ANÁLISIS ESTADÍSTICO MULTIVARIADO

Análisis Multivariado

Análisis Multivariado

Análisis Multivariado

1. Programa

2. Introducción
   - Breve repaso de álgebra lineal
   - Descripción de datos univariados
   - Descripción de datos multivariados
   - Ejemplo: Algunos indicadores socioeconómicos en centros urbanos de Argentina

3. Anexo: algo más sobre dependencia entre variables

1. **Preliminares**: Objetivos y alcance del análisis multivariado. Nociones útiles de álgebra matricial:
2. **Descripción de datos multivariados**: La matriz de datos, vector de medias, matriz de variancias y covariancias, medidas globales de variabilidad, medidas de distancia, matriz de precisión.
3. **Componentes Principales**: Enfoques descriptivo, estadístico y geométrico del problema. Calculo de las componentes. Propiedades. Los componentes como predictores óptimos. Escalado multidimensional.
4. **Inferencia con datos multivariados**: Variables aleatorias vectoriales. Propiedades. Distribuciones de probabilidad multivariadas.
5. **Análisis Factorial**: Hipótesis del modelo factorial. Propiedades. Unicidad del modelo. Estimación de factores. Rotación de los factores.
6. **Análisis Discriminante**: Clasificación entre dos poblaciones. Poblaciones normales: función lineal discriminante para clasificar dos o más poblaciones. Variables canónicas discriminantes: dos o mas grupos.
7. **Análisis de Conglomerados**: Métodos clásicos de partición. Métodos jerárquicos y no jerárquicos. Algoritmos de partición.
8. **Correlaciones Canónicas**: Construcción de variables canónicas. Propiedades. Estimación. Tests de hipótesis de interés.
9. **Detección de outliers**.

Introducción

- Llamaremos forma cuadrática a una expresión escalar del tipo $x'Ax$ donde x es un vector y A es una matriz cuadrada y simétrica.
- La forma cuadrática es siempre un escalar y su expresión general es la siguiente:

$$\sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_{ij}x_i x_j$$

- Diremos que una matriz A es semidefinida positiva si cualquier forma cuadrática formada a partir de ella es un número no negativo, para todo $x \neq 0$.
- Si la forma cuadrática es siempre un número positivo diremos que la matriz A es definida positiva.
- El determinante y la traza de matrices definidas positivas son números positivos.

Introducción

- Las matrices ortogonales son matrices cuadradas que pueden representar un giro en el espacio o simetría con respecto a un plano.
- Dado un vector x, si lo multiplicamos por matriz no singular C obtenemos un nuevo vector $y = Cx$.
- Si esta operación es un giro, la norma de y debe ser idéntica a la de x que implica la condición $y'y = x'C'Cx = x'x$, es decir, se debe verificar que $C'C = I$.
- A partir de y deducimos que $x = C^{-1}y$.
- Luego, premultiplicando a y por C' obtenemos $C'y = C'Cx = x$.
- De ambas condiciones se obtiene la **condición de ortogonalidad:**

$$C' = C^{-1}$$

Introducción

- Dada una matriz cuadrada hay determinadas propiedades invariantes ante transformaciones lineales que preservan la información existente en la matriz.
- Los **autovalores** son las medidas básicas de tamaño de una matriz. Se puede demostrar que las medidas globales de tamaño de la matriz (traza, determinante), sólo dependen de los autovalores y en consecuencia serán también invariantes ante transformaciones que preservan los autovalores.
- Los **autovectores** representan las direcciones características de la matriz. Para cada matriz cuadrada existen ciertos vectores que al transformarlos por la matriz sólo se modifica su norma y no su posición en el espacio.

Introducción

- Llamaremos autovectores de una matriz cuadrada de orden n a aquellos vectores cuya dirección no se modifica al transformarlos (multiplicarlos) por la matriz.
- El vector u es un autovector de la matriz A si se verifica:

$$Au = \lambda u$$

donde λ es un escalar que llamamos autovalor de la matriz.

- Para calcular los autovectores se resuelve $(A - \lambda I)u = 0$ que es un sistema homogéneo de ecuaciones que tendrá solución no nula si y solo si la matriz del sistema $(A - \lambda I)$ es singular.
- Este sistema tiene solución no nula si se verifica $|A - \lambda I| = 0$

Introducción

- Cuando la matriz tiene n autovalores propios distintos, a cada autovalor le podemos asociar un autovector bien definido y esos autovectores son linealmente independientes.
- Algunas propiedades que satisfacen los autovalores de una matriz:
  - Si λ es un autovalor de A ,entonces λ^r es un autovalor de A^r. En particular, si A es no singular, $\lambda \neq 0$ y λ^{-1} es un autovalor de A^{-1}.
  - Los autovalores de una matriz y su transpuesta son iguales.
  - La suma de los autovalores de A es igual a la traza de A.
  - El producto de los autovalores de A es igual al determinante de A.
  - Si A es una matriz simétrica, los autovalores son siempre reales y los autovectores son siempre ortogonales.
  - Si P es no singular, las matrices A y $P^{-1}AP$ tienen los mismos autovalores.

Introducción

- Las matrices simétricas pueden diagonalizarse mediante una transformación ortogonal.
- Sea A una matriz cuadrada y simétrica de orden n. Esta matriz tiene autovalores reales y autovectores ortonormales.
- Los autovectores $u_1, u_2, ..., u_n$ son linealmente independientes y forman una base en R^n.

$$A[u_1, u_2, ...u_n] = [\lambda_1 u_1, \lambda_2 u_2, ..., \lambda_n u_n]$$

donde λ son los autovalores reales que pueden ser no todos diferentes, de hecho algunos autovalores pueden ser nulos.

Introducción

- Llamando D a la matriz diagonal de términos λ_i a ecuación anterior puede escribirse como $AU = UD$ donde U es ortogonal.
- Multiplicando por $U' = U^{-1}$ se obtiene que $U'AU = D$, transformando así la matriz original en una matriz diagonal, mediante la matriz ortogonal U.
- Al diagonalizar una matriz simétrica:
  - los autovalores representan las constantes por las que se han multiplicado los vectores ortonormales iniciales.
  - los autovectores indican el giro o rotación realizado.

Introducción

Analicemos el siguiente determinante:

$$|U'||A||U| = |D|$$

como $|U'| = |U| = 1$ el determinante de A sera el producto de sus raíces características.

