálculo de esta variable se ha denominado *personascovidfechasexo* y se calcula como se presenta a continuación:

by grupofechasexo, sort: egen personascovidfechasexo=sum(resultado)

Creada la variable que representa el número de pacientes infectados por el virus SARS-CoV-2, tanto para hombres como para mujeres, se presenten los resultados en forma de un gráfico de barras, en donde se solicita la media de la variable *personascovidfechasexo* y se clasifica por sexo. La sintaxis queda definida de la siguiente manera:

```
graph bar (mean) personascovidfechasexo, over(fecha_sintomas, label(angle(vertical) labsize(half_tiny))) ytitle(Personas con SARS-CoV-2) ylabel(, angle(horizontal) gpattern(dash)) ymtick(, angle(vertical)) by(, title(Personas con SARS-CoV-2 en México por sexo) subtitle(Datos de enero a septiembre de 2020) note(Fuente:Elaboración propia con datos de la Secretaría de Salud)) scheme(s1mono) name(Méxicosexo, replace) by(sexo)
```

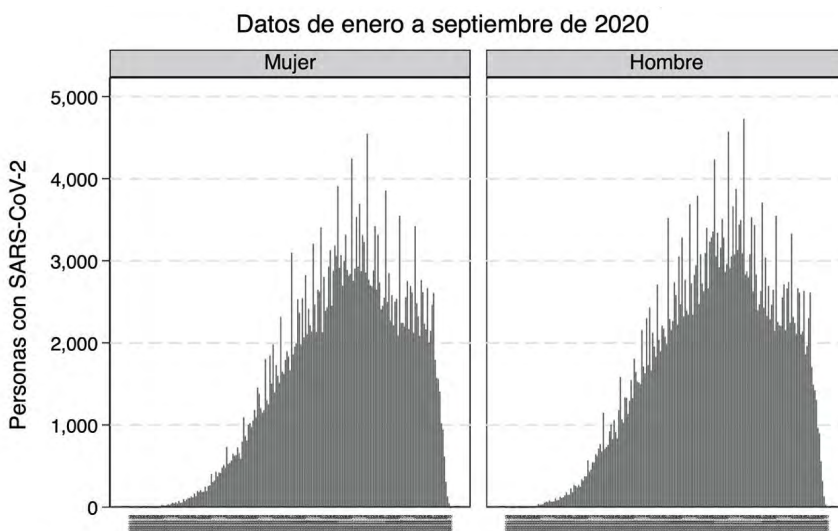
El resultado se presenta en la gráfica 2.2, donde observamos que la proporción de hombres contagiados por el virus SARS-CoV-2 es mayor al número de mujeres durante el periodo de enero a septiembre de 2020, aunque cabe destacar que de la gráfica visualizamos que la trayectoria es la misma para ambos.

Por otro lado, de acuerdo con la información que se proporciona en la base de datos, la variable *tipo de paciente* especifica si las personas contagiadas requirieron o no hospitalización. En este sentido, la categoría 1 de la variable *tipo pacientes* hace referencia a aquellos que no requirieron hospitalización; por su parte, el valor 2 a los que fueron hospitalizados y 99 al no especificado. Con esta información se crea la etiqueta y se asigna a la variable correspondiente. Para evitar confusión, se recomienda que ambas se nombren de la misma manera. Además, dado que el interés es mostrar la evolución diaria de casos de pacientes tanto de aquellos que fueron o no hospitalizados es necesario categorizar y agrupar a cada paciente para que coincida con la fecha de síntomas y el tipo de paciente y, posteriormente, crear la variable que pondere el número de pacientes que diariamente se

contagió por SARS-CoV-2. El conjunto de comandos que se requiere para realizar este procedimiento es el siguiente:

```
label define Tipo_pacientes 1 "Ambulatorio" 2 "Hospitalizado" 99
"No especificado"
label values tipo_paciente Tipo_pacientes
egen grupofechatipopaci=group(fecha_sintomas tipo_paciente)
by grupofechatipopaci,sort: egen personascovidfechatipopaci=
sum(resultado)
```

Gráfica 2.2. Gráfica de barras sobre el contagio diario por el virus SARS-CoV-2 por sexo



Fuente:Elaboración propia con datos de la Secretaría de Salud

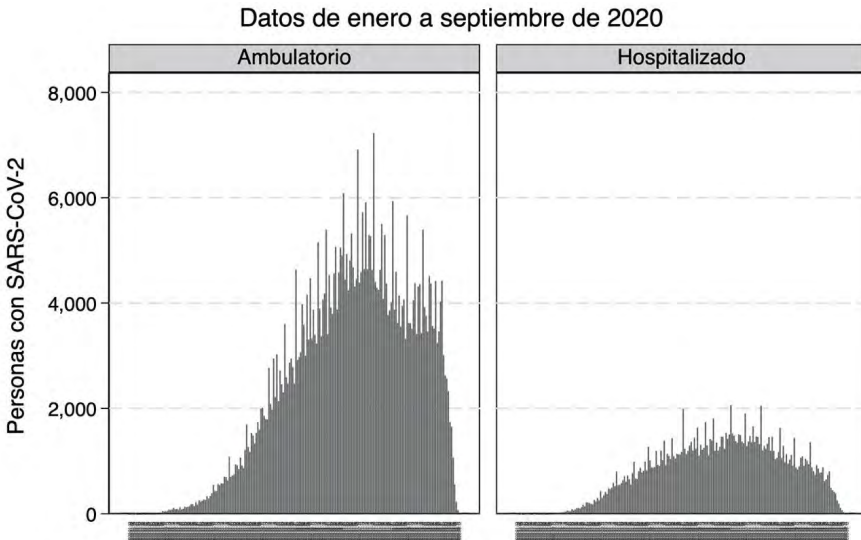
Fuente: Elaboración propia con datos de la Secretaría de Salud.

Definida la variable *personascovidfechatipopaci*, se presenta de forma gráfica siguiendo la misma estructura que se ha señalado previamente, empleando la siguiente sintaxis:

```
graph bar (mean) personascovidfechatipopaci, over(fecha_sinto-
mas, label(angle(vertical) labsize(half_tiny))) ytitle(Personas con
SARS-CoV-2) ylabel(, angle(horizontal) gpattern(dash)) ymtick(,
```

angle(vertical)) by, title(Personas con SARS-CoV-2 en México por tipo de paciente) subtitle(Datos de enero a septiembre de 2020) note(Fuente:Elaboración propia con datos de la Secretaría de Salud)) scheme(s1mono) name(Méxicotipopaci, replace) by(ti-po_paciente)

Gráfica 2.3. Gráfica de barras de contagio diario por SARS-CoV-2 por tipo de paciente



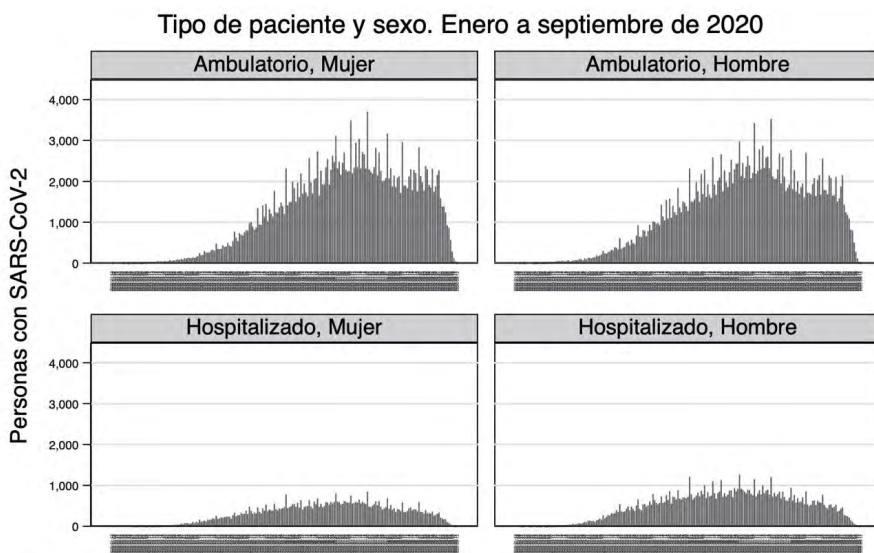
Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

La gráfica 2.3 muestra la distribución diaria de los tipos de pacientes contagiados por SARS-CoV-2, en donde está claro que el número de pacientes ambulatorios supera en promedio en al menos dos veces al número de pacientes hospitalizados. Al igual que en los gráficos anteriores, el mes de julio es donde observamos el mayor número de pacientes con SARS-CoV-2 que requirieron hospitalización. El máximo de registros diarios de hospitalización fue ligeramente superior a 2 000 pacientes. Ahora bien, podríamos estar interesados en los tipos de pacientes y el sexo, con el propósito de comparar entre hombres y mujeres respecto a quienes han requerido hospitalización. Para construir el gráfico con estas características, es necesario

agrupar la fecha de síntomas, el tipo de pacientes y el sexo para, posteriormente, crear la variable *personascovidfechatipopacisexo* a partir de sumar los valores de la variable resultado de acuerdo con los grupos que se han conformado en el paso previo. Finalmente, construimos el de barras siguiendo la sintaxis de los anteriores, es decir, las medias de la variable *personascovidfechatipopacisexo* con respecto a la fecha de síntomas, junto con la condición **by**; clasificamos cada uno de los cuadrantes de acuerdo con el tipo de paciente y el sexo, de tal forma que se generaran un total de cuatro cuadrantes. A continuación, mostramos la sintaxis que se requiere para construir el gráfico de pacientes diarios por tipo de paciente y por sexo, los resultados se presentan el cuadro 2.4.

Figura 2.4. Gráfico de barras de contagio diario por SARS-CoV-2 por tipo de paciente y sexo



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

```
egen grupofechatipopacisexo=group(fecha_sintomas tipo_paciente sexo)
```

```
by grupofechatipopacisexo, sort: egen personascovidfechatipopacisexo=sum(resultado)
```

```
graph bar (mean) personascovidfechatipopacisexo, over(fecha_
sintomas, label(angle(vertical) labsize(half_tiny))) ytitle(Personas
con SARS-CoV-2) ylabel(, labsize(vsmall) angle(horizontal))
ymtick(, angle(horizontal)) by(, title(Personas con SARS-CoV-2 en
México) subtitle(Tipo de paciente y sexo. Enero a septiembre de
2020) note(Fuente:Elaboración propia con datos de la Secretaría
de Salud)) scheme(s1mono) by(tipo_paciente sexo)
```

Al observar el total diario de pacientes hospitalizados y los ambulatorios por sexo, destacamos que tanto para hombres y mujeres, la mayoría no han requerido hospitalización; sin embargo, en los casos que sí se ha requerido, la mayoría son hombres, incluso en la gráfica observamos que hay días en que la razón de hospitalizados hombres-mujeres es aproximadamente de 2 a 1.

Con estos resultados, evidenciamos que los hombres han mostrado mayor vulnerabilidad por el virus SARS-CoV-2. Adicionalmente, es importante identificar tanto a las diferencias por sexo como los efectos que ha tenido el virus en la población por grupo de *edad*. Para ello, recodificamos la variable edad de acuerdo con los intervalos que consideremos pertinentes. Aquí hemos definido los siguientes cohortes: infantes de 0 a 11 años, jóvenes de 12 a 29, adultos jóvenes de 30 a 45, adultos de 46 a 59 y adultos mayores de 60 y más años. Así, se emplea el comando **recode** para reconfigurar los datos de la edad de acuerdo con la clasificación señalada. Es de notar que, para identificar a cada grupo de edad empleamos el símbolo de diagonal asignándole un valor distinto; recomendamos, además, crear una nueva variable que contenga esta clasificación de los pacientes por edad, mediante el comando **gen** aplicado en la misma línea de instrucción como una opción. A continuación, especificamos la sintaxis correspondiente:

```
recode edad (0/11=1) (12/29=2) (30/45=3) (46/59=4) (60/max=5),
gen(grupoedad)
```

Posteriormente, para identificar a cada uno de los grupos creamos las etiquetas que corresponden a cada uno de ellos de acuerdo con las edades de la siguiente manera:

label define grupoedad 1 "Infantes" 2 "Jóvenes" 3 "Adultos jóvenes" 4 "Adultos" 5 "Adultos mayores"

Luego, asignamos la etiqueta a la variable codificada que contiene los diferentes grupos de la población:

label value grupoedad grupoedad

En el cuadro 2.6 presentamos la distribución de pacientes de acuerdo con los grupos de edad; para construirlo utilizamos el comando **tabulate** seguido del nombre de la variable. Como podemos apreciar, la mayor proporción de pacientes contagiados con el virus son los adultos jóvenes que representan 35.2 % del total de pacientes, el segundo grupo son los adultos que representan 26.3 % y en tercero los adultos mayores con 19.5 %. Los pacientes contagiados por el virus, en su mayoría, son mayores de 30 años, mientras que los menores de 30 años que están contagiados solamente representan 19 %.

tabulate grupoedad

Cuadro 2.6. *Distribución de pacientes por grupo de edad*

<i>RECODE of edad (EDAD)</i>	<i>Freq.</i>	<i>Percent</i>	<i>Cum.</i>
Infantes	9 196	1.38	1.38
Jóvenes	117 848	17.63	19.01
Adultos jóvenes	234 980	35.16	54.16
Adultos	176 004	26.33	80.5
Adultos mayores	130 353	19.5	100
Total	668 381	100	

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Ahora bien, para analizar el número de pacientes que han requerido hospitalización por grupos de edad, podemos elaborar un cuadro condicionando por los distintos grupos etarios. Los resultados están en el cuadro 2.7. La población que se ha contagiado y que ha requerido ser hospitalizada son los adultos mayores con una incidencia en 44.5 % del total de los pa-

cientes, mientras que la población de los adultos constituye el segundo grupo con mayor incidencia, lo que refleja que la población mayor a 45 años de edad es la más vulnerable al virus SARS-CoV-2 y que generalmente son lo que requieren hospitalización.

tabulate grupoedad if tipo_paciente==2

Cuadro 2.7. Distribución de pacientes por grupo de edad y hospitalizados

<i>RECODE of edad (EDAD)</i>	<i>Freq.</i>	<i>Percent</i>	<i>Cum.</i>
Infantes	1 709	1.04	1.04
Jóvenes	7 043	4.28	5.32
Adultos jóvenes	30 418	18.5	23.83
Adultos	52 016	31.64	55.47
Adultos mayores	73 197	44.53	100
Total	164 383	100	

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Desde la perspectiva gráfica para analizar la incidencia de pacientes hospitalizados a través de los distintos grupos de edad, es necesario crear el grupo de pacientes por fecha de síntoma, tipo de paciente y grupo de edad, para posteriormente crear la variable a través de sumar el número de pacientes de acuerdo a estas tres condiciones con el propósito de crear el gráfico de barras empleando la media de los datos (dado que al tratarse de múltiples casos por días, la variable que contiene el total de cada uno se repite de acuerdo con el número de casos que se observaron diariamente) y condicionándolo para pacientes hospitalizados de acuerdo con la fecha de síntomas de los pacientes. Para presentar de forma separada los gráficos, al final de la sintaxis agregamos la condición **by** con la respectiva variable que permitirá clasificar los gráficos, que para este caso será el grupo de edad.

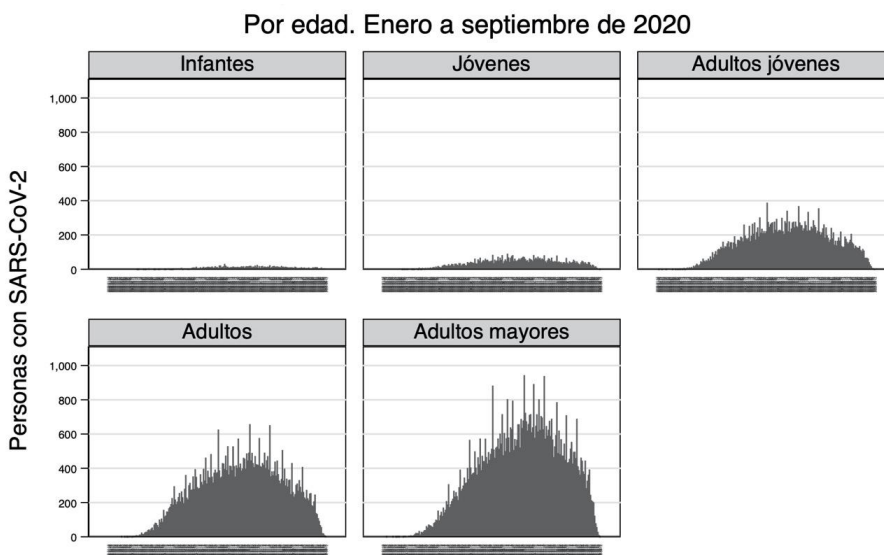
```
egen grupofechatipopacidad=group(fecha_sintomas tipo_paciente grupoedad)
by grupofechatipopacidad, sort: egen personascovidfechatipopacidad=sum(resultado)
```

```

graph bar (mean) personas_covid_fecha_tipo_pacientes if tipo_paciente==2,
over(fecha_sintomas, label(angle(vertical) labsi-
ze(half_tiny))) ytitle(Personas con SARS-CoV-2) ylabel(, labsi-
ze(vsmall) angle(horizontal)) ymtick(, angle(horizontal)) by(,
title(Personas hospitalizadas con SARS-CoV-2 en México) subtit-
le(Por edad. Enero a septiembre de 2020) note(Fuente:Elabora-
ción propia con datos de la Secretaría de Salud)) scheme(s1 mo-
no) by(grupoedad)

```

Gráfica 2.5. Gráfica de barras de contagio diario por SARS-CoV-2 por grupo de edad



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

En la gráfica 2.5 presentamos el número de casos diarios de pacientes hospitalizados de acuerdo con el grupo de edad al que pertenecen, en donde, a diferencia del cuadro 2.7, resulta más sencilla la forma de visualizar el comportamiento de los pacientes infectados. En este mismo gráfico observamos cómo a mayor edad se incrementan los pacientes hospitalizados, siendo los adultos mayores los que mayor incidencia tienen; este grupo de la población alcanzó niveles cercanos a los 1 000 pacientes hospitalizados

por día, mientras que en el caso de los adultos plenos en los meses de mayor número de infectados por el virus se alcanzó un aproximado de 700 hospitalizados; para los adultos jóvenes alrededor de 400 hospitalizados, en el caso de los jóvenes durante este periodo nunca fue mayor a 100 el total de hospitalizados. Una de las ventajas de presentar en gráfica los datos del cuadro 2.7 es que es posible visualizar el comportamiento diario de los casos, sin necesidad de cuadros tan extensos.

La información de hospitalizados diarios por grupo de edad también se puede presentar por sexo. Es importante que al incluir una nueva categoría de agregación creamos un grupo que integre el total de categorías. En este caso, debemos crear el grupo que integre la fecha de síntomas, tipo de pacientes, grupo de edad y sexo, tal y como sigue:

```
egen grupofechatipopacidadsexo=group(fecha_sintomas tipo_
paciente grupoedad sexo)
```

Posteriormente, empleamos la variable *resultado* que, como se mencionó previamente, es una variable que contiene el valor de uno, que representa cada uno de los casos de los pacientes observados a lo largo del periodo de análisis. Esta variable se suma de acuerdo con los grupos que se han creado considerando las cuatro categorías mencionadas, la sintaxis quedaría definida de la siguiente manera:

```
by grupofechatipopacidadsexo, sort: egen percovidfetipopaci-
gesex=sum(resultado)
```

Para graficar estos resultados, seguimos la estructura que ya hemos aplicado donde se presenta el número de casos diarios a través de la gráfica de barras para lo cual se utiliza la media de la variable *percovidfetipopacigesex*, condicionada a los pacientes hospitalizados. Cabe hacer mención de que para obtener un gráfico separado por grupo de edad y sexo, es necesario al final de la siguiente sintaxis en la opción **by** incluir ambas variables, tal como sigue.

```
graph bar (mean) percovidfetipopacigesex if tipo_paciente==2,
```

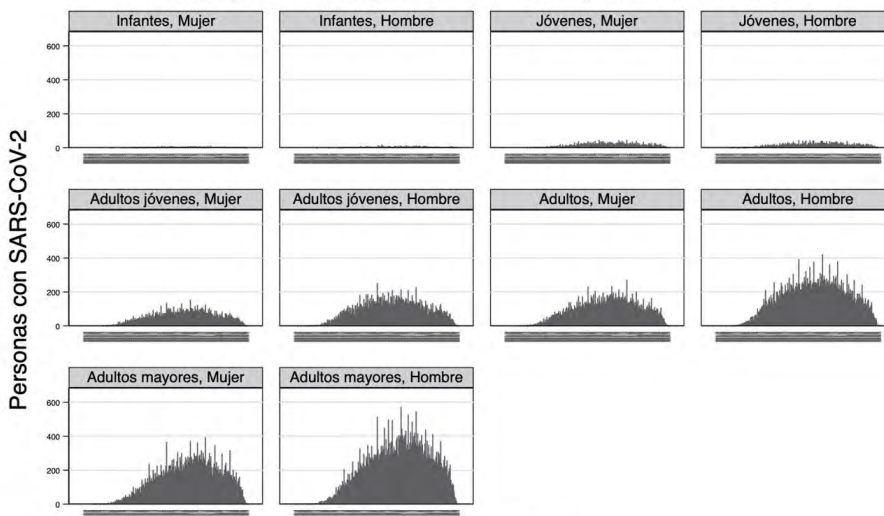
```

over(fecha_sintomas, label(angle(vertical) labsize(half_tiny)))
ytitle(Personas con SARS-CoV-2) ylabel(, labsize(vsmall) an-
gle(horizontal)) ymtick(, angle(horizontal)) by(, title(Personas
hospitalizadas con SARS-CoV-2 en México) subtitle(Por grupo de
edad y sexo. Enero a septiembre de 2020) note(Fuente:Elabora-
ción propia con datos de la Secretaría de Salud)) scheme(s1mo-
no) by(grupoedad sexo)

```

Como resultado obtenemos la gráfica 2.6, en donde es posible identificar que ser hombre y adulto mayor ha sido un factor de los pacientes que han requerido hospitalización; sin embargo, esta situación de vulnerabilidad de los hombres no es particular de los adultos mayores, debido a que en los otros grupos de edad en donde también se observa alta incidencia de hospitalización, prevalece el número de hombres que se han hospitalizado. En los últimos tres grupos de edad, observamos que en promedio los hombres hospitalizados superan 1.5 veces el total de casos diarios de mujeres hospitalizadas en México.

Gráfica 2.6. Gráfica de barras de contagio diario por SARS-CoV-2 por grupo de edad y sexo
Por grupo de edad y sexo. Enero a septiembre de 2020



Fuente:Elaboración propia con datos de la Secretaría de Salud

La representación gráfica permite identificar, a lo largo del tiempo, la evolución del comportamiento de los pacientes contagiados con el virus, pero no es posible establecer puntualmente la cantidad de personas hospitalizadas por sexo. En este sentido, es más conveniente recurrir a la tabulación de los datos. En el cuadro 2.8 presentamos la sintaxis y los resultados de la tabulación de los pacientes por grupo de edad y sexo. Al incorporar en la sintaxis la opción de **col** y **row**, se genera el porcentaje tanto de la variable que se encuentra en la fila, que en este caso es el *grupo de edad*, como por columna que es la variable de *sexo*. Esta forma de presentar la información permite al analista visualizar el dato absoluto y relativo para cada una de las variables. Es decir, se podría interpretar el total de pacientes por sexo de acuerdo con cada uno de los grupos de edad, ya sea en valor absoluto o relativo, esto implicaría un análisis vertical. Por otro lado, el análisis se podría realizar de forma horizontal, es decir, considerando el total de casos para cada uno de los grupos de edad y su distribución por sexo, de igual forma se puede interpretar con datos absolutos o relativos.

De los resultados presentados en el cuadro 2.8, el total de mujeres contagiadas con el virus es 319 760 representando 47.8 % de toda la población contagiada. De esta población, 35.6 % son mujeres adultas jóvenes, 25.9 % mujeres adultas, 19 % jóvenes y 18.1 % adultas mayores. En el caso de los hombres, la población contagiada representa 52.2 %, es decir, en México al 13 de septiembre existían 348 621 hombres contagiados con el virus SARS-CoV-2, la mayoría de la población es mayor de los 30 años. Analizando el cuadro de forma horizontal identificamos que en casi todos los casos la proporción de hombres contagiados supera a las mujeres, solamente en la clasificación de jóvenes es mayor la proporción de mujeres contagiadas que los hombres, superándolos por alrededor de tres puntos porcentuales, el total de mujeres contagiadas es 51.6 % y de hombres es 48.4 %. En valor absoluto, observamos que el total de mujeres jóvenes contagiadas es 60 763 y de hombres jóvenes es 57 085.

tabulate grupoedad sexo, col row

Cuadro 2.8. *Distribución de los pacientes contagiados por SARS-CoV-2 por grupo de edad y sexo*

<i>RECODE of edad (EDAD)</i>	<i>Mujer</i>	<i>Hombre</i>	<i>Total</i>
Infantes	4,355	4,841	9,196
	47.36	52.64	100
	1.36	1.39	1.38
Jóvenes	60,763	57,085	117,848
	51.56	48.44	100
	19	16.37	17.63
Adultos jóvenes	113,957	121,023	234,980
	48.5	51.5	100
	35.64	34.71	35.16
Adultos	82,830	93,174	176,004
	47.06	52.94	100
	25.9	26.73	26.33
Adultos mayores	57,855	72,498	130,353
	44.38	55.62	100
	18.09	20.8	19.5
Total	319,760	348,621	668,381
	47.84	52.16	100
	100	100	100

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Por su parte, el análisis vertical y horizontal que revisamos en el cuadro 2.8 se puede representar en forma de gráfica de pie o de pastel (como tradicionalmente se conoce). Para ello es importante que se defina cuál de las dos formas de análisis se desea graficar. En este caso, elegimos la forma vertical empleando la variable *sexo* y analizaremos su distribución porcentual a través de los *grupos de edad*. Empleamos el comando **graph pie** y se grafica sobre los grupos de edad y con la opción **by** se clasifica esa distribución por sexo, la sintaxis para aplicarse en el programa Stata quedaría definida de la siguiente manera:

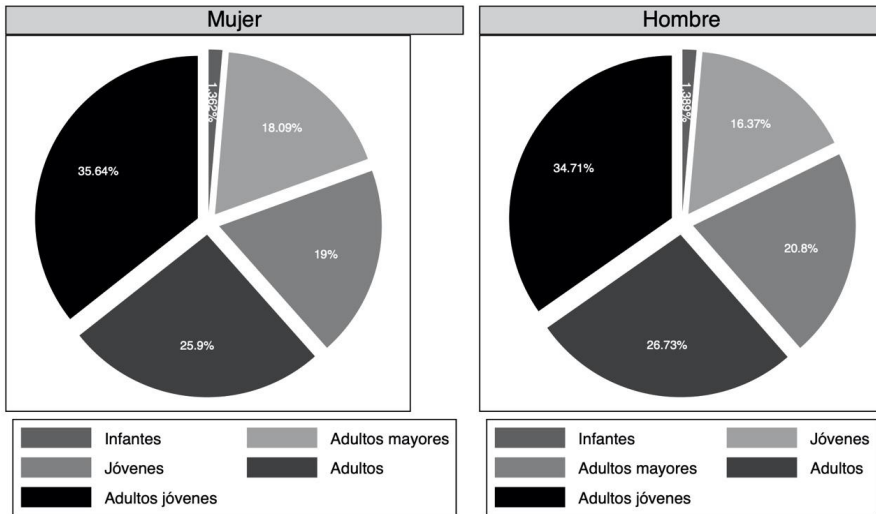
```

graph pie, over(grupoedad) sort pie(_all, explode) plabel(_all
percent, color(white) size(vsmall) orientation(rvertical)) line(la
lign(inside)) by(, title(Personas con SARS-CoV-2 en México) sub
title(Por grupo de edad y sexo. Enero a septiembre de 2020)
note(Fuente:Elaboración propia con datos de la Secretaría de Sa
lud.)) legend(size(small)) scheme(s1mono) by(sexo)

```

En la gráfica 2.7 se muestran los resultados, esta forma de representarlos constituye una alternativa a la forma en que se presenta en el cuadro 2.7, que contribuye en muchas ocasiones a tener un mayor entendimiento de los resultados. En la gráfica observamos que en realidad la distribución de contagiados por el virus SARS-CoV-2 varía relativamente poco entre hombres y mujeres, prácticamente la distribución de los contagiados por edad es la misma entre ambos sexos en México en el periodo que se analiza.

Gráfica 2.7. Gráfica de pastel del porcentaje de contagio por SARS-CoV-2 por grupo de edad y sexo Por grupo de edad y sexo. Enero a septiembre de 2020



Fuente:Elaboración propia con datos de la Secretaría de Salud.

Fuente: Elaboración propia con datos de la Secretaría de Salud.

El contagio y la propagación del virus SARS-CoV-2 en el mundo es un factor de mortandad, y en México no es la excepción. La información de los pacientes que han fallecido se especifica en la variable *fecha_def*, cuando esta variable tiene una fecha especificada debemos entender que es la fecha en que la persona falleció, caso contrario encontraremos el código 9999-99-99. Lo primero que haremos es transformar los datos de los pacientes no fallecidos en datos perdidos (*missing*) de la siguiente manera (las comillas reflejan que es un dato cualitativo):

```
replace fecha_def="" if fecha_def=="9999-99-99"
```

Ahora, nuestra variable *fecha_def* tiene registros únicamente de personas fallecidas. Lo que resta por hacer es contar estos casos utilizando la siguiente sintaxis en donde generamos la variable *muertos* con el valor de 1 para cuando existe una fecha de defunción; es importante asignarle este valor que representa a la persona muerta, de tal forma que, si desea conocer el total de muertes por día o mes o entre otras formas más, simplemente se puede sumar esta variable. Con la siguiente sintaxis le asignaremos el valor de la variable *muertos* para todos aquellos casos distintos de *missing*, es decir a todos aquellos que tienen fecha de defunción, y posteriormente con el comando **count** de la variable *muertos* cuando vale uno es posible conocer el total de muertos contagiados por el virus SARS-CoV-2:

```
gen muertos=1 if fecha_def!=""  
count if muertos==1
```

o directamente también se podría utilizar:

```
count if fecha_def!=""
```

El resultado señala que en México, al 13 de septiembre de 2020, han muerto 70 821 personas de 668 381 contagiados con el virus, esto representa a 10.6% de esta población. Para identificar cómo se distribuyen por grupo de edad y sexo, realizamos un gráfico de barras, pero antes de ello se debe crear el grupo que integre a las tres categorías que se emplearán en la gráfi-

ca, el grupo se denominará *grupomugesex* y agrupará a las variables *muertos*, *grupoedad* y *sexo*, esto se realiza a través del comando **egen** y la opción **group**:

```
egen grupomugesex=group(muertos grupoedad sexo)
```

Definidos los grupos, procedemos a crear la variable que sumará a los contagiados con el virus y que han muerto de acuerdo con los grupos que se formaron por edad y sexo. Para realizarlo empleamos la condición **by**, ordenamos con el comando **sort**, generamos la variable con el comando **egen** y la opción *sum*, y finalmente recodificamos los valores de la variable que sumaron cero por *missing*, de la siguiente forma:

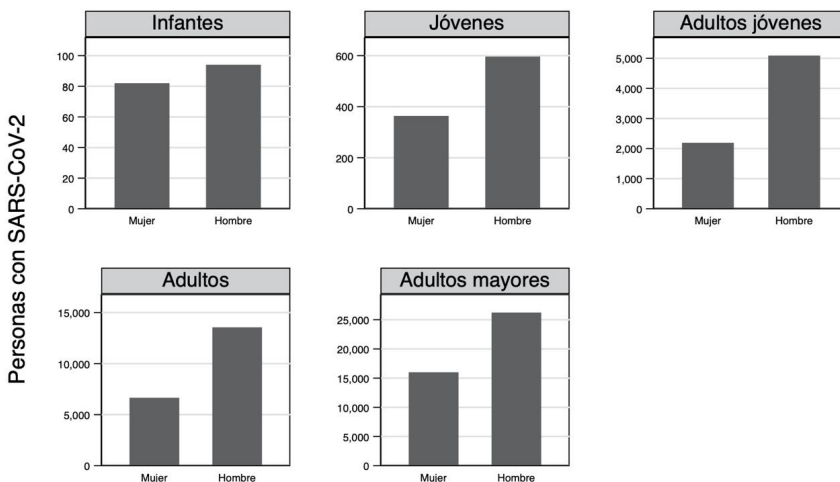
```
by grupomugesex, sort: egen percovidmugesex=sum(muertos)  
replace percovidmugesex=. if percovidmugesex==0
```

Creada la variable construimos la gráfica de barras respecto al total de personas que se contagiaron con el virus SARS-CoV-2 y fallecieron. Sobre la variable creada *percovidmugesex* se utiliza la media, debido a que como en los casos anteriores la suma del total de fallecidos por cada grupo se asigna a cada una de las personas que pertenece al grupo (por esta razón decidimos convertir los valores en *missing*, ya que de no hacerlo se consideraría un dato más, provocando que la media fuera menor a la real). Graficamos sobre el eje de las abscisas la variable *sexo* y para generar un gráfico para cada uno de los grupos de edad se emplea la opción **by** y la respectiva variable:

```
graph bar (mean) percovidmugesex, over(sexo, label(angle(horizontal) labsize(vsmall))) ytitle(Personas con SARS-CoV-2) ylabel(, labsize(vsmall) angle(horizontal)) ymtick(, angle(horizontal)) by(, title(Personas muertas por SARS-CoV-2 en México) subtitle(Por grupo de edad y sexo. Enero a septiembre de 2020) note(Fuente: Elaboración propia con datos de la Secretaría de Salud)) scheme(s1mono) by(grupoedad, style(rescale))
```

La gráfica 2.8 presenta el total de la población que ha fallecido a causa del virus SARS-CoV-2, por grupo de edad y sexo; en este observamos que en todos los casos el número de muertes siempre es mayor para los hombres, aunque en el caso de los adultos jóvenes el número de hombres muertos supera en alrededor de 2.5 veces el de las mujeres. En el caso de la población de adultos mayores es donde mayor número de muertes existen, para los hombres se tiene un registro de aproximadamente 26 000 fallecidos y en las mujeres alrededor de 16 000 fallecidas. El segundo grupo de la población en donde observamos incidencia de mortalidad por el virus es en los adultos, es decir, de aquellos que se encuentran entre 45 y 59 años, el total de hombres fallecidos se aproxima a los 14 000 y en el caso de las mujeres a los 7 000. Los casos de fallecimientos en la edad infantil durante el periodo de análisis se aproximan a los 170 casos en total.

Gráfica 2.8. Gráfica de barras de contagio por SARS-CoV-2 por grupo de edad y sexo
Por grupo de edad y sexo. Enero a septiembre de 2020



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

La mortalidad en los pacientes con el virus SARS-CoV-2, generalmente se ha asociado con padecimientos crónicos, en este sentido, realizamos el análisis de los datos relacionando el fallecimiento de los pacientes con

padecimiento como la diabetes, enfermedad pulmonar obstructiva crónica, asma, inmunosupresión e hipertensión, los más comunes entre los pacientes contagiados. A continuación, se remplazan los valores diferentes a la unidad de las variables debido a que únicamente se consideran los casos de pacientes que presentaron el padecimiento.

```
replace diabetes=. if diabetes!=1  
replace epoc=. if epoc!=1  
replace asma=. if asma!=1  
replace inmusupr=. if inmusupr!=1  
replace hipertension=. if hipertension!=1
```

Para representar los resultados empleamos el comando **table** por grupo de edad y, debido a que cada uno de los casos está representado por la unidad, por tal motivo se solicita la sumatoria de los casos que corresponda con cada uno de los padecimientos que se señalaron previamente y se ordenan por sexo. Los resultados y la sintaxis se presentan en el cuadro 2.9, en donde mostramos el total de casos que existen de padecimientos entre los pacientes contagiados con el virus, predominando aquellos que padecen hipertensión y diabetes. Los hombres reportan mayor incidencia con relación a las mujeres. Los datos muestran que 130 181 pacientes contagiados con el virus padecen hipertensión y 104 219 diabetes. Tanto en mujeres como en hombres, la población de adultos mayores es el grupo donde observamos más casos con estos dos padecimientos. Respecto a ellos, corroboramos que después de los 45 años se incrementa de forma más acelerada su incidencia, tanto para hombres como para mujeres. Por su parte, el principal problema que afronta la población de jóvenes e infantes es el asma.

```
table (sexo grupoedad), stat(total diabetes epoc asma hipertension inmusupr)
```

Cuadro 2.9. Distribución de pacientes por grupo de edad, sexo y padecimientos

SEXO and RECODE of edad (EDAD)	sum(diabetes)	sum(epoc)	sum(ssma)	sum(hipertension)	sum(inmusupr)
Mujer					
Infantes	35	3	123	29	105
Jóvenes	940	84	1 985	1 120	351
Adultos jóvenes	7 526	399	4 052	9 094	892
Adultos	18 356	1 006	2 940	22 480	1 288
Adultos mayores	22 070	3 307	1 639	29 923	1 240
Total	48 927	4 799	10 739	62 646	3 876
Hombre					
Infantes	33	8	151	35	120
Jóvenes	789	96	1 699	1 587	398
Adultos jóvenes	8 835	391	2 418	11 509	811
Adultos	20 881	1 081	1 447	23 249	962
Adultos mayores	24 664	3 477	1 059	31 155	1 184
Total	55 202	5 053	6 774	67 535	3 475

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Por otro lado, también se reportan otras condiciones de riesgo o enfermedades de los pacientes infectados por el virus SARS-CoV-2, tales como problemas cardiovasculares, obesidad, insuficiencia renal (*Renal_cronica*) y tabaquismo. Para clasificar a los pacientes que presentan estos factores de riesgo, partimos de que las variables que contienen esta información son binarias donde 1 es sí presenta tal condición y 2 que no; creamos nuevas variables que representan cada uno de los factores de riesgo y condicionamos cada una de ellas al valor de 1 de las variables originales.

gen Cardiovascular=1 if cardiovascular==1

gen Obesidad=1 if obesidad==1

gen Renal_cronica=1 if renal_cronica==1

gen Tabaquismo=1 if tabaquismo==1

Cada una de las variables creadas se compone por el valor de uno, con la finalidad de que sumemos el total de pacientes de cada uno de los facto-

res de riesgo. En el cuadro 2.9 mostramos el total de pacientes que padece algún factor de riesgo, tanto por sexo como grupo de edad. En la sintaxis, empleamos nuevamente el comando `table` y la variable a utilizar para enlistar los resultados, en este caso, por grupo de edad. En la parte de opciones especificamos la instrucción de que calcule la suma de los pacientes que presentan cada uno de los factores de riesgo, esta información se clasificó por sexo a través de `by(sexo)`.

table (sexo grupoedad), stat(total Cardiovascular Obesidad Renal_cronica Tabaquismo)

Cuadro 2.9. Distribución de pacientes por grupo de edad, sexo y otros factores de riesgos

SEXO and RECODE of edad (EDAD)	sum(Cardiovasc~r)	sum(Obesidad)	sum(Renal_cron~a)	sum(Tabaquismo)
Mujer				
Infantes	55	111	17	16
Jóvenes	301	6881	390	3126
Adultos jóvenes	849	21283	1019	6059
Adultos	1452	19589	1616	3388
SEXO and RECODE of edad (EDAD)				
Adultos mayores	3122	13123	2500	2036
Total	5779	60987	5542	14625
Hombre				
Infantes	57	161	20	37
Jóvenes	319	6910	506	6189
Adultos jóvenes	976	23458	1356	12291
Adultos	1845	19213	2117	7770
Adultos mayores	4374	11152	3183	7341
Total	7571	60894	7182	33628

Fuente: Elaboración propia con datos de la Secretaría de Salud.

