UNIVERSIDAD ABIERTA INTERAMERICANA

Carrera: Licenciatura en Matemática

Modelo de análisis multivariado: un ejemplo de aplicación en el ámbito empresarial

Autor: Javier Alejandro Di Salvo

Director de Tesis: Dr. Guillermo Federico Umbricht

Tesis presentada para optar al título de Licenciado en Matemática.

Diciembre, 2022.

A todos aquellos que creen en la ciencia.

AGRADECIMIENTOS

En primer lugar quisiera agradecer a todos los docentes que me acompañaron en mi formación académica, en especial, destacando el aporte de la Dra. Samira Abdel Masih, por su empeño, enseñanza y por infundir su pasión por la matemática.

A mi director de tesis, el Dr. Guillermo Federico Umbricht, por su paciencia, dedicación y acompañamiento, quién me transmitió todas las herramientas necesarias para realizar mi trabajo académico.

A mis familiares y amigos que me apoyaron durante esta última etapa de mi carrera.

Javier A. Di Salvo

RESÚMEN

En esta tesis se busca comprender cuales son las causas que inciden en el volumen de ventas de la actividad productiva ganadera, que se desarrolla en diferentes regiones de la República Argentina.

Para ello, se utilizan diversas herramientas (tanto gráficas como analíticas) del análisis multivariado o la estadística multivariante. Estas herramientas permiten conocer cuáles son las variables influyentes en el volumen de ventas y de qué manera influyen. Este análisis resulta de suma importancia, dado que al conocer las variables significativas; estas pueden ser controladas para obtener estados deseados con respecto al volumen de ventas.

Palabras clave: Análisis multivariado, análisis de componentes principales, análisis factorial.

ÍNDICE GENERAL

Introducción……. Pág. 1

Capítulo 1. CONCEPTOS MATEMÁTICOS PRELIMINARES

1.1 Modelos matemáticos……..……. Pág. 3

1.1.1. Tipos de modelos…….……..... Pág. 4

1.2 Modelo de regresión lineal simple……..……..…..….. Pág. 5

1.2.1. Recta de regresión lineal…….…… Pág. 9

1.2.2. Caracterización de los residuos…….……..……..…..... Pág. 10

Capítulo 2. REGRESIÓN LINEAL MÚLTIPLE.

2.1. Modelo de regresión lineal múltiple…….…….....……..... Pág. 13

2.2. Ecuación de regresión lineal múltiple…..……..…..……..... Pág. 18

2.3. Caracterización de los residuos……..…….…….……..... Pág. 21

Capítulo 3. ANÁLISIS MULTIVARIADO.

3.1. Introducción al análisis multivariado……..... Pág. 23

3.2. Herramientas de análisis multivariado……..... Pág. 24

3.3. Análisis multivariado de la varianza……..……..……..……... Pág. 26

3.4. Análisis de componentes principales…….……..……..… Pág. 28

3.4.1. Representación de los componentes principales…….…… Pág. 30

3.4.2. Test de esfericidad de Bartlett…….……...……..… Pág. 32

3.4.3. Prueba de Kaiser-Meyer-Olkin (KMO) ..……..…..……...… Pág. 32

3.5. Método de análisis factorial...……..……..…..…… Pág. 33

3.5.1. Comunalidad……..……..…… Pág. 34

3.5.2. Determinación del número de factores……..…. Pág. 34

3.5.3. Pruebas de adecuación y de fiabilidad de los datos…… Pág. 35

3.5.4. Rotación de factores…….……..…. Pág. 36

Capítulo 4. ESTUDIO DE CASO: MODELO DE ANÁLISIS MULTIVARIADO.

4.1. Presentación del problema……..... Pág. 41

4.2. Análisis descriptivo preliminar……..... Pág. 41

4.3. Método 1. Análisis de componentes principales (ACP)…….…….. Pág. 45
4.3.1. Pruebas de adecuación muestral……..…..... Pág. 50
4.3.2. Obtención de los componentes principales……..… Pág. 50
4.3.3. Representación gráfica de los componentes principales…..… Pág. 53
4.4. Método 2. Análisis factorial (AF)……..……... Pág. 55
4.4.1. Obtención de factores…….. Pág. 56
4.4.2. Representación gráfica de los factores.……... Pág. 58
4.5. Discusión de los resultados…….…… Pág. 61
Conclusión general……...…..…… Pág. 65
Bibliografía…….. Pág. 67

INTRODUCCIÓN

La estadística multivariante o multivariada es una rama de las estadísticas que abarca la observación y el análisis simultáneo de más de una variable respuesta. La aplicación de la estadística multivariante es llamada análisis multivariante.

La estadística multivariante trata de comprender los diferentes objetivos y antecedentes de cada una de las diferentes formas de análisis multivariante y cómo se relacionan entre sí. La aplicación práctica de la estadística multivariante a un problema particular puede involucrar varios tipos de análisis univariados y multivariados para comprender las relaciones entre las variables y su relevancia para el problema que se está estudiando.

En este trabajo se utilizan diferentes herramientas del análisis multivariado para comprender cuales son las causas que inciden en el volumen de ventas de la actividad productiva ganadera, en la República Argentina.

Esta tesis está organizada de la siguiente manera. En el Capítulo 1 se incluyen brevemente una serie de conceptos fundamentales para la comprensión de este trabajo.

En el Capítulo 2 se explican las características principales del modelo de regresión lineal múltiple (o análisis univariado). En el Capítulo 3 se extiende este estudio al análisis multivariado y se introducen diversos conceptos que serán de suma utilidad a nuestros fines.

Por último, en el Capítulo 4 se aborda el problema de interés a partir de las herramientas introducidas en el Capítulo 3.

Capítulo 1

CONCEPTOS MATEMÁTICOS PRELIMINARES

Sin pretender ser exhaustivo y con la finalidad de que el trabajo sea claro y auto-contenido se presentan a continuación una serie de conceptos necesarios para comprender el desarrollo de esta tesis. En todos los casos se define el concepto en cuestión, se muestran ejemplos y se cita bibliografía para aquellos lectores que necesiten mayor especificación.

1.1. Modelos matemáticos

Un modelo es útil para representar un fenómeno de la naturaleza dentro de un marco teórico específico. Modelar matemáticamente un problema permite, entre otras cosas, poder comprender el grado en que se relacionan las variables objeto de estudio, predecir su comportamiento ante variaciones imprevistas de su entorno y resolver problemas bajo una mirada cuantitativa.

Desde una perspectiva genérica, un modelo puede ser interpretado como la representación de un fenómeno de la naturaleza, cuya simbolización puede adquirir diferentes formatos dependiendo del problema que se quiera analizar. Entre las formas más utilizadas se puede mencionar: gráficos, esquemas, fórmulas y ecuaciones [1] [2] [3].

En el caso en el que un modelo pueda ser representado a través de ecuaciones y fórmulas matemáticas se estará haciendo referencia a un *modelo matemático,* donde el pensamiento analítico juega un rol importante ya que permite interpretar una situación particular de la vida cotidiana [2].

El modelado de un problema concreto se desarrolla mediante diferentes etapas. En la primera etapa, se debe observar el problema en forma integral analizando las variables involucradas en el objeto observable, interpretando el grado de asociación que pudieran presentar las variables al igual que el contexto en el cual se manifiestan. Adicionalmente, se puede tener en cuenta la forma de medirlas y de representarlas, ya sea a través de tablas, gráficos o esquemas [1].

Posteriormente, se procede a abordar el análisis de la situación problemática mediante rigurosidad matemática haciendo uso de ecuaciones, definiciones y teoremas con el fin de poder describir de manera cuantitativa el fenómeno considerado como objeto de estudio.

Finalmente, luego de observar el fenómeno, de analizar las variables y de representarlas mediante ecuaciones matemáticas, se procede a contrastar los valores obtenidos con los preexistentes a fin de poder obtener conclusiones [3] [4].

A continuación se muestra un esquema representativo con las diferentes etapas de un modelo matemático.

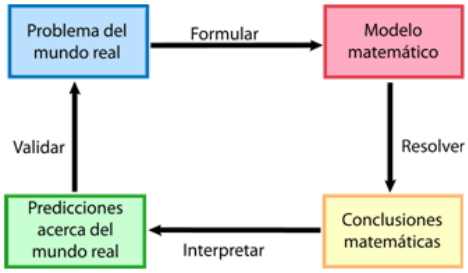


Figura 1.1.1. Etapas que se realizan a la hora de modelar matemáticamente [3].

1.1.1. Tipos de modelos

Además de analizar las diferentes etapas por la que se atraviesa a la hora de modelar un problema, se debe tener en cuenta el tipo de modelización a utilizar dependiendo del problema que se quiere estudiar.

Los modelos matemáticos pueden ser clasificados según criterios muy disímiles entre sí. Según la información en la que se basa el modelo es posible discriminar *entre modelo heurístico,* en el que la información tiene sostén en las definiciones teóricas as de las causas o motivos naturales que generan el fenómeno que se pretende estudiar y *modelo empírico* basado en el estudio de resultados experimentales. Para comprender mejor la diferencia mencionada en estas líneas puede ver [5] [6] [7]. Según el tipo de resultado que se pretende obtener, es posible discriminar entre *modelos cualitativos* y *modelos cuantitativos.* Los modelos cualitativos no buscan un resultado exacto sino una tendencia, como por ejemplo determinar si cierto parámetro del problema se incrementa o disminuye, este tipo de modelos suele valerse de información gráfica. En cambio los modelos cuantitativos buscan como resultado valores numéricos lo más preciso posible. Para comprender mejor esta diferencia ver por ejemplo: [8] [9] [10].

Otro criterio de clasificación que merece ser mencionado divide a los modelos según la aleatoriedad de la situación inicial, los *modelos estocásticos* [11] [12], en los que se tiene como solución la probabilidad de que ocurra determinado suceso o de obtener

determinado valor y modelos deterministas [13] [14], en los cuales se conocen los datos, estando éstos bien determinados. Actualmente existe una rama entera de la matemática dedicada al estudio de procesos estocásticos, es por ese motivo que esta clasificación es muy importante. Por último, la clasificación más importante desde el punto de vista matemático, es según el objetivo que persigue el modelo. Según esta clasificación existen *modelos de simulación* [15] [16] que buscan prever el resultado de un suceso ya sea que éste se pueda medir de manera precisa o aleatoria, *modelos de optimización* [17] [18] que buscan minimizar o maximizar con la finalidad de encontrar la configuración más satisfactoria y modelos de control [19] [20] [21] que buscan determinar los ajustes necesarios para obtener un resultado particular. Estos tipos de modelos son de suma importancia en Matemática Aplicada, hay investigadores que se dedican exclusivamente al estudio del modelado de problemas de control como así también de optimización y/o de simulación.

Dentro de los modelos matemáticos existe en la bibliografía un modelo utilizado en muchos fenómenos naturales [22] [23] el cual es conocido como *modelo de regresión lineal simple*.

1.2. Modelo de regresión lineal simple

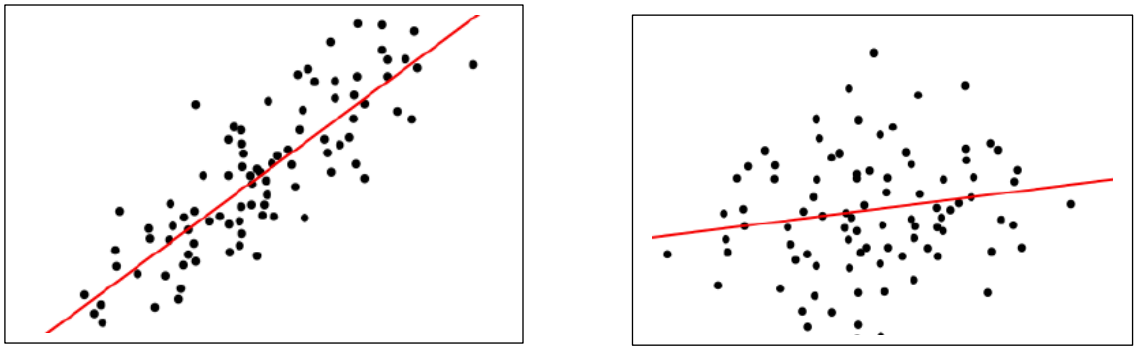
La regresión lineal o el ajuste lineal es un modelo matemático utilizado para aproximar la relación de dependencia entre dos variables, la variable independiente o *variable observable* que es la que surge de la observación de un proceso a estudiar y la variable dependiente o *variable respuesta* que expresa los valores que se infieren a través de la variable observable.

Dado un conjunto de datos, se busca la recta que mejor los modela, denominada recta de regresión lineal. Esta permite describir cuantitativamente el grado de asociación, o la correlación existente, entre las dos variables involucradas en el fenómeno estudiado. Este método es ampliamente utilizado en las diferentes áreas de la ciencia y la ingeniería. A manera de ejemplo, en [24] [25] [26] se pueden ver algunas aplicaciones concretas en biología, economía e ingeniería.

Con la finalidad de visualizar como son los comportamientos de las variables para los diferentes tipos de correlación, se incluyen en la Figura 1.2.1 dos gráficos de dispersión o dispersogramas y la respectiva recta de regresión lineal. En el Gráfico 1.2.1a) se aprecia

que la recta aproxima muy bien a la mayoría de los datos, es decir que la correlación es fuerte; por otro lado en el Gráfico 1.2.1b) existen varios datos que distan mucho de la recta de regresión, en este caso la relación es débil.

a) b)



Figuras 1.2.1. Diagramas de dispersión.[1]

A partir de un conjunto de datos es posible decidir analíticamente el grado de relación existente entre las variables de la muestra. Esta decisión se puede tomar a partir del coeficiente de *correlación lineal de Pearson* (r) que se define según,

$$r = \frac{n.\sum_{i=1}^{n} x_i \,.\, y_i \; - \; \sum_{i=1}^{n} x_i \,.\, \sum_{i=1}^{n} y_i}{\sqrt{n.(\sum_{i=1}^{n} {x_i}^2) - (\sum_{i=1}^{n} x_i)^2}\,\sqrt{n.(\sum_{i=1}^{n} {y_i}^2) - (\sum_{i=1}^{n} y_i)^2}}\,, \qquad (1)$$

donde n representa la cantidad de datos de la muestra y x_i e y_i el i-esimo valor de la variable independiente y dependiente respectivamente [27] [28].

Observación 1.2.1.

El coeficiente r dado por la Ecuación (1) satisface:

$$|r| \leq 1.$$

La información que ofrece el valor de r es de suma utilidad ya que el signo del coeficiente coincide con el signo de la pendiente de la recta de regresión. Por otra parte, su valor define el grado de correlación entre las variables, cuando $|r|$ se acerque a 1 se dice que la correlación es fuerte, caso contrario, es débil [29] [30].

[1] Recuperado de: https://vivaelsoftwarelibre.com/coeficiente-de-correlacion-en-r/

En la bibliografía existe otro valor que permite concluir sobre el desempeño de un modelo. Este valor es el *coeficiente de determinación* (R^2), que en el caso de un modelo lineal resulta simplemente el cuadrado *del coeficiente de correlación lineal de Pearson.* El *coeficiente de determinación* es, en algún sentido, un parámetro más general ya que permite decidir sobre la bondad de un modelo aunque este no sea lineal.

Para visualizar como los coeficientes de correlación de Pearson y de determinación se relacionan con el grado de asociación entre dos variables, se considera un ejemplo concreto.

Ejemplo 1.2.1.

Se toman los datos censales de la ciudad de Buenos Aires en el 2010. Se utiliza el método de Pearson para estudiar el grado de correlación entre las siguientes variables: Población total y total de viviendas por comunas.[2] Los datos utilizados para llevar a cabo el ejemplo se pueden ver en la Tabla 1.2.1., correspondientes a la fracción y radio 1 de cada comuna.

Comuna	**Población total (x_i)**	**Total de viviendas (y_i)**
1	336	82
2	518	124
3	1026	667
4	770	231
5	553	321
6	667	253
7	601	266
8	1147	271
9	530	238
10	641	300

Tabla 1.2.1. Datos censales CABA 2010.

A partir de los datos de la Tabla 1.2.1 se obtienen los diferentes términos del coeficiente de correlación de Pearson. Estos se pueden apreciar en la siguiente tabla:

[2] Recuperado de: https://data.buenosaires.gob.ar/en/dataset/informacion-censal-por-radio

x_i	y_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}).(y_i - \overline{y})$
336	82	-342,9	117580,41	-193,3	37364,89	66282,57
518	124	-160,9	25888,81	-151,3	22891,69	24344,17
1026	667	347,1	120478,41	391,7	153428,89	135959,07
770	231	91,1	8299,21	-44,3	1962,49	-4035,73
553	321	-125,9	15850,81	45,7	2088,49	-5753,63
667	253	-11,9	141,61	-22,3	497,29	265,37
601	266	-77,9	6068,41	-9,3	86,49	724,47
1147	271	468,1	219117,61	-4,3	18,49	-2021,83
530	238	-148,9	22171,21	-37,3	1391,29	5553,97
641	300	-37,9	1436,41	24,7	610,09	-936,13
6789	**2753**	**0**	**537032,9**	**0**	**220340,1**	**220382,3**

Tabla 1.2.2. Términos del coeficiente de correlación de Pearson

Utilizando los datos de la Tabla 1.2.2. y reemplazándolos en la expresión analítica del coeficiente de Pearson dado por la Ecuación (1), se obtiene:

$$\mathrm{r} = \frac{220382,3}{\sqrt{(537032,9\,.\,220340,1)}} = 0,64066278$$

O equivalentemente

$$R^2 = r^2 = \frac{(22039,13)^2}{53703,29\,.\,22034,01} = 0,41048232.$$

Debido a que $r > 0$ la recta de regresión tendrá pendiente positiva. Como r y R^2 son cercanos a 0,5 la correlación es media. En este caso, no existe una buena correlación entre las variables pero tampoco es mala. Este modelo, bajo diferentes circunstancias, podría ser aceptable. Para apreciar este hecho, se incluye en la Figura 1.2.2 la respectiva recta de regresión para el Ejemplo 1.2.1.