Por lo tanto,

- si un autovalor es nulo el determinante es igual a 0 y la matriz A es singular (no invertible)
- el rango de A sera igual al de D, que al ser diagonal será igual al número autovalores no nulos.
- El rango de una matriz simétrica es igual al número de autovalores no nulos.

Estadísticos descriptivos

- El análisis descriptivo univariado habitualmente comienza con el cálculo de la media:

$$\bar{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

- Para una variable binaria es la frecuencia relativa de aparición del atributo de interés y en el caso de variables numéricas, representa el centro geométrico de los datos.
- También es habitual calcular una medida de variabilidad con relación a la media. Para ello se calculan las desviaciones:

$$d_{ij} = (x_{ij} - \bar{x_j})^2$$

Estadísticos descriptivos

- El desvío estándar representa un promedio de desviaciones

$$s_j = \left(\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \right)^{1/2} = \left(\frac{1}{n} \sum_{i=1}^{n} d_{ij} \right)^{1/2}$$

- Para comparar la variabilidad de distintas variables conviene calcular una medida de variabilidad relativa, que no dependa de las unidades de medida:

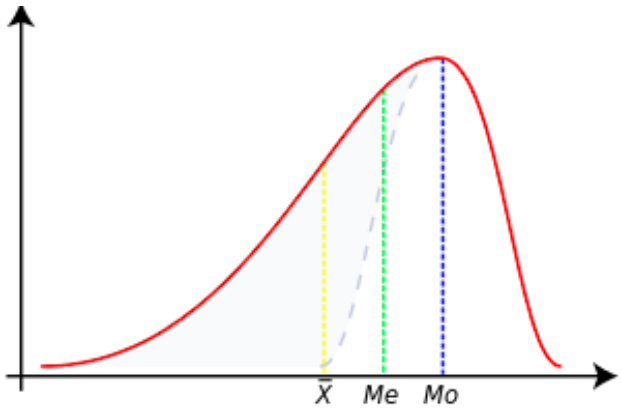
$$CV_j = \frac{s_j}{\bar{x}_j}$$

Estadísticos descriptivos

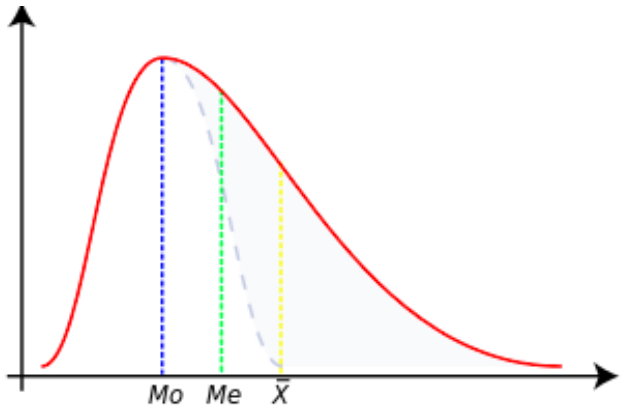
- Por otro lado conviene calcular algún coeficiente que mida la asimetría de los datos con respecto a su centro, la media:

$$A_j = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^3}{s_j^3}$$

- Este coeficiente es igual a cero para una variable simétrica.
- Si el valor de este coeficiente es positivo y mayor a 1 los datos tienen una distribución claramente asimétrica hacia la derecha.
- Si el valor de este coeficiente es negativo y menor a -1 los datos tienen una distribución claramente asimétrica hacia la izquierda.



Asimetría negativa



Asimetría positiva

Estadísticos descriptivos

- Otra característica importante de un conjunto de datos es su homogeneidad. Si las desviaciones d_{ij} son grandes o muy distintas, esto sugiere que hay datos que se separan mucho de la media y por lo tanto tenemos alta heterogeneidad. Una posible medida de homogeneidad es la variancia de las d_{ij}:

$$\frac{1}{n}\sum_{i=1}^{n}(d_{ij} - s_j^2)^2$$

- Por lo que vimos antes, s_j^2 es el promedio de las desviaciones. Podemos calcular una medida de homogeneidad análoga al coeficiente de variación, que no dependa de las unidades de medida.

$$H_j = \frac{1}{n}\frac{\sum_{i=1}^{n}(d_{ij} - s_j^2)^2}{(s_j^2)^2}$$

Estadísticos descriptivos

$$H_j = \frac{1}{n} \frac{\sum_{i=1}^{n} (d_{ij} - s_j^2)^2}{(s_j^2)^2}$$

- Desarrollando el cuadrado del numerador:

$$\sum_{i=1}^{n} (d_{ij} - s_j^2)^2 = \sum_{i=1}^{n} d_{ij}^2 + n s_j^4 - 2 s_j^2 \sum_{i=1}^{n} d_{ij} = \sum_{i=1}^{n} d_{ij}^2 - n s_j^4$$

- Podemos calcular el coeficiente de homogeneidad de la siguiente manera:

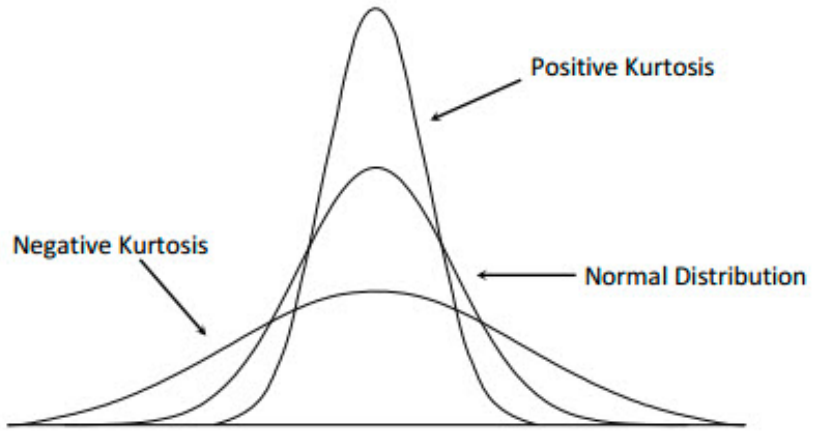
$$H_j = \frac{1}{n} \frac{\sum_{i=1}^{n} (d_{ij} - s_j^2)^2}{s_j^4} = \frac{1}{n} \frac{\sum_{i=1}^{n} d_{ij}^2 - n s_j^4}{s_j^4} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^4}{s_j^4} - 1$$