De los factores de riesgos identificados, el de mayor incidencia es el de obesidad, el número de pacientes con esta condición durante el periodo de análisis fue de 121,881 siendo ligeramente mayor el número de mujeres.

De los grupos de edad, el de adultos jóvenes que se ubican entre los 30 y 45 años son la mayor proporción de pacientes que presentan obesidad. En los que respecta a la condición cardiovascular y la insuficiencia renal, ser hombre y adulto mayor es la población de pacientes donde existe una mayor incidencia de estas condiciones de riesgo. En lo correspondiente al tabaquismo, es la segunda condición de riesgo con mayor prevalencia en los pacientes contagiados con el virus SARS-CoV-2, existe un total de 33 628 hombres y 14 625 mujeres, siendo los adultos jóvenes, tanto en hombres como en mujeres, los que presentan en mayor medida esta condición de riesgo.

Ahora bien, considerando los padecimientos y factores de riesgos, establecemos el total de pacientes que han fallecido derivado de estas situaciones. En cuadro 2.10, clasificamos de acuerdo con la edad, por sexo y padecimiento de los pacientes que han fallecido. Salvo el padecimiento de asma, en el resto de los padecimientos se asocia un mayor número de fallecidos a los hombres, siendo los adultos mayores en donde más se observan, tanto en hombres como en mujeres. La hipertensión y la diabetes son los padecimientos donde observamos el mayor número de fallecidos, los pacientes con hipertensión que fallecieron representaron 25 % de todos los que manifestaron tener el padecimiento. En relación con los que presentan diabetes, la proporción fue de 0.26. Los infectados que presentaban la enfermedad pulmonar obstructiva crónica, asma e inmunosupresión, no representaron ni la mitad del total de fallecidos con diabetes o por hipertensión.

table (sexo grupoedad) if muertos==1, stat(total diabetes epec asma hipertension inmusupr)

Cuadro 2.10. *Distribución de pacientes por grupo de edad, sexo y padecimientos de personas que fallecieron*

SEXO and RECODE of edad (EDAD)	sum(Diabetes)	sum(EPOC)	sum(Asma)	sum(Hipertension)	sum(inmusupr)
Mujer					
Infantes	3	0	3	4	8
Jóvenes	58	6	11	81	34
Adultos jóvenes	662	28	94	613	76

Adultos	2983	200	266	3035	231
Adultos mayores	7345	1283	399	9472	427
Total	11051	1517	773	13205	776
Hombre					
Infantes	4	1	0	4	5
Jóvenes	66	5	22	101	40
Adultos jóvenes	1118	65	108	1143	136
Adultos	4740	290	181	4695	270
Adultos mayores	10063	1537	335	12475	524
Total	15991	1898	646	1818	975

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Respecto a las condiciones de riesgo de los pacientes que han fallecido, hay un total de 17 335 que presentaban obesidad, siendo mayor en los hombres, particularmente en los adultos mayores. El segundo factor de riesgo que se asocia a los fallecimientos es el tabaquismo, que afectó 5 634 pacientes, donde 84 % fueron hombres. La insuficiencia renal y los problemas cardiovasculares son los dos factores de menor incidencia en los pacientes que fallecieron por el virus, al igual que en los casos anteriores la mayoría de la población fueron hombres. En todos los factores de riesgo, los casos de fallecimiento se presentaron principalmente en la población de adultos mayores.

table (sexo grupoedad) if muertos==1, stat(total Cardiovascular Obesidad Renal_cronica Tabaquismo)

Los resultados previos muestran que en un determinado número de pacientes fallecidos se presentaron diferentes padecimientos y factores de riesgo, principalmente en la población de adultos mayores. Para establecer la relación entre el número de fallecidos con los principales padecimientos, es decir, hipertensión y diabetes, emplearemos una gráfica de dispersión. Para esto necesitamos el número total de pacientes que presentaban algún tipo de padecimiento y que fallecieron. Calcularemos este dato de manera diaria considerando la fecha de síntomas. Dado que presentaremos la información

a lo largo de todo el periodo de análisis y a través de los meses, construimos la variable *mes* por medio del comando **gen** y la opción **substr** condicionada a la variable de fecha de síntoma, la cual tiene una estructura de año mes y día (2020-09-12), por lo que especificamos que a partir del espacio 6 se extraigan los dos siguientes números, que corresponderían al mes, dado que se obtiene como resultado un dato que es considerado letra (string), lo transformamos en número a través del comando **destring** y se le asigna el nombre de *mes2*:

Figura 2.11. *Distribución de pacientes por grupo de edad, sexo y otros factores de riesgo de personas que fallecieron*

SEXO and RECODE of edad (EDAD)	sum(Cardiovasc~r)	sum(Obesidad)	sum(Renal_cron~a)	sum(Tabaquismo)
Mujer				
Infantes	6	5	0	0
Jóvenes	14	99	62	12
Adultos jóvenes	55	782	199	90
Adultos	231	2279	562	211
Adultos mayores	1 102	4 132	1 168	588
Total	1 408	7 297	1 991	901
Hombre				
Infantes	2	3	2	4
Jóvenes	21	155	79	57
Adultos jóvenes	96	1 721	338	471
Adultos	420	3 633	870	1 243
Adultos mayores	1 830	4 526	1 628	2 958
Total	2 369	10 038	2 917	4 733

Fuente: Elaboración propia con datos de la Secretaría de Salud.

```
gen mes=substr(fecha_sintomas, 6,2)
destring mes, generate(mes2)
```

Posteriormente agrupamos la información por día tanto del total de fallecidos (muertes), así como los padecimientos y factores de riesgo, para lograrlo se emplea el comando **collapse** que suma el total de casos por día

de acuerdo con la fecha de *síntomas*, es por eso que se emplea la opción **sum**. Los valores de la variable *mes2* representan el mes, de tal forma que a lo largo de cada uno de los meses se repite el mismo número de acuerdo con los días que tenga el mes, es por ello por lo que la media regresará el valor correspondiente al mes:

```
collapse (sum) muertos diabetes epoc asma inmusupr hipertension Cardiovascular Obesidad Renal_cronica Tabaquismo (mean) mes2, by(fecha_sintomas)
```

Agrupados los datos por días, asociamos cada número de mes con su nombre mediante la definición de etiquetas y la asignamos a la variable correspondiente de la siguiente manera:

```
label define Mes2 1 "Enero" 2 "Febrero" 3 "Marzo" 4 "Abril" 5 "Mayo" 6 "Junio" 7 "Julio" 8 "Agosto" 9 "Septiembre"  
label value mes2 Mes2
```

Para establecer la relación diaria entre el total de pacientes fallecidos por el virus SARS-CoV-2 y el número de fallecidos con el padecimiento de diabetes, realizamos el gráfico de dispersión a través del comando **twoway** (dado que existe ahora un dato por día, no es necesario definir alguna opción como suma o media), así como la recta de regresión, que se presentan en la gráfica 2.9 y que obtenemos a través de la siguiente sintaxis:

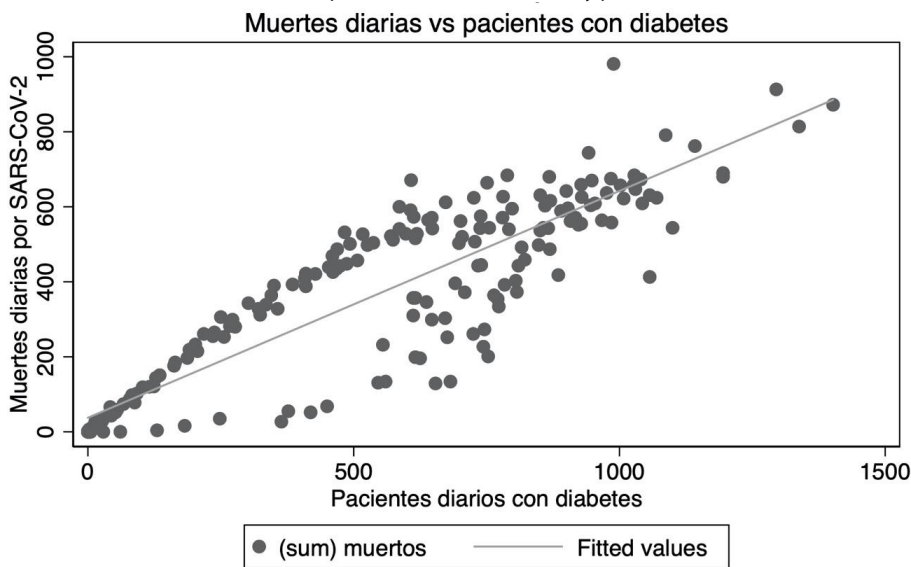
```
twoway (scatter muertos diabetes) || lfit muertos diabetes, ytitle(Muertes diarias por SARS-CoV-2) xtitle(Pacientes diarios con diabetes) title(Gráfico de dispersión) subtitle(Muertes diarias vs pacientes con diabetes) note(Fuente:Elaboración propia con datos de la Secretaría de Salud) scheme(s1mono)
```

La gráfica de dispersión 2.9 muestra una relación positiva entre el número de pacientes fallecidos y la diabetes. Esto significa que conforme el total de pacientes con diabetes aumente, el número de pacientes fallecidos

también incrementará. Se evalúa esta situación, pero ahora por mes, para lo cual se incluye en la sintaxis la extensión **by(mes2)**:

```
twoway (scatter muertos diabetes, msize(vsmall) msymbol(circle_hollow)) || lfit muertos diabetes, ytitle(Muertes diarias por SARS-CoV-2) xtitle(Pacientes diarios con diabetes) by(, title(Gráfico de dispersión por mes) subtitle(Muertes diarias vs pacientes con diabetes) note(Fuente:Elaboración propia con datos de la Secretaría de Salud)) scheme(s1 mono) by(mes2)
```

Gráfica 2.9. Gráfica de dispersión de muertes diarias y personas con diabetes



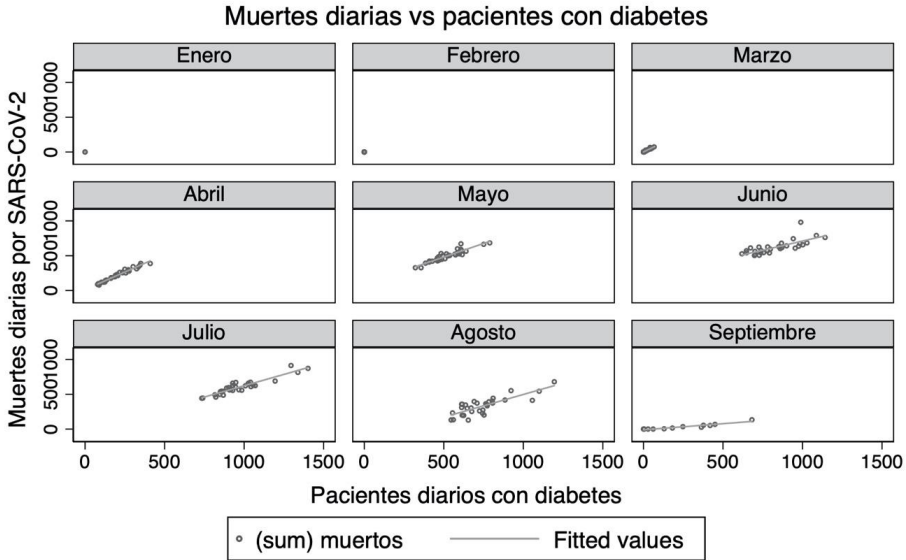
Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

En la gráfica 2.10, presentamos los resultados de la relación entre el número de pacientes fallecidos por mes que padecían diabetes. En los meses de enero y febrero el número de fallecimientos fueron de cero y dos respectivamente. Para el resto de los meses, prevalece la relación positiva entre estas variables, incrementándose la relación de casos, siendo más notoria para el mes de junio. En los meses de julio y agosto comienza un

descenso en el número de fallecimientos, aunque se mantiene el total de pacientes que presentan diabetes. En todos los meses se mantiene una relación positiva entre el total de fallecidos y que presentaban diabetes.

Gráfica 2.10. Gráfica de dispersión de muertes diarias y personas con diabetes por mes



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

En términos de los pacientes que fallecieron y que padecían hipertensión, su comportamiento se presenta en la gráfica 2.11, junto con la recta de regresión que se ajusta al conjunto de datos a lo largo del periodo. Para realizar la gráfica sólo sustituimos la variable de diabetes del ejemplo anterior por hipertensión, y utilizamos círculos huecos para las observaciones (`circle_hollow`):

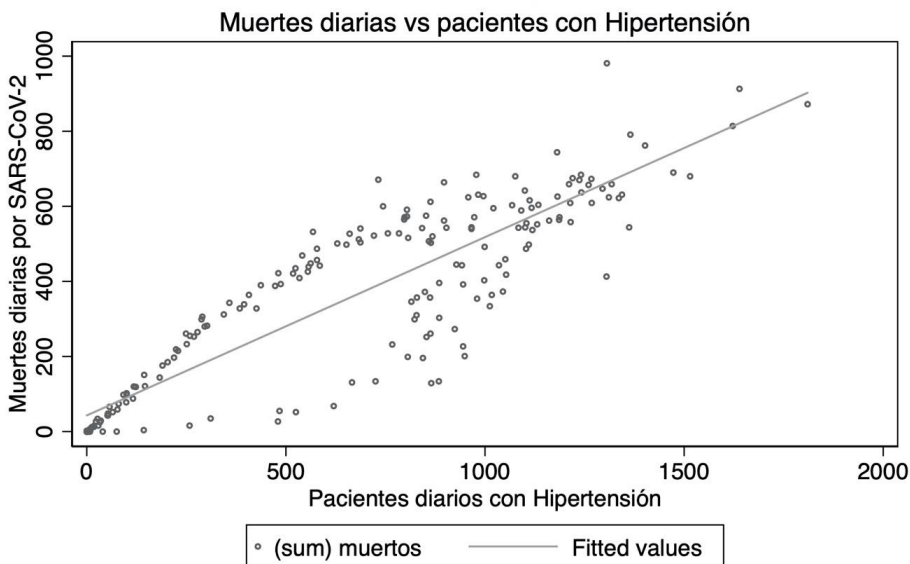
```

twoway (scatter muertos hipertension, msize(vsmall) msymbol(-
circle_hollow)) || lfit muertos hipertension, ytitle(Muertes diarias
por SARS-CoV-2) xtitle(Pacientes diarios con Hipertensión) title(-
Gráfico de dispersión) subtitle(Muertes diarias vs pacientes con
Hipertensión) note(Fuente:Elaboración propia con datos de la

```

Secretaría de Salud) scheme(s1mono)

Gráfica 2.11. Gráfico de dispersión de muertes diarias y personas con hipertensión



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

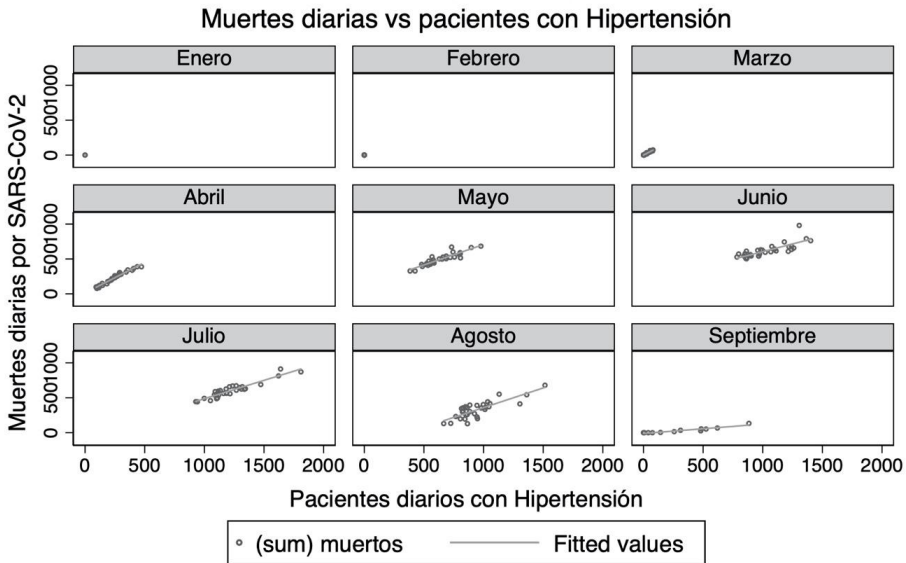
De la gráfica anterior, corroboramos que al igual que con el padecimiento de diabetes, los pacientes que presentan hipertensión se asocian directamente con el número de pacientes infectados de SARS-CoV-2 que han fallecido, de hecho, la inclinación que forma el conjunto de puntos permite visualizar una relación cuadrática entre estas dos variables.

Al graficar el padecimiento de hipertensión con el total de fallecidos por el virus a través de los meses de enero a septiembre de 2020 (aunque al igual que en el caso del padecimientos de diabetes, se analiza desde el mes marzo), encontramos un comportamiento positivo de acuerdo con la gráfica 2.12, con el paso del tiempo la relación se mantiene, pero se desplaza hacia abajo, esto podría reflejar que con el tiempo se ha tenido un conocimiento mayor del virus, así como también de tratamientos más efectivos, evitando

que el ritmo de crecimiento de fallecidos con padecimientos como diabetes e hipertensión incremente. El procedimiento para graficar por mes es similar al empleado en el padecimiento de diabetes.

```
twoway (scatter muertos hipertension, msize(vsmall) msymbol(circle_hollow)) || lfit muertos hipertension, ytitle(Muertes diarias por SARS-CoV-2) xtitle(Pacientes diarios con Hipertensión) by(, title(Gráfico de dispersión por mes) subtitle(Muertes diarias vs pacientes con Hipertensión) note(Fuente:Elaboración propia con datos de la Secretaría de Salud)) scheme(s1mono) by(mes2)
```

Figura 2.12. Gráfica de dispersión de muertes diarias y personas con hipertensión por mes



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Las gráficas de dispersión proporcionan un panorama respecto a la relación entre dos variables, sin embargo, si deseamos tener un conocimiento puntual sobre la relación entre estas, se emplea el coeficiente de correlación simple o de orden cero, cuyo valor se encuentra entre -1 y 1, entre más cercano se encuentre del -1 o 1 más fuerte es la relación, mientras que entre

más cercano sea del cero la relación es más débil. Además, el signo del coeficiente de correlación proporciona información sobre el tipo de relación que existe entre las variables, es decir, es positiva cuando ambas variables aumentan o disminuyen en la misma dirección, y negativa cuando una de ellas aumenta y la otra disminuye y viceversa. Para estimar la correlación entre las variables de muertos con tipo de padecimiento y factor de riesgo, utilizamos el comando **correlation** (en su forma reducida **corr**). En este caso los resultados se presentan en el cuadro 2.11 para los meses comprendidos entre abril a agosto. La variable de obesidad es la que mejor se relaciona con la variable muertos, mientras que la variable diabetes es la segunda y la tercera es la hipertensión. Todas estas variables se relacionan positivamente con la variable muerte, de alguna manera es consistente con lo observado en los gráficos de dispersión analizados previamente. El coeficiente de correlación es un estadístico apropiado para establecer el grado de asociación de las variables, sin embargo, comparado con los análisis de regresión, suele estar limitado, debido a que únicamente se vincula a través de cada par de variables, sin considerar simultáneamente el efecto de otras variables, algo que suele afectar de manera recurrente los modelos de regresión múltiple.

```
corr muertos Diabetes Obesidad hipertension if mes2>3 &
mes2<9
(obs=153)
```

Cuadro 2.11 *Correlación simple de fallecimientos y padecimientos*

<i>Variables</i>	<i>muertos</i>	<i>Diabetes</i>	<i>Obesidad</i>	<i>Hipert-n</i>
muertos	1			
Diabetes	0.7957	1		
Obesidad	0.8061	0.9928	1	
Hipertension	0.7845	0.9963	0.9938	1

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Un estadístico apropiado para dimensionar no solo el efecto de un par de variables, sino que también contempla el efecto simultáneo con otras

variables, es el coeficiente de correlación parcial (**pcorr**). De acuerdo con la correlación parcial que se presenta en el cuadro 2.12, el número de muertos está asociado con la diabetes, la obesidad y la hipertensión. De estas tres condiciones, la población con obesidad es la que se relaciona de forma directa y su grado de asociación con el número de muertes por SARS-CoV-2, es más alto cuando se eliminan los efectos de diabetes e hipertensión, de hecho, de acuerdo a la correlación cuadrática semiparcial en un modelo de regresión que considere los tres padecimientos como explicativas de los pacientes fallecidos, la exclusión de la variable obesidad provocaría que el coeficiente de determinación o R cuadrada (R^2) disminuyera en mayor proporción que si se excluyera cualquiera de las dos variables. En este caso, excluir del modelo obesidad disminuye en 4 % R^2 , en cambio excluir del modelo a la variable diabetes (dejando únicamente a obesidad e hipertensión como variables explicativas) disminuirá el R^2 en 1 %.

pcorr muertos Diabetes Obesidad hipertension if mes2>3 & mes2<9 (obs=153)

Cuadro 2.12. Correlación parcial de fallecimientos y padecimientos

Variable	Partial	Semipartial	Partial	Semipartial	Significance
	Corr.	Corr.	Corr.^2	Corr.^2	Value
Diabetes	0.1769	0.1013	0.0313	0.0103	0.0298
Obesidad	0.3353	0.2006	0.1124	0.0402	0
hipertensión	-0.2986	-0.1763	0.0892	0.0311	0.0002

Fuente: Elaboración propia con datos de la Secretaría de Salud.

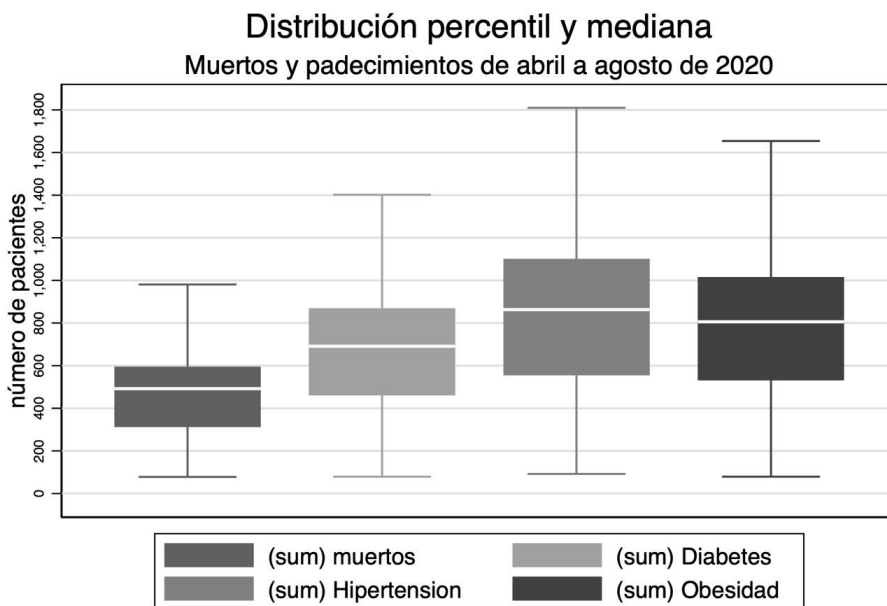
Otro de los aspectos que observamos en la correlación parcial, es que la relación de hipertensión con el número de muertes por SARS-CoV-2 (se excluyen los efectos de las variables diabetes y obesidad) es negativa, y es la segunda en importancia, pero, además, por el resultado obtenido se infiere que si aumentaran los casos con hipertensión, manteniendo constante la obesidad y la diabetes el número de muertes disminuiría.

Por otro lado, para analizar conjuntamente la distribución percentil del total de pacientes fallecidos, pacientes con padecimiento de diabetes, hiper-

tensión y con factor de riesgo de obesidad para los meses de abril a agosto de 2020, emplearemos el gráfico de caja a través del comando (**graph box**). Para especificar los meses que deseamos analizar solo condicionamos con **if** la variable *mes2* y definimos el rango de meses. Los resultados los presentamos en la gráfica 2.13.

graph box muertos Diabetes hipertension Obesidad if mes2<9 & mes2>3, ytitle(número de pacientes) ylabel(#9, labsize(vsmall)) title(Distribución percentil y mediana) subtitle(Muertos y padecimientos de enero a septiembre de 2020) note(Fuente:Elaboración propia con datos de la Secretaría de Salud) scheme(s1 mono)

Gráfica 2.13. Gráfico de caja de muertes y padecimientos



Fuente:Elaboración propia con datos de la Secretaría de Salud

Fuente: Elaboración propia con datos de la Secretaría de Salud.

La gráfica de caja proporciona la distribución percentil de los pacientes en cada una de las variables que analizamos. La hipertensión es la que pre-

senta mayor número de pacientes y una mayor variabilidad en el conjunto de datos; para complementar este gráfico y conocer los resultados puntuales de cada una de las variables empleamos el comando **summarize** (el cual se puede escribir de forma abreviada como **sum**) y la opción **detail** después de la coma, los resultados los mostramos en el cuadro 2.12. De esta forma, la mediana de pacientes con hipertensión es de 863, mientras que 75 % del conjunto de datos se identifica que representa un total de 1 101 pacientes. La segunda problemática con mayor incidencia es la obesidad seguido de la diabetes. Con respecto al total de pacientes fallecidos, observamos que la mediana es de 492 pacientes que mueren diariamente, mientras que 75 % de los datos señala que el total de fallecidos es de 595 pacientes. Los datos puntuales de las demás variables se pueden identificar en el cuadro 2.12.

summarize muertos Diabetes hipertension Obesidad if mes2<9 & mes2>3, detail

Cuadro 2.12. Estadístico sobre fallecimientos y padecimientos
(a)

<i>(sum) muertos</i>				
	<i>Percentiles</i>	<i>Smallest</i>		
1%	88	78		
5%	121	88		
10%	196	98	Obs	153
25%	312	98	Sum of Wgt.	153
50%	492		Mean	456.1503
		Largest	Std. Dev.	187.2547
75%	595	814		
90%	670	872	Variance	35064.33
95%	690	913	Skewness	-0.0976909
99%	913	981	Kurtosis	2.575902

(b)

(sum) Diabetes

	<i>Percentiles</i>	<i>Smallest</i>		
1%	83	79		
5%	124	83		
10%	206	86	Obs	153
25%	461	88	Sum of Wgt.	153
50%	691		Mean	657.6013
		Largest	Std. Dev.	298.3454
75%	869	1195		
90%	1027	1295	Variance	89009.96
95%	1087	1338	Skewness	-0.1502033
99%	1338	1402	Kurtosis	2.439776

(c)

(sum) Hipertension

	<i>Percentiles</i>	<i>Smallest</i>		
1%	99	92		
5%	144	99		
10%	249	100	Obs	153
25%	555	100	Sum of Wgt.	153
50%	863		Mean	820.9412
		Largest	Std. Dev.	381.3218
75%	1101	1515		
90%	1268	1622	Variance	145406.3
95%	1362	1639	Skewness	-0.181437
99%	1639	1810	Kurtosis	2.405881

(d)

<i>(sum) Obesidad</i>				
	<i>Percentiles</i>	<i>Smallest</i>		
1%	83	79		
5%	137	83		
10%	214	86	Obs	153
25%	531	94	Sum of Wgt.	153
50%	806		Mean	768.6471
		Largest	Std. Dev.	346.4788
75%	1016	1341		
90%	1192	1518	Variance	120047.6
95%	1277	1519	Skewness	-0.2584594
99%	1519	1654	Kurtosis	2.496029

Fuente: Elaboración propia con datos de la Secretaría de Salud.

Con estos resultados estadísticos, los cuales hemos presentado de forma puntual y gráfica, sintetizamos que el virus SARS-CoV-2 afecta principalmente a los hombres, de hecho, el mayor número de hospitalizados pertenecen a este sexo. Las poblaciones de adultos y adultos mayores son las más afectadas por el virus, nuevamente en estos grupos son los hombres los más afectados. En el caso de los padecimientos de los pacientes con el virus SARS-CoV-2, observamos que la diabetes y la hipertensión tienen mayor incidencia sobre los infectados, al igual que el factor de obesidad es uno de los más recurrentes. De estos últimos, los que más se asocian a la mortandad son la obesidad y la diabetes.

3. Análisis del ingreso a través de cuantiles

En muchas ocasiones surge el interés de identificar la distribución de algún fenómeno social o económico como, por ejemplo, podríamos estar interesados en establecer cómo se distribuye el número de delitos a través de las colonias en un municipio, con el objetivo de identificar si su incidencia se presenta en un número específico de colonias o si sucede de forma similar en cada una de las colonias del municipio. Otro ejemplo podría asociarse a la forma en que se distribuye el presupuesto público a través de cada uno de los municipios en México, donde el interés radica en determinar si el presupuesto se concentra en unos cuantos municipios del país o se encuentra perfectamente distribuido.

Tradicionalmente, el ingreso ha sido una de las variables que por lo general se analiza su distribución, con el propósito de identificar la concentración o desigualdad a través de la población. La forma jerárquica en que se puede subdividir a la población para analizar su distribución dependerá del interés que exista, en algunas ocasiones podría ser de interés ordenar de menor ingreso a mayor ingreso por grupos conformados por 10 % de la población, de tal forma que se tendrían en total 10 grupos; el primer grupo representaría el total de ingresos que acumularía 10 % de la población con los menores ingresos, mientras que el total de ingresos que acumularía el último grupo de 90 a 100 % de la población estaría representado por la población con mayores ingresos.

En el cuadro 3.1 se presenta un ejemplo de la distribución del ingreso para 10 grupos de la población, el cual se conoce como deciles; el primer

grupo de la población, es decir, 10 % de la población con los menores ingresos, concentra 5 % de todos los ingresos que se generan, mientras que en el segundo decil que va de 10 a 20 % de la población acumula 7 % de todos los ingresos que genera la población. El decil de 50 a 60 % de la población se queda con 9 % de los ingresos que genera la población, en el caso de los que generan más ingresos correspondientes al último decil de 90 a 100 % de la población, son los que se quedan con la mayor cantidad de ingresos acumulando un total de 21 %. Este ejemplo permite identificar la forma en que se concentran los ingresos a lo largo de los diferentes grupos de la población.

Cuadro 3.1. *Distribución del ingreso por deciles*

<i>Grupo</i>	<i>Ingresos acumulados (%)</i>
0-10	5
10-20	7
20-30	6
30-40	8
40-50	8
50-60	9
60-70	12
70-80	11
80-90	13
90-100	21

Fuente: Elaboración propia.

El ejemplo anterior considera agrupar a la población en 10 grupos, sin embargo, esto podría cambiar y preferir subdividir a la población en cinco partes, lo que se conocería como distribución quintil, o bien podría considerarse agrupar en cuatro partes lo que se conocería como distribución cuartil, o simplemente podría analizarse por cada punto porcentual, de tal manera que la población se clasificaría en 100 partes. A final de cuentas, la forma de agrupar a la población para conocer la distribución del ingreso dependerá de los objetivos del estudio que se esté realizando.

Este tipo de análisis se puede realizar en Stata a través del módulo *ps-hare*, en caso de no tenerlo instalado se puede obtener en la barra de comando por medio de la siguiente sintaxis:

sc install pshare

Para conocer con mayor detalle cada una de las opciones con la cuenta este comando, es posible acceder en la barra de comando a la ayuda mediante la sintaxis

help pshare

Para mostrar las principales herramientas de este módulo en el análisis de la distribución, utilizaremos la base de datos sobre salarios mensuales de trabajadores formales en los municipios de Altamira, Ciudad Madero y Tampico en el mes de febrero de 2022. Para conocer las características de la base de datos empleamos el comando *describe* que arroja la siguiente información:

Cuadro 3.2. Descripción de la base de datos

obs:	3752			
vars:	5			
Nombre de la variable	Tipo de almacenamiento	Formato	Valor de la etiqueta	Etiqueta de la variable
sector_economico_1	str44	%44s		
sexo	str6	%9s		
mes	str11	%11s		
salmen	float	%9.0g		
mun	str8	%9s		

Fuente: Elaboración propia.

La base de datos contiene un total de cinco variables donde cada una de estas contiene 3752 datos. De estas variables, cuatro son categóricas: en el caso de la variable *sector_economico_1* contiene las seis actividades más relevantes en los tres municipios estudiados; la variable *sexo* representa la categoría de hombre y de mujer; la variable *mes* contiene únicamente el dato de febrero 2022; y la variable *municipio* contiene los nombres de los tres municipios analizados. La única variable cuantitativa es *salmen*, esta variable representa el nivel de salario promedio de los trabajadores formales. La

información de la base de datos es un extracto de la información que es proporcionada por el Instituto Mexicano del Seguro Social (IMSS) sobre el empleo formal en México.

En principio se analiza la distribución de los ingresos de los trabajadores formales de forma general, clasificada en cinco categorías, por lo que se refiere a un análisis de cuantiles de los ingresos, la información se presenta en el siguiente cuadro a través de la sintaxis:

pshare estimate salmen

Por *default* los resultados obtenidos representan la proporción de trabajadores distribuidos en cada uno de los cinco grupos que se presentan en el cuadro 3.3, de tal forma que la suma de los cinco grupos representa la unidad. En el caso del grupo de 0 a 20 % de los trabajadores que menos ganan acumulan 0.029 de todos los ingresos que generan los trabajadores. En el siguiente grupo de 20 a 40 % de la población que obtiene un poco más de ingresos que el grupo previo, estos acumulan 0.055 de todos los ingresos que generan los trabajadores. Entre estos dos grupos o, dicho de otra forma, 40 por ciento de los trabajadores que menos obtienen acumulan 0.084 de todos los ingresos que generan 100 % de los trabajadores. En el grupo extremo se encuentra 20 % de los trabajadores que cuentan con los salarios mensuales más altos, representado por el grupo de 80 a 100 %, este grupo acumula en términos porcentuales 0.641 de todos los ingresos que generan los trabajadores en la zona de análisis. En este cuadro la información se complementa considerando los rangos o intervalos en los que se colocaría la proporción de ingresos que acumula cada uno de estos grupos.

Cuadro 3.3. *Distribución del ingreso por quintil en proporciones*

salmen	<i>Percentile shares (proportion)</i>		<i>Number of obs = 3,752</i>	
	Coefficient	Std. err.	[95% conf. interval]	
0-20	0.0288	0.0010	0.0268	0.0307
20-40	0.0554	0.0019	0.0517	0.0590
40-60	0.0962	0.0031	0.0902	0.1022
60-80	0.1786	0.0053	0.1682	0.1889
80-100	0.6411	0.0106	0.6203	0.6618

Fuente: Elaboración propia.

Los resultados previos también pueden ser descritos de forma porcentual a través de la siguiente sintaxis:

pshare estimate salmen, percent

En esta sintaxis se anexó la coma y se especificó que los resultados presentados estén en porcentaje. Los resultados que se presentan en el cuadro 3.4, prácticamente son los mismos del cuadro previo, la única diferencia es que la interpretación y su presentación se facilitan cuando se presentan en forma porcentual.

Cuadro 3.4. Distribución del ingreso por quintil en porcentaje

salmen	Percentile shares (percent)		Number of obs = 3,752	
	Coefficient	Std. err.	[95% conf. interval]	
0-20	2.878962	0.0996327	2.683623	3.074302
20-40	5.538616	0.1864561	5.173051	5.904181
40-60	9.619629	0.3076208	9.016509	10.22275
60-80	17.85699	0.5264662	16.82481	18.88918
80-100	64.1058	1.057048	62.03335	66.17824

Fuente: Elaboración propia.

La representación en quintiles de la distribución del ingreso se presentó en los dos últimos cuadros, sin embargo, también es importante establecer el nivel promedio de salario mensual de los trabajadores formales en México para febrero 2022, para lograr este objetivo mantenemos la estructura de la sintaxis que teníamos antes de la coma, y únicamente se sustituye *percent* por *average* de la siguiente manera:

pshare estimate salmen, average

Los resultados se presentan en el cuadro 3.5, en cada uno de los grupos se puede identificar el salario promedio de los trabajadores, en el caso del grupo de 0 a 20 que está representado por los trabajadores formales que tienen los menores salarios, su salario promedio mensual es de \$6 215.86; para el siguiente grupo es de \$11 958.22 y para los trabajadores que más

ganan su salario promedio mensual es de \$138 408.5; como se puede observar, los trabajadores que más ganan tienen 22 veces el salario de los que menos ganan, lo que expone de una forma más puntual las desigualdades que se observaron en los dos cuadros previos.

