SEVEN
EDITORA
2025

HALLAZGOS DE POLÍTICA SOCIAL USANDO STATA:
Casos Prácticos

Emmanuel Olivera Perez

EDITORA CHEFE
Profº Me. Isabele de Souza Carvalho

EDITOR EXECUTIVO
Nathan Albano Valente

AUTOR DO LIVRO
Emmanuel Olivera Pérez

PRODUÇÃO EDITORIAL
Seven Publicações Ltda

EDIÇÃO DE ARTE
Evellyn Thais de Souza

EDIÇÃO DE TEXTO
Stephanie Caroline Meyer de Quadros

BIBLIOTECÁRIA
Bruna Heller

IMAGENS DE CAPA
Evellyn Thais de Souza

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

P438h Pérez, Emmanuel Olivera.
Hallazgos de Política Social Usando Stata [recurso eletrônico] : Casos Prácticos / Emmanuel Olivera Pérez. – São José dos Pinhais, PR: Editora Seven, 2025.
Dados eletrônicos (1 PDF).

Inclui bibliografia.
ISBN 978-65-6109-246-3

1. Política social. 2. Ciências políticas. 3. Sociologia. I. Título.

CDU 32

Bruna Heller - Bibliotecária - CRB10/2348

Índices para catálogo sistemático:

1 Política 32

DOI: 10.56238/livrosindi202564-001

Seven Publicações Ltda
CNPJ: 43.789.355/0001-14
editora@sevenevents.com.br
São José dos Pinhais/PR

AUTOR DEL LIBRO

Emmanuel Olivera Pérez

Profesor Investigaron en Universidad Popular Autónoma del Estado de Puebla (UPAEP)

SEVEN
publicações acadêmicas

PRESENTACIÓN

El presente Libro se realiza con el compromiso social de cumplir dos principales propósitos. Por un lado, presentar y orientar al lector en el uso del paquete de software estadístico Stata, dotado con diversas funciones para aplicarse en cualquier estudio que involucre datos cuantitativos, abarcando desde un análisis de estadística descriptiva hasta tareas específicas para ciencia de datos. Cuenta con el potencial de procesar grandes volúmenes de información para obtener visualizaciones, resumen estadístico y reportes con diferentes niveles de desagregación.

Esta herramienta es ampliamente usada tanto en el ámbito académico, como en el sector privado, dado su gran capacidad, facilidad de uso, diversidad y constante actualización de paquetes complementarios para las necesidades de los análisis requeridos y su compatibilidad con otras herramientas, ya que los resultados procesados por el programa pueden tener diferentes formatos de salida, lo que facilita la exportación a otros programas o plataformas, en caso de requerirlo.

El segundo propósito es presentar un compendio de técnicas de análisis estadísticos bivariado y multivariado, las cuales se emplean para diferentes fines, como la descripción, agrupamiento o búsqueda de relaciones entre los datos, así como también la tarea de predicción. El uso de cada una de estas debe estar determinado por el objetivo de estudio y las variables que lo compongan, ya que cada método cuanta con sus propios requisitos que deben de tomarse en cuenta para obtener un resultado estadísticamente representativo y correcto.

En resumen, le presente libro aborda desde la preparación de los datos, incluyendo limpieza y operacionalización de los mismos, pasando por el análisis descriptivo, llegando hasta la inferencia estadística con un enfoque multivariado, diferenciando las aplicaciones en cuanto al uso de ciertos métodos basado en modelos de variables métricas y no métricas. Tales técnicas multivariadas como regresión múltiple, análisis de componentes principales, análisis factorial y análisis discriminante, análisis de clúster y correlación canónica, actualmente son las más usadas en ciencias sociales en diferentes aplicaciones temáticas.

SUMÁRIO

RESUMÉN

El presente libro muestra la aplicación de una diversidad de métodos estadísticos cuyo propósito es doble para la investigación científica: el primero es ofrecer una guía detallada sobre el uso de STATA para el desarrollo de los métodos estadísticos y el segundo, ofrece una metodología para el seguimiento, control y evaluación de políticas públicas, donde el principal interés es una visión desde la ciencia de datos. En este sentido, por medio de bases de datos reales, se integran casos prácticos desarrollando preguntas de investigación para visualizar por medio del software los principales instrumentos técnicos que permitan obtener resultados confiables en un contexto de política pública con el objetivo de hallar evidencia empírica que oriente el mejor funcionamiento y retroalimentación de acciones y programas promovidos desde el gobierno como principal generador de políticas públicas para el bien común de la sociedad. La importancia de este tipo de análisis es favorecer la toma de decisiones de manera técnica, de tal forma que se reduzcan ineficiencias atribuibles a la medición de los resultados de programas o acciones públicas.

El documento se presenta por capítulos cuyo orden busca acompañar al lector en cada parte del proceso del análisis, que inicia desde la introducción de los datos a Stata, la preparación y limpieza de los mismos, hasta la descripción de cada una de las técnicas que se abordan, mostrando detalladamente cómo ejecutarlas en el programa y qué es lo que se obtiene como resultado.

También se puede encontrar una reseña general y puntual de los métodos, con el fin de, además de comprender su funcionamiento, conocer los requerimientos, detección de problemas al momento de ejecutarse y sus posibles soluciones, para su ejecución. Esto ayuda a determinar si es viable hacer uso de la técnica elegida o si se tiene que recurrir a otra, ya que para cualquier trabajo de investigación es importante garantizar la fiabilidad de los resultados y el procedimiento para llegar a ellos.

Con el objetivo de facilitar y mostrar el desarrollo detallado del uso las técnicas estadísticas, estas se explican paso a paso por medio de ejemplos con bases de datos abiertos y de acceso libre de casos reales, de esta manera se considera que la compresión e interpretación de los resultados se observe claramente.

De igual manera se exponen los hallazgos obtenidos y qué posible alcance pueden tener si fueran tomados en cuenta en materia de creación, evaluación y seguimiento de política pública. Esto es debido a que los conjuntos de datos utilizados están relacionados con el área demografía, salud y ambiental, dentro de las cuales las técnicas apoyaron para encontrar *insights* de interés que bien funcionan como puntos de atención para atender posibles problemáticas detectadas.

Por lo anterior en el capítulo uno se hace una breve descripción del área de trabajo que tiene Stata, así el lector puede conocer las partes que lo componen, junto con una serie de comandos comúnmente empleados para hacer la exploración de los datos que se analizan, comprensión de las variables y acciones de limpieza básica para preparar los datos para la siguiente fase del proceso.

El capítulo dos está enfocado en la realización del análisis descriptivo del conjunto de datos, por medio de técnicas estadísticas dirigidas a observar la distribución de los datos, cómo acercarla a un comportamiento normal, fraccionar variables de interés en quintiles, análisis de correlación entre estas, análisis de contingencia y pruebas de independencia. Con esta primera parte del compendio se pretende que el analista ya tenga un conocimiento basto de los datos que este trabajando y tenga claridad de cuáles son las particularidades de estos, ya que con ello se determina qué técnicas son posibles de ejecutar según de acuerdo a las características ya detectadas.

En el capítulo tres se describe detalladamente la técnica estadística de análisis de regresión lineal y análisis de regresión múltiple, las estudian la posible relación entre dos o más variables donde una se asume como independiente y el resto como dependientes. Con el análisis se hace posible hacer inferencias sobre la distribución de probabilidad de influencia de la o las variables dependientes sobre la independiente. También se explica cómo hacer y el para qué se utilizan las pruebas de hipótesis, que son posibles gracias a los hallazgos obtenidos por los análisis de regresión.

El capítulo cuarto está dedicado a la explicación del análisis de regresión logística, ampliamente usado en el área de salud y social. En esencia este método calcula la probabilidad de éxito o fracaso de un evento dado, con la peculiaridad que sólo puede ejecutarse si la variable dependiente es categórica con dos opciones de respuesta únicamente, las cuales dependerá del analista determinar cuál opción represente el éxito o fracaso del evento.

La siguiente parte del compendio introduce de forma general qué es el análisis multivariado y el paso a paso de cómo ejecutar e interpretar algunas técnicas de esta naturaleza: componentes principales, análisis factorial, correlación canónica, análisis discriminante y análisis clúster. La información que ofrecen la aplicación de estas técnicas dependerá de lo que se requiera y pretenda encontrar en los datos, ya que cada una de ellas se emplea para objetivos específicos, pero evidentemente muestran una perspectiva más completa para entender al fenómeno de estudio, al tener la capacidad de poder integrar más variables al análisis.

El uso de cualquier técnica estadística descrita en este compendio puede ayudar al alcance de los diferentes objetivos que se planteen, el resultado obtenido sin duda ofrece información de valor que contribuya al entendimiento mayor al fenómeno de estudio, para la toma de decisiones respaldadas con información veraz y segura, que en materia de políticas pública es muy importante tener información de calidad que oriente ya sea al planteamiento de iniciativas, programas o cualquier acción institucional en favor de la sociedad.

El análisis bivariado permite estudiar simultáneamente dos variables que describan un fenómeno, con el objetivo de extraer información tanto a nivel descriptivo como inferencial.

En este apartado muestra la aplicación de diversos métodos descriptivas e inferenciales, dentro de los que podemos destacar el uso del histograma, correlación lineal, regresión lineal, R2 de Shapley y diversas pruebas para validación de la significancia estadística y bondad de ajuste de los modelos.

A través de ejemplos prácticos y detallados, usando bases de datos públicas, esta sección demostrará cómo aplicar estas técnicas de análisis bivariado, subrayando su relevancia y utilidad en el diseño y la implementación de políticas públicas más informadas y efectivas, con base al uso del software estadística STATA 16.

Para una mayor referencia del contenido de este capítulo ver:

Cameron, A. C., & Trivedi, P. K. (2010). Microeconometrics Using Stata (Rev. ed.). Stata Press.

La base de datos utilizada para esta sección se denomina *nacimientos_2021*, obtenida del repositorio de datos abiertos de la Secretaría de Salud de México, la cual contiene el registro de los nacimientos ocurridos en 2021 en el país. Está compuesta por 64 variables y 1,639,479 observaciones, haciendo que su visualización en una hoja de cálculo de Excel se límite a la capacidad máxima de filas (1,048,576), por lo que este tipo de *data set* es necesario trabajarlas en softwares con mayor capacidad de volumen de información, de lo contrario los resultados de cualquier análisis serán incorrectos.

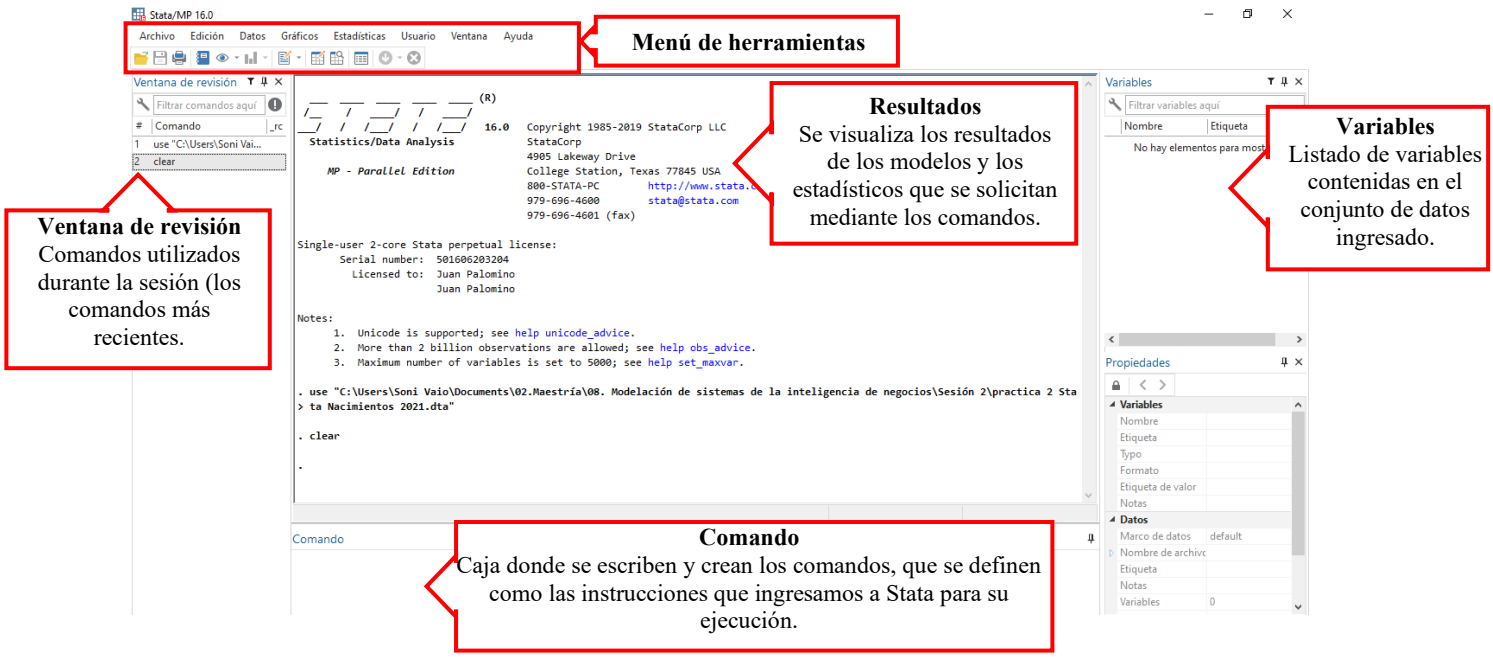
El conjunto de variables se divide en 3 categorías, datos de la madre o gestante, datos del nacido vivo y del nacimiento y datos del certificante. Sin embargo, con el fin de ejemplificar los pasos sobre el procedimiento y los futuros métodos estadísticos, se extrajeron las variables de interés que ayudaran a ejemplificar y desarrollas los análisis seleccionados a ejecutarse, reduciendo a un conjunto de datos con 21 variables. Estas son de tipo cuantitativo y 10 son categóricas binarias, las cuales se enlistan a continuación junto con su descripción:

Nombre de la variable	Descripción	Valor de registro
Madre_extranj~a	Extranjera	1 – Sí 0 – No
Edad	Edad de la Madre	
Se_cond_Indig~a	Si la madre se considera indígena	1 – Sí 0 – No
Habla_indigena	Habla alguna lengua indígena	1 – Sí 0 – No
Reside_Ext	Si la madre reside en el extranjero	1 – Sí 0 – No
No_embarazos	Número de embarazos	
Hijos_nac_mue~s	Hijos nacidos muertos	
Hijos_nac_vivos	Hijos nacidos vivos	
Hijos_sobrev	Hijos sobrevivientes	
Orden_nac	Orden de hijo del recién nacido	
Atencion_Pren	Recibió atención prenatal	1 – Sí 0 – No
Total_Consultas	Total de consultas recibidas durante el embarazo	
Sobrevivio_pa~o	Si el recién nacido sobrevivió el parto	1 – Sí 0 – No
Interrumpio_Est	Si la madre interrumpió sus estudios	1 – Sí 0 – No
Trabaja_Actual	Si la madre trabaja actualmente	1 – Sí 0 – No
Edad_padre	Edad del Padre	
Genero_nac	Genero del recién nacido	0 – Hombre

		1 – Mujer
Edad_gest	Edad gestacional del embarazo	
Talla	Talla del recién nacido en cm	
Peso	Peso del recién nacido en gramos	
Producto_emb	Si el recién nacido fue producto del embarazo	1 – Sí 0 – No

Con el conocimiento de lo que significa cada variable y su registro, es posible proceder con el listado de comandos usualmente utilizados para ingresar conjuntos de datos a Stata y realizar adecuaciones necesarias para su tratamiento.

Partes del área de trabajo de Stata:



¿Cómo ingresar datos a Stata?

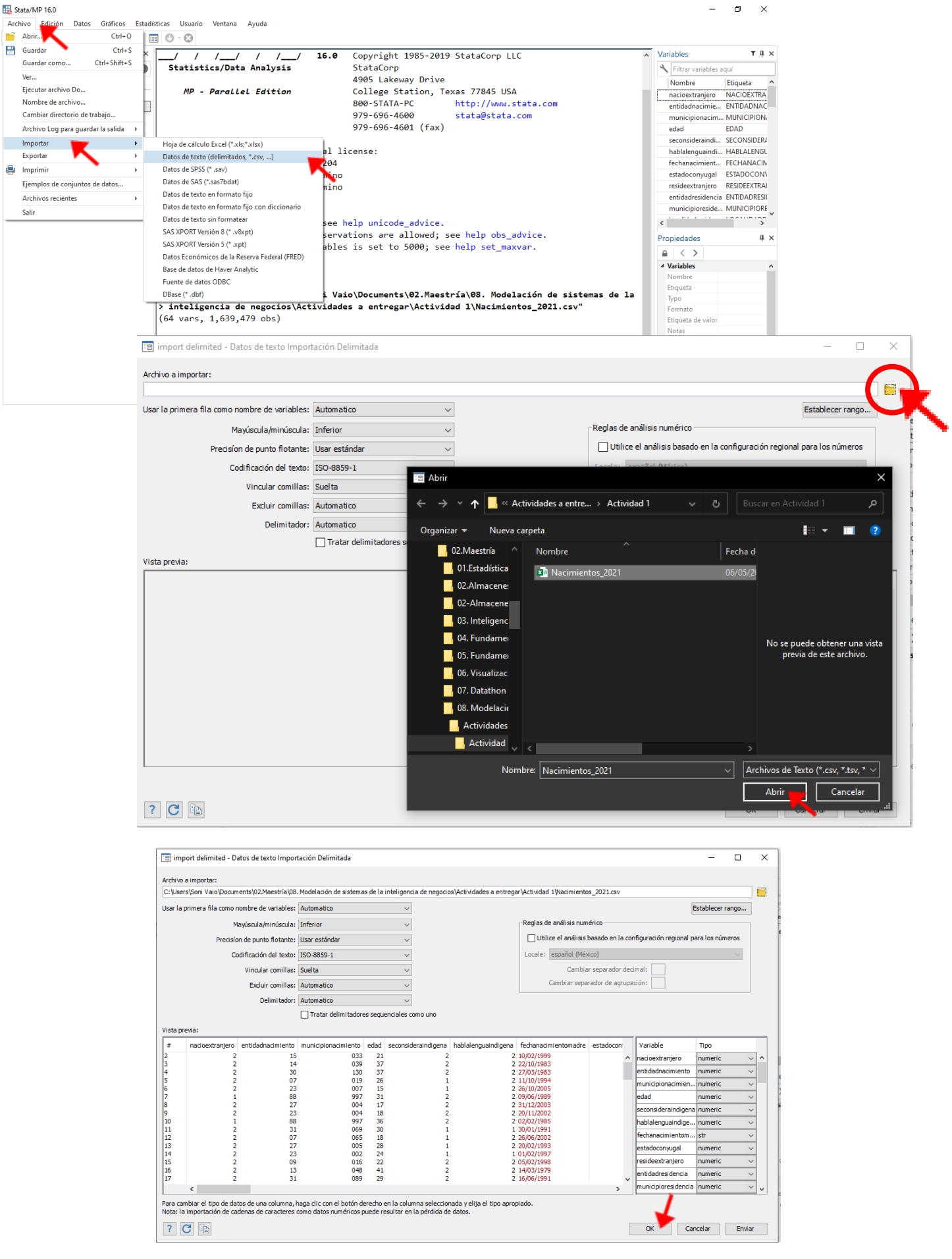
Una de las ventajas de trabajar con este programa es que se pueden ejecutar ciertas acciones por medio de comandos o apoyarse con el menú de herramientas, por ejemplo, el cargar un conjunto de datos de diferentes formatos (.csv, .xls, .txt, .sav, etc.).

La ruta para ingresar archivos a Stata desde el menú de Archivo es:

Archivo > Importar > Selección el tipo de archivo que se cargará

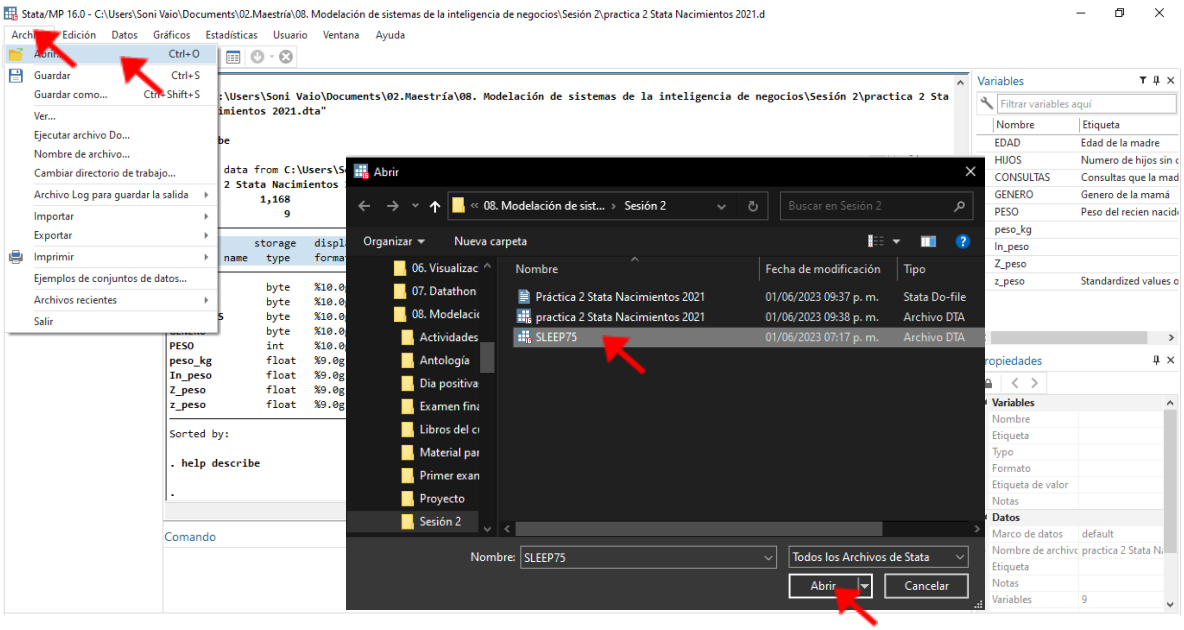
Como ejemplo se muestra el proceso para cargar un archivo delimitado por comas:

Archivo > Importar > Dato de texto delimitado por comas > Ubicar el archivo a cargar > Abrir > OK


Stata/MP 16.0
Archivo Edición Datos Gráficos Estadísticas Usuario Ventana Ayuda
Abrir... Ctrl+O
Guardar Ctrl+S
Guardar como... Ctrl+Shift+S
Ver...
Ejecutar archivo Do...
Nombre de archivo...
Cambiar directorio de trabajo...
Archivo Log para guardar la salida
Importar
Exportar
Imprimir
Ejemplos de conjuntos de datos...
Archivos recientes
Salir
Hoja de cálculo Excel (*.xls;*.xlsx)
Datos de texto (delimitados, *.csv, ...)
Datos de SPSS (* .sav)
Datos de SAS (*.sas7bdat)
Datos de texto en formato fijo
Datos de texto en formato fijo con diccionario
Datos de texto sin formatear
SAS XPORT Versión 8 (* .v8xpt)
SAS XPORT Versión 5 (* .xpt)
Datos Económicos de la Reserva Federal (FRED)
Base de datos de Haver Analytic
Fuente de datos ODBC
DBase (* .dbf)
Statistics/Data Analysis
MP - Parallel Edition
Copyright 1985-2019 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Variables
Propiedades
import delimited - Datos de texto Importación Delimitada
Archivo a importar:
Usar la primera fila como nombre de variables: Automatico
Establecer rango...
Mayúscula/minúscula: Inferior
Precisión de punto flotante: Usar estándar
Codificación del texto: ISO-8859-1
Vincular comillas: Suelta
Excluir comillas: Automatico
Delimitador: Automatico
Tratar delimitadores s
Reglas de análisis numérico
Utilice el análisis basado en la configuración regional para los números
Vista previa:
Abrir
Actividades a entre... Actividad 1
Buscar en Actividad 1
Organizar Nueva carpeta
02.Maestría
01.Estadística
02.Almacene
02-Almacene
03. Inteligenc
04. Fundame
05. Fundame
06. Visualizac
07. Datathon
08. Modelaci
Actividades
Actividad
Nombre
Fecha d
Nacimientos_2021
06/05/2
No se puede obtener una vista previa de este archivo.
Nombre: Nacimientos_2021
Archivos de Texto (*.csv, *.tsv, *
Abrir
Cancelar
C:\Users\Soni Vaio\Documents\02.Maestría\08. Modelación de sistemas de la inteligencia de negocios\Actividades a entregar\Actividad 1\Nacimientos_2021.csv
Locale: español (México)
Cambiar separador decimal:
Cambiar separador de agrupación:
Para cambiar el tipo de datos de una columna, haga clic con el botón derecho en la columna seleccionada y elija el tipo apropiado.
Nota: la importación de cadenas de caracteres como datos numéricos puede resultar en la pérdida de datos.
OK
Cancelar
Enviar

En el caso de que se requiera abrir un archivo mismo de Stata, con la extensión .**dta**, en lugar de seleccionar la opción *Importar* del menú Archivo, se debe elegir la opción *Abrir y* seguir la siente ruta:

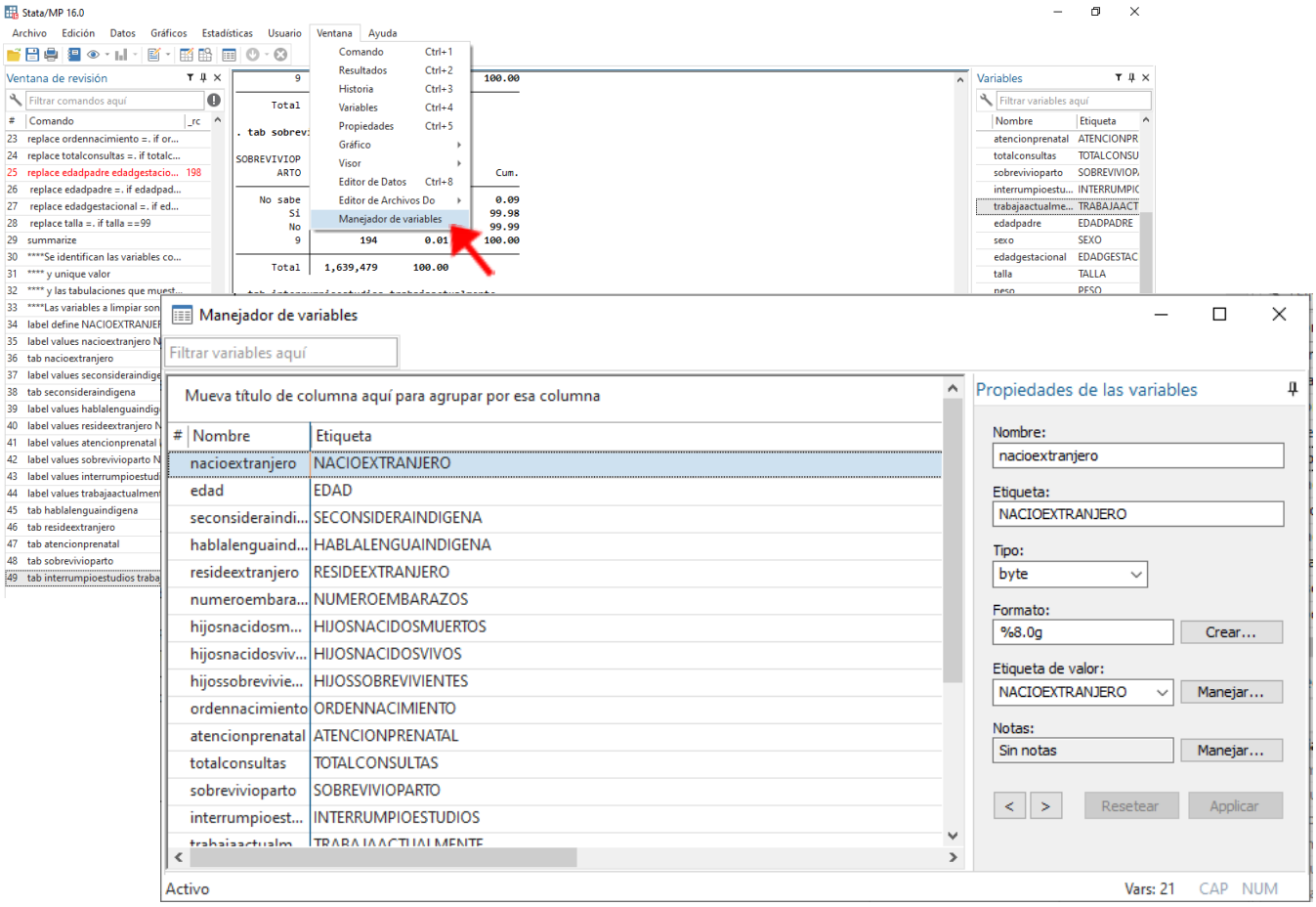
Archivo > Abrir > Ir a la ubicación del archivo en la ventana de explorador > Seleccionar el archivo > Abrir



Limpieza de variables

Se requiere agregar pequeñas descripciones acerca de lo que contiene las variables y en algunos casos agregar etiquetas a los registros que por medio de números expresen algún dato cualitativo, por ejemplo, en el caso de las variables categóricas o las variables *dummys*.

Para hacer esta limpieza nuevamente se apoya de barra de herramientas, la pestaña de *Ventana*, la opción *Manejador de variables*:



En la *Ventana manejador de variables* se encuentran las opciones para editar las variables que se cargan al importar o abrir el conjunto de datos:

Nombre: Aparece el nombre de la variable de acuerdo al conjunto de datos o se puede cambiar desde esta opción.

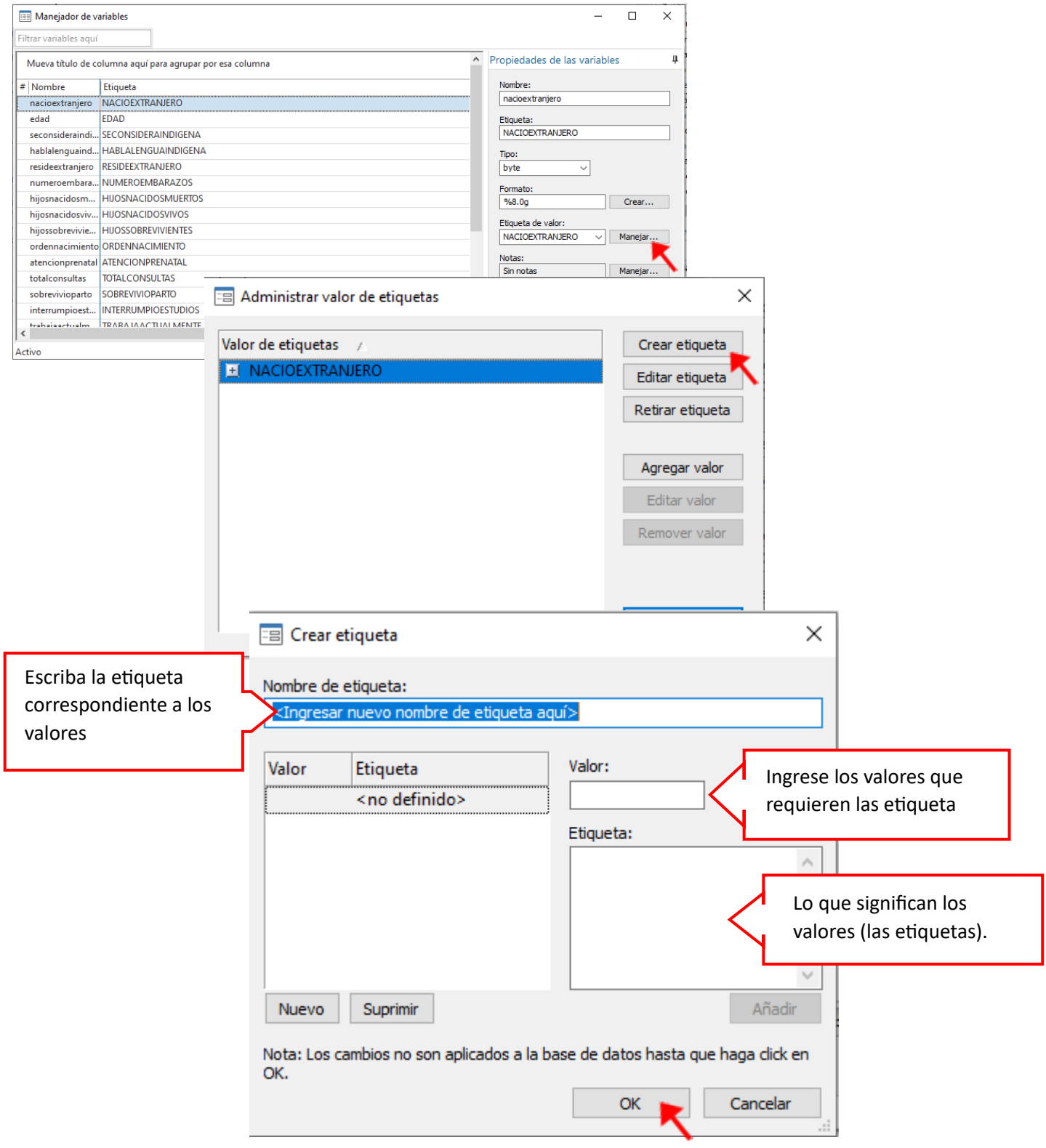
Etiqueta: En una pequeña descripción de lo que quiere decir el nombre de la variable.