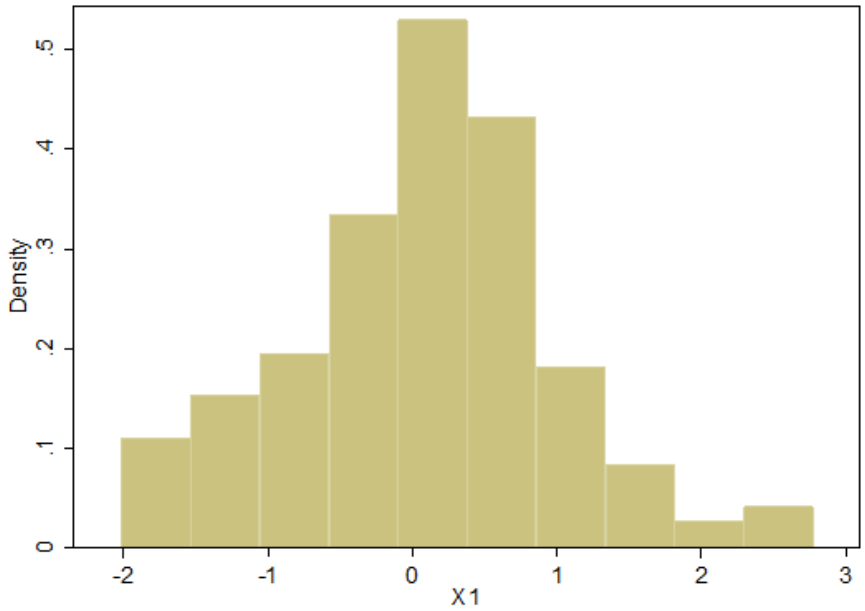
$$H_j = K_j - 1$$

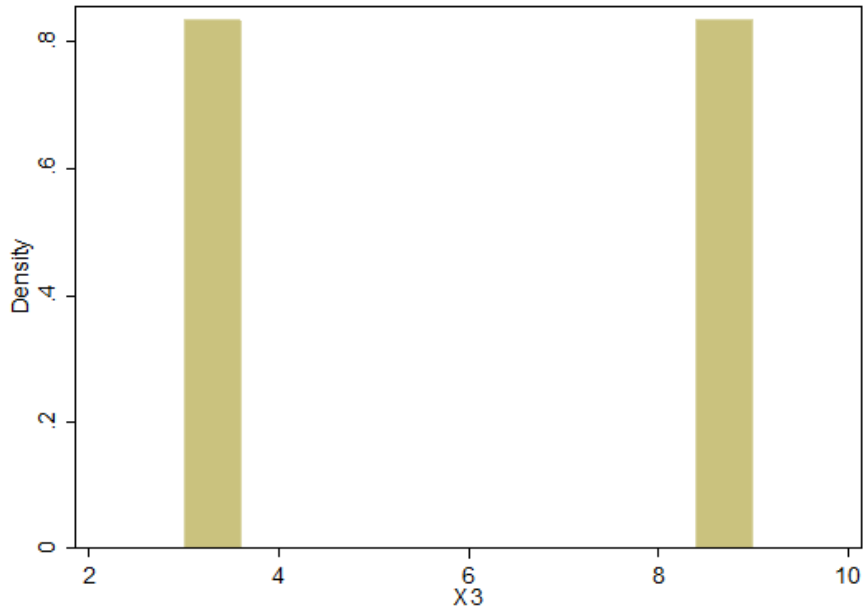
Estadísticos descriptivos

$$H_j = K_j - 1$$

- K_j es el coeficiente de Kurtosis. $K_j \geq 1$ ya que $H_j \geq 0$.
- Ambos coeficientes miden la relación entre la variabilidad de las desviaciones y la desviación media.
- Si hay unos pocos datos atípicos muy alejados del resto, los coeficientes de homogeneidad o kurtosis serán altos.
- En el extremo, si los datos se separan en dos grupos, H_j será pequeño.







Estadístico	X1	X1 bis X1 con outlier = 1000	X1 bis2 X1 con outlier = 100	X1 bis3 X1 con outlier = 10	X3
mínimo	-2.014	-2.014	-2.014	-2.014	3
máximo	2.774	1000.000	100.000	10.000	9
media	0.070	6.747	0.747	0.147	6
desvío estándar	0.928	81.648	8.210	1.225	3.010
CV	13.206	12.102	10.992	8.335	0.502
m4	2.38	6,488,563,620	646,975	65.19	81
asimetría - Aj	0.05	12.25	12.02	3.47	0.00
kurtosis - Kj	3.26	147.97	144.33	29.37	1
homogeneidad - Hj	2.26	146.97	143.33	28.37	0

Datos multivariados

- El principal objeto de trabajo en análisis multivariado es el conjunto de n observaciones para p variables.
- Sea x_i la i-ésima observación o elemento del conjunto ($i = 1, 2, \ldots, p$), el vector x se representa como:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

- Un vector $x_{p\times 1}$ es un conjunto ordenado de p números reales que representan una posición en un espacio p-dimensional V_p.

La matriz de datos

- Supondremos que hemos observado p variables numéricas en un conjunto de n elementos. El conjunto de las p variables conforman una variable vectorial o vector de variables.
- Esta información se representa en la matriz $X(n \times p)$, que llamaremos matriz de datos. Cualquier elemento genérico x_{ij} de esta matriz representa el valor de la variable j sobre el individuo i.
- La matriz de datos se puede representar de dos maneras diferentes. Por filas:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \ddots & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$$

- o por columnas:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix}$$

Ejemplo: Algunos indicadores socioeconómicos en centros urbanos de Argentina

- Fuente: Encuesta Permanente de Hogares, cuarto trimestre y segundo semestre de 2020.