Cuadro 3.5. *Salario promedio mensual por quintil*

	<i>Percentile shares (average)</i>		<i>Number of obs = 3 752</i>	
	Coefficient	Std. err.	[95% conf. interval]	
salmen				
0-20	6 215.86	102.9807	6 013.958	6 417.766
20-40	11 958.22	232.423	11 502.54	12 413.91
40-60	20 769.39	436.4262	19 913.73	21 625.04
60-80	38 554.38	904.3712	36 781.27	40 327.49
80-100	138 408.5	6 679.873	125 311.9	151 505.00

Fuente: Elaboración propia.

La representación de la información que se presentó en los cuadros previos podría parecer poco desagregada, el comando por default arroja los resultados en forma de cuantil, sin embargo, es posible presentar los resultados en forma de percentil, decil, cuartil, o cualquier otra forma de agrupar los datos que sea de interés a los objetivos del trabajo de investigación. Para este ejemplo, representaremos la información del salario promedio mensual en decil, es decir, la información del salario mensual se presentará en 10 grupos ordenados de los que menos ganan a los que más ganan, la sintaxis queda definida:

pshare estimate salmen, average nquantiles(10)

A diferencia de las sintaxis anteriores, solamente se agregó después de la coma la instrucción **nquantiles(10)** la cual permite señalar al programa que los datos se agruparán en 10 partes, entre paréntesis se pondrá el número de grupos que se pretendan construir, los resultados de esta sintaxis se presentan a continuación en el cuadro 3.6. Esta forma de presentar los datos facilita identificar la forma en que se distribuyen los ingresos a lo largo de los diferentes trabajadores, hasta el decil 80-90 se observa que la distancia entre los grupos no es tan marcada como la que se observa en el último decil, la

dieferencia salarial entre ambos deciles es de aproximadamente 140,000 pesos mensuales; en el cuadro anterior la distancia entre los dos últimos grupos era aproximadamente \$100 000, es por ello que al desagregar los grupos, podemos identificar con mayor precisión entre aquellos de menores salarios y de los de mayores salarios. En este cuadro, hasta 40 % de los trabajadores percibe en promedio salarios de \$13 405.45; en el cuadro anterior esta misma proporción de trabajadores percibía en promedio \$11 958.22, en este sentido, entre mayor desagregación, podríamos conocer de forma más precisa la composición de los grupos.

Cuadro 3.6. *Salario promedio mensual por decil*

salmen	Percentile shares (average)		Number of obs = 3,752	
	Coefficient	Std. err.	[95% conf. interval]	
0-10	5,417.555	3.992097	5,409.728	5,425.382
10-20	7,014.169	203.1562	6,615.862	7,412.476
20-30	10,511	139.6783	10,237.14	10,784.85
30-40	13,405.45	344.5133	12,730	14,080.9
40-50	17,844.47	402.4118	17,055.5	18,633.44
50-60	23,694.31	498.0268	22,717.88	24,670.73
60-70	32,008.51	778.3365	30,482.5	33,534.51
70-80	45,100.26	1,098.475	42,946.59	47,253.92
80-90	67,235.24	1,791.786	63,722.27	70,748.21
90-100	209,581.7	12,497.57	185,079	234,084.4

Fuente: Elaboración propia.

Revisando la estructura de los salarios a través de los grupos, surge el interés de establecer el total de ingresos que acumula cada uno de los grupos en lugar de sólo conocer el salario promedio. Esto se puede lograr a través de sustituir en la sintaxis **average** por **sum**, de la siguiente manera:

pshare estimate salmen, sum gini n(10) cformat(%9.0fc)

Además, a la sintaxis se anexó el comando **Gini** que es un estadístico que permite conocer el grado de concentración de los ingresos a través de los

diferentes grupos, su valor oscila entre 0 y 1, entre más cercano a 1 los ingresos están concentrados, mientras que entre más se acerque a cero el ingreso se distribuye de mejor forma entre los diferentes grupos, en este caso deciles. Al solicitar el total de ingresos en lugar del salario promedio, se incrementa el total de dígitos que se emplean en la tabla; para evitar que los resultados se presenten en forma de notación científica y facilitar su interpretación es necesario incluir en la sintaxis `cformat(%9.0fc)`, la cual fija el número de dígitos que debe tener cada uno de los resultados separados por comas, los resultados se presentan en el cuadro 3.7. Algunos de los datos no contienen comas, lo cual parecería un error, pero no es así debido a que el número de dígitos considerando las comas, se marcó en 9 de tal manera que aquellos que tengan más dígitos, considerando las comas, tendrán un formato distinto; en nuestro caso de los deciles de 60-70 en adelante se han omitido las comas para cumplir con esta regla. Recordemos que el objetivo es evitar que los resultados aparezcan en forma de notación científica, por ello se han fijado a 9 dígitos.

Cuadro 3.7. Total de ingresos mensuales por decil

	Percentile shares (sum)		Number of obs = 3,752	
	Coef.	Std. Err.	[95% Conf.	Interval]
salmen				
0-10	2 032 667	1 498	2 029 730	2 035 603
10-20	2 631 716	76 224	2 482 271	2 781 161
20-30	3 943 726	52 407	3 840 977	4 046 476
30-40	5 029 724	129 261	4 776 295	5 283 153
40-50	6 695 245	150 985	6 399 225	6 991 266
50-60	8 890 103	186 860	8 523 747	9 256 460
60-70	12 009 592	292 032	11 437 035	12 582 149
70-80	16 921 616	412 148	16 113 560	17 729 672
80-90	25 226 663	672 278	23 908 597	26 544 729
90-100	78 635 057	4 689 090	69 441 642	87 828 471
	Gini			
salmen	0.598189			

Fuente: Elaboración propia.

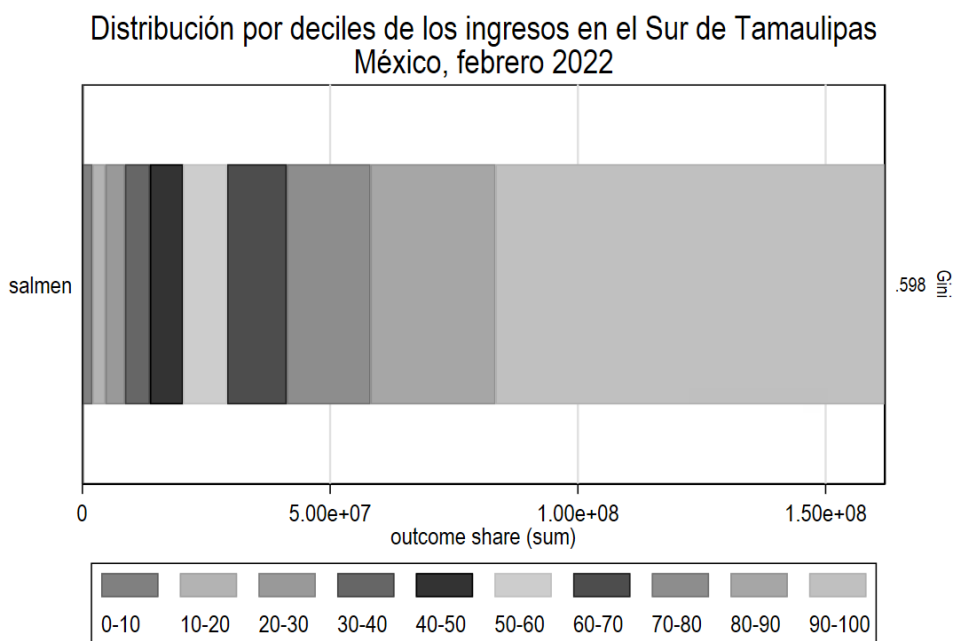
De los resultados del cuadro 3.7 se observa que la concentración es alta debido a que se obtuvo un estadístico de Gini de 0.60 siendo más cercano a 1 que a cero. Los trabajadores que perciben menores salarios concentran

\$2 032 667 mensuales, mientras que los trabajadores que se ubican en el decil 60-70 concentran \$12 009 592 mensuales, en el último grupo se ubican los trabajadores que ganan los más altos salarios, estos concentran \$78 635 057 mensuales. La suma de todos los ingresos es el valor mensual de los ingresos que generan los trabajadores formales en la zona analizada.

Estos últimos resultados se pueden observar de manera gráfica utilizando la siguiente sintaxis:

```
pshare stack, barwidth(.6) scheme(s1mono) title("Distribución por deciles de los ingresos en el Sur de Tamaulipas" "México, febrero 2022")
```

Gráfica 3.1. Total de ingresos acumulados mensuales por decil



Fuente: Elaboración propia.

En este caso se omite la variable de interés, debido a que el comando tomará como referencia los resultados del último cuadro que generó la gráfica 3.1 se representa en forma acumulada en orden del primer decil al

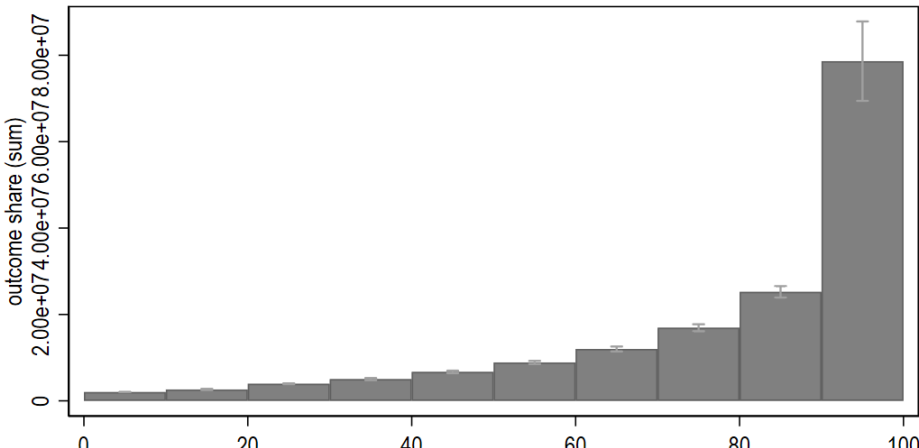
décimo decil. Las opciones después de la coma permiten especificar las características de las gráficas, la instrucción **barwidth(6)** permite asignar la amplitud de las barras, mientras que **scheme(s1mono)** especifica el tipo de gráfica a emplearse, **title** permite asignarle nombre al gráfico, en este caso las comillas permiten subdividir el nombre en dos partes para que no se vea aglomerado en una sola línea. La gráfica 3.1 es complementario al cuadro 3.7, ya que permite de forma visual establecer la distribución de los ingresos por cada uno de los deciles. En esta gráfica también se incluye el valor del estadístico de **Gini**.

La gráfica previa también se puede representar en forma de histograma, esto debido a que el objetivo podría ser analizar de forma independiente cada decil y poder comparar la altura de los deciles. Para construir un histograma de los deciles se presenta la siguiente sintaxis:

```
pshare histogram, scheme(s1mono) title("Distribución por deciles de los ingresos en el Sur de Tamaulipas" "México, febrero 2022")
```

Solamente se sustituye **stack** por **histogram**, la gráfica muestra los resultados. En este gráfico se observa claramente la desproporcionalidad de ingresos que acumula el décimo decil con respecto al resto. Otra de las diferencias de esta gráfica con la previa es que es posible establecer los intervalos de confianza de cada uno de los deciles.

Figura 3.2. *Histograma de total de ingresos acumulados mensuales por decil*
 Distribución por deciles de los ingresos en el Sur de Tamaulipas
 México, febrero 2022
 Gini = .598



Fuente: Elaboración propia

Otro de los aspectos que podría ser importante cuando se analiza la distribución de los ingresos, es el de establecer cómo se comporta su distribución a través de los diferentes sectores económicos o, tal vez, el interés se centre en las diferencias entre los municipios. Para este análisis, estableceremos la variable sexo para identificar las diferencias en la distribución del ingreso entre hombres y mujeres. Un aspecto que se tiene que considerar es que las variables categóricas no deben ser variables **string**, deben ser variables numéricas para ser consideradas en la sintaxis; dado que nuestra variable de sexo en la base de datos es **string**, la convertiremos a través de la siguiente sintaxis en numérica donde el valor de 1 representa hombres y el valor 2 representa mujeres:

```
encode sexo, gen(sex)
```

Esta sintaxis crea una nueva variable con estos valores llamada *sex*. Ahora se incorpora en la sintaxis para crear el cuadro por grupo quintil, en donde la variable de análisis sigue siendo *salmen*, en la parte de las opciones, se solicita el promedio del salario, el estadístico de la **Gini**, también se señala el tipo de formato de los resultados incluyendo la coma para separación de miles, y posteriormente a través de la opción de *over* se incorpora la variable categórica de *sexo* y se incluye la instrucción *total*, con el objetivo de incluir la distribución por sexo y la distribución global, la sintaxis queda definida como:

```
pshare estimate salmen, average gini n(5) cformat(%9.0fc)  
over(sex) total
```

Los resultados aparecen en el cuadro 3.8, no sólo existen diferencias importantes entre los cuantiles, sino que también es notable la diferencia entre los salarios promedio mensuales que perciben las mujeres con respecto a los hombres. En el quintil más bajo donde se ubican los trabajadores con menores ingresos podemos observar que las mujeres tienen un salario promedio de 5,715 mientras que el de los hombres es de 6,715, es decir,

entre ambos existe una diferencia de \$1,000. Para los trabajadores que perciben los salarios más altos la situación es similar, las mujeres tienen un salario promedio de 95,759 y los hombres de 161,916, la diferencia es de aproximadamente \$65,000. De esta manera es posible establecer que las diferencias no sólo se acrecentan entre los que menos ganan y más ganan, sino que también entre hombres y mujeres existe una brecha salarial importante. Los resultados totales que aparecen en el cuadro 3.8 hacen referencia a los resultados que se señalaron en el cuadro 3.5. También, en la parte final se presentan los resultados del estadístico de **Gini** tanto para los hombres como mujeres, y el valor global. Este estadístico señala que las diferencias en el salario promedio mensual para los hombres son mayores, mientras que el caso de las mujeres es menor.

Cuadro 3.8. *Salario promedio mensual por quintil y sexo*

<i>Percentile shares (average)</i>		<i>Number of obs = 3,752</i>			
1: sex = Hombre					
2: sex = Mujer					
salmen	Coefficient	Std. Err.	[95% conf.	nterval]	
1					
0-20	6,715	169	6,383	7,046	
20-40	13,328	321	12,698	13,958	
40-60	24,146	684	22,805	25,487	
60-80	45,140	1,357	42,479	47,800	
80-100	161,916	9,572	143,149	180,683	
2					
0-20	5,715	78	5,561	5,869	
20-40	10,259	283	9,703	10,814	
40-60	16,690	574	15,565	17,816	
60-80	29,384	1,015	27,395	31,373	
80-100	95,759	7,597	80,865	110,653	
total					
0-20	6,216	103	6,014	6,418	
20-40	11,958	232	11,503	12,414	
40-60	20,769	436	19,914	21,625	
60-80	38,554	904	36,781	40,327	
80-100	138,408	6,680	125,312	151,505	

	Gini
1	0.602803
2	0.5597082
total	0.598189

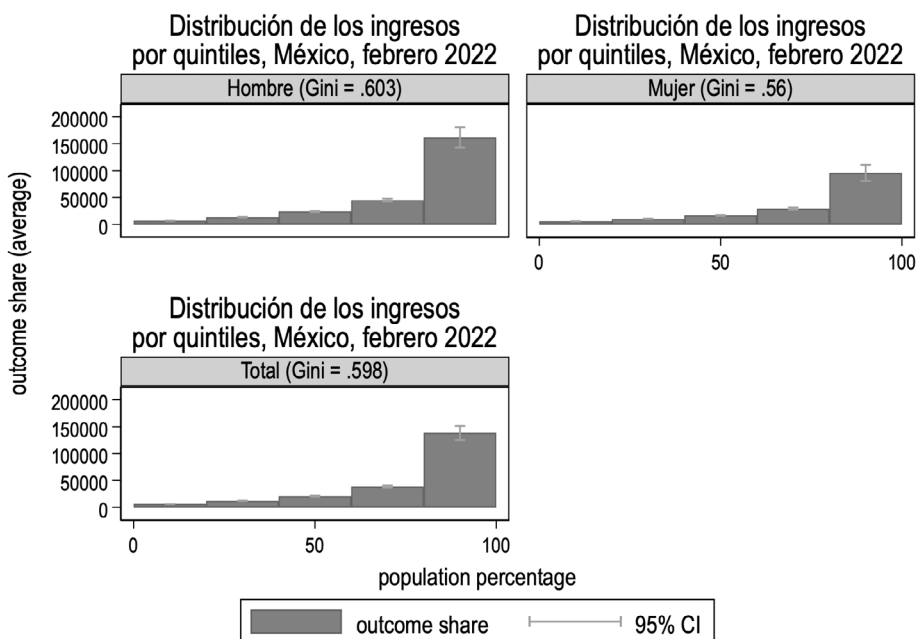
Fuente: Elaboración propia.

La clasificación de la distribución de los ingresos por sexo también se puede graficar tanto por el gráfico acumulado como por el gráfico de histograma; a continuación, se presenta la sintaxis para la gráfica de histograma:

pshare histogram, scheme(s1mono) title("Distribución de los ingresos" "por quintiles, México, febrero 2022") ylabel(, angle(horizontal))

El resultado se presenta en la gráfica 3.3, para obtener el gráfico acumulado únicamente se sustituye **histogram** por **stack**. Es importante que al incorporar variables categóricas como el sexo, se considere que la presentación de los gráficos puede resultar aglomerada, es por ello que en este gráfico se consideró colocar la leyenda del los valores del eje de y en forma horizontal, a través de la instrucción **ylabel(, angle(horizontal))**.

Gráfica 3.3. Histograma del salario promedio mensual por decil y sexo



Fuente: Elaboración propia.

Los resultados, tanto en forma de cuadro como en forma gráfica en el análisis cuantílico, resultan muy útiles para identificar la distribución de los ingresos. El abanico de oportunidades que proporciona el comando **pshare** es mayor, en este apartado se han presentado los más relevantes.

4. El modelo de Solow con datos de corte transversal

Introducción

Uno de los pilares teóricos de la economía es la macroeconomía. Dentro de esta se desarrolla la teoría del crecimiento económico, cuyo objetivo es estudiar las fluctuaciones de la producción real y potencial a muy largo plazo. El estudio del crecimiento económico comenzó a ser de máxima relevancia a partir del primer cuarto del siglo XIX, a raíz de la Gran Depresión de 1929-1933 en los Estados Unidos. Una de las principales incógnitas de ese tiempo era la velocidad con la que se tenía que recuperar el capital desgastado y emplear a los desempleados para mantener o superar el nivel de producción de un año anterior.

En 1956, Robert Solow hizo una de las contribuciones más relevantes a la teoría del crecimiento económico en la que establecía que el progreso tecnológico era una de las causas que permitían acumular capital de forma sostenida a las economías (Solow, 1956). El trabajo de Solow es hoy en día uno de los trabajos más citados y constituye una lectura obligada para cualquier estudiante de economía. Posterior al trabajo de Solow, surgieron cualquier cantidad de trabajos que reutilizaron los postulados de este académico para darle forma a otras teorías, mientras que a finales del siglo XIX aparecieron las primeras estimaciones con datos reales con base en las variables de dicho modelo.

El éxito del modelo Solow puede atribuirse a dos factores principalmente: (a) permite transitar de un modelo teórico a uno empírico de manera

natural mediante manipulaciones algebraicas sencillas; (b) la sencillez en la especificación y el requerimiento de las variables que componen el modelo.

Casi cualquier libro de macroeconomía o de crecimiento económico contiene el desarrollo del modelo neoclásico de crecimiento de Solow, por ejemplo, puede consultarse Acemoglu (2009) y Weber (2010), mientras que, uno de los artículos más consultados es el de Mankiw *et al.* (1992).

En este documento estimaremos la ecuación (7) que puedes encontrar en Mankiw *et al.* (1992, p. 411) que se deriva de tomar logaritmos sobre la ecuación fundamental de Solow alrededor del estado estacionario, la cual se define de la siguiente forma:

$$\ln \left(\frac{Y}{L} \right) = \alpha + \frac{\alpha}{1 - \alpha} \ln (s) - \frac{\alpha}{1 - \alpha} \ln (n + g + \delta) + \epsilon$$

En donde s es la tasa de ahorro, mientras que $n + g$ es la tasa a la que crece la población y la tecnología de manera exógena, y δ es la tasa a la que se desgasta el capital. Este modelo permite corroborar algunos elementos de la teoría con datos reales, por ejemplo, la calibración del modelo de Solow predice que, si $\alpha = 1/3$, es decir, la elasticidad del ahorro respecto al producto es de 0.5, mientras que la elasticidad de la tasa exógena a la que crece la población y a la que se deprecia el capital respecto al producto se predice que sea de -0.5.

Fuente de datos

Con base en lo anterior, para llevar a cabo el ejercicio empírico, lo primero que necesitamos son los datos que conforman las variables del modelo. Comenzando por la variable dependiente (Y/L) es la producción por habitante, lo que quiere decir que esta variable se compone de dos variables: la producción, equivalente al producto interno Bruto (PIB), y la población. En algunos trabajos utilizan la población ocupada, puesto que es la única capaz de participar en el proceso productivo. Para simplificar el ejercicio utilizaremos el PIB per cápita. Posteriormente, la variable s que equivale al ahorro se obtiene de restar el consumo al PIB, en otros trabajos suele utilizarse la

proporción que representa la inversión en el PIB. Finalmente, la variable compuesta por $n + g + \delta$ se obtiene al sumar la tasa de crecimiento de la población a 0.05, siguiendo el procedimiento de Mankiw *et al.* (1992). Todas las variables antes mencionadas podemos obtenerlas en la página del Banco Mundial para conformar un conjunto de datos de corte transversal en donde las observaciones corresponden a los países (<https://databank.bancomundial.org/home>).

Al descargar la información, la página del Banco Mundial ofrece varias opciones respecto al formato en el que deseamos descargar el archivo con la información. Recomendamos que se prefiera el formato .csv separado por comas debido a que suelen ser archivos de tamaño reducido lo que facilita su importación en Stata.

Una vez descargado el archivo, es necesario importarlo en Stata, puesto que no es un archivo nativo de este software, no podemos simplemente abrirlo con el comando `use`, sino, más bien, es necesario que utilicemos el comando `import`.

```
import delimited "C:\Users\...\WorldBankDataset.csv"  
(9 vars, 217 obs)
```

Además, Stata siempre nos da el mensaje de cuántas variables se importaron y la cantidad de observaciones.

Descripción de los datos

Lo primero que debemos de tener en cuenta es que las variables se encuentran en niveles, ya que son los datos que descargamos directamente de la página de internet del Banco Mundial. No obstante, antes de hacer alguna transformación a las variables es recomendable que observemos sus principales características mediante la obtención de algunos estadísticos para conocer la distribución de las variables.

En este caso hemos seleccionado tres variables, que son el PIB per cápita (*gdppc16*), el ahorro interno bruto (*s16*) y la tasa de crecimiento de la población (*n16*). Estamos interesados en conocer la distribución de cada una, para

ello el valor mínimo, el valor en el primer cuartil, la media, la mediana, el valor en el tercer cuartil y el valor máximo son útiles para conocer este aspecto de las variables. Finalmente, también es útil conocer la cantidad de datos que tiene cada variable. Una manera sencilla de representar esta información es a través del comando **tabstat** por medio del que podemos obtener los estadísticos deseados.

tabstat gdppc16 s16 n16, s(min p25 mean p50 p75 max n)

. tabstat gdppc16 s16 n16, s(min p25 mean p50 p75 max n)

stats	gdppc16	s16	n16
min	794.6046	-60.37849	-3.066274
p25	4356.138	9.249532	.5004802
mean	20620.41	18.15926	1.294304
p50	12890.55	20.00003	1.144029
p75	29087.23	27.89573	2.155312
max	117335.6	63.19872	4.845614
N	192	171	215

Esta información permite prever algunos aspectos importantes. Primero, el cálculo de logaritmos está indefinido para variables con valores negativos, por lo que esperamos que algunas observaciones de las variables ahorro y crecimiento de la población no sean consideradas, dado que en los valores mínimos pueden observarse cantidades negativas. Otro aspecto de interés es la cantidad de observaciones por variable; dado que no todos los datos están disponibles en todos los países, tenemos algunas variables como el PIB per cápita que sólo estuvo disponible en 192 países, mientras que el ahorro estuvo en 171 y la tasa de crecimiento de la población en 215. Esta falta de datos tiene repercusiones a la hora de llevar a cabo estimaciones, debido a que sólo se consideran las observaciones cuyos valores están presentes en todas las variables.

Una vez hecha la descripción de las variables podemos proceder a la transformación de estas. Para este ejercicio es necesario tomar el logaritmo tanto para el PIB per cápita como para el ahorro. En este caso podemos recurrir a la función `for var` que nos permite ejecutar la misma instrucción para un conjunto de variables y utilizarla como sigue:

```
for var gdppc16 s16: gen lnX = log(X)
-> gen lngdppc16 = log(gdppc16)
(25 missing values generated)
-> gen lns16 = log(s16)
(67 missing values generated)
```

La función `for var` nos permite ahorrar algunas líneas de código mediante la ejecución de la misma acción a cualquier cantidad de variables. En este caso seleccionamos `gdppc16` y `s16` para transformarlas en logaritmos. Es importante notar que cada variable creada lleva el prefijo *ln* en el nombre, ya que en la instrucción así se especificó.

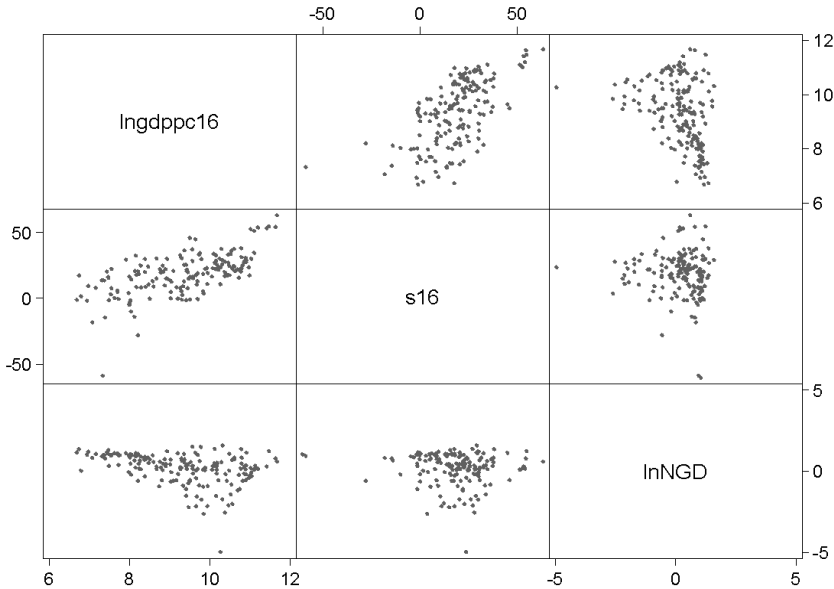
Para la variable de tasa de crecimiento de la población y depreciación es necesario agregar un elemento más a la transformación logarítmica. Recordemos que, siguiendo el procedimiento de Mankiw *et al.* (1992), esta variable se calcula mediante la transformación logarítmica de la suma de 0.05 más la tasa de crecimiento de la población. De tal forma que esta variable puede generarse como sigue:

```
gen lnNGD = log(n16 + 0.05)
(23 missing values generated)
```

Una vez que hemos generado todas las variables necesarias, también podemos analizar la relación entre estas mediante la correlación simple y representarla gráficamente con una gráfica de dispersión. Por medio de ésta podemos establecer la relación que guarda la variable dependiente con sus regresores, así como también la relación entre los mismos regresores. La utilidad de este ejercicio es anticipar una posible condición de colinealidad o multicolinealidad entre variables explicativas, lo cual es indeseable ya que provoca inestabilidad en la inferencia.

El comando que debemos utilizar para obtener la gráfica de matriz de correlaciones es **graph matrix** y lo implementamos mediante la siguiente sintaxis:

```
graph matrix lngdppc16 s16 lnNGD, scheme(s1mono) msiz(tiny)
```



Las dos opciones después de la coma, **scheme(s1mono)** la utilizamos para que la gráfica aparezca en tono monocromático (blanco y negro), mientras que la opción **msiz(tiny)** nos permite establecer el tamaño de los puntos, en este caso, pequeños.

Con todo lo anterior, hemos hecho un análisis descriptivo de nuestros datos y conocemos las generalidades de nuestras variables y su relación entre ellas. Ahora podemos pasar al ejercicio de la estimación.

Estimación

La estimación la llevaremos a cabo mediante el método de Mínimos Cuadrados Ordinarios (MCO), esto lo ejecutamos mediante la función **regress** o **reg** la cual nos devuelve un resultado como el siguiente:

reg lngdppc16 lns16 lnNGD

```
. reg lngdppc16 lns16 lnNGD
```

Source	SS	df	MS	Number of obs	=	134
				F(2, 131)	=	33.16
Model	62.614408	2	31.307204	Prob > F	=	0.0000
Residual	123.672505	131	.944064926	R-squared	=	0.3361
				Adj R-squared	=	0.3260
Total	186.286913	133	1.40065348	Root MSE	=	.97163

lngdppc16	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lns16	.5681673	.0893451	6.36	0.000	.3914214 .7449132
lnNGD	-.4243805	.0835577	-5.08	0.000	-.5896775 -.2590834
_cons	7.896213	.270566	29.18	0.000	7.360969 8.431457

Una estrategia para interpretar los resultados de una regresión es ir de lo general a lo particular. La parte superior de la salida de Stata nos muestra el análisis de varianza en donde nos provee información sobre la suma de los cuadrados tanto del modelo como de los residuales. Del lado derecho, tenemos información que también es de nuestro interés, como la cantidad de observaciones. Ya habíamos anticipado que hacían falta algunos datos en todas las variables. Es importante anotar que la estimación se ejecuta con las observaciones coincidentes en todas las variables. El siguiente dato es el estadístico F que permite probar la validez global del modelo. Este estadístico permite probar la hipótesis de que todos los coeficientes del modelo son cero, en otras palabras, en conjunto todas las variables ejercen un efecto nulo sobre la variable dependiente. Para este caso conviene fijarnos en el valor-p que se encuentra justo debajo del estadístico F, cuyo valor es 0.0000, lo que implica que se rechaza la nula a 1 % de significancia. En consecuencia, se corrobora que la especificación es globalmente válida.

El siguiente estadístico es la R-cuadrada, que mide la proporción de las variaciones de la variable dependiente que es explicada por las variaciones de las variables independientes. También es conocida como bondad de ajuste. Es común que estudiantes principiantes le den una importancia sobremedida

a la R-cuadrada; no obstante, debemos tener en mente dos aspectos a considerar: primero, que cuando la especificación del modelo viene precedida de una fundamentación teórica no hay mucho que hacer, sería inapropiado incorporar más regresores a la especificación; en segundo lugar, para el caso de datos de corte transversal, se presenta una mayor variabilidad en la información relativa a cada variable, lo que dificulta alcanzar valores altos de R-cuadrada. Por lo anterior, es importante apuntar que no existe un valor ideal para la R-cuadrada, más bien podemos pensar en la R-cuadrada como un valor deseado, que entre más se acerque a la unidad es más favorable para argumentar sobre una correcta especificación, aunque, como veremos más adelante, existen pruebas formales para esto. Lo que debemos mantener en mente es que la bondad de ajuste de nuestro modelo aquí estimado es de 0.33, que podría interpretarse como que aproximadamente 33.1 % de la variabilidad de $\ln dppc16$ está explicada por el conjunto de variables consideradas en el modelo. La R-cuadrada ajustada es similar al estadístico de la R-cuadrada usual, con la variante de que penaliza la inclusión de más variables, por lo que está “ajustada” por los grados de libertad. Es natural que la R-cuadrada ajustada sea menor que la R-cuadrada usual, siendo en los modelos de regresión lineal múltiple la adecuada a considerar.

El último estadístico de la columna superior derecha corresponde al error cuadrado medio de la regresión (RMSE, por sus siglas en inglés), el cual es aceptable debido a que se encuentra por debajo de la unidad, sin embargo, este estadístico es influenciado por las unidades de medida en las que se encuentra nuestra variable dependiente.

Pasando a la interpretación de los coeficientes de nuestra regresión tenemos que todos son estadísticamente significativos, incluyendo el intercepto. Esto lo sabemos mediante la observación del valor-p de cada uno de los coeficientes. Debido a que en todos los casos el valor es de 0.0000, rechazamos la hipótesis nula de que las variables no tienen impacto de manera puntual sobre la variable dependiente. El coeficiente de 0.568 corresponde a la elasticidad del PIB per cápita respecto al ahorro, esto lo podemos afirmar debido a que ambas variables se encuentran en logaritmos. Además, el modelo teórico predice que este valor sea igual a 0.5 si es que la participación del capital en el producto se corrobora que es igual a 1/3. Más adelante llevaremos una prueba para determinar esto. Por otro lado, el coeficiente de la

tasa de crecimiento de la población y la tasa de depreciación del capital es -0.424 y la teoría predice que esta elasticidad sea igual a -0.5 . El intercepto lo dejamos sin interpretación debido a que en este ejercicio no es relevante.

Es importante que guardemos los resultados de este modelo en la memoria de Stata, lo cual lo hacemos mediante la ejecución del comando **est store**.

est store SolowM

Hemos decidido ponerle el nombre de *SolowM* a las estimaciones del modelo. Con esto ya quedaron guardados los resultados de nuestra regresión y podremos recuperarlos más adelante sin la necesidad de estimarlo de nuevo.

Antes de pasar a otra versión del modelo de Solow, es conveniente llevar a cabo una prueba de hipótesis para determinar si los coeficientes estimados son iguales a los parámetros que predice el mismo. Dado que el modelo predice una elasticidad de 0.5 del producto respecto al ahorro, condicionamos el valor de $lns16 = 0.5$ y ejecutamos la prueba de Wald mediante el

comando **test**.¹

```
( 1) lns16 = .5          test lns16 = 0.5
```

```
F( 1, 131) = 0.58
Prob > F = 0.4469
```

Lo anterior también es conocido en la literatura de la econometría como prueba sobre restricciones lineales. Basándonos en el resultado de la prueba de restricción lineal, lo que estamos señalando es que la elasticidad del producto per cápita respecto al ahorro es exactamente igual a 0.5 . Considerando que el valor-p es mayor a 0.10 , no rechazamos esta hipótesis, lo cual quiere decir que, a pesar de que el coeficiente obtenido es 0.568 , estadísticamente hablando es igual a 0.5 . Es posible verificar que esta hipótesis también se cumple para el otro coeficiente. La prueba se lleva a cabo de la

¹ Si se tiene que el modelo viene dado por $lngdppc16 = b_1 + b_2 lns16 + b_3 lnNGD + u$, entonces se está realizando el siguiente contraste de hipótesis: $H_0: \beta_2 = 0.5$.

misma forma que la anterior, sin embargo, es importante que se tenga en mente que el valor que predice el modelo de Solow para la elasticidad de la tasa de crecimiento de la población y de la depreciación respecto al producto per cápita es -0.5, es decir, el modelo de Solow predice que la relación entre la tasa de crecimiento de la población y la depreciación del capital es negativa. Económicamente esto significa que mientras más crezca la población y se deprecie el capital, menor será la velocidad a la que la producción se acumule.

Dado que no podemos rechazar en ninguno de los dos casos la hipótesis nula, es posible que aseguremos que nuestro modelo predice los mismos coeficientes que indica la teoría. Ahora podemos hacer una prueba de hipótesis en donde condicionemos ambos valores de manera simultánea, es decir, que el coeficiente de la variable *lns16* es igual al negativo de la variable *lnNGD*, o bien, que $lns16 + lnNGD = 0$. Esto puede llevarse a cabo mediante la siguiente instrucción.

test lns16 = -lnNGD

```
. test lns16 = -lnNGD
```

```
( 1) lns16 + lnNGD = 0
```

```
F( 1, 131) = 1.38  
Prob > F = 0.2423
```

El resultado de la prueba de Wald indica que no rechazamos la hipótesis nula, por lo tanto, ambos coeficientes son estadísticamente iguales. Es posible obtener el valor de α mediante la estimación de una regresión restringida. Lo que tenemos que hacer es generar una variable de la diferencia entre *lns16* y *lnNGD* y correrla sobre el PIB per cápita. Tenemos que dividir el coeficiente entre $1 + \text{coeficiente}$ y el resultado es el valor de α . Con estos datos, al ejecutar el ejercicio, el valor de α es igual a 0.329, lo cual es razonable, dado que los coeficientes resultaron ser similares a lo que la teoría predice.

Diagnóstico de la regresión

La corroboración de la teoría es sólo una parte del trabajo empírico, ahora necesitamos saber qué tan confiable es nuestro modelo con base en las propiedades técnicas. Ya que tenemos los resultados de la regresión, es necesario llevar a cabo el diagnóstico para determinar si se cumplen los supuestos del modelo de regresión lineal múltiple.

Lo que nos interesa ahora es la inferencia, es decir, la confiabilidad con la que podemos sostener nuestras afirmaciones a partir de las estimaciones. Las pruebas de diagnóstico se basan en plantear de forma hipotética que no se ha cumplido, o que se ha cumplido alguno de los supuestos del MLC, por lo que si llegásemos a rechazar que se ha cumplido o que no rechazamos que no se ha cumplido alguno de los supuestos, estaríamos en una situación en donde la confiabilidad de nuestras estimaciones es cuestionable.

