CAPITULO IV

4. ANÁLISIS MULTIVARIADO DE LA POBLACIÓN INVESTIGADA

4.1. INTRODUCCIÓN

Luego de realizar el análisis de cada una de las variables tratadas en el capítulo anterior, en este capítulo se utilizará técnicas multivariadas que nos permitirá analizar las interacciones que tienen entre sí las variables y su correspondiente influencia en las unidades respectivas.

Para poder realizar este análisis utilizaremos en este capítulo cuatro técnicas idóneas:

1. Correlación lineal
2. Análisis Bivariado
3. Tablas de Contingencia
4. Componentes Principales
5. Correlación Canónica

La primera técnica estadística multivariada nos indicara si las variables analizadas se encuentran o no relacionadas linealmente entre sí; la segunda técnica nos permitirá realizar el análisis de las frecuencias obtenidas entre las diferentes categorías de las dos variables seleccionadas; con la tercera técnica se determinara si existe independencia o no entre las variables de estudio; por medio de la cuarta técnica nos permitirá concluir si las variables de estudio pueden ser expresadas a través de factores que agrupen características de estas variables y reduzcan considerablemente el número de variables de estudio; y por medio de la quinta técnica nos permitirá conocer si existe algún tipo de asociación entre 2 grupos de variables.

Para este análisis multivariado se ha decidido eliminar las variables de NACIONALIDAD (IP5) y LENGUA (IP6) de la información personal de los trabajadores y servidores públicos del MEC, se decidió esto en base de que más del 99% de la información proporcionada por estos trabajadores poseen nacionalidad ecuatoriana y hablan el idioma español, y no serán aleatorias en la práctica.

4.2 Método de imputación de datos

Antes de comenzar aplicar las técnicas estadísticas multivariada, se encontró que existe evidencia de que los datos obtenidos en el Censo del Magisterio no fueron completos por falta de respuesta o no declaración por parte del informante, para lo cual se procedió a imputar bs datos de tal forma que sí la variable tiene menos del 10% de ausencia de datos, se aplicó una técnica de simulación denominada Método de la Transformada Inversa, que consiste en generar números aleatorios (0-1) y observar en que intervalo de la distribución de frecuencia acumulada de cada variable encaja este número, y de esta manera asignar la codificación correspondiente al intervalo donde encajó dicho número, para lo cual se procedió hacer un software que permita realizar esta técnica debido a la gran

cantidad de información con la que se trabaja en cada variable de la base de datos del censo del Magisterio Fiscal en el litoral ecuatoriano; y si la variable tiene más del 10% de ausencia de información, se le asignó el valor de 0, en donde este valor representa para nuestro análisis información no proporcionada por el trabajador del MEC, en las variables que cumplan con este eventualidad.

4.3 Definiciones de estadística multivariada

El análisis estadístico puede obtener conclusiones significativas al hacer uso de las técnicas multivariadas; el presente estudio utilizará estas técnicas para determinar si existe algún efecto en la interacción de las variables, para esto se necesita algunas definiciones importantes que se detallarán a continuación.

4.3.1 Vector aleatorio

Sean X_1, X_2,..., X_p, p variables aleatorias o características sujetas a investigación, entonces se procede a definir un vector p variado $\mathbf{X} \in \mathbf{R}^p$, el cual está compuesto por p variables, esto es:

$$\mathbf{X}^t = [\ X_1\ \ X_2\ \ \ldots X_p\ ]$$

4.3.2 Matriz de datos

Se define a la matriz $\mathbf{X}$ como el arreglo compuesto por p filas y n columnas, en donde cada elemento de este arreglo está determinado por x_{ij}, que representa la i-ésima variable o característica sujeta a investigación y la j-ésima unidad de investigación.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} = [\ \boldsymbol{X_1}\ \boldsymbol{X_2}\ \ldots \boldsymbol{X_n}\ ]$$

Donde $\boldsymbol{X_1}\ \boldsymbol{X_2}\ \ldots \boldsymbol{X_n}$ representa una muestra de tamaño n tomada de una población p-variadas de tamaño N.