Lista de variables

X1	Tasa de actividad
X2	Tasa de empleo
X3	Tasa de desocupación
X4	Porcentaje de hogares por debajo de la línea de pobreza
X5	Porcentaje de personas por debajo de la línea de pobreza
X6	Porcentaje de hogares con acceso a computadora
X7	Porcentaje de hogares con acceso a internet

Estadísticos descriptivos

Vars	Promedio	Desv. Std	CV	Asimetría	Kurtosis
X1	44.11	3.63	0.08	-0.84	2.74
X2	40.60	3.02	0.07	-0.71	3.09
X3	7.84	3.10	0.40	0.42	-0.80
X4	28.81	5.81	0.20	-0.24	1.73
X5	38.25	7.24	0.19	-0.61	2.19
X6	63.90	8.40	0.13	0.59	-0.32
X7	89.95	4.39	0.05	-0.41	-0.52

Matriz de datos

Aglomerados urbanos	id	X1	X2	X3	X4	X5	X6	X7
Ciudad Autónoma de Buenos Aires	1	51.4	47.7	7.2	12.2	16.5	82.8	96.0
Partidos del GBA	2	43.4	37.3	14.1	40.9	51.0	59.1	88.6
Gran Mendoza	3	49.5	44.3	10.6	32.6	44.0	59.2	93.1
Gran San Juan	4	44.1	41.9	5.2	24.9	34.8	52.8	80.5
Gran San Luis	5	44.8	42.6	4.9	32.4	40.6	77.6	91.8
Corrientes	6	40.1	37.4	6.7	32.2	42.9	55.2	83.1
Formosa	7	32.1	30.7	4.2	25.7	36.4	52.0	84.6
Gran Resistencia	8	42.3	40.0	5.3	40.3	53.6	57.5	92.9
Posadas	9	46.0	43.1	6.4	27.6	37.7	59.9	90.1
Gran Catamarca	10	43.2	40.6	6.1	28.7	35.7	56.3	86.2
Gran Tucumán - Tafí Viejo	11	43.3	39.2	9.5	33.8	43.5	51.4	92.6
Jujuy - Palpalá	12	43.1	41.4	4.0	27.4	37.7	67.2	93.1
La Rioja	13	41.4	39.6	4.3	25.3	35.3	70.9	93.4
Salta	14	44.8	40.7	9.0	31.2	41.7	61.9	91.3
Santiago del Estero - La Banda	15	41.1	39.5	3.9	31.4	39.4	60.1	92.6
Bahía Blanca - Cerri	16	48.1	43.4	9.7	18.7	24.0	63.5	86.5
Concordia	17	39.7	36.3	8.6	39.3	49.5	54.3	84.5
Gran Córdoba	18	47.9	41.7	13.0	29.5	40.8	62.7	90.5
Gran La Plata	19	48.0	43.6	9.1	24.0	31.7	69.9	91.0
Gran Rosario	20	46.7	40.3	13.6	29.1	38.3	61.0	83.8
Gran Paraná	21	41.4	39.8	4.0	30.4	40.9	76.4	93.2
Gran Santa Fe	22	44.9	41.4	7.8	28.0	39.8	61.5	91.7
Mar del Plata	23	47.8	42.5	11.1	30.5	41.1	63.8	87.7
Río Cuarto	24	46.3	42.0	9.2	27.2	39.2	61.8	91.3
Santa Rosa - Toay	25	45.6	40.5	11.2	24.9	33.5	63.8	87.1
San Nicolás - Villa Constitución	26	41.1	37.2	9.5	32.4	43.6	54.5	82.1
Comodoro Rivadavia - Rada Tilly	27	41.4	40.0	3.3	24.0	31.7	71.9	94.7
Neuquen - Plottier	28	44.8	41.1	8.4	32.1	40.4	63.3	86.2
Río Gallegos	29	43.1	40.1	6.8	26.0	33.2	69.8	95.7
Ushuaia - Río Grande	30	44.4	38.7	12.8			80.9	98.4
Rawson - Trelew	31	48.0	45.4	5.4	25.2	32.0	74.8	92.3
Viedma - Carmen de Patagones	32	41.9	39.4	6.0	25.2	35.1	67.5	92.0

Estadísticos descriptivos

- En el caso multivariado, la medida descriptiva de posición central más utilizada es el vector de medias.

$$\bar{x} = \frac{1}{n} X'1 = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

- Se define la matriz de variancias y covariancias:

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & \ddots & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ s_{p1} & \dots & \dots & s_p^2 \end{pmatrix}$$

Es una matriz cuadrada y simétrica y semidefinida positiva.

Estadísticos descriptivos

- La matriz de datos centrada se define como la diferencia entre la matriz de datos X y la media de cada observación.

$$\tilde{X} = X - 1\bar{x}' = X - \frac{1}{n}11'X = (I - \frac{1}{n}11')X = PX$$

La matriz P es simétrica e idempotente (es decir, $PP = P$)

$$S = \frac{1}{n}\tilde{X}'\tilde{X} = \frac{1}{n}X'P'PX = \frac{1}{n}X'PX$$

- La matriz de variancias y covariancias es semidefinida positiva.
- Si y es cualquier vector entonces $y'Sy \geq 0$.
- La traza, el determinante y los autovalores de S son no negativos.

Matriz de datos centrada

Aglomerados urbanos	id	X1	X2	X3	X4	X5	X6	X7
Ciudad Autónoma de Buenos Aires	1	7.2	7.1	-0.6	-16.6	-21.7	18.9	6.1
Partidos del GBA	2	-0.7	-3.3	6.3	12.1	12.8	-4.8	-1.4
Gran Mendoza	3	5.4	3.7	2.8	3.8	5.8	-4.7	3.1
Gran San Juan	4	0.0	1.3	-2.7	-3.9	-3.4	-11.1	-9.5
Gran San Luis	5	0.7	2.0	-3.0	3.6	2.4	13.7	1.8
Corrientes	6	-4.0	-3.2	-1.1	3.4	4.7	-8.8	-6.8
Formosa	7	-12.0	-9.9	-3.6	-3.1	-1.8	-12.0	-5.3
Gran Resistencia	8	-1.8	-0.6	-2.5	11.5	15.4	-6.4	2.9
Posadas	9	1.9	2.5	-1.5	-1.2	-0.5	-4.0	0.2
Gran Catamarca	10	-0.9	0.0	-1.8	-0.1	-2.5	-7.6	-3.7
Gran Tucumán - Tafí Viejo	11	-0.8	-1.4	1.6	5.0	5.3	-12.5	2.7
Jujuy - Palpalá	12	-1.0	0.8	-3.8	-1.4	-0.5	3.2	3.2
La Rioja	13	-2.7	-1.0	-3.5	-3.5	-2.9	7.0	3.4
Salta	14	0.6	0.1	1.1	2.4	3.5	-2.0	1.3
Santiago del Estero - La Banda	15	-3.0	-1.1	-3.9	2.6	1.2	-3.8	2.6
Bahía Blanca - Cerri	16	4.0	2.8	1.9	-10.1	-14.2	-0.4	-3.5
Concordia	17	-4.4	-4.3	0.8	10.5	11.3	-9.6	-5.4
Gran Córdoba	18	3.8	1.1	5.2	0.7	2.6	-1.2	0.5
Gran La Plata	19	3.9	3.0	1.3	-4.8	-6.5	5.9	1.0
Gran Rosario	20	2.6	-0.3	5.8	0.3	0.1	-2.9	-6.2
Gran Paraná	21	-2.7	-0.8	-3.9	1.6	2.7	12.5	3.3
Gran Santa Fe	22	0.8	0.8	0.0	-0.8	1.6	-2.4	1.7
Mar del Plata	23	3.7	1.9	3.3	1.7	2.9	-0.1	-2.3
Río Cuarto	24	2.2	1.4	1.4	-1.6	1.0	-2.1	1.3
Santa Rosa - Toay	25	1.5	-0.1	3.4	-3.9	-4.7	-0.1	-2.8
San Nicolás - Villa Constitución	26	-3.0	-3.4	1.6	3.6	5.4	-9.4	-7.8
Comodoro Rivadavia - Rada Tilly	27	-2.7	-0.6	-4.6	-4.8	-6.5	8.0	4.7
Neuquen - Plottier	28	0.7	0.5	0.5	3.3	2.2	-0.6	-3.8
Río Gallegos	29	-1.1	-0.5	-1.0	-2.8	-5.0	5.9	5.8
Ushuaia - Río Grande	30	0.3	-1.9	5.0			17.0	8.5
Rawson - Trelew	31	3.9	4.8	-2.4	-3.6	-6.2	10.8	2.3
Viedma - Carmen de Patagones	32	-2.2	-1.2	-1.8	-3.6	-3.1	3.5	2.0