Uno de los supuestos del Modelo Lineal Clásico (MLC) es que no hay colinealidad perfecta entre las variables independientes, no obstante, es posible la existencia de colinealidad menos que perfecta, la cual afectará la confiabilidad de las estimaciones dependiendo del grado en que nuestras variables independientes estén correlacionadas.

Para verificar si la colinealidad puede afectar de manera considerable la confiabilidad de las estimaciones, utilizamos el Factor Inflacionario de la Varianza (VIF, por sus siglas en inglés), el cual es un estadístico comúnmente utilizado para dicho fin. Su implementación en Stata se lleva a cabo después de la estimación tal como sigue:

```
estat vif
```

Como regla práctica, si el VIF de una variable es superior a 10, decimos que esa variable es muy colineal; en consecuencia, tendría que implementarse alguna corrección. En el caso de nuestro ejemplo, la colinealidad no es un problema, por lo que no es necesario llevar a cabo alguna acción adicional en este rubro (Gujarati y Porter, 2010).

Sabemos que, teóricamente, a nuestro modelo no le hacen falta variables dado que estamos considerando aquellas que establece la teoría económica,

lo cual es un gran alivio porque incurrir en este problema implica que nuestras estimaciones sean sesgadas.

Una prueba de heteroscedasticidad sirve para determinar si la varianza de los residuales condicionados en las variables independientes es constante, y es de alguna forma una prueba general para identificar si se ha omitido alguna variable relevante en el modelo. Una prueba comúnmente usada en el trabajo empírico lleva el apellido de sus creadores, Breusch-Pagan y puede ejecutarse en Stata mediante el comando `estat hettest` justo después de la regresión, para cuyo caso agregaremos la opción `rhs`, para que condicione los residuales en las variables independientes y no en los valores estimados de la variable dependiente.²

Debido a que la prueba Breusch-Pagan requiere el cumplimiento de la normalidad de los residuales, se puede proceder de dos maneras. La primera es considerar que dado el tamaño de la muestra es de 134 observaciones, y aludiendo al teorema central del límite, este número de observaciones es suficiente para dar sustento a la normalidad de los residuales. La otra forma de probar la normalidad de los residuos es mediante la aplicación de una prueba estadística formal. Al respecto, si bien en la bibliografía estadística existe una cantidad importante de contraste para ello, los economistas han preferido la prueba Jarque-Bera. Para ello la sintaxis correspondiente es:

```
predict res, res  
jb res
```

```
Jarque-Bera normality test: 1.615 Chi(2) .4459
```

```
Jarque-Bera test for Ho: normality:
```

El resultado de la prueba Jarque-Bera muestra que se cumple el supuesto de normalidad por lo que ahora se puede verificar el supuesto de homocedasticidad.

```
estat hettest, rhs
```

```
_____hettest, rhs
```

² Dado que Stata ejecuta la prueba sobre los resultados de la última regresión, es necesario estimar el modelo sobre el cual quiere llevarse a cabo la prueba de hipótesis.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: lnsl6 lnNGD
```


El resultado es claro, más porque Stata, en la salida de la prueba, nos informa cuál es la hipótesis nula. Dado que establecemos que la varianza es constante, al observar el valor-p sabemos que debemos de rechazar esa hipótesis, por lo tanto, la varianza no es constante, y si la varianza no es constante, los residuales no son homoscedásticos o lo que es equivalente, los residuales tienen heteroscedasticidad.

Una vez que sabemos que se tiene un problema de heteroscedasticidad, es necesario llevar a cabo una acción para corregirlo, ya que la inferencia hecha con los resultados del modelo es inválida, es decir, los errores estándar de nuestros coeficientes están influenciados por la correlación de los residuales con las variables explicativas, y, dado que el estadístico t se calcula al dividir el coeficiente entre el error estándar, y a su vez el estadístico t permite determinar si se rechaza la hipótesis nula de que los coeficientes son estadísticamente igual a cero, es necesario garantizar que los errores estándar están calculados de forma apropiada. Existe un método para solucionar este problema de manera sencilla en Stata utilizando errores estándar robustos para heteroscedasticidad. Lo único que hay que hacer es agregar la opción **robust** en la regresión en la que deseamos corregir dicho problema.

reg lngdppc16 lns16 lnNGD, robust

```
. reg lngdppc16 lns16 lnNGD, robust
```

```
Linear regression               Number of obs   =       134
                              F(2, 131)      =       19.31
                              Prob > F             =       0.0000
                              R-squared            =       0.3361
                              Root MSE         =       .97163
```

lngdppc16	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lns16	.5681673	.1823512	3.12	0.002	.2074331	.9289015
lnNGD	-.4243805	.0893393	-4.75	0.000	-.601115	-.247646
_cons	7.896213	.5471181	14.43	0.000	6.813883	8.978543

En esta última regresión se ha corregido el problema de heteroscedasticidad; dado que éste no genera sesgo, los coeficientes son idénticos a los de la primera regresión. La diferencia entre ambos resultados puede notarse en los errores estándar y, por consiguiente, en los estadísticos *t*.³

El modelo de Solow con capital humano

Tal y como se señala en MRW (1992), al modelo de Solow se le puede agregar el capital humano con el objetivo de considerar un factor que también contribuye a explicar la producción, debido a que no sólo el capital físico es importante, sino que también el capital humano lo es. Esto implica agregar una nueva variable al modelo, que denominaremos *school16* y consiste en el porcentaje de la población en edad de trabajar que tiene educación secundaria. Consideramos que esta variable captura de mejor forma la acumulación de capital que la que utilizan en Mankiw *et al.* (1992).

Nuestra estimación del modelo de Solow con capital humano sufre algunos cambios con respecto a la regresión original. La primera diferencia la notamos en el tamaño de la muestra, que ahora es 71, también la R-cuadrada ha disminuido. Los coeficientes han cambiado su magnitud, sobre todo el que se asocia a la variable de crecimiento de la población y la depreciación. Si bien el coeficiente del capital humano tiene el signo esperado, éste no es significativo, por lo que parece ser poco relevante para el modelo.

reg lngdppc16 lnS16 lnNGD lnSchool16

```
. reg lngdppc16 lnS16 lnNGD lnSchool16
```

Source	SS	df	MS	Number of obs	=	71
				F(3, 67)	=	7.67
Model	14.3886043	3	4.79620142	Prob > F	=	0.0002
Residual	41.909159	67	.625509836	R-squared	=	0.2556
				Adj R-squared	=	0.2222
Total	56.2977633	70	.804253761	Root MSE	=	.79089

lngdppc16	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnS16	.4019152	.0955785	4.21	0.000	.2111398 .5926907
lnNGD	-.2082695	.0863481	-2.41	0.019	-.3806211 -.035918
lnSchool16	.2994927	.7922531	0.38	0.707	-1.281851 1.880837
_cons	7.455444	3.340945	2.23	0.029	.7868903 14.124

³ Para una discusión más amplia y técnica sobre este tema, sugerimos revisar Wooldridge (2020).

Vamos a guardar la estimación anterior con el nombre “SolowHC” para utilizarlo posteriormente. Para ello, utilizamos el comando **est store**:

```
est store SolowHC
```

Al realizar las pruebas de normalidad, y Breusch-Pagan para heteroscedasticidad, los resultados muestran que los residuales tienen una distribución normal, así como también varianza constante.

```
jb res1
```

```
Jarque-Bera normality test: 1.08 Chi(2) .5827
```

```
Jarque-Bera test for Ho: normality:
```

Para terminar con nuestro ejercicio de análisis de regresión, complementamos esta sección mostrando una acción que forma parte del análisis de resultados, y es la comparación de nuestros modelos estimados.

Comparación de modelos

Partimos del hecho de que hemos guardado las estimaciones de las regresiones mediante el comando **est store**, por lo que ahora mostraremos una forma de recuperarlas y mostrarlas en un cuadro que nos permita comparar de manera simultánea ambas estimaciones a través del comando **est tab**.

```
est tab SolowM SolowHC, stats(N aic bic) b(%5.3f) star
```

```
. est tab SolowM SolowHC, stats(N aic bic) b(%5.3f) star
```

Variable	SolowM	SolowHC
lns16	0.568***	0.402***
lnNGD	-0.424***	-0.208*
lnschool16		0.299
_cons	7.896***	7.455*
N	134	71
aic	375.528	172.060
bic	384.222	181.111

En el cuadro obtenido vemos los coeficientes de los dos modelos, del lado izquierdo el modelo tradicional de Solow y del lado derecho el modelo de Solow con capital humano. El cuadro permite comparar los coeficientes de manera simple, aunque debemos tener precaución porque cada uno de los modelos están estimados con dos muestras distintas. Mientras el modelo tradicional tiene 134 observaciones, el modelo con capital humano tiene 71 lo que perjudica la comparabilidad entre ambos modelos. No obstante, el modelo con capital humano reúne los requisitos estadísticos para considerarse como una estimación robusta en el sentido de que se corrobora el cumplimiento de los supuestos que garantizan tener los mejores estimadores linealmente insesgados, debido a que no tenemos variables omitidas y los residuales son homocedásticos. Además, en el cuadro se incluyen los denominados criterios de información AIC y BIC. El primero se conoce en la literatura como el criterio de Akaike y el segundo como el criterio bayesiano, ambos nos proveen información similar y nos ayudan a elegir entre dos o más modelos a través de elegir aquel modelo en el que se obtenga el valor más pequeño de estos criterios de información. Con base en lo anterior, el modelo de Solow con capital humano sería nuestro modelo preferido.

5. La utilidad de los modelos de suavizamiento exponencial en el pronóstico de la curva de infectados por COVID-19 en México

Introducción

Los seres humanos siempre hemos estado interesados en conocer lo que nos depara el futuro. Los pronósticos de eventos que aún no han sucedido pueden basarse en evaluaciones subjetivas o en procedimientos numéricos. Los primeros como, por ejemplo, la lectura astrológica o de cartas del tarot; la visión de hechos que aún no han ocurrido a través de una bola de cristal; la interpretación de las burbujas de una copa de champagne o de una taza de café; las prácticas oraculares por parte de pitonisas o sacerdotes son, entre otras, muchas formas que aún existen para pronosticar. Desde un punto de vista científico, la combinación de la experiencia histórica con la aguda visión que proporciona el manejo de herramientas estadístico-matemáticas han ofrecido diferentes métodos como alternativa razonada para lograr adelantarse a ciertos eventos en el futuro. Sin lugar a duda, la predicción se ha convertido en una tarea relevante en muchas áreas de la sociedad en general.

Gran parte de los modelos de predicción involucran el uso de datos de serie de tiempo, esto es una secuencia cronológica de observaciones de una variable de interés por conocer sus valores futuros. Es aquí donde la modelación estadística ofrece representaciones muy simplificadas de cómo opera el mundo real para entonces predecir, aunque sea de manera aproximada, los hechos que han de ocurrir. Toda predicción es un intento de anticipar el futuro y los métodos estadísticos reconocen la posibilidad de cometer errores en sus pronósticos, siendo quizás la mayor ventaja que tiene sobre aquellos

otros procedimientos subjetivos, como los antes señalados, donde no existe ninguna medida posible de error.

Este capítulo ilustra tres métodos de suavizamiento exponencial con el propósito de predecir la evolución del número de personas infectadas con SARS-CoV-2, a nivel nacional y por entidad federativa. La fuente de información empleada corresponde a las cifras oficiales del Instituto Nacional de Estadística y Geografía (INEGI) cubriendo el periodo del 22 de enero al 15 de septiembre de 2020. Se busca ofrecer al lector los elementos básicos para entender con claridad suficiente el análisis que aquí se presenta utilizando las herramientas que ofrece el paquete estadístico Stata.

Tipos de modelos estadísticos de pronóstico

Desde la óptica de los métodos de series de tiempo, se distinguen dos tipos de predicciones: condicionales e incondicionales. Las primeras se realizan mediante modelos causales donde una o varias variables provocan o condicionan el comportamiento de otra variable en estudio. Para ponerlo en un sentido matemático, la(s) variable(s) independiente(s) condicionan a la variable de respuesta o dependiente. Por su parte, las predicciones incondicionales son aquellas que se realizan sin consideración de ninguna otra(s) variable(s) para conocer el comportamiento de la que desea pronosticar, es decir, sólo es necesario conocer las propiedades mismas de la serie que desea pronosticar para conocer su evolución futura.

Las predicciones incondicionales suelen estar basadas en dos enfoques alternativos: el determinista (clásico) y el estocástico (moderno). El enfoque determinista utiliza tendencias, promedios y otras herramientas similares aplicadas sobre los valores actuales, pasados y futuros de la propia serie a predecir. Este tipo de abordaje resulta ser de una naturaleza más exploratoria donde su uso no considera ningún tipo de modelo paramétrico y su justificación puede no emplear totalmente aspectos probabilísticos e inferenciales.¹ Por su parte, el enfoque estocástico también utiliza a los valores

¹ Algunos autores como Gardner (2006) y Hyndman *et al.* (2008) han incorporado diversos aspectos probabilísticos e inferenciales a los métodos de suavizamiento exponencial tratando de mejorar a esta clase de modelos.

actuales, pasados y futuros de la propia serie a pronosticar, sin embargo, su análisis utiliza en gran medida el marco probabilístico e inferencial considerando representaciones paramétricas de la serie en estudio.²

Esta distinción sobre la inclusión o no de elementos probabilísticos y de la inferencia estadística en los modelos ha hecho que algunos autores consideren a los primeros, los determinísticos, como métodos, mientras que los segundos, los estocásticos, realmente como modelos, distinción que a menudo se mantiene y que conlleva importantes implicaciones en la forma en que se aborda la tarea de pronosticar³. Sin la intención de entrar en un debate, en este capítulo consideramos que ambas representaciones corresponden a las de un modelo estadístico. De cualquiera forma, en ambos enfoques —determinista o estocástico—, la capacidad de hacer predicciones sobre los valores futuros de la serie descansa en gran medida en descubrir regularidades, y en la importante decisión que debe realizar el investigador para determinar correctamente el tipo de abordaje que utilizará en un problema de pronóstico concreto.

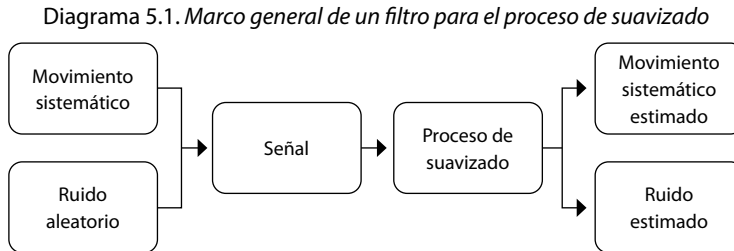
Modelos de suavizamiento exponencial

Los modelos de suavizado exponencial están clasificados dentro de los modelos deterministas. Se distinguen por considerar un algoritmo adaptativo de pronóstico o filtro que pondera geoméricamente a los valores de la serie en estudio. Un filtro, en términos del análisis de series de tiempo, es precisamente un algoritmo que toma como entradas a ciertos valores de la serie y separa las señales que presenta la serie, como pueden ser: la tendencia, la estacionariedad, y el ruido o movimientos aleatorios proporcionando salidas con valores estimados de estos componentes. Existen diferentes tipos de filtros que buscan

² El mejor ejemplo de la aplicación del enfoque estocástico está dado por la metodología de Box-Jenkins en el análisis de series de tiempo, y que sirvió de base, en su momento, para el desarrollo de métodos más avanzados de pronóstico.

³ El hecho de incluir aspectos probabilísticos conlleva a la consideración del proceso de generación de datos del que se puede derivar un método de pronóstico. De esta manera, las previsiones que se realizan utilizando una función de pronóstico es la base necesaria para la construcción de intervalos de predicción.

separar la forma funcional intrínseca del ruido de la serie. El diagrama 5.1 ofrece una representación esquemática de la operación de un filtro.



Fuente: Elaboración propia.

El principio básico de los modelos de suavizamiento exponencial es que los datos son importantes para el pronóstico, pero los datos recientes son aún más importantes que los del pasado. Algunas de las ventajas de estos modelos son el bajo almacenamiento de datos que requieren el mínimo esfuerzo computacional para su funcionamiento, la identificación clara de las señales de la serie, además de servir como herramienta pedagógica para entender otro tipo de modelos más complejos.

Son tres los modelos de suavizamiento exponencial más utilizados para la elaboración de pronósticos: el que se emplea para modelar series sin tendencia ni variación estacional, denominados modelos de suavizamiento exponencial simple (SES); el que considera que la serie en estudio muestra una tendencia lineal, pero ninguna variación estacional que corresponde al de modelo suavizamiento exponencial de Holt (SEH); y el modelo de Holt-Winters (SEHW) donde la serie presenta una tendencia lineal y una variación estacional (ya sea aditiva o multiplicativa).

(a) Suavizamiento Exponencial Simple (SES)

Este es un método de pronóstico que aplica ponderaciones distintas a las observaciones que sirven de base para realizar sus predicciones. Parte de la idea de que las observaciones más recientes ejercen una mayor influencia que las más remotas por lo que las ponderaciones utilizadas decrecen ex-

ponencialmente. Lo noción anterior se describe adecuadamente con la denominada ecuación de actualización, sin tendencia, siguiente:

$$L_t = a y_t + (1-a)L_{t-1}$$

donde

L_t = valor pronosticado en el periodo t mediante SES.

y_t = valor observado en el periodo $t-1$.

L_{t-1} = valor pronosticado en el periodo $t-1$ mediante SES.

a = constante de suavizamiento.

Dos elementos de la ecuación anterior requieren especial atención. Primero, la estimación inicial, esto es, L_{t-1} ; en segundo lugar, el valor de a que es la constante de suavizamiento y la cual debe ubicarse entre 0 y 1. Para determinar L_{t-1} se ha propuesto calcular el valor promedio a partir de una fracción de la muestra en análisis⁴ (Bowerman *et al.*, 2011; DeLurgio, 2008; Otero, 1993; Peña, 2010). Para la constante de suavizamiento, esto es a , su valor suele definirse considerando si los valores más recientes ejercen mayor influencia en la estimación, entonces la constante de suavizamiento debe ser cercana a cero (en caso contrario deberá ser próxima a 1) o bien, partir de algunos criterios convencionales de bondad de ajuste de un modelo estadístico.⁵

⁴ La fracción considerada para el cálculo del promedio puede variar: algunos autores consideran que con tres o cuatro valores es suficiente (Otero, 1993); otros señalan que el promedio puede oscilar entre 25 y 50% de la muestra.

⁵ Estas medidas son el error medio absoluto (MAD), la suma de los errores al cuadrado (SSE), el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE). Quizás sea importante recordar la distinción entre el proceso de ajuste y el proceso de pronóstico. Como DeLurgio (1998, p. 51) señala, "el proceso de ajuste involucra utilizar los valores pasados para ajustar los parámetros del modelo. En contraste, el proceso de pronóstico se emplea para predecir valores futuros, no conocidos, de la serie". Otros autores señalan a esta diferencia a través de la noción de predicción *dentro* y *fuera* de la muestra. La primera se trata de un ajuste, ya que se cuenta con valores observados reales, mientras que la predicción hace referencia a valores aún no observados.

(b) Suavizamiento Exponencial de Holt (SEH)

Es común observar en las series de tiempo la existencia de ciertos patrones de comportamiento que dan cuenta a componentes tales como la tendencia (T), el ciclo (C) y la estacionalidad (E). La T suele definirse como el incremento (o decremento) persistente en los valores de la serie. Un C se identifica con movimientos de mediano y largo plazo en los valores de la serie (*i. e.* de entre 2 y 5 años, respectivamente) los cuales son recurrentes, pero no periódicos. Por su parte, la E es aquella que sufren los valores de la serie de manera sistemática a lo largo de la misma, es decir, corresponde a un patrón que se repite con una periodicidad conocida (por ejemplo, de manera mensual).

En este sentido, el modelo de suavizamiento exponencial de Holt, SEH, pronostica de manera separada el nivel y la tendencia que muestra la serie, razón por la cual utiliza una constante de suavizamiento para el nivel a , y otra con el objetivo de alisar a la tendencia lineal que muestra la serie b . Así, el método de SEH ajusta tanto al nivel como a la tendencia para cada valor de t , de manera que para pronosticar m periodos futuros se pronostica el nivel, S_p , y se adiciona a la tendencia, b_p , dentro de una ecuación de pronóstico.

De esta manera, la implementación del método de SEH considera dos ecuaciones de actualización (una para el nivel y otra para la tendencia lineal), y otra ecuación de pronóstico, que corresponde al caso una tendencia aditiva, esto es:

$$\begin{aligned} L_t &= ay_t + (1 - a) (L_{t-1} + b_{t-1}) \\ b_t &= b (L_t - L_{t-1}) + (1 - b) b_{t-1} \\ F_{t+m} &= L_t + b_t m \end{aligned}$$

donde

L_t= valor pronosticado para el nivel en el periodo **t**.

y_t= valor observado en el periodo **t**.

L_{t-1}= valor pronosticado en el periodo **t-1**.

a= constante de suavizamiento del nivel.

b= constante de suavizamiento de la tendencia.

bt = tendencia suavizada en el periodo t

m = horizonte del pronóstico

F_{t+m} = pronóstico para m periodos futuros

Como en el caso del modelo SES, se presentan dos cuestiones básicas a ser resueltas: (a) ¿cómo se determinan los valores de las constantes de suavizamiento a y b ? y (b) cuáles deben ser los valores iniciales, L_{t-1} y b_{t-1} , para comenzar el proceso de pronóstico? Para encontrar los valores de las dos constantes de suavizamiento (nivel, a , y tendencia, b), el método más utilizado consiste en buscar la combinación de ambas constantes que minimice el valor de la raíz del error cuadrático medio (RMSE). Por otra parte, para inicializar el proceso de predicción, valores L_1 y b_1 , es común considerar al primer valor de la serie, $L_1 = y_1$, mientras que para b_1 es posible obtener un promedio de las primeras observaciones de y , calcular la pendiente de la serie, entre algunas propuestas.

(c) Suavizamiento Exponencial de Holt-Winters (SEHW)

El modelo de Holt fue extendido para poder capturar el componente de estacionalidad que puede presentarse en una serie de tiempo, dando lugar al llamado modelo de suavizamiento exponencial Holt-Winters (SEHW). Este modelo se sustenta en tres ecuaciones de suavizamiento: una para el nivel, otra para la tendencia y otra para la estacionalidad. La ecuación adicional que se agrega es para considerar a la estacionalidad. Existen dos versiones diferentes del modelo de SEHW, dependiendo de si la estacionalidad se modela de forma aditiva o multiplicativa.⁶

Para el caso multiplicativo las ecuaciones de actualización son las siguientes:⁷

⁶ Si la amplitud del patrón estacional es independiente al nivel de la serie, entonces es apropiado un modelo aditivo. Sin embargo, como apunta Winters (1960, p. 327), con mayor frecuencia ocurre que la amplitud del patrón estacional es proporcional al nivel de serie correspondiendo así a la versión multiplicativa del modelo la cual es considerada en este trabajo.

⁷ Una explicación de cada ecuación de SEHW puede encontrarse en Makidrakis (1997, p. 468).

$$\begin{aligned}
 L_t &= a(y_t / S_{t-s}) + (1-a) (L_{t-1} + b_{t-1}) \\
 b_t &= b(L_t - L_{t-1}) + (1-b)b_{t-1} \\
 S_t &= g (y_t / L_t) + (1-g)S_{t-s} \\
 F_{t+m} &= (L_t + b_{tm}) S_{t-s+m}
 \end{aligned}$$

donde

s= la duración de la estacionalidad (i.e. el número de meses o trimestres en un año). L_t = valor pronosticado para el nivel en el periodo t.

b_t = tendencia suavizada en el periodo t.

S_t = valor del componente estacional en t.

y_t = valor observado en el periodo t.

m= horizonte del pronóstico

F_{t+m} = pronóstico para m períodos futuros

Para poder inicializar las ecuaciones de actualización se requiere conocer los parámetros a , b y g , lo cual suele realizarse minimizando el error cuadrático medio (MSE) para diferentes valores de estos parámetros; también se necesitan estimaciones de L_0 , b_0 y S_0 , esto es, del nivel estacionalizado,⁸ la tendencia suavizada y los factores estacionales de la serie de tiempo en el periodo de tiempo 0. Para L_0 se han propuesto, principalmente, dos formas: la primera, empleando algún procedimiento que permita identificar a los índices de estacionalización,⁹ o bien, utilizar el análisis de regresión con variables *dummy* para desestacionalizar a la serie.¹⁰ El cálculo de la tendencia, b_0 , puede realizarse mediante la pendiente de una regresión lineal con los datos desestacionalizados. Los índices estacionales encontrados puede ser la estimación inicial de S_0 .

⁸ Nótese que la primera ecuación hace referencia a la estimación actual obtenida al desestacionalizar los valores de la serie original.

⁹ En esta parte puede consultarse Anderson *et al.*, 2016, p. 846; Bowerman *et al.*, 2011, p. 708; Lind *et al.*, 2019, p. 581).

¹⁰ Al respecto puede consultarse Gujarati y Porter (2010, p. 290).

La aplicación de los modelos SES, SEH y SHW en el pronóstico de los confirmados por SARS-CoV-2 con Stata

(a) Preparación de la base de datos

El paso inicial consiste en importar el archivo de observaciones diarias de confirmados, que se encuentra en el portal del Instituto Nacional de Estadística y Geografía (INEGI).¹¹ Al extraer los datos se observa que estos se encuentran en formato csv por lo que es necesario transferirla a formato Stata, lo cual se puede lograr utilizando el comando **insheet** el cual lee a un conjunto de datos que no está en formato Stata.¹² Utilizaremos este último procedimiento.

```
cd "<directorio en donde se encuentra el archivo *csv.>"
pwd
insheet using Casos_Diarios_Estado_Nacional_Confirmados_20200918.csv, clear
edit
```

La opción **clear** después de la coma se utiliza para limpiar la memoria de Stata. Así, en caso de que estuviera cargada otra base de datos, al utilizar esta opción se borraría para importar la nueva base de datos.

Al explorar la base de datos mediante el comando **browse**, se observa la existencia de las variables: *cve_ent*, que hace referencia a la clave otorgada por INEGI a cada entidad del país; *poblacion*, que da cuenta del volumen de efectivos en cada estado; *nombre*, que nos ofrece el número de personas confirmadas con SARS-CoV-2, así como una serie de variables, que aumentan en forma unitaria, desde *v4* hasta *v254*. Estas variables corresponden a los casos diarios registrados por fecha, es decir, *v4* corresponde al 12/01/20, *v5* al 13/01/20, y así sucesivamente hasta *v254* al 18/09/20. Lo anterior se debe a que al importar el archivo csv, Stata no pudo identificar la etiqueta

¹¹ <https://coronavirus.gob.mx/datos/#DownZCSV>

¹² Otra forma posible es manipulando el archivo csv en excel para luego importar a Stata dicho archivo.

numérica asignada al tiempo, por lo que le ubica por defecto la etiqueta de variable `var#`, como se muestra en el cuadro siguiente:

Cuadro 5.1. Base de datos en su forma original

Nombre	v4	v5	v6
Aguascalientes	0	0	...
Baja California	0	0	...
Baja California Sur	0	0	...
Campeche	0	0	...
...
...

Este formato de la base de datos en donde el tiempo se encuentra en las columnas y las entidades federativas en renglones no resulta adecuado para el análisis de series de tiempo. Es necesario reacomodar este formato a otro donde los nombres de las entidades se encuentren sobre las columnas y cada renglón corresponda a un valor del tiempo, de manera tal que la base de datos tome la forma siguiente:

Cuadro 5.2. Base de datos ideal para el análisis de series de tiempo

Tiempo	Aguascalientes	Baja California	Baja California Sur	Campeche
12/01/20	0	0	0	0
13/01/20	0	0	0	0
14/01/20

Antes de ello modificaremos la base un poco. En primer lugar, la variable `cve_ent` no está ordenada en forma ascendente por lo que lo ordenamos con el comando `sort`. Eliminamos a la variable `población` con el comando `drop` pues no la utilizaremos en nuestro análisis. También eliminaremos al registro correspondiente a `nacional`, ya que sólo trabajaremos con las entidades federativas, utilizaremos el comando `drop in` para ello.¹³

edit

¹³ Observe la diferencia entre los comandos `drop` y `drop in`. El primero elimina variables, mientras que el segundo elimina registros u observaciones.


```
sort cve_ent  
drop poblacion  
drop in 1
```

Hecho lo anterior, el cambio del formato se realizará utilizando el comando **xpose** el cual permite transponer al conjunto de datos cambiando a las variables en observaciones,¹⁴ esto es:

```
xpose, clear
```

Ahora los renglones son columnas y las que eran columnas son renglones. Al observar la base datos con **browse** se podrá notar que los dos primeros renglones muestran los valores de la variable *cve_ent* y como valores perdidos los nombres de las entidades. Este es un inconveniente del comando **xpose**, pero que se puede resolver fácilmente eliminando estos dos renglones con el comando **drop in**; posteriormente utilizando el comando **rename** para renombrar a la *v1* hasta *v32* por las entidades federativas, y generando la variable de tiempo adecuada. Observe que las etiquetas de las variables *v1* a *v32* se hacen más cortas considerando únicamente dos letras que permitan identificar de manera abreviada y clara a cada una de las entidades federativas. Por su parte, la variable tiempo, *t*, toma un formato diario. A continuación, se muestra parte de la sintaxis para lograr esto:

```
drop in 1/2  
rename v1 ag  
rename v32 zt
```

Después de renombrar las variables correspondientes a cada entidad federativa, generamos la variable que representará el tiempo a través del comando **gen** y utilizando el formato de tiempo de Stata. En este caso, dado que tenemos casos diarios, comenzando el 12 de enero de 2020, incorporamos esta información, tal como sigue:

¹⁴ Aunque también puede emplearse cambiando en sentido inverso, es decir, las observaciones en variables.

```
gen t=mdy(1, 12, 2020)+_n-1
```

El comando anterior sólo genera una variable que tiene poco sentido para nosotros, sin embargo, Stata reconoce perfectamente que se trata de un índice de tiempo. Con el objetivo de que el formato de esta variable sea amigable, ya que también es útil a la hora de elaborar gráficas, le asignamos el formato correspondiente utilizando la siguiente sintaxis:

```
format t %d
```

Finalmente, declaramos a Stata que la base de datos es una serie de tiempo mediante el comando **tsset**:

```
tsset t
```

Por último, los epidemiólogos consideran que durante las primeras etapas de una pandemia, el número de casos diagnosticados, es decir, de personas confirmadas con la enfermedad, sigue en el tiempo un crecimiento exponencial. Por tanto, si se conoce el número de diagnósticos positivos, se puede calcular su número acumulado a lo largo del tiempo t . Es por ello que calculamos la serie acumulada para cada entidad federativa. Para ello se genera una variable cuyo prefijo es la letra c seguido del nombre corto de la entidad respectiva con los comandos **gen** y **sum**. Parte de la sintaxis es la siguiente (observe que todas las entidades federativas siguen el mismo formato de sintaxis):

```
gen cag=sum(ag)  
gen czt=sum(zt)
```

Una forma alternativa para obtener el mismo resultado es mediante la ejecución de **loops**,¹⁵ que son útiles para cuando deseamos repetir una ac-

¹⁵ Los loops o bucles en la programación sirven para codificar tareas repetitivas. En Stata las formas más usuales de hacer loops es a través de los comandos **foreach** y **forvalues**.

ción varias veces. Esto nos ahorra varias líneas de código. La forma de realizarlo es como sigue:

```
foreach var of varlist ag-zt{
gen c`var' = sum(`var')
}
```

(b) Análisis exploratorio de las series

Con la base de datos en forma adecuada, en el sentido del análisis de series de tiempo, lograr una correcta comprensión de la información requiere explorar las series en estudio. Esto se puede realizar mediante gráficos, así como recurriendo a la estadística descriptiva con el fin de identificar patrones de comportamiento en los datos. La aplicación del comando **summarize** (o su abreviación **sum**) nos ofrece la estadística descriptiva básica de las diferentes series, con el propósito de simplificar la introducción de cada una de las variables en la sintaxis a través de seleccionar únicamente la primera y la última variable separadas por un guion.

sum ag-zt

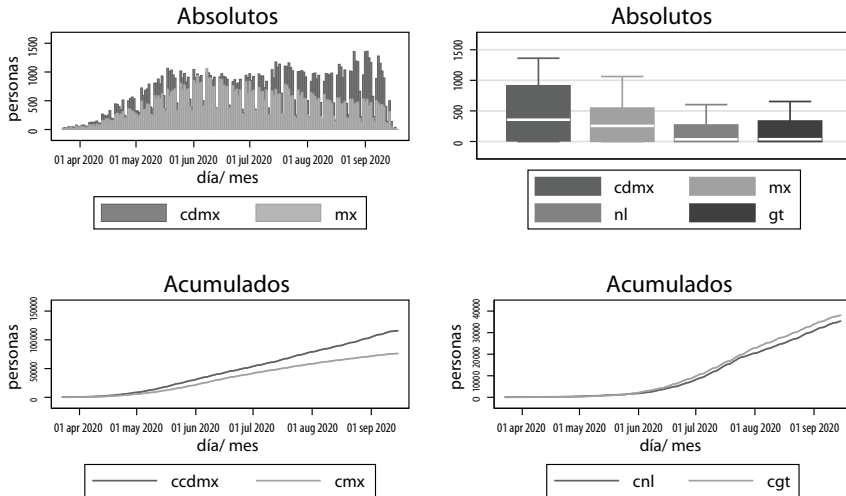
obteniéndose

Variable	Obs	Mean	Std. Dev.	Min	Max
ag	251	26.12351	28.78772	0	119
bc	251	73.56574	65.30818	0	230
bs	251	37.07171	47.052	0	173
cp	251	23.25498	30.39116	0	114
co	251	98.23108	122.8981	0	424
cl	251	17.33466	23.58761	0	80
cs	251	25.58167	36.12584	0	146
ch	251	37.3745	35.10059	0	116

cdmx	251	461.3187	430.2991	0	1361
dg	251	32.03984	40.87941	0	144
-----+					
gt	251	151.5618	189.64	0	654
gr	251	68.45418	73.05681	0	303
hg	251	47.10757	47.32027	0	176
jl	251	96.17131	104.971	0	333
mx	251	303.5538	286.7727	0	1062
-----+					
mi	251	73.40637	83.4167	0	282
mo	251	22.50598	21.5617	0	88
ny	251	22.31076	24.22228	0	85
nl	251	141.4223	177.4956	0	603
ox	251	60.82072	68.33286	0	263
-----+					
pu	251	118.7371	133.3723	0	503
qt	251	32.03586	35.9807	0	149
qr	251	44.78486	48.65233	0	196
sl	251	85.21514	116.7295	0	420
si	251	70.39442	68.32961	0	291
-----+					
so	251	94.33068	114.7279	0	482
tb	251	121.996	133.021	0	462
tm	251	109.6175	136.0715	0	508
tx	251	28.30677	29.98956	0	105
vz	251	125.741	136.4673	0	522
-----+					
yu	251	68.0239	77.81012	0	369
zt	251	26.44223	35.58741	0	172

A partir de estos resultados se seleccionarán aquellas entidades con los números de casos más altos, siendo: Ciudad de México (CDMX), Estado de México (MX), Nuevo León (NL) y Guanajuato (GT).

Gráfica 5.1. *Confirmados SARS-CoV-2 por entidades seleccionadas*
(23 de marzo al 15 de septiembre de 2020)



Fuente: Elaboración propia a partir de INEGI.

Las gráficas que se realizan corresponden a: (a) gráfica de barras (**twoway bar**) y (b) diagrama de caja (**graph box**), a partir de sus casos diarios; y (c) serie de tiempo (**tsline**), considerando los casos acumulados (gráfica 5.1). A fin de ahorrar espacio se combinan estas figuras en un solo gráfico mediante el comando **combine**. Además, dado que sólo nos enfocaremos en un conjunto de días específico, del 23 de marzo al 15 de septiembre, seleccionaremos este rango mediante la condicionante **tin**. La sintaxis para ello es la siguiente:¹⁶

```
twoway bar cdmx mx t in 72/251, title(Absolutos) ytitle(personas)
xtitle(día/ mes) saving(g1) scheme(s1mono)
graph box cdmx mx nl gt, title(Absolutos) saving(g2) scheme(s-
1mono)
tsline ccdmx cmx if tin(23mar2020, 15sep2020), title(Acumula-
dos) ytitle(personas) xtitle(día/ mes) saving(g3)scheme(s1mono)
```

¹⁶ Observe que el condicionante *tin*, permite seleccionar un intervalo de tiempo, de forma tal que *tin(d1, d2)* indicaría considerar sólo datos que cumplen con $d1 \leq t \leq d2$, donde *t* es la variable de tiempo previamente establecida en la instrucción *tsset*.

```

tsline cnl cgt if tin(23mar2020, 15sep2020), title(Acumulados)
ytitle(personas) xtitle(día/ mes) saving(g4) scheme(s1mono)
graph combine g1.gph g2.gph g3.gph g4.gph, title(Confirmados
SAR2-COVID19 por entidades seleccionadas) subtitle((23 de mar-
zo al 15 de septiembre de 2020)) note(Fuente: Elaboración propia
a partir de INEGI) scheme(s1mono)

```

(c) La estimación de los modelos

Se inicia con el modelo `SES` utilizando el comando **tssmooth exponential**. Por lo ya señalado, el parámetro de suavizado α controla la velocidad a la que se ajusta el pronóstico. Valores pequeños de α ajustan los pronósticos lentamente. Debido a la forma de la gráfica de los casos acumulados en las entidades de CDMX, MX, GT y NL se puede sospechar que un valor de $\alpha = 0.5$ podría funcionar adecuadamente. Por otra parte, para cada estado se tomará como valor inicial al promedio de la muestra a partir de la observación 72 correspondiente al 23 de marzo (comando **samp0(72)**) o simplemente **s0(72)**. Una forma de investigar si este primer modelo funciona adecuadamente consiste en comparar de manera gráfica los valores observados con los pronosticados (dentro de la muestra).

La sintaxis para realizar estas acciones se explica a continuación: las primeras cuatro líneas corresponden al suavizamiento exponencial (comando **tssmooth exponential**) de las series para cada una de las entidades federativas consideradas (CDMX, MX, GT y NL). El valor de $\alpha = 0.5$ queda inserto con el comando **p(.5)**. Note además que se genera una nueva serie en cada entidad agregando la letra *c* al inicio de cada nombre. Las siguientes cuatro líneas efectúan las gráficas de serie de tiempo (**tsline**) y se guardan en formato monocromático (comandos **scheme** opción **s1 mono**) bajo los nombres de **g6**, **g7**, **g8** y **g9**. Posteriormente, se combinan las gráficas en una sola imagen (comando **combine**) ubicando título, subtítulo y fuente a esa imagen.