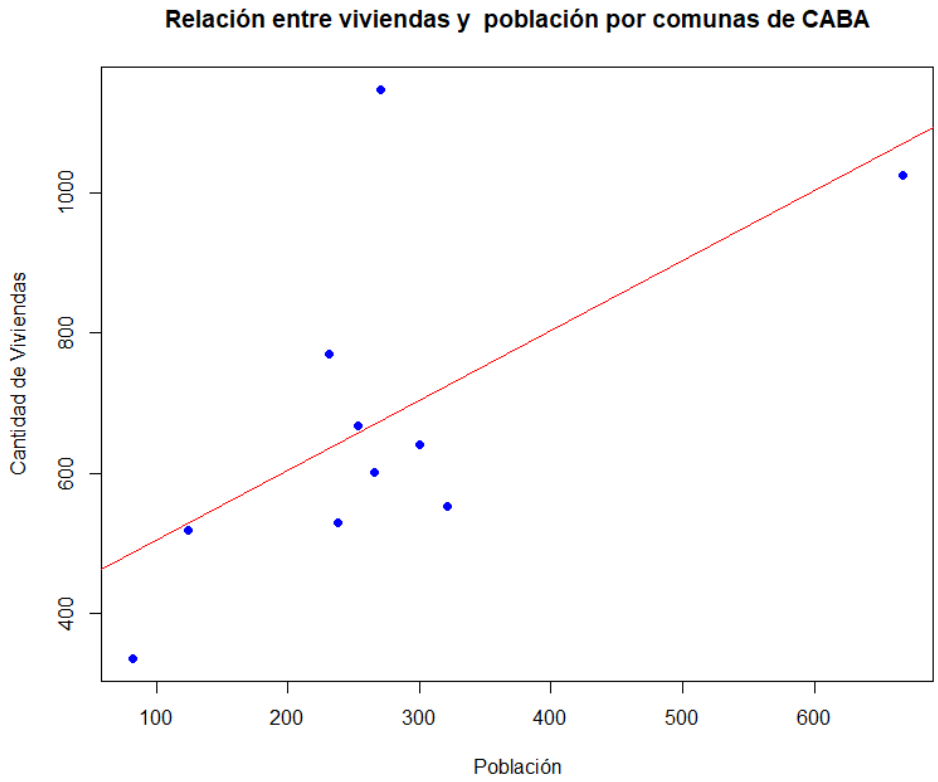


Figura 1.2.2. Recta de regresión lineal en base a datos censales en CABA.

1.2.1. Recta de regresión lineal

Como se mencionó en la Sección 1.2. un modelo de regresión lineal consiste en encontrar la ecuación de la recta que mejor aproxima a los datos de la muestra. Se busca determinar los parámetros reales β_o , β_1 tal que la recta dada por

$$y = \beta_o + \beta_1 . x \ + \ \epsilon, \qquad (2)$$

sea la que mejor aproxima a los datos x_i e y_i.

El parámetro ϵ dado en la Ecuación (2) es una variable aleatoria normalmente distribuida [29] [33] [34] con media cero y varianza σ^2, que indica el error o residuo cometido en la aproximación. Este se determina a través de las diferencias entre los valores observados y los valores predichos. Si $\epsilon \ < \ 0$ la mayoría de los puntos predichos están por debajo de la recta de regresión, caso contrario, si $\epsilon > \ 0$ la mayoría de los puntos están por arriba de dicha recta.

Dado un conjunto de datos observados o simulados (x_i, y_i) con $\ i = 1,2, \dots . , n$ es posible determinar los parámetros β_o y β_1 que minimizan el error cuadrático medio. Este problema de minimización puede abordarse de manera numérica con diferentes estrategias, en esta tesis se utilizará el método de mínimos cuadrados [27][29][37], a partir

del cual se obtiene una expresión analítica para cada uno de los parámetros de interés, estas son:

$$\beta_{\mathrm{o}} = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot (\sum_{i=1}^{n} x_i \cdot y_i)}{n. \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \tag{3}$$

$$\beta_1 = \frac{(\sum_{i=1}^{n} x_i \cdot y_i) - \sum_{i=1}^{n} x_i \cdot (\sum_{i=1}^{n} x_i \cdot y_i)}{n. \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}. \tag{4}$$

Con la finalidad de visualizar como se determina la recta de regresión lineal, se considera el siguiente ejemplo concreto.

Ejemplo 1.2.2.

Se aborda nuevamente los datos del Ejemplo 1.2.1. Se utilizan las Ecuaciones (3)-(4) para calcular los parámetros β_{o} y β_1 en este caso particular. Operando se obtiene:

$$\beta_{\mathrm{o}} = \frac{2167121083}{5370329} = 403{,}536, \qquad \beta_1 = \frac{5381069{,}658}{5370329} = 1{,}002\,.$$

Debido a esto, la ecuación de regresión lineal correspondiente al Ejemplo 1.2.1 que relaciona la cantidad de viviendas en relación con la población, viene dada por:

$$y = 1{,}002.\, x + 403{,}5360.$$

1.2.2. Caracterización de los residuos

En todo modelo predictivo es de suma importancia conocer el tipo de errores existentes entre el valor observado y el predicho. En la Sección 1.2.1. se afirmó que el error o el residuo ϵ es una variable aleatoria distribuida normalmente con media cero y varianza σ^2. Esto significa que la varianza de cada error se supone constante; es decir, que los residuos se distribuyen de manera regular y presentan un bajo grado de dispersión con respecto a la media. Los modelos cuyos residuos tienen varianza constante se conocen en la bibliografía como homocedásticos [17] [25]. Si se considera, nuevamente, la Figura 1.2.1 se puede observar que el caso del grafico a) los residuos tienen un comportamiento homocedástico; situación que no ocurre para el gráfico b) cuyo comportamiento es heterocedástico.

En este contexto y con la finalidad de caracterizar los residuos, resulta de utilidad poder decidir sobre el grado de normalidad que estos tienen. Para ello, existen en la literatura, muchas herramientas estadísticas. En esta tesis se utilizará el test de Shapiro-Wilk [27], cuyo coeficiente se representa a través de la siguiente expresión:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \, x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{5}$$

donde $x_{(i)}$ indica las posiciones de los datos de una muestra ordenados en forma creciente, $\bar{x}$ la media de los datos, y a_i son los coeficientes tabulados de Shapiro-Wilk, los cuales dependen del nivel de significación.

Observación 1.2.2.

El coeficiente W dado por la Ecuación (5) satisface:

$$W \leq 1.$$

La información que ofrece el valor de W permite decidir sobre "que tan" normalmente distribuidos están los elementos de una muestra. Cuando más cercano a 1 sea el valor de W mayor será el grado de normalidad [38] [39].

Para mostrar esta idea, se consideran los datos del Ejemplo 1.2.1. y se aplica el test de Shapiro-Wilk. Para la variable total de viviendas; el valor del coeficiente es $W = 0{,}79979$ y para la variable población total, resulta que, $W = 0{,}91097$. En ambos casos las variables presentan un grado de normalidad considerable siendo este, mayor para la variable total de viviendas.

Capítulo 2

REGRESIÓN LINEAL MÚLTIPLE

En el capítulo anterior se abordaron los conceptos preliminares referentes al modelo de regresión lineal simple, se presentaron herramientas, tanto gráficas como analíticas y se caracterizaron los errores que se comete al aproximar, a un conjunto de datos, con un modelo lineal. En este capítulo se generalizarán esos conceptos con la finalidad de relacionar más de una variable observable a una única variable respuesta.

2.1. Modelo de regresión lineal múltiple

El modelo que aproxima linealmente a un conjunto de datos con n variables independientes $(x_1, x_2, \dots x_n)$, se conoce en la bibliografía con el nombre de modelo de regresión lineal múltiple o análisis univariado [27] [28] [41] [42], donde la ecuación de aproximación es la de un hiperplano. Es decir,

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon\,, \tag{6}$$

donde $\beta_o, \beta_1, \beta_2, \dots, \beta_n$ son los coeficientes de la ecuación del hiperplano, Y representa la variable predictiva y ε indica el error o residuo cometido en la aproximación. Un caso particular de sumo interés es cuando se relacionan dos variables observables a una única variable predictiva; aquí el hiperplano se reduce a un plano de predicción. A manera de ejemplo, consideremos como varía el volumen de ventas de una determinada casa de electrodomésticos en función de la cantidad de radios y televisores vendidos. Ver la Figura 2.1.1.

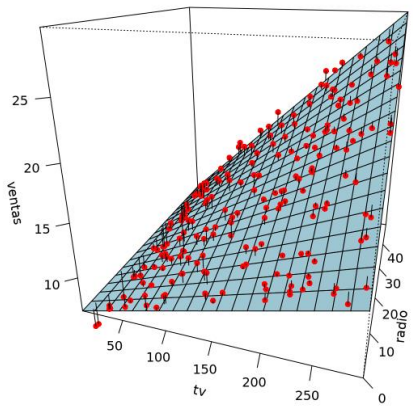


Figura 2.1.1. Plano de regresión lineal múltiple. [3]

[3] www.rpubs.com/Joaquin_AR/226291

Para conocer el grado de asociación entre un conjunto de *n* variables de una muestra, se puede calcular el *coeficiente de correlación múltiple* [27] [28], el cual representa una extensión del coeficiente de correlación lineal de Pearson que se presentó en el Capítulo 1, cuya fórmula se muestra a continuación:

$$r_{ij} = \frac{n.\sum_{i,j=1}^{n} x_i \,.\, x_j - \sum_{i=1}^{n} x_i \,.\, \sum_{j=1}^{n} x_j}{\sqrt{n.(\sum_{i=1}^{n} {x_i}^2) - (\sum_{i=1}^{n} x_i)^2}\sqrt{n.(\sum_{j=1}^{n} {x_j}^2) - (\sum_{j=1}^{n} x_j)^2}}\ , \tag{7}$$

donde *n* representa la cantidad de datos de la muestra y x_i e x_j el par de variables de las cuales se pretende analizar si existe algún grado de correlación.

Observación 2.1.1.

Al igual que el coeficiente de correlación de Pearson, el coeficiente r_{ij}, cuya expresión está dada por la ecuación (7), cumple:

$$|r_{ij}| \leq 1. \tag{8}$$

El valor r_{ij} indica el grado de asociación entre las variables ij; cuando el valor de r_{ij} se aproxima a 0 la correlación es débil, mientras que cuanto más se acerque a 1 el grado de asociación será más fuerte. Estas propiedades se verifican también al evaluar r_{ji}, dado que se cumple la propiedad conmutativa con respecto al coeficiente de Pearson. Además se puede verificar que el coeficiente de correlación calculado sobre la misma variable es igual, es decir, r_{ii} =1 para todo i=1,2,…,n [29] [30].

A manera de ejemplo se estudia un caso particular para ver cómo se relacionan entre sí, tres variables presentes en el modelo.

Ejemplo 2.1.1.

Se considera, nuevamente, el Ejemplo 1.2.1 abordado en el Capítulo 1, donde se incluye una nueva variable al modelo, llamada viviendas particulares. Se pretende calcular y analizar el nivel de asociación entre las siguientes variables: población total, total de viviendas y cantidad de viviendas particulares en distintas comunas de la ciudad de Buenos Aires, como se muestra a continuación:

Comuna	Población total (x_1)	Total de viviendas (x_2)	Viviendas particulares (x_3)
1	336	82	80
2	518	124	124
3	1026	667	667
4	770	231	231
5	553	321	319
6	667	253	253
7	601	266	266
8	1147	271	271
9	530	238	238
10	641	300	300

Tabla 2.1.1. Datos censales. CABA 2010.[4]

A partir de los datos de la Tabla 2.1.1 se obtienen los valores del coeficiente de correlación lineal múltiple entre las variables analizadas lo cual permite cuantificar de qué manera se relacionan las variables. Los valores entre las variables: población total y total de viviendas ya fueron calculados en el Ejemplo 1.2.1 del Capítulo 1, cuyo resultado fue:

$$r_{12} = 0{,}6407.$$

Luego, considerando las variables x_1: población total y x_3: cantidad de viviendas particulares, se obtiene la siguiente tabla:

x_1	x_3	$x_1 - \overline{x_1}$	$(x_1 - \overline{x_1})^2$	$x_3 - \overline{x_3}$	$(x_3 - \overline{x_3})^2$	$(x_1 - \overline{x_1}).(x_3 - \overline{x_3})$
336	80	-342,9	117580,41	-194,9	37986,01	66831,21
518	124	-160,9	25888,81	-150,9	22770,81	24279,81
1026	667	347,1	120478,41	392,1	153742,41	136097,91
770	231	91,1	8299,21	-43,9	1927,21	-3999,29
553	319	-125,9	15850,81	44,1	1944,81	-5552,19
667	253	-11,9	141,61	-21,9	479,61	260,61
601	266	-77,9	6068,41	-8,9	79,21	693,31
1147	271	468,1	219117,61	-3,9	15,21	-1825,59
530	238	-148,9	22171,21	-36,9	1361,61	5494,41

[4] Recuperado de: https://data.buenosaires.gob.ar/en/dataset/informacion-censal-por-radio

641	300	-37,9	1436,41	25,1	630,01	-951,29
6789	**2749**	**0**	**537032,9**	**0**	**220936,9**	**221328,9**

Tabla 2.1.2. Términos del coeficiente de correlación entre las variables x_1 y x_3 .

Se aplica la fórmula de correlación lineal múltiple a los datos de la Tabla 2.1.2. y se obtiene:

$$r_{13} = \frac{221328{,}9}{\sqrt{(537032{,}9\,.\,220936{,}9)}} = 0{,}6425\,.$$

Finalmente, siendo las variables x_2: total de viviendas y x_3: cantidad de viviendas particulares, resulta:

$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_2 - \overline{x_2}}$	$\boldsymbol{(x_2 - \overline{x_2})^2}$	$\boldsymbol{x_3 - \overline{x_3}}$	$\boldsymbol{(x_3 - \overline{x_3})^2}$	$\boldsymbol{(x_2 - \overline{x_2}).(x_3 - \overline{x_3})}$
82	80	-193,3	37364,89	-194,9	37986,01	66282,57
124	124	-151,3	22891,69	-150,9	22770,81	24344,17
667	667	391,7	153428,89	392,1	153742,41	135959,07
231	231	-44,3	1962,49	-43,9	1927,21	-4035,73
321	319	45,7	2088,49	44,1	1944,81	-5753,63
253	253	-22,3	497,29	-21,9	479,61	265,37
266	266	-9,3	86,49	-8,9	79,21	724,47
271	271	-4,3	18,49	-3,9	15,21	-2021,83
238	238	-37,3	1391,29	-36,9	1361,61	5553,97
300	300	24,7	610,09	25,1	630,01	-936,13
2753	**2749**	**0**	**220340,1**	**0**	**220936,9**	**220382,3**

Tabla 2.1.3. Términos del coeficiente de correlación entre las variables x_2 y x_3 .

Se sustituye en la Euación (7), de acuerdo a la Tabla 2.1.3., el coeficiente de correlación entre las variables es:

$$r_{23} = \frac{220382{,}3}{\sqrt{(220936{,}9\,.\,220340{,}1)}} = 0{,}9988.$$

Como se mencionó en la observación 2.1.1. se comprueba la propiedad conmutativa con respecto a los coeficientes de correlación, dado que $r_{12} = r_{21}$, $r_{13} = r_{31}$ y $r_{23} = r_{32}$. Además se puede notar que: $r_{11} = r_{22} = r_{33} = 1$.

De forma reducida se representan los valores obtenidos mediante una matriz de correlación, como se muestra a continuación:

	Población total	Total de viviendas	Viviendas particulares
Población total	1,0000	0,6407	0,6425
Total de viviendas	0,6407	1,0000	0,9988
Viviendas particulares	0,6425	0,9988	1,0000

Tabla 2.1.4. Matriz de correlación.

En función a los resultados obtenidos en la Tabla 2.1.4. se observa que el valor de los coeficientes es positivo; esto implica que el tipo de relación es directa. Además, se puede notar que el grado de asociación entre las variables total de viviendas y viviendas particulares es cercano a uno, a diferencia del coeficiente de correlación entre las variables población total y total de viviendas, cuyo valor es medio.

Con la finalidad de representar el grado de asociación entre las variables de manera gráfica, se utilizan diferentes dispersogramas ubicados en forma de matriz, cuyos puntos representan las coordenadas de los datos relacionados.