Ejemplo: Algunos indicadores socioeconómicos en centros urbanos de Argentina

$$\bar{x} = \frac{1}{n} X'1 = \begin{pmatrix} 44.11 \\ 40.60 \\ 7.84 \\ 28.81 \\ 38.25 \\ 63.90 \\ 89.95 \end{pmatrix}$$

- Matriz S de variancias y covariancias.

S: Matriz de variancias y covariancias

	X1	X2	X3	X4	X5	X6	X7
X1	13.60						
X2	10.36	9.31					
X3	5.25	0.84	9.10				
X4	-7.09	-8.17	3.84	33.72			
X5	-9.48	-10.75	4.58	41.09	52.47		
X6	11.90	13.92	-6.35	-23.60	-31.77	63.06	
X7	4.00	5.33	-3.65	-4.72	-6.01	21.47	17.49

Traza	198.74
Traza/p	28.39
Determinante	8707.76
Variancia efectiva	3.65

Estadísticos descriptivos

- Si la matriz S es singular existe una relación lineal exacta entre las variables.
- Para cada observación i se verifica $w'(x_i - \bar{x}) = 0$, es decir $\tilde{X}w = 0$.
- Multiplicando por $\tilde{X}'$ y dividendo por n:

$$\frac{1}{n}\tilde{X}'\tilde{X}w = Sw = 0$$

- La matriz S tiene un autovalor igual a 0 y w es el autovector asociado a ese autovalor nulo. Si multiplicamos la expresión anterior por w' se obtiene:

$$\frac{1}{n}w'\tilde{X}'\tilde{X}w = w'Sw = 0$$

.

- Es posible reducir la dimensión del conjunto de datos eliminando esa variable. Las coordenadas del vector w nos indican la combinación lineal redundante.

Estadísticos descriptivos

Generalizando, si S tiene rango $h < p$:

- Existen $p - h$ variables redundantes que pueden eliminarse.
- Si S tiene h autovalores distintos de 0 y existirán $r = p - h$ vectores no nulos que representan r combinaciones lineales exactas entre las variables y verifican:

$$Sw_i = 0 \qquad \forall i = 1, 2, ..., r$$

- Es posible representar las observaciones de la matriz de datos a partir de $h = p - r$ variables.
- Cuando hay más de un autovalor nulo, las relaciones lineales entre las variables no están definidas unívocamente:

$$S(a_1 w_1 + a_2 w_2 + ... + a_r w_r) = 0$$

Estadísticos descriptivos

- Una forma alternativa de analizar el problema. Como

$$S = \frac{1}{n}\tilde{X}'\tilde{X}$$

El rango de S coincide con el de la matriz de datos centrada, ya que para cualquier matriz A, si llamamos $rg(A)$ al rango de A, siempre se verifica que:

$$rg(A) = rg(A') = rg(A'A) = rg(AA')$$

Por lo tanto, si la matriz de datos centrada tiene rango p, este será también el rango de S.
Si existen r combinaciones lineales entre las variables X, el rango de la matriz de datos centrada será $h = p - r$ y este será también el rango de la matriz S.

Medidas de variabilidad conjunta

- Cuando las variables están expresadas todas en la misma unidad de medida puede resultar interesante encontrar una medida de la variabilidad global o promedio que permitan comparar distintos conjuntos de datos.
  Se define la **varianza total** de los datos por medio de la traza de S:

$$T = traza(S) = \sum_{i=1}^{p} s_i^2$$

Mientras que la **variancia media** es:

$$\bar{s}^2 = \frac{1}{p} \sum_{i=1}^{p} s_i^2 = \frac{T}{p}$$

Medidas de variabilidad conjunta

- Un inconveniente con este tipo de medida es que no se tiene en cuenta la estructura de dependencia entre las variables.
- Si la dependencia entre las variables es muy alta, la variabilidad conjunta de los datos es pequeña ya que conociendo una variable podemos determinar aproximadamente los valores de las demás.
- Una medida alternativa es la **variancia generalizada**, debida a Wilks. Se define como el determinante de la matriz de variancias y covariancias:

$$VG = |S|$$

Su raíz cuadrada es el desvío estándar generalizado.

Estadísticos descriptivos - Medidas de Variabilidad

La variancia generalizada cuenta con las siguientes propiedades:

- Está bien definida, ya que el determinante de la matriz de variancias y covariancias es siempre no negativo
- Es una medida de área (si $p = 2$), volumen (si $p = 3$) o hipervolumen (para $p > 3$) ocupado por el conjunto de datos.
- Si $p = 2$, el desvío estándar generalizado es

$$|S|^{1/2} = s_x s_y (1 - r^2)^{1/2}$$

- Si las variables son independientes, la mayoría de sus valores estarán dentro de un rectángulo cuyos lados tienen longitud $6s_x$ y $6s_y$, ya que por el teorema de Tchebychev entre la media y 3 desvíos estándar podemos encontrar al menos el 90 por ciento de los datos.