Tipo: Tipo de variable.

Formato: Formato de la variable.

Etiqueta de valor: El significado de los valores registrados en algunas variables.

El asignar etiquetas a los valores se debe seguir la siguiente serie de pasos:



Listado de comandos

- **clear:** Comando utilizado para limpiar la memoria de Stata.

- **describe:** Se utiliza para obtener una descripción general del conjunto de datos en forma de tabla detallando: Nombre de la variable, tipo de almacenamiento, el formato o en qué unidades se encuentra la variable, etiquetas de los valores y la descripción de la variable.

```
. describe

Contains data from C:\Users\Soni Vaio\Documents\02.Maestría\08. Modelación de sistemas de la inteligencia de negocios\Sesión 2\pr
> actica 2 Stata Nacimientos 2021.dta
  obs:         1,168
 vars:             9                          1 Jun 2023 21:38
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
EDAD            byte    %10.0g                Edad de la madre
HIJOS           byte    %10.0g                Numero de hijos sin contar al recien nacido
CONSULTAS       byte    %10.0g                Consultas que la madre tomo
GENERO          byte    %10.0g     Sexo       Genero de la mamá
PESO            int     %10.0g                Peso del recien nacido
peso_kg         float   %9.0g
In_peso         float   %9.0g
Z_peso          float   %9.0g
z_peso          float   %9.0g                 Standardized values of ( PESO )
-------------------------------------------------------------------------------
```

- **replace:** Remplaza un valor por otro, el cual es indicado por quien trabaje con el conjunto de datos. Este comando se utiliza para señalar los valores nulos (missing values), que en el programa de Stata se indican por un punto.

  Sintaxis:

  replace *variable* **=. if** *variable* == *valor del conjunto de datos que señale valor nulo*

  Ejemplo:

  replace peso =. **if** peso == 9999

Análisis descriptivo

Las acciones ejecutadas en la etapa del análisis descriptivo buscan sintetizar o convertir datos en información descriptiva para tener un panorama más claro de lo se está estudiando. Para ello, se muestran algunos comandos que arrojan estadísticas representativas de las variables:

- **summarize - sum:** Muestra las estadísticas descriptivas básicas de cada una de las variables que conforma el conjunto de datos. Por medio de una tabla se detalla estadísticas básicas de cada variable, iniciando por la columna *Obs* se señala el total de observaciones, en la columna *Mean* arroja el resultado del promedio calculado para cada variable, continúa la columna *Std. Dev.* que es la desviación estándar y finalmente se encuentran las columnas de valores mínimos *(Min)* y máximos *(Max)* de registro de cada variable.

- También se puede especificar qué variable se requiere conocer sus estadísticos descriptivos básicos agregándolo a comando, por ejemplo, **sum** EDAD.

- **mdesc**: Contabiliza por variable los valores nulos y el porcentaje que representan.

Variable	Obs	Mean	Std. Dev.	Min	Max
EDAD	1,168	25.76199	6.504091	13	52
HIJOS	1,168	2.100171	1.140243	0	9
CONSULTAS	1,168	6.287671	2.795722	0	27
GENERO	1,168	.4708904	.4993657	0	1
PESO	1,107	3110.298	523.6411	545	4440
peso_kg	1,107	3.110298	.5236411	.545	4.44
In_peso	1,107	3110.298	523.6411	545	4440
Z_peso	1,107	5.57e-10	1	-4.898962	2.539338
z_peso	1,107	5.57e-10	1	-4.898962	2.539338

Variable	Missing	Total	Percent Missing
EDAD	0	1,168	0.00
HIJOS	0	1,168	0.00
CONSULTAS	0	1,168	0.00
GENERO	0	1,168	0.00
PESO	61	1,168	5.22
peso_kg	61	1,168	5.22
In_peso	61	1,168	5.22
Z_peso	61	1,168	5.22
z_peso	61	1,168	5.22

- **tabulate - tab**: Muestra en forma de tabla cada uno de los valores que se encuentran registrados en una variable, junto con su frecuencia, porcentaje y porcentaje acumulado. En la imagen de ejemplo el comando es **tab** EDAD.

```
. tabulate EDAD
```

Edad de la madre	Freq.	Percent	Cum.
13	2	0.17	0.17
14	5	0.43	0.60
15	11	0.94	1.54
16	28	2.40	3.94
17	41	3.51	7.45
18	50	4.28	11.73
19	80	6.85	18.58
20	76	6.51	25.09

- **codebook**: El resultado de este comando es una descripción detallada de los estadísticos descriptivos de cada variable, agrupando la información antes mencionada y se agrega los respectivos percentiles.

```
. codebook
```

```
EDAD                                                    Edad de la madre

                  type:  numeric (byte)

                 range:  [13,52]                  units:  1
         unique values:  37                   missing .:  0/1,168

                  mean:   25.762
             std. dev:   6.50409

           percentiles:        10%       25%       50%       75%       90%
                                18        20        25        30        35
```

Análisis de valores extremos

Esta acción detecta entre que valores se encuentra el grueso de la distribución de los datos y lo que se encuentre fuera de este rango se decide que es mejor para el análisis retirarlos o conservarlos.

En Stata, el comando **extremes** permite detectar valores atípicos basándose en diversas métricas. La sintaxis del comando para identificar estos valores por medio del rango intercuartílico (IQR) con un umbral de 3 es:

extremes *nombre de la variable* **, iqr(3)** Ejemplo: **extremes** *Orden_nac*, **iqr(3)**

```
. extremes Orden_nac, iqr(3)
```

obs:	iqr:	Orden_~c
727.	3.000	9
3345.	3.000	9
3788.	3.000	9
6224.	3.000	9
7373.	3.000	9
7522.	3.000	9
8323.	3.000	9
10665.	3.000	9
11491.	3.000	9
11722.	3.000	9
11727.	3.000	9
12776.	3.000	9
14207.	3.000	9
14565.	3.000	9
15201.	3.000	9
15296.	3.000	9
15763.	3.000	9
18573.	3.000	9
20516.	3.000	9
20897.	3.000	9

—more—

Con un valor de iqr de 3, el análisis arroja que 9 es un valor extremo y no afectaría que puedan retirarse.

Generar quintiles

Se le conoce como quintil a la fracción de una distribución cuando esta es dividida en 5 partes iguales (Q1, Q2, Q3, Q4 y Q5). El comando para hacer esta división es **xtile** con la sintaxis:

xtile *nombre de la variable nueva a generar = nombre de la variable que se dividirá* **, nquantiles (5)**

Ejemplo: **xtile** *q5_Peso = Peso*, **nquantiles(5)**

Lo que genera una variable extra con el nombre de ***q5_Peso***

```
. xtile q5 Peso=Peso, nquantiles(5)
```

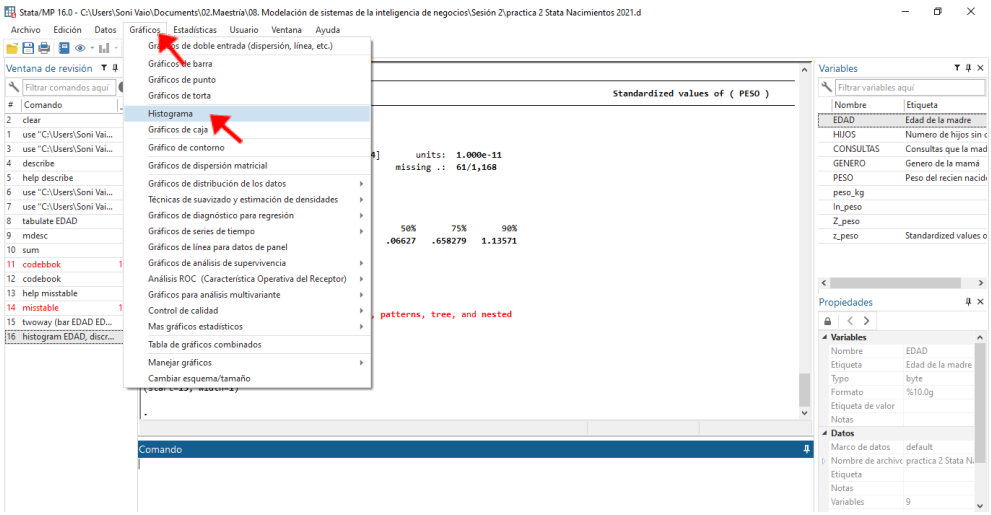

```
. tab q5_Peso
```

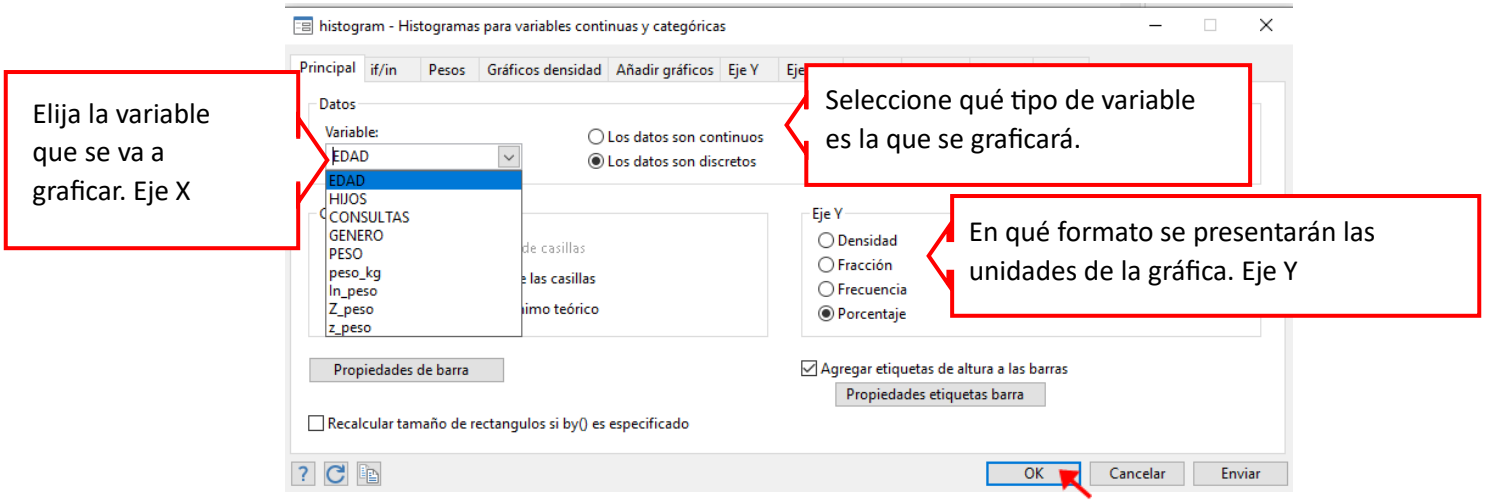
5 quantiles of Peso	Freq.	Percent	Cum.
1	310,475	20.05	20.05
2	316,025	20.41	40.46
3	302,869	19.56	60.02
4	334,533	21.60	81.62
5	284,601	18.38	100.00
Total	1,548,503	100.00	

Gráficos exploratorios

En la barra de herramientas se ubica el botón de *Gráficos*, donde se encuentran las diferentes opciones para realizar visualizaciones de los datos seleccionados. También se pueden hacer por medio de comandos, pero en estos primeros acercamientos al programa se recomienda usar el asistente del botón mencionado.

- **Histograma**

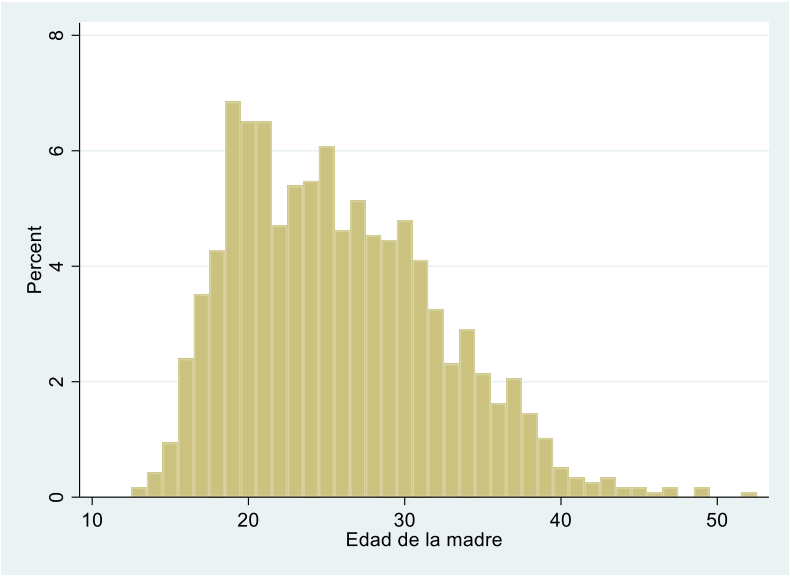




Elija la variable que se va a graficar. Eje X

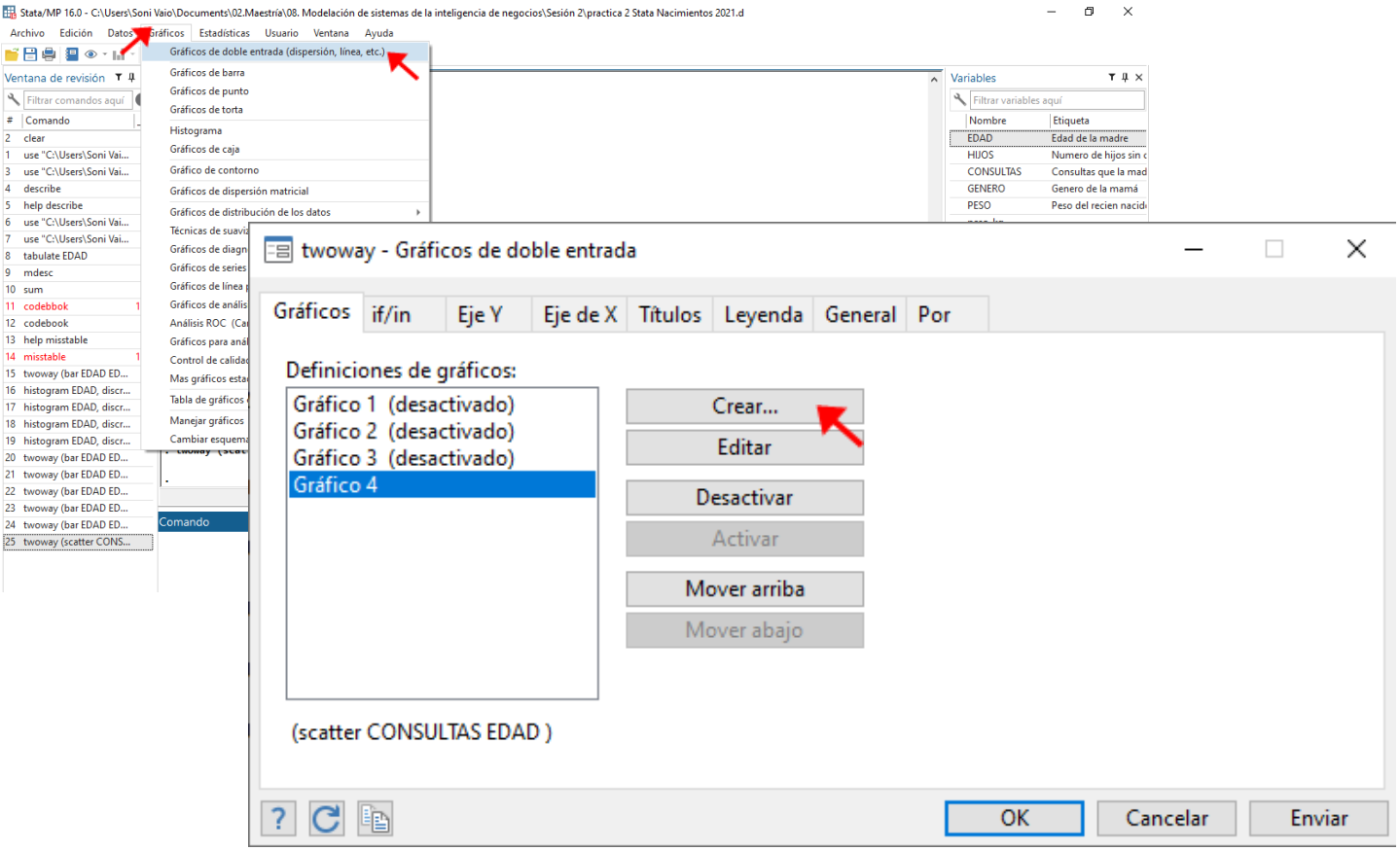
Seleccione qué tipo de variable es la que se graficará.

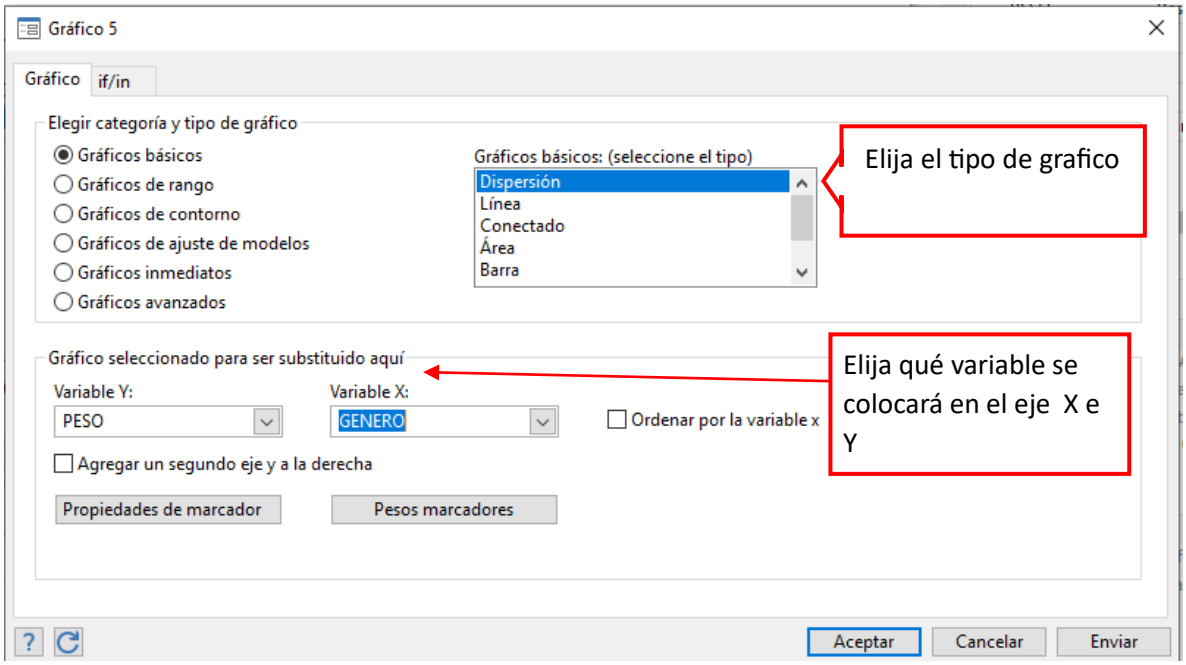
En qué formato se presentarán las unidades de la gráfica. Eje Y



- **Gráfico de doble entrada**

  Con este tipo de gráfica se podrá contrastar dos variables en una sola visualización.





Análisis de Correlación de Pearson

El análisis de correlación de Pearson indica qué tan fuerte o débil es la relación lineal entre variables, así como también si es positiva o negativa. Es importante destacar que este indicador sólo es posible calcularlo entre variables exclusivamente numéricas.

El comando para conocer los coeficientes de correlación es **corr** y para saber si estos coeficientes son significativamente estadísticos se utiliza **pwcorr** … **,sig**

Cálculo de los coeficientes de correlación	Cálculo de los coeficientes de correlación con significancia estadística
Sintaxis: corr *listado de variables numéricas a correlacionar…*	Sintaxis: pwcorr *listado de variables numéricas a correlacionar…*,sig
Ejemplo: corr Talla Peso	Ejemplo: pwcorr Edad Total_Consultas , sig

```
. corr Talla Peso
(obs=1,548,365)
```

	Talla	Peso
Talla	1.0000	
Peso	0.6613	1.0000

La tabla muestra los coeficientes de correlación de Pearson, los cuales al ser más cercanos a 1 es fuerte y cercanos a 0 es débil.

```
. pwcorr Edad Total_Consultas, sig
```

	Edad	Total_~s
Edad	1.0000	
Total_Cons~s	0.1673	1.0000
	0.0000	

El *P value* es el indicador de significancia estadística, el cual al ser menor o igual a .05 se demuestra que la presencia de significancia estadística y da pauta rechazar o no la hipótesis de correlación.

PLANTEAMIENTO DE HIPÓTESIS PARA CORRELACIÓN

H_0 = NO hay correlación entre las variables

vs

H_1 = Sí hay correlación entre las variables

Análisis de Correlación de Spearman

El coeficiente de correlación de rango de Spearman es una herramienta estadística utilizada para evaluar la hipótesis de que no existe una asociación significativa y la identificación de presencia de una relación lineal entre dos variables categóricas.

En este trabajo se usará el estadístico para la buscar la relación lineal entre variables *dummy*, donde la respuesta se registra en 0 y 1 refiriéndose a una respuesta cualitativa, como si y no u hombre y mujer. Según el diccionario de datos del dataset *nacimientos_2021*. El comando para este análisis es **spearman**, obteniendo el coeficiente de correlación de spearman y el P value, el cual indica si el análisis cuenta con significancia estadística. La sintaxis es:

spearman *variable a analizar variable a analizar…*

Ejemplo:

spearman Se_cond_Indigena Trabaja_Actual

```
. spearman Se_cond_Indigena Trabaja_Actual

    Number of obs =   1533417
Spearman's rho =     -0.1135

Test of Ho: Se_cond_Indigena and Trabaja_Actual are independent
    Prob > |t| =      0.0000
```

Número de observaciones tomadas para hacer la correlación

Resultado del cálculo de la correlación obteniendo el coeficiente de Spearman

P value resultado de la correlación entre las variables seleccionadas. Este valor indica si el cruce tiene significancia estadística.

En el caso de que se requiera agregar más de una variable al análisis, se requerirá hacer una matriz de correlación de Spearman, donde al comando sólo debe agregarse las variables extras. Para la obtención de los respectivos Pvalues se debe agregar **stats(p)** al final del listado de variables, creando el siguiente comando.

spearman *listado de variables categóricas a correlacionar*, **stats(p)**

Ejemplo:

spearman Se_cond_Indigena Trabaja_Actual Genero_nac

```
. spearman Se_cond_Indigena Trabaja_Actual Genero_nac
(obs=1533403)

             | Se_con~a Trabaj~l Genero~c
-------------+---------------------------
Se_cond_In~a |   1.0000
Trabaja_Ac~l |  -0.1135   1.0000
  Genero_nac |  -0.0007  -0.0003   1.0000
```

Al correr el comando sin la instrucción **stats(p)** arroja sólo los coeficientes de correlación de Spearman.

Cálculo de p-value de la correlación de Spearman

Ejemplo:

spearman *Se_cond_Indigena Trabaja_Actual Gen_nac*, **stats(p)**

```
. spearman Se_cond_Indigena Trabaja_Actual Genero_nac , stats(p)
(obs=1533403)
```

Key
Sig. level

	Se_con~a	Trabaj~l	Genero~c
Se_cond_In~a			
Trabaja_Ac~l	0.0000		
Genero_nac	0.3709	0.7247	

En la matriz se muestra únicamente los P value de la correlación de Spearman

Test de normalidad de los datos – Test de Shapiro Wilk

Las pruebas de bondad de ajuste se emplean para evaluar si los datos de una muestra pueden ser considerados como provenientes de una distribución normal.

El test se ejecuta con el comando **swilk** *variable o listado de variables*.

Ejemplo:

swilk *Edad*

El título de la tala muestra el planteamiento de hipótesis

```
. swilk Edad

                    Shapiro-Wilk W test for normal data

    Variable |        Obs        W           V          z       Prob>z
-------------+-------------------------------------------------------
        Edad |  1,639,411    0.98146   2423.650     22.112     0.00000

Note: The normal approximation to the sampling distribution of W'
      is valid for 4<=n<=2000.
```

Pvale como indicador para la prueba de hipótesis de normalidad.

PLANTEAMIENTO DE HIPÓTESIS PARA EL TEST DE NORMALIDAD

H_0 = Los datos son normales

vs

H_1 = Los datos NO son normales

Análisis de contingencia

Otras formas que se le conoce a este procedimiento en los datos son tablas de contingencia, tablas de probabilidad o tablas cruzadas, y es comúnmente utilizado para examinar la relación entre dos o más variables. A través de pruebas estadísticas como el *chi-cuadrado*, este análisis ayuda a determinar si existe una asociación significativa entre las variables.

Con el fin de ilustrar de mejor manera su aplicación se hace uso de una nueva base de datos abiertos y acceso libre denominada *sleep75*, la cual cuenta con información sobre los patrones de sueño de diferentes personas, conformada por 34 variables cuantitativas y 706 registros.

Nombre de la variable	Descripción
age	in years
black	=1 if black
case	identifier
clerical	=1 if clerical worker
construc	=1 if construction worker
educ	years of schooling
earns74	total earnings, 1974
gdhlth	=1 if in good or excel. health
inlf	=1 if in labor force
leis1	sleep - totwrk
leis2	slpnaps - totwrk
leis3	rlxall - totwrk
smsa	=1 if live in smsa
lhrwage	log hourly wage
lothinc	log othinc, unless othinc < 0
male	=1 if male
marr	=1 if married
prot	=1 if Protestant
rlxall	slpnaps + personal activs
selfe	=1 if self employed
sleep	mins sleep at night, per wk
slpnaps	minutes sleep, inc. naps
south	=1 if live in south
spsepay	spousal wage income
spwrk75	=1 if spouse works
totwrk	mins worked per week
union	=1 if belong to union
worknrm	mins work main job
workscnd	mins work second job
exper	age - educ - 6

yngkid	=1 if children < 3 present
yrsmarr	years married
hrwage	hourly wage

La indicación de tabla cruzada entre dos variables se hace con el comando **tab,** agregando instrucciones para arrojar porcentajes. El comando es:

tab *variable a cruzar variable a cruzar* **, nofreq cell**

Comando tabla.

Comando para indicar que NO arroje frecuencias, si no porcentajes.

Comando que solicita la columna de porcentajes totales.

Ejemplo:

tab *sleep5q age5q*, **nofreq cell**

(Como resultado se obtiene a lo que en estadística se denomina tabla de probabilidad)

```
. tab sleep5q age5q, nofreq cell
```

Edad joven — Edad vieja

5 quantiles of sleep	5 quantiles of age 1	2	3	4	5	Total
1 (Duerme)	5.38	4.96	3.26	3.54	2.97	20.11
2 (Duerme)	3.40	4.82	4.53	4.11	3.26	20.11
3	3.97	3.26	4.11	4.67	3.97	19.97
4	4.82	2.97	3.26	4.25	4.67	19.97
5	5.38	3.26	3.26	2.97	4.96	19.83
Total	22.95	19.26	18.41	19.55	19.83	100.00

Pruebas de independencia

Las pruebas de independencia funcionan para corroborar si dos eventos son independientes, a través de contrastar las frecuencias observadas con las frecuencias esperadas de acuerdo con la hipótesis nula. Para ello se aplica la prueba de **Chi2.** El comando para aplicar esta prueba se agrega a la instrucción de un análisis de contingencia:

tab *variable a cruzar variable a cruzar* **, nofreq cell chi2**

Ejemplo:

```
. tab sleep5q age5q, nofreq cell chi2
```

5 quantiles of sleep	5 quantiles of age 1	2	3	4	5	Total
1	5.38	4.96	3.26	3.54	2.97	20.11
2	3.40	4.82	4.53	4.11	3.26	20.11
3	3.97	3.26	4.11	4.67	3.97	19.97
4	4.82	2.97	3.26	4.25	4.67	19.97
5	5.38	3.26	3.26	2.97	4.96	19.83
Total	22.95	19.26	18.41	19.55	19.83	100.00

Pearson chi2(16) = 22.6186 Pr = 0.124 → P value de la prueba de Chi_2

PLANTEAMIENTO DE HIPÓTESIS PARA LA PRUEBA DE CHI_2

H_0 = Los datos son independientes

vs

H_1 = Los datos NO son independientes

Para una mayor referencia del contenido de este capítulo ver:

Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Se utiliza para predecir el valor de una variable según el valor de otra. La variable que desea predecir se denomina variable dependiente. La variable que está utilizando para predecir el valor de la otra variable se denomina variable independiente. En Stata para realizar el análisis de regresión lineal simple se hace por el comando **reg** con la siguiente sintaxis:

reg *variable dependiente variable independiente*

Retomando la base de datos *sleep75*, donde se analiza los ciclos de sueño de un grupo de personas, a modo de ejemplo se selecciona como variable dependiente las horas de sueño que duermen a la semana los sujetos de estudio (sleep), y como variable independiente la edad de cada uno de ellos y ellas (age). Para realizar un análisis de regresión con las mencionadas variables el comando a ejecutar es:

reg *sleep age*

```
. reg sleep age

      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(1, 704)     =    5.80
       Model |  1137207.85     1  1137207.85            Prob > F      =  0.0163
    Residual |  138102628    704  196168.506           R-squared     =  0.0082
-------------+------------------------------           Adj R-squared =  0.0068
       Total |  139239836    705  197503.313           Root MSE      =  442.91

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   3.540881   1.470639     2.41   0.016     .6535177    6.428244
       _cons |   3128.913   59.46811    52.61   0.000     3012.157    3245.669
------------------------------------------------------------------------------
```

R-cuadrada → Bondad de ajuste

Constante de B0 o intercepto

Coeficiente de la variable age (independiente)

P value de B0

P value de B1

PLANTEAMIENTO DE HIPÓTESIS PARA REGRESIÓN LINEAL

B1 variable independiente (beta uno)	B0 variable dependiente (beta cero)
$H_0 = B1 = 0$ vs $H_1 = B1 \neq 0$	$H_0 = B0 = 0$ vs $H_1 = B0 \neq 0$
Planteamiento de B1 Se busca que B1 sea diferente a 0, para que si explique a BETA CERO (FIJANDOSE EN EL COEFICIENTE DE LA VARIABLE INDEPENDIENTE). Si un parámetro es 0 no sería estadísticamente significativo y no valdría la pena hacer un estudio de estos datos.	Planteamiento de B0 Se busca que B0 sea diferente a 0 Si BETA CERO ES IGUAL A CERO (FÍJANDOSE EN LA CONSTANTE) quiere decir que B0 (variable dependiente) no es válida para explicar a B1 (variable independiente). No es constante.

Estimación de Valores Ajustados y Residuales

Valores ajustados - *Fitted values*

Es la estimación que se realiza previo a un análisis de regresión. Con el comando **predict** *variable dependiente* **_e** se genera una nueva variable donde se encuentra una estimación para cada observación del conjunto de datos tomando en cuenta la variable independiente.

Ejemplo:

predict sleep_**e**

```
. predict sleep_e
(option xb assumed; fitted values)
```

sleep_e	Fitted values

Valores residuales

Los valores residuales son la diferencia entre los valores observados y el estimado. El comando **predict u** *este valor puede cambiarse por cualquier cosa*, **resid**, con lo se genera una nueva variable con los valores residuales.

Ejemplo:

predict *error*, **resid**

error	Residuals

Análisis con Logaritmos (Modelos de Elasticidad Constante)

Este tipo de análisis ayuda a entender a los fenómenos de estudio desde la unificación de las unidades. Algunos de los beneficios de este procedimiento es que se adquiere el valor de R2 y a veces se gana significancia estadística, así como también se facilita la interpretación, ya que las unidades de logaritmo es porcentaje.