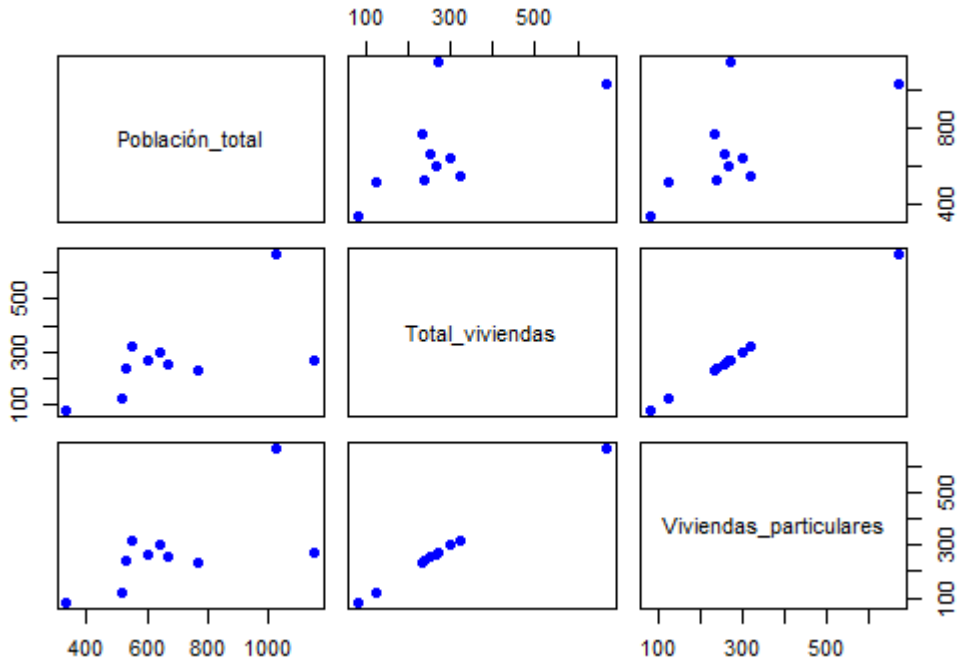


Figura 2.1.2.Correlograma múltiple.

En la Figura 2.1.2. se puede apreciar que la nube de puntos muestra una tendencia ascendente lo cual indica que las variables se relacionan de manera directa. En los gráficos correspondientes a viviendas particulares y total de viviendas los puntos se encuentran alineados en forma creciente, mientras que en los demás correlogramas la nube de puntos están más dispersos, cuyo hecho es atribuible a un menor grado de asociación entre las variables.

A continuación se muestran las ecuaciones que permiten determinar los coeficientes de regresión lineal múltiple.

2.2. Ecuación de regresión lineal múltiple

En el caso particular que el modelo de regresión lineal considere dos variables independientes y una variable dependiente, la ecuación de regresión estará representada por un plano como se mostró en la Figura 2.1.

Se generaliza la interpretación del modelo para *n* variables independientes, donde gráficamente se obtiene un *hiperplano* en un espacio de $n + 1$ dimensiones. Un modelo lineal de *n* variables permite estimar el comportamiento de la variable dependiente a partir de la observación de n variables independientes representadas por $x_1, x_2, \dots, x_{\mathrm{n}}$. Analíticamente, la ecuación de regresión lineal múltiple queda expresada por la ecuación (6).

Al igual que en el modelo de regresión lineal simple, se utiliza el método de mínimos cuadrados con el fin de obtener los valores de los parámetros representativos por la ecuación (6). Al operar algebraicamente se obtienen las *ecuaciones normales* $(*)$ correspondientes al *hiperplano de regresión* de mínimos cuadrados [27] [28] [29] [40]:

$$\sum_{i:1}^{n} y_i = n.\beta_{\mathrm{o}} + \beta_1.\sum_{i:1}^{n} x_1 + \beta_2.\sum_{i:1}^{n} x_2 + \cdots + \beta_n.\sum_{i:1}^{n} x_{\mathrm{n}} ,$$

$$\sum_{i:1}^{n} x_1.y_i = \beta_{\mathrm{o}}.\sum_{i:1}^{n} x_1 + \beta_1.\sum_{i:1}^{n} {x_1}^2 + \beta_2.\sum_{i:1}^{n}(x_1.x_2) + \cdots + \beta_n.\sum_{i:1}^{n}(x_1.x_n) \ (*) ,$$

$$\sum_{i:1}^{n} x_2.y_i = \beta_{\mathrm{o}}.\sum_{i:1}^{n} x_2 + \beta_1.\sum_{i:1}^{n}(x_1.x_2) + \beta_2.\sum_{i:1}^{n} {x_2}^2 + \cdots + \beta_n.\sum_{i:1}^{n}(x_2.x_n) \ (*) ,$$

……..

$$\sum_{i=1}^{n} x_{\mathrm{n}} \,.y_i = \beta_{\mathrm{o}}.\sum_{i=1}^{n} x_{\mathrm{n}} + \beta_1.\sum_{i=1}^{n}(x_{\mathrm{n}-1}.x_{\mathrm{n}}) + \beta_2.\sum_{i=1}^{n}(x_{n-2}.x_{\mathrm{n}}) \dots + \\ + \beta_n.\sum_{i=1}^{n} {x_{\mathrm{n}}}^2 \ (*) \ .$$

Donde $x_1, x_2, \ldots, x_n$ representan las variables independientes, "y" la variable dependiente, β_o , β_1 , $\beta_2 \ldots$, β_n los parámetros de la ecuación y ε el error cometido en la aproximación.

Ejemplo 2.2.1.

Se considera nuevamente los datos del Ejemplo 2.1.1. donde se calcula la ecuación de regresión lineal múltiple para estimar el valor de la cantidad de viviendas particulares en función de la población total y del total de viviendas.

Para ello, se plantean los términos que permiten hallar las ecuaciones normales con el fin de calcular los valores de los parámetros β_o, β_1 y β_2, como se muestra en la siguiente tabla:

x_1	x_2	x_3	$x_1{}^2$	$x_2{}^2$	$x_3{}^2$	$x_1.x_2$	$x_1.x_3$	$x_2.x_3$
336	82	80	112896	6724	6400	27552	26880	6560
518	124	124	268324	15376	15376	64232	64232	15376
1026	667	667	1052676	444889	444889	684342	684342	444889
770	231	231	592900	53361	53361	177870	177870	53361
553	321	319	305809	103041	101761	177513	176407	102399
667	253	253	444889	64009	64009	168751	168751	64009
601	266	266	361201	70756	70756	159866	159866	70756
1147	271	271	1315609	73441	73441	310837	310837	73441
530	238	238	280900	56644	56644	126140	126140	56644
641	300	300	410881	90000	90000	192300	192300	90000
6789	**2753**	**2749**	**5146085**	**978241**	**976637**	**2089403**	**2087625**	**977435**

Tabla 2.1.4. Términos para hallar las ecuaciones normales.

Se aplican los datos de la Tabla 2.1.4. a las ecuaciones normales y se obtiene:

$$6789 = 10.\beta_o + \beta_1.2753 + \beta_2.2749\,,$$

$$2089403 = \beta_o.2753 + \beta_1.978241 + \beta_2.977435,$$

$$2087625 = \beta_o.2749 + \beta_1.977435 + \beta_2.976637.$$

El sistema resultante se puede resolver utilizando cualquier método analítico o numérico. De esto resulta:

$$\beta_o = 485{,}7818\ ; \qquad \beta_1 = -106{,}1180\ ; \qquad \beta_2 = 106{,}9749.$$

En función de los valores hallados, la ecuación de regresión lineal múltiple que muestra la relación entre viviendas particulares con la cantidad de viviendas y población total, se expresa a través de la siguiente ecuación:

$$y = 485{,}7818 - 106{,}1180.x_1 + 106{,}9749.x_2 + \varepsilon,$$

donde "y" es la cantidad de viviendas particulares, x_1 la población total, x_2 el total de viviendas y ε el error cometido en la estimación.

Gráficamente se puede representar un plano de regresión de mínimos cuadrados para predecir como varía la cantidad de viviendas particulares en relación a la población y cantidad de hogares en la ciudad de Buenos Aires, como se muestra a continuación:

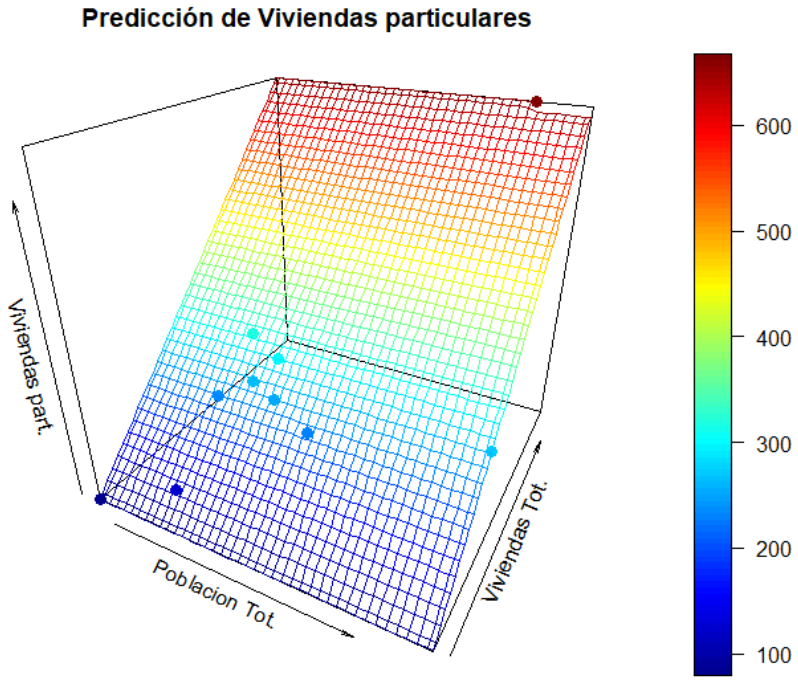


Figura 2.1.3.Plano de regresión sobre datos censales 2010.

En la Figura 2.1.3. se puede observar que el plano de regresión se ubica de forma ascendente. Esto indica una correlación directa entre las variables población total y viviendas totales con respecto a la variable viviendas particulares. Por otra parte, los puntos se ubican muy próximos al plano de regresión lo cual explica que el error cometido en la estimación sea pequeño.

En la siguiente sección se describen las principales características del error presentes en este tipo de modelo.

2.3. Caracterización de los residuos

Al desarrollar un análisis aplicando un modelo de regresión lineal múltiple es posible cometer errores tanto en la selección de la muestra como en la selección de las variables. A fin de poder minimizar el residuo, se deben tener presente el efecto de: la heterocedasticidad, especificidad y multicolinealidad [29] [35] [42].

Al igual que en el modelo de regresión lineal simple las situaciones que favorecen un comportamiento *heterocedástico* de los residuos, donde la varianza no es constante y por lo tanto la distribución de los residuos es irregular presentan un alto grado de dispersión de los datos al implementar el modelo. Esto podría surgir, por ejemplo, al haber eliminado alguna variable que era verdaderamente significativa para el modelo de regresión, cuyo concepto es conocido como *especificidad.*

Otro aspecto que afecta la variación de los residuos es la *multicolinealidad*, la cual se verifica cuando el grado de asociación entre las variables independientes seleccionadas para el modelo de regresión es alta. En el caso que dos variables estén correlacionadas se dice que una variable absorbe correlación del resto de las variables afectando el resultado esperado e incrementando el error cometido en la medición. Bajo esta circunstancia, podría ocurrir que una variable sea interpretada como relevante para el análisis mientras que en realidad es poco significativa.

En caso que no se hayan incluido variables importantes para el modelo o se haya omitido el efecto de la mutlicolinealidad, se deben redefinir las variables y volver a aplicar el modelo evaluando, en qué medida se ha logrado disminuir el residuo.

Con la finalidad de ampliar sobre el grado de asociación entre un conjunto de variables observables y más de una variable respuesta, se abordará en el próximo capítulo un concepto llamado *análisis multivariado*.

Capítulo 3

ANÁLISIS MULTIVARIADO

A diferencia del análisis univariado tratado en el capítulo 2, el método multivariado surge de la necesidad de explicar y predecir cómo se comportan un conjunto de varias variables respuestas, en relación a una cantidad determinada de variables observables.

3.1. Introducción al análisis multivariado

Entre los diferentes métodos de análisis multivariado se encuentran los análisis descriptivos y predictivos. Mientras que los métodos descriptivos se utilizan para identificar el nivel de asociación entre variables independientes, los métodos predictivos se emplean para inferir cual será el comportamiento de una o de un grupo de variables a partir de la observación de un conjunto de parámetros. Entre los métodos descriptivos más utilizados se encuentra, por ejemplo: *análisis de componentes principales, análisis factorial* y *análisis de correspondencia simple o múltiple* [47] [48] [49]. Por otra parte, los métodos predictivos más aplicados son: *análisis de correlación canónico, regresión múltiple y análisis multivariante de la varianza* [48] [49].

El *análisis de componentes principales* se utiliza para reducir la dimensión de un conjunto de variables no correlacionadas en un número pequeño de variables llamadas componentes principales. Estas permiten describir el comportamiento de casi toda la información presente en las variables originales. En segundo lugar, el *análisis factorial* busca estudiar el grado de asociación entre las variables a partir de un número reducido de variables que no están correlacionadas y que son llamadas factores. Por su parte, en el *análisis de correspondencia* intervienen dos o más variables categóricas con la finalidad de identificar algún grado de relación entre las variables, las cuales pueden ser representadas a través de una tabla de contingencia [63] [68].

El *análisis de correlación canónico* consiste en relacionar un conjunto de variables respuestas con otro grupo de variables independientes para luego analizar el grado de asociación entre ambos grupos. Desde un punto de vista teórico es considerado como una ampliación del modelo de *regresión lineal múltiple*, el cual estudia la relación entre una variable respuesta con respecto a un conjunto de variables observables. Por último, el *análisis multivariante de la varianza* es un método que permite determinar el nivel de correlación entre n parámetros y más de una variable respuesta.

3.2. Herramientas de análisis multivariado

La *matriz de correlación* [44] [45] permite analizar el grado de asociación entre las variables de un conjunto de datos. Si bien existen diferentes métodos para hallar un nivel de correlación entre las variables, como el método de Kendall o de Spearman, en este caso particular, se utilizará el método de Pearson [43] debido a la mayor precisión en sus resultados. A continuación, se abordará un ejemplo con el fin de poder analizar el grado de asociación entre las siguientes variables: año, Peso Neto (en Kilos) y Valor CIF (en dólares).

Ejemplo 3.2.1.

En base a la información recolectada de la página oficial del gobierno de la provincia de Santa Fe[5]*, se presentan los datos referentes al nivel de importaciones en la provincia entre los años 2016 y 2018 en relación a las variables: precio de importación, Peso Neto importado, siendo el valor CIF el costo de la mercadería vendida más el costo del flete y del seguro internacional.*

Los datos suministrados constan de 2.713 observaciones, cuyas variables son:

	año	aduana	pais	cod_sección	seccion	capitulo	Valor CIF (en Dólares)	Peso Neto (en Kilos)
1	2018	Rosario	Brasil	I	Animales Vivos Y Productos Del Reino Animal	Carne y despojos comestibles	293473928	1136096851
2	2018	Rosario	Brasil	II	Productos Del Reino Vegetal	Hortalizas, plantas, raíces y tubérculos alimenticios	8580	546
3	2018	Rosario	Brasil	II	Productos Del Reino Vegetal	Frutas y frutos comestibles; cortezas de agrios, melones o s...	9774392	11340
4	2018	Rosario	Brasil	II	Productos Del Reino Vegetal	Café, te, yerba y especias	2318503402	101202382
5	2018	Rosario	Brasil	II	Productos Del Reino Vegetal	Semillas y frutos oleaginosos, semillas y frutos diversos; pla...	1528837	1
6	2018	Rosario	Brasil	III	Grasas Y Aceites	Grasas y aceites animales o vegetales; productos de su desd...	166542784	3866426
7	2018	Rosario	Brasil	IV	Productos Alimenticios, Bebidas Y Tabaco	Preparación de carne, pescado o de crustáceos, moluscos o ...	330962	1524
8	2018	Rosario	Brasil	IV	Productos Alimenticios, Bebidas Y Tabaco	Azúcares y artículos de confitería	35158032	231556
9	2018	Rosario	Brasil	IV	Productos Alimenticios, Bebidas Y Tabaco	Cacao y sus preparaciones	24504847	65820
10	2018	Rosario	Brasil	IV	Productos Alimenticios, Bebidas Y Tabaco	Preparaciones a base de cereales, harinas, almidón, fécula o ...	34627446	1678375

Showing 1 to 10 of 2,713 entries, 8 total columns

Tabla 3.2.1. Datos sobre importaciones en la Provincia de Santa Fe.

Con el fin de mostrar el grado de asociación entre las variables: año, Valor CIF y Peso Neto, se construye una matriz de correlación en la cual se introducen dentro de una tabla de doble entrada los coeficientes de correlación de Pearson de todas las variables numéricas, como se observa en la Tabla 3.2.2:

[5] Instituto Provincial de Estadística y Censos (IPEC).
datos.santafe.gob.ar/dataset/importaciones-por-aduanas-de-santa-fe.

	Año	Valor CIF (u$s.)	Peso Neto (kg.)
Año	1,00	0,18	0,09
Valor CIF (u$s.)	0,18	1,00	0,36
Peso Neto (kg.)	0,09	0,36	1,00

Tabla 3.2.2. Matriz de correlación, método de Pearson.

Entre los resultados obtenidos prevalece un grado de asociación positivo entre las variables. El mayor grado de asociación está presente entre las variables: *"Valor CIF"* y *"Peso Neto"*, no obstante, su grado de asociación es medio. Por otra parte, las variables *"Peso Neto"* y *"Año"* presentan el menor grado de correlación.