Estadísticos descriptivos - Medidas de Variabilidad

- En el otro extremo, si las variables están relacionadas linealmente y el coeficiente de correlación es distinto de cero, la mayoría de los puntos tenderán a situarse alrededor de una recta de regresión y habrá una reducción del área tanto mayor cuanto mayor sea R^2.
- En el límite, si R^2 es exactamente igual a 1, todos los puntos están ubicados sobre una linea recta, la relación entre las variables es exacta y el área ocupada es igual a 0.
- Recordar que en este caso S es singular ya que $p = 2$ pero su rango es igual a 1.

Estadísticos descriptivos - Medidas de Variabilidad

- Un inconveniente de la variancia generalizada es que no sirve para comparar conjuntos de datos con distinta cantidad de variables.
- Si a un conjunto de datos con p variables le agregamos una variable más, no correlacionada con las anteriores y de variancia igual a s^2_{p+1} es fácil comprobar que se satisface lo siguiente:

$$|S_{p+1}| = |S_p| s^2_{p+1}$$

y eligiendo las unidades de medida de la variable $p+1$ podemos hacer que la variancia generalizada aumente o disminuya a voluntad.

- Para evitar estos inconvenientes, se ha propuesto otra medida global de variabilidad denominada **variancia efectiva**.

$$VE = |S|^{1/p}$$

Estadísticos descriptivos - Medidas de Variabilidad

- Para matrices diagonales (p variables no correlacionadas entre si), la variancia efectiva es la media geométrica de las variancias de las variables.
- El determinante de S es el producto de sus autovalores, por lo tanto la variancia efectiva también es la media geométrica de los autovalores que por ser semidefinida positiva serán siempre no negativos.
- La variancia efectiva tiene en cuenta la estructura de dependencia de las variables ya que si una variable fuera combinación lineal de las restantes, al existir un autovalor nulo se obtendrá $VE = 0$, mientras que la variancia media será no nula.
- La variancia efectiva siempre es menor a la variancia media.

Ejemplo: Algunos indicadores socioeconómicos en centros urbanos de Argentina

- Matriz S de variancias y covariancias.

S: Matriz de variancias y covariancias

	X1	X2	X3	X4	X5	X6	X7
X1	13.60						
X2	10.36	9.31					
X3	5.25	0.84	9.10				
X4	-7.09	-8.17	3.84	33.72			
X5	-9.48	-10.75	4.58	41.09	52.47		
X6	11.90	13.92	-6.35	-23.60	-31.77	63.06	
X7	4.00	5.33	-3.65	-4.72	-6.01	21.47	17.49

Traza	198.74
Traza/p	28.39
Determinante	8707.76
Variancia efectiva	3.65

Medidas de dependencia

Uno de los objetivos del análisis multivariado es comprender la estructura de dependencias entre las variables. Estas dependencias pueden darse:

- entre pares de variables $\rightarrow$ matriz de correlación
- entre una variable con respecto a las restantes $\rightarrow$ regresión
- entre pares de variables, eliminando el efecto de las demás variables $\rightarrow$ correlaciones parciales
- entre el conjunto de todas las variables $\rightarrow$ coeficiente de dependencia conjunta

Dependencia entre pares de variables

- El coeficiente de correlación entre dos variables X e Y se define como

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- La dependencia entre pares de variables se mide a traves de la matriz de correlaciones. Es una matriz cuadrada, simétrica y semidefinida positiva (al igual que S).

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

$$R = D^{-1/2} S D^{-1/2}$$

donde $D = D(S)$ es una matriz diagonal que contiene los elementos diagonales de la matriz de variancias y covariancias S.

Dependencia conjunta

- Finalmente se puede obtener una medida conjunta de la dependencia entre las variables. Podemos utilizar el determinante de la matriz de correlación, que mide el alejamiento del conjunto de variables de la situación de perfecta dependencia lineal.
- Se puede demostrar $0 \leq |R| \leq 1$ y además:
  - Si las variables están todas no correlacionadas, R es una matriz identidad de orden p y $|R| = 1$.
  - Si una variable es combinación lineal del resto, S y R son singulares y por lo tanto $|R| = 0$
  - En el caso general, se puede demostrar que:

$$|R_p| = (1 - R^2_{p/1,2,3,\ldots,p-1})(1 - R^2_{p-1/1,2,3,\ldots,p-2}) \ldots (1 - R^2_{1/2})$$

Dependencia conjunta

- De acuerdo con esta propiedad, $|R_p|^{\frac{1}{p-1}}$ representa la media geométrica de la proporción de variabilidad explicada por todas las regresiones anteriores.
- Se puede observar que también es la media geométrica de los autovalores de R_p, teniendo en cuenta que solo tenemos $p-1$ autovalores independientes ya que están ligados por la relación:

$$\sum_{i=1}^{p} \lambda_i = p$$

donde λ_i son los autovalores de R. Se define el *coeficiente de dependencia efectiva* como:

$$D(R_p) = 1 - |R_p|^{1/(p-1)}$$

- Si $p = 2$ $|R_2| = 1 - r_{12}^2$ y este coeficiente coincide con el cuadrado del coeficiente de correlación entre las dos variables.

Ejemplo: Algunos indicadores socioeconómicos en centros urbanos de Argentina

- Matriz R de correlaciones.

R: Matriz de correlaciones

	X1	X2	X3	X4	X5	X6	X7
X1	1.00						
X2	0.92	1.00					
X3	0.47	0.09	1.00				
X4	-0.33	-0.46	0.22	1.00			
X5	-0.36	-0.49	0.21	0.98	1.00		
X6	0.41	0.57	-0.27	-0.51	-0.55	1.00	
X7	0.26	0.42	-0.29	-0.19	-0.20	0.65	1.00

Traza	7
Traza/p	1
Determinante	0.000003876
Dependencia efectiva	0.87

Dependencia conjunta

- La matriz S^{-1} se denomina **matriz de precisión** y contiene información sobre la relación multivariada entre cada una de las variables y el resto. Puede demostrarse que esta matriz contiene información sobre:
- Por filas y fuera de la diagonal principal, los coeficientes de regresión múltiple de la variable correspondiente a esa fila, explicada por todas las demás.
- En la diagonal, las inversas de las variancias residuales de la regresión de cada variable con el resto.
- Si estandarizamos los elementos de esta matriz, los elementos fuera de la diagonal principal son los coeficientes de correlación parcial entre estas variables.
- Por lo tanto S^{-1} contiene toda la información sobre las regresiones de cada variable sobre las restantes.