En términos de elasticidades el nivel en el que se pueden empatar y analizar las variables es:

Level – level (nivel - nivel),

In (logaritmo) – level (nivel),

Level - In (logaritmo)

In (logaritmo) - In (logaritmo).

En Stata para generar logaritmos de las variables es a través del siguiente comando:

gen *nombre de la variable nueva con logaritmo* **= in(***variable a calcular logaritmo***)**

Ejemplo de un análisis logaritmo - level:

gen log_sleep **= in(**sleep**)**

Con la ejecución de esta instrucción se genera la variable con el logaritmo de la variable *sleep*. Para obtener un resultado, el siguiente paso sería realizar el análisis de regresión lineal

reg *log_sleep* (variable con logaritmo) *age* (variable normal)

```
. reg log_sleep age
```

Source	SS	df	MS	Number of obs	=	706
				F(1, 704)	=	6.33
Model	.14286797	1	.14286797	Prob > F	=	0.0121
Residual	15.8859268	704	.022565237	R-squared	=	0.0089
				Adj R-squared	=	0.0075
Total	16.0287948	705	.022735879	Root MSE	=	.15022

log_sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.001255	.0004988	2.52	0.012	.0002758	.0022343
_cons	8.032312	.0201692	398.25	0.000	7.992713	8.071911

Variable dependiente

R^2

Bondad de ajuste

Coeficiente de la variable age (independiente)

P value del análisis de regresión en la variable age

Resultado:

- Se observa un mejoramiento de la R^2, porque cuando se calculan logaritmos aumenta la variabilidad, lo que causa un efecto positivo en los indicadores de R^2 y la significancia estadística.
- La variable dependiente está en logaritmos (que son aproximaciones porcentuales).

Con estos puntos clave a observar el en resultado obtenido se puede concluir que si la edad aumenta en un año más (porque está en nivel, es decir conserva sus unidades), los minutos de sueño que duerme una persona aumentan en .1255%.

Ejemplo de un <u>level – logaritmo (efectos en niveles cambios porcentuales)</u>:

gen *log_age* **= in(***age***)**

reg sleep log_age

```
. reg sleep log_age
```

Source	SS	df	MS		
Model	891303.042	1	891303.042	Number of obs	= 706
Residual	138348533	704	196517.802	F(1, 704)	= 4.54
				Prob > F	= 0.0335
Total	139239836	705	197503.313	R-squared	= 0.0064
				Adj R-squared	= 0.0050
				Root MSE	= 443.3

R^2 Bondad de ajuste

Variable dependiente

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
log_age	122.9174	57.71672	2.13	0.034	9.599897	236.2349
_cons	2821.777	209.4207	13.47	0.000	2410.613	3232.941

Coeficiente de la variable log_age (independiente)

P value de la regresión en logaritmo age

Resultado:

El cambio se hace en la interpretación en porcentaje de la variable independiente que es la edad y la variable dependiente que ahora se encuentra en unidades de la misma, que para el ejemplo sería minutos. Por ello es válido inferir que si la persona aumenta en 1% de edad, se estima que en promedio aumenta 1.229174 los minutos de sueño.

Análisis de Regresión Múltiple

La decisión de subir de nivel el análisis de regresión lineal a regresión múltiple es para fortalecer la explicación de la variable dependiente. A medida que se incluyan más variables, la bondad de ajuste aumenta y el error se reduce, lo que ayuda a robustecer el análisis.

El comando es igual que para el análisis de regresión lineal, **reg** al que se le deberá agregar el listado de variables independientes con las cuales se quiera explicar el fenómeno a estudiar. Para el ejemplo de este análisis se retoma el *data set nacimientos_2021*, con el registro de nacimientos ocurridos en México y bajo el planteamiento de la pregunta ¿cuáles son los factores que influyen en el peso de un recién nacido?

Ejemplo de planteamiento del modelo:

reg *Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Hijos_nac_vivos Orden_nac Atencion_Pren Total_Consultas*

La variable *Peso* es la dependiente y después se enlistan una serie de variables explicativas (variables independientes) que se buscan demostrar que tanto influyen en el peso de un recién nacido.

Source	SS	df	MS		
Model	1.1835e+11	12	9.8626e+09	Prob > F	0.0000
Residual	1.0938e+11	1,062,578	102940.126	R-squared	0.5197
Total	2.2773e+11	1,062,590	214319.042	Adj R-squared	0.5197
				Root MSE	320.84

	Number of obs	=	1,062,591
	F(12, 1062578)	=	95809.22
	Prob > F	=	0.0000
	R-squared	=	0.5197
	Adj R-squared	=	0.5197
	Root MSE	=	320.84

R-cuadrada → Bondad de ajuste

Peso	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Talla	78.27424	.1356128	577.19	0.000	78.00844	78.54004
Edad	1.453334	.0751132	19.35	0.000	1.306115	1.600554
Edad_gest	96.49551	.2231838	432.36	0.000	96.05808	96.93294
Genero_nac	32.02111	.6242899	51.29	0.000	30.79753	33.2447
Edad_padre	.9454787	.0559297	16.90	0.000	.8358584	1.055099
Trabaja_Actual	17.62199	.8345975	21.11	0.000	15.98621	19.25777
Se_cond_Indigena	-30.34506	1.136263	-26.71	0.000	-32.5721	-28.11802
No_embarazos	116.0391	2.346669	49.45	0.000	111.4397	120.6385
Hijos_nac_vivos	7.205051	.7141546	10.09	0.000	5.805332	8.60477
Orden_nac	-107.198	2.370981	-45.21	0.000	-111.845	-102.5509
Atencion_Pren	18.75484	2.280198	8.23	0.000	14.28573	23.22396
Total_Consultas	3.415552	.1076855	31.72	0.000	3.204492	3.626612
_cons	-4655.141	8.051556	-578.17	0.000	-4670.922	-4639.36

Constante o intercepto

Coeficiente de la variable independiente

P value de las variables independientes

Conclusión:

Es importante señalar que, en el ejemplo el único valor negativo del coeficiente de la variable independiente es resultado de una variable *dummy* si la madre se considera indígena o no. En lo que respecta a la interpretación, se toma en cuenta el valor equivalente a 1, significando que la madre se autopercibe como indígena, que de acuerdo a lo resultado y concretándose en que el recién nacido de una madre con esta característica su bebé pesa -30.345 gr.

Con el ejemplo desarrollado se busca demostrar como el análisis de regresión múltiple proporciona una comprensión detallada de cómo diferentes factores influyen en un resultado específico, como el peso al nacer de un bebé en relación con la identidad étnica de la madre. El hallazgo de si una madre se autopercibe como indígena, el peso de su bebé tiende a ser aproximadamente 30.345 gramos menor en comparación con el de los bebés de madres que no se identifican como indígenas, puede funcionar como fundamento para la elaboración de políticas públicas al proporcionar evidencia empírica de diferencias en la salud neonatal asociadas con la identidad étnica.

R de Shapley

Existen diferentes métodos a aplicar sobre el modelo para rectificar que no haya incumplimiento de alguno de los supuestos de validación del modelo de regresión (linealidad, normalidad de errores, independencia de errores y homocedasticidad), así como también algo que nos demuestre que tan adecuadamente esta ajustado el modelo y cuál es la aportación de cada variable para explicar el fenómeno. El que se explica en este compendio es la R de Shapley, que es la descomposición de Shapley y Owen del R-cuadrada.

Esté análisis, se ejecuta con el comando **rego**, el cual utiliza los resultados de la regresión y descompone la parte de la varianza explicada (medida por R-cuadrada) en contribuciones por cada variable independiente o grupos de variables independientes. Dicho de otra manera, con el resultado se podrá observar porque la R-cuadrada tiene el valor que arroja.

La sintaxis del comando es similar al de regresión lineal, sólo se sustituye por el comando mencionado:

rego *variable dependiente grupo de variables independientes*

Ejemplo:

rego *Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas*

```
. rego Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas
```

Gr	Regressor	Coef.		Std.Err.	P>|t|	Std.Coef.	Shapley %R2
1	Talla	78.58779	***	.1355483	0.000	0.4642	56.0405
2	Edad	1.461493	***	.0749641	0.000	0.0199	0.1330
3	Edad_gest	97.36831	***	.2225177	0.000	0.3512	42.6058
4	Genero_nac	31.88848	***	.6248332	0.000	0.0344	0.4643
5	Edad_padre	.9508109	***	.0559551	0.000	0.0156	0.1163
6	Trabaja_Actual	17.04587	***	.8321216	0.000	0.0148	0.0351
7	Se_cond_Indigena	-29.82967	***	1.135853	0.000	-0.0179	0.0578
8	No_embarazos	14.29078	***	.3048432	0.000	0.0377	0.3206
9	Atencion_Pren	19.05987	***	2.282138	0.000	0.0060	0.0489
10	Total_Consultas	3.173693	***	.1069946	0.000	0.0222	0.1777
-	Intercept	-4702.122		7.98929	0.000		
	Observations	1062778					
	Overall R2	0.51884					
	Root MSE	321.1539					
	F-stat. Model	114597.9	***		0.000		
	Log Likelihood	-7642281					

Porcentaje de aportaciones de cada variable a la varianza explicada

Resultado:

Con lo demostrado por el resultado de la descomposición de R^2 la talla y la edad gestacional de un recién nacido son las variables con mayor aportación de variabilidad en la variable dependiente, que es Peso, lo que significa que estas dos variables contribuyen más al ajuste del modelo. Por el lado contrario la variable Trabaja actualmente es la que menor aportación al modelo, lo que significa que podría prescindirse y no habría gran afectación al resultado.

Estandarización de los coeficientes de beta

Al estandarizar los coeficientes resultantes, con el análisis de regresión de las variables independientes, funciona para conocer el efecto de cada variable sobre la variable dependiente, la cual representa al fenómeno de estudio. Lo que es importante detectar es la relación, donde a mayor número de beta mayor efecto sobre la variable dependiente. Para realizar la estandarización al modelo de regresión, al final del grupo de variables independientes se debe agregar la instrucción de beta (**, b**).

Retomando el ejemplo anterior:

reg *Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas*, **b**

```
. reg Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas,
> b
```

Source	SS	df	MS		
Model	1.1820e+11	10	1.1820e+10	Number of obs	= 1,062,778
Residual	1.0961e+11	1,062,767	103139.843	F(10, 1062767)	> 99999.00
				Prob > F	= 0.0000
				R-squared	= 0.5188
				Adj R-squared	= 0.5188
Total	2.2781e+11	1,062,777	214353.294	Root MSE	= 321.15

Peso	Coef.	Std. Err.	t	P>|t|	Beta
Talla	78.58779	.1355483	579.78	0.000	.4642186
Edad	1.461493	.0749641	19.50	0.000	.0198632
Edad_gest	97.36831	.2225177	437.58	0.000	.3511973
Genero_nac	31.88848	.6248332	51.04	0.000	.0344225
Edad_padre	.9508109	.0559551	16.99	0.000	.0156492
Trabaja_Actual	17.04587	.8321216	20.48	0.000	.0148383
Se_cond_Indigena	-29.82967	1.135853	-26.26	0.000	-.0178848
No_embarazos	14.29078	.3048432	46.88	0.000	.0376917
Atencion_Pren	19.05987	2.282138	8.35	0.000	.0059858
Total_Consultas	3.173693	.1069946	29.66	0.000	.0222067
_cons	-4702.122	7.98929	-588.55	0.000	.

Efectos estandarizados

Resultado:

De acuerdo a la estandarización se demuestra que la variable *Talla,* que tiene la mayor influencia relativa sobre el peso de un recién nacido en comparación con el resto de las variables del modelo, refleja su importancia significativa en la predicción del peso.

Variance inflation factor – Prueba de multicolinealidad o colinealidad

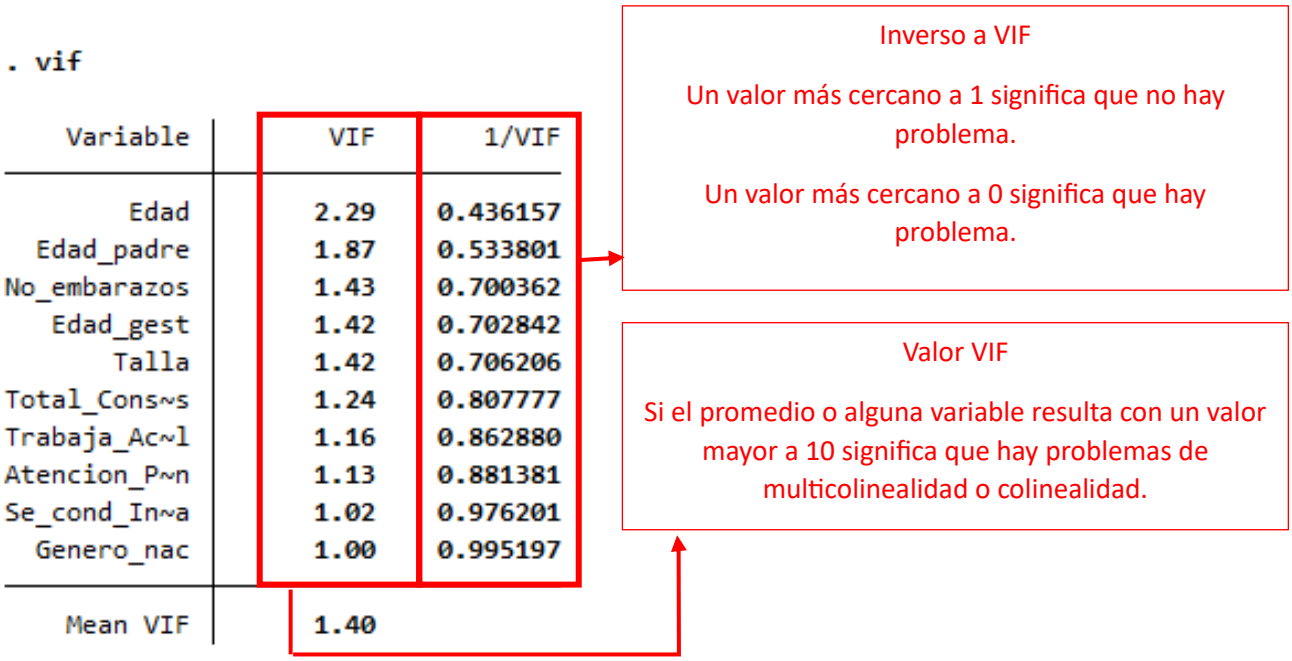
El comando para ejecutar este análisis se debe realizar después de correr el modelo (comandos *postestimation*) y se hace por medio de la instrucción **vif**. Si en el resultado de la evaluación en cada variable es mayor a 10 estará indicando que existe un problema. El cual puede solucionarse omitiendo las variables que nos señale esta prueba y volver a ejecutar el comando.

Ejemplo:

Se ejecuta el modelo que se ha estado trabajando en las demostraciones anteriores.

reg *Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas*

Después se ejecuta el comando **vif**



Resultado:

Los VIFs obtenidos están altamente por debajo de 5, lo que indica que no hay problemas significativos de multicolinealidad en el modelo. Los valores de VIF varían entre 1.00 y 1.42, que están dentro de un rango saludable. El promedio de VIF es de 1.20 confirma que, en general, las variables independientes en el modelo no presentan un problema de multicolinealidad significativa, lo que asegura que es posible interpretar los coeficientes de las variables independientes con un grado razonable de confianza.

Prueba de heterocedasticidad y variables omitidas

La heterocedasticidad se presenta cuando los errores no son constantes a lo largo de toda la muestra y para conocer si existe su presencia se hace una prueba con el comando **hettest**, que se ejecuta después de correr el modelo de análisis.

Ejemplo:

reg *Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas*

Comando de prueba: **hettest**

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Peso

         chi2(1)      =  6088.32
         Prob > chi2  =   0.0000
```

P value indicador de la prueba de heterocedasticidad

Menor a .05 se rechaza la hipótesis nula

Mayor a .05 NO se rechaza la hipótesis nula.

Resultado

La hipótesis nula de esta prueba dice que la varianza es constante y con un resultado de 0.0000 se rechaza esta premisa, evidenciando que sí hay presencia de heterocedasticidad, significando que los errores se mueven a medida que X (la variable independiente) o el grupo de variables explicativas se mueven.

Con el fin de ajustar el modelo es posible caer en el error de omisión de variables importantes, lo cual puede traer problemas de sesgo en los coeficientes estimados, reducción de la precisión predictiva del modelo y conclusiones incorrectas.

El comando **ovtest** también se ejecuta después de correr el modelo y su interpretación se realiza por medio del P value, con el cual se determina la presencia de este tipo de error.

Ejemplo:

reg *Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas*

Comando de prueba: **ovtest**

```
. ovtest

Ramsey RESET test using powers of the fitted values of Peso
       Ho:  model has no omitted variables
            F(3, 1062764) =    3633.27
                  Prob > F =      0.0000
```

P value indicador de la prueba de variable omitida

Menor a .05 se rechaza la hipótesis nula

Mayor a .05 NO se rechaza la hipótesis nula.

Resultado:

Como lo menciona la misma respuesta de Stata, la hipótesis nula de esta prueba es que el modelo no tiene variables omitidas, por lo que, con un valor de 0.000 demostrando que si hay

significancia estadística, se rechaza H0 demostrando que sí hay omisión de variables importantes para el modelo.

Pruebas de hipótesis lineales

El análisis de regresión lineal no solo proporciona estimaciones de las relaciones entre variables, sino que también hace posible la formulación de preguntas adicionales basadas en los hallazgos. Estas preguntas pueden ser exploradas y respondidas mediante la aplicación de pruebas de hipótesis lineales, que permiten evaluar la validez de supuestos específicos sobre los parámetros del modelo.

La dirección del planteamiento de la hipótesis lineal a corroborar, se define bajo el planteamiento si el coeficiente de una variable independiente puede tomar ciertos valores, o en forma de pregunta, ¿es posible que el coeficiente ocupe un valor dado por quien esté realizando el análisis?

El comando para realizar pruebas de hipótesis lineales es **test**, donde**:**

test *nombre de la variable independiente = valor especificado por quien realiza el análisis*

Ejemplo:

reg Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas

Prueba de hipótesis lineal: **test** Talla = 80

```
. test Talla = 80

 ( 1)  Talla = 80

       F(  1,1062767) =  108.55
            Prob > F =    0.000
```

Resultado de la prueba de hipótesis

Menor a .05 se rechaza la hipótesis nula

Mayor a .05 NO se rechaza la hipótesis nula.

La solicitud ingresada a Stata es que evalúe la hipótesis nula de que el coeficiente de la variable *Talla* en el modelo de regresión es igual a 80, maquetándose de la siguiente manera:

H0: Talla = 80

HA: Talla ≠ 80

Dado el resultado obtenido se rechaza la hipótesis nula con un valor de 0.0000, demostrando que la prueba de hipótesis sugiere que la relación entre *Talla* y *Peso* (controlando por las demás variables incluidas en el modelo) no se describe adecuadamente por un coeficiente de 80, que de acuerdo al contexto de los datos, se señala la influencia de la talla sobre el peso del recién nacido no puede ser representada correctamente por un valor de 80.

Presentación de resultados de regresión

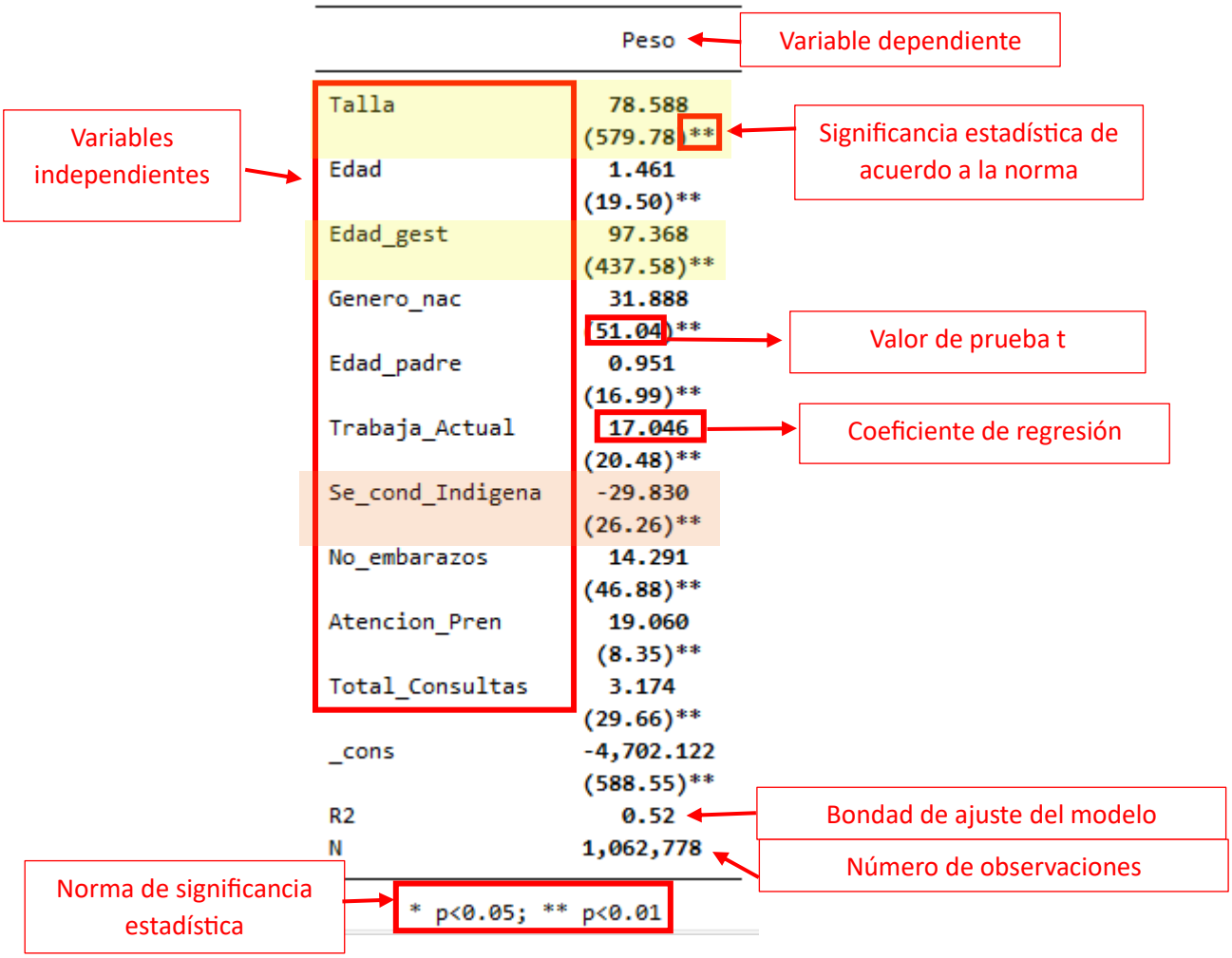
La forma de presentar lo hallado en un análisis de regresión, se acorta a sólo lo más relevante del todo el proceso que conllevó llegar a las conclusiones, por lo que se requiere hacer una síntesis de los comandos realizados en Stata. El programa cuenta con un comando que hace esta acción de concretar el proceso, a modo concentrar la visualización de resultados de los coeficientes de regresión y la significancia estadística, y este es **outreg**.

Ejecutar este comando se debe hacer después de correr el modelo ya adecuadamente equilibrado y así arroja los resultados que conducen a las conclusiones del proceso de análisis.

Ejemplo:

reg Peso Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas

outreg



Conclusión del capítulo de Regresión

El modelo de regresión múltiple planteado busca explicar el peso de los recién nacidos con respecto a las variables independientes señaladas, *Talla Edad Edad_gest Genero_nac Edad_padre Trabaja_Actual Se_cond_Indigena No_embarazos Atencion_Pren Total_Consultas*. En lo que respecta al ajuste del modelo, el valor de R^2 indica que el 52% de la variabilidad en el peso del bebé se demuestra por dichas variables, señalando que se explica un poco más de la mitad del fenómeno, que para cuestiones de este ejercicio de ejemplo se considera suficiente la bondad de ajuste.

En la tabla sintetizada del proceso análisis señala que las variables *edad gestacional* y talla son las que mayormente influyen en la variable peso , dado sus altos coeficientes, 97.368 y 78.588 respectivamente, igualmente se puede ver en el valor de la significancia estadística con un P value menor que .01.

Otra de las variables que causan interés es la que registra si la madre se considera indígena, con un coeficiente que describe un efecto negativo alto de -29.830 y una significancia

estadística de un P value menor a .01, lo que puede sugerir un evento de interés que está perjudicando a este sector de la población. Esto es prueba suficiente para hacer un llamado de solicitud de atención para la intervención de las instituciones relacionadas al tema, ya sea instancias del área de salud pública, dependencias de asuntos indígenas, instituciones de cualquier nivel específicas de la mujer, entre otras.

Conforme se fue desarrollando la ejemplificación del modelo, específicamente en el caso de la base de datos *nacimientos_2021*, el modelo diseñado para el uso de la técnica de regresión lineal múltiple se adecua hasta finalizarlo con la inclusión de 10 variables, que para fines del presente trabajo funciona suficientemente para ilustrar como estos factores influyen en el peso del recién nacido.

Para una mayor referencia del contenido de este capítulo ver:

Cameron, A. C., & Trivedi, P. K. (2013). Regression Analysis of Count Data (Vol. 53). Cambridge University Press.

Lo fundamental de este modelo es calcular probabilidades sobre planteamientos de problemáticas que se expresen por medio de variables categóricas, o también conocidas como variables *dummy*, donde su respuesta sea 0 y 1. De no contar con este registro, pero la variable tiene esta naturaleza se puede hacer la transformación de las observaciones.

La base de datos utilizada para demostrar el proceso de análisis de regresión múltiple funciona para trabajar regresión logística, *nacimientos_2021*, donde se pondrá especial atención en la variable *Se_cond_Indigena*, ya que en los resultados de los análisis anteriores se observa valores de interés y cumple con la condición de ser dicotómica.

El modelo logit es el logaritmo de una razón de momios, los cuales al final de proceso de análisis se interpreta como la probabilidad de éxito sobre la probabilidad de fracaso, que para la obtención de este estadístico se requiere calcular los valores ajustados resultantes de la ejecución del modelo, lo que significa que para todas las observaciones se va a sustituir la variable independiente y se calculará una estimación.

De manera general, los pasos para realizar este análisis son:

1. Ejecutar el modelo de análisis *logit*.
2. Realizar la transformación a razón de momios elevándolo a la exponencial.
3. Calcular probabilidades con el modelo propuesto.
4. Evaluar el modelo logístico.

El proceso se inicia con él con el comando para ejecutar la regresión *logit* que es **logit**, con la siguiente sintaxis:

logit *variable* dummy (binaria) *variable independiente (numérica)*

Ejemplo:

logit *Se_cond_Indgena Edad*

```
. logit Se_cond_Indigena Edad

Iteration 0:   log likelihood = -461803.31
Iteration 1:   log likelihood = -461536.67
Iteration 2:   log likelihood = -461536.44
Iteration 3:   log likelihood = -461536.44

Logistic regression                             Number of obs     =  1,621,541
                                                LR chi2(1)        =     533.75
                                                Prob > chi2       =     0.0000
Log likelihood = -461536.44                     Pseudo R2         =     0.0006

------------------------------------------------------------------------------
Se_cond_Indigena |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
            Edad |  -.0104449   .0004539   -23.01   0.000    -.0113345   -.0095552
           _cons |  -2.138582     .01202  -177.92   0.000     -2.16214   -2.115023
------------------------------------------------------------------------------
```

Estas son las interacciones 0, 1, 2, 3 con las derivadas hasta encontrar las probabilidades que maximizan la función.
El log de verosimilitud se detiene en -461536.44, valor que ayuda a comprender el ajuste del modelo.

Método de evaluación Chi^2

Método de evaluación Seudo R^2

Coeficiente de regresión logit en momio logístico.

El coeficiente de logit y la constante (el percepto) se lee en unidades momios logísticos y como tal así se debe interpretar. Al aumento de las unidades de "Y" (variable dependiente), los momios logísticos de esta toman el valor de 1, es decir que muestre un caso de éxito, aumenta o disminuya, según sea el valor del coeficiente en momios logísticos.

El siguiente paso es hacer la transformación a razón de momios elevándolo a la exponencial, agregando a la instrucción de **logit** el comando **or:**

logit *variable dummy (binaria) variable independiente (numérica)***, or**

Ejemplo:

logit Se_cond_Indigena Edad**, or**

```
. logit Se_cond_Indigena Edad, or

Iteration 0:   log likelihood = -461803.31
Iteration 1:   log likelihood = -461536.67
Iteration 2:   log likelihood = -461536.44
Iteration 3:   log likelihood = -461536.44

Logistic regression                             Number of obs     =  1,621,541
                                                LR chi2(1)        =     533.75
                                                Prob > chi2       =     0.0000
Log likelihood = -461536.44                     Pseudo R2         =     0.0006

------------------------------------------------------------------------------
Se_cond_Indigena | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
            Edad |   .9896095   .0004492   -23.01   0.000     .9887295    .9904903
           _cons |   .1178218   .0014162  -177.92   0.000     .1150785    .1206305
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

El resultado es basado en datos anteriores, pero transformado en unidades de razón de momios logísticos. Lo que significa que ya es la probabilidad de éxito sobre la probabilidad de fracaso, pero en la interpretación se describe como posibilidades de que suceda el evento de éxito.

En este ejemplo la inferencia a la que se concluye es que no es conveniente continuar con el análisis, porque el resultado de momios salió negativo, es decir que se identifica un efecto de disminución, demostrando que la probabilidad de fracaso es mayor que la probabilidad de éxito. Sin embargo, en función de explicar cómo sería la interpretación del coeficiente logit, se sugiere la siguiente redacción: A un año más de edad las posibilidades de que una mamá se considera indígena aumentan en .9896095.

Obteniendo la razón de momio se puede continuar al siguiente paso que es calcular probabilidades con el modelo planteado. Dicha acción para realizarla en cada una de las observaciones del conjunto de datos, se tiene que calcular los valores ajustados por medio del comando **predict**. Esta instrucción debe correrse después de haber ejecutado los pasos anteriores.

predict *nombre para la nueva variable que se generará*

Ejemplo:

predict *p_ind*

```
. predict p_ind2
(option pr assumed; Pr(Se_cond_Indigena))
(68 missing values generated)
```

Este comando indica el cálculo de la variable dependiente para que esta tome el valor de 1, es decir pondera la probabilidad de que se cumpla el evento de éxito.

La instrucción genera una nueva variable, la cual para visualizarse se puede introducir los comandos **sum**, **tab** o **list**. Para continuar con la representación del proceso de análisis se utiliza el comando **sum**:

sum p_ind2

```
. sum p_ind2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
p_ind2	1,639,411	.0824859	.0049752	.0598145	.0968622

Con el tabulado de estadísticas descriptivas es posible identificar que las probabilidades varían entre .0598145 (5.98%) y .0968622 (9.68%) a favor del evento que la mamá se considere indígena en la media del total de la variable edad. En el caso de que se requiera conocer la probabilidad en determinadas observaciones, se utiliza el comando **list**, agregando la instrucción **in:**

list *nombre de la variable generada* **in** *rango observaciones*

list *p_ind* **in** *500/530*

```
. list p_ind in 500/530
```

	p_ind
500.	.0808363
501.	.0785385
502.	.0839955
503.	.0748412
504.	.0748412
505.	.0889447
506.	.0808363
507.	.086438
508.	.0719993
509.	.0816158
510.	.0734076
511.	.0785385
512.	.0800636
513.	.0777859
514.	.0816158

El resultado que se obtiene es el listado de observaciones indicadas con su respectiva probabilidad de que el evento de éxito se cumpla, que en el caso del ejemplo es que la mamá se considere indígena.

De acuerdo a lo que se observa en la tabla que van de la observación 500 a la 514, la probabilidad de que la mamá se autoperciba como indígena oscila entre .0712 (7.01%) y .0890 (8.90%).

Otra opción para el cálculo de probabilidades se puede hacer con el comando **margins**, con el cual se estima la frecuencia relativa promedio, es decir la probabilidad media del evento de éxito. Esta instrucción sólo se puede ejecutar si previamente se corrió el modelo de regresión *logit*.

Ejemplo:

Modelo: **logit** *Se_cond_Indgena Edad*

margins

```
. margins

Predictive margins                              Number of obs     =  1,621,541
Model VCE    : OIM

Expression   : Pr(Se_cond_Indigena), predict()
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.0824833	.000216	381.87	0.000	.0820599	.0829066

Probabilidad media la frecuencia relativa de éxito

El resultado obtenido para la constante con un valor de 0.0825 representa la probabilidad predicha de que la madre se autodenomine como indígena cuando todas las variables independientes del modelo están en su valor de la media. Dicho en otras palabras, cuando se presente una mamá con una edad de 26 años, que es el promedio de acuerdo a este conjunto de datos, la probabilidad predicha de que ella se considere indígena es de aproximadamente 8.25%.