4.3.3 Vector de medias

Sea $\mathbf{X}^t = [\ X_1\ \ X_2\ \ \ldots X_p\ ]$, un vector p variado, se define al vector de medias como el valor esperado de $\mathbf{X}$.

$$\boldsymbol{\mu} = E\left[\mathbf{X}\right] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_p] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

4.3.4 Matriz de varianzas y covarianzas

Sea $\mathbf{X}^t = [\ X_1\ X_2\ \ldots X_p\ ]$, un vector p variado, se define la matriz de varianzas y covarianzas como el valor esperado del producto de la diferencia del vector p-variado y su vector de medias con su traspuesta.

$$\boldsymbol{\Sigma} = \boldsymbol{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t\right]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_p) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p, X_1) & \operatorname{cov}(X_p, X_2) & \cdots & \operatorname{cov}(X_p, X_p) \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

donde σ_{ii} es la varianza de la i-ésima variable y σ_{ij} es la covarianza entre la i-ésima y j-ésima variable, además $\sigma_{ij} = \sigma_{ji}$, por lo tanto Σ es simétrica y diagonalizable ortogonalmente.

i = 1, 2, ..., p

j = 1, 2, ..., p

4.3.5 Matriz de correlación

Sea $\mathbf{\Sigma}$ la matriz de varianza y covarianza de un vector aleatorio $\mathbf{X}$ p-variado, se define $\mathbf{V^{1/2}}$ como una matriz diagonal.

$$\mathbf{V^{1/2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

donde $\sqrt{\sigma_{ii}}$ es la desviación estándar de la variable aleatoria $\mathbf{X}_i$, entonces se define a la matriz de correlación como:

$$\rho = (\mathbf{V}^{1/2})^{-1}\Sigma(\mathbf{V}^{1/2})^{-1} = \begin{bmatrix} \sigma_{11}/\sigma_1\sigma_1 & \sigma_{12}/\sigma_1\sigma_2 & \cdots & \sigma_{1p}/\sigma_1\sigma_p \\ \sigma_{21}/\sigma_2\sigma_1 & \sigma_{22}/\sigma_2\sigma_2 & \cdots & \sigma_{2p}/\sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}/\sigma_p\sigma_1 & \sigma_{p2}/\sigma_p\sigma_2 & \cdots & \sigma_{pp}/\sigma_p\sigma_p \end{bmatrix}$$

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

donde ρ_{ij} es el coeficiente de correlación entre las variables $\mathbf{X}_i$ y $\mathbf{X}_j$

i = 1, 2, ..., p

j = 1, 2, ..., p

4.3.6 Componentes principales

Componentes principales es una técnica estadística multivariada que permite la reducción de datos, algebraicamente es una combinación lineal de p variables aleatorias X_1 X_2 ... X_p, geométricamente componentes principales representa la elección de un nuevo sistema de coordenadas obtenidas al rotar el sistema original con el nuevo ejes de coordenadas formado por cada

componente. Los nuevos ejes representan la dirección de máxima variabilidad obtenida.

Sea $\mathbf{X}^t$ = [X_1 X_2 ... X_p] un vector p variado, y cada una de las variables que lo componen son variables aleatorias observables y no necesariamente normales; sea Σ la matriz de varianza y covarianza del vector p variado $\mathbf{X}$, y sea $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ los valores propios correspondientes a Σ.

Considere las siguientes combinaciones lineales:

$$Y_1 = \mathbf{a}_1{}^t\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

$$Y_2 = \mathbf{a}_2{}^t\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = \mathbf{a}_p{}^t\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

Se define a las varianzas y covarianzas de las componentes principales como:

$$Var\,(Y_i) = \mathbf{a}_i{}^t\Sigma\,\mathbf{a}_i \qquad i = 1, 2, \ldots, p$$

$$Cov\,(Y_i, Y_j) = \mathbf{a}_i{}^t\Sigma\,\mathbf{a}_j \qquad j = 1, 2, \ldots, p$$

Se definen a $Y_1, Y_2, ..., Y_p$ como las componentes principales, no correlacionadas y ortonormales entre sí, y además $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq ... \geq \text{Var}(Y_p) \geq 0$. Por lo cual, éstas deben cumplir con:

$$\| \mathbf{a}_i \| = 1 \qquad \textit{para } i = 1, 2, ..., p$$

$$\langle \mathbf{a}_i, \mathbf{a}_j \rangle = 0 \qquad \textit{para } i \neq j$$

donde $\| \mathbf{a}_i \|$ es la norma del vector $\mathbf{a}_i$ y $\langle \mathbf{a}_i, \mathbf{a}_j \rangle$ es el producto interno entre los vectores $\mathbf{a}_i$ y $\mathbf{a}_j$.