```

tssmooth exponential ccdmx1=ccdmx, p(.5) s0(72)
tssmooth exponential ccmx1=cmx, p(.5) s0(72)

```

```

tsmooth exponential ccgt1=cgt, p(.5) s0(72)
tsmooth exponential ccnl1=cnl, p(.5) s0(72)
tsline ccdmx ccdmx1 if tin(23mar2020, 15sep2020), title(CCD-
MX5) ytitle(casos acumulados) xtitle(día/ mes) saving(g6) sche-
me(s1mono)
tsline ccmx ccmx1 if tin(23mar2020, 15sep2020), title(CMX) ytit-
le(casos acumulados) xtitle(día/ mes) scheme(s1mono) savin-
g(g7)
tsline ccgt ccgt1 if tin(23mar2020, 15sep2020), title(CGT) ytitle(-
casos acumulados) xtitle(día/ mes) scheme(s1mono) saving(g8)
tsline ccnl ccnl1 if tin(23mar2020, 15sep2020), title(CNL) ytitle(casos
acumulados) xtitle(día/ mes) scheme(s1mono) saving(g9)
graph combine g6.gph g7.gph g8.gph g9.gph, title(SES con
alfa=0.5) subtitle((23 de marzo al 15 de septiembre de 2020)) no-
te(Fuente: Elaboración propia a partir de INEGI) scheme(s1mono)

```

Con todo esto se obtuvieron los siguientes resultados:

CCDMX	exponential coefficient =	0.5000
	sum-of-squared residuals =	367521051
	root mean squared error =	1210.1
CMX	exponential coefficient =	0.5000
	sum-of-squared residuals =	163049039
	root mean squared error =	805.98
CGT	exponential coefficient =	0.5000
	sum-of-squared residuals =	53922748
	root mean squared error =	463.5
CNL	exponential coefficient =	0.5000
	sum-of-squared residuals =	47904431
	root mean squared error =	436.87

Las gráficas de los modelos SES muestran resultados aparentemente favorables. Sin embargo, no existe punto de comparación si no se tienen modelos alternativos para cada entidad federativa, donde el valor de la constante de suavizamiento, a , pueda variar. Para tener un punto de referencia

dejaremos que Stata proponga a este parámetro de suavizado realizando todos los modelos que minimice a la raíz cuadrada del error cuadrático medio (RMSE), variante que sólo requiere eliminar de la sintaxis el valor asignado inicialmente a (esto es, $p(0.5)$), así como la opción `s0(72)` que considera entonces a la mitad de la muestra para la estimación del valor inicial (note que se generan los nombres de las entidades con doble cc):

```
tssmooth exponential ccdmx2=ccdmx
tssmooth exponential ccmx2=cmx
tssmooth exponential ccgt2=cgt
tssmooth exponential ccnl2=cnl
```

Las siguientes líneas que se muestran mantienen la estructura que se explicó anteriormente.

```
tsline cdmx ccdmx2 if tin(23mar2020, 15sep2020), title(CCDMX5)
ytile(casos acumulados) xtitle(día/ mes) saving(g11) scheme(s-
1mono)
tsline cmx ccmx2 if tin(23mar2020, 15sep2020), title(CMX) ytile(-
casos acumulados) xtitle(día/ mes) scheme(s1mono) saving(g12)
tsline cgt ccgt2 if tin(23mar2020, 15sep2020), title(CGT) ytile(casos
acumulados) xtitle(día/ mes) scheme(s1mono) saving(g13)
tsline cnl ccnl2 if tin(23mar2020, 15sep2020), title(CNL) ytile(casos
acumulados) xtitle(día/ mes) scheme(s1mono) saving(g14)
graph combine g11.gph g12.gph g13.gph g14.gph, title(SES
con alfa Min. MSE) subtitle((23 de marzo al 15 de septiembre de
2020)) note(Fuente: Elaboración propia a partir de INEGI) sche-
me(s1mono)
```

Los resultados ahora presentan el coeficiente exponencial (es decir, α) que se calcula como óptimo bajo el rubro *optimal exponential coefficient*, los cuales se basan en la minimización del RMSE.

```
ccdmx2      computing optimal exponential coefficient (0,1)
             optimal exponential coefficient =      0.9998
```


	sum-of-squared residuals	=	105361023
	root mean squared error	=	647.89276
ccmx2	computing optimal exponential coefficient (0,1)		
	optimal exponential coefficient	=	0.9998
	sum-of-squared residuals	=	46088410
	root mean squared error	=	428.50807
ccgt	computing optimal exponential coefficient (0,1)		
	optimal exponential coefficient	=	0.9998
	sum-of-squared residuals	=	14772379
	root mean squared error	=	242.59864
ccnl	computing optimal exponential coefficient (0,1)		
	optimal exponential coefficient	=	0.9998
	sum-of-squared residuals	=	12920375
	root mean squared error	=	226.88234

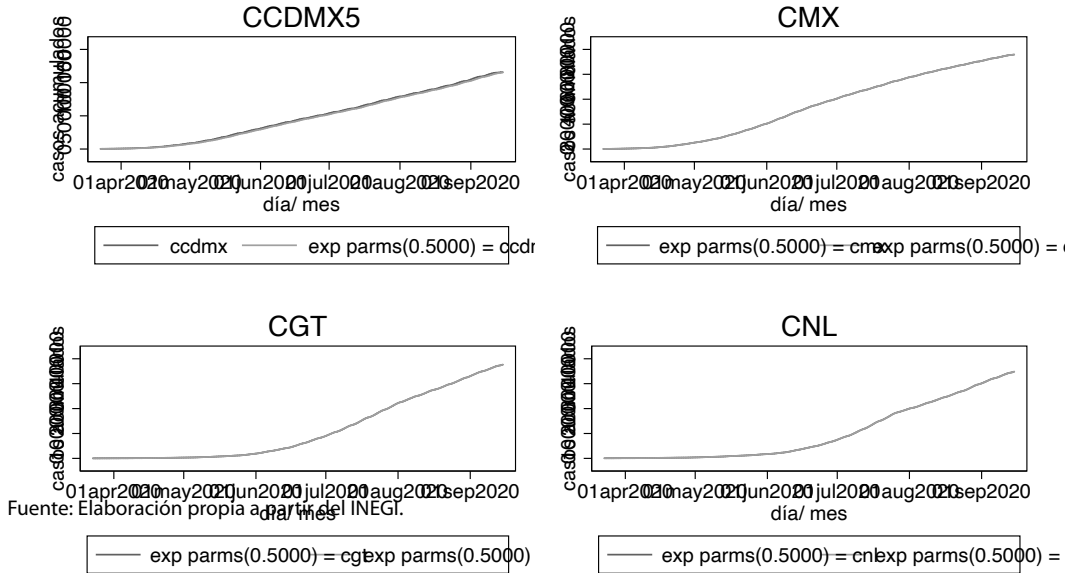
Nuevamente las gráficas 5.3 señalan que los pronósticos de los modelos estimados, optimizando el valor de la RMSE, son adecuados. Es de notar que todos muestran el mismo valor de $a=0.1$ señalando una alta influencia del pasado reciente. El cuadro 5.3 muestra una comparación de los valores de la RMSE, en cada entidad federativa, con los dos valores de la constante de suavización considerados en las estimaciones realizadas anteriormente.

Cuadro 5.3. Comparación de los modelos SES a partir del valor del RMSE

Entidad	RMSE	
	a= 0.5	a= 0.1
Federativa		
CDMX	1210.10	647.89
MX	805.98	428.51
GT	463.50	242.60
NL	436.87	226.88

Gráfica 5.2. SES con $\alpha = 0.5$

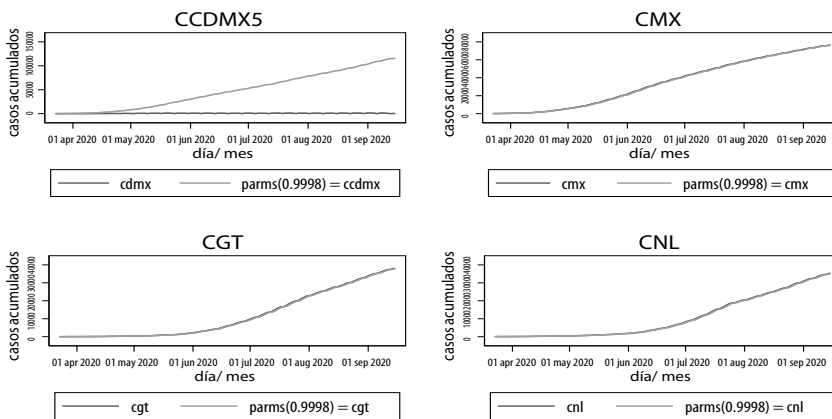
SES con $\alpha = 0.5$ (23 de marzo al 15 de septiembre de 2020)



Gráfica 5.3. SES con α Min. MSE

Fuente: Elaboración propia a partir de INEGI

(23 de marzo al 15 de septiembre de 2020)



Fuente: Elaboración propia a partir de INEGI

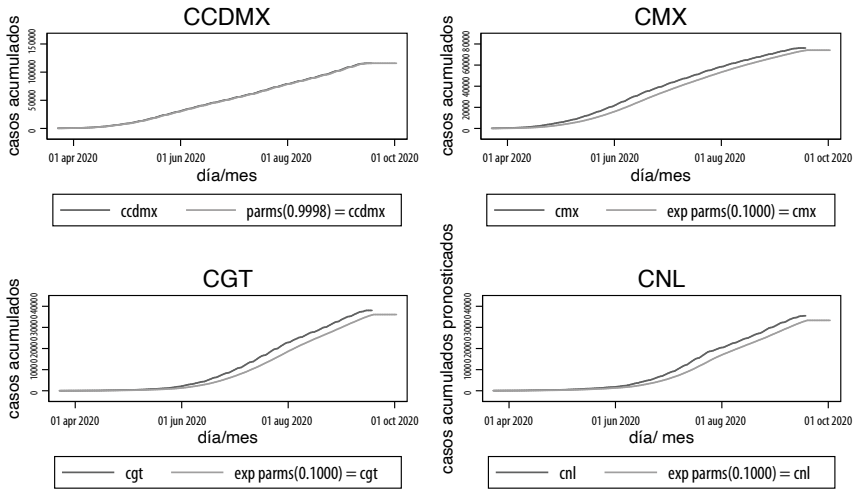
Pasaremos ahora el pronóstico fuera de la muestra con el modelo SES para 15 periodos considerando un $a = 0.1$. Observe que la serie que se generará toma el nombre con una c al principio de la entidad y el número 3 al final de esta. Por otra parte, el comando `forecast` lleva entre paréntesis el número de observaciones que se pronosticarán (en este caso 15). La sintaxis en Stata para todo ello conservando la estructura para la creación de gráficas y combinarlas en una sola imagen es la siguiente:

```

tssmooth exponential ccdmx3=ccdmx, p(.1) forecast(15)
tssmooth exponential ccmx3=cmx, p(.1) forecast(15)
tssmooth exponential ccgt3=cgt, p(.1) forecast(15)
tssmooth exponential ccnl3=cnl, p(.1) forecast(15)
tsline ccdmx ccdmx3 if tin(23mar2020, 2oct2020), title(CCDMX)
ytitle(casos acumulados) xtitle(día/mes) saving(g16) scheme(s-
1mono)
tsline cmx ccmx3 if tin(23mar2020, 2oct2020), title(CMX) ytit-
le(casos acumulados) xtitle(día/mes) scheme(s1mono) savin-
g(g17)
tsline cgt ccgt3 if tin(23mar2020, 2oct2020), title(CGT) ytitle(ca-
sos acumulados) xtitle(día/mes) scheme(s1mono) saving(g18)
tsline cnl ccnl3 if tin(23mar2020, 2oct2020), title(CNL) ytitle(ca-
sos acumulados pronosticados) xtitle(día/ mes) scheme(s1mono)
saving(g19)
graph combine g16.gph g17.gph g18.gph g19.gph, title(Pronós-
ticos a partir alfa 0.1) subtitle((23 de marzo al 15 de septiembre
de 2020)) note(Fuente: Elaboración propia a partir de INEGI)
scheme(s1mono)

```

Gráfica 5.4. Pronóstico a partir de alfa 0.1
(23 de marzo al 15 de septiembre de 2020)



Fuente: Elaboración propia a partir de INEGI

Como se puede observar, los pronósticos proporcionados por el modelo SES, fuera de la muestra, no resultan adecuados ya que mantienen un valor constante para los periodos que se predice. Lo anterior explica por qué todas las series presentan un componente de tendencia que no ha sido considerado.

El modelo SEH ofrece la posibilidad de incluir a la tendencia observada en las series. La estimación de los modelos SEH, como fue señalado anteriormente, requiere de la determinación de las constantes de suavizamiento a y b donde Stata permite la posibilidad de encontrarlos bajo el comando `tssmooth hwinters`, optimizando los valores de la RMSE. A continuación, se muestra la sintaxis para la estimación, dentro de la muestra, de las series bajo estudio considerando la estructura ya señalada anteriormente en cuanto al nombre de las series que se generan, el guardado de las gráficas y la combinación en una sola imagen.

```
tssmooth hwinters ccdmx4=ccdmx
tssmooth hwinters ccmx4=cmx
tssmooth hwinters ccgt4=cgt
```

```

tssmooth hwinters ccnl4=cnl
tsline ccdmx ccdmx4 if tin(23mar2020, 15sep2020), title(CCDMX)
ytitle(casos acumulados) xtitle(día/mes) saving(g21) scheme(s-
1 mono)
tsline cmx ccmx4 if tin(23mar2020, 15sep2020), title(CMX) ytitle(-
casos acumulados) xtitle(día/mes) scheme(s1mono) saving(g22)
tsline cgt ccgt4 if tin(23mar2020, 15sep2020), title(CGT) ytitle(-
casos acumulados) xtitle(día/mes) scheme(s1mono) saving(g23)
tsline cnl ccnl4 if tin(23mar2020, 15sep2020), title(CNL) ytitle(ca-
sos acumulados pronosticados) xtitle(día/ mes) scheme(s1mono)
saving(g24)
graph combine g21.gph g22.gph g23.gph g24.gph, title(Pronós-
ticos de SEH optimizando alfa y beta) subtitle((23 de marzo al 15
de septiembre de 2020)) note(Fuente: Elaboración propia a partir
de INEGI) scheme(s1mono)

```

Los resultados son los siguientes:

Para CDMX	(not concave) Iteration 7 Optimal weights: alpha = 1.0000 beta = 0.1525 root mean squared error = 324.6467
Para CMX	(not concave) Iteration 4 Optimal weights: alpha = 1.0000 beta = 0.1487 root mean squared error = 202.2471
Para CGT	(not concave) Iteration 6 Optimal weights: alpha = 1.0000 beta = 0.1990 root mean squared error = 98.42167
Para CNL	(not concave) Iteration 7 Optimal weights: alpha = 1.0000

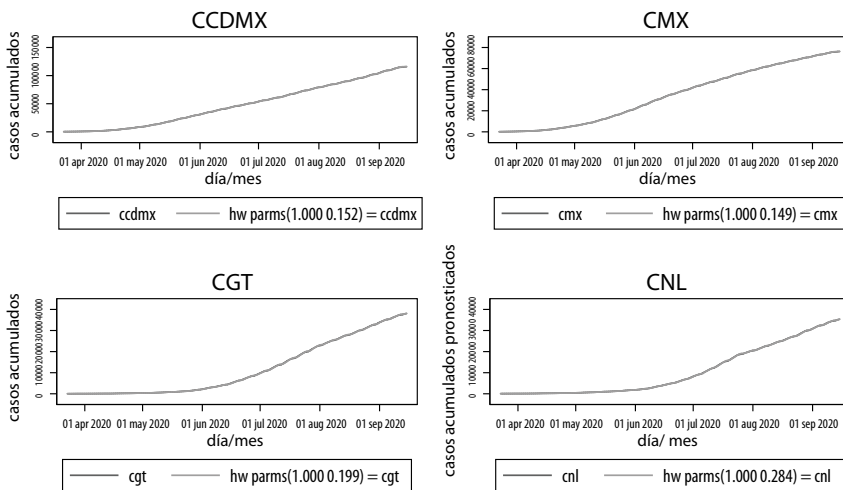
$$\text{beta} = 0.2838$$

$$\text{root mean squared error} = 85.74363$$

Aquí se señala que $a = 1$, en todas las entidades federativas, mientras que b tiene valores pequeños que oscilan entre 0.1525 y 0.2838, siendo esta la estimación de la pendiente en el tiempo. Como se señala en la literatura, la información sobre la pendiente proviene de dos fuentes: la primera, es de la diferencia de las medias estimadas, y, en segundo lugar, de la estimación inmediata anterior de la pendiente. De esta manera, los valores de b pueden dar cuenta de que la tendencia lineal que observan las series no depende principalmente de las últimas observaciones de estas y podrían no cambiar rápidamente.

Por otra parte, la gráfica 5.5 revela que el modelo SHE no necesariamente presentan un buen ajuste dentro de la muestra ya que estos se encuentran encima de los valores registrados por las series.

Gráfica 5.5. Pronósticos de SEH optimizando alfa y beta
(23 de marzo al 15 de septiembre de 2020)



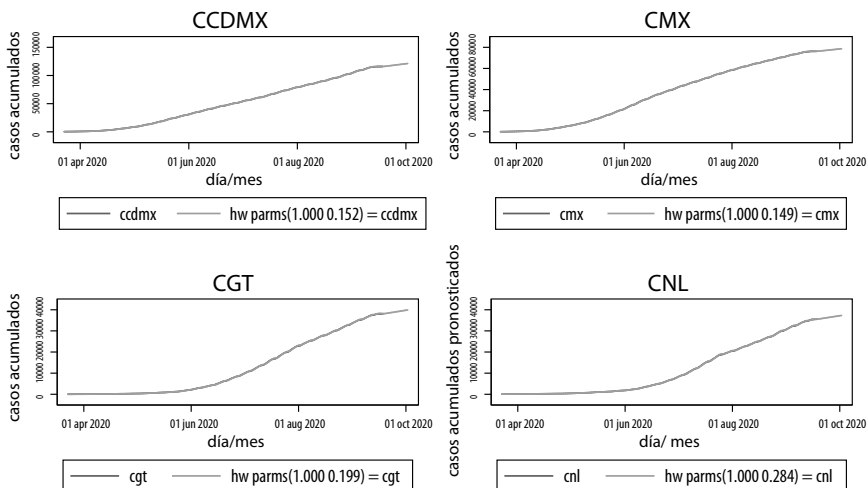
Fuente: Elaboración propia a partir del INEGI.

La mayor prueba de la bondad de este modelo proviene de los pronósticos que se realicen fuera de la muestra. Para realizarlo se emplea la siguiente sintaxis, la cual sólo se ejemplifica para el caso de la CDMX, al ser similar la estructura para las otras entidades federativas consideradas (note, sin embargo, que el comando **combine** considera a las gráficas generadas siguiendo la numeración correspondiente):

```
tssmooth hwinters ccdmx5=ccdmx, forecast(15)
tpline ccdmx ccdmx5 if tin(23mar2020, 2oct2020), title(CCDMX)
ytile(casos acumulados) xtitle(día/mes) saving(g26) scheme(s-1mono)
graph combine g26.gph g27.gph g28.gph g29.gph, title(Pronósticos de SEH fuera de la muestra) subtitle((23 de marzo a 2oct2020)) note(Fuente: Elaboración propia a partir de INEGI)
scheme(s1mono)
```

Sin embargo, a diferencia del modelo SES, las estimaciones de SEH pueden ser considerados como mejores. La gráfica 5.6 da cuenta de esto.

Gráfica 5.6. Pronósticos de SEH fuera de la muestra
(23 de marzo al 2 oct 2020)



Fuente: Elaboración propia a partir del INEGI.

Continuando con la aplicación de modelos en estudio, el SEHW permite añadir el componente estacional al modelo SEH. Esto es, además de ser un procedimiento de suavizamiento exponencial que considera la posibilidad de que la serie tenga tendencia lineal, agrega el hecho de que ésta muestre movimientos temporales de manera sistemática a lo largo del tiempo. Si bien la inspección visual de las gráficas de las series bajo estudio no permite identificar a este tipo de componente —ya sea en forma aditiva o multiplicativa— para llevar a cabo la implementación se recurrió a considerar la serie de tiempo de número de casos diarios para la estimación. Lo anterior se realizó con el propósito de ejemplificar el uso del modelo SEHW en Stata.

La asignación de las constantes de suavizamiento α , b y g —para este caso— consideró un modelo SEHW multiplicativo con un establecimiento óptimo basado en la minimización del RMSE. Cabe notar que ante la ausencia del componente estacional se utilizó la opción `period` que especifica el periodo de la estacionalidad y la cual se hipotetizó en el valor de 4. Adicionalmente, para el caso de Guanajuato fue necesario emplear la opción `altstarts` que utiliza un método alternativo para calcular los valores iniciales de los términos constante, lineal y estacional, calculado los factores de estacionalidad con base a una regresión con variables indicadoras estacionales. La sintaxis en Stata para su aplicación, considerando los valores pronosticados fuera de la muestra para cuatro periodos, viene dada por:

```
tssmooth shwinters ccdmx6=cdmx, period(4)forecast(4)
tssmooth shwinters ccmx6=mx, period(4) forecast(4)
tssmooth shwinters ccgt6=gt, period(4) altstarts forecast(4)
tssmooth shwinters ccnl6=nl, period(4) forecast(4)
tsline cdmx ccdmx6 if tin(23mar2020, 15sep2020), title(CCDMX)
ytitle(casos diarios) xtitle(día/mes) saving(g31) scheme(s1mono)
tsline mx ccmx6 if tin(23mar2020, 15sep2020), title(CMX) ytitle(-
casos diarios) xtitle(día/mes) scheme(s1mono) saving(g32)
tsline gt ccgt6 if tin(23mar2020, 15sep2020), title(CGT) ytitle(ca-
sos diarios) xtitle(día/mes) scheme(s1mono) saving(g33)
tsline nl ccnl6 if tin(23mar2020, 15sep2020), title(CNL) ytitle(ca-
sos diarios pronosticados) xtitle(día/ mes) scheme(s1mono) sa-
ving(g34)
```



```
graph combine g31.gph g32.gph g33.gph g34.gph, title("Pronósticos de SEHW optimizando alfa, beta y gamma") subtitle((23 de marzo al 15 de septiembre de 2020)) note(Fuente: Elaboración propia a partir de INEGI) scheme(s1 mono)
```

Los resultados fueron los siguientes:

```
Para CDMX      Iteration 6
                Optimal weights:
                    alpha = 0.7030
                    beta = 0.0000
                    gamma = 0.0669
                root mean squared error = 255.1267

Para MX        Iteration 8
                Optimal weights:
                    alpha = 0.1518
                    beta = 0.0000
                    gamma = 0.1018
                root mean squared error = 152.5402

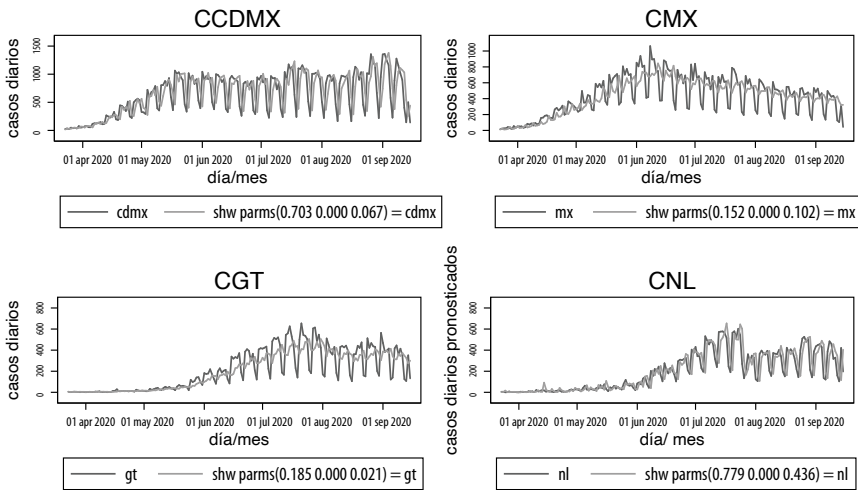
Para GT        Iteration 9
                Optimal weights:
                    alpha = 0.1847
                    beta = 0.0000
                    gamma = 0.0209
                root mean squared error = 99.55393

Para NL        Iteration 9
                Optimal weights:
                    alpha = 0.7793
                    beta = 0.0000
                    gamma = 0.4357
                root mean squared error = 89.73398
```

Como se puede observar, en este caso el valor de la constante de suavizamiento $b=0$ lo que deriva que las series no muestran ningún patrón de tendencia. Por su parte, los pronósticos realizados no resultan ser adecuados debido a la aplicación artificial que se realizó del factor estacional tal y como se muestra en la gráfica 5.7.

Gráfica 5.7. Pronósticos de SEHW optimizando alfa, beta y gamma

(23 de marzo al 15 de septiembre de 2020)



Fuente: Elaboración propia a partir de INEGI

En términos generales, puede considerarse que SEHW no se ajusta lo suficientemente bien a los valores de número de confirmados de forma diaria.

6. Un modelo con variable dependiente trunca en combinación con la técnica de diferencia en diferencias para evaluar una política pública

Introducción

En el presente capítulo se atienden dos retos comunes en el estudio de las ciencias sociales, principalmente en política pública y economía. El primero de los retos consiste en evaluar una decisión de política pública, en particular una intervención del gobierno en un área geográfica determinada. El segundo de los retos tiene que ver con la distribución de la variable dependiente, lo que comúnmente se denomina en la literatura como variable trunca o truncada, también llamada solución de esquina. “Dicha variable es cero para una fracción no trivial de la población, pero es distribuida continua sobre valores positivos” (Wooldridge, 2020, p. 571).

Este capítulo describe el procedimiento para estimar un modelo con variable trunca implementando la técnica de diferencia en diferencias en Stata. Este procedimiento fue utilizado en el estudio de Valdez y Hernández (2019), quienes evalúan el impacto en el consumo de la homologación del impuesto al valor agregado (IVA) en estados fronterizos en el año 2013.

La motivación del estudio surgió porque antes de 2013 en la franja fronteriza norte se pagaba una tasa de IVA equivalente a 11 %, mientras que en el resto del país la misma tasa se cobraba a 16 %. No obstante, a partir de enero de 2013 el Gobierno de México decidió homologar las tasas de IVA a 16 % en todo el país. Debido a que un aumento en un impuesto encarece los productos de forma inmediata, ya que dicho impuesto lo paga directamente el consumidor, surge la necesidad de evaluar el impacto en el consumo de esta decisión del gobierno.

El problema descrito en el párrafo anterior reúne las características necesarias para llevar a cabo un cuasiexperimento, en donde un cambio exógeno, como una decisión de gobierno, afecta sólo a una parte de la población. Esto requiere que se dispongan datos de la población antes de que sucediera el cambio de política pública, así como también después de esta, de tal forma que se pueda comparar si en la población intervenida hubo un cambio posterior a la entrada en vigor de la nueva política, así como también en comparación con la población que no se vio afectada por el cambio. Este tipo de experimentos son comunes en la medicina, principalmente cuando se pretende evaluar la efectividad de algún medicamento o vacuna, en estos se elige una población, a una fracción de esta se le suministra el medicamento, y a otra un placebo. Después se evalúa si la población intervenida sufrió cambios con respecto a cómo estaban antes, y con respecto a los que se les suministró el placebo.

En este sentido, la decisión de homologar el IVA sólo afectó a los residentes de los estados fronterizos, mientras que al resto de entidades no les afectó. Por lo anterior, los residentes de la frontera componen el grupo de tratamiento, mientras que el resto del país compone el grupo de control, y, evidentemente, debe existir un antes y un después.

Sea C el grupo de control y sea T el grupo de tratamiento, sea dT una variable binaria que identifica el grupo de tratamiento asignando el valor de 1 si pertenece al conjunto de observaciones sujetas al cambio en la política pública y cero para el resto. Por su parte, sea $d2$ una variable binaria que identifica el segundo periodo o el periodo que contiene las observaciones una vez entrada en vigor la nueva política, con lo anterior es definida la siguiente ecuación

$$y = \beta_0 + \delta_0 d2 + \beta_1 T + \delta_1 d2 \cdot dT + \mathbf{X}\boldsymbol{\beta} + \varepsilon \quad (7)$$

En donde y es la variable de interés, esa que es afectada por el cambio en la política pública, por su parte, X es un conjunto de variables explicativas del comportamiento de la variable dependiente, ε es un término de error estocástico. El parámetro de interés es δ_1 , denominado el estimador de *diferencia en diferencias (DD)*, el cual mide el efecto de la política (Wooldridge, 2020).

Debido a que pretendemos evaluar el impacto en el consumo, la información de esta variable se encuentra en la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) que levanta el INEGI cada 2 años con representatividad nacional, para el ámbito urbano y rural, y desde la edición 2016 con representatividad estatal.

Unos de los primeros pasos necesarios son unir y depurar las bases de datos ya que cada ejercicio estadístico de las ENIGH genera una estructura de datos de sección cruzada, por lo que es imprescindible unirlos e identificarlas mediante el año al cual se asocian. Mostrar cómo se unen está fuera del alcance del presente capítulo, no obstante, puede explorarse el uso de la función **append** en Stata para llevar a cabo la tarea de unirlos.

Una vez que se tiene la base de datos lista, el primer paso es generar las variables necesarias para el análisis econométrico. Podemos comenzar por las variables binarias requeridas para la técnica de diferencia en diferencias. Una variable binaria que requerimos es la que identifica a los dos grupos, en este caso de entidades federativas.

Utilizamos entonces el comando habitual para generar una variable que llamaremos **fn** (frontera norte) que va a tomar el valor uno si la entidad pertenece al grupo de tratamiento, es decir, si fue afectada por la homologación del IVA, y cero si no lo fue. Una estrategia simple es primero crear la variable con ceros de la siguiente forma:

```
gen fn = 0
```

La instrucción anterior generará una columna con ceros. Ahora reemplazaremos por unos en las entidades que pertenecen a la frontera norte, por ejemplo, para Baja California:

```
replace fn = 1 if ent == 2 // BC
```

En este caso, lo que hicimos fue condicionar el valor a que la clave de la entidad federativa **ent** fuera dos, que corresponde al estado de Baja California. La instrucción debe ser análoga para el resto de las entidades.

Una vez que identificamos los grupos, referimos los diferentes periodos, el que corresponde con el momento antes de la homologación del IVA, y

después de este. Para ello, debemos tener una variable que asocie cada observación con el año, en nuestro caso tenemos sólo dos de estos últimos: 2012 y 2014. Por lo tanto, la variable binaria puede generarse mediante la siguiente función:

```
gen d14 = (2014.yr), after(yr)
```

La instrucción anterior generará una variable binaria que tendrá el valor uno para todas las observaciones de 2014, y cero para las de 2012. La opción `after(yr)` sólo es para ubicarla después de la variable `yr`.

Por último, necesitamos la variable que capturará el impacto del cambio en la política pública en el grupo de tratamiento. Para ello generaremos una variable que denominaremos `did`, y corresponderá al producto de dos variables binarias; las dos que hemos generado, de tal forma que utilizamos la siguiente sintaxis:

```
gen did = fn*d14
```

Adicional a lo anterior, es conveniente deflactar las variables monetarias para eliminar el efecto del aumento de los precios para aislar de manera más convincente lo que deseamos medir. Requerimos primero el Índice Nacional de Precios al Consumidor (INPC) para deflactar las variables. La idea detrás de deflactar las variables consiste en dividir cada valor de nuestra variable monetaria entre el INPC del año. En Stata, esto puede hacerse de la siguiente forma para la variable de consumo:

```
gen consum2 = (gasto_mon/105.279)*100 if yr == 2012  
replace consum2 = (gasto_mon/113.438)*100 if yr == 2014
```

Las dos líneas de código anteriores generan la variable `consum2`, producto de deflactar la variable `gasto_mon` tanto para el año 2012, como del 2014. En el caso del artículo que estamos replicando, deben deflactarse todas las variables que sean necesarias. Por ejemplo, todos los conceptos de gasto que se deseen analizar con el modelo Tobit.

Debido a que vamos a estimar bastantes modelos, lo más conveniente es automatizar el proceso para no repetir la misma instrucción por varias veces, sino que con una instrucción pueda hacerse. En este caso, una vez que deflactamos todas las variables de conceptos de gasto, las vamos a agrupar en un objeto de la siguiente forma:

```
gl depvars = "bebidas2 alifuera tabaco2 vesti2 agua2 energia2  
cuida2 utens2 enseres2 salud2 atenc hospital2 transpub transfor  
adqvehi mtto-otrosrast"
```

Puede hacerse lo mismo con las variables independientes:

```
gl indepvars = "d14 fn did ing2 edad_jefe educjefe tot_integ"
```

Con los objetos definidos procedemos a realizar las estimaciones del modelo Tobit para todas las que se asignaron en el objeto `depvars`. Para ello realizamos un *loop* para repetir la misma acción en repetidas ocasiones.