El mismo comando también funciona para conocer la probabilidad de un valor específico, **margins** agregando la instrucción **at**

margins, at (*variable independiente = valor específico***)**

margins, atmeans → Este comando muestra la frecuencia relativa exactamente de la media.

Ejemplo para obtener el valor exacto de la media:

margins, atmeans

```
. margins, atmeans

Adjusted predictions                            Number of obs     =  1,621,541
Model VCE    : OIM

Expression   : Pr(Se_cond_Indigena), predict()
at           : Edad            =     26.07276 (mean)
```

Valor de la media

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.0823447	.0002161	381.09	0.000	.0819212	.0827682

Probabilidad de la media exacta

Ejemplo para obtener el valor exacto en algún valor de la variable independiente:

margins, at(*Edad=28***)**

```
. margins, at(Edad=28)

Adjusted predictions                            Number of obs     =  1,621,541
Model VCE    : OIM

Expression   : Pr(Se_cond_Indigena), predict()
at           : Edad            =          28
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.0808363	.0002256	358.30	0.000	.0803941	.0812785

Probabilidad de la variable independiente cuando toma el valor de 0

La edad en la que se desea saber la probabilidad de que la madre se considere indígena es de 28 años, y con el análisis de regresión logística se sabe que es de .0808 (8.08%). En comparación con el resultado anterior sobre la probabilidad de evento de éxito sobre la media, se señala que es mayor probable que la madre pertenezca a la comunidad indígena si es de menor edad.

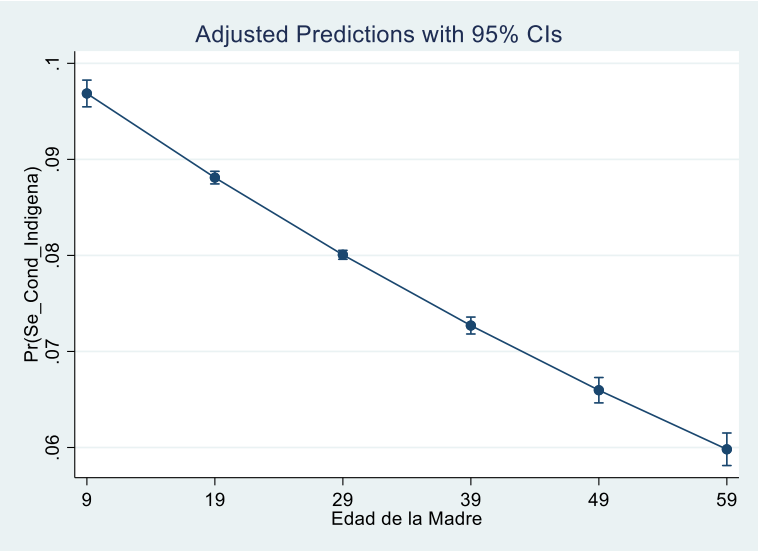
Para una mejor interpretación se puede graficar las estimaciones de probabilidad y nuevamente con el comando **margins**, siendo capaz de realizar la instrucción agregando el comando **plot**.

El comando **marginsplot** sólo es posible ejecutarlo si previamente se realiza los comandos de margins, lo que tendría que seguir la siguiente serie de pasos

margins, atmeans o margins, at(*Edad=0***) o margins, at (**Edad=**(**9 **(**10**)** 59**))**

marginsplot

Ejemplo:



El graficar los resultados es útil para observar las tendencias que suelen tomar los datos conforme los valores aumenten o disminuyan. En el caso del ejemplo desarrollado es evidente la tendencia descendente en la probabilidad de pertenecer a una comunidad indígena por parte de las madres a medida que la edad de ellas aumenta, hecho que coincide al comparar las probabilidades calculadas en la edad de 28 y la edad media de 26 años.

Algunos de los valores relevantes que se muestran en el gráfico pueden observar en los extremos de los ejes. A los 9 años la probabilidad de que la madre sea indígena es de aproximadamente 10%, en contraste con madres que se presenten entre los 49 y 39 años donde la probabilidad cae hasta el 4%.

Finalmente, después de haber obtenido las probabilidades por medio del modelo logístico, el último paso es evaluarlo. El método que se utiliza es la **Prueba de Chi2 de Pearson**, ya que con este se demuestra si el modelo se encuentra ajustado de manera correcta. El comando **estat gof** se debe ejecutar después de haber corrido el modelo para que arroje su resultado.

PLANTEAMIENTO DE HIPÓTESIS PARA LA PRUEBA DE EVALUACIÓN DEL MODELO LOGIT

```
. estat gof

Logistic model for Se_cond_Indigena, goodness-of-fit test

      number of observations =   1348613
 number of covariate patterns =   1217179
       Pearson chi2(1217168) =   1219187.01
                 Prob > chi2 =         0.0979
```

P value resultante de la prueba de evaluación, el valor que determina si rechazamos o no la prueba de hipótesis.

H_o: El modelo NO está adecuadamente ajustado

vs

H_1: El modelo está adecuadamente ajustado

Ante la evidencia dada por el P value resultante de la prueba de Chi2 de Pearson, con 0.0979, demostrando que no hay significancia estadística lo que significa que no se rechaza la hipótesis nula, es decir que con esta propuesta de modelo las variables independientes no tienen un efecto significativo en la variable dependiente, por lo que se sugiere tomar otros métodos de evaluación para confirmar este primer resultado.

Conclusión del capítulo Análisis Logístico

En el análisis logístico realizado para evaluar la relación entre la variable que registra si una mamá se considera indígena y la edad de la misma, resulta de interés la tendencia a la baja observada en la gráfica, la cual indica que la probabilidad predicha de pertenecer a la una comunidad indígena disminuye a medida que aumenta la edad de la madre. Sin embargo, también existe la evidencia de que la variable edad no muestra un efecto significativo en la variable dependiente, por lo que es necesario hacer análisis más profundos y complementarse, asó como explorar el modelo con otras variables que puedan aportan mayor información.

Un análisis multivariado permite estudiar simultáneamente un conjunto de variables que describan un fenómeno, lo que hace posible extraer más información acerca de este para apoyar a una mejor toma de decisiones en todos los sentidos estadísticos, como prueba de hipótesis, estimación, predicciones, etc.

En este apartado muestra el desarrollo de la aplicación de algunas técnicas de esta naturaleza, como son: Análisis por Componentes Principales (PCA), Análisis Factorial, Correlación Canónica, Análisis Discriminante y Análisis Clúster. Cada una de ellas cuenta con propiedades y características diferentes, por lo que su implementación con los datos debe adecuarse a la función de las capacidades de cada una. Es por ello que se desarrollarán ejemplos con diferentes conjuntos de datos los cuales fueron seleccionados cuidadosamente para demostrar la capacidad de cada método.

También es importante destacar que las técnicas multivariantes tienen aplicaciones únicas y valiosas en materia de diseño, implementación y diagnóstico de políticas públicas. Por ejemplo, el PCA y el análisis Factorial son herramientas poderosas para reducir la dimensionalidad de los datos, permitiendo a las instituciones encargadas de llevar registros, estudios poblacionales, geográficos o de cualquier índole donde se deba hacer recolección de datos, estás técnicas son perfectas para posteriores tratamientos

El análisis discriminante permite clasificar y predecir categorías, lo cual es útil en la segmentación de poblaciones objetivo. De igual forma, otra técnica empleada para división de grupos es el análisis de clúster, pero la forma en cómo lo hacen es diferente. Es por ello que se resalta la importancia de conocer exhaustivamente los datos que serán sometidos al proceso, que se encuentren adecuadamente preprocesador y tener claridad en el objetivo de la implementación para obtener un resultado óptimo.

Una de las propiedades compartida de los análisis multivariados es la varianza compartida y la interdependencia, lo que significa es que a través de estas propiedades se estudia la variabilidad conjunta del fenómeno, determinando que se explica a través de múltiples variables.

A través de ejemplos prácticos y detallados, esta sección demostrará cómo aplicar estas técnicas de análisis multivariado, subrayando su relevancia y utilidad en el diseño y la implementación de políticas públicas más informadas y efectivas.

Para una mayor referencia del contenido de este capítulo ver:

Cleff, T. (2020). "Applied statistics and multivariate data analysis for business and economics: A modern approach using SPSS, Stata, and Excel". Cham, Switzerland: Springer.

Este método, como uno de los objetivos principales es hacer un índice multivariado de valores característicos (*eigenvalues*), el cual ayuda a dar un orden a la información por medio del puntaje calculado por la técnica y poder comparar las observaciones entre ellas. Además, se enfoca en la reducción de la dimensión de las variables, pero no de la información, es decir que con la aplicación de esta técnica es posible explicar un fenómeno en la menos cantidad de variables posibles, pero con la construcción de nuevas variables que contengan información de las variables existentes.

Un proceso necesario en la aplicación de componentes principales, y el resto de métodos multivariantes que se verán a lo largo de la presente antología, es el proceso de rotación Varimax, el cual es necesario para maximizar los resultados sobre la variabilidad conjunta. En el caso específico de esta técnica, el proceso de rotación lo que hace es ampliar en cada componente la variabilidad que puede explicar, la que se calcula y se encuentra disponible.

La base de datos abiertos y acceso libre que se utiliza para desarrollar dicha técnica contiene datos sobre contaminación del aire registrada por medio de las emisiones de so2 en diferentes ciudades de Estado Unidos de América. Este conjunto de datos es de pocas variables y únicamente de 41 registros, que pudieran parecer contrastante contra la función y capacidad que tiene este método, pero el objetivo del trabajo es demostrar su desarrollo hasta la aplicación de generación de un índice.

Las variables que conforman al conjunto de datos denominado *contaminación* son:

Variable	Definición
ciudad	Nombre de la ciudad.
so2	Registro de emisiones de dióxido de azufre.
temp	Registro de temperatura.
manuf	Registro de contaminación emitido por manufactura.
pop	Cantidad de habitantes.
wind	Velocidad del viento.
precip	Registro de precipitación.
days	Cantidad de días de registro.

Los pasos generales a seguir para ejecutar componentes principales en Stata son:

- Proponer un diseño de modelo y ejecutar el comando para componentes principales.

- Calcular la bondad de ajuste.
- Interpretar el resultado del comando pca con las variables sugeridas para un modelo ajustado.
- Extraer los componentes que están dando mayor explicación acerca del fenómeno.
- Ejecutar rotación Varimax.
- Extraer componentes y cálculo de scores.
- Reescalar para generar un índice.

Como se ha demostrado anteriormente, un conjunto de datos puede contener diferentes variables que se expresan en diferentes unidades, lo que al momento de interpretar puede causar confusión o errores. Por ello es importante estandarizar las unidades de las variables antes de iniciar cualquier estudio. El método de componentes principales al ejecutarse por default estandariza las unidades de las variables que se especifican en el modelo, haciendo que el resultado se encuentre en unidades estandarizadas.

También es importante mencionar que existen estadísticos que indican si las variables seleccionadas para componer el modelo de análisis son las adecuadas para obtener resultados estadísticamente correctos. Para el caso de componentes principales el indicador correspondiente es Índice de Kaiser Mayer Olkin (KMO) es el que señala si es posible realizar el procedimiento ya que mide la adecuación de la muestra (Rodrígues & Giménez, 2017), a lo que se ha mencionado anteriormente como bondad de ajuste.

El primer paso es realizar la propuesta de modelo a analizar con el comando correspondiente de la técnica, que es **pca**, el cual tiene la siguiente sintaxis:

pca *listado de variables*

Ejemplo:

pca *so2 temp manuf pop wind precip days*

```
. pca so2 temp manuf pop wind precip days

Principal components/correlation                  Number of obs    =         41
                                                  Number of comp.  =          7
                                                  Trace            =          7
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.72812	1.21578	0.3897	0.3897
Comp2	1.51233	.117362	0.2160	0.6058
Comp3	1.39497	.502982	0.1993	0.8051
Comp4	.891991	.545213	0.1274	0.9325
Comp5	.346779	.246491	0.0495	0.9820
Comp6	.100288	.0747727	0.0143	0.9964
Comp7	.0255149	.	0.0036	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
so2	0.4897	0.0846	0.0144	-0.4042	0.7304	0.1833	0.1495	0
temp	-0.3154	-0.0886	0.6771	0.1852	0.1625	0.6107	-0.0237	0
manuf	0.5412	-0.2259	0.2672	0.0263	-0.1641	-0.0427	-0.7452	0
pop	0.4876	-0.2820	0.3448	0.1134	-0.3491	-0.0879	0.6491	0
wind	0.2499	0.0555	-0.3113	0.8619	0.2683	0.1501	0.0158	0
precip	0.0002	0.6259	0.4920	0.1839	0.1606	-0.5536	-0.0103	0
days	0.2602	0.6780	-0.1096	-0.1098	-0.4400	0.5049	0.0082	0

Después de haber corrido el modelo se recomienda hacer el cálculo de bondad de ajuste (KMO), el cual se expresa entre 0 y 1 y se espera que se ubique más cercano a 1, ya que demuestra que la variabilidad explicada es mejor porque significa que la variabilidad compartida es muy elevada. Un valor que rebase .5 es aceptable y favorable para el análisis del modelo diseñado. Si este indicador se aleja de 1 quiere decir que no se presta para un análisis multivariado.

Existen dos formas de calcular dicho indicador:

El total entre todas las variables	Calcular el indicador variable por variable
factortest *listado de variables del modelo*	**factortest** *listado de variables del modelo* Después comando **estat kmo**
Ejemplo: **factortest** *so2 temp manuf pop wind precip days*	Ejemplo: **factortest** *so2 temp manuf pop wind precip days* **estat kmo**

Con la ejecución de la serie de comandos Stata muestra una tabla resultante como se muestra a continuación:

```
. factortest so2- days

Determinant of the correlation matrix
Det                =     0.005


Bartlett test of sphericity

Chi-square         =           198.584
Degrees of freedom =                21
p-value            =             0.000
H0: variables are not intercorrelated


Kaiser-Meyer-Olkin Measure of Sampling Adequacy
KMO                =     0.430
```

Aquí se ubica el valor del indicador KMO, quien indica el nivel de la bondad de ajuste del modelo.

Con respecto al resultado obtenido de 0.430 el valor de KMO es bajo para lo establecido, por lo que se sugiere hacer un cambio en el modelo que contribuya a la variabilidad explicada.

La otra forma de conocer el indicador es la que se sugiere calcular, por la ventaja de conocer el indicador por variables, con ello se reconoce cuál de estas ayuda a mejorar el ajuste y cuál es conveniente retirar del modelo, demostrándose con el valor más bajo.

factortest so2 temp manuf pop wind precip days

estat kmo

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
so2	0.5708
temp	0.3248
manuf	0.5330
pop	0.5252
wind	0.4025
precip	0.2269
days	0.3461
Overall	0.4297

De acuerdo al ejemplo que se muestra, la variable precipitación tiene poca contribución al tener un valor bajo de KMO, por lo que conveniente es retirarla del modelo.

Como ya se demuestra en la imagen, la variable que menor aporte tiene al modelo es la que contiene el registro de precipitación en las ciudades con un KMO de 0.2269, arrojando la evidencia suficiente para retirarla y volver a correr el análisis, esperando una mejor ponderación de este indicador.

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
so2	0.5942
temp	0.6412
manuf	0.5289
pop	0.5163
wind	0.5756
days	0.7602
Overall	0.5650

Con la omisión de la variable que menos aportación presentaba a la variabilidad compartida, el indicador KMO Overall arroja una bondad de ajuste superior a .5, lo que es favorable para continuar con el análisis. Además, de que el resto de las variables también están por arriba de .5, demostrando que estas variables son las suficientes para ayudar a entender el fenómeno y el modelo se encuentra adecuadamente ajustado.

pca so2 temp manuf pop wind days

estat kmo

Al obtener un modelo ajustado adecuadamente junto con la prueba, se continúa con el análisis por componentes principales y su interpretación. El resultado que arroja Stata se compone de los siguientes elementos:

Modelo adecuadamente ajustado: **pca** *so2 temp manuf pop wind days*

Al obtener un modelo ajustado adecuadamente junto con la prueba, se continúa con el análisis por componentes principales y su interpretación. El resultado que arroja Stata se compone de los siguientes elementos:

```
. pca so2 temp manuf pop wind days

Principal components/correlation                  Number of obs    =         41
                                                  Number of comp.  =          6
                                                  Trace            =          6
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Componentes calculados

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.72812	1.28669	0.4547	0.4547
Comp2	1.44143	.506765	0.2402	0.6949
Comp3	.934666	.379604	0.1558	0.8507
Comp4	.555062	.239879	0.0925	0.9432
Comp5	.315183	.289645	0.0525	0.9957
Comp6	.0255378	.	0.0043	1.0000

Porcentaje correspondiente de variabilidad conjunta de cada componente

Porcentaje acumulado en los 2 primeros componentes

Eigenvalue (valores característicos)

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Unexplained
so2	0.4897	-0.0507	-0.3832	-0.2786	0.7152	0.1472	0
temp	-0.3154	0.5644	-0.0410	0.5962	0.4730	-0.0329	0
manuf	0.5412	0.3501	0.0139	0.0469	-0.1667	-0.7445	0
pop	0.4876	0.4464	0.0824	0.1435	-0.3360	0.6502	0
wind	0.2499	-0.2551	0.8561	0.1695	0.3327	0.0138	0
days	0.2601	-0.5403	-0.3341	0.7179	-0.1158	-0.0001	0

Matriz de coeficientes / eigenvectors para todas las variables y componentes.

Los ponderadores de los componentes.

Component	Eigenvalue	Difference
Comp1	2.72812	1.28669
Comp2	1.44143	.506765
Comp3	.934666	.379604
Comp4	.555062	.239879
Comp5	.315183	.289645
Comp6	.0255378	.

Los eigenvalue ayudan a determinar en a qué componente se inclinan las observaciones en la solución ortogonal

El objetivo de la técnica de componentes principales es reducir el número de variables a estudiar sin perder información. Como se muestra en el resultado obtenido, el primer y segundo componente explican el 69.49% de la variabilidad conjunta del fenómeno.

Enfocándose en los eigenvalue se resalta que los componentes 1 y 2 están explicando un 4.16955 (2.72812 + 1.44143) de las variables del modelo. Algo positivo porque se está explicando 4.16955 variables de 6.

Lo siguiente en el proceso es extraer los componentes que están dando mayor explicación del fenómeno, que generalmente se ubican entre los primeros 2 o 3 componentes resultantes. Para ello se utiliza el comando del modelo agregando la instrucción **comp**

pca *listado de variables del modelo*, **comp (***número de componentes a extraer***)**

Ejemplo: **pca** *so2 temp manuf pop wind days*, **comp (2)**

Algunos puntos a considerar para la interpretación:

```
. pca so2 temp manuf pop wind days, comp(2)

Principal components/correlation                  Number of obs    =         41
                                                  Number of comp.  =          2
                                                  Trace            =          6
    Rotation: (unrotated = principal)             Rho              =     0.6949
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.72812	1.28669	0.4547	0.4547
Comp2	1.44143	.506765	0.2402	0.6949
Comp3	.934666	.379604	0.1558	0.8507
Comp4	.555062	.239879	0.0925	0.9432
Comp5	.315183	.289645	0.0525	0.9957
Comp6	.0255378	.	0.0043	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
so2	0.4897	-0.0507	.3421
temp	-0.3154	0.5644	.2694
manuf	0.5412	0.3501	.02432
pop	0.4876	0.4464	.06415
wind	0.2499	-0.2551	.7358
days	0.2601	-0.5403	.3946

Número de observaciones

Número de componentes extraídos

Traza: Diagonal principal de la matriz de correlaciones

Variabilidad explicada del fenómeno con los componentes extraídos. Es como la R^2, la bondad de ajuste.

Variabilidad no explicada por cada variable.

Variabilidad ponderada por cada variable a cada componente

- Los valores que son importantes en un componente, en el segundo no lo serán porque son ortogonales, es decir que no hay correlación entre ellos y en su conjunto ayudan a comprender todo el fenómeno de estudio.
- Los resultados obtenidos se interpretan de forma separada entre los componentes. Se debe hacer una interpretación para el componente 1 y otra interpretación para el componente 2.
- Los eigenvector son los valores que indican a qué componente corresponde cada variable (marcadas en diferentes colores para cada componente). A partir de esto se determina la descripción de la clasificación que hace el análisis, es decir el cómo se integra cada componente.

Para obtener un mejor resultado de los eigenvectors se requiere hacer el proceso de rotación por Varimax, ya que existen observaciones en las fronteras de los vectores, y aunque los eigenvaluen hacen la separación entre cada componente, es probable que algunas observaciones no hayan sido tomadas en cuenta en la variabilidad.

El comando para aplicar la rotación varimax es **rot**:

pca *listado de variables del modelo*, **comp (***número de componentes a extraer***)**

Después de correr el modelo se ejecuta el comando **rot**

Ejemplo:

pca *so2 temp manuf pop wind days*, **comp (2)**

rot

```
. rot

Principal components/correlation                  Number of obs    =         41
                                                  Number of comp.  =          2
                                                  Trace            =          6
Rotation: orthogonal varimax (Kaiser off)         Rho              =     0.6949
```

Component	Variance	Difference	Proportion	Cumulative
Comp1	2.34922	.528888	0.3915	0.3915
Comp2	1.82033	.	0.3034	0.6949

Rotated components

Variable	Comp1	Comp2	Unexplained
so2	0.3838	0.3083	.3421
temp	0.0414	-0.6452	.2694
manuf	0.6446	-0.0004	.02432
pop	0.6518	-0.1104	.06415
wind	0.0714	0.3499	.7358
days	-0.0747	0.5950	.3946

Component rotation matrix

	Comp1	Comp2
Comp1	0.8400	0.5427
Comp2	0.5427	-0.8400

El Kaiser off significa que no hay observaciones que no contribuyan, pero en este ejemplo no es el caso.

Variabilidad explicada del fenómeno no cambia porque se determina desde el principio del análisis

La matriz de

La matriz de componentes rotados indica las ecuaciones lineales, tanto él intercepto como la pendiente de los vectores.

Al comparar los resultados del análisis antes y después de la rotación se observa un cambio en los eigenvalues y los eigenvectors porque se ajustaron para aportar mayor variabilidad, con lo que se puede llegar a una mejor interpretación dado con estos nuevos valores por variable.

Antes de rotación

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
so2	0.4897	-0.0507	.3421
temp	-0.3154	0.5644	.2694
manuf	0.5412	0.3501	.02432
pop	0.4876	0.4464	.06415
wind	0.2499	-0.2551	.7358
days	0.2601	-0.5403	.3946

Después de rotación

Rotated components

Variable	Comp1	Comp2	Unexplained
so2	0.3838	0.3083	.3421
temp	0.0414	-0.6452	.2694
manuf	0.6446	-0.0004	.02432
pop	0.6518	-0.1104	.06415
wind	0.0714	0.3499	.7358
days	-0.0747	0.5950	.3946

Con los componentes seleccionados y rotados por el método Varimax, se puede seguir con el cálculo de scores, lo cuales su función es asignar un puntaje a cada una de las observaciones que integran al conjunto de datos, haciendo posible identificar posiciones dentro de una escala. Dicha acción requiere que se repita el paso de extracción de los componentes seleccionados, sólo que para hacer la ponderación de scores se utiliza el comando **predict:**

predict *nombre de la nueva variable correspondiente al componente 1 nombre de la nueva variable correspondiente al componente 2…*

Ejemplo:

predict *c1 c2*

```
. predict c1 c2
(score assumed)

Scoring coefficients
    sum of squares(column-loading) = 1
```

Variable	Comp1	Comp2
so2	0.4897	-0.0507
temp	-0.3154	0.5644
manuf	0.5412	0.3501
pop	0.4876	0.4464
wind	0.2499	-0.2551
days	0.2601	-0.5403

- Como resultado se obtiene los coeficientes basados en la rotación varimax ortogonal, quienes son los ponderadores para calcular las combinaciones lineales.

- Con estos se calculan los scores de los valores ajustados.

Los *scores* (puntaje) permiten la obtención de mayor información, ya que hace posible solicitar información de estadística descriptiva de estos mismos, así como graficas que ayuden a dar respuesta a la problemática, a dar respuesta, con la aplicación de otros comandos como **scoreplot.**

Como se había comentado, uno de los propósitos de la técnica componentes principales era la obtención de un índice, el cual se realiza por medio de los scores ponderados para cada

observación, de igual manera se mencionó el poder asignarles una posición dentro de un rango, que sería el paso final de este proceso de análisis. El reescalar las variables hace que los scores se transformen en una escala que quien realice el análisis lo determine, con el fin de hacer a la interpretación más sencilla e intuitiva. Vuelve al rango del índice a una escala más entendible, como entre 0 y 10, 0 y 1, 0 y 100, etc. Usualmente, se utiliza la escala entre 0 y 100. Para ello se utiliza el comando **gen index** en la siguiente fórmula:

gen index=((*nombre del componente* **- r(min))/(r(max)-r(min)))*100**

Donde:

- **gen index=**: Es la instrucción de realizar un índice.
- **r**: Son los escalares. Los escalares son los valores mínimos y máximos del rango de scores que tiene un componente.
- **r(min)**: Es la indicación que tome en cuenta el escalar r y el valor mínimo del componente que se está reescalando.
- **r(max)-r(min)**: Restar los escalares mínimo y máximo del componente a reescalar.
- ***100**: Todo lo anterior se calcule por 100. Este valor puede cambiar de acuerdo al rango en el que se desee convertir la escala.

Se recomienda hacer la consulta de los escalares del componente a reescalar, que se puede hacer con el comando **sum**, así los valores que se utilizarán en la fórmula el programa de Stata los tendrá inmediatamente en la memoria de trabajo.

Ejemplo:

Consultar los escalares de los componentes extraídos y ponderados:

sum *c1 c2*

```
. sum c1 c2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          c1 |         41    2.18e-09    1.651702  -2.682108    7.23104
          c2 |         41   -4.02e-09    1.200596   -2.38225   3.205482
```

Aplicar la fórmula:

gen index=((*c1* **- r(min))/(r(max)-r(min)))*100**

Hacer una vista rápida de las estadísticas básicas del índice para corroborar que se haya hecho el cálculo correctamente:

sum index

```
. sum index
```

Variable	Obs	Mean	Std. Dev.	Min	Max
index	41	27.05607	16.66173	0	100

Al reescalar los scores con el objetivo de que se distribuyan en una escala entre 0 y 100, también hace que se reescale la media. En cuanto a la interpretación, podría redactarse: La media de contaminación de acuerdo a los registros de las variables dadas es de 27.056, lo que permite entender más o menos como se encuentra el grado promedio de esta categoría de contaminación (la cual es identificada por la descripción que dan las variables que integran al componente c1), presente en todas las ciudades.

Gracias al índice generado se pueden realizar consultas de información que puedan responderse por medio de las variables que integran a los componentes. En el caso del ejemplo se puede consultar al índice qué información detallada tienen las ciudades y qué posición ocupan, ejecutándose por medio de uno de los comandos **list** o **table**, antecediendo la instrucción **sort** (indica que el resultado lo ordene de forma ascendente de acuerdo a la variable seleccionada).

Ejemplo:

sort *index*

list *ciudad index so2 temp manuf pop wind days*

```
. sort index

. list ciudad index so2 temp manuf pop wind days
```

	ciudad	index	so2	temp	manuf	pop	wind	days
1.	Phoenix	0	10	70.3	213	582	6	36
2.	Alburq	8.959139	11	56.8	46	244	8.9	58
3.	Lrock	9.938818	13	61	91	132	8.2	100
4.	Miami	9.956051	10	75.5	207	335	9	128
5.	NewO	10.56318	9	68.3	204	361	8.4	113
6.	Jackson	13.34227	14	68.4	136	529	8.8	116
7.	Richmond	16.95509	26	57.8	197	299	7.6	115
8.	Charlest	16.97202	31	55.2	35	71	6.5	148
9.	Nashville	17.51365	18	59.4	275	448	7.9	119
10.	Sfran	17.70378	12	56.7	453	716	8.7	67
11.	Memphis	17.86481	10	61.6	337	624	9.2	105
12.	SLC	18.11021	28	51	137	176	8.7	89

Como resultado se obtiene una tabla con los registros de las variables indicadas en el comando (que son las planteadas en el modelo de análisis) junto con su posición en la escala que va del 1 al 100, de acuerdo a la variable *index*.

Con respecto a lo que muestra la tabla la ciudad de Phoenix ocupa el lugar 0 del índice lo que sugiere que esta ciudad tiene los valores más bajos en la combinación ponderada de las variables al compararse con las otras ciudades, acercándose a la idea de que es posiblemente la que cuenta con menos contaminación.

Conclusión del capítulo Componentes Principales:

El análisis de componentes principales (PCA) ha sido efectivo para realizar una comparación más sencilla entre las ciudades a través del índice generado, logrando hacer un ranking entre estas. La transformación de las variables originales en un índice normalizado entre 0 y 100 proporciona una forma clara de identificar qué ciudades tienen las mayores y menores contribuciones de las variables que conforman el modelo y nos permite entender el fenómeno de la contaminación compuesto por cada factor que la compone.

Este ejemplo, de tener que emplearse a un caso de la vida real, sería de gran utilidad para la formulación de políticas públicas en materia ambiental, ya que permite identificar rápidamente qué ciudades pueden necesitar más atención o intervenciones por parte de las instituciones, dependencias gubernamentales, organizaciones y sociedad en general. Retomando del caso de la ciudad de Phoenix, con el índice más bajo, podría ser una prioridad menor en comparación con Salt Lake City, que tiene el índice más alto y, aparentemente, mayores desafíos ambientales.

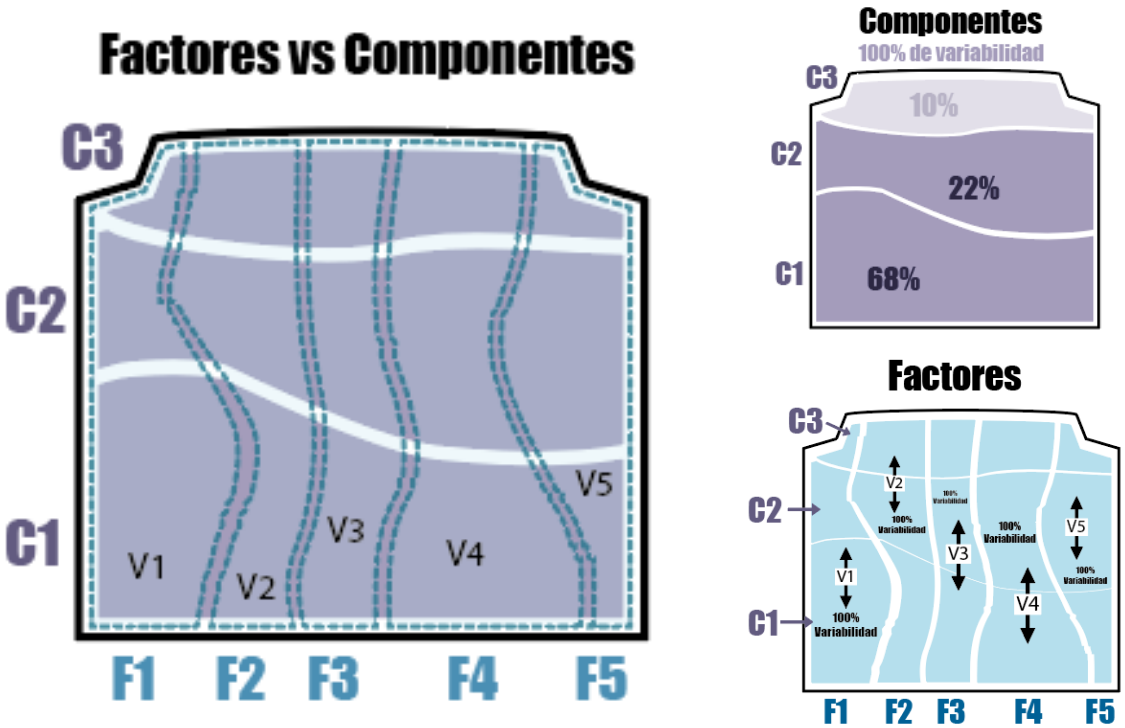
Para una mayor referencia del contenido de este capítulo ver:

Geldhof, J., Arnold, M., & Constantine, N. (2016). Better crunching: Recommendations for multivariate data analysis approaches for program impact evaluations. "Journal of Extension", (June).