La primera componente principal es la combinación lineal $Y_1 = \mathbf{a}_1{}^t\mathbf{X}$ de máxima varianza, esto es que maximiza la varianza de Y_1, sujeta a que la norma del vector $\mathbf{a}_1$ sea unitaria.

La segunda componente principal es la combinación lineal $Y_2 = \mathbf{a}_2{}^t\mathbf{X}$, sujeta a que la norma del vector $\mathbf{a}_2$ sea unitaria, la $\text{Cov}(Y_1, Y_2) = 0$ y a que $\text{Var}(Y_2) \leq \text{Var}(Y_1)$.

La i-ésima componente principal es la combinación lineal $Y_i = \mathbf{a}_i{}^t\mathbf{X}$, sujeta a que la norma del vector $\mathbf{a}_i$ sea unitaria, la $\text{Cov}(Y_{i-1}, Y_i) = 0$ y a que $\text{Var}(Y_i) \leq \text{Var}(Y_{i-1})$.

para i = 2, 3, ..., p

Como resultados obtenemos que si $\boldsymbol{\Sigma}$ es la matriz de covarianzas asociada con el vector $\mathbf{X}^t = [\ X_1\ \ X_2\ \ \ldots X_p\ ]$, donde $\boldsymbol{\Sigma}$ tiene los pares de valores y vectores propios (λ_1,e_1), (λ_2,e_2), ..., (λ_p,e_p), donde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Entonces se puede probar que la i-ésima componente principal Y_i es igual a:

$$Y_i = \mathbf{e}_i^t\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p \quad i = 1, 2, ..., p$$

Y además:

$$Var\ (Yi) = \mathbf{e}_i^t\Sigma\ \mathbf{e}_i = \lambda_i \qquad i = 1, 2, \ldots, p$$

$$Cov\ (Yi,Yj) = \mathbf{e}_i^t\Sigma\ \mathbf{e}_j = 0 \qquad j = 1, 2, \ldots, p$$

El porcentaje total de la varianza contenida por la i-ésima componente principal, está dada por:

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

4.3.7 Correlación canónica

El análisis de correlación canónica es una técnica estadística multivariada que mide el grado de asociación entre 2 grupos de variables.

Sea $\mathbf{X}$ un vector aleatorio p variado ($\mathbf{X} \in \mathrm{R}^p$) el cual podemos particionarlo en 2 grupos de vectores de tamaño q y p-q respectivamente, de la siguiente forma:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_q \\ ..... \\ X_{q+1} \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} \\ ..... \\ \mathbf{X}^{(2)} \end{bmatrix}$$

Donde el vector de medias correspondiente a $\mathbf{X}^{(1)}$ y $\mathbf{X}^{(2)}$ es la partición del vector de medias $\mathbf{X} \in \mathrm{R}^p$, que es el siguiente:

$$\boldsymbol{\mu} = E\left[\mathbf{X}\right] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \\ \cdots\cdots \\ \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \hline \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

Donde:

$\mathbf{X}^{(1)}$ y $\boldsymbol{\mu}^{(1)} \in \mathrm{R}^q$

$\mathbf{X}^{(2)}$ y $\boldsymbol{\mu}^{(2)} \in \mathrm{R}^{p-q}$

Las matrices de varianzas para $\mathbf{X}^{(1)}$ y $\mathbf{X}^{(2)}$ son Σ_{11} y Σ_{22} respectivamente y la matriz de covarianza de $\mathbf{X}^{(1)}$ con $\mathbf{X}^{(2)}$ es $\Sigma_{12} = \Sigma_{21}^{t}$.