A fin de poder visualizar el grado de asociación entre las variables mencionadas, se utiliza un *correlograma* donde el color de la esfera manifiesta el tipo de correlación [31].

La *correlación es positiva* si el color de la esfera es azul y si el color es rojo, la *correlación es negativa*. Por otra parte, el tamaño de las esferas y la intensidad del color es proporcional al grado de asociación, es decir, esferas grandes de color intenso indican *correlación fuerte* y esferas pequeñas de color claro expresan *correlación débil*.

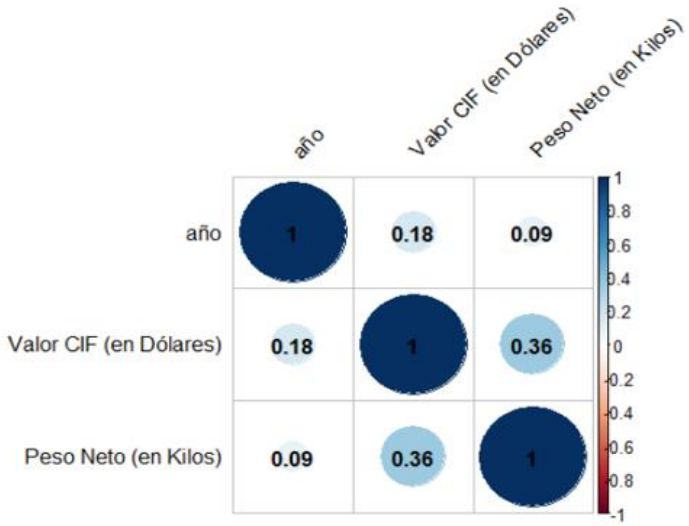


Figura 3.2.1. Correlograma datos de importación.

En el gráfico representado en la Figura 3.2.1. se observa en general un grado de asociación positivo entre las variables, debido a que el color que prevalece en las esferas es azul, siendo el mayor valor de correlación el representado en la diagonal principal cuyo valor es igual a uno. Por otro lado, el tamaño de las esferas entre las variables: *Valor CIF* y *Peso Neto* es medio, lo cual muestra un grado de asociación menor entre estas magnitudes, mientras que las regiones de menor intensidad muestran menor grado de correlación entre la variable *año* con respecto a las variables *Peso neto* y *Valor CIF*.

Además del correlograma, otra representación gráfica muy utilizada para analizar datos multivariados son los *gráficos de panel* [46]. Este tipo de gráfico permite visualizar en un mismo esquema, la matriz de correlación, los gráficos de dispersión y los histogramas de cada variable numérica analizada, como se presenta a continuación:

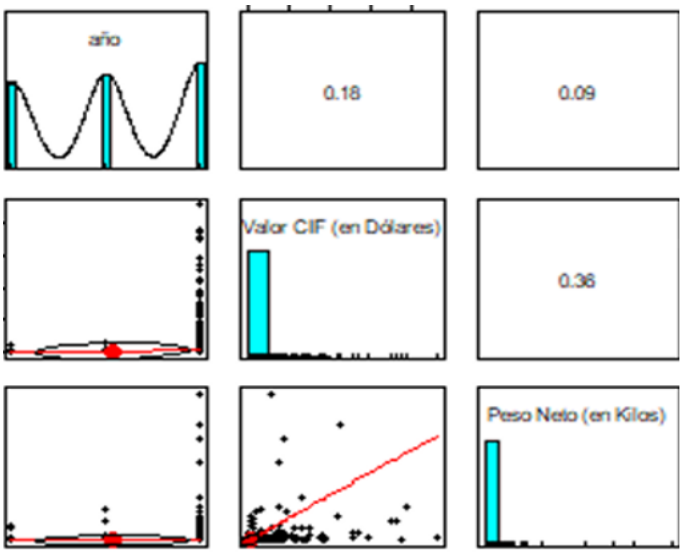


Figura 3.2.3. Gráfico de panel datos de importación.

En el histograma representado en la Figura. 3.2.3 se puede notar que las operaciones comerciales se han incrementado levemente al pasar los años. En cuanto al *"Valor CIF"* se puede constatar que los valores se concentran en un monto significativo y luego disminuyen abruptamente al igual que sucede con el "Peso Neto" de las mercaderías importadas.

Con respecto a los gráficos de dispersión se observa, en general, una correlación positiva entre las variables analizadas, cuyos valores se representan a la derecha de la diagonal principal. El mayor valor que alcanza el coeficiente de correlación es de 0,36, correspondientes al "*Peso Neto*" y al *"Valor CIF",* cuyo grado de asociación es bajo. Por otra parte, las relaciones entre las magnitudes: "*años*" - "*Peso Neto*" y "*año*" -"*Valor CIF*" representan los niveles más bajos de correlación cuyos coeficientes son: 0,009 y 0,18 respectivamente.

3.3. Análisis multivariado de la varianza

Si se pretende analizar el grado de asociación entre un conjunto de variables independientes, también llamadas factores y más de una variable dependiente cuantitativa, se puede realizar un análisis multivariado de la varianza, concepto conocido como método MANOVA [50] [51] [52].

Este modelo permite bridar un análisis global sobre la naturaleza de las magnitudes intervinientes, teniendo en cuenta la interrelación entre las diferentes variables respuestas. Es por ello que, dependiendo de la cantidad de variables dependientes y_j que se desean analizar en relación a las variables observables denotadas por x_i presentes en el modelo, el cual se expresa como [53]:

$$y_1, y_2, \dots, y_j \sim x_1, x_2, \dots, x_i .$$

Un método utilizado para realizar un análisis MANOVA se denomina *estadístico o traza de Pillai-Bartlett* denotado por v [55] [56], el cual es un indicador que se obtiene al sumar las varianzas de las variables independientes con el fin de determinar si su valor presenta algún grado de correlación con respecto a las variables respuestas. Dado que el cálculo de la distribución del test de Pillai es complejo, se suele utilizar un software para obtener su resultado, por ejemplo, R Studio.

Por otra parte, la traza de Pillai-Bartlett cumple:

$$0 \leq v \leq 1. \qquad (9)$$

El valor que toma v oscila entre 0 y 1. Los valores próximos a uno indican que las variables independientes tiene un efecto significativo con respecto a las variables respuestas. Por otro lado, si el valor de la traza de Pillai es próximo a cero indica baja relación entre las variables independientes y dependientes [57] [58] [59]. Una desventaja que presenta el modelo consiste en que es eficiente para muestras de más de 30 elementos [54].

Con el fin de mostrar un caso de aplicación del modelo MANOVA se retoma el Ejemplo 3.2.1. abordado en la Sección 3.2. sobre las importaciones de mercaderías en la provincia de Santa Fe.

Ejemplo 3.3.1.

Se quiere determinar cuál es el grado de asociación entre la variable observable "año" con respecto a las variables respuestas: "Peso Neto", "Valor CIF" de importación.

Siendo y_1, y_2 las variables dependientes correspondientes a los parámetros: "*Peso Neto*", "*Valor CIF*" y con respecto a x_1 la variable explicativa "*año*", cuyo modelo multivariado se muestra a través de la siguiente expresión:

$$y_1, y_2 \sim x_1.$$

Con el propósito de evaluar el grado de asociación entre las variables mencionadas; se calcula, a continuación, el valor de la traza de Pillai-Bartlett utilizando el software R Studio.

El resultado obtenido es el siguiente:

$$v = 0{,}031917.$$

Como el valor del estadístico de Pillai-Bartlett es próximo a cero, se concluye que el grado de asociación entre los parámetros: año, valor CIF y el peso neto de importación es muy bajo. Dicho resultado coincide con las conclusiones obtenidas en el Ejemplo 3.2.1 del Capítulo 3.

En las próximas secciones se abordarán dos métodos utilizados en el análisis multivariado para reducir la dimensionalidad de las variables observables. Estos son: *análisis de componentes principales y análisis factorial*.

3.4. Análisis de componentes principales

Una de las técnicas más utilizadas para analizar la variabilidad de un conjunto de variables es conocida como *análisis de componentes principales (PCA)* [62] [63].

El método consiste en reducir la dimensión de un conjunto de variables observables, las cuales no están correlacionadas entre sí, en un grupo con menor cantidad de variables, llamadas *componentes principales* de modo que su análisis permita explicar el comportamiento de casi toda la información presente en las variables originales. Para poder interpretar este concepto gráficamente, se representa en la siguiente imagen una taza en tres dimensiones, cuyas proyecciones se reducen a dos dimensiones.

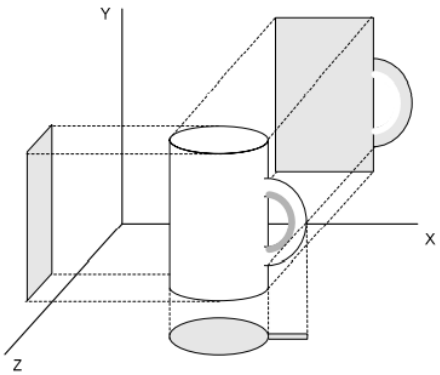


Figura 3.4.1.Proyección y reducción de las dimensiones de una tasa [61].

Al reducir la dimensionalidad de la taza en tres planos de dos dimensiones se pierde valor de los datos con respecto al ancho y largo de la figura original, por lo tanto, se deben seleccionar las proyecciones de los planos que mejor se aproximen a las dimensiones del objeto real.

De manera análoga trabajando con *k* parámetros observables, al aplicar el método de reducción de variables se obtienen nuevas magnitudes llamadas componentes principales, las cuales permiten describir el comportamiento de los valores originales presentes en una muestra de datos, cuyo objetivo es poder captar la mayor cantidad de variabilidad total de los datos reduciendo la perdida de información presente en los valores originales.

En primer lugar, se parte de un conjunto de *n* datos para obtener una *matriz de varianza y covarianza.* Esta matriz, contienen en la diagonal la varianza de cada variable independiente y en las demás entradas la covarianza de cada par de elementos [61] [67]. Es decir,

$$\sum = \begin{pmatrix} \boldsymbol{\sigma_{11}}^{\mathbf{2}} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \boldsymbol{\sigma_{22}}^{\mathbf{2}} & \cdots & \sigma_{2k} \\ \vdots & \ddots & & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \boldsymbol{\sigma_{kk}}^{\mathbf{2}} \end{pmatrix} .$$

Donde $\sum$ indica la matriz de varianza y covarianza, ${\sigma_{11}}^2, {\sigma_{22}}^2, \ldots, {\sigma_{kk}}^2$ son los elementos de la diagonal principal, los cuales muestran la inercia proyectada de cada variable, cuya suma representa la varianza total de la muestra, también llamada *traza de la matriz de covarianza o inercial total* $Tr(\sum)$ [64] [74], la cual se expresa por:

$$Tr(\sum) = {\sigma_{11}}^2 + {\sigma_{22}}^2 + \cdots {\sigma_{kk}}^2.$$

Luego, con el fin de eliminar la distorsión entre las variables causadas por el uso de diferentes escalas de medidas, se estandarizan las variables obteniendo así la *matriz de correlación R* [61] [68] [70] expresada por:

$$R = \begin{pmatrix} \boldsymbol{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\sigma_{22}} & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\sigma_{kk}} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \boldsymbol{\sigma_{11}}^{\mathbf{2}} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \boldsymbol{\sigma_{22}}^{\mathbf{2}} & \cdots & \sigma_{2k} \\ \vdots & \ddots & & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \boldsymbol{\sigma_{kk}}^{\mathbf{2}} \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\sigma_{22}} & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\sigma_{kk}} \end{pmatrix}^{-1} .$$

Equivalentemente, se obtiene:

$$R = \begin{pmatrix} \mathbf{1} & r_{12} & \dots & r_{1k} \\ r_{21} & \mathbf{1} \dots & & r_{2k} \\ \vdots & \ddots & & \vdots \\ r_{k1} & r_{k2} & \dots & \mathbf{1} \end{pmatrix}.$$

Donde R es una matriz cuadrada y simétrica, cuyos coeficientes indican las covarianzas estandarizadas de las variables relacionadas entre cada una de ellas [61] [72]. Si se suman los elementos de la diagonal principal se obtiene la traza de la matriz de correlación $Tr(R)$ [72] ,la cual refleja la cantidad de variables totales involucradas en la muestra. Una forma de corroborar que los resultados obtenidos son correctos consiste en verificar que la traza de R es igual a la cantidad de variables.

Entonces, a partir de la matriz de correlación, se obtienen los componentes principales Dim_j, los cuales se calculan mediante la combinación lineal entre las variables independientes, quedando expresados como:

$$Dim_j = a_{1j}.x_1 + a_{2j}.x_2 + \dots + a_{pj}.x_k,$$

donde $a_{1j}, a_{2j}, \dots, a_{pj}$ expresan los pesos de cada componente principal correspondientes a los autovectores de la matriz de correlación [68].

Los valores obtenidos de las varianzas de cada componente $Dim_1, Dim_2, \dots Dim_j$, se ordenan de manera decreciente de modo que vaya disminuyendo el grado de variabilidad al incorporar mayor cantidad de componentes principales [72] [73].

$$\sigma^2(Dim_1) \geq \sigma^2(Dim_2) \geq \cdots \geq \sigma^2\left(Dim_j\right), con\ j < k$$

Es así que la primera componente Dim_1 captura la mayor proporción de la varianza total, logrando explicar con mayor relevancia el comportamiento de las variables observables.

3.4.1. Representación de los componentes principales

Luego de obtener las nuevas variables llamadas componentes principales, se debe definir un nuevo eje de coordenadas que permita reflejar de manera aproximada las características de las variables observables, es decir, captando la mayor cantidad de varianza total.

Entonces, con el fin de encontrar un nuevo eje de coordenadas que permita reducir las distancias ortogonales de los puntos con respecto al nuevo eje, se debe diagonalizar la

matriz de correlación conservando la traza de la matriz. A continuación, se muestra gráficamente como se proyecta un punto sobre un nuevo eje de coordenadas:

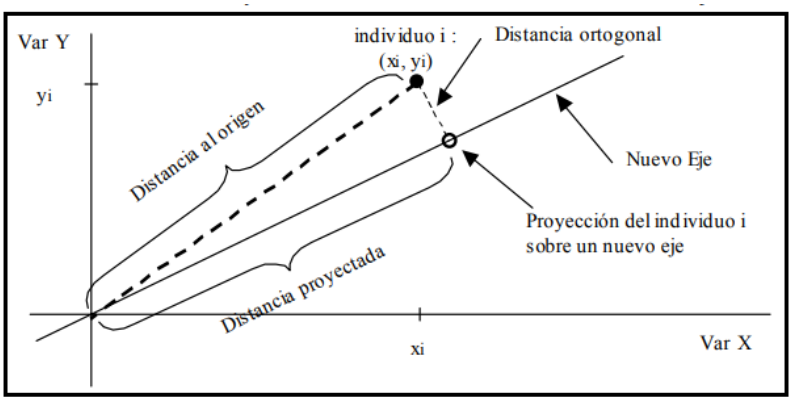


Figura 3.4.2. [61] Proyección de un punto sobre un nuevo eje.

En la Figura 3.4.2. se puede observar la proyección ortogonal de un punto perteneciente a las coordenadas (x_i, y_i) de una variable observable, sobre un nuevo eje, cuya distancia al origen debe ser lo más grande posible para capar mayor varianza y reducir de esta manera las distancias ortogonales [61] [72] [74]. Entonces, la distancia al origen (marcada en la Figura 3.4.2. con líneas punteadas) y la distancia proyectada lograrán aproximarse cuando el coseno cuadrado del ángulo que se forma entre ambas medidas tienda a 1.
Para cumplir con este propósito, se debe diagonalizar la matriz de correlación obteniendo así una nueva matriz, cuyos valores propios presentes en la diagonal principal reflejan las varianzas asociadas a cada componente principal denotados por: $\delta_1, \delta_2, \dots, \delta_k$, mientras que los valores fuera de la diagonal son nulos indicando carencia de variabilidad entre las magnitudes. En particular, a partir de k variables observables se obtiene la siguiente matriz diagonalizada:

$$D = \begin{pmatrix} \delta_1 & 0 & \dots & 0 \\ 0 & \delta_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \delta_k \end{pmatrix},$$

con $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_k$ lo cual muestra una distribución de la inercia proyectada ordenada de manera decreciente [61].

Entonces, con el fin de conservar la mayor cantidad de la inercia total de la matriz de correlación se verifica que $\delta_1 + \delta_2 + \cdots + \delta_k = n$. Es por ello que al sumar los autovalores presentes en la matriz de correlación se obtiene la *cantidad total de variables observables* pertenecientes a la muestra [71] [73].

Luego, al dividir los valores propios de la matriz de correlación por la inercia total, se obtiene la *proporción de varianza total* explicada por cada componente principal, cuyos valores se ordenan de manera decreciente dado que las primeras componentes presentan mayor grado de variabilidad. Si se multiplican cada proporción de varianza por cien se obtienen los *valores porcentuales de la variación total* [61] [73], es decir,

$$\frac{\delta_1}{\boldsymbol{tr(R)}}.100 \geq \frac{\delta_2}{tr(R)}.100 \ldots \geq \frac{\delta_k}{tr(R)}.100 \ .$$

A continuación se presentan dos métodos para evaluar si la muestra se adecua al método de análisis de componentes principales.