```
foreach var of varlist $depvars {  
  tobit `var' $indepvars, ll(0)  
  margins, eydx(did)  
}
```

La opción `ll(0)` se especifica debido a que es el valor inferior que se repite en una proporción importante de la variable. En este caso, dado que hay muchos hogares que reportan cero gastos en los distintos bienes o servicios que corresponde con las variables dependientes.

Los resultados que se obtienen de una de las regresiones son como sigue:

Refining starting values:

Grid node 0: log likelihood = -190265.08

Fitting full model:

Iteration 0: log likelihood = -190265.08

Iteration 1: log likelihood = -189284.77

Iteration 2: log likelihood = -189239.21

Iteration 3: log likelihood = -189239.09

Iteration 4: log likelihood = -189239.09

Tobit regression

Number of obs = 28,481

Uncensored = 22,360

Limits: Lower = 0

Left-censored = 6,121

Upper = +inf

Right-censored = 0

LR chi2(7) = 2511.14

Prob > chi2 = 0.0000

Log likelihood = -189239.09

Pseudo R2 = 0.0066

bebi das2	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
d14	-21.78784	13.51875	-1.61	0.107	-48.28522	4.709537
fn	393.5741	24.46956	16.08	0.000	345.6126	441.5356
did	-53.47929	29.55179	-1.81	0.070	-111.4022	4.443625
ing2	.003383	.0001382	24.48	0.000	.0031122	.0036539
edad_jefe	-3.776409	.3965486	-9.52	0.000	-4.553663	-2.999155
educ_jefe	30.34585	2.556744	11.87	0.000	25.33451	35.35719
tot_integ	53.16332	3.030139	17.54	0.000	47.2241	59.10254
_cons	103.027	32.76504	3.14	0.002	38.80596	167.248
var(e. bebi das2)	828585	8094.584			812870.2	844603.7

No obstante, las estimaciones del cuadro anterior no deben interpretarse directamente debido a que el modelo Tobit estima sobre una variable latente y^* , lo cual deja sin interpretación práctica los coeficientes de la regresión.

Los coeficientes que nos interesa interpretar son los efectos marginales promedio, los cuales obtenemos en Stata mediante el comando **margins**. Es por eso, por lo que en el *loop* se especifica la siguiente instrucción:

margins, eydx(did)

Con esa instrucción obtenemos el efecto marginal promedio, en otras palabras, el cambio en proporcional en la variable dependiente ante un cambio infinitesimal de la variable independiente, que, en nuestro caso, nos interesa la variable **did**. Con ello nos da el siguiente resultado:

Average marginal effects Number of obs = 28,481
Model VCE: OIM

Expression: Linear prediction, predict()
ey/dx wrt: did

	Delta-method					
	ey/dx	std. err.	t	P> t	[95% conf. interval]	
did	-.3200989	106.9721	-0.00	0.998	-209.9905	209.3503

Debido a que la variable dependiente de la estimación anterior corresponde al consumo de bebidas, el efecto marginal promedio indica que, la homologación del IVA en los estados fronterizos tuvo un efecto negativo, reduciendo 32 % el consumo promedio de las bebidas en estas entidades con respecto a las entidades que no se vieron afectadas.

Con esto podemos dar por concluido el ejercicio en donde hemos evaluado el impacto de una política pública en el consumo de diversos bienes en los hogares fronterizos. Las posibilidades de utilizar esta técnica y la metodología aquí descrita son variadas, y conforman una opción dentro del amplio espectro de retos que surgen dentro de las ciencias sociales.

7. Modelo de datos panel para estimar el papel de la educación en la desigualdad por ingresos

Introducción

Datos panel es una estructura de datos que consiste en la observación longitudinal de las unidades de análisis. Como bien sabemos, las unidades de análisis pueden ser personas, empresas, escuelas, industrias, ciudades, municipios, entidades federativas, países, por mencionar algunas. Conformarían un panel de datos si de cada una de estas unidades se tiene información para al menos dos periodos distintos. Estos periodos pueden ser semanas, meses, trimestres, semestres, años, lustros, décadas. Si, por ejemplo, se tiene una muestra de familias a las que se les tomaron datos de ingreso familiar, condiciones de la vivienda, número de integrantes, etc., para conformar un panel de datos tendría que visitarse, en un segundo momento, exactamente a las mismas familias que se visitaron en el primer momento.

La estructura de datos en forma de panel tiene la característica de que todos los individuos de la muestra se observan en diferentes periodos; a diferencia de los datos agrupados, los individuos en los datos panel son exactamente los mismos en todos los momentos del tiempo. Con individuos nos referimos de manera general al objeto de estudio, que también puede estar compuesto por países, estados, ciudades, sectores económicos, actividades económicas, empresas, escuelas, por mencionar algunos. En síntesis, una estructura de datos de panel combina datos de sección cruzada y de series de tiempo.

Algunas de las ventajas de trabajar con datos panel es que el tamaño de la muestra es mayor, esto nos permite tener más grados de libertad y, en

consecuencia, podemos estimar más parámetros. Aunado a lo anterior, los datos de panel también nos permiten estimar efectos dinámicos asociados al tiempo, ampliando el alcance del análisis de forma sustancial.

Una de las características más importantes de lo que se podría denominar el método clásico de estimar datos de panel es que se cumpla el supuesto de exogeneidad estricta, lo cual en términos técnicos significa que ninguna de las variables se correlacione con el término de error, ya sea de manera contemporánea o en el tiempo. En términos prácticos, este supuesto refiere que si alguna de las variables explicativas que hemos incluido en nuestro modelo de regresión se relaciona con alguna otra que también es relevante para explicar el fenómeno que estamos intentando explicar, pero que a su vez se deja fuera del modelo, ya sea por ignorancia o por falta de datos, entonces estaríamos violando el supuesto de exogeneidad estricta.

Existen dos clasificaciones de las estimaciones de datos panel dentro de la literatura de la econometría: (a) los supuestos de efectos fijos y (b) los supuestos de efectos aleatorios. Normalmente, los ejercicios econométricos de esta índole se centran en determinar qué clase de supuestos usaremos para nuestras estimaciones, basándonos en criterios de consistencia y eficiencia de las estimaciones, para lo cual existen pruebas formales, una de ellas, la prueba de Hausman.

En este capítulo se muestra cómo realizar estimaciones utilizando datos con estructura de panel con el *software* Stata para analizar el papel que ha tenido la educación en la desigualdad por ingresos en los países, con una muestra de 38 de estos, los cuales observan las mismas variables durante 24 años.

Metodología

La taxonomía general de los modelos de datos panel es la siguiente (Wooldridge, 2020):

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}$$

En donde se destaca la presencia de subíndices i y t , los cuales indican que las variables cambian tanto entre unidades como en el tiempo. Por todo

lo demás, no es diferente a una especificación típica de modelo de regresión lineal. Un componente adicional de especial interés es a_i , es un conjunto de factores inobservables para cada una de las unidades de observación que no cambia en el tiempo. Si habláramos de países, ese componente podría representar, por ejemplo, su ubicación geográfica, ya que esa nunca cambia. Si en la muestra tuviéramos personas, ese componente de factores fijos puede ser el sexo o el color de ojos de la persona, ya que son elementos que se mantienen constantes a lo largo del tiempo.

El modelo de regresión que utilizaremos para ilustrar el caso práctico con datos panel se especifica a continuación:

$$\ln(Gini_{it}) = \beta_0 + \beta_1 Enroll_{it} + \beta_2 Trade_{it} + \beta_3 Inflation_{it} + \beta_4 Unempl_{it} + \beta_5 \ln(GDPpc_{it}) + \beta_6 \ln(FDI_{it}) + a_i + d_t + u_{it}$$

En este, la variable dependiente es el índice de Gini expresado en logaritmos $\ln(Gini_{it})$, mientras que como variables independientes ocuparemos la tasa de matriculación neta (*Enroll*), el grado de apertura comercial (*Trade*), la inflación (*Inflation*) el desempleo (*Unempl*), el PIB per cápita ($\ln GDPpc$), y la inversión extranjera directa ($\ln FDI$).

En este caso debemos prestar atención a los subíndices de las variables, ya que a diferencia de los datos de sección cruzada y de series de tiempo, los subíndices indican que la variable cambia tanto entre las i (corte transversal) como en t (tiempo). Por lo tanto, los elementos antepenúltimo y penúltimo sólo cambian entre las secciones cruzadas y entre el tiempo respectivamente, o, dicho de otra forma, son efectos constantes para cada sección cruzada y son efectos constantes en el tiempo que tienen el mismo efecto sobre las secciones cruzadas. En el primer caso, pueden ser aspectos relacionados con las características y particularidades de cada país, como por ejemplo la posición geográfica, el clima, el tamaño del territorio. Estos elementos son constantes a lo largo del tiempo, pero diferentes entre las secciones cruzadas. En el segundo caso, tenemos como ejemplo las crisis o las pandemias, que afectan a todas las secciones cruzadas por igual y que sólo cambian en el tiempo.

Dado que en este ejercicio utilizaremos diversas variables explicativas, un atajo útil es crear objetos que contengan a las variables que fungirán

como regresores en el modelo de regresión. Esto facilita la introducción de código, reduce errores y ahorra tiempo debido a que no tenemos que escribir los nombres de las variables en cada modelo que estimamos.

En este caso utilizaremos datos que provienen de la página de Banco de Datos del Banco Mundial, la cual puede consultarse en el siguiente enlace: <https://databank.bancomundial.org/home.aspx> siendo de libre acceso. De esta fuente se obtuvo información de 38 países. El estudio completo derivado de estos datos, y de este análisis, están publicados en Hovhannisyan, Castillo-Ponce y Valdez (2019). El formato del archivo de descarga, en este caso es con extensión *.xls para abrirse con Excel o un programa similar.

Caso práctico

Por tal motivo, es necesario importar la base de datos en Stata mediante el comando **import**, tal como se muestra a continuación:

```
import excel "GINI38.xlsx", sheet("Sheet1") firstrow clear
```

Las opciones después de la coma le indican a Stata que debe buscar los datos en el libro con el nombre "Sheet1", mientras que "firstrow" sirve para que la primera fila la lea como el nombre de las variables.

Un aspecto importante en los datos de panel es identificar a las secciones cruzadas de la muestra, que en este caso corresponden a los países. En la base de datos sólo se encuentran los nombres de los países en la variable *Country*, la cual está en formato *string* o texto. Esto implica que Stata sea incapaz de reconocer a las secciones cruzadas dado el formato en que se encuentra la variable *Country*. Una manera sencilla de lidiar con esta situación es codificar cada nombre de país y asignarle un número. Afortunadamente, Stata puede hacerlo de manera automática sin la necesidad de tantas líneas de comando a través de la función **encode**.

```
encode Country, gen(id)
```

Con la instrucción anterior se le ordenó a Stata que asigne a cada nombre de país un número y que genere una variable de nombre *id* con esta información. Mediante el uso del comando **browse** puede abrirse la parrilla de datos para corroborar que se ha creado la variable y que ahora contiene los nombres de los países, pero asociados a un número, comenzando en el 1 y terminando en el 38.

Una vez codificados los nombres de los países, comenzamos a preparar la base de datos para el análisis. En primer lugar, el coeficiente de Gini en la base de datos original está expresado en porcentaje, no obstante, es necesario recordar que este índice no tiene un significado porcentual, sino que está en un rango entre 0 y 1. Por esta razón, es necesario transformar dicha variable dividiendo cada valor entre 100, de tal forma que nos queden valores en el rango que el coeficiente de Gini está definido. La instrucción debe de ser de la siguiente forma:

```
replace gini = (gini/100)
```

Por otro lado, la inversión extranjera directa (*fdi*) puede transformarse a millones de dólares para reducir la cantidad de valores antes del punto decimal. Esto puede realizarse con una instrucción similar a la anterior:

```
replace fdi = (fdi/1000000)
```

Una transformación adicional consistiría en expresar en logaritmos las variables con valores estrictamente positivos, tales como el coeficiente de Gini, el PIB per cápita y la inversión extranjera directa. Una manera sencilla de llevar a cabo esto es mediante la función *for var*.

Después de esto, lo ideal sería obtener algunos estadísticos descriptivos o estadísticos resumen que nos permitan conocer el comportamiento de las variables de manera general. También podemos obtener listados específicos, por ejemplo, si quisiéramos conocer el comportamiento del índice de Gini para México, podemos generar una lista con estos valores mediante el comando *list*.

```
list yr gini if Country == "Mexico"
```

Por otro lado, si nuestro interés se encuentra en saber cuáles son los países con menor coeficiente de Gini y en qué año, lo primero que necesitamos hacer es ordenar la base de datos de manera ascendente en función del índice de Gini con el comando **sort**. Posteriormente, introducimos la instrucción para que muestre en una lista los primeros 10 países y el año respectivo:

```
sort gini
list id yr gini in 1/10
```

1. Sweden	1995.211
2. Bulgaria	1990.213
3. Sweden	1994.221
4. Norway	2012.225
5. Norway	2013.227
6. Norway	2011.229
7. Romania	1990.229
8. Norway	1990.233
9. Belarus	1990.233
10. Norway	2014.235

Si, por el contrario, quisiéramos saber cuáles son los países con un mayor índice de Gini y en qué año, tendríamos que utilizar operadores lógicos condicionantes, por ejemplo:

```
list id yr gini in -25/L if gini != .
```

Lo que le estamos pidiendo a Stata con la última instrucción es que nos muestre una lista con las variables *id*, *yr* y *Gini* con los últimos 25 datos siempre y cuando no incluya valores perdidos. Dado que la variable Gini está ordenada de forma ascendente, los últimos datos de la muestra contendrán necesariamente los de valores más altos, aunque también al final el *software* posiciona los valores perdidos. Estos últimos no nos interesan, por esa razón es necesario incluir en la instrucción que omita estas observaciones.

Dado que los datos de panel tienen un componente temporal, en ocasiones es útil calcular valores promedio para todos los periodos ya que esto facilita el

análisis gráfico. Debemos de tener cuidado a la hora de mostrar gráficos cuando utilizamos datos panel, debido a que se combinan secciones cruzadas con tiempo, esto vuelve indistinguible en un gráfico de dispersión el año al que pertenece cada observación, aunque puedan distinguirse los individuos o viceversa. En este caso, los promedios para todos los años son útiles. Una forma simple de efectuar estos cálculos es mediante la función **egen**.

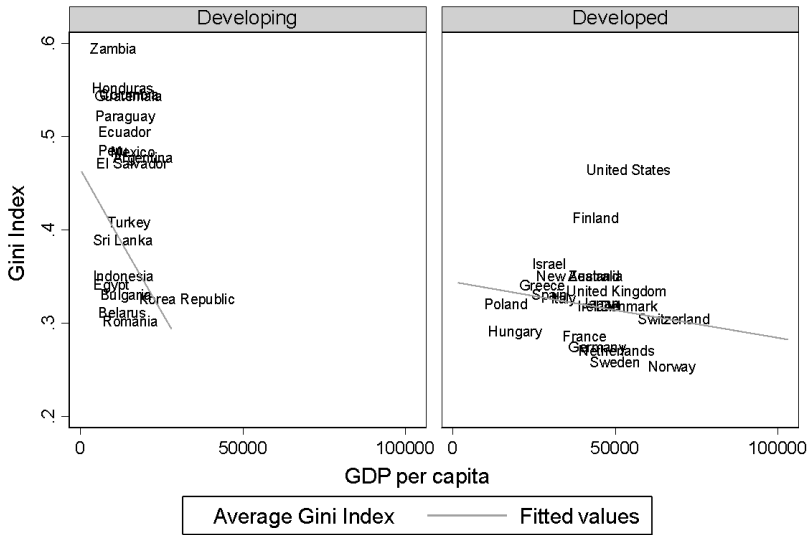
```
egen av_gdppc = mean(gdppc), by(id)
```

La función anterior generará una variable con el nombre *av_gdppc* que corresponderá al PIB per cápita promedio de cada país. Esta es la razón por la que utilizamos la instrucción **by**. Dado que cada *id* corresponde a un país, el promedio se calcula sobre los años en cada uno de los países.

Descripción y análisis de datos

Una vez calculados los promedios para las variables, podemos utilizarlos para generar gráficos de dispersión que nos muestren los promedios por país, por ejemplo, la siguiente instrucción le indica a Stata que genere, en primer lugar, un gráfico de dispersión (*scatter*) y que utilice las etiquetas que se encuentran en la variable *id*, eso es para que en lugar de puntos muestre los nombres de los países. La siguiente instrucción le indica a Stata que estime un ajuste lineal entre las dos variables en niveles, es decir, la relación lineal entre el índice de Gini y el PIB per cápita. Finalmente, se le indica a Stata que separe el gráfico por grupos, los cuales están identificados por la variable binaria *devcountry* la cual separa a países desarrollados de países en desarrollo. El resto de las instrucciones tienen que ver con la parte cosmética del gráfico, como por ejemplo el título del eje de las ordenadas y de las abscisas, así como la paleta de colores a utilizar, que para el caso de este gráfico hemos definido el esquema monocromático.

```
tw (scatter avgini av_gdppc, ml(id) m(i)) ///  
> (lfit gini gdppc), by(devcountry) ytitle(Gini Index) ///  
> xtitle(GDP per capita) scheme(s1mono)
```



Graphs by Developed Countries

En la gráfica anterior observamos que el PIB per cápita guarda una relación diferente con el índice de gini según el nivel de desarrollo del país. Para los países en desarrollo la desigualdad medida mediante el coeficiente de Gini es superior que para los países desarrollados. Sobre todo, son países latinoamericanos los que se encuentran en la parte alta de la distribución del índice.

Además de las gráficas anteriores, los cuadros con estadísticos resumen para cada una de las secciones cruzadas y son útiles en los datos de panel. Podemos obtener un cuadro con los valores promedio del índice de Gini, la inflación y del desempleo con la siguiente instrucción:

```
tabstat gini gdppc infla unem, by(id) s(mean) nototal long /// format(%9.3f)
```

La instrucción previa le indica al *software* que muestre el promedio por sección cruzada, que omita los totales y que utilice el formato largo (*long*) para la tabla. Además, que los valores los cierre en tres dígitos después del punto decimal.

Lo hecho hasta aquí constituye un conjunto de opciones para llevar a cabo la exploración y el manejo de la base de datos, así como también la descripción de la información que contiene dicha base de datos con estructura de panel. Es importante tener en mente que no estamos tratando con una estructura de datos de corte transversal o de serie de tiempo, sino con una combinación de ambas, por lo que la complejidad del ejercicio se incrementa. Existen, por supuesto, otras posibilidades para el análisis descriptivo, sin embargo, hemos decidido mostrar las más utilizadas y funcionales.

Resultados

En la presente sección llevaremos a cabo el análisis estadístico de los datos y la estimación del modelo de regresión.

En total son seis regresores para el modelo lineal, así que crearemos un objeto que contenga a estas seis variables mediante el comando `global`, mismo que también puede utilizarse con su abreviación `gl` con la siguiente sintaxis:

`gl X1 "enroll tradeop infla unem lngdppc lnfdi"`

Es importante advertir el uso de las comillas al principio y al final del conjunto de variables que utilizaremos como regresores. En este ejemplo hemos decidido nombrar como *X1* al conjunto de esas seis variables. De tal forma que puedo utilizar dicho objeto en cualquier momento, no sólo para hacer estimaciones, sino también para obtener estadísticos descriptivos o gráficos de esas seis variables con el simple hecho de llamar al objeto *X1* sin la necesidad de escribir los seis nombres de las variables.

El primer modelo que estimaremos será de datos agrupados por medio de MCO. Para esto necesitamos un conjunto de variables binarias para cada año de la muestra. Dado que tenemos datos desde 1990 hasta 2014 es un total de 24 años, esto implica que debemos generar 24 variables binarias. Hacerlo de forma manual implica la generación de no pocas líneas de código, además de que es posible cometer errores. No obstante, existe una forma sencilla de crear estas variables binarias con una sola línea de código utilizando la función `tab` de la siguiente forma:

tab yr, g(y)

Con el simple hecho de agregar la opción **generate** (**g** de forma abreviada) crea las 24 variables que requerimos para identificar a cada año y así agrupar los datos para cada uno de estos.

Source	SS	df	MS	Number of obs	=	866
				F(30, 835)	=	19.34
Model	21.7770015	30	.725900049	Prob > F	=	0.0000
Residual	31.3434394	835	.037537053	R-squared	=	0.4100
				Adj R-squared	=	0.3888
Total	53.1204408	865	.061410914	Root MSE	=	.19374

Ingini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
enroll	-.0072143	.001503	-4.80	0.000	-.0101644 -.0042642
tradeop	-.0023396	.0002184	-10.71	0.000	-.0027682 -.0019109
infla	-.0000545	.0000538	-1.01	0.311	-.00016 .000051
unem	-.0011391	.0016356	-0.70	0.486	-.0043494 .0020712
Ingdppc	-.0046513	.0075163	-11.26	0.000	-.0094044 -.00698981
lnfdi	-.005906	.0048652	-1.21	0.225	-.0154555 .0036435
yr					
1991	-.0253912	.0508305	-0.50	0.618	-.1251617 .0743793
1992	.0086011	.0512969	0.17	0.867	-.092085 .1092872
1993	.0111259	.0494808	0.22	0.822	-.0859954 .1082472
1994	.0387234	.0486445	0.80	0.426	-.0567564 .1342032
1995	.0783007	.0487911	1.60	0.109	-.017467 .1740684
1996	.0787208	.0494422	1.59	0.112	-.0183248 .1757663
1997	.0811203	.0488059	1.66	0.097	-.0146764 .176917
1998	.1500373	.0505171	2.97	0.003	.0508818 .2491928
1999	.14376	.0496204	2.90	0.004	.0463647 .2411552
2000	.1402752	.051402	2.73	0.006	.0393829 .2411675
2001	.1244156	.0499352	2.49	0.013	.0264024 .2224288
2002	.1289284	.0497111	2.59	0.010	.0313355 .2265019
2003	.1365567	.0499144	2.74	0.006	.0385844 .2345291
2004	.1552667	.0500916	3.10	0.002	.0569466 .2535869
2005	.1677352	.0500012	3.35	0.001	.0695924 .265878
2006	.1785911	.0503059	3.55	0.000	.0798503 .2773319
2007	.1909034	.0499614	3.82	0.000	.0928388 .2889681
2008	.200264	.0501146	4.00	0.000	.1018986 .2986295
2009	.1712786	.0500263	3.42	0.001	.0730866 .2694707
2010	.1937348	.0501537	3.86	0.000	.0952928 .2921769
2011	.2143318	.0505965	4.24	0.000	.1150205 .3136431
2012	.1922852	.0495322	3.88	0.000	.0950629 .2895075
2013	.2181083	.0520409	4.19	0.000	.115962 .3202546
2014	.2076948	.052191	3.98	0.000	.1052537 .3101358
_cons	.5524938	.1312403	4.21	0.000	.2948941 .8100934

Ya podemos estimar la regresión utilizando datos agrupados por el método de MCO. En este caso seguimos utilizando la función **reg**, sin embargo, no escribiremos todas las variables explicativas, sino que las llamaremos con el objeto *XI* anteponiendo el signo de pesos. Adicional a lo anterior,

también incluiremos las variables binarias para agrupar los datos por tiempo, por lo que utilizaremos el operador **i**, donde se excluye el año inicial el cual será representado por el intercepto.

reg lngini \$X1 i.yr

El resultado es una salida similar a todas las que se llevan a cabo por el método de MCO. Lo que hay que destacar en este ejemplo es la incorporación de las variables explicativas mediante el objeto *X1* así como también de las variables binarias con el operador **i**. Prestando atención a los coeficientes, podemos notar dos cosas evidentes: (a) los coeficientes son pequeños y (b) todos tienen signo negativo. El que los coeficientes sean valores pequeños es razonable dado que la variable dependiente es el logaritmo del índice de Gini. Este índice toma valores entre 0 y 1 por lo que una variación del 1 % de este índice implica un cambio en centésimas de dicho coeficiente. Dado que el coeficiente *Enroll* está en niveles, pero representa porcentajes, este se interpreta como un aumento de 1 %, en la matriculación implica que el coeficiente de Gini disminuya 0.72 %. Por otro lado, los coeficientes para las variables *lngdppc* y *lnfdi* se interpretan de forma directa. Para el caso de la primera, la interpretación es que un aumento de 1 % en el PIB per cápita disminuye 0.08 % el índice de Gini. Un caso análogo sucede con la variable de la Inversión Extranjera Directa. Respecto a la inferencia estadística, tenemos tres coeficientes que no son estadísticamente significativos, la inflación, el desempleo y la inversión extranjera directa.

Necesitamos hacer una prueba de heterocedasticidad para corroborar que la inferencia sea válida. En este caso procedemos con la función `estat hettest`.

estat hettest, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

H0: Constant variance

Variab es: enroll tradeop infla unem lngdppc lnfdi 1990b.yr 1991.yr 1992.yr 1993.yr 1994.yr 1995.yr
 1996.yr 1997.yr 1998.yr 1999.yr 2000.yr 2001.yr 2002.yr 2003.yr 2004.yr 2005.yr 2006.yr 2007.yr
 2008.yr 2009.yr 2010.yr 2011.yr 2012.yr 2013.yr 2014.yr

chi2(30) = 64.09

Prob > chi2 = 0.0003

El resultado es que el modelo presenta problemas de heterocedasticidad, lo cual invalida los estadísticos de t así como también el estadístico F . Sin embargo, podemos resolver este problema mediante el cálculo de errores estándares robustos agregando la opción **robust**.

reg lngini \$X1 i.yr, robust

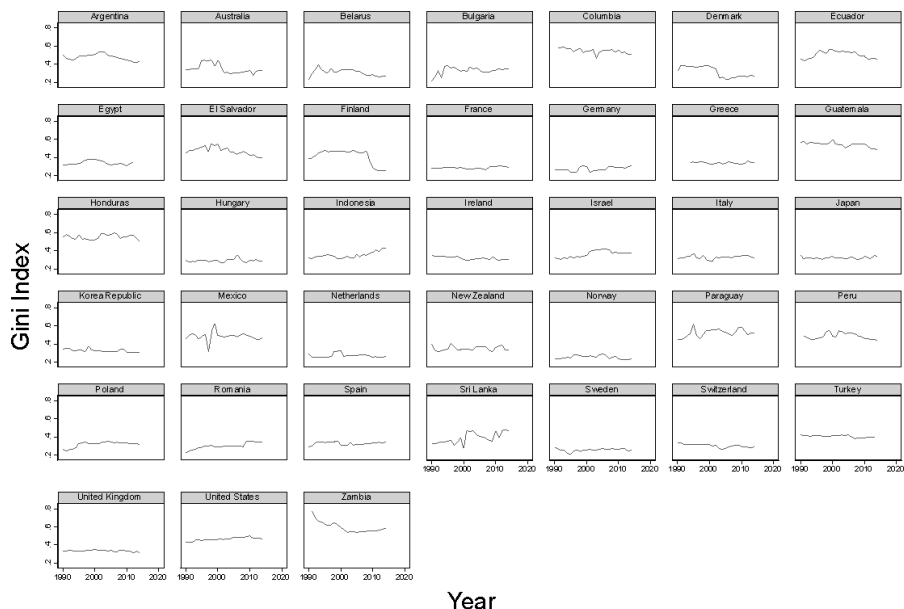
Cuando realizamos trabajos empíricos en donde tenemos observaciones que cambian entre sí, pero que además estas mismas también cambian en el tiempo, no sólo debemos preocuparnos por la heterocedasticidad, sino también por la correlación serial, es decir, del problema en el que los errores se correlacionan en el tiempo. Existen pruebas formales de correlación serial para los datos de series de tiempo, no obstante, en datos panel todavía se encuentran en proceso de desarrollo. Lo que tenemos disponible es la posibilidad de hacer las pruebas de correlación serial *a mano*.

Lo que necesitamos para probar correlación serial, en primer lugar, es declarar a Stata la estructura de los datos, que en este caso son datos de panel, de lo contrario el programa considerará que las observaciones pertenecen a una estructura de secciones cruzadas. Además, el programa necesita *conocer* la variable del tiempo sobre la cual calculará las primeras diferencias de las variables. Esto lo hacemos mediante la función **xtset**. La sintaxis se muestra a continuación, en donde después del comando debe especificarse el identificador para las secciones cruzadas, que en nuestro caso se denomina *id* y después el que permita identificar a la variable tiempo, que en nuestro caso es *yr*.

xtset id yr, yearly
panel variable: id (strongly balanced)
time variable: yr, 1990 to 2014
delta: 1 year

Ahora Stata ya *conoce* que la estructura de datos es un panel. Lo que podemos hacer, una vez declarado lo anterior, es un gráfico de panel de forma sencilla mediante el comando **xtline** cuya sintaxis es como sigue:

xtline gini, scheme(s1 mono)



Graphs by Country

Lo que hacemos a continuación es estimar la ecuación en primeras diferencias, lo cual logramos en Stata mediante el operador **D.** de la siguiente forma:

```
reg D.(lngini $X1 y2-y25), nocons tsscons
```

En el caso anterior debemos de incluir las variables binarias de manera explícita ya que el operador **D.** no permite incluirlas con el operador **i.** Además, debemos de especificar que no se incluya intercepto con la opción **nocons** y que se calcule la suma total de cuadrados como si se tuviera intercepto **tsscons.**

Una vez estimada la regresión, debemos generar los residuales mediante la función **predict** utilizando la opción **residuals (res).**

```
predict resid1, res
```

Con lo anterior creamos una variable llamada *resid1* que corresponde con los residuales del modelo. Para probar la correlación serial regresamos la variable *resid1* sobre ella misma, pero con un rezago, sin constante y con errores estándar robustos, tal como se muestra a continuación:

reg resid1 L.resid1, nocons robust

```

Linear regression              Number of obs   =       726
                              F(1, 725)       =       7.79
                              Prob > F               =     0.0054
                              R-squared              =     0.0694
                              Root MSE           =     .06586
    
```

resid1	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
resid1 L1.	-.2597399	.0930338	-2.79	0.005	-.4423877	-.0770921

Con base en la prueba, podemos afirmar que los errores están correlacionados en el tiempo, ya que para afirmar lo contrario tendríamos que encontrar evidencia de que el residual rezagado tiene un efecto nulo sobre sí mismo.

Una forma de resolver el problema de correlación serial es utilizar la transformación de primeras diferencias, pero haciendo una estimación de errores estándar robustos tanto para la heterocedasticidad, como para la correlación serial. Esto lo logramos mediante la opción de **cluster**.

reg D.(Ingni \$X1 y2-y25), nocons cluster(id)

Existe una transformación propia de los datos panel que puede utilizarse para obtener una estimación más eficiente. Utilizando el supuesto de que existe cierta heterogeneidad, que no observamos, que hace única a cada observación de la muestra, podemos utilizar la transformación de efectos fijos. Para esta estimación ya no usaremos la función **reg**, sino que utilizaremos la función **xtreg** que es propia de la estructura de datos panel.

xtreg lngini \$X1 i.yr, fe

Estimamos la misma ecuación, además, incorporamos variables binarias de tiempo para controlar efectos asociados a este. Finalmente, utilizamos la opción **fe** que es la abreviación de (fixed effects). Podemos notar que la salida de Stata cambia en comparación con la que se obtiene mediante el comando **reg**, además, esta salida provee más información.

Lo que podemos destacar de estos resultados es que nos proporcionan tres diferentes coeficientes de determinación, uno correspondiente para cada transformación, no obstante, debemos de fijarnos sólo en el que corresponde con la transformación *within*. También nos informa sobre la cantidad de grupos que contiene la muestra, los cuales ya conocíamos desde antes.

La estimación de efectos fijos también provee información sobre la correlación que suponemos que existe entre el error y las variables explicativas, que en este caso es de -0.6421 . Respecto a los coeficientes, sólo no tenemos significancia puntual en la apertura comercial y en la Inversión Extranjera Directa.

Es importante que guardemos las estimaciones anteriores debido a que más adelante las necesitaremos. Esto puede hacerse como ya hemos visto en capítulos anteriores, mediante la función `est store`.

est store m1fe

Por otro lado, la estimación con supuestos de efectos fijos no resuelve el problema de correlación serial, sólo nos permite obtener mejores estimadores que por el método de MCO y en algunos casos podría resolverse este problema al suponer una correlación arbitraria del error con las variables independientes. La principal preocupación cuando utilizamos estimaciones en datos panel es el cumplimiento del supuesto de exogeneidad estricta.

En este caso, no se cuenta con una prueba formal para hacer esto, por lo que tenemos que proceder de forma manual. Lo que debemos hacer es utilizar un adelantamiento de los regresores mediante el operador **F**, tal como se muestra a continuación:

Fixed-effects (within) regression
 Group variable: id

Number of obs = 866
 Number of groups = 38

R-sq:

within = 0.1091
 between = 0.2728
 overall = 0.1647

Obs per group:

min = 17
 avg = 22.8
 max = 25

corr(u_i, Xb) = -0.6421

F(30,798) = 3.26
 Prob > F = 0.0000

Ingin	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
enroll	-.0038234	.0008714	-4.39	0.000	-.0055339	-.0021129
tradeop	-.0002128	.0002906	-0.73	0.464	-.0007831	.0003575
infla	.0000649	.0000257	2.53	0.012	.0000145	.0001153
unem	.0064559	.0013048	4.95	0.000	.0038947	.0090171
Ingdppc	.0593319	.0163903	3.62	0.000	.0271588	.0915051
Infdi	.0000725	.0037888	0.02	0.985	-.0073647	.0075097
yr						
1991	-.0123721	.0228194	-0.54	0.588	-.0571653	.032421
1992	-.0068516	.0231011	-0.30	0.767	-.0521977	.0384944
1993	-.0009162	.0224425	-0.04	0.967	-.0449694	.0431371
1994	.0071525	.0221698	0.32	0.747	-.0363656	.0506706
1995	.0297404	.0225401	1.32	0.187	-.0145044	.0739852
1996	.0272573	.0229415	1.19	0.235	-.0177755	.0722902
1997	.0180719	.0227168	0.80	0.427	-.0265199	.0626636
1998	.0676532	.0235824	2.87	0.004	.0213623	.113944
1999	.0686989	.0231949	2.96	0.003	.0231686	.1142292
2000	.0563831	.0242267	2.33	0.020	.0088275	.1039387
2001	.0444861	.023321	1.91	0.057	-.0012916	.0902637
2002	.0488815	.0233121	2.10	0.036	.0031212	.0946418
2003	.030658	.0240022	1.28	0.202	-.016457	.0777729
2004	.0230407	.0249052	0.93	0.355	-.0258467	.0719282
2005	.0185345	.0255036	0.73	0.468	-.0315277	.0685966
2006	.0080315	.0264388	0.30	0.761	-.0438664	.0599293
2007	.0013946	.0273576	0.05	0.959	-.0523068	.0550961
2008	-.0120139	.0284478	-0.42	0.673	-.0678552	.0438274
2009	-.0039613	.0271919	-0.15	0.884	-.0573374	.0494147
2010	-.01246	.0282877	-0.44	0.660	-.0679871	.0430671
2011	-.0281422	.0299396	-0.94	0.348	-.0869118	.0306274
2012	-.0416018	.0294308	-1.41	0.158	-.0993726	.0161691
2013	-.0267523	.0306914	-0.87	0.384	-.0869976	.0334931
2014	-.0466598	.0308655	-1.51	0.131	-.1072468	.0139273
_cons	-1.220001	.148742	-8.20	0.000	-1.511973	-.9280295
sigma_u	.27956067					
sigma_e	.08667872					
rho	.91229799	(fraction of variance due to u_i)				

F test that all u_i=0: F(37, 798) = 91.18

Prob > F = 0.0000

xreg Ingini \$X1 F.(\$X1) i.yr, fe cluster(id)

```

Fixed-effects (within) regression          Number of obs   =       794
Group variable: id                       Number of groups =        38

R-sq:                                     Obs per group:
    within = 0.1233                        min =          12
    between = 0.1982                       avg =         20.9
    overall = 0.1100                       max =          24

corr(u_i, Xb) = -0.5555                    F(35,37)       =       395.74
                                                Prob > F        =       0.0000
    
```

(Std. Err. adjusted for 38 clusters in id)

Ingini	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
enroll	-.001731	.0022622	-0.77	0.449	-.0063147	.0028526
tradeop	-.000206	.0005536	-0.37	0.712	-.0013276	.0009157
infla	.0000356	.0000159	2.24	0.031	3.39e-06	.0000679
unem	.0085358	.0033541	2.54	0.015	.0017398	.0153318
lngdppc	.0309842	.0570181	0.54	0.590	-.0845453	.1465137
lnfdi	.003396	.0044701	0.76	0.452	-.0056613	.0124533
enroll F1.	-.0021965	.0022496	-0.98	0.335	-.0067546	.0023615
tradeop F1.	-.0002135	.0005543	-0.39	0.702	-.0013366	.0009096
infla F1.	.0000826	.0000338	2.45	0.019	.0000142	.0001511
unem F1.	-.0028362	.0028884	-0.98	0.333	-.0086887	.0030162
lngdppc F1.	.0224805	.0563442	0.40	0.692	-.0916836	.1366446
lnfdi F1.	-.0071417	.0064163	-1.11	0.273	-.0201423	.0058588

En este caso debemos de prestar atención a los coeficientes de las variables adelantadas. Si alguno de estos fuera estadísticamente significativo implicaría que la variable a la que se asocia el coeficiente en cuestión no es estrictamente exógena. Esto quiere decir que la variable está correlacionada con el término de error.

En el ejemplo anterior identificamos que la variable *inflación* viola el supuesto de exogeneidad estricta, mientras que las otras variables pueden considerarse como exógenas. Ahora, en lugar de suponer que existe una

correlación arbitraria entre el término de error con las variables explicativas, supondremos que esa correlación se da de forma aleatoria. En este caso decimos que estimamos bajo el supuesto de efectos aleatorios, seguimos utilizando la función **xtreg**, sin embargo, cambiamos la opción de **fe** a **re** tal como se muestra a continuación:

```
xtreg lngini $X1 i.yr, re
```

Estas estimaciones también las guardamos mediante la función **est store**.

```
est store m1re
```

Una vez que hemos llevado a cabo las estimaciones utilizando tanto supuesto de efectos fijos, como de supuestos aleatorios, debemos decidir cuál de estos dos provee estimadores más eficientes. La prueba de Hausman (1978) nos ayuda a tomar dicha decisión ya que compara los coeficientes de efectos fijos con los coeficientes de efectos aleatorios y calcula si estas diferencias son estadísticamente significativas. Si no hay diferencias, entonces es más eficiente utilizar el supuesto de efectos aleatorios, no obstante, si las diferencias son significativas, efectos fijos provee las estimaciones más eficientes.

Para llevar a cabo la prueba de Hausman, utilizamos la función que lleva el mismo apellido del autor: **hausman**, seguido de las estimaciones que guardamos de forma previa.