Medidas de distancia

- Una familia de medidas de distancia muy habituales en R^p es la familia de métricas o distancias de Minkowski:

$$d_{ij}^{(r)} = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^r \right)^{1/r}$$

Cuando r = 2 obtenemos la distancia euclídea:

$$d_{ij}^{(2)} = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right)^{1/2} = [(x_i - x_j)'(x_i - x_j)]^{1/2}$$

Es la distancia más utilizada pero depende de las unidades de medida de las variables.

Medidas de distancia

- Una manera de evitar este problema es dividir cada variable por un término que elimine el efecto de escala. Esto conduce a la familia de métricas euclídeas ponderadas, que se definen como:

$$d_{ij} = [(x_i - x_j)' M^{-1} (x_i - x_j)]^{1/2}$$

Donde M es una matriz diagonal que se utiliza para estandarizar las variables y hacer la medida invariante ante cambios de escala. Por ejemplo, podemos colocar en la diagonal de M los desvíos estándar de las variables:

$$d_{ij} = \left(\sum_{h=1}^{p} \frac{(x_{ih} - x_{jh})^2}{s_h} \right)^{1/2}$$

Medidas de distancia

- Se define la distancia de Mahalanobis entre un punto y su vector de medias de la siguiente manera:

$$d_i = [(x_i - \bar{x})' S^{-1} (x_i - \bar{x})]^{1/2}$$

Consideremos que p=2:

$$S^{-1} = \frac{1}{(1-r^2)} \begin{pmatrix} s_1^{-2} & -r s_1^{-1} s_2^{-1} \\ -r s_1^{-1} s_2^{-1} & s_2^{-2} \end{pmatrix}$$

Y la distancia al cuadrado entre dos puntos (x_1, y_1), (x_2, y_2) es:

$$d_M^2 = \frac{1}{(1-r^2)} \left[\frac{(x_1 - x_2)^2}{s_1^2} + \frac{(y_1 - y_2)^2}{s_2^2} - 2r \frac{(x_1 - x_2)(y_1 - y_2)}{s_1 s_2} \right]$$

- Si $r = 0$, la distancia de mahalanobis es la distancia euclídea estandarizando las variables por sus desvíos estándar.

Asimetría y Kurtosis

- La generalización de estos coeficientes para el caso multivariado no es inmediata. A continuación vemos una de las propuestas más utilizadas y se debe a Mardia (1970).
- Se propuso calcular las distancias de Mahalanobis para cada par de elementos muestrales (i, j):

$$d_{ij} = [(x_i - \bar{x})' S^{-1} (x_j - \bar{x})]$$

Coeficiente de asimetría multivariante:

$$A_p = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^3$$

Coeficiente de kurtosis multivariante:

$$K_p = \frac{1}{n} \sum_{i=1}^{n} d_{ii}^2$$

Asimetría y Kurtosis

Propiedades:

- Para variables escalares, $A_p = A^2$:

$$A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{(x_i-\bar{x})(x_j-\bar{x})}{s^2}\right]^3 =$$

$$A_p = \frac{1}{n^2 s^6}\left[\sum_{i=1}^{n}(x_i-\bar{x})^3\right]^2 = A^2$$

- El coeficiente de asimetría es no negativo, y solo será igual a cero si los datos son simétricos.

Asimetría y Kurtosis

Propiedades:

- Para variables escalares $K = K_p$.

$$d_{ii} = \left[\frac{(x_i - \bar{x})(x_i - \bar{x})}{s^2}\right]^2 = \frac{(x_i - \bar{x})^4}{s^4}$$

- Los coeficientes son invariantes ante transformaciones lineales de los datos.
- Si $y = ax + b$, los coeficientes de asimetría y kurtosis de x e y son idénticos.

Dependencia entre una variable con respecto al resto

- Hemos mencionado que si una variable es una combinación lineal exacta de las restantes, es posible predecir sus valores sin error.
- Consideremos que haya variables muy relacionadas con el resto y queremos medir su grado de dependencia.
- Sea x_p una de las variables del vector que puede estár relacionada con las restantes variables $(x_1, \ldots, x_{p-1})$
- El mejor predictor lineal de x_p a partir de las restantes variables es:

$$\hat{x}_p = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_{p-1}$$

siendo $\hat{\alpha} = \bar{x}_p - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 - \cdots - \hat{\beta}_p \bar{x}_{p-1}$.

Dependencia entre una variable con respecto al resto

- Los coeficientes estimados se determinan de manera que la ecuación proporcione la mejor predicción posible de los valores de x_p, minimizando los residuos de la estimación.
- Una medida de bondad de ajuste habitual que calculamos a partir de un modelo de regresión es el R^2, que indica cuanto de la variabilidad de la variable y se puede explicar por el modelo de regresión.
- Cuando el modelo tiene una sola variable explicativa, se puede demostrar fácilmente que el R^2 es igual al coeficiente de correlación entre x e y elevado al cuadrado.
- Esta medida también es un coeficiente de correlación múltiple (elevado al cuadrado).

$$R^2_{j/1,2,\ldots,j-1,j+1,\ldots,p} = 1 - \frac{SSError}{SSTotal} = 1 - \frac{s_r^2(j)}{s_j^2}$$