3.4.2. Test de esfericidad de Bartlett

Una forma de comprobar que las variables de la muestra están fuertemente correlacionadas consiste en utilizar el *test de esfericidad de Bartlett* [65] [71], el cual compara la matriz de correlación con la matriz identidad. Si los valores de los coeficientes de las matrices son distintos, entonces existe correlación significativa entre las variables, caso contrario, las variables observables no presentan ningún nivel de asociación y por lo tanto, no se podría utilizar el *método de análisis de componentes principales* [69] [73].

3.4.3. Prueba de Kaiser-Meyer-Olkin (KMO)

Otra técnica utilizada para evaluar la implementación de componentes principales, es llamada prueba de *Kaiser-Meyer-Olkin (KMO)*, la cual mide el grado de adecuación a la muestra. El método consiste en analizar si las correlaciones obtenidas entre las variables son poco significativas.

Los valores del coeficiente (KMO) varían entre 0 y 1. Los valores próximos a 1 indican mejor grado de adecuación muestral, mientras que los valores cercanos a 0 indican bajo nivel de adecuación, es decir, las correlaciones parciales son pequeñas o poco significativas. En particular, los coeficientes superiores a 0,7 muestran buena adecuación a la muestra, mientras que los valores inferiores a 0,7 presentan mala adecuación, por lo tanto, no se podría aplicar el método de análisis de componentes principales [64] [66].
Además del método presentado en esta sección, se desarrolla a continuación otra forma de reducir la dimensionalidad de las variables llamado *análisis factorial*.

3.5. Método de análisis factorial

A diferencia del análisis de componentes principales, el método de análisis factorial permite explicar el grado de correlación entre las variables observables considerando el efecto de las *variables latentes o no observables,* llamadas factores. Para efectuar un análisis factorial se parte de: una matriz de variables observables (X), una matriz de factores (F) y una matriz de carga (Λ).

La matriz de variables observables, cuya dimensión es $n \times p$ está formada por los elementos de la muestra, mientras que *la matriz de factores* de dimensión $n \times m$ representa las variables latentes o no observables. Por otra parte, *la matriz de carga* de dimensión $p \times m$, muestra las correlaciones entre las variables observables p y los factores m. Analíticamente, la relación entre la matriz de carga y la de factores, se expresa mediante la siguiente ecuación [68] [77]:

$$X = \mu^T + F\Lambda^T + U, \tag{10}$$

siendo μ^T la traspuesta de la media aritmética de las variables x_i de dimensión $n \times p$ y Λ^T la traspuesta de la matriz de carga $m \times p$ Por otra parte, U representa una matriz que reúne el total de perturbaciones o desvíos sobre las variables observables de dimensión $(n\ x\ p)$ [68] [77] [78].

Con el fin de expresar la varianza en términos de la matriz de carga y de los residuos de las variables (ψ), se presenta a continuación, la propiedad fundamental:

$$V = \Lambda\Lambda^T + \psi, \tag{11}$$

cuya igualdad expresa que la matriz de covarianza V de los datos observados de la muestra, es igual al producto de la matriz de carga Λ por su traspuesta Λ^T más la matriz diagonal, cuyos elementos de la diagonal principal son los residuos de cada variable, mientras que los elementos fuera de la diagonal son iguales a cero. [75] [76].

Usando p variables y m factores la ecuación (12) quedan expresadas como:

$$V = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pm} \end{bmatrix} \cdot \begin{bmatrix} \rho_{11} & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{1m} & \rho_{2m} & \cdots & \rho_{pm} \end{bmatrix} + \begin{bmatrix} \psi_{11} & 0 & \ldots & 0 \\ 0 & \psi_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \psi_{pp} \end{bmatrix},$$

siendo *(ρ)* las cargas factoriales.

3.5.1. Comunalidad y unicidad

La *comunalidad* h_2 mide la proporción de varianza explicada por cada variable, la cual se obtiene a partir de la suma de los pesos factoriales, los cuales reflejan las correlaciones entre las variables observables y los factores. Los valores próximos a uno muestran mayor comunalidad, mientras que los valores próximos a cero indican menor variabilidad explicada por cada variable [68] [76].

Por otra parte, la *unicidad* u_2 representa la varianza explicada que no ha podido ser representada por los factores. Su valor varía entre 0 y 1, siendo los valores cercanos a cero aquellos que representan mayor unicidad [68] [75] [77].

Dado que la comunalidad depende de la cantidad de variables independientes y de los factores que se encuentran en una muestra, a continuación se presentan diferentes mecanismos para hallar el número de factores utilizados en el análisis factorial.

3.5.2. Determinación del número de factores

Entre los diferentes métodos presentes en la bibliografía para hallar la cantidad de factores factibles se encuentran el *análisis factorial confirmatorio* y el *análisis factorial exploratorio*. El *análisis factorial confirmatorio* [79] [80] [81] se utiliza para determinar la cantidad de factores que están asociados a un número de variables observables, donde el número de factores se supone conocido y se establecen restricciones con respecto a la matriz de carga. A continuación se muestra gráficamente un caso particular de dos factores mediante un diagrama de rutas:

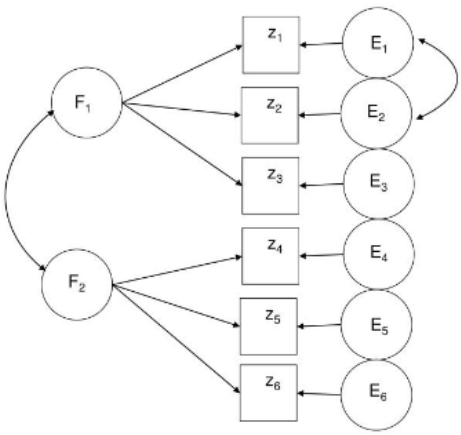


Figura 3.5.1. Análisis Factorial Comprobatorio[6]

[6] https://www.slideserve.com/mervyn/introducci-n-al-an-lisis-factorial-confirmatorio

Siendo la cantidad de factores igual a: F_1 y F_2, $z_1, \dots, z_6$ las variables observables y $E_1, \dots, E_6$ el error específico de cada observación, las flechas unidireccionales indican los pesos atribuibles a cada par F_i-Z_j o Z_j-E_k, mientras que las flechas bidireccionales muestran las covarianzas entre los errores o entre los factores.

En general, para hallar el número de factores de un conjunto de variables se utiliza el *estimador de saturación de factores Omega* [82] [83] [84] el cual permite determinar la estimación más acertada, es decir, el mejor ajuste con menor error en la medición al momento de seleccionar la cantidad de factores. El estimador (ω) se obtiene a partir de la suma de las cargas factoriales (ρ) , como se muestra en la siguiente ecuación:

$$\omega = \frac{(\sum_{i=1}^{n} \rho_{\mathrm{i}})^2}{(\sum_{i=1}^{n} \rho_{\mathrm{i}})^2 + (\sum_{i=1}^{n} 1 - {\rho_{\mathrm{i}}}^2)}.$$

Los valores del coeficiente (ω) superiores a 0,70 presentan confiabilidad aceptable. Por otra parte, *el análisis factorial exploratorio* [85] [86] no requiere que el investigador considere a priori la cantidad de factores ni su grado de correlación, sino que el número de factores se determina en función de aquellos que logren explicar la mayor proporción de varianza acumulada. Para poder visualizar la cantidad de factores en relación a los autovalores, se puede utilizar un *gráfico de sedimentación.*

3.5.3. Pruebas de adecuación y de fiabilidad de los datos.

Al igual que en análisis de componentes principales se utilizan el *test de esfericidad de Bartlett* para analizar el grado de correlación entre las variables observables.

Otro test muy utilizado en la bibliografía es la prueba de *prueba de Káiser-Meyer-Olkin*, cuya finalidad es evaluar el grado de adecuación muestral de los datos. Ambas técnicas son condición necesaria para aplicar el método de análisis factorial, caso contrario, se debería descartar su implementación [76].

Con el fin de determinar el grado de confianza de los datos de la muestra se analizan los indicadores: *Alfa de Crombach* y *Lambda de Guttman*.

El indicador Alfa de Crombach [87] [88] es una prueba de confiabilidad que permite estimar el nivel de consistencia de los datos observados, el cual se calcula a través de la siguiente ecuación:

$$\alpha = \frac{n\,p}{1 + p(n-1)},$$

siendo p el promedio entre las correlaciones de las variables observables y n la cantidad total de variables presentes en la muestra. Los valores aceptables del indicador son aquellos que se encuentran por encima de 0,70. Por otra parte, los valores por debajo del intervalo de aceptación indican que los datos presentan baja consistencia.

Otro indicador muy utilizado para evaluar la consistencia de los datos de una muestra se denomina Lambda de Guttman [89] [90], el cual permite estimar la correlación entre un conjunto de variables como se ve reflejado en la siguiente expresión:

$$\lambda_2 = 1 - \frac{\sum_{j=1}^{n} \sigma_j^2}{\sigma_x^2} + \sqrt{\frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2}{\frac{n-1}{\sigma_x^2}}}\ , \text{con } i \neq j\ ,$$

donde ${\sigma_j}^2$ representa la varianza de la variable j, ${\sigma_x}^2$ la varianza total de los datos observados y $n.\sum_{i=1}^{n} \sum_{j=1}^{n} {\sigma_{ij}}^2$expresa la sumatoria de las covarianzas entre las variables i y j multiplicada por la cantidad de variables. Al igual que en indicador Alpha de Crombrach, los valores mayores a 0,70 muestran un alto grado de confiabilidad mientras que los valores por debajo de este indicador presentan una baja aceptación con respecto a la confianza de los datos de la muestra.

3.5.4. Rotación de factores

En la sección anterior se mencionó que un factor está asociado a un conjunto de variables observables, cuya relación se ve reflejada en una matriz de carga. Entonces, para poder facilitar la interpretación de los factores, se deben alinear los componentes de la matriz de carga con el sistema de coordenadas mediante una rotación.

Existen diferentes criterios para rotar los factores, entre los más utilizados se encuentran: *la rotación ortogonal* o *varimax y la oblicua* u *oblimin*. El primero es un método de rotación que minimiza el número de variables que tienen altas cargas en cada factor, busca evitar que los factores estén correlacionados. Para ello se produce una rotación de 90° en los ejes. El Grafico 3.5.2 esquematiza la situación explicada:

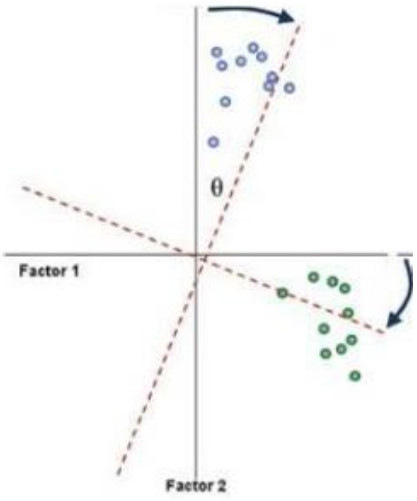


Figura 3.5.2. Rotación varimax.[7]

Para reducir las cargas elevadas en los factores, se debe maximizar la suma de las varianzas entre los factores y las variables observables, como se muestra en la siguiente ecuación:

$$\frac{1}{p}\sum_{i=1}^{p}(\delta_{ij}{}^2 - \overline{\delta_{.j}})^2 = \frac{1}{p}\sum_{i=1}^{p}\delta_{ij}{}^4 - \left(\frac{1}{p}\right)^2 \cdot (\sum_{i=1}^{p}\delta_{ij}{}^2)^2 \ ,$$

donde δ_{ij} representa los componentes de la matriz de carga , los cuales están relacionados al factor j, con $i = 1,..,p$. Por otra parte, $\overline{\delta_{.j}}$ indica el promedio del vector columna j de la matriz de carga después de la rotación [68].

Por otra parte el método de rotación oblicua se produce cuando los vectores giran en un ángulo menor a 90°. El objetivo principal de esta rotación es aumentar el grado de correlación entre los factores y las variables [61] [68] [91]. El Gráfico 3.5.3 representa este caso:

[7] Recuperado de: https://www.slideshare.net/rajdeepkraut/factor-analysis-fa

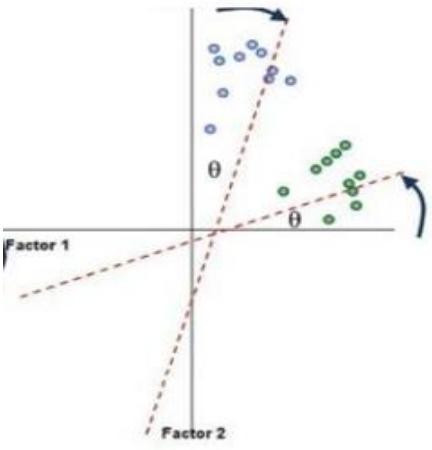


Figura 3.5.3. Rotación Oblimin.[8]

Para entender la diferencia entre ambas rotaciones se considera un ejemplo de aplicación.

Ejemplo 3.5.1.

Se suponen, en este caso, dos factores llamados Factor 1 y Factor 2 y un conjunto de 9 variables representadas por puntos en los ejes cartesianos, se aplica una rotación varimax y oblimin con la finalidad de comparar los resultados obtenidos. En la Figura 3.5.4., se muestran los puntos de cada variable sin sufrir rotación y los vectores que salen del origen, los cuales representan la dirección y sentido de cada variable. Los vectores que están próximos entre sí, indican mayor grado de correlación entre las magnitudes, mientras que si están alejados la correlación es débil.

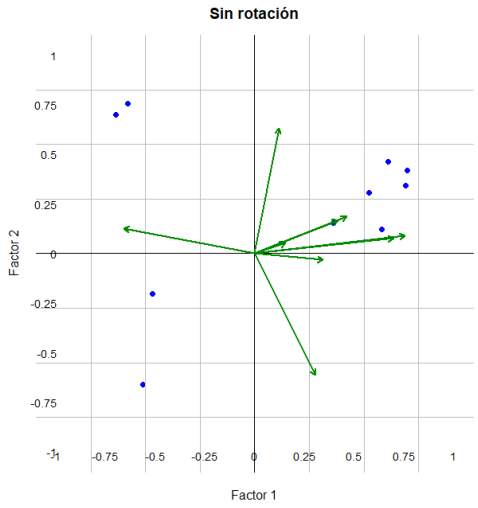


Figura 3.5.4. Puntuaciones sin rotación de factores.

[8] Recuperado de: https://www.slideshare.net/rajdeepkraut/factor-analysis-fa

Posteriormente se realiza una rotación ortogonal con el objetivo de reducir el número de variables que presentan cargas altas en cada factor y se obtiene el siguiente gráfico:

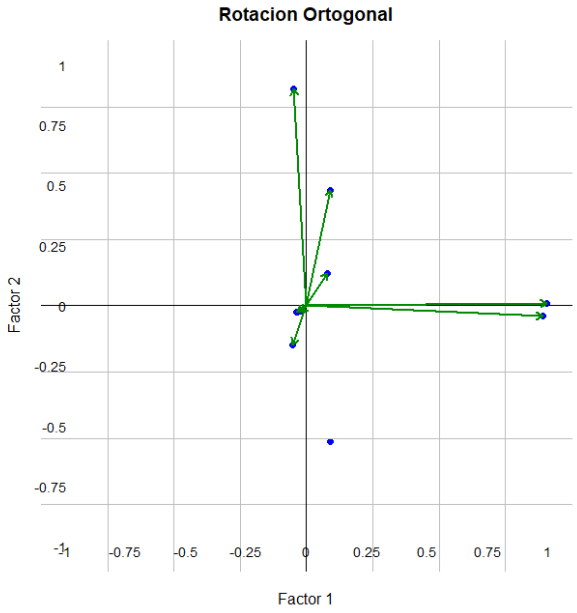


Figura 3.5.5. Puntuaciones con rotación ortogonal de factores.

En el Gráfico 3.5.5. se puede observar que luego de la rotación varimax los puntos se alinean con los vectores y además, se ubican próximos a los semiejes positivos de abscisas y ordenada pertenecientes a los factores F1 y F2. Posteriormente, con el fin de utilizar otro método de rotación y de contrastar sus resultados se aplica una rotación oblicua y se obtiene la siguiente Figura:

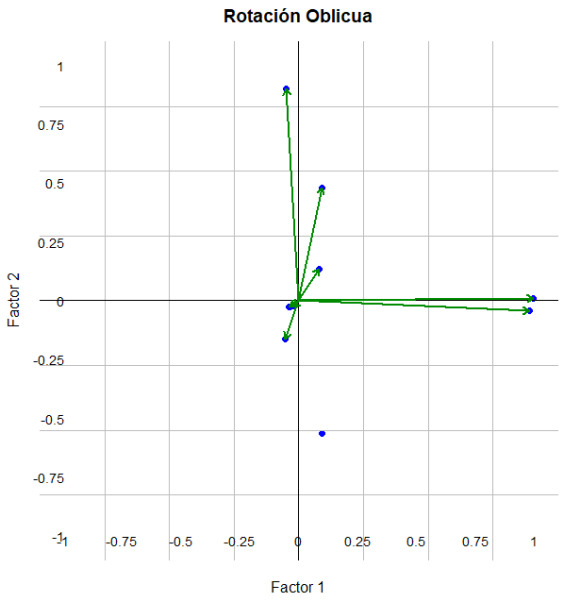


Figura 3.5.6. Puntuaciones con rotación oblicua de factores.