```
hausman m1fe m1re
```

Con lo anterior obtendremos una prueba como la que sigue:

Test: Ho: difference in coefficients not systematic

$$\text{chi2}(29) = (\mathbf{b}-\mathbf{B})'[(\mathbf{V}_b-\mathbf{V}_B)^{-1}](\mathbf{b}-\mathbf{B})$$

$$= 28.14$$

$$\text{Prob}>\text{chi2} = 0.5105$$

(V_b-V_B is not positive definite)

La hipótesis nula de la prueba de Hausman puede resultar algo confusa¹, sin embargo, es más fácil utilizar el siguiente criterio para decidir entre supuestos de efectos fijos o efectos aleatorios: utilizaremos los supuestos de efectos aleatorios a menos que rechacemos la prueba de Hausman. Con base en los resultados de la prueba no rechazamos la hipótesis nula, lo cual significa que preferimos los efectos aleatorios.

Para finalizar, sabemos que debemos de utilizar efectos aleatorios, por lo tanto, repetimos la estimación con esta especificación y además agregamos la opción para calcular errores estándar robustos a la heterocedasticidad y a la correlación serial, es decir, la opción **cluster**.

xtreg lngini \$X1 i.yr, re cluster(id)

```

Random-effects GLS regression           Number of obs   =   866
Group variable: id                     Number of groups =   38

R-sq:                                  Obs per group:
    within = 0.1012                      min =   17
    between = 0.0077                     avg =  22.8
    overall = 0.0000                      max =   25

corr(u_i, X) = 0 (assumed)              Wald chi2(30)   =  295.35
                                           Prob > chi2     =   0.0000
    
```

(Std. Err. adjusted for 38 clusters in id)

lngini	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
enroll	-.0036784	.0016446	-2.24	0.025	-.0069018	-.0004549
tradeop	-.0004012	.0006498	-0.62	0.537	-.0016747	.0008723
infla	.0000574	.0000247	2.32	0.020	8.97e-06	.0001058
unem	.0053655	.0021608	2.48	0.013	.0011304	.0096007
lngdppc	.0173444	.0310134	0.56	0.576	-.0434407	.0781296
lnfdi	.0013195	.0061215	0.22	0.829	-.0106785	.0133175

Con lo anterior obtenemos un resultado como el previo que se muestra en donde hemos omitido de forma deliberada los coeficientes para las va-

¹ La hipótesis nula de la prueba de Hausman es: La diferencia en los coeficientes no es sistemática

riables binarias de tiempo. Una situación para destacar es que en todas las estimaciones el coeficiente de *enroll* es consistente en signo y también en magnitud. Cabe señalar que este es el coeficiente de interés para el estudio en cuestión. Este se interpreta como que un aumento de 1 % en la matriculación disminuye 0.3 % el coeficiente de Gini. El estadístico es significativo a 5 % porque su valor-p es menor a 0.05, pero mayor a 0.01. Por otro lado, la apertura comercial, el PIB per cápita y la Inversión Extranjera Directa no son significativas para explicar la desigualdad en el ingreso para los países de esta muestra.

Conclusiones

De esta forma concluimos el ejercicio para realizar estimaciones con datos de panel, hemos visto los procedimientos típicos que tienen que llevarse a cabo cuando se tiene esta estructura de datos. Es importante puntualizar que este procedimiento de estimación es válido para cuando tenemos regresores estrictamente exógenos y la cantidad de periodos es menor que la cantidad de observaciones, es decir, cuando $i < t$.

Las estimaciones con datos panel pueden considerarse como las más contemporáneas, en parte gracias a una mayor disponibilidad de datos públicos y a la observación longitudinal de variables de las instituciones tanto nacionales como internacionales. Las técnicas de datos panel junto, con las de econometría espacial, dominan actualmente la literatura de las aplicaciones estadísticas de disciplinas como la economía, finanzas, contabilidad, administración, por mencionar algunas.

8. Modelo de Ecuaciones Estructurales con variables latentes: análisis de la satisfacción en programas sociales

Todos los casos presentados en los demás capítulos de este libro tienen una característica en común: las variables que utilizamos en el análisis son observables. En jerga estadística, cuando decimos observable nos estamos refiriendo a que tenemos un dato que puede ser un número, un nombre, una fecha, una característica. Variables observables existen muchas, el Producto Interno Bruto de un país, el sexo de una persona, la población total de un municipio, el monto de ventas de una empresa, el porcentaje de mujeres en el sector manufacturero, por mencionar sólo unas pocas. El lector puede advertir que una variable observable es tal que podemos representar mediante un número.

No obstante, en las Ciencias Sociales existen aspectos que deseamos analizar, pero que no necesariamente podemos representar mediante un número, tal como la felicidad o la satisfacción. Es decir, si el día de hoy te preguntan que indiques tu nivel de felicidad, posiblemente dudes en la respuesta que darías. Puedes responder simplemente que te sientes feliz, o que te sientes muy feliz o que te desbordas de felicidad. Si bien pueden identificarse escalas de felicidad o grados, no es posible afirmar que estoy 9.512 feliz o que tengo 99 % de felicidad. Al final del día, el nivel que indiques estará basado en tu propia experiencia y en cómo te sientas en el día en particular que te realizan la pregunta. También es necesario reconocer que cada uno de nosotros tiene comprensiones distintas de la felicidad. Para algunos los aspectos intangibles como la salud o la compañía de un ser querido pueden ser motivo de sentirse feliz, mientras que para otros esto

puede no ser así. Si realizáramos un análisis bajo estas circunstancias con las herramientas tradicionales, enfrentaríamos grandes retos respecto a la precisión de nuestras mediciones como a la eficiencia de estas.

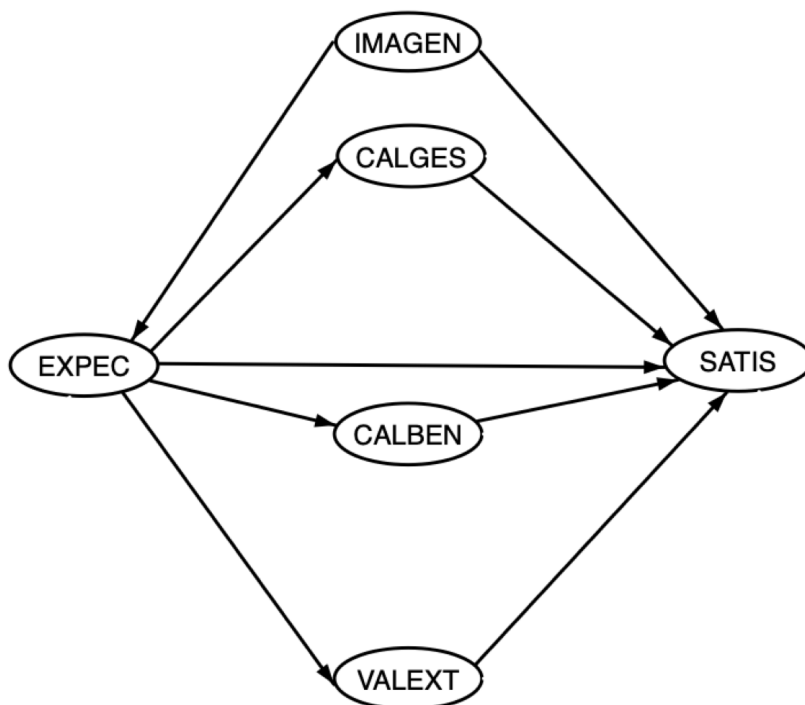
En el área de los negocios, alcanzar la satisfacción del cliente es fundamental para que los productos o servicios que proveen las empresas puedan seguir teniendo demanda en los mercados, lo que significaría que se pretende retener a los clientes. Por su parte, desde la visión del sector público, la idea de la satisfacción de la población radica en que el gasto público que se ejerce a través de los programas sociales para atender necesidades de la población pueda cumplir su objetivo. A diferencia del sector privado, en el sector público no se pretende retener a los ciudadanos, sino cambiar las condiciones de la población y que esta pueda mejorar su nivel de vida.

Con objeto de analizar la satisfacción de los beneficiarios del programa público, en este capítulo retomaremos parte del modelo teórico del Índice Mexicano de Satisfacción de los Beneficiarios de Programas Sociales (IMSAB) (Rodríguez *et al.*, 2014), que ha sido diseñado para este propósito. Dado que no existe una forma específica de medición de la satisfacción, partimos de esta propuesta metodológica considerando como base el uso de variables latentes, e indicadores de percepción, con el propósito de establecer los factores que inciden en la percepción de satisfacción de beneficiarios de programas sociales en México.

La figura 8.1 presenta el modelo a partir de constructos latentes considerados para la determinación de la satisfacción. La primera variable es IMAGEN, la cual es la única variable exógena en el modelo. Ella representa la información que reciben los beneficiarios del programa e impacta de forma directa a la variable expectativa (*EXPC*) y satisfacción (*SATIS*). Con respecto a la variable expectativa, ésta recoge información sobre lo que considera podría representar el programa en su condición social, en este caso se le solicita al beneficiario que conteste ubicándose en la línea del tiempo previo a recibir el apoyo social. La variable de expectativa impacta la calidad de gestión (*CALGES*), calidad del beneficio (*CALBEN*), valoración de las externalidades (*VALEXT*) y la satisfacción (*SATIS*). *CALGES* es una variable que capta la información de los beneficiarios respecto a la operación y administración del proceso que implica el recibir el beneficio del programa, impacta directamente a *SATIS*. Por su parte, *CALBEN* refleja la percepción de los

beneficiarios sobre las características y valoración que representa el apoyo recibido del programa social, e impacta de forma directa a la satisfacción. Respecto a *VALEXT*, esta representa la percepción del beneficiario sobre los costos con la comunidad a través de la exclusión social, la cual se presenta como resultado de ser beneficiario del programa público, esta variable impacta únicamente a la satisfacción. La relación que guardan las variables previas sobre la satisfacción son positivas; en lo referente a la valoración de las externalidades su relación con la satisfacción es negativa. Finalmente, la variable satisfacción representa la percepción sobre la valoración que ha tenido el apoyo social sobre su condición social y si se corresponde con las expectativas que se había construido previo a recibir el apoyo.

Figura 8.1. Modelos de satisfacción de programas sociales en México



Fuente: Elaborado con base en los planteamientos del IMSAB de Rodríguez *et al.* (2014).

Para las variables que presentamos en el diagrama 8.1, no existe de manera puntual alguna medida que permita dimensionar su valor y el impacto que ejerce sobre el nivel de satisfacción de los beneficiarios de los programas sociales. Por lo tanto, es necesaria la elaboración de constructos a través de los cuales se pueda identificar, transversalmente, cómo cada una de las variables impacta de manera individual y conjuntamente sobre la percepción que tienen los beneficiarios de los programas sociales en México. El método de estimación para un modelo con estas características es el de Ecuaciones Estructurales con Variables Latentes (EEVL), el cual ha sido recurrente en los últimos años en el ámbito empresarial para evaluar la percepción del cliente sobre diferentes aspectos de la operación de las empresas.

La fuente de datos que emplearemos para la estimación del modelo es la muestra sobre la evaluación realizada en 2014 al Programa de Desarrollo de Zonas Prioritarias (PDZP), de la Secretaría de Desarrollo Social y que proporcionaba diferentes bienes y servicios que atendían los efectos de pobreza y vulnerabilidad social en México. Este programa se evaluó considerando cinco diferentes tipos de beneficios que proporcionaba el programa: muros, techos, estufas, baños y centros comunitarios de aprendizaje (Rodríguez *et al.*, 2014).

La obtención de la información se realizó mediante un cuestionario que se diseñó para cada uno de los cinco tipos de beneficios que se señalaron previamente, en donde cada sección del cuestionario correspondía a los constructos latentes que se presenta en el gráfico 3.1, cada sección contaba con un conjunto de preguntas de percepción en donde las respuestas están representadas en una escala de Likert entre el rango de 1 a 10. Las respuestas que mejor se ajusten con cada constructo latente serán las variables manifiestas¹, que permitirán estimar los constructos latentes. Se emplearon en total 3,532 observaciones.

En este sentido, para representar y dimensionar a las variables latentes se emplearán las variables manifiestas o proxy. La selección de dichas variables

¹ También se conocen como variables proxy, debido a que su valor representa una medida aproximada de una variable que no puede ser cuantificable, tal como en el caso de las variables latentes.

proxy se realiza mediante análisis factorial confirmatoria², debido a que se parte de un modelo que se ha empleado en la evaluación de otros programas sociales (véase Cogco *et al.*, 2013; Pérez *et al.*, 2017; Rodríguez *et al.*, 2012), por lo que se analiza y estima el modelo de satisfacción de programas públicos partiendo de la experiencia previa. La selección de las variables manifiestas que determinarán a los constructos latentes, se selecciona utilizando el método de máxima verosimilitud (MV) en la primera etapa, y posteriormente se rotan los factores a través del procedimiento de rotación oblicua, esto debido a que por la correlación que existe entre los factores se asume que no son ortogonales, lo cual se probará una vez que se rote la matriz de factores.

En el cuadro 8.1, se desarrolló el análisis factorial considerando las variables proxy o manifiestas que mejor se ajustan a cada uno de los seis factores, donde estos últimos representan a cada una de las variables latentes que se muestran en el cuadro 8.1. Esto se logra a través del comando **factor** juntamente con las variables proxy³ que consideramos (de acuerdo con los antecedentes de modelos estimados) son las que mejor se pueden asociar a cada una de las variables latentes. Debido a que la escala de Likert que se emplea presenta una distribución diferente a la normalidad, resulta más eficiente estimar este modelo mediante Máxima Verosimilitud (MV) empleando el comando **ml** con seis factores, se obtuvieron los primeros resultados sin rotar:

```
factor p17_claridad_info p18_apropiada_info p21_cambian_
condic p24_condic_vida p25_ayud_mejor p31_amable p33_satisf_trato
p38_apoyo_agrado p39_calif_apoyo p40_satisf_benefic p46_cambia_partic
p47_activ_comun p56_calif_apoyo p57_satisf_apoyo p58_mejora_condic, ml factor(6)
```

² Esta técnica permite replicar la estimación de un modelo que previamente ha sido estimado, de tal forma que se parte de factores definidos con las respectivas variables manifiestas.

³ El número de variables manifiestas que se emplean en la estimación de los seis factores podrían ser más, esto dependerá de los grados de libertad y la correcta identificación del modelo.

Cuadro 8.1. *Análisis factorial del modelo de satisfacción*

(obs=2,472)

Iteration 0: log likelihood = -521.58497

Iteration 16: log likelihood = -33.953277

Factor analysis/correlation

Method: maximum likelihood

Rotation: (unrotated)

Number of obs = 2,472

Retained factors = 6

Number of params = 75

Schwarz's BIC = 653.865

(Akaike's) AIC = 217.907

Log likelihood = -33.95328

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	5.18893	4	0.5776	0.5776
Factor2	1.46109	1	0.1626	0.7403
Factor3	0.92058	0	0.1025	0.8427
Factor4	0.81495	0	0.0907	0.9334
Factor5	0.31717	0	0.0353	0.9688
Factor6	0.28071	.	0.0312	1

LR test: independent vs. saturated: $\chi^2(105) = 1.6e+04$ Prob> $\chi^2 = 0.0000$

LR test: 6 factors vs. saturated: $\chi^2(30) = 67.64$ Prob> $\chi^2 = 0.0001$

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Uniqueness
p17_clarid~o	0.5336	-0.0981	0.6987	-0.1601	-0.0484	-0.0427	0.1876
p18_apropi~o	0.5719	-0.1061	0.4716	-0.0462	0.0226	0.0153	0.4364
p21_cambia~c	0.2311	0.0952	0.0594	0.3736	-0.0374	-0.0108	0.7929
p24_condic~a	0.4669	0.1759	0.1642	0.6628	-0.0388	-0.0209	0.2828
p25_ayud_m~r	0.4858	0.092	0.0759	0.4005	0.0446	0.0628	0.5834
p31_amable	0.2462	0.1045	-0.0179	-0.0346	0.1642	0.2531	0.8359
p33_satisf~o	0.5415	0.2062	0.1052	-0.0547	0.2833	0.3676	0.4348
p38_apoyo~o	0.6757	0.3544	-0.0681	-0.0346	0.1813	-0.0386	0.3776
p39_calif_~o	0.7103	0.3035	-0.1363	-0.0853	0.1561	-0.1122	0.3405
p40_satisf~c	0.6785	0.3949	-0.1255	-0.0712	0.1613	-0.2217	0.2876
p46_cambia~c	0.7157	-0.5934	-0.1687	0.0158	0.0027	-0.0069	0.1069
p47_activ_~n	0.5848	-0.5039	-0.1721	-0.0145	0.0057	0.0182	0.3739
p56_calif_~o	0.7046	0.3165	-0.1581	-0.1363	-0.1979	0.0377	0.3193
p57_satisf~o	0.7002	0.3971	-0.17	-0.1023	-0.2045	0.0827	0.2641
p58_mejora~c	0.6734	0.3175	-0.1068	-0.0267	-0.1949	0.0558	0.3926

Sin embargo, resulta más práctico rotar, debido a que facilita el análisis de las cargas para realizar agrupación de las variables en cada factor. Para seleccionar el método de rotación es importante establecer la presencia de correlación entre los factores⁴. En este caso, el método de rotación oblicuo permitirá seleccionar el grupo de proxy o manifiestas que representan a cada una de las seis variables latentes, utilizando el criterio de que las cargas sean las más altas (bajo el supuesto de que se preferirán las cargas que sean mayores a 0.5), los resultados aparecen a continuación:

rotate, oblimin(0) oblique factors(6)

Cuadro 8.2. Rotación de factores de satisfacción por el método oblicuo

Factor analysis/correlation			
Method: maximum likelihood			
Rotation: oblique oblimin (Kaiser off)			
Number of obs = 2,472			
Retained factors = 6			
Number of params = 75			
Schwarz's BIC = 653.865			
(Akaike's) AIC = 217.907			
Log likelihood = -33.95328			
Factor	Variance	Proportion	Rotated factors are correlated
Factor1	4.54654	0.5061	
Factor2	4.522	0.5034	
Factor3	2.94102	0.3274	
Factor4	2.62794	0.2925	
Factor5	2.56194	0.2852	
Factor6	2.4469	0.2724	
LR test: independent vs. saturated: $\chi^2(105) = 1.6e+04$ Prob> $\chi^2 = 0.0000$			
LR test: 6 factors vs. saturated: $\chi^2(30) = 67.64$ Prob> $\chi^2 = 0.0001$			
Rotated factor loadings (pattern matrix) and unique variances			

⁴ Los criterios de rotación ortogonal son: varimax, quartimax, equamax, parsimax, entropía, entre otros. Para rotación oblicuas: promax. Para rotación ortogonal y oblicua: oblimin, oblimax, quartimin, entre otros.

<i>Variable</i>	<i>Factor1</i>	<i>Factor2</i>	<i>Factor3</i>	<i>Factor4</i>	<i>Factor5</i>	<i>Factor6</i>
p17_clarid~o	0.0153	-0.0071	-0.0164	-0.0275	-0.0188	0.9218
p18_apropi~o	-0.024	0.0333	0.0853	0.1132	0.0734	0.621
p21_cambia~c	0.0298	-0.0213	-0.0274	-0.0059	0.471	-0.0299
p24_condic~a	0.0077	0.0067	-0.0244	-0.0193	0.8495	0.0132
p25_ayud_m~r	0.0069	0.0375	0.1288	0.1009	0.5265	-0.0096
p31_amable	0.0443	-0.0351	0.4243	0.0186	-0.0162	-0.069
p33_satisf~o	0.0294	0.051	0.6704	-0.0074	0.0145	0.0588
p38_apoyo_~o	0.0439	0.6228	0.1561	-0.0138	0.0498	0.0079
p39_calif_~o	0.0558	0.7091	0.053	0.0767	-0.0168	-0.0085
p40_satisf~c	-0.0137	0.8946	-0.061	-0.0257	0.0058	0.0094
p46_cambia~c	-0.0049	0.0118	-0.0149	0.9361	0.0166	0.0165
p47_activ_~n	0.0185	-0.0152	0.015	0.7976	-0.0248	-0.0172
p56_calif_~o	0.7962	0.0437	-0.0231	0.0411	-0.0478	0.0213
p57_satisf~o	0.8668	-0.0054	0.0273	-0.0307	-0.0032	-0.0227
p58_mejora~c	0.7466	-0.0174	-0.0049	0.009	0.0801	0.0231
Factor rotation matrix						
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
Factor1	0.8371	0.8269	0.65	0.748	0.5616	0.6167
Factor2	0.4279	0.4471	0.2542	-0.6345	0.1945	-0.1206
Factor3	0.0754	-0.2075	0.5645	0.0055	-0.0067	-0.0392
Factor4	-0.1867	-0.1508	0.1039	-0.1942	0.183	0.7597
Factor5	-0.2488	0.2061	0.4221	0.007	-0.0321	-0.0397
Factor6	-0.1174	-0.0896	-0.0727	0.0052	0.7824	-0.1576

Fuente: Estimaciones propias

Antes de seleccionar cada uno de los grupos de variables que se asocian a los seis constructos latentes, es importante establecer que el procedimiento de rotación seleccionado fue el correcto. Para verificar esto, posterior a la rotación, empleamos el comando **estat common** el cual muestra la matriz de correlación de los factores rotados. Los resultados de esta aplicación se presentan en el cuadro 8.3, donde observamos la existencia de correlación entre los factores, lo cual justifica el haber rotado la matriz a través del método oblicuo.

estat common

Cuadro 8.3. Matriz de correlación de los factores rotados

<i>Factors</i>	<i>Factor1</i>	<i>Factor2</i>	<i>Factor3</i>	<i>Factor4</i>	<i>Factor5</i>	<i>Factor6</i>
Factor1	1					
Factor2	0.8552	1				
Factor3	0.5796	0.6118	1			
Factor4	0.3889	0.364	0.3104	1		
Factor5	0.4348	0.4484	0.3593	0.265	1	
Factor6	0.3482	0.3555	0.4217	0.389	0.3401	1

Fuente: Estimaciones propias.

La selección de las variables manifiestas para cada uno de los factores (variables latentes) la realizamos a través de las columnas y las filas del cuadro 8.4, en el caso de las columnas hace referencia a los factores y las filas, a las variables proxy, los valores de la intersección del factor y variable proxy se conoce como carga factorial, y este valor permite establecer con qué variable latente se relaciona más cada una de las variables proxy. Para el primer factor, las preguntas 56, 57 y 58 cumplen el criterio de que las cargas factoriales son mayores a 0.5, lo que justifica que estas variables determinarán el comportamiento de la variable latente *SATIS*.⁵ Para el factor 2, las preguntas 38, 39 y 40 son las variables que mejor se ajustan en la explicación del factor y hace referencia a la parte de la calidad del beneficio (*CALBEN*). Para el factor 3, la pregunta 33 supera el criterio de que las cargas factoriales deben ser superiores a 0.5. Sin embargo, no es factible emplear solamente una variable manifiesta en la determinación del constructo latentes, en este sentido, holgamos el criterio de las cargas y tomamos la que más se aproxime a la carga de 0.5, en este caso es la pregunta 31, por lo que el constructo latente que se constituye es calidad de la gestión (*CALGES*).

⁵ Revisando el cuestionario que se presenta en el informe: https://www.gob.mx/cms/uploads/attachment/file/28288/IF_ESB_PDZP_2014.pdf, es posible identificar que esas preguntas corresponden a la sección de Satisfacción de los beneficiarios, por lo tanto, el factor que asocia a esas preguntas corresponde al constructo latente de Satisfacción. Esto ocurre para cada conjunto de preguntas que se asocia a un determinado factor, se tendrá que determinar a qué sección del cuestionario corresponde ese conjunto de preguntas para así determinar qué variable latente representa.

Continuando con los resultados del cuadro 8.4, el cuarto factor está constituido con las preguntas 46 y 47, las cuales corresponden al apartado de valoración de las externalidades (*VALEXT*), por lo que estas explicarán dicho constructo latente. El quinto factor se constituye por las preguntas 21, 24 y 25, en donde se analizan las expectativas (*EXPEC*) de los beneficiarios del programa antes de que recibieran el programa. Finalmente, el sexto factor se asocia con la variable latente de imagen con las preguntas 17 y 18, las cuales explicarán dichos constructos latentes.

En esta primera etapa del análisis factorial confirmatorio, seleccionamos las variables manifiestas o proxy que se utilizarán para la estimación del modelo que se presenta en el cuadro 8.1. En el cuadro 8.4 se presentan las variables manifiestas agrupadas para cada uno de los constructos latentes, con el número y la pregunta correspondiente al cuestionario, así como su respectiva carga factorial correspondiente a la matriz rotada por el método oblicuo. Esta forma de presentar los resultados facilita la interpretación de los primeros resultados obtenidos.

Posterior a la selección de las variables manifiestas asociadas a cada constructo latente, establecemos los estadísticos correspondientes a cada una de las variables manifiestas, las cuales se obtendrán en el programa Stata a través del uso del comando **sum** y la opción de **detail**. Utilizamos este comando con esta opción debido a que para el análisis de los modelos es necesario conocer la media de la variable, la desviación estándar y los coeficientes de simetría y curtosis que se utilizan para justificar la distribución de cada una de las variables, así como la variabilidad que existe entre las respuestas que se obtuvieron de aplicar el cuestionario a beneficiario del programa PDZP. A continuación, se presenta la sintaxis para obtener estas medidas estadísticas, y los resultados se muestran de forma resumida en el cuadro 8.5, es decir, sólo se presentan el resultado de la pregunta 17 y la pregunta 58, con el propósito de ilustrar lo obtenido, los resultados completos se presentan en el cuadro 8.7.

```
sum p17_claridad_info p18_apropiada_info p21_cambian_condic p24_condic_vida p25_ayud_mejor p31_amable p33_satisftrato p38_apoyo_agrado p39_calif_apoyo p40_satisf_benefic
```


Cuadro 8.4. Determinación de las variables manifiestas y de los constructos latentes a partir del análisis factorial

Variable Latente	Símbolo	Indicador Proxy	Factor					
			1	2	3	4	5	6
Imagen del programa	P17	¿Cómo calificaría la claridad de la información que recibió?	0.0153	-0.0071	-0.0164	-0.0275	-0.0188	0.9218
	P18	¿Qué tan apropiada considera la información que recibió?	-0.024	0.0333	0.0853	0.1132	0.0734	0.621
Expectativas del programa	P21	¿En qué medida podría cambiar sus condiciones de vida?	0.0298	-0.0213	-0.0274	-0.0059	0.471	-0.0299
	P24	¿Ayudaría a mejorar sus condiciones de vida?	0.0077	0.0067	-0.0244	-0.0193	0.8495	0.0132
	P25	¿En qué medida considera que ayuda a su familia?	0.0069	0.0375	0.1288	0.1009	0.5265	-0.0096
Calidad de gestión	P31	¿Qué tan amable fue con usted el personal que le entregó el apoyo?	0.0443	-0.0351	0.4243	0.0186	-0.0162	-0.069
	P33	¿Qué tan satisfecho está con la forma en cómo lo tratan?	0.0294	0.051	0.6704	-0.0074	0.0145	0.0588
Calidad del beneficio	P38	¿En qué medida las características del apoyo son de su agrado?	0.0439	0.6228	0.1561	-0.0138	0.0498	0.0079
	P39	¿Cómo califica el apoyo del programa?	0.0558	0.7091	0.053	0.0767	-0.0168	-0.0085
	P40	¿Qué tan satisfecho está con los beneficios del programa?	-0.0137	0.8946	-0.061	-0.0257	0.0058	0.0094
Valoración de las externalidades	P46	¿Cómo ha cambiado su nivel de participación en las actividades de la comunidad a partir de que recibió el beneficio?	-0.0049	0.0118	-0.0149	0.9361	0.0166	0.0165
	P47	¿Cómo ha cambiado su nivel de participación en las actividades de la comunidad a partir de que recibió el beneficio?	0.0185	-0.0152	0.015	0.7976	-0.0248	-0.0172
Satisfacción	P56	¿Qué calificación le da al programa?	0.7962	0.0437	-0.0231	0.0411	-0.0478	0.0213
	P57	¿Qué tan satisfecho está con el programa?	0.8668	-0.0054	0.0273	-0.0307	-0.0032	-0.0227
	P58	¿En qué medida le ha ayudado el programa a mejorar sus condiciones de vida?	0.7466	-0.0174	-0.0049	0.009	0.0801	0.0231

Fuente: Estimaciones propias.

p46_cambia_partic p47_activ_comun p56_calif_apoyo p57_satisf_apoyo p58_mejora_condic, detail

Cuadro 8.5. *Medidas descriptivas, dispersión y de distribución de probabilidad*

<i>Claridad de la información que recibió</i>				
	<i>Percentiles</i>	<i>Smallest</i>		
1%	1	1		
5%	2	1		
10%	4	1	Obs	2,765
25%	7	1	Sum of Wgt.	2,765
50%	9		Mean	7.895841
		Largest	Std. Dev.	2.438387
75%	10	10		
90%	10	10	Variance	5.945732
95%	10	10	Skewness	-1.280952
99%	10	10	Kurtosis	3.794835
<i>P17 ¿Qué tan apropiada considera la información que recibió?</i>				
	<i>Percentiles</i>	<i>Smallest</i>		
.
.
.
<i>¿En qué medida le ha ayudado el programa a mejorar sus condiciones de vida?</i>				
	<i>Percentiles</i>	<i>Smallest</i>		
1%	1	1		
5%	3	1		
10%	6	1	Obs	2,776
25%	8	1	Sum of Wgt.	2,776
50%	9		Mean	8.323847
		Largest	Std. Dev.	2.108059
75%	10	10		
90%	10	10	Variance	4.443914
95%	10	10	Skewness	-1.845988
99%	10	10	Kurtosis	6.311299

Fuente: Estimaciones propias.

Cuando únicamente se emplea el comando **sum**, solamente se pueden obtener medidas de desviación estándar y media; con la opción **detail**, se amplía la información estadística, proporcionando un panorama de la distribución percentil y de la forma de la distribución de probabilidad de cada una de las variables que se estimen. En el caso de la pregunta 17, en donde analizamos la claridad de la información, que recibió en una escala del 1 al 10, la respuesta media fue 7.9 con una desviación estándar de 2.4 y una mediana de 9 según la distribución de 50 % de los datos. Para que una variable presente una distribución normal, la regla es que su simetría (*skewness*) sea cero y su curtosis (*Kurtosis*) sea de 3, sin embargo, para esta variable sus valores son de -1.3 y 3.8, respectivamente, lo que significa que no presenta un comportamiento normal, así que estimar el modelo a través de Mínimos Cuadrados Ordinarios (MCO) no resulta eficiente. Por lo que, al analizar el resto de las variables y coincidir en este resultado, se probará un método de estimación no paramétrico.

Por otro lado, en los MEE resulta fundamental evaluar la pertinencia de la escala de las variables manifiestas a través del estadístico de prueba alfa de Cronbach. El criterio que se utiliza, y que ha sido comúnmente aceptado en los trabajos sobre análisis factorial, es obtener un valor del estadístico por lo menos de 0.70. El estadístico se emplea en dos momentos, en principio para determinar que el uso de las 17 variables proxy emplean una escala que hace factible utilizarse para estimar el modelo de satisfacción, todas ellas agrupadas en seis factores. En el cuadro 8.6 se presenta el estadístico de alfa de Cronbach de las 17 variables manifiestas que se emplearán en la estimación del modelo de satisfacción. El comando empleado es **alpha** con la opción **asis** que proporciona el signo de la manifiesta, la opción **item**⁶ se emplea para identificar el efecto que se genera por excluir ese item, y la opción **label** para que proporcione la etiqueta de cada una de las variables.

```
alpha p17_claridad_info p18_apropiada_info p21_cambian_
condic p24_condic_vida p25_ayud_mejor p31_amable p33_sa
tisf_trato p38_apoyo_agrado p39_calif_apoyo p40_satisf_be
```

⁶ Las variables proxy o manifiestas también pueden recibir este nombre.

Cuadro 8.6. Estadístico alfa de Cronbach para el total de las variables empleadas para estimar el modelo de satisfacción

Test scale = mean(unstandardized items)							
Item	Obs	sign	item-test corr.	item-rest corr.	interitem. Cov	alpha	Label
p17_clarid~o	2765	+	0.547	0.4438	1.347448	0.8574	Claridad de la información que recibió
p18_apropi~o	2736	+	0.6099	0.5243	1.330324	0.8526	¿Qué tan apropiada considera la información que recibió?
p21_cambia~c	2749	+	0.317	0.2041	1.462413	0.8686	¿En qué medida podrían cambiar sus condiciones de vida?
p24_condic~a	2766	+	0.4999	0.4229	1.406047	0.8569	¿Ayudará a mejorar sus condiciones de vida?
p25_ayud_m~r	2741	+	0.5223	0.447	1.398043	0.8559	¿En qué medida el ser beneficiario del programa ayudaría a su familia?
p31_amable	2753	+	0.3595	0.2398	1.445464	0.8686	¿Qué tan amable fue el personal que entregó el apoyo?
p33_satisf~o	2771	+	0.6436	0.5669	1.324697	0.8508	¿Qué tan satisfecho está con la forma en cómo lo trataron?
p38_apoyo_~o	2742	+	0.7209	0.6662	1.325175	0.8474	¿En qué medida las características del apoyo son de su agrado?
p39_calif_~o	2725	+	0.7165	0.6665	1.345722	0.8486	¿Cómo califica el apoyo del programa?
p40_satisf~c	2776	+	0.7299	0.6716	1.308154	0.8466	¿Qué tan satisfecho está con los beneficios del programa?
p46_cambia~c	2675	+	0.6374	0.5429	1.309719	0.8532	¿Cómo ha cambiado su nivel de participación en las actividades de la comunidad?
p47_activ_~n	2655	+	0.5546	0.4323	1.335186	0.8608	¿Las actividades que realiza en la comunidad son con beneficios del mismo programa?
p56_calif_~o	2771	+	0.7472	0.6934	1.30508	0.8457	¿Qué calificación le da al beneficio?
p57_satisf~o	2774	+	0.7365	0.6781	1.302467	0.846	¿Qué tan satisfecho está con el beneficio?
p58_mejora~c	2776	+	0.7294	0.6667	1.296082	0.8463	¿En qué medida le ha ayudado el programa a mejorar su condiciones de vida?
Test scale					1.349466	0.8623	mean(unstandardized items)

Fuente: Estimaciones propias.

**nefic p46_cambia_partic p47_activ_comun p56_calif_apoyo
p57_satisf_apoyo p58_mejora_condic, asis item label**

El total de observaciones que se muestra en el cuadro 8.6 varía de acuerdo con los datos faltantes (missings) en cada una de las variables, esto sucede debido a que algunas de las personas encuestadas no contestaron la pregunta. La columna de *sign* refiere que todas los items que se emplean se asocian de forma directa. La columna de *item-test correlations* y *item-rest correlations* contribuye en establecer la escala del item con el resto de las escalas, en este caso la pregunta 21 y la 31 no se ajustan de forma adecuada con el resto de las escalas de los items, observando los resultados de ambos estadísticos, se aprecia que son los más pequeños de todos los items. El alfa de Cronbach general es de 0.8623, un valor que justifica la confiabilidad de las escalas de todos los items, esto es así ya que en el caso de la pregunta 21 y 31 que tenían la menor relación con el resto de los items, su exclusión no contribuye en un notable ajuste del modelo, eliminando esos items, el estadístico de alfa de Cronbach global se incrementa a 0.8686, por lo que al final se ha decidido dejarlos en el modelo.

Posteriormente, evaluamos la consistencia de la escala de los items calculando el estadístico del alfa de Cronbach para cada grupo de variables proxy que determinan al constructo latente; a continuación se señala la sintaxis para el cálculo del estadístico para cada grupo de variables, los resultados se omiten y se presentan de forma resumida en el cuadro 8.7, donde además se integran con los resultados descriptivos, dispersión, distribución de probabilidad y con las cargas factoriales, es una forma mucho más sencilla de realizar el análisis de las variables manifiestas que permiten estimar el modelo de satisfacción.

**alpha p17_claridad_info p18_apropiada_info , asis item label
alpha p21_cambian_condic p24_condic_vida p25_ayud_mejor ,
asis item label
alpha p31_amable p33_satisf_trato , asis item label
alpha p38_apoyo_agrado p39_calif_apoyo p40_satisf_benefic,
asis item label
alpha p46_cambia_partic p47_activ_comun, asis item label**

**alpha p56_calif_apoyo p57_satisf_apoyo p58_mejora_condic,
asis item label**

(se omiten todos estos resultados)

Del cuadro 8.7, en lo que respecta al análisis del estadístico del alfa de Cronbach para cada grupo de variables manifiestas que conforman los constructos latentes, se observa que la imagen, las expectativas, la calidad del beneficio, la valoración de las externalidades y la satisfacción tienen un valor estadístico superior al 0.70, lo que refleja la factibilidad de usarlos como grupo de variables que se asocian de manera consistente para la determinación de cada uno de estos constructos latentes. En el caso de la variable latente de calidad de la gestión, la situación es distinta, el valor del estadístico del alfa de Cronbach se ubica por debajo del mínimo aceptable, siendo la variable manifiesta (item o pregunta) que mide “que tan amable fue el personal que le brindó el apoyo”, la que origina la falta de consistencia en este constructo latente, de esta forma limita que el grupo de preguntas pueda ser consistente en la estimación del constructo latente de la calidad de la gestión. Sin embargo, este último resultado no afecta de manera global el alfa de Cronbach, el cual continúa siendo alto, por lo que se sugiere que el constructo latente de calidad de la gestión con sus respectivas variables proxy se mantengan en el modelo.

Revisando la media de cada variable manifiesta, se observa que en la variable latente, imagen del programa, las variables tienen una media muy próxima a 8, lo que significa que en promedio los beneficiarios perciben de una buena manera la imagen del programa. También en la variable latente de valoración de las externalidades observamos los valores más bajos, esto debido a que la población reconoce que el ser beneficiario del programa social ha impactado su participación con su comunidad de forma negativa. El revisar la simetría y curtosis de todas las variables manifiestas, en este mismo cuadro, observamos que en ninguno de los casos la simetría es cero y la curtosis es 3, lo que permite establecer que el método lineal tradicional no es factible de emplearse. Por otro lado, las cargas factoriales de los indicadores sobre la variable latente en casi todos los casos superan el criterio de 0.5. Ante tal escenario, se puede establecer que tanto la escala como los indicadores son pertinentes para desarrollar estimación del modelo de sa-

Cuadro 8.7. Estadístico descriptivo de los indicadores de las variables que componen el modelo de satisfacción del PDZP

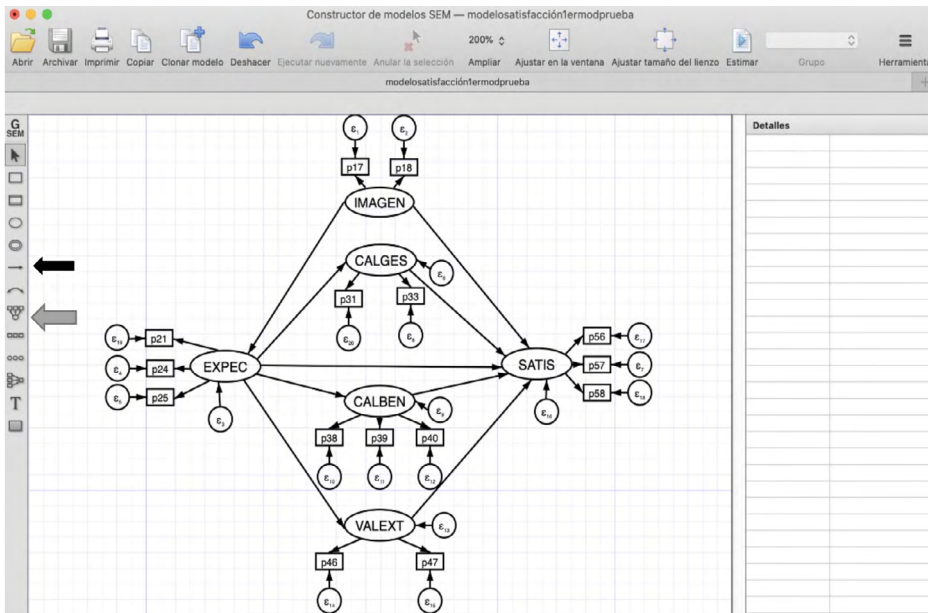
Variable latente	Símbolo de la proxy	Indicador proxy	Media	Desviación estándar	Simetría	Curtosis	Cargas factoriales	
							(IMAGE)	(las mayúsculas-STAND)
Imagen del Programa	P17	¿Cómo calificaría la claridad de la información que recibió?	7.93	2.40	-1.31	3.92	0.74	0.78
	P18	¿Qué tan apropiada considera la información que recibió?	8.02	2.26	-1.37	4.34	0.84	
Expectativas del programa	P21	¿En qué medida podría cambiar sus condiciones de vida?	8.35	2.23	-1.86	6.08	0.36	0.62
	P24	¿Ayudaría a mejorar sus condiciones de vida?	8.57	1.76	-1.96	7.73	0.56	
	P25	¿En qué medida considera que ayuda a su familia?	8.52	1.74	-1.77	6.95	0.61	
Calidad de Gestión	P31	¿Qué tan amable fue con usted el personal que le entregó el apoyo?	8.57	2.36	-2.11	6.69	0.30	0.45
	P33	¿Qué tan satisfecho está con la forma en cómo lo trataron?	8.52	2.15	-1.90	6.18	0.72	
Calidad del beneficio	P38	¿En qué medida las características del apoyo son de su agrado?	8.50	1.86	-1.76	6.30	0.81	0.83
	P39	¿Cómo califica el apoyo del programa?	8.64	1.72	-2.08	8.44	0.82	
	P40	¿Qué tan satisfecho está con los beneficios del programa?	8.50	2.00	-1.96	6.87	0.84	
	P46	¿Cómo ha cambiado su nivel de participación en las actividades de la comunidad a partir de que recibió el beneficio?	7.15	2.52	-0.89	3.05	0.99	
Valoración de las externalidades	P47	¿Cómo ha cambiado su nivel de participación en las actividades de la comunidad a partir de que recibió el beneficio?	6.72	2.79	-0.74	2.50	0.74	0.84
	P56	¿Qué calificación le da al programa?	8.44	1.96	-1.91	6.83	0.85	
Satisfacción	P57	¿Qué tan satisfecho está con el programa?	8.50	2.00	-2.06	7.38	0.88	0.89
	P58	¿En qué medida le ha ayudado el programa a mejorar sus condiciones de vida?	8.33	2.11	-1.84	6.29	0.79	

Fuente: Estimaciones propias.

tatisfacción del programa social PDZP, empleando el método de estimación de Máxima Verosimilitud (MV).

Una vez definidas las variables latentes y sus respectivas variables manifiestas, especificamos el modelo en su forma estructural como se muestra en la figura 8.2, sólo que puede ser diseñado haciendo uso del editor del modelo de ecuaciones estructurales que se encuentra en la pestaña de estadísticas del programa Stata (a partir de la versión 12). La construcción gráfica del modelo es sencilla, principalmente haciendo uso en la barra de herramientas que aparece en la figura 8.2 donde empleamos el icono que representa las variables latentes con sus variables manifiestas y el icono de la flecha que permite vincular cada una de las variables latentes. Asignamos cada una de las variables manifiestas que emplearemos para estimar los constructos latentes, siendo esta una alternativa para hacer la estimación del modelo. Sin embargo, la estimación del modelo también se puede realizar en forma de ecuación a través de la barra de comando.

Figura 8.2. Modelo de satisfacción de los beneficiarios del PDZP



Fuente: Estimaciones propias

Previa a la estimación, establecemos *a priori* la relación de variable que se espera entre los constructos latentes. En el caso de la variable *imagen* (variable exógena) se relaciona directamente con *expectativas* y *satisfacción*, esperando que exista una relación positiva con ambas. En lo que respecta a la variable *expectativas* se relaciona con las variables *calidad de la gestión*, *calidad del beneficio*, *valoración de las externalidades* y la *satisfacción*, las cuales se espera se relacionen positivamente con esta variable de *expectativas*. Por su parte, la *calidad de la gestión*, *calidad del beneficio* y *valoración de las externalidades* se relacionan de forma directa con la *satisfacción*.

Definido el modelo de satisfacción para los beneficiarios del PDZB, estimamos a través del Sistema de Ecuaciones Estructurales (SEE), mediante el método de estimación de MV. Utilizamos este método de estimación debido a que permite suponer normalidad multivariable de los reactivos utilizados como indicadores; por lo tanto, los parámetros estimados son consistentes, eficientes y asintóticamente insesgados (Levy y Varella, 2003; Manzano y Zamora, 2009). Además, el método de MV posee la ventaja de que las estimaciones obtenidas no dependen de la escala de medición de las variables empleadas en el análisis.

La estimación se puede realizar directamente en el constructor del modelo SEM del programa Stata, en el icono de estimar, considerando que previamente se elaboró el modelo y asignaron las variables manifiestas a cada constructo latente. La otra opción para estimar el modelo es a través de la barra de comando, en donde se emplea el comando **sem** y entre paréntesis se colocan cada una de las variables latentes con la respectiva variable manifiesta separadas a través de un guion medio (–) y el signo de mayor que (>), de la misma forma se plantea la relación causal entre las variables latentes. En la sintaxis, en la parte de las opciones (después de la coma), se establece que la matriz de varianza y covarianza sea robusta, lo que significa que los errores estándar de los coeficientes sean robustos. También, en la parte de opciones se establece quiénes son las variables latentes (entre paréntesis) y se utiliza la opción de **nocapslatent** para indicarle que no asocie las variables con letras mayúsculas como la variable latente. En la regresión se establece que los resultados de los coeficientes que proporciona el programa sean los estandarizados. A continuación se presenta la sintaxis, y en el cuadro 8.8 los resultados de la estimación del modelo de satisfacción.