Es probable que exista confusión al momento de contrastar los resultados obtenidos por un análisis por componentes principales y análisis factorial, ya que ambas técnicas buscan simplificar la información por medio del agrupamiento de variables. Sin embargo, se comportan de forma diferente los componentes y los factores. En el ámbito de análisis multivariante, un factor es el conjunto de variables que definen un concepto, el cual se puede entender como un constructo.

Para entender mejor la diferencia entre ambos a continuación se presenta el diagrama tinaco:

Imagine un tinaco lleno al 100% de capacidad (lo que representa al conjunto de datos) y se quiere explicar lo que hay en su interior.



FACTORES vs COMPONENTES

Factor basado en componentes principales	Componente
Grupo de variables que definen al concepto. * Examinar la máxima variabilidad que se pueda identificar por variable (no en conjunto). * Los factores se analizan de forma vertical. * No se omite variabilidad, lo que significa que las variables que son tomadas en cuenta en el resultado del análisis estarán explicadas en un 100%. * Automáticamente, la técnica define los factores que se asocian a todas las variables. Lo hace de forma óptima (con los menos factores posibles) que explican la máxima variabilidad del fenómeno que se está estudiando. *Carga factorial: el grado de correspondencia entre la variable y el Factor, es decir, cargas altas indican que dicha variable es representativa para dicho factor.	Un conjunto de variables que ponderan un concepto. * Buscamos encontrar la máxima variabilidad explicada concentrada en el conjunto de observaciones (tinaco). Esto es lo importante. *Después de haber encontrado la máxima variabilidad explicada, se concentra la atención en las variables. *Los componentes se analizan de forma horizontal. * Se puede omitir variabilidad porque solo se utiliza los primeros componentes para explicar el fenómeno.

Existen dos tipos de análisis factorial:

Exploratorio

- Se tiene un conjunto de variables que no están organizadas, ni clasificadas, haciendo que se dificulte visibilizar un concepto.
- Es posible saber los conceptos, pero no las variables, o no conocer ni variables, ni conceptos.
- En la práctica, cuando se tiene este panorama, se ingresa el conjunto de datos al software y se corre el análisis factorial para que este identifique libremente los conceptos por el agrupamiento de las variables.

Confirmatorio

- Se conocen las variables que definen cada concepto o se tiene identificado el concepto, pero no las variables.
- Ejemplo de ello podría ser cuando se tiene un conjunto de datos sobre violencia, y se sabe que hay diferentes tipos de violencia, psicológica, física, económica, patrimonial, etc. Entonces se requiere saber los patrones entre los variables y cómo se comportan en estos conceptos.
- Funciona para confirmar o replantear los conceptos y compararlos con la literatura.

De forma general se identifican 4 pasos principales para elaborar un análisis por medio de esta técnica:

1. Hacer un análisis por matriz de correlaciones. * Se requiere una adecuada matriz de correlaciones para un buen análisis factorial, donde las correlaciones deben ser mayores a 0.3.	2. Extraer los factores. (Recomendable de 4 a 5 factores, para después escalarlos si se requiere). * El proceso se iguala al obtener valores ajustados (método de mínimos cuadrados) o el score (método de componentes principales). Se obtendrán los valores ajustados de cada una de las combinaciones que permiten extraer a todos los factores, por lo que se obtienen nuevas variables.
3. Realizar rotación. Los factores son ortogonales, por lo que no están correlacionados. *Rotación Varimax Teóricamente, los conceptos se pueden relacionar entre ellos, por lo que se requiere rotar los factores de forma oblicua, lo que se denomina rotación Promax. *Rotación promax: Se va a rotar tratando de maximizar la mayor correlación entre dos o más factores. Este método ayuda a una mejor explicación del fenómeno.	4. Tomar en cuenta consideraciones finales para determinar el número de factores. * Si se encuentra dos cargas factoriales mayores a 0.5 en dos factores de una variable. La decisión de a qué factor corresponde esa variable, se inclina al que aporte mayor valor. * Depende mucho la interpretación de quien realice el análisis. * Puede suceder que un factor se pueda omitir, aunque en el análisis se muestre identificado, pero si no aporta gran explicación de variabilidad se puede omitir.

Notas relevantes para el análisis:

- Para efectos de la técnica no afecta / no es un problema que las variables de análisis sean de tipo numéricas discretas, continuas ni categóricas, ya que las unidades de las variables se estandarizan.
- En Stata se pueden hacer diferentes métodos de análisis factoriales:
  - Principal factor
  - Principal component factor (análisis factorial por componente principal).
  - Iterated principal factor (factores principales integrados).
  - Maximum-likelihood factor (máxima verosimilitud)
- Los dos últimos métodos son poco usados y son más sofisticados, pero se usan cuando los métodos de factor principal y por componentes principales no aportan buenos resultados. Sin embargo, lo principal de la ejecución de estas técnicas es explicar la máxima variabilidad explicada.
- El análisis factorial por componentes principales determina conceptos que se puedan generar a partir de los datos dados. El análisis por factor principal tiene un mejor rendimiento en la creación de índices.
- Un análisis factorial eficiente te identifica entre 4 o 5 factores porque pueden explicarse adecuadamente.

La base de datos a usar es construida por la selección de algunas variables de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) de México, del módulo bienes del hogar y características sociodemográficas, específicamente del estado de Puebla. La fuente de información es Instituto Nacional de Estadística y Geografía (INEGI), lo cual nos garantiza el libre uso del conjunto del conjunto de datos. Un estudio como este permite conocer la distribución, monto y estructura de los ingresos y gastos de los hogares mexicanos, así como sus características y la de los integrantes que conforman cada uno de ellos.

El objetivo de hacer uso de esta base de datos es generar una perspectiva de la distribución de la riqueza en el estado por medio de las 25 variables selectas y los 3,003 registros. Explorar la posibilidad de conocer cuáles son las características de un hogar de esta ciudad que es considerado como mayor capacidad económica y cómo se contrasta con un hogar que puede considerarse con menor riqueza. Las variables son:

Variable	Descripción	Respuesta
CVE_EDO	Clave estado	
CVE_MPIO	Clave municipio	
edad	3.5. ¿Cuántos años cumplidos tiene? (Integrante 1)	
computadora	¿El hogar cuenta con computadora?	1 - Sí 2 - No
estufa	¿El hogar cuenta con estufa?	1 - Sí 2 - No
lavadora	¿El hogar cuenta con lavadora?	1 - Sí 2 - No
refrigerador	¿El hogar cuenta con refrigerador?	1 - Sí 2 - No
DVD	¿El hogar cuenta con DVD?	1 - Sí 2 - No
televisor	¿El hogar cuenta con televisor?	1 - Sí 2 - No
boiler	¿El hogar cuenta con boiler?	1 - Sí 2 - No
celular	¿El hogar cuenta con celular?	1 - Sí 2 - No
microondas	¿El hogar cuenta con microondas?	1 - Sí 2 - No
tostador	¿El hogar cuenta con tostador?	1 - Sí 2 - No
internet	¿El hogar cuenta con internet?	1 - Sí 2 - No
agua_entub~a	¿El hogar cuenta con agua entubada?	1 - Sí 2 - No

banio	¿El hogar cuenta con banio?	1 - Sí 2 - No
electricidad	¿El hogar cuenta con electricidad?	1 - Sí 2 - No
telefono	¿El hogar cuenta con telefono?	1 - Sí 2 - No
tv_satelital	¿El hogar cuenta con tv_satelital?	1 - Sí 2 - No
servicio_dom	¿El hogar cuenta con servicio_dom?	1 - Sí 2 - No
mun	Municipio	
zm	Su hogar se encuentra en una zona metropolitana	1 - Sí 2 - No
genero	Género del encuenstado	
hrs_lab	Horas laborales	
ing_mens	Ingreso mensual	

En la práctica, antes de ejecutar el modelo diseñado para esta técnica, se requiere conocer algún indicador de nivel variabilidad conjunta que demuestre si es viable proceder con el análisis. Para ello se utiliza el indicador KMO, el cual se determina que si es mayor a .5 es el valor suficiente para continuar.

Además, es necesario evaluar la intercorrelación existente entre las variables, que es determinado por la hipótesis nula de la prueba de esfericidad de Bartlett, la cual "comprueba si la matriz de correlaciones es una matriz identidad. Se puede dar como válidos aquellos resultados que nos presenten un valor elevado del test y cuya fiabilidad sea menor a 0,05.(Chacón et al., 2021).

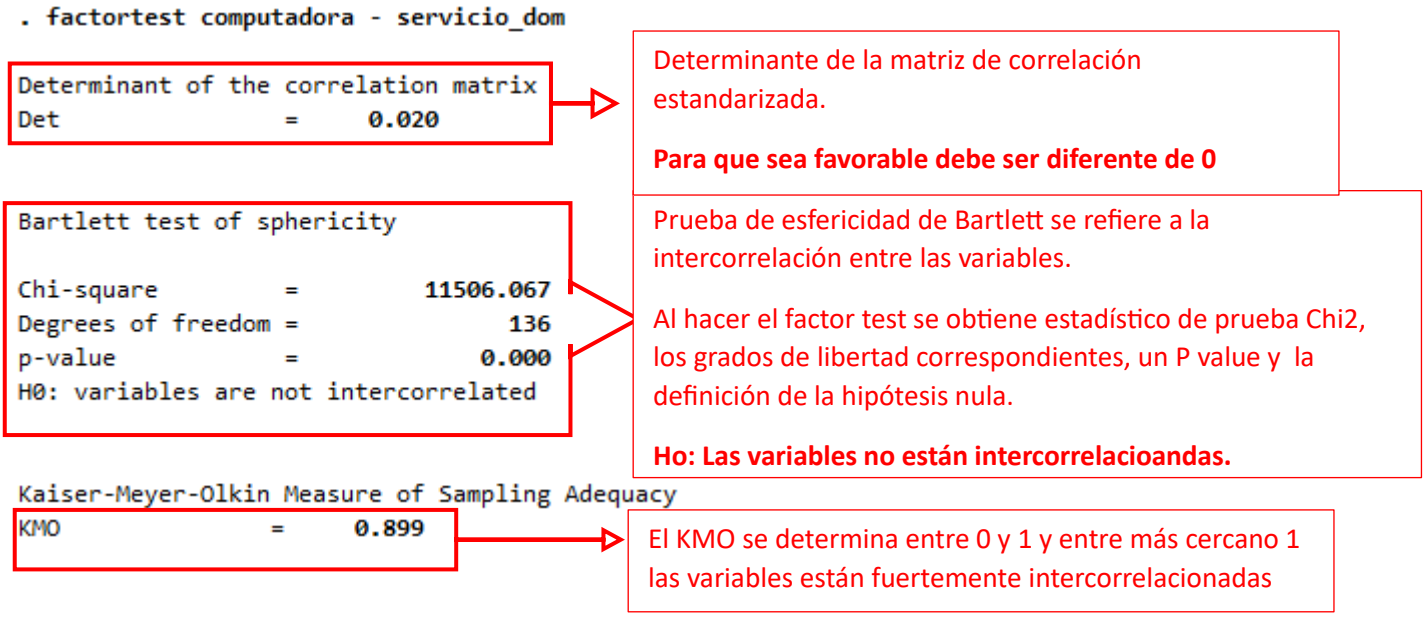
Por último, se requiere obtener el determinante de la matriz de correlación, el cual para este tipo de técnicas debe ser diferente de 0, de lo contrario sería el resultado de una matriz singular (y no cuadrada)

Estos parámetros son calculados gracias al comando **factortest:**

factortest *listado de variables pertenecientes al modelo*

Ejemplo:

factortest *computadora - servicio_dom*



.

Sea cual sea el tipo de análisis factorial que se va a realizar, se necesita los valores de los indicadores anteriores y que estos determinen si es conveniente seguir con el análisis o buscar un planteamiento de modelo distinto. Se necesitan las pruebas de que las variables que componen al modelo se encuentren intercorrelacionadas, que se genere una matriz de correlación cuadrada y un valor de KMO favorable.

De haber obtenido valores convenientes, se prosigue a ejecutar el comando de análisis factorial. Dentro de este documento se presentará el análisis factorial por componente principal y análisis factorial por factor principal, entre los cuales habrá similitudes y diferentes durante el proceso.

La explicación se inicia por el **análisis factorial por componente principal**, el cual se enfoca en clasificar y generar nuevas variables que contribuyan a explicar de mejor manera el fenómeno, mediante la agrupación de variables en conceptos (factores). El comando para realizar un análisis factorial bajo este método es:

factor *listado de variables que componen al modelo* **, pcf**

Ejemplo:

factor *computadora - servicio_dom*, **pcf**

```
. factor computadora - servicio_dom, pcf
(obs=2,960)

Factor analysis/correlation                    Number of obs    =      2,960
    Method: principal-component factors        Retained factors =          3
    Rotation: (unrotated)                      Number of params =         48
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.88012	3.34001	0.2871	0.2871
Factor2	1.54011	0.47839	0.0906	0.3777
Factor3	1.06171	0.06533	0.0625	0.4401
Factor4	0.99638	0.08549	0.0586	0.4987
Factor5	0.91090	0.08567	0.0536	0.5523
Factor6	0.82522	0.02158	0.0485	0.6008
Factor7	0.80364	0.06159	0.0473	0.6481
Factor8	0.74205	0.02362	0.0437	0.6918
Factor9	0.71843	0.02420	0.0423	0.7340
Factor10	0.69423	0.02119	0.0408	0.7749
Factor11	0.67304	0.04135	0.0396	0.8145
Factor12	0.63169	0.03389	0.0372	0.8516
Factor13	0.59780	0.03013	0.0352	0.8868
Factor14	0.56766	0.00234	0.0334	0.9202
Factor15	0.56532	0.04037	0.0333	0.9534
Factor16	0.52495	0.25822	0.0309	0.9843
Factor17	0.26674	.	0.0157	1.0000

```
LR test: independent vs. saturated: chi2(136) = 1.2e+04 Prob>chi2 = 0.0000
```

```
Factor loadings (pattern matrix) and unique variances
```

Variable	Factor1	Factor2	Factor3	Uniqueness
computadora	0.7153	-0.3228	0.0496	0.3817
estufa	0.4624	0.5242	-0.0807	0.5049
lavadora	0.6064	0.1463	-0.0980	0.6013
refrigerador	0.5364	0.3286	-0.1138	0.5914
DVD	0.5714	0.0381	-0.0536	0.6692
televisor	0.2794	0.5311	0.4318	0.4534
boiler	0.6736	0.0431	-0.1231	0.5292
celular	0.5256	0.1758	-0.2276	0.6411
microondas	0.6611	-0.1719	-0.0361	0.5321
tostador	0.4516	-0.3633	0.2491	0.6020
internet	0.7497	-0.3283	0.0659	0.3258
agua_entub~a	0.2520	0.2525	-0.1796	0.8405
banio	0.5652	0.2116	-0.2419	0.5772
electricidad	0.1225	0.4218	0.6699	0.3582
telefono	0.5984	-0.2353	0.1290	0.5700
tv_satelital	0.5570	-0.0739	-0.0002	0.6843
servicio_dom	0.3089	-0.3144	0.3873	0.6558

Número de observaciones tomadas para el análisis

Número de factores retenidos

Número de parámetros que se requieren para estimar

Factores retenidos por arrojar un eigenvalue mayor a 1.

Porcentaje correspondiente de variabilidad conjunta de cada factor

Porcentaje acumulado en los 3 primeros factores

Prueba de independencia likelihood ratio (LR)

Chi^2 con sus respectivos grados de libertad

Pvalue sobre la hipótesis nula

Ho: las variables son independientes

CARGAS FACTORIALES

Identifica los patrones de la matriz

Unicidad

Lo que no se explica de cada variable

Comunalidad

Es la suma de las cargas factoriales elevadas al cuadrado específicamente de una variable, en cada uno de los factores.

Esta técnica en automático identifica los conceptos por medio de los factores retenidos. En el este caso del ejemplo ha retenido 3 factores (dentro de los 17 resultantes), con los cuales se explica la mayor variabilidad, demostrándose por los eigenvalues mayores o igual a 1, haciendo que se extraigan aquello que cuentan con esta característica. Lo demás lo omite por no considerarlo concluyente en la covariabilidad.

Una información de gran valor resultante de este análisis son las cargas factoriales, las cuales ponderan de una forma especial al eigenvalue permitiendo entender mejor porque una variable se asocia con un cierto nivel al factor. Además, son las que hacen posible la conceptualización gracias esta propiedad, ya que de acuerdo a los valores obtenidos, se identifica que variables corresponden a qué factores.

El siguiente paso en el proceso de elaboración de este análisis es hacer rotación, la cual se puede ejecutar por método varimax o método promax. Para el caso de Varimax el comando es igual que en el método anterior y debe realizarse igualmente después de haber corrido el modelo:

factor *listado de variables que componen al modelo* **, pcf**

Después de correr el modelo se ejecuta el comando **rot**

Ejemplo:

factor *computadora - servicio_dom*, **pcf**

rot

```
. rot

Factor analysis/correlation                    Number of obs    =      2,960
    Method: principal-component factors        Retained factors =          3
    Rotation: orthogonal varimax (Kaiser off)  Number of params =         48
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	3.61105	1.03062	0.2124	0.2124
Factor2	2.58043	1.28997	0.1518	0.3642
Factor3	1.29046	.	0.0759	0.4401

```
LR test: independent vs. saturated: chi2(136) = 1.2e+04 Prob>chi2 = 0.0000
```

Con la rotación cambian los eigenvalues de los factores.

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
computadora	0.7608	0.1971	-0.0252	0.3817
estufa	0.0508	0.6401	0.2877	0.5049
lavadora	0.3752	0.4998	0.0905	0.6013
refrigerador	0.2128	0.5796	0.1656	0.5914
DVD	0.4189	0.3891	0.0632	0.6692
televisor	0.0178	0.2720	0.6872	0.4534
boiler	0.4813	0.4884	0.0240	0.5292
celular	0.2652	0.5371	-0.0130	0.6411
microondas	0.6129	0.3028	-0.0225	0.5321
tostador	0.6200	-0.0868	0.0777	0.6020
internet	0.7949	0.2055	-0.0094	0.3258
agua_entub~a	0.0158	0.3981	0.0267	0.8405
banio	0.2730	0.5901	0.0007	0.5772
electricidad	0.0093	-0.0112	0.8010	0.3582
telefono	0.6365	0.1413	0.0701	0.5700
tv_satelital	0.4830	0.2836	0.0445	0.6843
servicio_dom	0.5105	-0.2120	0.1967	0.6558

También con la rotación cambian las cargas factoriales, pero la unicidad no.

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.7921	0.5912	0.1520
Factor2	-0.5670	0.6203	0.5419
Factor3	0.2261	-0.5154	0.8266

Matriz de rotación

Contiene los coeficientes de las combinaciones lineales de los factores.

La rotación no ayuda a aumentar la variabilidad explicada, este método hace que se maximice la variabilidad compartida y separarla (enviar las observaciones que quedaron fuera de los factores), hacia donde contribuya más. Es por ello que cambia el valor de los eigenvalue y por consecuencia las cargas factoriales, pero la variabilidad no explicada.

Con las cargas factoriales rotadas ya es posible determinadas qué variables corresponden a cada factor, determinándose a que sólo serán tomadas en cuenta las cargas factoriales con un valor igual o mayor a .5, pero de la forma en como se presenta resulta complejo seleccionar las variables. Para ubicar las cargas factoriales que permiten categorizar las variables se utiliza nuevamente el comando **rot**, pero agregando la instrucción **blank()**:

rot, blank(.5)

El valor que se encuentra dentro del paréntesis idealmente debe ser .5, pero puede ajustarse de acuerdo a los resultados y el objetivo de la investigación. Es posible variarse entre .4 y .5 con el fin de tomar todas las variables para definir conceptos

Ejemplo:

- Debido a que en el resultado se encuentra variabilidad no explicada (unicidad) mayor que .5, hay variables que no cuentan con el valor de carga factorial mínima de .5 y por lo tanto, no aparecen en este primer filtro, como es lavadora, DVD, agua embotellada, televisión satelital.

```
. rot, blank(.5)

Factor analysis/correlation                    Number of obs    =      2,960
    Method: principal-component factors        Retained factors =          3
    Rotation: orthogonal varimax (Kaiser off)  Number of params =         48
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	3.61105	1.03062	0.2124	0.2124
Factor2	2.58043	1.28997	0.1518	0.3642
Factor3	1.29046	.	0.0759	0.4401

LR test: independent vs. saturated: chi2(136) = 1.2e+04 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
computadora	0.7608			0.3817
estufa		0.6401		0.5049
lavadora				0.6013
refrigerador		0.5796		0.5914
DVD				0.6692
televisor			0.6872	0.4534
boiler				0.5292
celular		0.5371		0.6411
microondas	0.6129			0.5321
tostador	0.6200			0.6020
internet	0.7949			0.3258
agua_entub~a				0.8405
banio		0.5901		0.5772
electricidad			0.8010	0.3582
telefono	0.6365			0.5700
tv_satelital				0.6843
servicio_dom	0.5105			0.6558

(blanks represent abs(loading)<.5)

Con la indicación **blank(.5)** se puede observar que carga factorial de cada variable se ubica en

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.7921	0.5912	0.1520
Factor2	-0.5670	0.6203	0.5419
Factor3	0.2261	-0.5154	0.8266

- Con el propósito de que todas las variables se ubiquen en algún factor se puede modificar el valor mínimo solicitado, por ejemplo, bajarlo a .4 o .38. Sin embargo, esto debe hacerse con cautela porque de ello dependerá el resultado e interpretación requerido.
- **El valor de .4 es lo mínimo (desde la perspectiva teórica) que se podría esperar para asociar una variable a algún factor.**

Con el fin de acomodar cada variable en un factor, es este ejemplo se vuelve a ejecutar el comando **rot**, pero ahora que visualice las cargas factoriales iguales o mayores a .39:

rot, blank(.39**)**

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
computadora	0.7608			0.3817
estufa		0.6401		0.5049
lavadora		0.4998		0.6013
refrigerador		0.5796		0.5914
DVD	0.4189			0.6692
televisor			0.6872	0.4534
boiler	0.4813	0.4884		0.5292
celular		0.5371		0.6411
microondas	0.6129			0.5321
tostador	0.6200			0.6020
internet	0.7949			0.3258
agua_entub~a		0.3981		0.8405
banio		0.5901		0.5772
electricidad			0.8010	0.3582
telefono	0.6365			0.5700
tv_satelital	0.4830			0.6843
servicio_dom	0.5105			0.6558

(blanks represent abs(loading)<.39)

- Con el ajuste del comando **rot,** la variable boiler sus cargas factoriales se ubican en el factor 1 y 2, pero una variable sólo puede asignarse a un solo factor. Por ello la variable se coloca en el factor donde su carga factorial sea más elevada que para el presente ejemplo boiler se asignará al factor 2.

Con las variables identificadas en cada factor se alista la información para poder distinguir y describir cómo se componen los factores extraídos, es decir hacer el trabajo de conceptualización:

Factor 1	Factor 2	Factor 3
Lujo El lujo este asociado en común a que las personas tengan en casa los siguientes servicios y electrónicos	**Servicios básicos / indispensables** Activos de primera necesidad	**Ocio / Entretenimiento** Activos relacionados con cuestiones de entretenimiento.
Computadora (0.7608) DVD (0.4189) Microondas (0.6129) Tostadora (0.6200) Internet (0.7949) Televisión (0.6365) TV Satelital (0.4830) Servicio doméstico (0.5105)	Estufa (0.6401) Lavadora (0.4998) Refrigerador (0.5796) Boiler (0.4884) Celular (0.5371) Agua entubada (0.3981) Baño (0.5901)	Televisión (0.6872) Electricidad (0.8010)

Para definir el concepto que proponen los factores es importante tener conocimiento del tema y el fenómeno que se está estudiando. Para el caso de la práctica importa saber si la información dada ayuda a determinar diferentes niveles de riqueza indirecta, es por ello que los conceptos están con las menos palabras posibles, denominan 3 grados diferentes de posesión de los bienes enlistados, lujo, servicios básicos y entretenimiento.

El otro método de rotación es promax, que a diferencia de Varimax, este hace rotación oblicua de las cargas originales. El comando para ejecutar este método de rotación es:

rot, promax (*número de potencia que asigna quien realiza el análisis***)**

Si no se indica algún número de potencia promax, Stata por default asigna el número 3.

Ejemplo:

rot, promax (3**)**

```
. rot, promax (3)

Factor analysis/correlation                    Number of obs    =      2,960
    Method: principal-component factors        Retained factors =          3
    Rotation: oblique promax (Kaiser off)      Number of params =         48
```

Factor	Variance	Proportion	Rotated factors are correlated
Factor1	4.07489	0.2397	
Factor2	3.54660	0.2086	
Factor3	1.30432	0.0767	

LR test: independent vs. saturated: chi2(136) = 1.2e+04 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
computadora	0.7476	0.0940	-0.0694	0.3817
estufa	-0.1336	0.6878	0.2209	0.5049
lavadora	0.2539	0.4828	0.0276	0.6013
refrigerador	0.0582	0.5954	0.0998	0.5914
DVD	0.3315	0.3560	0.0103	0.6692
televisor	-0.0825	0.3015	0.6594	0.4534
boiler	0.3713	0.4530	-0.0410	0.5292
celular	0.1319	0.5386	-0.0762	0.6411
microondas	0.5627	0.2313	-0.0729	0.5321
tostador	0.6735	-0.1891	0.0670	0.6020
internet	0.7805	0.0981	-0.0556	0.3258
agua_entub~a	-0.0939	0.4280	-0.0145	0.8405
banio	0.1250	0.5949	-0.0682	0.5772
electricidad	-0.0178	-0.0014	0.8024	0.3582
telefono	0.6284	0.0545	0.0356	0.5700
tv_satelital	0.4287	0.2318	0.0003	0.6843
servicio_dom	0.5881	-0.3055	0.2024	0.6558

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.8741	0.7863	0.1735
Factor2	-0.4479	0.5091	0.5166
Factor3	0.1882	-0.3500	0.8385

Al realizar la rotación promax, las cargas factoriales se modificaron para maximizar la posible correlación entre los factores, lo cual para demostrarlo se requiere extraer los factores y realizar un análisis de correlación donde también se solicite la significancia estadística.

El comando para extraer los factores es el mismo que la técnica pasada:

predict *nombre de la nueva variable correspondiente al factor 1 nombre de la nueva variable correspondiente al factor 2…*

Ejemplo:

predict f1 f2 f3

(El presente resultado se muestra con rotación varimax)

```
. predict f1 f2 f3
(option regression assumed; regression scoring)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3
computadora	0.24551	-0.06745	-0.05272
estufa	-0.13511	0.30630	0.13604
lavadora	0.02367	0.17999	-0.00593
refrigerador	-0.05816	0.25258	0.04376
DVD	0.06730	0.11060	-0.01056
televisor	-0.05825	0.03812	0.53175
boiler	0.06725	0.15873	-0.05967
celular	-0.02788	0.24498	-0.09900
microondas	0.16290	0.02838	-0.06805
tostador	0.26010	-0.21255	0.08017
internet	0.25661	-0.07345	-0.04084
agua_entub~a	-0.09030	0.21939	-0.04313
banio	-0.03768	0.27114	-0.09623
electricidad	0.00722	-0.14048	0.67381
telefono	0.21122	-0.08493	0.03629
tv_satelital	0.11757	0.03781	-0.00878
servicio_dom	0.24833	-0.27720	0.20051

Coeficientes estandarizados de las combinaciones lineales de cada uno de los factores extraídos.

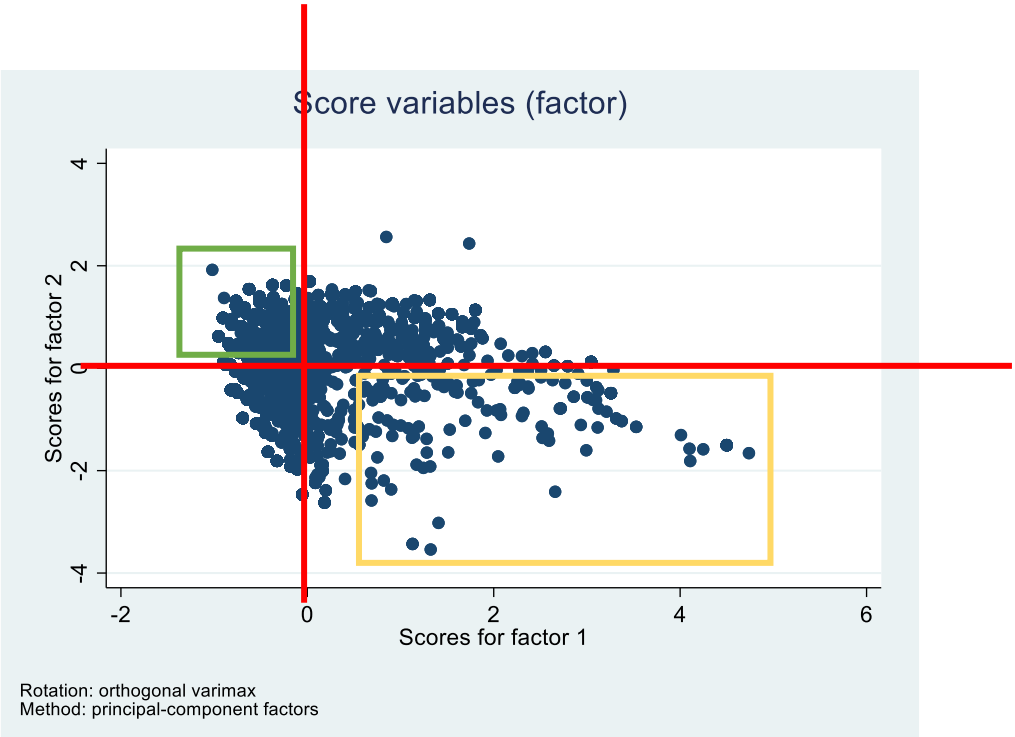
Scores

El método para determinar estos scores es por medio de una regresión, por lo que estos son también los fitted values (valores ajustados)

Al ejecutar esta extracción también se hace el cálculo de scores lo que hace posible la adquisición de más información gracias a esta ponderación. Una forma de visualización útil para observar cómo es el comportamiento de las observaciones en los factores es un diagrama de dispersión, que se puede hacer por medio del comando **scoreplot**

scoreplot

Esta acción se ejecuta después de la extracción de los factores, ya que no acepta lista de variables después del comando. En automático arroja la gráfica entre el factor 1 y 2.

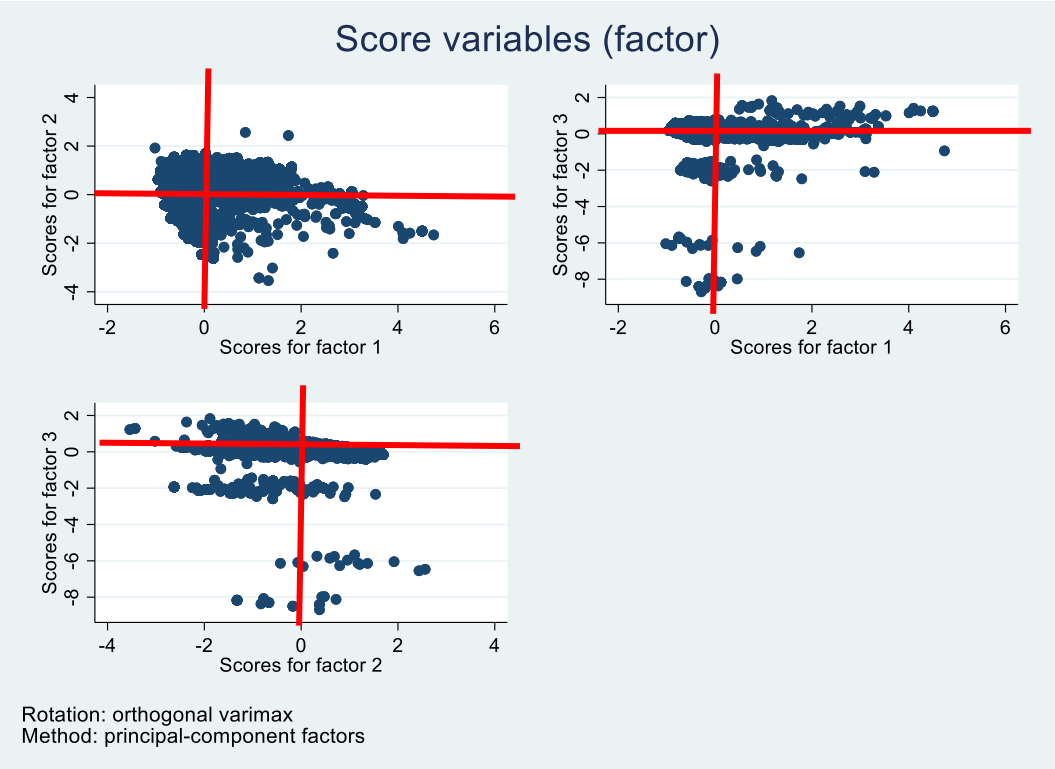


- La gráfica muestra los scores del factor 1 contra los scores del factor 2, o dicho con los conceptos definidos, los scores del lujo contra los de primera necesidad.
- En el encuentro entre el cruce de las medias de 0, se distingue que hay mayor concentración de scores, visualizando un posible equilibrio entre el factor que habla sobre lujo y el factor de activos de primera necesidad.
- También resalta que existen observaciones que se encuentran en una riqueza asociada al lujo con altos valores de scores, es decir con mucho lujo, pero faltantes de riqueza básica o los activos de primera necesida
- Por el otro lado, hay una alta concentración de observaciones en el área ascendente, a lo que se refiere de riqueza de primera necesidad, pero con pocos lujos al ubicarse por debajo de la media de forma vertical y ascendente de forma horizontal.