En la Figura 3.5.6. se puede notar que los vectores no han sufrido variación con respecto a la rotación de los ejes presentada en el Gráfico 3.5.6, dado que el nivel de correlación permanece invariable. Por lo tanto, en este caso, con el objetivo de ajustar los coeficientes de la matriz de carga con el sistema de coordenadas, se puede utilizar tanto la rotación varimax como la rotación oblimin.

En el próximo capítulo se muestra el desarrollo de un modelo multivariado aplicado a un caso concreto del ámbito empresarial.

Capítulo 4

ESTUDIO DE CASO: MODELO DE ANÁLISIS MULTIVARIADO

4.1. Presentación del problema.

Con el objetivo de estudiar cuales son las causas que inciden en el volumen de ventas de la actividad productiva ganadera que se desarrolla en diferentes regiones de la República Argentina, se utiliza el programa "R Studio" para analizar una base de datos obtenida a partir del repositorio público del gobierno nacional [9] llamada "producción carne bovina".

En primera instancia se parte de un análisis exploratorio, recopilando información estadística descriptiva de la base de datos. Luego, se plantean dos tipos de modelos utilizados para reducir la dimensión de las variables observables con el fin de determinar cuál de ellos describe mejor el caso de estudio propuesto.

4.2. Análisis descriptivo preliminar.

Con el fin de analizar la naturaleza de los datos se efectúa un análisis estadístico descriptivo a partir de 2.598 registros, tomando las siguientes variables cuantitativas:

Margen_bruto($/ha), resultado_neto($/ha), ingreso_neto($/ha), gastos_directos($/ha), costos_indirectos($/ha), eficiencia_stock(%), producción(Kg/ha) y carga(Kg/ha).

El ingreso neto, expresa el volumen de ventas de carne bovina producida por hectárea, a partir del cual, al deducir el costo de la mercadería vendida se obtiene el margen bruto o beneficio económico por hectárea. Luego, descontando del margen bruto, gastos operativos y financieros se obtiene el resultado neto, el cual refleja la ganancia real obtenida por una empresa.

Por otra parte, los gastos directos son aquellos asociados a la producción ganadera por hectárea, por ejemplo: costos de insumos o gastos de transporte, mientras que los gastos o costos indirectos representan erogaciones que involucran diferentes procesos productivos dentro de una empresa ganadera, como es el caso del costo de alquiler del campo o el seguro de producción [94] [95]. Finalmente, la carga comprende la cantidad de animales que permanecen en una superficie determinada por periodo de tiempo,

[9] Recuperado de: www. https://datos.gob.ar/dataset/agroindustria-ganaderia---produccion-carne-bovina/archivo/agroindustria_cabb1226-e84b-4b0b-a1c4-6a98f854760e

mientras que la producción o rinde, muestra la cantidad de kilos de carne producidos por hectárea a diferencia de la eficiencia stock, la cual representa el porcentaje de kilos de carne producido por animal [96] [97].

A continuación se muestra una tabla que contiene los siguientes indicadores: valores máximos, mínimos, los cuartiles, el coeficiente de variación y el valor de la media, mediana y moda de cada parámetro analizado.

	Margen bruto ($/Ha.)	Resultado neto ($/Ha.)	Ingreso neto ($/Ha.)	Gastos directos ($/Ha.)	Costos indirectos ($/Ha.)	Eficiencia stock (%)	Producción (Kg/Ha.)	Carga (Kg/Ha.)
Mínimo	-107,00	-1299,00	205,00	46,00	75,00	20,00	7,00	19,00
Máximo	3148,00	1761,00	5980,00	4424,00	2522,00	59,00	284,00	492,00
Media	1001,076	414,787	1585,558	584,457	586,319	32,755	78,708	209,244
Mediana	645,00	241,00	1013,00	228,00	401,00	31,00	46,00	182,00
Moda	341,00	1330,00	2900,00	256,00	466,00	31,00	92,00	276,00
Rango	3255,00	3060,00	5775,00	4378,00	2447,00	39,00	277,00	473,00
Desvío	725,40	532,00	1344,485	864,814	486,093	9,402	76,291	134,132
Varianza	526219,47	283023,69	1807638,7	747903,1	236286,51	88,406	5820,347	17991,41
1° cuartil	477,00	119,00	664,00	137,00	241,00	27,00	28,00	101,00
3° cuartil	1400,00	750,00	2286,00	423,00	777,00	35,00	92,00	293,00
Coeficiente Variación (%)	72,463	128,259	84,796	147,969	82,906	28,705	96,929	64,103

Tabla. 4.1.1. Análisis estadístico descriptivo.

A partir de la Tabla. 4.1.1. se puede observar que el valor con mayor varianza está representado por la variable ingreso_neto, mientras que la magnitud producción, muestra la menor varianza. Por otro lado, el parámetro que presenta mayor coeficiente de variación es gastos_directos, siendo eficiencia_stock la que contiene el coeficiente más bajo. Por último, el valor del rango más representativo proviene de la variable ingreso_neto, en tanto, la magnitud producción presenta menor valor de rango por hectárea.

Con el objetivo de visualizar la distribución de los datos de la variable ingreso_neto y gastos_directos se muestran a continuación los siguientes gráficos:

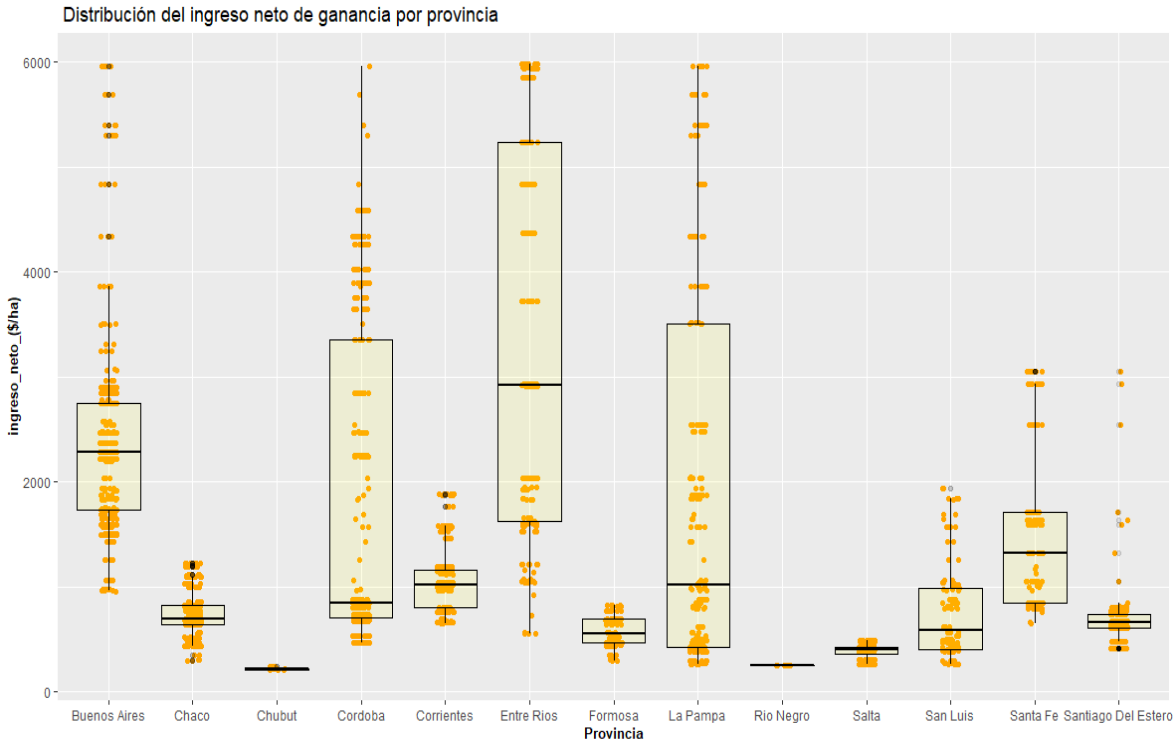


Figura. 4.1.1. Ingreso neto de ganancia por provincia.

La Figura 4.1.1. representa un gráfico de boxplot o de caja y bigote, el cual muestra la distribución del ingreso neto de ganancia por provincia en la relación a la mediana, primer cuartil, tercer cuartil, puntos máximos y mínimos de cada categoría. Las provincias con mayor nivel de dispersión de datos son: Buenos Aires, Entre Ríos, La Pampa y Córdoba. Por otra parte, las categorías que presentan mayor concentración de datos son: Chaco, Formosa, Corrientes, Salta, San Luis, Santa Fe y Santiago del Estero. Las provincias con mayor nivel de ingreso neto son: Entre Ríos, Buenos Aires y Santa Fe, mientras que, Chubut, Rio Negro y Salta tienen menor ingreso neto por hectárea.

Por otro lado, la distribución de los gastos directos por provincia, se representa en el siguiente gráfico:

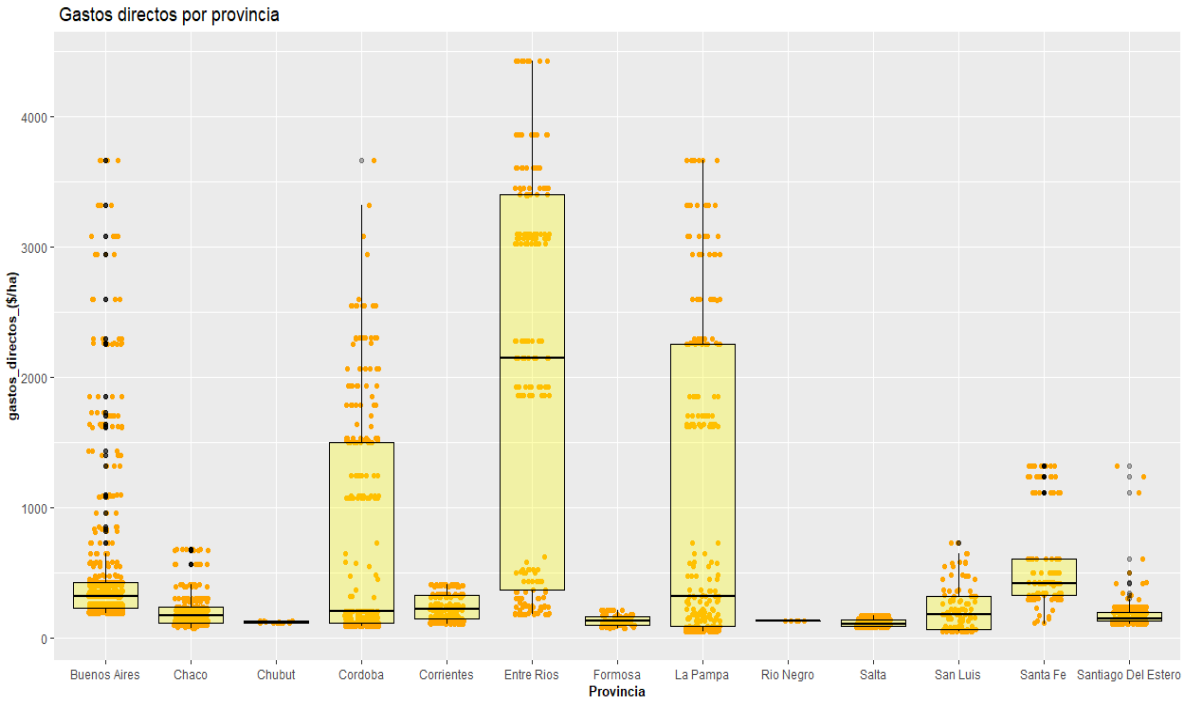


Figura. 4.1.2. Gastos directos de producción por provincia.

En el gráfico de boxplot representado en la Figura 4.1.2. se muestran los gastos directos asociados a la producción bovina por provincia, siendo las categorías con mayor grado de dispersión de gastos: Entre Ríos, La Pampa, Córdoba y Buenos Aires. Por otra parte, las provincias con menor dispersión de datos son: Corrientes, Formosa y Salta, mientras que, Chaco, San Luis, Santa Fe y Santiago del Estero presentan un nivel de dispersión medio. Las provincias con mayor gasto directo de producción son: Entre Ríos, Santa Fe y La Pampa. En tanto, Chubut, Rio Negro y Salta tienen el menor gasto directo de producción bovina por hectárea.

Además de representar la distribución de los datos a través de un gráfico de Boxplot, se puede analizar cómo se distribuyen las frecuencias de las variables, mediante el siguiente *histograma*:

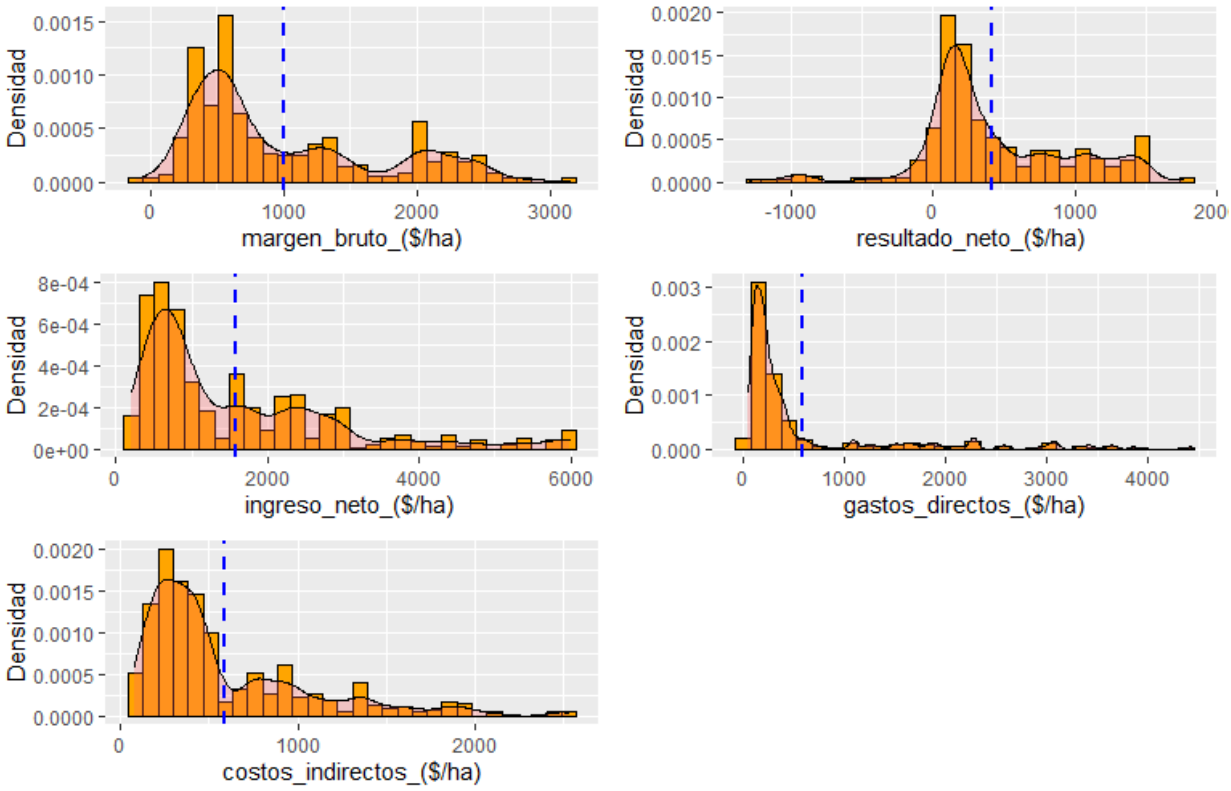


Figura 4.1.3. Histograma del mercado ganadero.

Con respecto a la Figura 4.1.3. se puede apreciar que la distribución de datos con respecto al ingreso_neto no es normal dado que las curvas de distribución no son simétricas. El mismo comportamiento se observa en las variables: margen_bruto, resultado_neto, gastos_directos y costos_indirectos, los cuales presentan una distribución levemente sesgada hacia la derecha, es decir, los datos se encuentran más alejados de su media aritmética.

4.3. Método 1. Análisis de componentes principales (ACP)

A partir de la base de datos presentada en 4.1.1. se obtiene la matriz A, la cual representa de los datos de una muestra, cuyas variables observables son:

Margen bruto ($/Ha.)	Resultado neto ($/Ha.)	Ingreso neto ($/Ha.)	Gastos directos ($/Ha.)	Costos indirectos ($/Ha.)	Eficiencia stock (%)	Producción (Kg/Ha.)	Carga (Kg/Ha.)
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

Tabla 4.1.2. Selección de variables cuantitativas.