$$\Sigma = \left[\begin{array}{ccc|ccc} \sigma_{11} & \cdots & \sigma_{1q} & \sigma_{1,q+1} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} & \sigma_{q,q+1} & \cdots & \sigma_{qp} \\ \hline \sigma_{q+1,1} & \cdots & \sigma_{q+1,q} & \sigma_{q+1,q+1} & \cdots & \sigma_{q+1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \\ \sigma_{p1} & \cdots & \sigma_{pq} & \sigma_{p,q+1} & \cdots & \sigma_{pp} \end{array}\right] = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$$

Ahora consideremos las siguientes combinaciones lineales:

$$U = \mathbf{a}^t \mathbf{X}^{(1)}$$

$$V = \mathbf{b}^t \mathbf{X}^{(2)}$$

Y su respectiva varianza y covarianza son:

$$Var\ (U) = \mathbf{a}^t \Sigma_{12} \mathbf{a}$$

$$Var\ (V) = \mathbf{b}^t \Sigma_{12} \mathbf{b}$$

$$Cov\ (U,V) = \mathbf{a}^t \Sigma_{12} \mathbf{b}$$

Nosotros buscaremos coeficientes de **a** y **b** tal que:

$$Corr(U,V) = \frac{\mathbf{a}^t \Sigma_{12} \mathbf{b}}{\sqrt{\mathbf{a}^t \Sigma_{11} \mathbf{a}} \sqrt{\mathbf{b}^t \Sigma_{22} \mathbf{b}}}$$

Basándonos en esto, definimos:

El primer par de variables canónicas, es el par de combinaciones lineales U_1, V_1 que tiene varianza unitaria y que maximiza la correlación entre ambas.

El segundo par de variables canónicas, es el par de combinaciones lineales U_2, V_2 que tiene varianza unitaria y que maximiza la correlación entre ambas, y además en todos los casos no esta correlacionada con el primer par de variables canónicas.

El k-ésimo par de variables canónicas, es el par de combinaciones lineales U_k, V_k que tiene varianza unitaria y que maximiza la correlación entre ambas, y además en todos los casos no esta correlacionada con las k-1 pares de variables canónicas previas.

Se denomina a la correlación entre el k-ésimo par de variables canónicas la k-ésima correlación canónica.

Para encontrar los vectores **a** y **b** nos basamos en los siguientes resultados:

Suponga $q \leq p$ y que los vectores $\mathbf{X}^{(1)}$ y $\mathbf{X}^{(2)}$ tienen:

$$Cov\ (\mathbf{X}^{(1)}) = \Sigma_{11}$$
$$Cov\ (\mathbf{X}^{(2)}) = \Sigma_{22}$$
$$Cov\ (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \Sigma_{12} = \Sigma_{21}^{t}$$

Los coeficientes de los vectores **a** y **b**, para la combinación lineal

$$U = \mathbf{a}^t \mathbf{X}^{(1)}$$

$$V = \mathbf{b}^t \mathbf{X}^{(2)}$$

max Corr (U,V) = ρ_1^*

son: **a**, **b**

Logrando el k-ésimo par de variables canónicas:

$$U_k = \mathbf{e}^t_k \boldsymbol{\Sigma}_{11}^{-1/2} \mathbf{X}^{(1)}$$

$$V_k = \mathbf{f}^t_k \boldsymbol{\Sigma}_{22}^{-1/2} \mathbf{X}^{(2)}$$

Con:

Corr $(U_k, V_k) = \rho_k^*$

Donde ρ_k^* representa la k-ésima correlación canónica entre las k-ésimas variables canónicas, y además $\rho_1^{*2} \geq \rho_2^{*2} \geq .... \geq \rho_p^{*2}$ son los valores propios de la matriz resultado de la multiplicación de: $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ y $e_1, e_2, ..., e_p$ son los vectores propios asociados a ésta, y $f_1, f_2, ..., f_p$ son los vectores propios de la matriz obtenida de la multiplicación de $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$.

Las variables canónicas tienen las siguientes propiedades:

$$Var(U_k) = Var(V_k) = 1$$

$$Cov\,(U_k, U_l) = Cov\,(U_l, U_k) = 0 \quad k \neq l$$

$$Cov\,(V_k, V_l) = Cov\,(V_l, V_k) = 0 \quad k \neq l$$

$$Cov\,(U_k, V_l) = Cov\,(U_l, V_k) = 0 \quad k \neq l$$

$$para\ k, l = 1, 2, \ldots, p$$