El comando para visualizar las gráficas de los 3 factores resultantes en este ejemplo es:

scoreplot, factor(3) combined

El número que está dentro del paréntesis puede cambiarse de acuerdo al número de gráficas que se requiera hacer.



El segundo método que se explicará es el **análisis factorial por factor principal**, que a diferencia del método anterior, el mayor interés de estos resultados se concentra en un solo factor principal que permite realizar índices para explicar el fenómeno, de una forma muy similar al análisis de componentes principales, ponderando por medio de scores en una escala que va de lo más alto a lo más bajo.

El comando para ejecutar este análisis es concretamente **factor**:

factor *listado de variables que componen al modelo*

Ejemplo:

factor *computadora - servicio_dom*

```
. factor computadora - servicio_dom
(obs=2,960)

Factor analysis/correlation                 Number of obs    =      2,960
    Method: principal factors               Retained factors =          7
    Rotation: (unrotated)                   Number of params =         98
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.22697	3.41956	0.9260	0.9260
Factor2	0.80741	0.57164	0.1769	1.1029
Factor3	0.23577	0.04478	0.0516	1.1545
Factor4	0.19099	0.06460	0.0418	1.1964
Factor5	0.12639	0.06239	0.0277	1.2240
Factor6	0.06400	0.05723	0.0140	1.2381
Factor7	0.00677	0.02138	0.0015	1.2395
Factor8	-0.01460	0.02187	-0.0032	1.2363
Factor9	-0.03647	0.02030	-0.0080	1.2284
Factor10	-0.05677	0.02670	-0.0124	1.2159
Factor11	-0.08347	0.02525	-0.0183	1.1976
Factor12	-0.10872	0.01312	-0.0238	1.1738
Factor13	-0.12184	0.01237	-0.0267	1.1471
Factor14	-0.13421	0.00761	-0.0294	1.1177
Factor15	-0.14182	0.03480	-0.0311	1.0867
Factor16	-0.17662	0.04232	-0.0387	1.0480
Factor17	-0.21894	.	-0.0480	1.0000

```
LR test: independent vs. saturated: chi2(136) = 1.2e+04 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Uniqueness
computadora	0.7012	-0.3000	0.0635	0.0255	-0.1009	0.0703	-0.0334	0.3974
estufa	0.4161	0.3901	0.0569	-0.0130	-0.0172	0.0320	-0.0117	0.6698
lavadora	0.5547	0.1443	-0.0631	0.0032	-0.0325	-0.0524	0.0073	0.6636
refrigerador	0.4856	0.2648	0.0008	-0.0065	-0.0662	-0.0934	0.0033	0.6809
DVD	0.5195	0.0660	-0.1639	0.1251	-0.0514	-0.0150	0.0072	0.6803
televisor	0.2443	0.3292	0.1663	0.1913	-0.0091	-0.0064	-0.0027	0.7676
boiler	0.6286	0.0745	-0.0334	-0.1454	0.1069	-0.0221	0.0089	0.5651
celular	0.4740	0.1594	-0.1714	0.0247	-0.0722	0.0889	-0.0003	0.7068
microondas	0.6168	-0.0942	-0.1437	0.0422	0.0519	-0.0535	-0.0225	0.5821
tostador	0.4073	-0.2147	-0.0958	0.1427	0.1587	0.0029	-0.0149	0.7330
internet	0.7458	-0.3259	0.1600	-0.0337	-0.0958	0.0364	-0.0006	0.3003
agua_entub~a	0.2181	0.1518	0.0517	-0.1490	0.1006	0.1330	0.0042	0.8767
banio	0.5147	0.1890	-0.0369	-0.1737	0.0650	-0.0075	-0.0160	0.6632
electricidad	0.1040	0.2213	0.1985	0.1510	0.0926	0.0294	-0.0103	0.8685
telefono	0.5604	-0.1814	0.2014	-0.0754	0.0371	-0.1102	0.0112	0.5932
tv_satelital	0.5055	-0.0309	-0.0043	0.0358	-0.0610	0.0456	0.0578	0.7331
servicio_dom	0.2715	-0.1666	-0.0078	0.1037	0.1614	0.0296	0.0273	0.8600

Número de observaciones tomadas para el análisis

Número de factores retenidos

Número de parámetros que se requieren para estimar

Factor útil por arrojar un eigenvalue mayor a 1.

El único factor.

Porcentaje correspondiente de variabilidad conjunta de cada factor

Unicidad

Lo que no se explica de cada variable

- Al aplicar el método se retiene 7 factores y se obtiene una unicidad mayor (mayor variabilidad NO explicada) porque todos ellos son positivos en eigenvalue, pero sólo el primer factor es el de interés al concentrar el .9260 de la variabilidad / varianza explicada (92.6%). Sin embargo, al retener tantos factores representa una desventaja sobre el análisis anterior por componentes principales.
- La razón por la cual el primer factor explica tanta variabilidad es por un truco técnico al no restringir la varianza, la maximiza de forma gradual en cada factor (la máxima en el primer factor, luego un poco menos en el segundo y así sucesivamente), haciendo que se concentre la variabilidad al máximo en el primer factor, hasta donde le sea posible.
- Es importante resaltar que el ajuste de variabilidad llega hasta 1 por la naturaleza de la matriz de correlaciones, lo que hace que la secuencia de los factores posteriores al primero haya disminución de variabilidad debido a este ajuste.

El paso siguiente de la obtención de los factores con sus respectivos eigenvalues y cargas factoriales es ejecutar el comando de rotación Varimax y seleccionar las cargas factoriales que correspondan a cada factor por medio del comando **rot, blank(**.5**)**:

Ejemplo:

rot, blank(.5**)**

```
. rot, blank(.5)

Factor analysis/correlation                    Number of obs    =      2,960
    Method: principal factors                  Retained factors =          7
    Rotation: orthogonal varimax (Kaiser off)  Number of params =         98
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	3.04108	1.30237	0.6662	0.6662
Factor2	1.73871	1.38432	0.3809	1.0471
Factor3	0.35439	0.07755	0.0776	1.1247
Factor4	0.27684	0.15062	0.0606	1.1854
Factor5	0.12623	0.01359	0.0277	1.2130
Factor6	0.11264	0.10424	0.0247	1.2377
Factor7	0.00840	.	0.0018	1.2395

Se destaca que la covariabilidad se ajustó con la rotación del primer factor.

Pero el segundo factor rebasa a 1.

No se ajusta a 1.

```
LR test: independent vs. saturated: chi2(136) = 1.2e+04 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Uniqueness
computadora	0.7579							0.3974
estufa								0.6698
lavadora								0.6636
refrigerador								0.6809
DVD								0.6803
televisor								0.7676
boiler								0.5651
celular								0.7068
microondas	0.5118							0.5821
tostador								0.7330
internet	0.8263							0.3003
agua_entub~a								0.8767
banio								0.6632
electricidad								0.8685
telefono	0.5847							0.5932
tv_satelital								0.7331
servicio_dom								0.8600

Con la indicación blank(.5) desaparecieron las cargas factoriales que son distintas o menores a .5

Unicidad

Lo que no se explica de cada variable

Factor rotation matrix

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Factor1	0.8153	0.5479	0.1072	0.1507	0.0223	0.0025	0.0170
Factor2	-0.5197	0.7356	0.4018	-0.1551	0.0429	0.0372	-0.0013
Factor3	0.2108	-0.2933	0.6480	-0.4774	-0.4112	0.2294	-0.0055
Factor4	-0.0167	-0.2493	0.6287	0.5003	0.3499	-0.4099	0.0398
Factor5	-0.1339	0.0152	0.0665	0.6868	-0.4787	0.5254	-0.0241
Factor6	0.0461	-0.1012	0.0836	-0.0575	0.6892	0.7083	0.0229
Factor7	-0.0170	0.0026	-0.0232	-0.0074	-0.0439	0.0141	0.9985

Con la finalidad de explorar si se pueden integrar las más variables posibles al primer factor, se disminuye el valor de la carga factorial al comando:

rot, blank(.4**)**

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Uniqueness
computadora	0.7579							0.3974
estufa		0.4980						0.6698
lavadora		0.4327						0.6636
refrigerador		0.4707						0.6809
DVD								0.6803
televisor								0.7676
boiler	0.4537	0.4491						0.5651
celular		0.4109						0.7068
microondas	0.5118							0.5821
tostador	0.4002							0.7330
internet	0.8263							0.3003
agua_entub~a								0.8767
banio		0.4769						0.6632
electricidad								0.8685
telefono	0.5847							0.5932
tv_satelital	0.4360							0.7331
servicio_dom								0.8600

(blanks represent abs(loading)<.4)

- Al disminuir la carga factorial, se agrega un segundo factor implicando que se tenga que extraer e interpretar, como el proceso al método anterior, pero esta técnica su naturaleza no es clasificar, sino realizar índices. Además, el modelo arroja que deben contemplarse el resultado completo de 7 factores y de acuerdo a los principios de análisis multivariante, que es explicar al fenómeno de estudio por medio de una reducción de variables, no es eficiente que se busque interpretar todos los factores.

Como se había mencionado con anterioridad, lo importante de este análisis es enfocarse en el primar factor resultante, que en el caso del ejemplo se contemplaran 2 debido a buscar contemplar la mayor cantidad de variables posibles, por lo que el siguiente paso es extraer los factores de interés y hacer la ponderación de scores, usando nuevamente el comando **predict**:

Ejemplo:

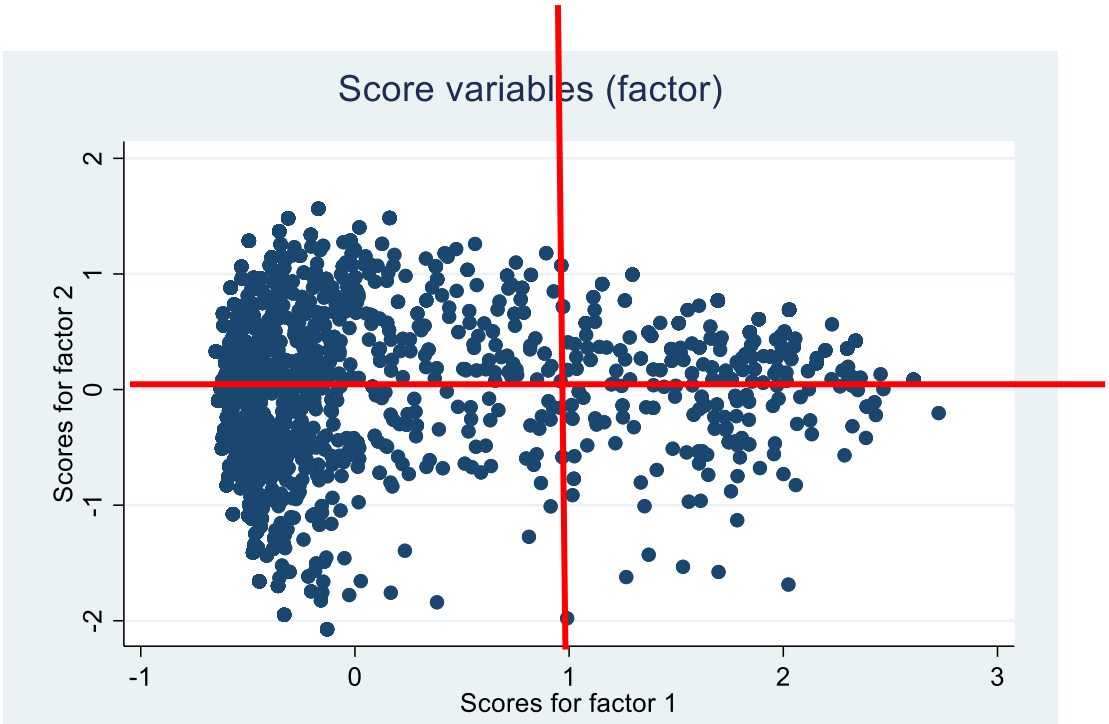
predict *g1 g2*

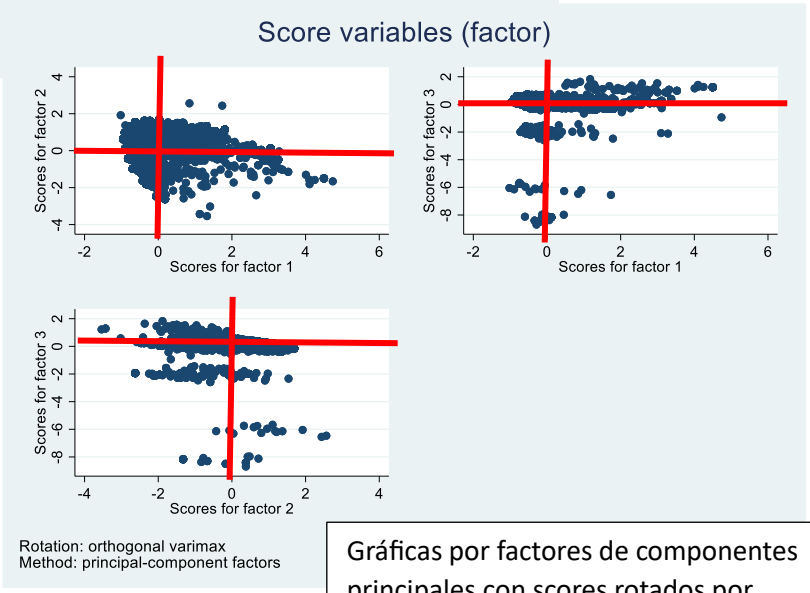
```
(option regression assumed; regression scoring)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
computadora	0.28409	-0.11809	-0.00998	-0.03018	0.15749	-0.01083	-0.05972
estufa	-0.04501	0.21317	0.15100	-0.07892	0.02313	0.05003	-0.01391
lavadora	0.01657	0.16307	0.00709	0.01549	0.00743	-0.08059	0.01340
refrigerador	-0.00826	0.18960	0.06693	-0.06434	-0.03410	-0.11240	0.00528
DVD	0.01702	0.10837	0.00318	0.11980	0.13501	-0.13757	0.01979
televisor	-0.02842	0.07049	0.29842	-0.01007	0.00919	-0.04080	0.00345
boiler	0.04090	0.19381	-0.09338	0.05025	-0.12962	0.12661	0.00571
celular	-0.01114	0.15822	-0.03764	0.03339	0.19309	-0.01166	0.00848
microondas	0.07467	0.07951	-0.09135	0.20317	0.00271	-0.08307	-0.03003
tostador	0.06508	-0.05633	0.00188	0.26889	0.00027	0.00968	-0.01099
internet	0.44416	-0.19191	0.02185	-0.20455	-0.01993	0.03723	0.01992
agua_entub~a	-0.00962	0.08316	0.00361	-0.03467	-0.01101	0.20751	0.00055
banio	-0.00504	0.20977	-0.07413	-0.02851	-0.08011	0.10777	-0.02902
electricidad	-0.02165	0.01360	0.24897	0.02843	-0.03424	0.05559	-0.00844
telefono	0.14317	-0.03413	0.04657	-0.04738	-0.26190	-0.00221	0.00966
tv_satelital	0.06602	0.03816	0.02326	0.00075	0.08342	-0.01255	0.08173
servicio_dom	0.04630	-0.05738	0.03484	0.17917	-0.02097	0.05934	0.03161

Obtenido los scores también se puede hacer una exploración de estos resultados por medio de un diagrama de dispersión:

scoreplot





Gráficas por factores de componentes principales con scores rotados por Varimax.

Al hacer la comparativa de gráficos resultantes por ambos métodos del análisis factorial se observa que son similares. Sin embargo, entre los factores arrojados por factor principal sí existe correlación porque maximiza la variabilidad en general, tratando al no ajustarse a 1 provocando ligeros cambios en los eigenvectors.

Conclusión del capítulo Análisis Factorial

De acuerdo a la matriz reflejada en el diagrama de dispersión entre el factor 1 y 2, extraídos por la técnica de factor principal, se sabe que el factor 1 está compuesto por la variable computadora, boiler, microondas, tostador, internet, teléfono y tv satelital; mientras que el factor 2 se encuentra los activos de estufa, lavadora, refrigerador, celular y baño. Estos resultados son cercanos a los obtenidos con la técnica anterior, lo que permite que más o menos se pueda llegar a la misma conclusión sobre identificar la riqueza indirecta, aunque esto es debido a que se está forzando la interpretación. Es importante resaltar que la naturaleza del método por factor principal no es clasificar conceptos, si no es para generar índices para ordenar las observaciones de acuerdo a los scores obtenidos.

Para una mayor referencia del contenido de este capítulo ver:

Gudmundsson, G. (1977). Multivariate analysis of economic variables. "Journal of the Royal Statistical Society: Series C (Applied Statistics), 26"(1), 48–59.

Esta técnica es especialmente útil en contextos donde la complejidad y la multidimensionalidad de los datos requieren un análisis más profundo, ya que le método obtiene las varianzas canónicas, de donde resultan los eigenvalues y los eigenvectors, siendo la base fundamental sobre las cuales se estiman las combinaciones lineales (del mismo modo que componentes principales), indicando la correlación entre los dos conjuntos de variables que componen el modelo.

Es importante que el cruce que hace la técnica entre los grupos de variables se haga ante una explícita dependencia entre las variables porque de lo contrario la interpretación de resultados sería difícil y no congruente con la técnica. Al diseñar el modelo se debe tener claridad de cuáles son las variables dependientes y que están sean explicadas de forma coherente por las variables independientes.

El resultado de este método obtiene una variable que describe la combinación lineal que se va a conformar por las variables dependientes y por las variables independientes. De tal forma que por un lado, se va a medir la variabilidad canónica en "X" y la variabilidad canónica en "Y".

Algunos aspectos que se deben cumplir para poder realizar un análisis por el método de relación canónica son:

- Se requiere que el número de variables dependientes sea al menos igual al número de variables independientes.
- No corre si no se tiene al menos 10 observaciones.
- Se recomienda que el KMO sea más o menos de .8
- La correlación canónica al elevarse al cuadrado resulta un coeficiente de determinación, el cual se toma como el valor de la bondad de ajuste del modelo (R2), el cual indica qué porcentaje de variabilidad conjunta explican de las variables dependientes.
- Algo que infla la correlación lineal es la multicolinealidad. Es por ello que debe aplicarse una prueba que indique la existencia de este problema, así evitar un sesgo en la combinación canónica y un resultado erróneo. En caso de tener presencia de multicolinealidad, esta se soluciona omitiendo la variable que pudiera causar el problema.
- Una correlación canónica por arriba de .3 en términos absolutos puede considerarse relevante.

- Si correlación canónica se acerca 1, se sospecha de multicolinealidad provocando un sesgo.

El data set empleado para mostrar el funcionamiento y resultados del método se le denomina *WAGE2*, los cuales contiene datos de acceso libre con registro de remuneraciones mensuales, educación, varias variables demográficas y puntuaciones de IQ de 935 hombres en 1980. El planteamiento que se busca explorar es la posible relación entre la remuneración económica de estos individuos y su nivel de preparación e influencia intelectual. La estructura del conjunto de datos es:

Variable	Descripción
age	Ingresos mensuales
hours	Promedio de horas semanales
IQ	Puntuación de IQ
KWW	Conocimiento de la puntuación del trabajo mundial
educ	Años de educación
exper	Años de experiencia laboral
tenure	Años con el empleador actual
age	Edad en años
married	=1 si está casado
black	=1 si es negro
south	=1 si vive en el sur
urban	=1 si vive en SMSA
sibs	Numero de hermanos
brthord	Orden de nacimiento
meduc	Educación de la madre
feduc	Educación del padre
lwage	Logaritmo natural del salario

Los pasos generales a seguir para realizar el análisis de correlación canónica son:

1. Ejecutar el comando de análisis al modelo diseñado.
2. Ejecutar el comando al modelo diseñado con la instrucción de cálculo de variables estandarizadas.
3. Interpretación de la correlación canónica.
4. Búsqueda de multicolinealidad por medio del indicador VIF.
5. Realizar el test KMO.

El comando que se utiliza para hacer aplicación de esta técnica es:

canon (variables dependientes del modelo**) (**variables independientes del modelo**)**

Ejemplo:

canon (wage hours KWW **) (** educ meduc feduc IQ **)**

```
. canon ( wage hours KWW ) ( educ meduc feduc IQ )

Canonical correlation analysis                    Number of obs =       722

Raw coefficients for the first variable set

                        1          2          3
-------------------------------------------------
        wage     0.0011     0.0016    -0.0017
       hours     0.0160     0.1051     0.0885
         KWW     0.0954    -0.0829     0.0561
-------------------------------------------------

Raw coefficients for the second variable set

                        1          2          3
-------------------------------------------------
        educ     0.2245     0.1210     0.2748
       meduc     0.0599     0.1475     0.2681
       feduc     0.0197     0.1692    -0.3431
          IQ     0.0343    -0.0615    -0.0263
-------------------------------------------------

Canonical correlations:
  0.5333  0.0809  0.0493

Tests of significance of all canonical correlations

                       Statistic      df1      df2          F     Prob>F
          Wilks' lambda     .70922       12     1892    21.8643     0.0000 a
         Pillai's trace    .293344       12     2151    19.4269     0.0000 a
Lawley-Hotelling trace      .40639       12     2141    24.1689     0.0000 a
    Roy's largest root     .397362        4      717    71.2272     0.0000 u

                                e = exact, a = approximate, u = upper bound on F
```

Coeficientes fila del segundo conjunto de variables

Combinaciones lineales para el primer grupo, es decir para las variables dependientes.

Coeficientes de combinaciones lineales

Coeficientes fila del primer conjunto de

Combinaciones lineales para el segundo grupo, que son las variables independientes.

Entre los primeros resultados sin estandarizar se puede observar los coeficientes de las combinaciones lineales, los cuales funcionan para ayudarnos a sacar la correlación canónica. Estos se interpretan como se hace en un análisis de regresión lineal.

Un mejor resultado se obtiene con unidades estandarizadas y para ello se ejecuta el comando:

canon (*variables dependientes del modelo***) (***variables independientes del modelo***) ,stdc**

Ejemplo:

canon (*wage hours KWW* **) (** *educ meduc feduc IQ* **), stdc**

El modelo hace un planteamiento sobre el desempeño laboral, que está en función o depende de los antecedentes de educación, dando como resultado:

```
. canon ( wage hours KWW ) ( educ meduc feduc IQ ), stdc

Canonical correlation analysis                    Number of obs =       722

Standardized coefficients for the first variable set
```

	1	2	3
wage	0.4690	0.6344	-0.7052
hours	0.1169	0.7657	0.6448
KWW	0.7324	-0.6369	0.4309

Combinaciones lineales para el primer grupo, es decir para las variables dependientes

Coeficientes estandarizados de combinaciones lineales

Denominadas con la letra "u"

```
Standardized coefficients for the second variable set
```

	1	2	3
educ	0.5021	0.2707	0.6146
meduc	0.1693	0.4173	0.7585
feduc	0.0651	0.5593	-1.1342
IQ	0.5073	-0.9089	-0.3889

Combinaciones lineales para el segundo grupo, que son las variables independientes.

Coeficientes estandarizados de combinaciones lineales

Denominadas con la letra "v"

```
Canonical correlations:
 0.5333  0.0809  0.0493
```

CorrCan 1. CorrCan 2 CorrCan 3

Correlaciones canónicas resultantes de las combinaciones lineales.

```
Tests of significance of all canonical correlations

                    Statistic     df1      df2          F       Prob>F
        Wilks' lambda     .70922      12     1892    21.8643    0.0000 a
       Pillai's trace    .293344      12     2151    19.4269    0.0000 a
Lawley-Hotelling trace     .40639      12     2141    24.1689    0.0000 a
    Roy's largest root    .397362       4      717    71.2272    0.0000 u

                e = exact, a = approximate, u = upper bound on F
```

P value para probar la hipótesis nula de Rc = 0

- De acuerdo a los coeficientes estandarizados, se puede decir que, si wage se incrementa en una unidad, la variabilidad canónica aumenta en .4691 desviaciones estándar.
- La segunda combinación lineal absorbe la covariabilidad de la combinación del grupo de variables de manera ortogonal de la primera combinación. De tal forma que las combinaciones lineales no tengan correlación entre sí para cada grupo, pero entre los

grupos sí. Es decir, que esa combinación lineal 1 no tiene correlación con la combinación lineal 2, ni la 1 con la 3, ni la 2 con la 3, ni la 3 con la 2, etc. Explican cosas diferentes.

- La primera correlación canónica es: correlación canónica 1, de la combinación 1 de las variables dependientes con 1 de la combinación lineal de las variables independientes, que es igual a .5333, y así sucesivamente.
- Se destaca que las correlaciones canónicas 2 y 3 son muy bajas, prácticamente 0, porque la primera es la que maximiza la correlación entre cada una de las combinaciones lineales.

La HIPÓTESIS NULA dice que no hay correlación o que la correlación canónica es igual a 0.

Después de haber obtenido la correlación canónica de la primera combinación lineal, que es quien maximiza la covariabilidad, se puede complementar la interpretación con el cálculo de la bondad de ajuste canónica. Para ello se eleva la correlación canónica al cuadrado

Ejemplo:

De acuerdo a los valores encontrados en los resultados obtenidos anteriormente, el valor de la correlación canónica es .53332. Para elevar al cuadrado un valor se aplica la siguiente instrucción en Stata:

dis .5333^**2** →

```
. dis (0.5333^2)
.28440889
```

El resultado de .28440889 es el valor de ajuste del modelo, el cual nos dice en términos de porcentaje cuánto es lo que explica las variables independientes sobre las variables dependientes. Tomando en cuenta la situación planteada, se puede observar que la historia académica de las personas explica un .2844 (28.40%) del desempeño laboral. Además, se demuestra que el desempeño laboral cuenta con una fuerte correlación lineal con el historial académico hasta con un valor de .5333 de correlación canónica.

Como se mencionó anterioridad, es importante descartar la existencia de multicolinealidad en el resultado obtenido en el análisis y para ello se ejecuta la instrucción que calcula el indicador Variance inflation factor (VIF). Para ello se requiere extraer la covarianza canónica de cada grupo de variables, que de acuerdo a los resultados obtenidos, la matriz del grupo de variables independientes, el programa los interpreta con la letra "v" y la matriz del grupo de variables dependientes con la letra "u". El comando para extraer la covarianza canónica es:

predict *nombre de la nueva variable* , **u corr(1)**

Donde:

- **predict** es el comando para extraer e indicar que se genere una nueva variable
- Nombre de la nueva variable que indica la correlación canónica 1 de las variables dependientes (por eso se agrega la "Y" al nombre de la variable).
- **"u"** señala que extraiga la información de la matriz denominada "u", que son las variables dependientes.
- **corr(1)** significa que se está solicitando la correlación 1 (la combinación lineal 1).

  Ejemplo:

  predict *CVY1*, **u corr(1)**

  Y para observar el resultado se ejecuta el comando **sum** con la nueva variable generada:

Ajustes de las combinaciones lineales

```
. sum RC1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         RC1 |       935     5.21534    .9996769   2.555416   9.349903
```

La covariabilidad que es reflejada por la desviación estándar, que al ser maximizada en la primera combinación lineal se acerca lo más posible a 1

Se repite el proceso de extracción, pero ahora indicando la matriz del grupo de variables independientes, cambiando en el comando la letra "u" por "v" y cambiando el nombre de la nueva variable, por ejemplo CVX1.

Después de haber extraído la covarianza canónica de cada grupo de variables, para hacer la prueba de multicolinealidad se requiere hacer una regresión lineal con las variables extraídas y el grupo de variables que le corresponden, variable de la matriz del grupo de variables independientes con este grupo de variables, igualmente para las variables dependientes, luego se calcula el VIF:

Ejemplo:

Variables dependientes

reg *CVY1 wage hours KWW*

```
. reg CVY1 wage hours KWW
```

Source	SS	df	MS	Number of obs	=	935
				F(3, 931)	>	99999.00
Model	933.396564	3	311.132188	Prob > F	=	0.0000
Residual	1.6200e-11	931	1.7401e-14	R-squared	=	1.0000
				Adj R-squared	=	1.0000
Total	933.396564	934	.999353923	Root MSE	=	1.3e-07

Bondad de ajuste 1

CVY1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wage	.0011494	1.13e-11	1.0e+08	0.000	.0011494	.0011494
hours	.0160495	6.02e-10	2.7e+07	0.000	.0160495	.0160495
KWW	.0953793	6.02e-10	1.6e+08	0.000	.0953793	.0953793
_cons	-6.14e-08	3.21e-08	-1.91	0.056	-1.25e-07	1.65e-09

vif

Variables independientes

```
. vif
```

Variable	VIF	1/VIF
KWW	1.14	0.879951
wage	1.12	0.891434
hours	1.02	0.984594
Mean VIF	1.09	

No hay multicolinealidad al no obtener un valor igual o por arriba a 10.

reg *CVX1 educ meduc feduc IQ*

```
. reg CVX1 educ meduc feduc IQ
```

Source	SS	df	MS	Number of obs	=	722
				F(4, 717)	>	99999.00
Model	720.999991	4	180.249998	Prob > F	=	0.0000
Residual	2.4960e-11	717	3.4811e-14	R-squared	=	1.0000
				Adj R-squared	=	1.0000
Total	720.999991	721	.999999987	Root MSE	=	1.9e-07

Bondad de ajuste 1

CVX1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.2244891	3.89e-09	5.8e+07	0.000	.2244891	.2244891
meduc	.0598676	3.07e-09	1.9e+07	0.000	.0598675	.0598676
feduc	.0196802	2.69e-09	7.3e+06	0.000	.0196802	.0196802
IQ	.0343395	5.68e-10	6.0e+07	0.000	.0343395	.0343395
_cons	4.92e-08	5.30e-08	0.93	0.353	-5.48e-08	1.53e-07

vif

```
. vif
```

Variable	VIF	1/VIF
feduc	1.64	0.611493
educ	1.57	0.638544
meduc	1.56	0.640075
IQ	1.46	0.686385
Mean VIF	1.56	

No hay multicolinealidad al no obtener un valor igual o por arriba a 10.

factortest *las variables que integran el modelo*

Ejemplo:

factortest *wage hours KWW educ meduc feduc IQ*

```
. factortest wage hours KWW educ meduc feduc IQ

Determinant of the correlation matrix
Det                =     0.228


Bartlett test of sphericity

Chi-square         =          1060.184
Degrees of freedom =                21
p-value            =             0.000
H0: variables are not intercorrelated