Luego con el fin de calcular los desvíos de los valores de la matriz *A*, se obtiene el *vector de medias*, el cual, está compuesto por las medias aritméticas de cada magnitud, como se indica a continuación [52]:

$$\bar{x} = \frac{1}{n} . \sum_{i=1}^{n} x_i = \begin{bmatrix} \overline{x_1} \\ \vdots \\ \overline{x_p} \end{bmatrix}.$$

Es así que el vector de medias para el caso analizado queda representado por:

$$\bar{x} = \begin{bmatrix} \overline{x_1} \\ \overline{x_2} \\ \overline{x_3} \\ \overline{x_4} \\ \overline{x_5} \\ \overline{x_6} \\ \overline{x_7} \\ \overline{x_8} \end{bmatrix}, \text{ es decir, } \quad \bar{x} = \begin{bmatrix} 1001{,}0762 \\ 414{,}7868 \\ 1585{,}5581 \\ 584{,}4565 \\ 586{,}3195 \\ 32{,}7552 \\ 78{,}7082 \\ 209{,}2444 \end{bmatrix}$$

Geométricamente el vector columna $\bar{x}$ representa el baricentro o centro de gravedad de la nube de puntos de los datos multivariados presentes en un espacio de 8 dimensiones [45]. Entonces, con el fin de obtener la variabilidad de los datos con respecto a las medias aritméticas de cada parámetro, se calcula la *matriz de varianzas y covarianzas*, cuyos valores se presentan en la siguiente tabla:

	Margen bruto ($/Ha.)	Resultado neto ($/Ha.)	Ingreso neto ($/Ha.)	Gastos directos ($/Ha.)	Costos indirectos ($/Ha.)	Eficiencia Stock (%)	Producción (Kg/Ha.)	Carga (Kg/Ha.)
Margen bruto ($/Ha.)	526219,47	286461,46	792986,29	266775,76	239765,49	2671,66	28888,62	67502,80
Resultado neto ($/Ha.)	286461,46	283023,69	178940,35	-107517,19	3462,23	-770,57	-2041,15	10978,18
Ingreso neto ($/Ha.)	792986,29	178940,35	1807638,72	1014670,01	613987,204	9644,961	90295,73	163822,18
Gastos directos ($/Ha.)	266775,76	-107517,19	1014670,01	747903,17	374226,68	6973,42	60406,87	96317,92
Costos indirectos ($/Ha.)	239765,49	3462,23	613987,20	374226,68	236286,51	3441,48	31926,07	56522,29
Eficiencia stock (%)	2671,66	-770,57	9644,96	6973,42	3441,48	88,40	662,15	1015,67
Producción (Kg/Ha.)	29888,62	-2041,15	90295,73	60406,87	31926,07	662,15	5820,34	9756,54
Carga (Kg/Ha.)	67502,80	10978,18	163822,18	96317,92	56522,29	1015,67	9756,51	17991,41

4.1.3. Matriz de varianza y covarianza.

Los valores presentes en la diagonal principal de la Tabla 4.1.3. muestran la varianza de cada variable explicativa, mientras que los valores por encima y por debajo de la diagonal expresan las covarianzas entre cada par de parámetros. En general se puede observar una correlación directa entre los parámetros, dado que las covarianzas son positivas. Por otro lado, las covarianzas entre: el resultado_neto y los gastos_directos; la eficiencia_stock y la producción, son negativas mostrando un nivel de correlación inversa. En especial, analizando el comportamiento del ingreso_neto con el resto de las variables, se puede notar que las covarianzas son positivas describiendo una correlación directa entre las mismas.

Con el fin de eliminar distorsiones entre las magnitudes causadas por el uso de diferentes unidades de medida, se estandarizan los datos de la matriz de varianza y covarianza obteniendo así la siguiente *matriz de correlación*:

	Margen bruto ($/Ha.)	Resultado neto ($/Ha.)	Ingreso neto ($/Ha.)	Gastos directos ($/Ha.)	Costos indirectos ($/Ha.)	Eficiencia stock (%)	Producción (Kg/Ha.)	Carga (Kg/Ha.)
Margen bruto ($/Ha.)	**1,00**	0,74	0,81	0,43	0,68	0,39	0,54	0,69
Resultado neto ($/Ha.)	0,74	**1,00**	0,25	-0,23	0,01	-0,15	-0,05	0,15
Ingreso neto ($/Ha.)	0,81	0,25	**1,00**	0,87	0,94	0,76	0,88	0,91
Gastos directos ($/Ha.)	0,43	-0,23	0,87	**1,00**	0,89	0,86	0,92	0,83
Costos indirectos ($/Ha.)	0,68	0,01	0,94	0,89	**1,00**	0,75	0,86	0,87
Eficiencia stock (%)	0,39	-0,15	0,76	0,86	0,75	**1,00**	0,92	0,81
Producción (Kg/Ha.)	0,54	-0,05	0,88	0,92	0,86	0,92	**1,00**	0,95
Carga (Kg/Ha.)	0,69	0,15	0,91	0,83	0,87	0,81	0,95	**1,00**

Tabla 4.1.4. Matriz de correlación.

En la Tabla 4.1.4. se obtiene una matriz, donde se comprueba que la mayoría de los coeficientes de correlación entre el ingreso_neto y las demás variables son positivos

indicando correlación positiva fuerte, mientras que el par ingreso_neto y resultado_neto es positiva débil.

Para poder visualizar la intensidad y el grado de asociación entre las variables analizadas utilizando el coeficiente de correlación de Pearson, se muestra a continuación el siguiente *correlograma*:

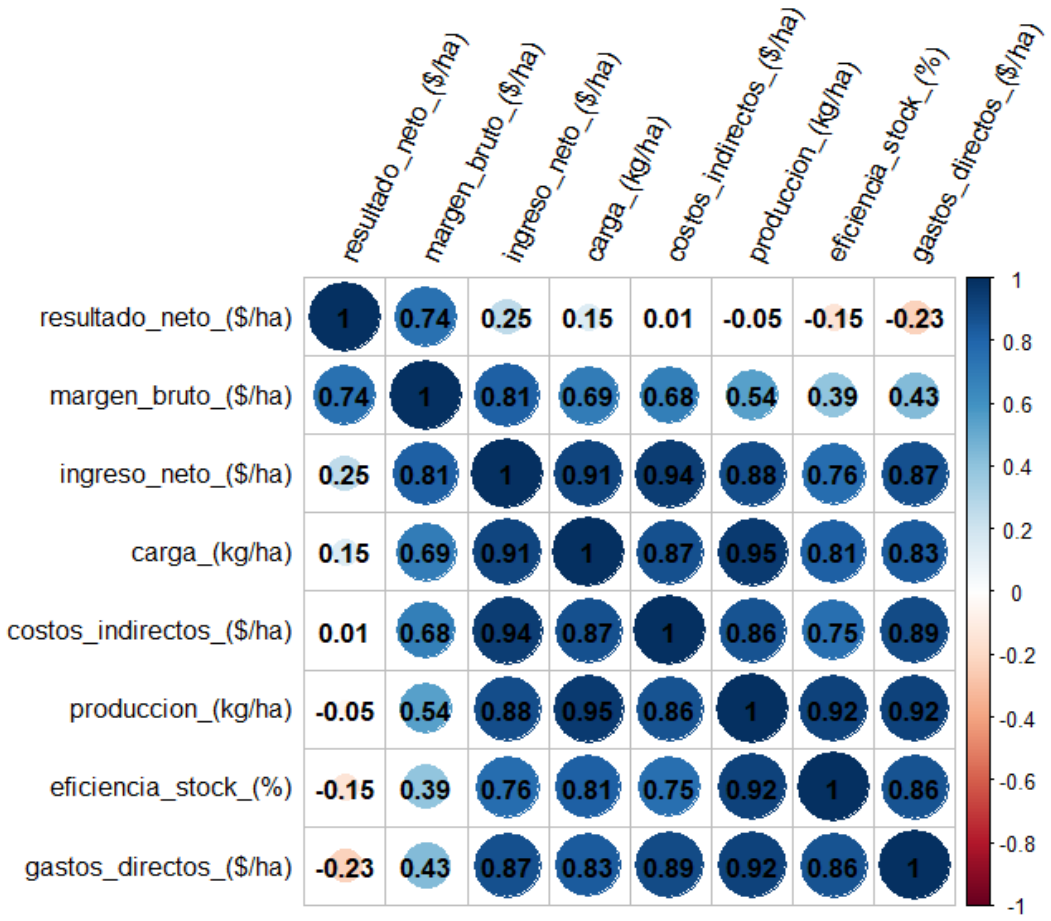


Figura 4.1.4. Correlograma del mercado ganadero.

En la Figura 4.1.4. se puede notar que la mayoría de las correlaciones entre el ingreso_neto y las demás variables están representadas por esferas azules de color intenso indicando correlación positiva fuerte. No obstante, la magnitud ingreso_neto correlaciona de manera positiva pero con menor intensidad con la variable resultado_neto.

En base al correlograma, con el propósito de mostrar cómo se distribuyen los datos de las variables analizadas, se presenta el siguiente *grafico de dispersión*:

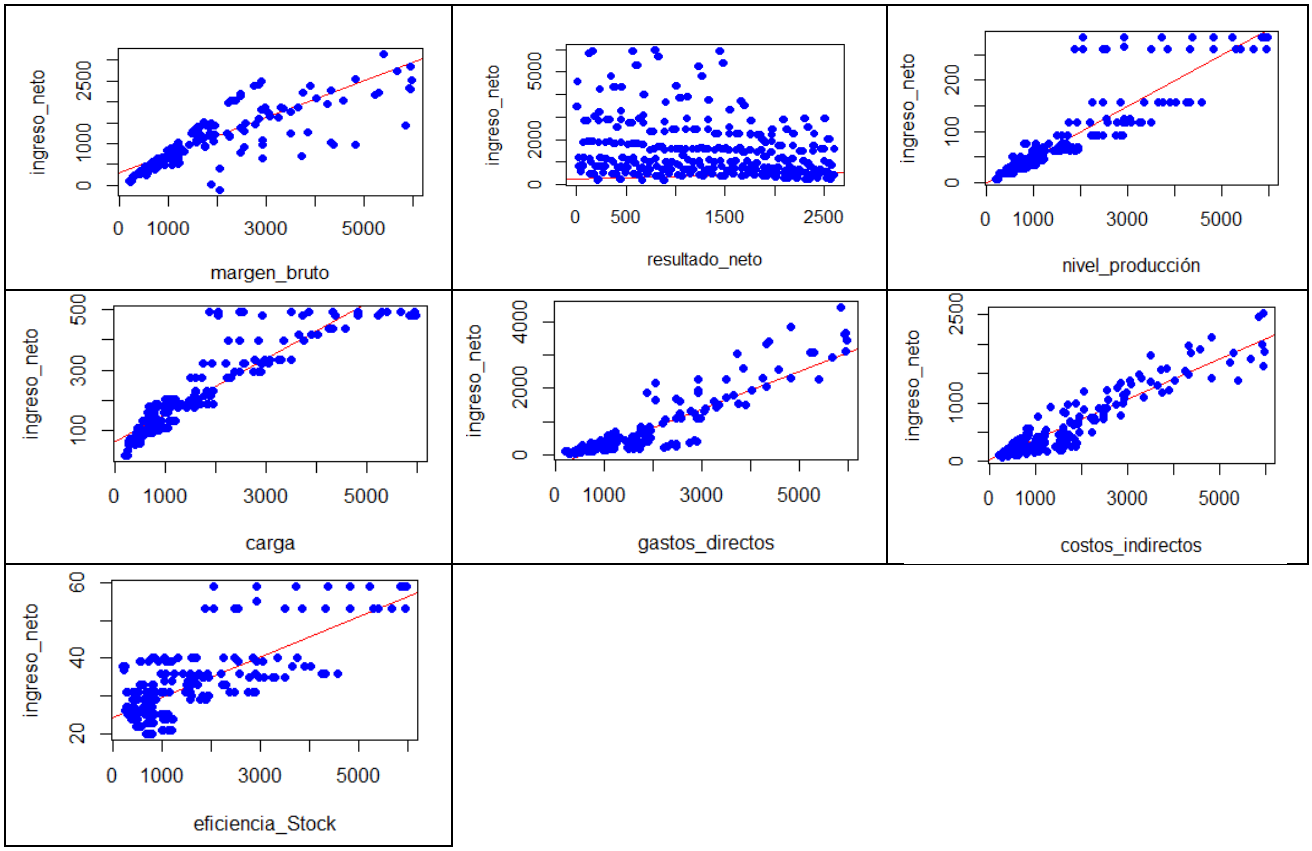


Figura 4.1.5. Dispersograma del mercado ganadero.

Con respecto a la Figura 4.1.5. se observan que en general las nubes de puntos se encuentran concentradas y próximas a la recta de regresión indicada en color rojo, con excepción de los gráficos representados entre las magnitudes: ingreso_neto y resultado_neto e ingreso_neto y eficiencia_Stock, donde prevalece mayor dispersión de datos. Por otra parte, todas las rectas de regresión que relacionan el ingreso_neto y los demás parámetros son crecientes, lo cual indica un grado de asociación positivo entre las magnitudes analizadas. Es decir, en la figura 4.1.5. se observa que los gastos_directos y los costos_indirectos son las variables que resultan más significativas a la hora de estudiar y de evaluar el ingreso neto. Además, el ingreso_neto aumenta conforme aumenten las variables ya mencionadas.

Además del correlograma y del dispersograma, se presenta a continuación otro método utilizado en el análisis multivariado para analizar el grado de asociación entre los parámetros, llamado *test de esfericidad de Bartlett*, el cual fue tratado en la Sección 3.4.2 del Capítulo 3. La matriz de correlación referente a la Tabla 4.1.4. difiere considerablemente de la matriz identidad, por lo que en términos del test de esfericidad de Bartlett, el método de análisis de componentes principales es utilizable.

4.3.1. Pruebas de adecuación muestral.

Con el fin de determinar el grado de adecuación muestral, es decir, evaluar si el valor de la muestra es representativo de manera tal que permita la utilización del análisis de componentes principales, se aplica la prueba de Kaiser-Meyer-Olkin (KMO). El resultado del coeficiente (KMO) es igual a 0,72 lo cual indica que la adecuación de la muestra es buena. Por lo tanto, en base al test de esfericidad de Bartlett y a la prueba de de Kaiser-Meyer-Olkin se concluye que el modelo de análisis de componentes principales puede ser implementado a la base de datos del mercado ganadero.

Por otra parte, para identificar que magnitud tiene mayor influencia con respecto a la adecuación muestral, se puede calcular la medida de adecuación muestral (MSA) de cada parámetro, cuyo valor se muestra en la siguiente tabla:

Margen_ bruto ($/Ha.)	Resultado_ neto ($/Ha.)	Ingreso_ neto ($/Ha.)	Gastos_ directos ($/Ha.)	Costos_ indirectos ($/Ha.)	Eficiencia_ stock (%)	Producción (Kg/Ha.)	Carga (Kg/Ha.)
0,64	0,32	0,76	0,74	0,73	0,77	0,75	0,78

Tabla 4.1.5. Medida de adecuación muestral (MSA).

A partir de la Tabla 4.1.5 se observa que la medida de adecuación muestral del ingreso_neto representa un valor intermedio entre el mayor coeficiente atribuible a la variable carga y el menor valor correspondiente al atributo resultado_neto.

En la próxima sección se obtienen los componentes principales (Dim_j), a partir de la matriz de correlación calculada en el Sección 4.1.3.

4.3.2. Obtención de los componentes principales.

Variable (x_k)	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5	Dim_6	Dim_7	Dim_8
Margen bruto ($/Ha.)	-0,1151	0,6107	0,0449	0,1655	-0,1775	0,2932	0,3869	-0,5630
Resultado neto ($/Ha.)	-0,4584	0,5128	-0,3742	0,1713	0,2685	-0,0578	-0,0381	0,5295
Ingreso neto ($/Ha.)	0,2322	0,3784	0,1809	0,1241	0,3052	-0,3401	-0,6669	-0,3177
Gastos directos ($/Ha.)	0,4569	0,0698	0,2489	0,1112	0,6513	-0,0095	0,5075	0,1732
Costos Indirectos ($/Ha.)	0,3303	0,3504	0,4501	0,0665	-0,5368	0,0471	-0,0670	0,5157

Eficiencia stock (%)	0,4136	-0,1050	-0,5351	0,6581	-0,2344	-0,1947	0,0674	-0,0288
Producción (Kg/Ha.)	0,3907	0,0929	-0,3664	-0,2332	0,1247	0,7351	-0,3005	0,0600
Carga (Kg/Ha.)	0,2844	0,2708	-0,3739	-0,6507	-0,1416	-0,4628	0,2170	-0,0328

Tabla 4.1.6. Vector de cargas de los componentes principales.

En la Tabla 4.1.6. las cargas de los componentes principales denotados por: Dim_1,…, Dim_8, están representadas por cada vector columna o autovectores, los cuales pueden ser esquematizados mediante un gráfico de biplot.

Posteriormente, se realiza una descomposición espectral de la matriz de correlación para obtener los autovalores asociados a los autovectores, cuyo resultado es:

Compo-nente	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5	Dim_6	Dim_7	Dim_8
Auto-valores	5,79461	1,70776	$2{,}9237.10^{-1}$	$1{,}2739.10^{-1}$	$7{,}0545.10^{-2}$	$7{,}3063.10^{-3}$	$2{,}2147.10^{-7}$	$8{,}7989.10^{-8}$

Tabla 4.1.7. Autovalores de cada componente principal.

En función de lo desarrollado en la Sección 3.4. del Capítulo 3, se obtienen los autovalores de cada componente principal, los cuales se plasman en la Tabla 4.1.7., donde se puede notar que las dos primeras componentes tienen mayor peso en relación a las demás componentes.

Luego sumando los autovalores de la matriz de correlación se obtiene la traza, cuyo resultado es:

$$5{,}7946 + 1{,}70776 + 2{,}9237.10^{-1} + 1{,}2739.10^{-1} + 7{,}0545.10^{-2} + +7{,}3063.10^{-3} + 2{,}2147.10^{-7} + 8{,}7989.10^{-8} \approx 8,$$

Siendo 8 la cantidad total de parámetros observados.