```

sem (IMAGEN -> p17_claridad_info) (IMAGEN -> p18_apropiada_
info) (IMAGEN -> EXPEC) (IMAGEN -> SATIS) (EXPEC -> p24_con-
dic_vida) (EXPEC -> p25_ayud_mejor) (EXPEC -> CALGES) (EX-
PEC -> CALBEN) (EXPEC -> VALEXT) (EXPEC -> SATIS) (EXPEC ->
p21_cambian_condic) (CALGES -> p33_satisf_trato) (CALGES ->
SATIS) (CALGES -> p31_amable) (CALBEN -> p38_apoyo_agrado
) (CALBEN -> p39_calif_apoyo) (CALBEN -> p40_satisf_benefic)
(CALBEN -> SATIS) (VALEXT -> p46_cambia_partic) (VALEXT ->
p47_activ_comun) (VALEXT -> SATIS) (SATIS -> p57_satisf_apo-
yo) (SATIS -> p56_calif_apoyo) (SATIS -> p58_mejora_condic),
vce(robust) standardized latent(IMAGEN EXPEC CALGES CALBEN
VALEXT SATIS) nocapslatent

```

En principio, observamos en el cuadro 8.8 que de las 3,532 observaciones que componen la base de datos, aproximadamente 30 % de éstas (1,060) contenían preguntas sin respuesta, por lo que el modelo se corrió con 2,472 observaciones. En la primera parte de los resultados también se presentan las interacciones necesarias para encontrar los coeficientes que mejor se ajustan a la relación; cuando se solicita la matriz de varianzas y covarianzas robustas, el estadístico cambia al logaritmo de la pseudoverosimilitud (log pseudolikelihood) el cual tiene un valor de -70698.018 . Posteriormente, se presentan las variables manifiestas que se normalizaron con respecto a cada uno de los constructos latentes.

En la primera parte de los resultados estructurales del cuadro 8.8, observamos con relación a la percepción de satisfacción que experimentan los beneficiarios del PDZP, que la variable *imagen* y la de *expectativa* presentan valores de Z a un nivel de 95 % de confianza menores a 2 en valor absoluto y el valor de la probabilidad del estadístico es mayor a 0.05, por lo que ninguna de estas dos variables es significativa en su determinación de la satisfacción. Observamos que el efecto de estas variables es indirecto, debido a que los beneficiarios visualizan al programa desde antes de recibir el beneficio y, por lo tanto, se construyen expectativas sobre dicho apoyo, impactando de manera directa y positiva la *calidad de gestión* (CALGES), la *calidad del beneficio* (CALBEN) y la *valoración de las externalidades* (VALEXT). En este sentido, sus expectativas sobre el programa no se construyen solamente con-

siderando el apoyo recibido, el beneficiario considera los otros aspectos para establecer su grado de satisfacción. La variable de *imagen* (IMAGEN) es significativa, por lo que sí impacta a la *variable expectativa* (EXPEC).

Cuadro 8.8. Resultado de la estimación del modelo de satisfacción

(1060 observations with missing values excluded)

Endogenous variables
 Measurement: p17_claridad_info p18_apropiada_info p24_condic_vida p25_ayud_mejor p21_cambian_condic p33_satisf_trato p31_amable p38_apoyo_agrado p39_calif_apoyo p40_satisf_benefic p46_cambia_partic p47_activ_comun p57_satisf_apoyo p56_calif_apoyo p58_mejora_condic
 Latent: EXPEC SATIS CALGES CALBEN VALEXT

Exogenous variables
 Latent: IMAGEN

Fitting target model:
 Iteration 0: log pseudolikelihood = -71702.113 (not concave)
 ...
 Iteration 6: log pseudolikelihood = -70698.018
 Number of obs = 2,472
 Structural equation model
 Estimation method = ml
 Log pseudolikelihood = -70698.018
 (1) [p24_condic_vida]EXPEC = 1
 (2) [p33_satisf_trato]CALGES = 1
 (3) [p38_apoyo_agrado]CALBEN = 1
 (4) [p46_cambia_partic]VALEXT = 1
 (5) [p57_satisf_apoyo]SATIS = 1
 (6) [p17_claridad_info]IMAGEN = 1

	Standardized	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
Structural						
EXPEC						
	IMAGEN	0.6148112	0.0242821	25.32	0	0.5672191 0.6624033
SATIS						
	EXPEC	-0.0087428	0.0823962	-0.11	0.915	-0.1702364 0.1527509
	CALGES	0.0964108	0.0438417	2.2	0.028	0.0104825 0.182339
	CALBEN	0.8143543	0.0421373	19.33	0	0.7317667 0.8969419
	VALEXT	0.0651998	0.019815	3.29	0.001	0.0263632 0.1040365
	IMAGEN	-0.0186421	0.0236737	-0.79	0.431	-0.0650417 0.0277576
CALGES						
	EXPEC	0.6951226	0.0368114	18.88	0	0.6229735 0.7672717
CALBEN						
	EXPEC	0.7312146	0.0243556	30.02	0	0.6834784 0.7789507
VALEXT						
	EXPEC	0.4921099	0.0230266	21.37	0	0.4469787 0.5372411

Measurement							
p17_claridad_info							
IMAGEN	0.7396001	0.0229983	32.16	0	0.6945243	0.784676	
_cons	3.277937	0.0729568	44.93	0	3.134944	3.42093	
p18_apropiada_info							
IMAGEN	0.8797413	0.0212099	41.48	0	0.8381706	0.9213119	
_cons	3.615658	0.0844131	42.83	0	3.450211	3.781105	
p24_condic_vida							
EXPEC	0.6049341	0.0299035	20.23	0	0.5463242	0.6635439	
_cons	4.946713	0.1482041	33.38	0	4.656238	5.237188	
p25_ayud_mejor							
EXPEC	0.6247591	0.028078	22.25	0	0.5697272	0.6797911	
_cons	5.04955	0.1452958	34.75	0	4.764776	5.334325	
p21_cambian_condic							
EXPEC	0.3451944	0.0324207	10.65	0	0.281651	0.4087378	
_cons	3.709052	0.100906	36.76	0	3.51128	3.906824	
p33_satisf_trato							
CALGES	0.8133762	0.0366258	22.21	0	0.741591	0.8851615	
_cons	4.292982	0.1218761	35.22	0	4.054109	4.531855	
p31_amable							
CALGES	0.3615188	0.0289815	12.47	0	0.3047161	0.4183214	
_cons	3.594315	0.1036324	34.68	0	3.391199	3.797431	
p38_apoyo_agrado							
CALBEN	0.7904081	0.0149322	52.93	0	0.7611415	0.8196746	
_cons	4.936389	0.1334925	36.98	0	4.674748	5.198029	
p39_calif_apoyo							
CALBEN	0.8139786	0.0146893	55.41	0	0.7851881	0.8427691	
_cons	5.307955	0.1627439	32.62	0	4.988983	5.626927	
p40_satisf_benefic							
CALBEN	0.8179973	0.0147535	55.44	0	0.7890809	0.8469137	
_cons	4.868765	0.1429077	34.07	0	4.588671	5.148859	
p46_cambia_partic							
VALEXT	0.9767527	0.0189478	51.55	0	0.9396157	1.01389	
_cons	2.916397	0.0584561	49.89	0	2.801825	3.030969	
p47_activ_comun							
VALEXT	0.7640924	0.0193151	39.56	0	0.7262355	0.8019493	
_cons	2.439058	0.0461774	52.82	0	2.348552	2.529564	
p57_satisf_apoyo							
SATIS	0.8484667	0.0171281	49.54	0	0.8148962	0.8820372	
_cons	4.900623	0.1491628	32.85	0	4.60827	5.192977	
p56_calif_apoyo							
SATIS	0.8222963	0.016294	50.47	0	0.7903606	0.8542321	
_cons	4.93387	0.1434171	34.4	0	4.652778	5.214962	
p58_mejora_condic							
SATIS	0.7775995	0.0163612	47.53	0	0.7455321	0.8096669	
_cons	4.602722	0.1311815	35.09	0	4.345611	4.859833	

Fuente: Estimaciones propias

Es importante señalar que, al ser los resultados estandarizados, los valores de los coeficientes que se asocian a los constructos latentes representan la magnitud de impacto que tiene cada uno de estos sobre los constructos que determinan. En el caso de la variable *expectativa* (EXPEC) el valor del coeficiente más alto se asocia a la variable de *calidad del beneficio* (CALBEN), lo que representa este resultado es que los beneficiarios del programa PDZP construyen sus expectativas principalmente sobre la calidad del apoyo que reciben incluso por encima de la gestión y los impactos de las externalidades, además, de todas las variables latentes que se asocian con la satisfacción, es la calidad del beneficio (CALBEN) la que tiene un mayor impacto en la determinación de la satisfacción de los beneficiarios del programa PDZP. La calidad de la gestión también contribuye en la explicación de la satisfacción de los beneficiarios, aunque el peso que le proporciona los beneficiarios es menor en comparación con la calidad del beneficio. La variable de valoración de las externalidades, tiene un impacto muy bajo sobre la determinación de la satisfacción, lo que significa que es a la que menos peso estadístico se le atribuye en la explicación de la satisfacción.

Ahora bien, es importante establecer el ajuste del modelo que acabamos de calcular, para ello es necesario estimar el modelo sin coeficientes estandarizados de la siguiente forma:

```
sem (IMAGEN -> p17_claridad_info ) (IMAGEN -> p18_apropiada_info ) (IMAGEN -> EXPEC ) (IMAGEN -> SATIS ) (EXPEC -> p24_condic_vida) (EXPEC -> p25_ayud_mejor) (EXPEC -> CALGES) (EXPEC -> CALBEN ) (EXPEC -> VALEXT) (EXPEC -> SATIS ) (EXPEC -> p21_cambian_condic ) (CALGES -> p33_satisf_trato ) (CALGES -> SATIS ) (CALGES -> p31_amable ) (CALBEN -> p38_apoyo_agrado ) (CALBEN -> p39_calif_apoyo) (CALBEN -> p40_satisf_benefic) (CALBEN -> SATIS ) (VALEXT -> p46_cambia_partic) (VALEXT -> p47_activ_comun ) (VALEXT -> SATIS ) (SATIS -> p57_satisf_apoyo) (SATIS -> p56_calif_apoyo) (SATIS -> p58_mejora_condic), latent(IMAGEN EXPEC CALGES CALBEN VALEXT SATIS ) nocapslatent
```

En este caso no es necesario exhibir los resultados de la estimación, por lo que se han omitido; lo importante es hacer la estimación y posteriormente calcular los estadísticos que permiten valorar el ajuste del modelo estructural de satisfacción a través del comando **estat gof, stats(all)**. Dentro de los estadísticos más importantes se encuentran:

- La prueba Chi-cuadrada se desea que sea lo menos posible y que su valor-p sea mayor de 0.05 esto significaría que los datos se ajustan muy bien al modelo.
- Índice de Aproximación de la Raíz de Cuadrados Medios del Error (RMSEA). Debe ser menor o igual a 0.09. Si es mayor a 0.1, el modelo puede ser mejorado. Un valor menor a 0.05 indica que el ajuste del modelo es aceptable, aunque es más deseable uno cercano a cero. Una limitación de este índice es que, como su expresión involucra al tamaño de muestra, para muestras pequeñas tiende a sobreestimar el ajuste del modelo.
- Criterio de Información Bayesiana (bic) y Criterio de información de Akaike (AIC). Son índices que toman en cuenta la complejidad del modelo y el grado de ajuste; lo atractivo de estos dos índices es que, cuando se cuenta con varias versiones del modelo original, se pueden comparar entre sí por medio de los valores obtenidos de estos índices en cada uno de los modelos, prefiriendo a aquel cuyos índices sean los de menor valor.
- Índice Ajustado Comparativo (cfi) y el Índice de Ajuste No Normado (TLI). Consideran los grados de libertad y el tamaño de muestra. Se prefiere que su valor sea mayor o igual a 0.9.
- Raíz Cruada Media del Residual Estandarizado (srmr). Se prefiere que su valor esté lo más proximo a cero sin que su valor exceda el umbral de 0.08.
- Coeficiente de Determinación (CD). Se asocia al estadístico de R cuadrada y representa el ajuste global del modelo, se prefiere que su valor se acerque a la unidad.

Estos estadísticos se estiman en el cuadro 8.9. En general, los resultados que se presentan permiten identificar que existe un buen ajuste del modelo

de satisfacción de beneficiarios del programa PDZP, por lo que el método de ecuaciones estructurales con variables latentes ha resultado pertinente para evaluar dicho modelo de satisfacción.

estat gof, stats(all)

Cuadro 8.9. Medidas de ajuste del modelo de satisfacción del programa PDZP

<i>Fit statistic</i>	<i>Value</i>	<i>Description</i>
Likelihood ratio		
chi2_ms(81)	875.033	model vs. saturated
p > chi2	0	
chi2_bs(105)	16230.241	baseline vs. saturated
p > chi2	0	
Population error		
RMSEA	0.063	Root mean squared error of approximation
90% CI, lower bound		
upper bound	0.067	
pclose	0	Probability RMSEA <= 0.05
Information criteria		
AIC	141504.04	Akaike's information criterion
BIC	141817.93	Bayesian information criterion
Baseline comparison		
CFI	0.951	Comparative fit index
TLI	0.936	Tucker-Lewis index
Size of residuals		
SRMR	0.048	Standardized root mean squared residual
CD	0.84	Coefficient of determination

Fuente: Estimaciones propias.

Para conocer variable por variable, tanto las manifiestas como las latentes, su contribución en el ajuste del modelo, se hace uso del comando **estat eqgof**, posterior a la estimación del modelo, los resultados se presentan en el cuadro 8.10. De los resultados previos se conoce que el R cuadrada global fue de 0.84, en este cuadro se puede identificar cuál de la variables latentes y manifiestas se ajusta mejor al modelo.

estat eqgof

Cuadro 8.10. *Descomposición de la varianza por variables en el modelo de satisfacción*

<i>Equation-level goodness of fit</i>						
depvars	fitted	variance predicted	residual	R-squared	mc	mc2
observed						
p17_clarid~o	5.85138	3.200754	2.650626	0.5470084	0.7396001	0.5470084
p18_apropi~o	4.971138	3.847386	1.123752	0.7739447	0.8797413	0.7739447
p24_condic~a	3.017307	1.104169	1.913138	0.3659452	0.6049341	0.3659452
p25_ayud_m~r	2.876878	1.122914	1.753964	0.390324	0.6247591	0.390324
p21_cambia~c	5.033708	0.5998125	4.433896	0.1191592	0.3451944	0.1191592
p33_satisf~o	4.03494	2.669439	1.365501	0.6615809	0.8133762	0.6615809
p31_amable	5.646388	0.7379592	4.908428	0.1306958	0.3615188	0.1306958
p38_apoyo_~o	3.01911	1.886174	1.132936	0.6247449	0.7904081	0.6247449
p39_calif_~o	2.670916	1.769645	0.9012706	0.6625612	0.8139786	0.6625612
p40_satisf~c	3.148806	2.106928	1.041878	0.6691196	0.8179973	0.6691196
p46_cambia~c	6.11526	5.834238	0.2810221	0.9540458	0.9767527	0.9540458
p47_activ_~n	7.719554	4.506962	3.212591	0.5838372	0.7640924	0.5838372
p57_satisf~o	3.098111	2.230317	0.8677943	0.7198957	0.8484667	0.7198957
p56_calif_~o	3.012509	2.036972	0.975537	0.6761713	0.8222963	0.6761713
p58_mejora~c	3.402961	2.057638	1.345323	0.604661	0.7775995	0.604661
latent						
EXPEC	1.104169	0.417368	0.6868011	0.3779928	0.6148112	0.3779928
SATIS	2.230317	1.720918	0.5093986	0.7716026	0.8784091	0.7716026
CALGES	2.669439	1.289861	1.379578	0.4831954	0.6951226	0.4831954
CALBEN	1.886174	1.008489	0.8776842	0.5346747	0.7312146	0.5346747
VALEXT	5.834238	1.41289	4.421348	0.2421722	0.4921099	0.2421722
overall				0.8397555		

Fuente: Estimaciones propias.

Una vez estimado el modelo, y dado el buen ajuste de éste, realizamos el cálculo del nivel de satisfacción que experimentaron los beneficiarios del programa PDZP, en sus diferentes vertientes, para lo cual se requieren los coeficientes estandarizados que se obtuvieron de la estimación previa relacionadas con el constructo latente de satisfacción, y de igual forma se re-

quieren las medias de las variables manifiestas de ese mismo constructo. En el caso de estas últimas, podríamos emplear las medias que aparecen en el cuadro 8.7, sin embargo, al existir un número alto de valores perdidos, es necesario calcular las medias empleando únicamente los datos que se utilizaron en la estimación del modelo de satisfacción, estos resultados se obtienen después de la estimación a través del comando **estat summarize**, los resultados se presentan en el cuadro 8.11.

estat summarize

Cuadro 8.11. *Estadísticas descriptivas postestimación del modelo de satisfacción*

<i>Estimation sample sem</i>				
Number of obs = 2,472				
Variable	Mean	Std. Dev.	Min	Max
p17_clarid~o	7.929207	2.419452	1	10
p18_apropi~o	8.061489	2.230056	1	10
p24_condic~a	8.592638	1.737392	1	10
p25_ayud_m~r	8.564725	1.69648	1	10
p21_cambia~c	8.321602	2.244047	1	10
p33_satisf~o	8.623382	2.009122	1	10
p31_amable	8.540858	2.376694	1	10
p38_apoyo~o	8.577265	1.73791	1	10
p39_calif_~o	8.674757	1.634624	1	10
p40_satisf~c	8.639563	1.774847	1	10
p46_cambia~c	7.211974	2.473406	1	10
p47_activ_~n	6.776699	2.778971	1	10
p57_satisf~o	8.625809	1.767423	1	10
p56_calif_~o	8.563511	1.742421	1	10
p58_mejora~c	8.490696	1.85118	1	10

Fuente: Estimaciones propias

Con los resultados de los cuadros 8.8 y 8.11 es factible establecer el valor de nivel de satisfacción de los beneficiarios, utilizando las medias de las variables manifiestas correspondiente a la variable latente de satisfacción y los coeficientes estimados del modelo, de la siguiente forma:

$$\begin{aligned} \text{satisfacción} &= \frac{(0.8223)(8.5635) + (0.8485)(8.6258) + (0.7776)(8.4907)}{0.8223 + 0.8485 + 0.7776} = \frac{20.9628}{2.4484} \\ &= 8.6 \end{aligned}$$

El resultado significa que los beneficiarios tienen un nivel de satisfacción alrededor de 8.6 en una escala de 1 a 10.

Referencias

- Acemoglu, D. (2009). *Introduction to modern economic growth*. Princeton University Press.
- Bowerman, B. L., O'Connell, R. T. y Murpree, E. S. (2011). Times Series Forecasting and Index Numbers. En *Business statistics in practice* (6ª ed.). Boston: McGraw-Hill.
- Cogco, A., Pérez, J., y Martínez, O. (2013) Satisfacción de programas sociales. El caso del programa de abasto de Leche Liconsa. *Revista de Economía del Rosario*, 16(1), 125-147.
- DeLurgio, S. A. (1998a). Simple Methods. En *Forecasting Principles and Applications* (pp. 147-173). Singapur: McGraw Hill.
- DeLurgio, S. A. (1998b). Descomposition Methods and Seasonal Indexes. En *Forecasting Principles and Applications* (pp. 173-203). Singapur: McGraw Hill.
- Gardner, E. S. (2006). Exponential Smoothing: The State of the Art: part II. *International Journal of Forecasting*, 22, 637-666.
- Gardner, E. S. (1985). Exponential Smoothing: The State of the Art: part I. *International Journal of Forecasting*, 4, (1-2S), 1-38.
- Gujarati, D., y Porter, D. (2010). *Econometría* (5ª ed.). McGraw-Hill.
- Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6), 1251-1271.
- Hovhannisyan, A., Castillo-Ponce, R., y Valdez, R. I. (2019). The Determinants of Income Inequality: The Role of Education. *Scientific Annals of Economics and Business*, 66(4), 451-464. <https://doi.org/10.47743/saeb-2019-0040>
- Lind, D. A., Marchal, W. G., y Wathen, S. A. (2019). Métodos no paramétricos: análisis de datos ordenados. En *Estadística aplicada a los negocios y la economía* (17ª ed.) (pp. 680-719). McGraw Hill.
- Makridakis, S. G., Wheelwright, S. C., y Hyndman, R. J. (1997). *Forecasting: Methods and Applications* (3ª ed.). Wiley.
- Mankiw, G. N., Romer, D., y Weil, D. N. (1992). A contribution to the empirics of economic growth. *Quarterly Journal of Economics*, 107(2), 407-437. <https://doi.org/10.2307/2118477>

- Otero, J. M. (1993). Modelos de alisado. En *Econometría. Series temporales y predicción*. Madrid: AC.
- Ord, K., y Fildes, R. (2013). *Principles of Business Forecasting*. South-Western Cengage Learning.
- Peña, D. (2005). Predicción con modelos ARIMA. En *Análisis de series temporales* (223-254). Madrid: Alianza.
- Pérez Cruz, J. A., Martínez-Martínez, O. A., y Cogco Calderón, A. R. (2017). ¿Satisfacción con programas de fomento a la artesanía en México? El caso de Fonart. *Investigación administrativa*, 46(120). http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2448-76782017000200002&lng=es&tlng=es
- Rodríguez Vargas, M., Cogno, A. R., Islas Camargo, A., Herrera Ramos, J. M., Martínez Martínez, O. A., Pérez Cruz, J. A., Canales Sánchez, A., y López Sandoval, I. M. (2012). *Índice Mexicano de Satisfacción de los beneficiarios de Programa Sociales implementados por la SEDESOL en México (IMSAB). Informe final*. México: Universidad Autónoma de Tamaulipas.
- Rodríguez Vargas, M., Cogco Calderón, A. R., y Pérez Cruz, J. A. (2014). *Evaluación de la satisfacción de los beneficiarios del Programa para el Desarrollo de Zonas Prioritarias (PDZP) 2014. Informe final*. México: Universidad Autónoma de Tamaulipas. https://www.gob.mx/cms/uploads/attachment/file/28288/IF_ESB_PDZP_2014.pdf
- Stata Corp. (2015). *Stata Time-Series. Reference manual release 14, College Station* (pp. 651-652). Texas: A Stata Press Publication StataCorp LP College Station.
- Solow, R. M. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics*, 70(1), 65-94. <https://doi.org/10.2307/1884513>
- Sontheimer, K., y Fuente, E. (1974). La predicción como finalidad y problema de la ciencia social moderna. *Revista española de la opinión pública*, (35), 7-21.
- Valdez, R. I., Pérez-Cruz, J. A., y Estrada-Danell, R. I. (2022a). Salario y crecimiento económico municipales en México, 1988-2018. *Paradigma económico. Revista de economía regional y sectorial*, 14(2), 85-108.
- Valdez, R. I., y García, F. F. (2022b). The distribution of wage inequality across municipalities in Mexico: a spatial quantile regression approach. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, 17(3), 669-697.
- Valdez Ramírez, R. I., y Hernández Gómez, E. (2019). Impacto de la homologación del IVA en el consumo de los hogares de Baja California, Baja California Sur y Quintana Roo, México. *Estudios regionales en economía, población y desarrollo: Cuadernos de trabajo de la UACJ*, 9(51). <https://revistas.uacj.mx/ojs/index.php/estudiosregionales/article/view/3137/5093>
- Weber, L. (2010). *Demographic Change and Economic Growth: Simulations on Growth Models*. Physica-Verlag.
- Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324-342.
- Wooldridge, J. M. (2020). *Introductory Econometrics: A modern approach* (7ª ed.). Cengage Learning.

Sobre los autores

Jorge Alberto Pérez Cruz

Doctor en Ciencias Económicas por la Universidad Autónoma de Baja California. Obtuvo la maestría en Economía aplicada por El Colegio de La Frontera Norte y tiene licenciatura en Economía por la Universidad Autónoma de Tamaulipas. Es profesor investigador de tiempo completo en la Universidad Autónoma de Tamaulipas, asimismo, es profesor invitado en la Universidad Anáhuac Online. Es miembro del Sistema Nacional de Investigadores, nivel I, docente con perfil deseable PRODEP y líder del Cuerpo Académico Consolidado “Bienestar Económico y Social”. Sus principales líneas de investigación son: crecimiento económico, aglomeración industrial y políticas de bienestar.

Publicaciones recientes:

Pérez Cruz, J. A., y Estrada Danell, R. I. (2022). Medición de la pobreza multidimensional desde una perspectiva local. En J. A. Pérez Cruz, G. I. Ceballos Álvarez, A. R. Cogno Calderón y R. I. Estrada Danell (Coords.), *De la medición de la pobreza a la estrategia de intervención comunitaria en el sur de Tamaulipas, 2022* (pp. 51-70). México: Ediciones Comunicación Científica.

Pérez Cruz, J. A., y Ceballos Álvarez, G. I. (2022). En J. A. Pérez Cruz, G. I. Ceballos Álvarez, A. R. Cogno Calderón y R. I. Estrada Danell (Coords.), *De la medición de la pobreza a la estrategia de intervención comunitaria en el sur de Tamaulipas, 2022* (pp. 89-145). México: Ediciones Comunicación Científica.

Huerta Wong, J. E., Martínez Martínez, O. A., y Pérez Cruz, J. A. (2023). In-

roducción. Políticas del Bienestar. ¿Nuevas preguntas, nuevas hipótesis, mismos desafíos? En O. A. Martínez Martínez, A. R. Cogno Calderón y J. A. Pérez Cruz (Coords.), *Política social en tiempos de la Cuarta Transformación. Continuidad o cambio de paradigma* (pp. 11-31). México: Ediciones Comunicación Científica.

Ceballos Álvarez, G. I., y Pérez Cruz, J. A. (2022). Experiencias docentes ante la enseñanza remota de emergencia ocasionada por el COVID-19. En A. M. Martínez Jerez, G. I. Ceballos Álvarez y J. A. Pérez Cruz (Coords.), *Retos y oportunidades de una pandemia. Reflexiones en torno a los aprendizajes a partir del COVID-19* (pp. 161-186). México: Universidad de Tamaulipas/Anáhuac online/ Ediciones Comunicación Científica.

ORCID: <https://orcid.org/0000-0003-4435-0339>

ResearchGate: <https://www.researchgate.net/profile/Jorge-Perez-Cruz-2>

Google Academic: <https://scholar.google.com/citations?user=FRdBB7gAAAA-J&hl=en>

Rolando Israel Valdez Ramírez

Doctor en Ciencias Económicas por la Universidad Autónoma de Baja California, maestro en Economía por la Universidad Autónoma de Puebla y licenciado en Economía por la Universidad Autónoma de Tamaulipas. Es profesor de tiempo completo en la Universidad Autónoma de Tamaulipas. Fue candidato a Investigador Nacional por parte del Consejo Nacional de Humanidades, Ciencia y Tecnología (Conahcyt) del 2018 al 2021 y recibió el Mérito Académico por parte de la Universidad Autónoma de Baja California. Sus principales líneas de investigación son: crecimiento económico, economía regional y seguridad alimentaria.

Publicaciones recientes:

Valdez-Ramírez, R. I., y Sobrevilla, V. (2021). The sectoral-regional structure of the wages in Mexico. *Papeles de Población*, 27(108), 185-209. <https://rppoblacion.uaemex.mx/article/view/12847>

Determinantes del tamaño de las empresas recién nacidas en México: Un análisis de datos panel (2020).

Desigualdad salarial en México: un enfoque sectorial y regional con regresión por cuantiles (2019).

ORCID: <https://orcid.org/0000-0002-1491-305X>

ResearchGate: <https://www.researchgate.net/profile/Rolando-Valdez-2>

Google Academic: <https://scholar.google.es/citations?user=4ou0FOgAAAAJ&hl=es>

Academia: <https://uat-mx.academia.edu/RolandoValdez>

Fortino Vela Peón

Maestro en Demografía por El Colegio de México y Licenciado en Economía por la Universidad Autónoma Metropolitana, campus Iztapalapa (UAM-I). Es profesor de economía en el Departamento de Producción Económica de la Universidad Autónoma Metropolitana, campus Xochimilco (UAM-X), donde imparte cursos de Estadística y Econometría.

Su investigación actual incluye temas de migración y relaciones internacionales, mortalidad infantil, mercados laborales, población y desarrollo, y sus diferentes respuestas a cambios en la estructura de edad. Se desempeñó como coordinador de la licenciatura en Economía (2016-2019) y de la maestría en Relaciones Internacionales (2020-2022), ambas en la UAMX.

Rodríguez Nava, A., Vela Peón, F., y García Villanueva, C. A. (Coords.) (2022), *Trabajo, pobreza, pobreza laboral*. México: UAM.

Vela Peón, F., y López Ponce, C. M. (2020). La utilidad del Análisis de Redes Sociales (ARS) cómo estrategia para el análisis social. *Constante Regional*, 8(15), 103-119.

Vela Peón, F., y García Álvarez, M. (2019). Dinámica del comercio: un análisis no paramétrico para México y Estados Unidos (1965-2011). En M. A. Correa Serrano y J. E. Mendoza Cota (Coords.), *Estrategia de los bloques regionales. Libre comercio y regulación internacional* (pp. 151-184). México: UAM/ Itaca.

Vela Peón, F., Rodríguez, A., y Raymundo, A. (2017). Democracia y derechos humanos: Avances a partir de la Constitución Política de los Estados Unidos Mexicanos. En J. Flores (Coord.), *100 años de la Constitución de 1917. Análisis interdisciplinarios* (2017). México: UAM-X.

ORCID: <https://orcid.org/0000-0003-3887-5534>

ResearchGate: <https://www.researchgate.net/profile/Fortino-Peon>

Google Academic: <https://scholar.google.com/citations?user=XdFQ7NsAAAAJ&hl=en>

Academia: <https://independent.academia.edu/FortinoVelaPe%C3%B3n>

Investigación Aplicada
a las Ciencias Sociales en Stata, de Jorge
Alberto Pérez Cruz, Rolando Israel Valdez
Ramírez y Fortino Vela Peón, editado y publicado
por Ediciones Comunicación Científica, S. A. de C. V., se
publicó en acceso abierto en los formatos PDF, Epub3 y HTML5,
en febrero de 2025.

En la actualidad, se ha observado que la cantidad de datos que se generan sobre prácticamente todo lo que nos rodea crece a un ritmo acelerado y gran parte de estos se encuentran disponibles en línea, lo que hace que estén al alcance para todos aquellos que tengan interés por realizar análisis cuantitativo, ya sea desde una perspectiva de la estadística descriptiva o inferencial. Esta forma tan vertiginosa en que se genera la información implica una serie de retos. Sin duda, existen más retos que afrontar en este escenario tan dinámico y globalizado de la información. La capacidad de adaptabilidad a este entorno demanda al analista cuantitativo el conocimiento de métodos de estimación, el manejo de bases de datos, la construcción de variables, el uso de software especializado y capacidad analítica para la toma de decisiones.

Este libro está diseñado para estudiantes, profesionistas, investigadores y todo aquel interesado en el uso de Stata para el análisis de datos. En los distintos capítulos se muestran los procesos sistemáticos que implican desarrollar e implementar análisis de estadística descriptiva, el uso de gráficos y la estimación de modelos de regresión lineal y logística a través de aplicaciones en donde se emplearán estructuras de datos como corte transversal, series de tiempo y de panel, el cual brinda la posibilidad de utilizarse a través de pantallas o comandos, además de que brinda la posibilidad de programar con un lenguaje sencillo. La selección de las aplicaciones responde a los principales retos que enfrentan en la actualidad los profesionistas e investigadores.



Jorge Alberto Pérez Cruz es Doctor en Ciencias Económicas por la Universidad Autónoma de Baja California. Profesor investigador de tiempo completo en la Universidad Autónoma de Tamaulipas y profesor invitado en la Universidad Anáhuac Online. Es miembro del SNII, nivel I, docente con perfil deseable PRODEP y líder del Cuerpo Académico Consolidado Bienestar Económico y Social.



Rolando Israel Valdez Ramírez es Doctor en Ciencias Económicas por la Universidad Autónoma de Baja California. Profesor de tiempo completo en la Universidad Autónoma de Tamaulipas. Fue candidato a Investigador Nacional por parte del Consejo Nacional de Ciencia y Tecnología del 2018 al 2021 y recibió el Mérito Académico por parte de la Universidad Autónoma de Baja California.



Fortino Vela Peón es Maestro en Demografía por El Colegio de México y Licenciado en Economía por la Universidad Autónoma Metropolitana-Iztapalapa. Profesor de economía en el Departamento de Producción Económica de la Universidad Autónoma Metropolitana campus Xochimilco, en donde imparte cursos de estadística y econometría.



Dimensions



RENIICYT
Registro Nacional de Publicaciones
y Programas Científicos y Tecnológicos
2000922



Google
Scholar



DOI.ORG/10.52501/CC.155