Kaiser-Meyer-Olkin Measure of Sampling Adequacy
KMO                =     0.781
```

Valor del KMO que para esta técnica se recomienda que sea por arriba de .8, pero no es algo determinante

Conclusión del capítulo Correlación Canónica

En el contexto específico analizado, se observa que la historia académica de las personas explica el 28.44% del desempeño laboral. Además, el análisis muestra una fuerte correlación lineal entre el desempeño laboral y el historial académico, con un valor de correlación canónica de .5333.

Con dichos resultados se destaca la importancia de la educación y el historial académico en el desempeño laboral. Estos resultados pueden ser esenciales para la formulación de políticas públicas orientadas a mejorar los resultados educativos y laborales. Al entender la influencia significativa del nivel educativo y el conocimiento en el mercado laboral, se pueden diseñar intervenciones más efectivas que promuevan la educación como una herramienta clave para mejorar las condiciones laborales y, por ende, el bienestar socioeconómico de la población.

Para una mayor referencia del contenido de este capítulo ver:

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (Pearson, Ed.; 6 th).

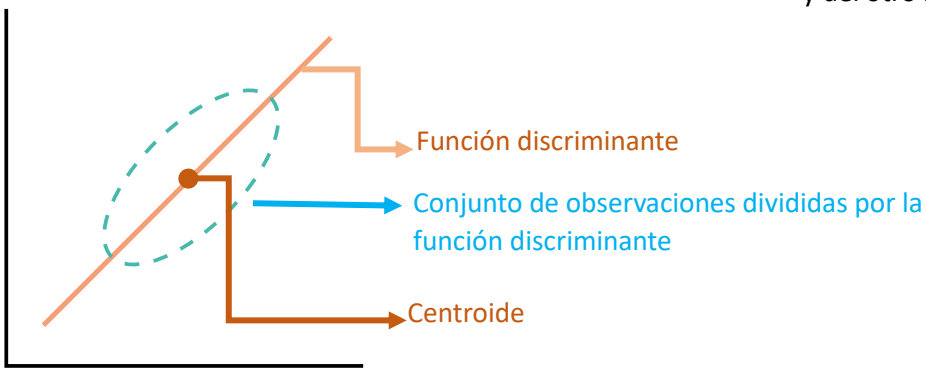
La idea general del uso de esta técnica multivariante, como bien dice su nombre, es separar las observaciones en grupos y conocer a estos para ver la posibilidad de sacar una observación de ese algún grupo y meterlo a otro grupo. Se discrimina observación por observación de acuerdo a los predictores que quien realiza el análisis decide.

En la práctica, este método se puede ejecutar para clasificar a las personas en grupos socioeconómicos basándose en sus ingresos, educación y otros factores demográficos. Lo importante aquí es que se debe contrastar variables que cuenten con en dos o más categorías diferentes contra variables métricas.

Una peculiaridad que se resalta es que esta técnica es empleada para hacer predicciones a priori (calculando estadísticas antes de que suceda el evento). Se busca obtener el dato sobre la pertenencia de ciertas observaciones a un grupo, a partir de un conjunto de predictores continuos, que son las variables que explican a los grupos.

La forma visual de representar esta técnica se hace por medio del siguiente diagrama:

Por medio de una función discriminante se dibuja una recta que separa a los grupos. De un lado de la recta hay un grupo y del otro lado hay otro grupo.



Se sabe cuántas funciones lineales discriminantes calcular con esta regla: Del número de grupos o predictores, el valor mínimo (el más chico, ya sea del grupo o de los predictores) menos 1. Cada función discriminante es ortogonal a la anterior y el número de dimensiones (funciones discriminantes) es igual a g - 1 o p -1, lo que sea menor.

- Cuando tenemos muchos grupos y un número menor de predictores, es favorable para el análisis.
- El problema es cuando tenemos muchos predictores y pocos grupos.

Ejemplo:

Cantidad de grupos	Cantidad de predictores	Resultado de funciones discriminantes
2 grupos	4 predictores	Función discrimínate 1
3 grupos	4 predictores	Función discrimínate 2
4 grupos	2 predictores	Función discrimínate 1

Predictores:

Son aquellas variables, numéricas o categóricas, que definen las características de los grupos o que pueden ayudar a identificarlas. El análisis es altamente sensible a los predictores, así que si se agregan o eliminan van a alterar el resultado, por lo que se debe tener claridad de cómo se integran los grupos y establecer previamente como se pueden describir. Es por esta razón que es necesario poner atención en que variables se determinan como predictores, siendo el elemento más importante de este método y de hallarse funciones no significativamente estadísticas, no vale la pena seguir con el análisis, porque esos predictores no logran explicar los grupos.

Con los resultados obtenidos se pueden validar estos grupos por medio de los parámetros poblacionales sobre los muestrales, se hace uso del análisis de parámetros muestrales para generalizarlos a los poblacionales.

La forma de como Stata presenta los resultados al ejecutar el comando de la técnica es por medio de una tabla (*clasification table*) y nos dice que observaciones entran y salen de los grupos. Se recomienda que los grupos estén muy bien definidos, de lo contrario no se tendrá una buena discriminación. Para determinar de forma correcta los grupos es obtenerlo de una sola muestra (de un solo conjunto de datos), así esta única selección de datos será partida, ya que, si se juntan varias muestras, metodológicamente hace que el análisis no funcione.

En lo que respecta a los *missing values* posiblemente existentes en el conjunto de datos no representa un problema para la técnica, pero en cuanto a los outliers se recomienda que se eliminen, para que la distribución se acerque a lo normal y se obtenga un mejor rendimiento. Por otro lado, se puede presentar el problema de multicolinealidad, ya que no nos ayuda a discriminar de forma favorable. Entonces se requiere detectarlo y eliminarlo.

El procedimiento en general para realizar un análisis discriminante es:

- Determinar cuáles son las variables que funcionarán para clasificar en grupos y los predictores.
- Ejecutar el comando del análisis.
- Estimar la función discriminante y el grupo discriminado (las clases).
- Calcular los centroides.
- Enlistar las observaciones para identificar cuáles fueron de estas fueron reclasificadas.

El caso de ejemplo a usarse para la explicación del análisis se hace con un conjunto de datos que describe la posición académica y de ingresos de un padre, y a partir de estos datos se desea conocer cuál es la relación de estas variables con el evento de que ellos vivan en una zona conurbada. La variable que clasifica a los grupos está denominada como *zm*, que indica si los entrevistados viven en una zona conurbada o no. Para conocer la descripción de esta variable se corre el comando **tab** *zm*.

```
. tab zm
```

zm	Freq.	Percent	Cum.
0	1,973	65.70	65.70
1	1,030	34.30	100.00
Total	3,003	100.00	

La variable cuenta con 1030 registros de personas que radican en una zona conurbada y otros 1973 que están en una zona NO conurbada. Demostrando que l base de datos este conformada por 3,003 observaciones que se utilizaran como información para previa para logar hacer predicciones de futuros personas que lleguen a establecerse en el lugar del estudio.

Los predictores que se utilizan para el análisis son *ingreso mensual, educación del entrevistado y educación del padre*. Son variables numéricas y de naturaleza categórica. Específicamente, estas últimas son aquellas las que indican el grado de estudios del padre y cuentan con un orden jerárquico, donde el nivel primaria es el más bajo y el posgrado es el más alto. Para conocer el contenido de los predictores se ejecuta el comando **sum** *ing_mens, educ, educ_padre*.

```
. sum ing_mens educ educ_padre
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ing_mens	1,492	3847.7	3934.46	0	70000
educ	2,998	2.366911	1.274529	0	6
educ_padre	2,521	.8659262	1.115312	0	6

El comando para ejecutar el análisis discriminante es **candisc**:

candisc *variables predictores* **, group (***variable que determina los grupos***)**

Ejemplo:

candisc *ing_mens educ educ_padre*, **group (** *zm* **)**

Canonical linear discriminant analysis

Fcn	Canon. Corr.	Eigen-value	Variance Prop.	Variance Cumul.	Like-lihood Ratio	F	df1	df2	Prob>F	
1	0.3288	.121184	1.0000	1.0000	0.8919	50.857	3	1259	0.0000	e

Ho: this and smaller canon. corr. are zero; e = exact F

Estadístico de prueba

P value de significancia estadística

Correlación canónica

Eigen value que explica toda la varianza acumulada

Función lineal discriminante que explica a todos los grupos.

Standardized canonical discriminant function coefficients

(Función discriminante)

	function1
ing_mens	-.3679477
educ	-.2577698
educ_padre	-.6915526

Cargas factoriales y dentro de la ecuación estos serías los coeficientes de la función discriminante

Canonical structure

	function1
ing_mens	-.5773121
educ	-.6742407
educ_padre	-.8875398

Correlación de la función discriminante y los predictores

La ecuación de la función discriminante es:

Di = -.36 ing_mens -.2577 educ -.6915 educ_padre

Al tener coeficientes negativos nos dice que está inversamente relacionada.

Group means on canonical variables

zm	function1
0	.2815792
1	-.4296898

Centroides

Son las medias de la función discriminante por cada grupo. Recuerda que la función pasa por todos los grupos y de acuerdo a los centroides es como va separando a los grupos.

Resubstitution classification summary

Key
Number Percent

True zm	Classified 0	1	Total
0	564	199	763
	73.92	26.08	100.00
1	228	272	500
	45.60	54.40	100.00
Total	792	471	1,263
	62.71	37.29	100.00
Priors	0.5000	0.5000	

Valores absolutos de observaciones pertenecientes al grupo verdadero

En la muestra hay 763 observaciones en el grupo de zona no metropolitana, de los cuales 564 están bien clasificadas y 199 no están correctamente clasificadas, por lo que son las observaciones discriminadas (de las 763). Es por ello que las coloca en el grupo de zona metropolitana.

Porcentajes

En la muestra hay 500 observaciones en el grupo de zona metropolitana, de los cuales 272 están bien clasificadas y 228 no, por lo que estas son discriminadas para pasar a la zona no metropolitana.

Probabilidad a priori: Se calcula bajo criterios no determinísticos. Se puede determinar cuál es la probabilidad de que las observaciones se inclinen a un

Clasificación resultante por el análisis:

792 observaciones corresponden al grupo de zona no metropolitana.

471 observaciones son de la zona metropolitana.

Con un p-value de 0.0000 se rechaza la hipótesis nula indicando que la relación canónica o la función discriminante es estadísticamente significativa.

- En el caso de requerirse presentar el modelo de la correlación canónica del ejemplo, quedaría primero los predictores y luego los grupos, porque se requiere explicar los grupos en función de los predictores. Haciendo que las variables dependientes fueran los grupos y las variables independientes los predictores (**canon (*zm*) (*ing_mens educ educ_padre*)**).

Es necesario comprender plenamente la tabla de grupo verdadero vs grupo discriminado, ya que es la información que se busca al aplicar esta técnica de análisis.

- En la lectura horizontal de la tabla, la columna de totales muestra el resultado total de observaciones correspondientes a zona metropolitana (1) con 763, y zona no metropolitana (0) con 500. Estos datos son los registros correspondientes al grupo verdadero (True zm).
- Los resultados en vertical corresponden a las observaciones que fueron reclasificadas, es decir aquellas que estaban en un grupo, pero dadas las características acordes en el modelo, deberán pertenecer al otro grupo.
- Del grupo real con 763 observaciones, el análisis ubica 564 correctamente dentro del grupo de zona no metropolitana (correctamente clasificadas), equivalente al 73.92%. Luego 199 fueron reclasificadas (discriminadas), reubicándolas en el grupo de zona metropolitana, correspondiendo al 26.08%.
- Con respecto a la probabilidad sobre las observaciones, se observa que se calculó que el 50% de las observaciones pertenezcan a la zona no metropolitana el 50% pertenezca a la zona metropolitana.

El siguiente paso es estimar la función discriminante y el grupo discriminando (las clases), por medio del siguiente comando:

predict *nombre de la nueva variable*, **dscore**

Ejemplo:

predict *df1*, **dscore**

```
. predict df1, dscore
(1,740 missing values generated)
```

El comando genera una nueva variable con los valores calculados por la función discriminante, formando los nuevos grupos.

Conocer las estadísticas básicas de esta nueva variable se pueden explorar con un **sum**:

sum *df1*

```
. sum df1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
df1	1,263	7.83e-10	1.05844	-7.653388	1.467964

True zm	Classified 0	1	Total
0	564 73.92	199 26.08	763 100.00
1	228 45.60	272 54.40	500 100.00
Total	792 62.71	471 37.29	1,263 100.00
Priors	0.5000	0.5000	

Se resalta que las observaciones resultantes de la tabla de grupos discriminados coincide con las observaciones consideradas en la nueva variable.

La variable *df1* que ahora aparece en el conjunto de datos del ejemplo, cuenta con valores de los scores asignados por el análisis de acuerdo a los predictores ingresados, ubicándose entre un número máximo 7.653388 y un número mínimo de 1.467964, con una desviación estándar equivalente a 1, lo que asegura que es una función discriminante estandarizada.

Ahora el procedimiento se repite para el grupo con las nuevas clasificaciones, el grupo discriminante, con el comando **predict _____, class**:

predict *nombre de la nueva variable* **, class**

Ejemplo:

predict *dclase*, **class**

```
. predict dclase, class
(1,740 missing values generated)
```

Nuevamente se genera una variable con los valores asignados por el cálculo para cada observación correspondiente al grupo verdadero y observaciones discriminadas.

Así como anteriormente, se recomienda hacer una visualización de la nueva variable generada, para lo cual se puede hacer por medio de un **sum** o **tab**:

tab dclase

```
. tab dclase
```

classification	Freq.	Percent	Cum.
0	792	62.71	62.71
1	471	37.29	100.00
Total	1,263	100.00	

True zm	Classified 0	1	Total
0	564 73.92	199 26.08	763 100.00
1	228 45.60	272 54.40	500 100.00
Total	792 62.71	471 37.29	1,263 100.00
Priors	0.5000	0.5000	

En la comparación de los resultados con la tabla obtenida por el análisis discriminante se remarca que el resultado coincide que las observaciones discriminadas de la clasificación resultante del análisis. Estos son los valores clasificados.

Después de haber obtenido las variables con los scores asignados para cada respectivo grupo, el siguiente paso del proceso es conocer los centroides de estos, por medio de un comando descriptivo como **sum**. La instrucción para realizar esta acción debe ir antecedida por un **bysort,** por medio del cual se indica que los comandos siguientes se ejecutarán separadamente para cada valor único de la variable señalada (que para el caso del análisis sería la variable que determina los grupos). Esto permite calcular las estadísticas descriptivas para cada grupo definido por los valores de la variable que se escriba en el comando.

Ejemplo:

bysort *zm*: **sum** *df1*

```
. bysort zm: sum df1
```

-> zm = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
df1	763	.2815792	.8527902	-3.302905	1.467964

-> zm = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
df1	500	-.4296898	1.190169	-7.653388	1.385085

True zm	Classified 0	1	Total
0	564 73.92	199 26.08	763 100.00
1	228 45.60	272 54.40	500 100.00
Total	792 62.71	471 37.29	1,263 100.00
Priors	0.5000	0.5000	

El resultado es el resumen de las estadísticas descriptivas de la variable con los scores de la función discriminante, en cada respuesta que hace que se distingan dos grupos. De acuerdo al

ejemplo que se ha mostrado, se obtuvieron las medias del grupo verdadero, el cual es la zona no metropolitana y la metropolitana, estos son los centroides con los que inicia el análisis:

- Los valores de .2815792 y -.4296898 son los centroides de los grupos verdaderos que ayudan a discriminar inicialmente. Mismos valores que aparecen en la tabla de Group means on canonical variables al ejecutar el comando del análisis discriminante:

Group means on canonical variables

zm	function1
0	.2815792
1	-.4296898

Lo que falta es calcular los **centroides con los que acaba el análisis**, es decir las medias del grupo con las observaciones reclasificadas/discriminadas, lo cual se realiza con el comando, pero se sustituye con la variable creada para el grupo discriminado, que en el caso del ejemplo se llamó *dclase*.

Ejemplo:

bysort *dclase*: **sum** *df1*

```
. bysort dclase: sum df1

-------------------------------------------------------------------------------
-> dclase = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         df1 |        792    .6514405    .3965254  -.0721596   1.467964

-------------------------------------------------------------------------------
-> dclase = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         df1 |        471   -1.095416    .9087836  -7.653388  -.0854899

-------------------------------------------------------------------------------
-> dclase = .

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         df1 |          0
```

Las medias resultantes del análisis discriminante son los valores .6514405 y -1.095416, los cuales ayudarán a determinar a qué grupo pertenecen las observaciones.

Uno de los resultados más valiosos de este método multivariante es identificar cuáles observaciones fueron reclasificadas de acuerdo al análisis y eso puede hacerse enlistando las observaciones del grupo verdadero contra las observaciones reclasificadas, lo cual se puede hacer con la instrucción **estat list**:

Ejemplo:

estat list

```
. set more on

. estat list
```

Obs.	Classification True	Class.	Probabilities 0	1
1	0	0	0.6943	0.3057
2	0	0	0.5342	0.4658
3	0	0	0.6295	0.3705
4	0	0	0.6699	0.3301
5	0	0	0.6205	0.3795
6	0	0	0.7243	0.2757
7	1	0 *	0.6064	0.3936
8	1	0 *	0.6568	0.3432
9	1	0 *	0.5060	0.4940
10	0	0	0.5060	0.4940
11	1	0 *	0.5342	0.4658
12	0	0	0.7174	0.2826
13	0	0	0.6279	0.3721

- La tabla consecuente muestra observación por observación cuál es el resultado después de haber pasado por el análisis discriminante. Ejemplo:
- Observación 1: En el grupo verdadero dice que está en la categoría 0 (zona no conurbada).
- En el grupo discriminado está en la categoría 0 (zona no conurbada). A esta observación no le pasó nada.
- Las marcadas con un asterisco son las que sí fueron reclasificadas, mostrando en qué lugar estaban en el grupo verdadero y en dónde quedaron después del análisis, como es la observación 7, 8, 9 y 11.

Se puede obtener las características de cada observación, que dicho lo anterior las que causan mayor interés son las que fueron reclasificadas. A través del comando **sum** con las

variables que fueron marcadas como predictores junto con un **in,** se apunta que posición ocupa la observación de interés.

Ejemplo:

sum ing_mens educ educ_padre **in 2/2**

```
. sum ing_mens educ educ_padre in 2/2
```

Variable	Obs	Mean	Std. dev.	Min	Max
ing_mens	1	999	.	999	999
educ	1	4	.	4	4
educ_padre	1	1	.	1	1

Conclusión del capítulo Análisis discriminante:

La observación corresponde a una persona con un ingreso mensual promedio de $1000, estudió hasta la preparatoria y su padre sin estudios. **Es una observación que en el grupo verdadero estaba en la zona metropolitana y dada sus características la reclasificó a la zona no conurbada.**

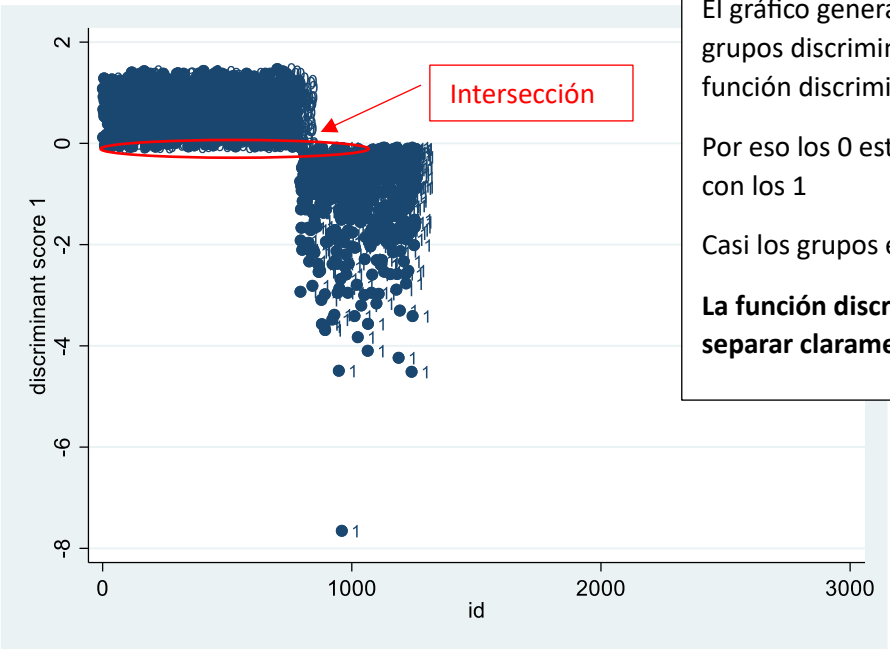
Es importante señalar que para determinar a qué grupo pertenecen las observaciones en el análisis, cada variable influye de manera distinta en el resultado. En el caso del ejemplo desarrollado, de acuerdo a las cargas factoriales, la variable de educación del padre es la que más peso tiene sobre el resultado, con un valor de -.6915526. (Las cargas factoriales se ven en la tabla *Standardized canonical discriminant function coefficients,* como uno de los valores resultantes al ejecutar un análisis discriminante).

También es muy interesante el comportamiento del conjunto de datos bajo el efecto de este método mediante un gráfico, que para ello se utiliza una combinación de comando, ya que esta acción es un poco forzada. Para hacerlo se siguen los siguientes pasos:

1. Generar un ID, una columna que enliste cada una de las observaciones, a través del comando **gen id=_n.**
2. Inserta el siguiente comando después de haber generado el id **twoway (scatter** *df1 id***, mlabel(** *dclase* **))**, donde:

- **twoway (scatter** es la indicación de hacer un gráfico de dispersión.
- *df1 id* es la orden de que se haga el gráfico con las variables df1 y id (las cuales son las que se generaron para este ejemplo).

- **mlabel(** *dclase* **))** es la instrucción dada para que etiquete a las observaciones con la variable *dclase*.



El gráfico generado muestra los grupos discriminados divididos por la función discriminante.

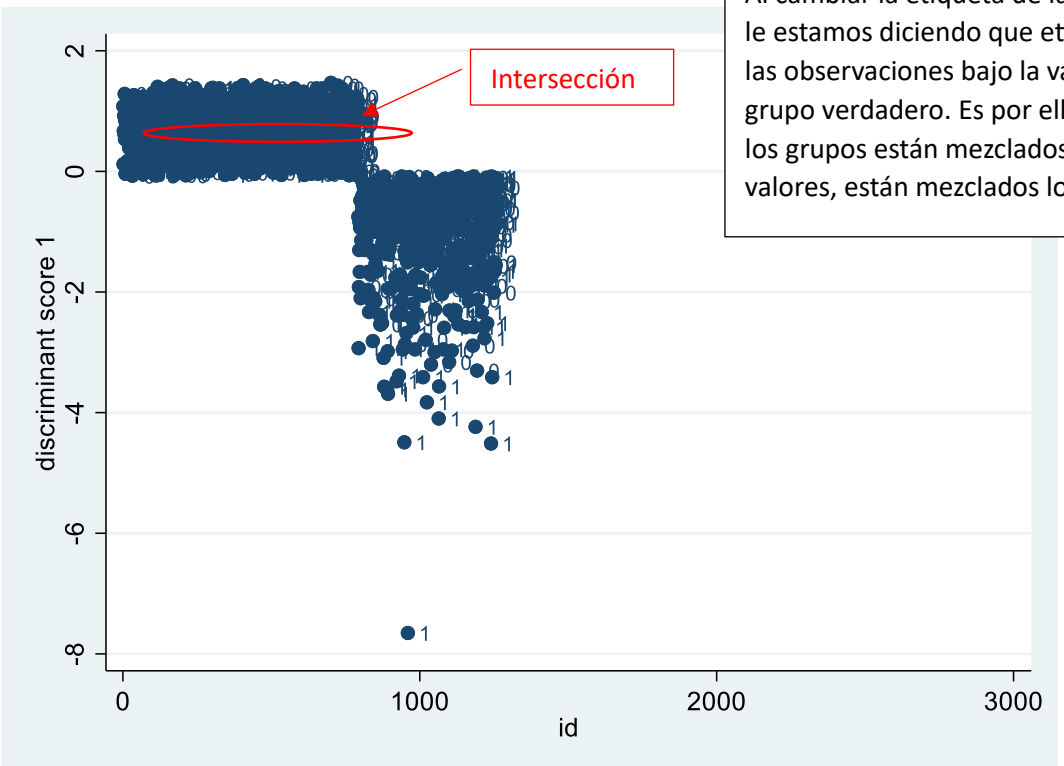
Por eso los 0 están con los 0 y los 1 con los 1

Casi los grupos están divididos por 0.

La función discriminante ayuda a separar claramente a los grupos.

Ahora se hace la gráfica para el grupo verdadero:

twoway (*scatter df1 id***, mlabel(** *zm* **))**



Al cambiar la etiqueta de la variable, le estamos diciendo que etiquete a las observaciones bajo la variable del grupo verdadero. Es por ellos que en los grupos están mezclados ambos valores, están mezclados los 0 y 1.

Para una mayor referencia del contenido de este capítulo ver:

Cleff, T. (2025). "Applied statistics and multivariate data analysis for business and economics" (2nd ed.). Cham, Switzerland: Springer.

El objetivo de esta técnica es que a través de ciertas propiedades se generen grupos/conjuntos de elementos u objetos a modo de que con estas características se logre identificar un criterio de homogeneidad. El clúster es cada uno de los grupos que forma la técnica con sus propios elementos, demostrando que estos cuentan con rasgos más o menos similares bajo el criterio de homogeneidad.

Como método multivariante busca reducir la dimensión de los datos, tratando de explicarlos con el menor número de grupos, los cuales a nivel teórico no son explícitos, pero de forma intuitiva es posible generarlos y conocer cómo se explican con ciertos predictores, cómo es que se agrupan los datos, por ejemplo. Los predictores son características que se eligen entre las variables del conjunto de datos a analizar, que ayudan a explicar el fenómeno de estudio o funcionarán para contestar alguna pregunta de la investigación.

El análisis de clúster ideal es que las variables se agrupen naturalmente, quien realice la investigación no forcé la partición en grupos, sino que se haga naturalmente con los indicadores que se determinen, así como se espera que los predictores dados al clúster lo hagan homogéneo, pero esta situación no es sencilla que suceda ya que, entre las mismas variables del fenómeno existe mucha heterogeneidad.

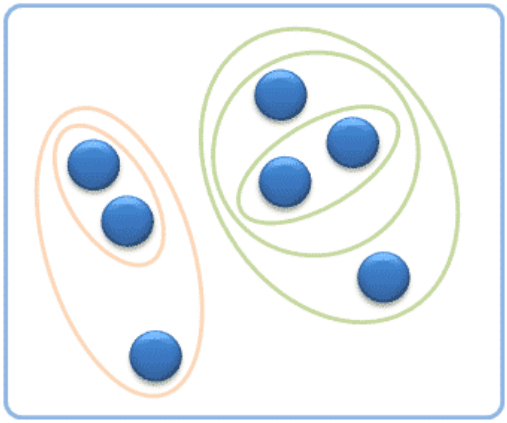
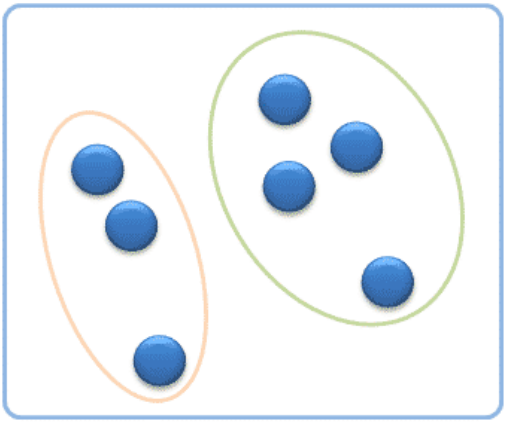
Una de las principales preguntas a satisfacer al aplicar el análisis de clúster es, ¿cuántos grupos se deben de formar? No hay nada que indique, ni en técnica, ni en teoría, exactamente cuántos grupos se deben generar. Por experiencia o por una investigación del fenómeno de estudio como antecedente se puede tener una noción de quizás cuantos grupos se pudieran obtener, pero no es una precisión. Por ello, este punto exige más fundamentación teórica que fundamente la explicación del resultado, dado que sólo se sabe la respuesta hasta que se ejecuta el análisis y eso podría representar una debilidad de la técnica en comparativa con otras.

Puntos importantes:

- Similitud (similarity): es una magnitud que se ocupa para medir el grado de homogeneidad, es decir qué tan comunes / similares pueden ser dos elementos de acuerdo a un sólo criterio. Se entiende como una medida de homogeneidad o heterogeneidad.
- El análisis de cluster debe estar basado en un solo criterio
- Criterio no es lo mismo que una característica o propiedades.

- No puede ser una variable que se utilice para agrupar.
- Un criterio no puede ser un predictor del modelo, una variable que explique el fenómeno con la cual determinen grupos.
- Tiene que ser una característica común en todo el contexto de las observaciones, pero que no sea un predictor del fenómeno.
- La distancia es un criterio que puede ser usado universalmente, a cualquier par de objetos en un diagrama de "n" dimensiones.
- Es necesario tener las variables apropiadas para poder distinguir los grupos. Si no se incluye las variables que pueden ayudar a particionar a la población y se dejan fuera todas estas opciones, no se conocería cuál es la variable que ayuda a agrupar mejor. Entre más cosas se dejen fuera mayor es el número de grupos que puedo formar. Sin embargo, esto no significa que para formar grupos ideales tenga yo que tener el mayor número de variables, más bien es importante conocer y tener cuáles son los predictores apropiados para poder distribuir los grupos. Si se deja muchas variables fuera, no se puede explicar lis grupos de manera adecuada.
- En la práctica suelen utilizarse entre 5 y 10 grupos en promedio, más usualmente 7.

Métodos de análisis de cluster que mayormente se utilizan

Método jerárquico	Método no jerárquico (como K-means)
* Se puede aplicar con un tamaño de la muestra menor de 1000 observaciones. * No más de 20 variables. * Idealmente 300 o 200 observaciones. * La técnica es altamente sensible a outliers. Por ello hay se deben analizarlos previamente, con el fin de que sea posible eliminarlos. Cuando se hace un análisis jerárquico se determina CLUSTER JERÁRQUICOS HIERARCHICAL CLUSTER 	Puede ser cualquier tamaño de muestra. CLUSTER NO JERÁRQUICOS PARTITIONING CLUSTER  Se establece un conjunto de características y con ellas se forman los clusters, pero de facto no hay observaciones que se unan primordialmente. No se asume que haya un par de objetos que su unión sea superior a otras agrupaciones. En la práctica, cuando se ejecuta este tipo de análisis se suele hacer dos cosas:

cuáles son las observaciones que con base al criterio de distancia de disimilitud, por ejemplo euclidiana, toman jerarquía en este tipo de grupos. Entonces se observa que se forma un grupo más grande, luego uno más pequeño, luego otro más pequeño y así sucesivamente.

En otras palabras, del conjunto de clusters formados, el cluster más amplio tiene jerarquía sobre el resto de los grupos, de tal forma que se pueda definir qué tipo de jerarquía se considera para poder particionar los objetos.

Entre más grupos allá y más grandes, son más heterogéneos. Entre más grupos pequeños haya, jerárquicamente hablando, hay más homogeneidad.

- Dar instrucción de que la primera observación sea la que tome para iniciar a agrupar.
- Otra consideración es que el inicio de dicha agrupación sea aleatorio.

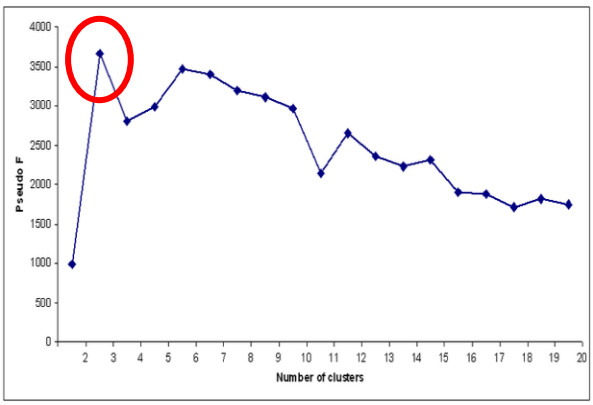
En este método no hay jerarquías, por lo que todas las observaciones son idénticas, es decir que todas tienen la misma probabilidad de agruparse en la metodología.

Las metodologías más usadas para calcular distancias:

- Single-linkage / vecino más cercano
- Complete linkage: Calcula la distancia completa entre la observación más lejana con respecto a otra observación, considerando que la distancia no es una razón de disimilitud, sino que se le da prioridad a ciertas propiedades multivariantes, es decir que en conjunto explicarse.
- Average linkage / intermedio entre las anteriores.
- Centroid linkage / es lo más parecido a usar medias, al utilizar la distancia media con respecto a la técnica más usada de métodos no jerárquicos (K-means).
- Ward's method

Determinando el número de clusters

Stopping rules / Reglas de clusters que determinan una compensación entre heterogeneidad y el número de grupos.



- Seudo F: funciona al inverso de la seudo t cuadrada, ya que no es una medida de heterogeneidad, es una medida de homogeneidad en donde se suele usar el más alto.
- Seudo T cuadrada: Al ir generando grupos, se identifica una mayor heterogeneidad y de repente baja. Después de ello se repite el comportamiento.
- Usualmente lo que se hace para identificar esta regla de la seudo T cuadrada, es que cuando llega a un pico inmediatamente al número menor que sigue, ese es el número de clusters más apropiado entre el número de homogeneidad.

De manera general, los pasos a seguir para realizar un análisis de cluster son:

1. Explorar las variables a utilizar como predictores para el modelo.
2. Seleccionar la metodología a usar para calcular distancias.
3. Realizar la gráfica de grupos de cluster:

a) Determinar la distancia a considerar.
b) Conocer la cantidad de observaciones que tendrá cada grupo.

4. Aplicar stopping rules para determinar la cantidad de clusters convenientes para el resultado.
5. Particionar los clusters.
6. Validar los grupos.

La base de datos de apoyo para ejemplificar esta metodología corresponde a un estudio de mercado y esta integrada por 100 observaciones y 24 variables, que al igual que todas las utilizadas en el documento son de acceso libre. Las variables que se usaron como predictores aparecen bajo el nombre de *x6, x8, x12, x15* y *x18,* las cuales sus respuestas están en una escala de 0 a 10.

Variable	Descripción
id	Número único de registro
x1	Tipo de cliente
x2	Tipo de industria
x3	Tamaño de la empresa
x4	Región
x5	Sistema de distribución
x6	Calidad del producto
x7	Actividades de comercio electrónico
x8	Apoyo técnico
x9	Resolución de quejas
x10	Publicidad

x11	Línea de producto
x12	Imagen de Salesforce
x13	Precio competitivo
x14	Reclamos de garantía
x15	Nuevos productos
x16	Pedido y facturación
x17	Flexibilidad de precios
x18	Velocidad de entrega
x19	Satisfacción
x20	Probabilidad de recomendación
x21	Probabilidad de compra
x22	Nivel de compra
x23	Consideración de alianza estratégica

El primer paso recomendado es conocer cómo están compuestas las variables seleccionadas como predictores, qué tipo de información poseen y cuáles son sus estadísticas básicas, para lo cual se utilizan los comandos **des, sum y pwcorr…, sig**:

des x6 x8 x12 x15 x18