Dependencia entre una variable con respecto al resto

- Es posible demostrar que los términos diagonales de la matriz S^{-1} son las inversas de las variancias residuales de la regresión de cada variable con el resto. Es posible entonces calcular rápidamente los coeficientes de regresión múltiple al cuadrado de cada variable con respecto a las restantes a partir de los elementos de S y S^{-1}. Por ejemplo para la primera variable con respecto a las $p-1$ restantes:
  - Tomar el primer elemento diagonal de S, que es la variancia de la primera variable, s_1^2,
  - Invertir la matriz S y tomar el primer elemento diagonal de S^{-1} que llamaremos s^{11}. Este término es $1/s_r^2(1)$, la variancia residual del ajuste de regresión de la variable 1 en función de las restantes,
  - Calcular R_1^2

$$R_1^2 = 1 - \frac{1}{s^{11}s_1^2}$$

Dependencia directa entre dos variables

- La dependencia directa entre dos variables se mide a través del coeficiente de correlación parcial.
- Definimos $r_{12,3\ldots p}$ como el coeficiente de correlación parcial entre las variables x_1 y x_2, dadas las variables $(x_3, x_4, \ldots, x_p)$. Es el coeficiente de correlación entre x_1 y x_2 cuando se eliminan de estas variables los efectos de las variables restantes.
- Se puede demostrar que los coeficientes de correlación parcial entre cada par de variables se obtienen estandarizando los elementos de la matriz S^{-1}.
- Si s^{ij} son los elementos de S^{-1}, el coeficiente de correlación parcial entre las variables x_1, x_2 se obtiene de la siguiente manera:

$$r_{1,2/3,\ldots,p} = -\frac{s^{12}}{\sqrt{s^{11}s^{22}}}$$

Relación entre variables: edad, salario hora, nivel educativo (medido a través de los años de escolaridad) e ingresos familiar.

Matriz de Variancias y Covariancias (S):

	SALARIO HORA	INGRESO FAMILIAR	EDAD	EDUCACION
SALARIO HORA	**88.92**	10896.51	30.62	9.67
INGRESO FAMILIAR	10896.51	5177649.96	231.93	2825.58
EDAD	30.62	231.93	211.79	-9.96
EDUCACION	9.67	2825.58	-9.96	13.52

Matriz de Correlaciones (R):

	SALARIO HORA	INGRESO FAMILIAR	EDAD	EDUCACION
SALARIO HORA	1.00	0.51	0.22	**0.28**
INGRESO FAMILIAR	0.51	1.00	0.01	0.34
EDAD	0.22	0.01	1.00	-0.19
EDUCACION	0.28	0.34	-0.19	1.00

Matriz de Precisión (S^{-1}):

	SALARIO HORA	INGRESO FAMILIAR	EDAD	EDUCACION
SALARIO HORA	0.01683993	-0.00003119	-0.00275598	-0.00755402
INGRESO FAMILIAR	-0.00003119	0.00000028	0.00000262	-0.00003365
EDAD	-0.00275598	0.00000262	0.00537023	0.00537918
EDUCACION	-0.00755402	-0.00003365	0.00537918	0.09035497

Inversa de los elementos diagonales de la matriz de precisión:

SALARIO HORA	**59.38**
INGRESO FAMILIAR	3609659.25
EDAD	186.21
EDUCACION	11.07

Correlación Múltiple: cálculo del R^2 correspondiente al modelo:

$$Salario\_hora = \beta_0 + \beta_1\, ingreso\_familiar + \beta_2\, edad + \beta_3\, educacion + u$$

Utilizando la información de las matrices S y S^{-1}

$$R^2 = 1 - \frac{59.38}{88.92} = 0.3322$$

Matriz de Precisión (S^{-1}):

	SALARIO HORA	INGRESO FAMILIAR	EDAD	EDUCACION
SALARIO HORA	0.01683993	-0.00003119	-0.00275598	-0.00755402
INGRESO FAMILIAR	-0.00003119	0.00000028	0.00000262	-0.00003365
EDAD	-0.00275598	0.00000262	0.00537023	0.00537918
EDUCACION	-0.00755402	-0.00003365	0.00537918	0.09035497

$$Salario\_hora = \beta_0 + \beta_1\, ingreso\_familiar + \beta_2\, edad + \beta_3\, educacion + u$$

Además es posible recuperar los coeficientes estimados del modelo:

Variables explicativas	*Coeficientes estimados*
INGRESO FAMILIAR	- (- 0.00003119) / 0.01683993 = 0.0018524
EDAD	- (-0.00275598) / 0.01683993 = 0.1636574
EDUCACION	- (-0.00755402) / 0.01683993 = 0.4485780

Ajustes de modelos de regresión

Dependent Variable: SALARIO_HORA Included observations: 3167				
	Coefficient	*Std. Error*	*t-Statistic*	*Prob.*
INGRESO_FAMILIAR	**0.001852**	6.42E-05	28.87243	0.0000
EDAD	**0.163657**	0.009610	17.02976	0.0000
EDUCACION	**0.448578**	0.040407	11.10151	0.0000
C	-8.735059	0.624737	-13.98198	0.0000
R-squared	**0.332150**	S.E. of regression		7.710885
Adjusted R-squared	0.331517	Sum squared resid		188064.9

Correlación parcial entre salario hora y nivel educativo a partir de la matriz de precisión:

	SALARIO HORA	INGRESO FAMILIAR	EDAD	EDUCACION
SALARIO HORA	**0.01683993**	-0.00003119	-0.00275598	**-0.00755402**
INGRESO FAMILIAR	-0.00003119	0.00000028	0.00000262	-0.00003365
EDAD	-0.00275598	0.00000262	0.00537023	0.00537918
EDUCACION	-0.00755402	-0.00003365	0.00537918	**0.09035497**

$$R^2_{salario\_hora,educación/ingreso\_familiar,edad} = -\frac{-0.00755402}{\sqrt{0.01683993 \times 0.09035497}} = 0.193656$$

Dependent Variable: SALARIO_HORA				
Included observations: 3176				
	Coefficient	Std. Error	t-Statistic	Prob.
INGRESO_FAMILIAR	0.002099	6.13E-05	34.24078	0.0000
EDAD	0.142136	0.009569	14.85423	0.0000
C	-3.536582	0.421325	-8.393944	0.0000
R-squared	0.305973			

Dependent Variable: EDUCACION				
Included observations: 3167				
	Coefficient	Std. Error	t-Statistic	Prob.
INGRESO_FAMILIAR	0.000548	2.65E-05	20.67855	0.0000
EDAD	-0.047638	0.004142	-11.49986	0.0000
C	11.57772	0.182172	63.55366	0.0000
R-squared	0.149587			

Dependent Variable: RESID_EDUCACION				
Included observations: 3167				
	Coefficient	Std. Error	t-Statistic	Prob.
RESID_SALARIO_HORA	0.083604	0.007528	11.10502	0.0000
C	1.40E-05	0.059134	0.000236	0.9998
R-squared	**0.037503**			

En este modelo auxiliar, el R^2 es el coeficiente de correlación al cuadrado entre las dos variables de residuos.

$$Correlación(resid\_salario\_hora, resid\_educación) = \sqrt{0.037503} = 0.193657$$