```
. des x6 x8 x12 x15 x18

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
x6              double  %5.1f      labels5    X6 - Product Quality
x8              double  %5.1f      labels5    X8 - Technical Support
x12             double  %5.1f      labels5    X12 - Salesforce Image
x15             double  %5.1f      labels5    X15 - New Products
x18             double  %5.1f      labels5    X18 - Delivery Speed
```

sum x6 x8 x12 x15 x18

```
. sum x6 x8 x12 x15 x18

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          x6 |        100        7.81    1.396279         5         10
          x8 |        100       5.365    1.530457       1.3        8.5
         x12 |        100       5.123     1.07232       2.9        8.2
         x15 |        100        5.15    1.493048       1.7        9.5
         x18 |        100       3.886    .7344372       1.6        5.5
```

Estos comandos se ejecutan con la intención de explorar la información que contienen las variables. Después se ejecuta el comando de correlación para explorar si existe correlación lineal entre ellas y si cuenta con significancia estadística.

pwcorr x6 x8 x12 x15 x18**, sig**

```
. pwcorr x6 x8 x12 x15 x18, sig
```

	x6	x8	x12	x15	x18
x6	1.0000				
x8	0.0956 0.3441	1.0000			
x12	-0.1518 0.1316	0.0170 0.8668	1.0000		
x15	0.0270 0.7898	-0.0736 0.4669	0.0316 0.7547	1.0000	
x18	0.0277 0.7843	0.0254 0.8016	0.2716 0.0063	0.1057 0.2950	1.0000

Hacer una matriz de correlación ayuda para orientarnos qué metodología de cálculo de distancia es conveniente usar, así se puede tener un mejor resultado, aunque para el clúster no es un criterio indispensable que entre las variables haya correlación, pero de igual forma se tendrá que explicar la disimilitud ya que, lo que se desea encontrar son las diferencias entre las variables.

Para el análisis es relevante que tipo de metodología se utiliza para hacer el agrupamiento y de acuerdo al resultado obtenido en las variables, ejemplo se observa que:

- Entre estas variables existe una correlación baja.
- En la práctica cuando se encuentra una **matriz de correlación donde no hay significancia estadística y bajas correlaciones**, la técnica que más ayuda a encontrar disimilitudes es **complete linkage**, porque se van a encontrar elementos con las mayores distancias que expliquen la disimilitud. Se empezará a agrupar a partir de las mayores distancias.
- Cuando se encuentra una **matriz donde las correlaciones son altas en general y estadísticamente significativas**, idealmente se escoge la metodología del **vecino más cercano**, porque le será más sencillo al análisis encontrar los elementos cercanos.
- A pesar de las recomendaciones mencionadas, se sugiere correr las 3 metodologías más usadas para métodos jerárquicos (Single-linkage / vecino más cercano, complete linkage y average linkage), y tomar en cuenta la que tenga mejor rendimiento.

Como ya se señaló, por las características obtenidas en la matriz de correlaciones arrojada por los predictores que se utilizan para el análisis, se decide ejecutar el método de complete linkage. El comando de dicha acción es **cluster completelinkage**

cluster completelinkage listado de variables **, name(** nombre del cluster **)**

Ejemplo:

cluster completelinkage *x6 x8 x12 x15 x18*, name(last)

```
. cluster completelinkage x6 x8 x12 x15 x18, name(last)
```

Obteniendo como resultado la creación de tres medidas automáticamente, que se pueden ver en la ventana de variables.

La información que contienen estas variables recientemente creadas es:

sum last_id last_ord last_hgt

```
. sum last_id last_ord last_hgt
```

Variable	Obs	Mean	Std. Dev.	Min	Max
last_id	100	50.5	29.01149	1	100
last_ord	100	50.5	29.01149	1	100
last_hgt	99	2.362777	1.635172	.4	8.938121

- Last_id: Es una columna que se genera con un valor autonumérico, otorgando un valor a cada una de las observaciones, que en caso del ejemplo va del 1 al 100 por ser el total de observaciones que tiene el conjunto de datos.
- Last_ord: Es el orden jerárquico con el que se fueron considerando las observaciones para ser agrupadas.
- Last_hgt: Es la medida de similitud. (la distancia euclidiana en múltiples dimensiones).

También es conveniente ejecutar el comando **list** para observar que ha pasado con las observaciones como resultado del análisis indicado.

list last_id last_ord last_hgt

```
. list last_id last_ord last_hgt

     +-------------------------------+
     | last_id   last_ord    last_hgt |
     |-------------------------------|
  1. |       1          1   1.8841444 |
  2. |       2         24   2.9308702 |
  3. |       3         41   4.0975602 |
  4. |       4          2   .50990195 |
  5. |       5         45   1.7635192 |
     |-------------------------------|
  6. |       6         16   .99498744 |
  7. |       7         52   3.0626786 |
  8. |       8         98   4.6636895 |
  9. |       9          3          .4 |
 10. |      10         94   1.1090537 |
     |-------------------------------|
 11. |      11         38   2.0808652 |
 12. |      12          5   .72801099 |
     +-------------------------------+
```

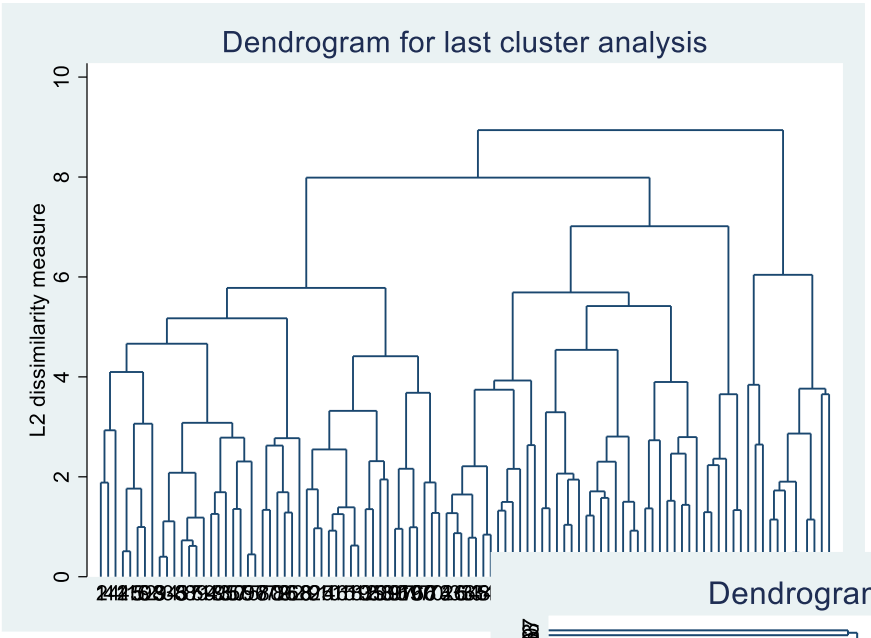
- La primera observación la tomó como jerarquía para iniciar a hacer los clusters.
- Entre las columnas last_id y last_ord cuentan con la misma media y desviación estándar porque contienen los mismos elementos, pero el orden de jerarquía con el que se consideró para hacer los clusters.

Lo siguiente a realizar es ejecutar el comando para hacer el dendrograma de cluster:

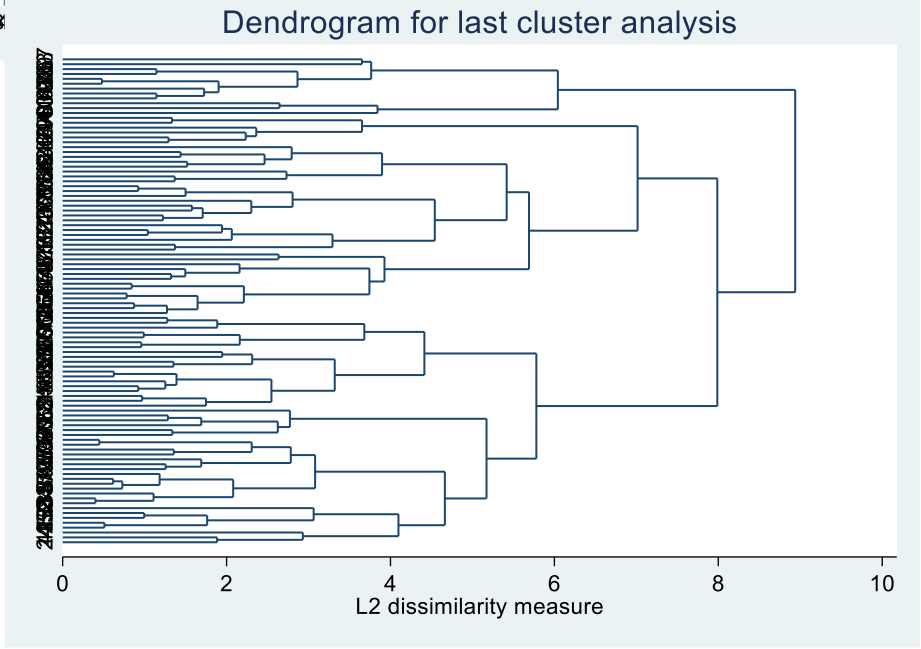
cluster dendrogram

(Este comando se ejecuta después de haber realizado el comando de metodología de cálculo de distancia)

cluster dendrogram o para visualizarlo de forma horizontal, cluster dendogram, horizontal

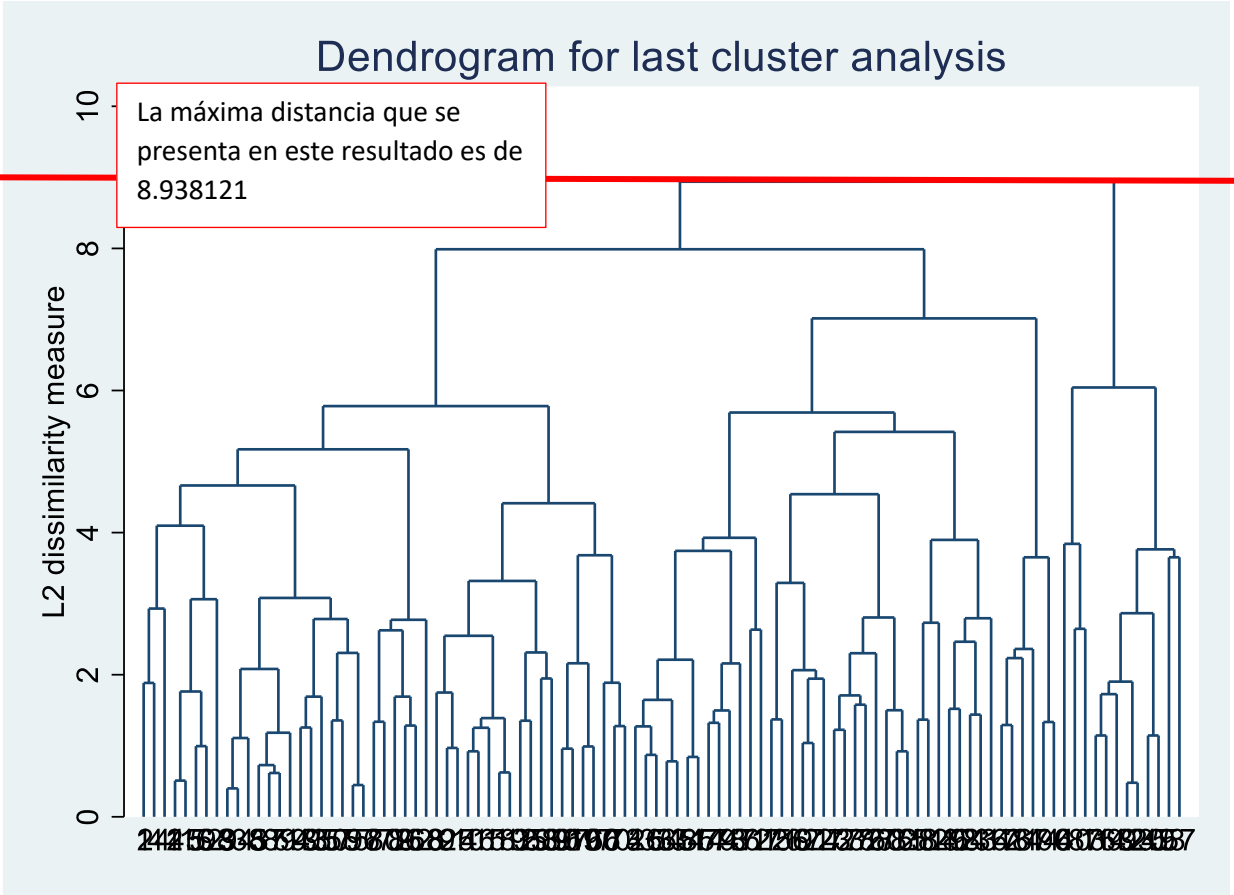


Dendograma vertical



Dendograma horizontal

- Se genera así porque se hace a partir del único cluster que se ha generado (last).
- La distancia con la que se genera es por de facto la distancia euclidiana (L2).
- Aquí se muestra claramente una de las primeras preguntas que se presenta al realizar un análisis de cluster, ¿cuántos cluster se deben tomar en cuenta? No hay un criterio que indique un número preciso.

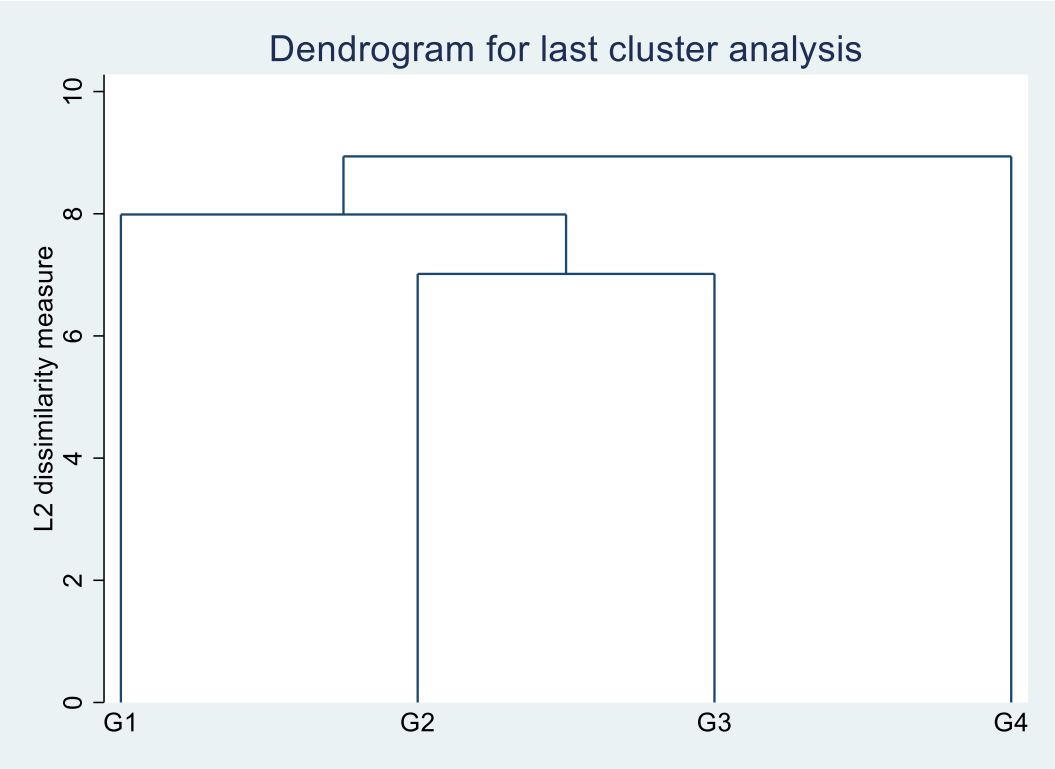


Este dendrograma es el resultado de un cluster por método jerárquico, donde se combinan observaciones en un determinado orden, que de acuerdo a la técnica de cálculo de distancia (vecino más cercano o complete linkage), se empezarán a generar grupos. Como se muestra en el resultado de la gráfica, se obtuvieron grupos heterogéneos, son muy pocos, pero heterogéneos.

Con este resultado conocer el número de grupos y su descripción es muy difícil, así que se afina el comando para cluster indicando hasta qué número de distancia toma en cuenta para generar el dendrograma, por medio de una instrucción más al comando que es **cutvalue.**

Ejemplo:

cluster dendrogram, cutvalue(*7*)



Con la indicación de tomar 7 como la máxima medida de distancia, el análisis arrojo 4 clusters, los cuales se desprenden del dendroagrama pasado.

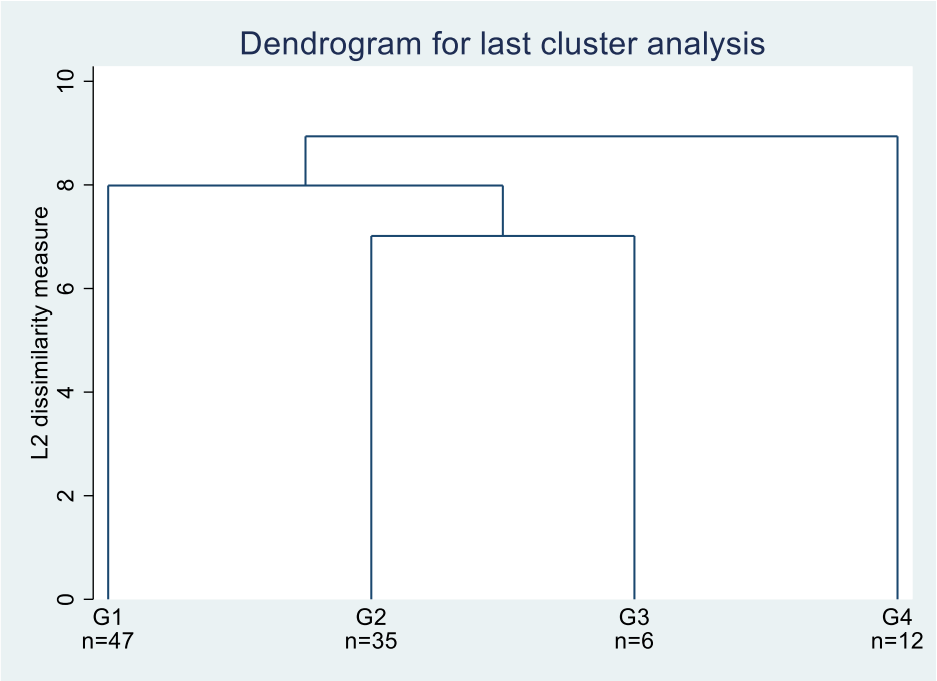
Lo que sigue del desarrollo del análisis es saber cuántas observaciones hay en cada grupo, para lo cual nuevamente se ejecuta el último comando, pero agregando la instrucción **showcount**

cluster dendrogram, cutvalue(*cantidad de máxima distancia*) showcount

(Arroja la creación de clusters de acuerdo a la cantidad de la máxima distancia especificada)

Ejemplo:

cluster dendrogram, cutvalue(*7*) showcount



Siguiendo el proceso dictado por la teoría y con el fin de responder uno de los principales planteamientos a atender por el análisis de cluster, se requiere definir ¿cuántos grupos son necesarios para dar explicación al fenómeno de estudio? O ¿cuántos grupos se deben generar para obtener un resultado adecuado para su interpretación? El cálculo de la cantidad de clusters a formar se puede orientar gracias a las stopping rules que se determinan a través de las medidas seudo F y seudo T cuadrada.

El comando para stoppin rule que de forma automática arroja la medida seudo F es:

cluster stop

(Medida de homogeneidad seudo F)

Ejemplo:

cluster stop

Number of clusters	Calinski/ Harabasz pseudo-F
2	14.51
3	18.82
4	17.02
5	14.99
6	17.27
7	17.80
8	17.13
9	17.61
10	17.52
11	17.83
12	18.25
13	17.93
14	17.16
15	16.63

Se debe ubicar el valor más alto para determinar el número de clusters.

En el caso del ejemplo se ubica en el número 3 de clusters. Sin embargo, no es recomendable que sea el único criterio a considerar, también es recomendable aplicar la medida de la seudo t cuadrada para tener una mejor estimación de los grupos a formar.

El comando para obtener la medida de heterogeneidad, que es la t cuadrada es:

```
cluster stop cluster stop, rule (duda)
(Medida de heterogeneidad)
```

Al ser una medida de heterogeneidad se comporta al inverso de la seudo F, con una cierta consideración. Se debe ubicar un pico en el resultado de la medida y se tiene que tomar en cuenta el siguiente número de clústeres a generar donde la distancia se registre más baja hasta que vuelva a subir.

Leyendo la columna de la seudo t cuadrada, el pico o el valor más alto se ubica en el número de clústeres a crear 2, luego desciende en el número 3, así como en el 4, en el número 5 sube nuevamente la medida representando un pico, por lo que el valor al que debemos tomar en cuenta es el 4.

```
. cluster stop, rule (duda)
```

Number of clusters	Duda/Hart	
	Je(2)/Je(1)	pseudo T-squared
1	0.8710	14.51
2	0.8087	20.35
3	0.7867	10.57
4	0.5276	8.95
5	0.7119	18.21
6	0.7572	10.58
7	0.7439	6.89
8	0.7229	9.97
9	0.6938	8.83
10	0.5168	11.22
11	0.6321	9.90
12	0.5237	5.46
13	0.8131	2.53
14	0.6134	3.78
15	0.3247	2.08

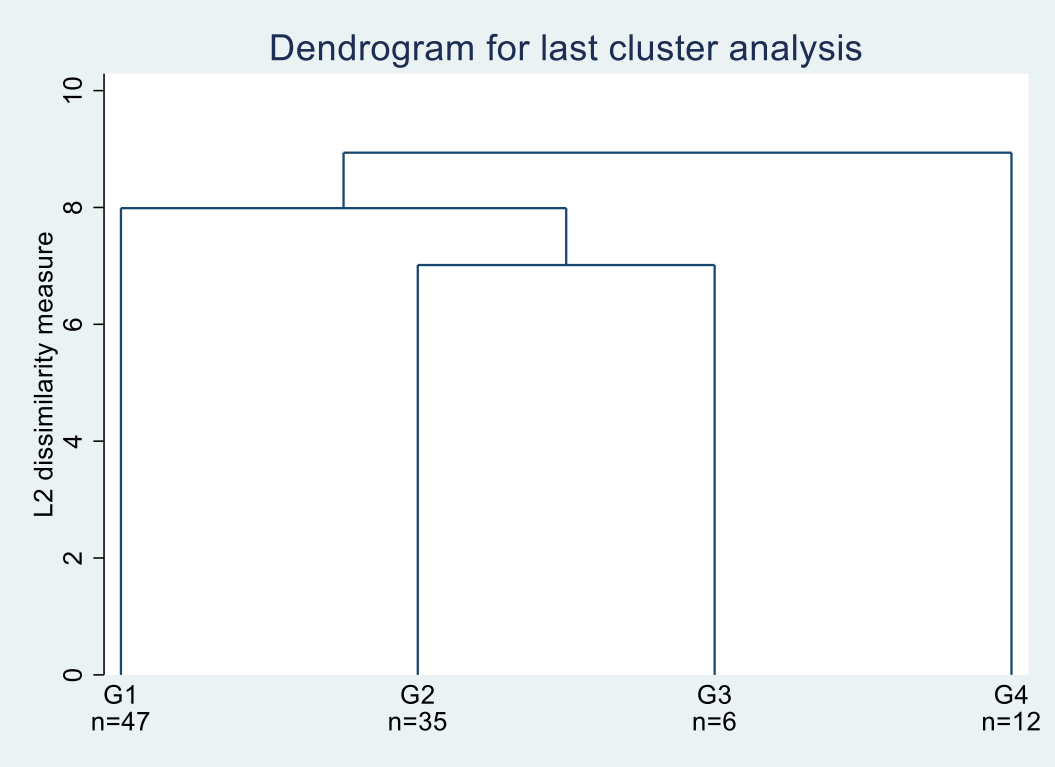
Para elegir qué número de clusters generar, después de haber obtenido ambas medidas, se recomienda hacer la prueba de ambos valores y evaluar qué número de grupos es mejor para el análisis. El comando para indicar la cantidad de cluster que se desea obtener es:

cluster dendrogram, cutnumber(*número de clusters a obtener*) showcount

(Arroja el número de clusters deseado especificando la cantidad)

Ejemplo:

cluster dendrogram, cutnumber(*4*) showcount



Con la decisión de cuántos grupos se formarán para interpretar en el resultado del análisis de clusters, junto con el número de observaciones que integra cada uno de ellos, lo que continúa es hacer la partición del conjunto de datos en los clusters determinados. Para realizar dicha acción se necesita generar una variable con el siguiente comando:

cluster generate grupos = group(*Cantidad de grupos a generar*)

Tomando en cuenta los resultados de las stopping rules del ejemplo se corre el comando indicando que particione en 4 grupos:

cluster generate grupos = group(*4*)

Con ello se creó la variable con los grupos dividiendo las observaciones: **tab** grupos

```
. tab grupos
```

grupos	Freq.	Percent	Cum.
1	47	47.00	47.00
2	35	35.00	82.00
3	6	6.00	88.00
4	12	12.00	100.00
Total	100	100.00	

El último paso del proceso es hacer la validación de los clusters obtenidos, lo cual se puede realizar con las variables que funcionaron como predictores o incluso se pueden agregar otras características que ayuden a explicar el fenómeno.

Ejemplo:

tabstat x6 x8 x12 x15 x18 **, by (** grupos **)**

```
. tabstat x6 x8 x12 x15 x18, by ( grupos )

Summary statistics: mean
  by categories of: grupos
```

El estadístico solicitado es la media

Para las categorías de la variable grupos

grupos	x6	x8	x12	x15	x18
1	7.665957	4.508511	5.176596	4.614894	3.961702
2	8.288571	6.74	4.782857	5.08	3.651429
3	5.483333	7.133333	6.2	5.216667	4.483333
4	8.141667	3.825	5.366667	7.416667	3.975
Total	7.81	5.365	5.123	5.15	3.886

Validar es verificar cómo son las medias de cada grupo:

- Qué tan diferentes son las medias de cada grupo.
- Qué patrones se visualizan: grupos en donde ciertas medias con las más bajas, las más altas, más o menos las que se encuentran en el inter.

Conclusiones:

En el grupo 1 se resalta que tiene la menor media de la variable x15 (4.614894), que corresponde a nuevos productos.

La media más grande la tiene la variable x6 (8.288571) ubicada en el grupo 2, quien tiene información sobre calidad de producto.

Para el x8 su media más alta se encuentra en el grupo 3 con un valor de 7.133333, refiriéndose a soporte técnico.

El grupo 3 es el que concentra mayor número de medias altas y por ello es el que cuenta con los mejores calificados en la muestra (recuerde que los valores de las variables son calificaciones que van de 10 a 0, donde 10 es lo mejor y 0 lo peor).

Lo que se busca hacer con estas descripciones es obtener hallazgos que nos muestren patrones, por ejemplo: El grupo 3, el cual muestra las medias más altas de todas las variables, se puede observar que la variable de políticas de soporte técnico es la que mejor está ranqueada y la calidad del producto, por lo que son los rubros que se pueden mejorar en los otros grupos para acercarse a los resultados del grupo 3.

Acock, A. C. (2014). “A gentle introduction to Stata” (4th ed.). College Station, TX: Stata Press.

Blundell, R., & Costa Dias, M. (2009). Alternative Approaches to Evaluation in Empirical Microeconomics. Journal of Human Resources, 44(3), 565-640.

Butt, N. S., Shahbaz, M. Q., & Kamal, S. (2008). MANOVA with summary statistics: A Stata program. “SSRN Electronic Journal, 4”(1).

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2018). Regression discontinuity designs using covariates. “arXiv”.

Cameron, A. C., & Trivedi, P. K. (2010). Microeconometrics Using Stata (Rev. ed.). Stata Press.

Cameron, A. C., & Trivedi, P. K. (2013). Regression Analysis of Count Data (Vol. 53). Cambridge University Press.

Chib, S. (1995). Marginal likelihood from the Gibbs output. “Journal of the American Statistical Association, 90”(4), 1313–1321.

Chib, S., & Greenberg, E. (1995). Bayesian analysis of binary and polychotomous response data. “Journal of the American Statistical Association, 88”(2), 669–679.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. “The American Statistician, 49”(4), 327–335.

Cleff, T. (2020). “Applied statistics and multivariate data analysis for business and economics: A modern approach using SPSS, Stata, and Excel”. Cham, Switzerland: Springer.

Cleff, T. (2025). “Applied statistics and multivariate data analysis for business and economics” (2nd ed.). Cham, Switzerland: Springer.

Deaton, A. (2013). The Great Escape: Health, Wealth, and the Origins of Inequality. Princeton University Press.

Fan, J., Liao, Y., & Liu, H. (2015). An overview on the estimation of large covariance and precision matrices. “arXiv”.

Geldhof, J., Arnold, M., & Constantine, N. (2016). Better crunching: Recommendations for multivariate data analysis approaches for program impact evaluations. “Journal of Extension”, (June).

Gudmundsson, G. (1977). Multivariate analysis of economic variables. “Journal of the Royal Statistical Society: Series C (Applied Statistics), 26”(1), 48–59.

Haug, S., & Stelzer, R. (2011). Multivariate ECOGARCH processes. “Econometric Theory, 27”(2), 344–371.

Heckman, J. J., & Vytlacil, E. (2007). Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. Handbook of Econometrics, 6, 4779-4874.

Heckman, J. J., & Vytlacil, E. (2007). Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Estimate the Average Treatment Effect. Handbook of Econometrics, 6, 4875-5143.

Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis (Pearson, Ed.; 6 th).

Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. “Review of Economic Studies, 65”, 361–393.

Michler, J. D., & Josephson, A. (2021). Recent developments in inference: Practicalities for applied economics. “arXiv”.

Neelon, B., Chang, H. H., Ling, Q., & Hastings, N. S. (2016). Spatiotemporal hurdle models for zero-inflated count data: Exploring trends in emergency department visits. “Statistical Methods in Medical Research, 25”(6), 2558–2576.

Seo, M. H., Kim, S., & Kim, Y.-J. (2019). Estimation of dynamic panel threshold model using Stata. “arXiv”.

StataCorp. (2025). “Stata 19 Multivariate Statistics Reference Manual”. College Station, TX: Stata Press.

Verardi, V., & Dehon, C. (2010). Multivariate outlier detection in Stata. “The Stata Journal, 10”(2), 313–329.

Weber, S. (2010). Bacon: An effective way to detect outliers in multivariate data using Stata (and Mata). “The Stata Journal, 10”(3), 403–420.

REALIZAÇÃO:

SEVEN
publicações acadêmicas

ACESSE NOSSO CATÁLOGO!

WWW.SEVENPUBLI.COM

CONECTANDO O **PESQUISADOR** E A **CIÊNCIA** EM UM SÓ CLIQUE.