A continuación, con el objetivo de visualizar la relación entre los autovalores y los componentes principales, se muestra el siguiente gráfico de sedimentación:

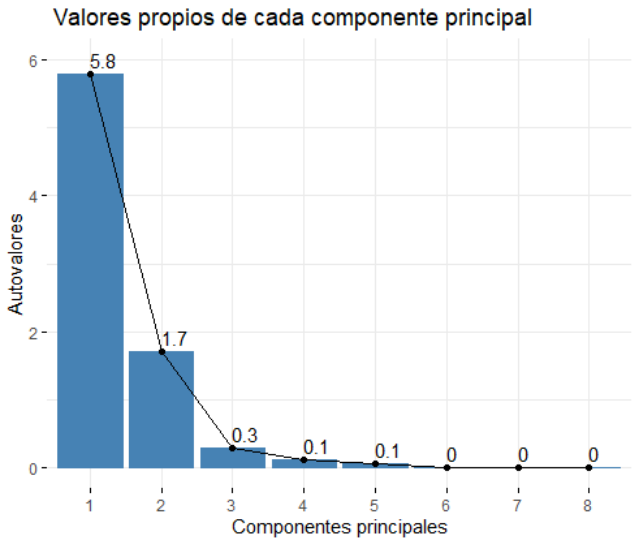


Figura 4.1.6. Valores propios en función de los componentes principales.

En el gráfico de barras esquematizado en la Figura 4.1.6. se observa que las dos primeras componentes logran captar mayor proporción de varianza representadas por los valores propios: 5,8 y 1,7 respectivamente. Por lo tanto, para realizar un análisis de componentes principales se pueden utilizar los dos primeros componentes denotados por Dim_1 y Dim_2. Posteriormente, al dividir cada valor propio de la matriz de correlación por su traza se obtiene la proporción de varianza de cada componente principal, como se muestra la siguiente tabla:

Componente	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5	Dim_6	Dim_7	Dim_8
Proporción de varianza	0,7243	0,2134	0,0365	0,0159	0,0088	0,0009	$2{,}7683.10^{-8}$	$1{,}0998.10^{-8}$

Tabla 4.1.8. Proporción de varianza de cada componente principal.

En la tabla 4.1.8. se puede observar que la primera componente principal captura 72,43% de varianza, seguida de la segunda componente cuyo valor es de 21,34 %.

En la próxima tabla se muestra la relación entre la varianza porcentual y acumulada de las componentes principales:

Componente	$\boldsymbol{Dim_1}$	$\boldsymbol{Dim_2}$	Dim_3	Dim_4	Dim_5	Dim_6	Dim_7	Dim_8
Varianza	5,795	1,708	0,292	0,127	0,071	0,007	0,000	0,000
Porcentaje de varianza	72,433	21,347	3,655	1,592	0,882	0,091	0,000	0,000
Porcentaje de varianza acumulada	72,433	**93,780**	97,434	99,027	99,909	100,00	100,00	100,00

Tabla 4.1.9. Autovalores de la matriz de componentes principales.

En la Tabla 4.1.9. se puede notar que las primeras componentes principales: Dim_1 y Dim_2 acumulan un total de 93,78% de variabilidad total; logrando explicar mejor el comportamiento de las variables observables. Con el objetivo de representar el porcentaje de varianza explicada y la relación entre los autovalores y cada componente principal, se presenta a continuación el siguiente gráfico:

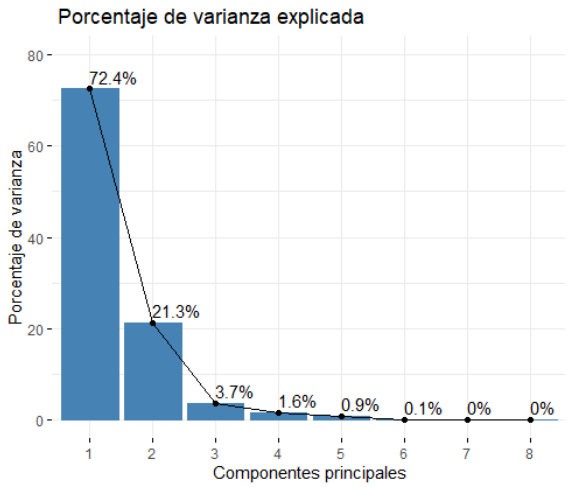


Figura 4.1.7. Varianza porcentual explicada por cada componente principal.

En la Figura 4.1.7. se utiliza un gráfico de sedimentación para mostrar la relación entre la cantidad de componentes principales y la proporción de varianza explicada, donde se puede observar que las dos primeras componentes logran captar mayor porcentaje de varianza, correspondientes a 93,7%. Por otra parte, si se consideran 5 componentes se obtiene 99,90% de varianza acumulada permitiendo minimizar la pérdida de información correspondiente a las magnitudes originales presente en este caso de estudio.

4.3.3. Representación gráfica de los componentes principales.

Para representar en el plano, el grado de correlación entre los parámetros y los componentes principales, se deben seleccionar los componentes que presentan mayor varianza. Entonces, se toman las dos primeras componentes principales denotadas por Dim_1 y Dim_2, dado que en base a la Figura 4.1.6 se observa que presentan mayor proporción de varianza correspondientes a: 72,4% y 21,3% logrando explicar con mayor precisión el comportamiento entre las variables observable y los componentes principales. Dicha relación se puede analizar mediante el siguiente gráfico:

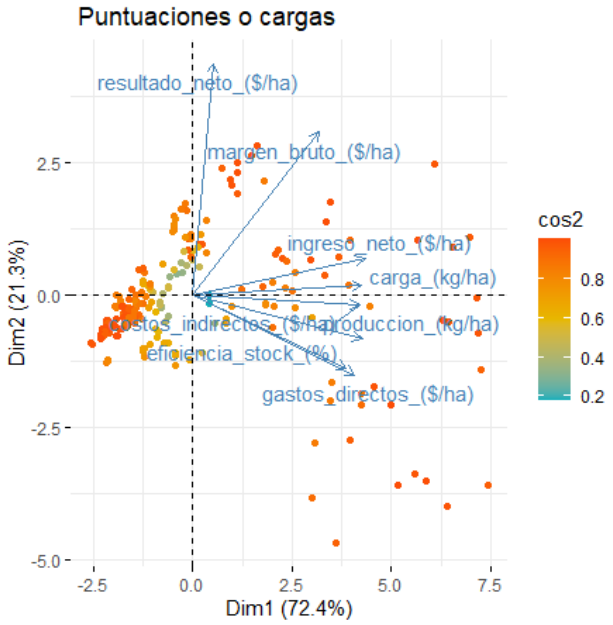


Figura 4.1.8. Puntuaciones.

En la Figura 4.1.8. los puntos representan las cargas, las cuales se obtienen a partir de las correlaciones entre los parámetros y los componentes principales Dim_1 y Dim_2. Por otra parte, se observa que el comportamiento del *resultado_ neto* está representado por el componente Dim_2, cuya varianza acumulada es de 21,3%, mientras que las siguientes magnitudes: *carga, costos_indirectos, producción, ingreso_neto, eficiencia_stock, gastos_directos y margen_bruto* son mejor explicadas por el componente principal Dim_1, donde la proporción de varianza es mayor alcanzando un 72,4%. Además, las puntuaciones de color rojo muestran mayor proporción de varianza explicada, mientras que los de color azul menor porcentaje. En particular, el ingreso_neto, al igual que la carga, la producción, la eficiencia stock, los gastos directos y los costos indirectos, se encuentran bien representados por los componentes principales, dado que el coseno cuadrado entre la distancia al origen y la distancia proyectada de cada variable varía entre 0,60 y 1 logrando capturar mayor proporción de varianza según lo expuesto en el Capítulo 3, Sección 3.4.1.

Luego, con el objetivo de analizar el grado de asociación ente las componentes principales y las cargas, teniendo en cuenta el comportamiento de los vectores, se muestra a continuación el siguiente gráfico:

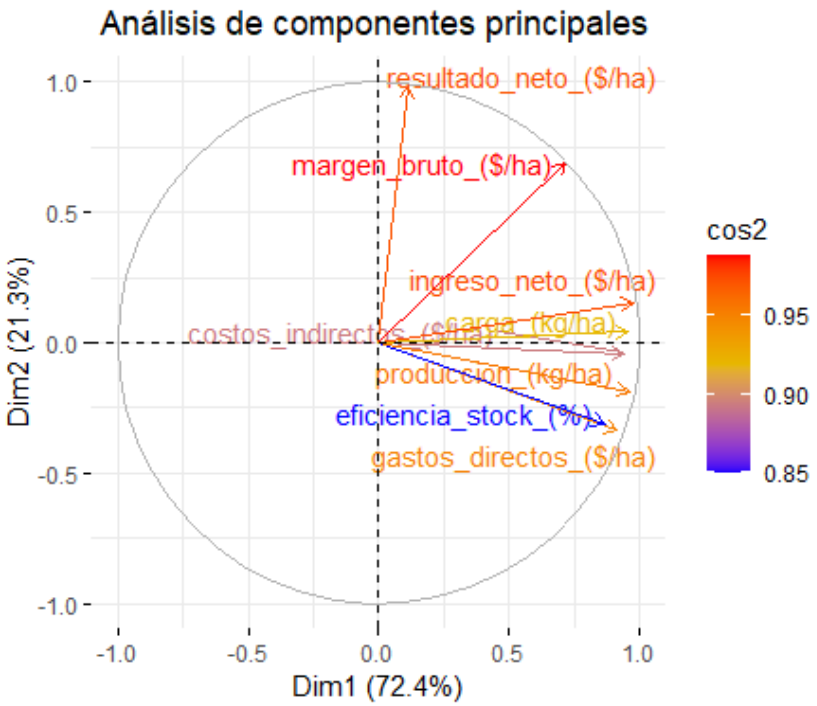


Figura 4.1.9. Proyección de las puntuaciones en relación a las dimensiones 1 y 2.

En la Figura 4.1.9. se muestra una *circunferencia de correlación* o gráfico de *"biplot*", cuyo eje de abscisas varía entre -1 y 1 representando el coeficiente de correlación de Pearson. De acuerdo al gráfico se observa que la variable ingreso_neto se encuentra correlacionada positivamente con la mayoría de las magnitudes, pero presenta muy poco grado de asociación con la magnitud *resultado_neto.*

Entonces, la variación del volumen de ingreso neto en la producción bovina podría estar asociada a las variables descriptas por el componente Dim_1, cuyos coeficientes de correlación son positivos.

4.4. Método 2. Análisis factorial (AF).

En esta sección se aplica el método de análisis factorial utilizando la misma base de datos, pero considerando además la incidencia de variables no observables o variables latentes. Con el fin de determinar si los datos se distribuyen normalmente, es decir, si la varianza entre la variable observable y los factores presentan un comportamiento homocedástico, según lo desarrollado en el Capítulo 1, Sección 1.2.2. se efectúa el test de normalidad de Shapiro-Wilk, cuyos resultados obtenidos son:

Variable observable	Test de normalidad de Shapiro-Wilk
Margen_bruto ($/ha)	W = 0,87159
Resultado_neto ($/ha)	W = 0,92206
Ingreso_neto ($/ha)	W = 0,81765
Gastos_directos ($/ha)	W = 0,59907
Costos_indirectos ($/ha)	W = 0,8164
Eficiencia_stock (%)	W = 0,83143
Producción (kg/ha)	W = 0,73285
Carga (kg/ha)	W = 0,87358

Tabla 4.1.10. Test de normalidad.

En la Tabla 4.1.10. se observa que el coeficiente de Shapiro-Wilk (W) es superior a 0,70 para la mayoría de las variables, por lo tanto, es similar a una distribución normal, es decir, los residuos presentan un comportamiento homocedástico, mientras que la magnitud gastos_directos presenta un coeficiente medio, cuyo valor es menor a 0,60 del cual se puede inferir que su distribución se encuentra un poco sesgada.

Luego, al igual que en el método de análisis de componentes principales, se evalúa la adecuación muestral utilizando la *prueba de esfericidad de Bartlett* y el coeficiente de *Kaiser-Meyer-Olkin* (KMO). De acuerdo a los resultados obtenidos en la Sección 4.3.1. del Capítulo 4, se observa que la adecuación de los datos a la muestra es satisfactoria.

A continuación, se realiza un análisis de fiabilidad de los datos para determinar el nivel de confianza de los valores analizados, utilizando los coeficientes: *Alfa de Crombach* y *Lambda de Guttman.* Los resultados obtenidos son: 95% y 99% respectivamente, lo cual comprueba un alto grado de confianza en los datos utilizados.

4.4.1. Obtención de factores.

Luego de analizar el nivel de adecuación muestral y la confiabilidad de los datos, se calcula el *estimador de saturación de factores Omega*, con el fin de obtener la cantidad de factores que serán utilizados en el análisis factorial, como se muestra en la siguiente tabla:

Estimador	g	F1	F2	F3	F4	F5	F6	F7	F8
Factor Omega	1,00	0,94	0,89	1,01	0,99	0,97	NA	NA	NA

Tabla 4.1.11. Estimador de saturación de factores.

De acuerdo al estimador Omega presente en la Tabla 4.1.11., para efectuar el análisis factorial se puede tomar hasta 5 factores, siendo F_1, F_2, F_3, F_4 y F_5; la cantidad de factores representativos de las variables observables, cuyos coeficientes varían entre 0,89 y 1,01. Dado que los coeficientes son superiores a 0,70 muestran un alto grado de aceptación, según la Sección 3.5.2 del Capítulo 3. Por otra parte los factores F_6, F_7 y F_8, presentan valores muy bajos tendientes a cero los cuales son denotados como NA, por lo tanto son excluidos del análisis factorial. Para visualizar la cantidad de factores en función de los valores propios, se presenta a continuación, un gráfico de sedimentación:

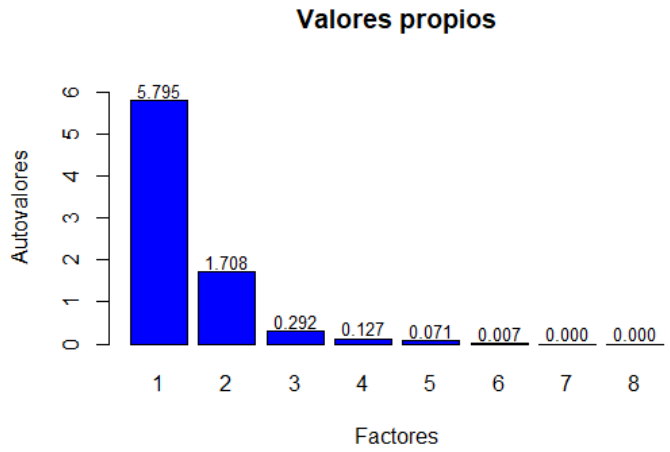


Figura 4.1.10. Gráfico de sedimentación de factores.

En la Figura 4.1.10. se observa que la proporción de varianza explicada por cada factor desciende al incrementarse la cantidad de factores. Por otro lado, las barras que representan los factores 6,7 y 8 tienden a cero. Por lo tanto, gráficamente, el número de factores que captura mayor varianza es cinco. Además del análisis factorial confirmatorio, se realiza un análisis factorial exploratorio, con el fin de calcular las cargas entre los factores y los parámetros según la siguiente tabla:

Variables	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	h_2	u_2
Margen_bruto ($/ha)	0,80	0,59	-0,08						0,99	0,0023
Resultado_neto($/ha)	0,22	0,97	0,09						0,99	0,0035
Ingreso_neto ($/ha)	1,00			-0,07					1,00	0,0018
Gastos_directos($/ha)	0,87	-0,45	0,08	-0,13					0,98	0.0037
Costos_indirectos($/ha)	0,95	-0,18	-0,22	0,07					0,98	0.0038
Eficiencia_stock (%)	0,78	-0,37	0,32	0,11	0,28				0,94	0,0063
Producción (kg/ha)	0,90	-0,28	0,28	0,16					0,99	0,0087
Carga (kg/ha)	0,93	-0,07	0,22	0,23	-0,11				0,98	0,0191

Tabla 4.1.12. Cargas estandarizadas.

La Tabla 4.1.12. muestra una matriz de cargas estandarizadas obtenidas a partir de la matriz de correlación, la cual representa la relación entre las variables observables y la carga de los factores $F_1, \dots, F_8$. En particular, las magnitudes con mayor comunalidad son: margen_bruto, resultado_neto, ingreso_neto, gastos_directos, costos_indirectos, producción y carga. Mientras que el parámetro con menor comunalidad es la eficiencia_stock. Otro indicador utilizado para medir la varianza que no ha podido ser explicada por los factores es la unicidad (u_2). En base a la Tabla 4.1.12. se observa que el ingreso_neto presenta un valor alto de unicidad dado que su valor es próximo a cero, al igual que en las variables: margen_bruto, resultado_neto, gastos_directos y costos_indirectos, lo cual comprueba que el nivel de varianza no explicada por los factores es alta.

4.4.2. Representación gráfica de los factores.

Con el objetivo de poder interpretar la relación entre el ingreso_neto y los demás parámetros en relación a los factores obtenidos a partir de la matriz de carga, se presentan a continuación tres escenarios. En el primero, se muestra la matriz de puntuaciones sin rotar. Luego, se efectúa una rotación ortogonal con el fin de ajustar las variables al nuevo eje de coordenadas y de esta manera identificar que variables presentan un comportamiento similar al Factor 1 o al Factor 2. Por último, se realiza una rotación oblicua con el fin de comparar que método describe mejor el comportamiento entre los factores y las magnitudes observables. A continuación se representa el primer supuesto:

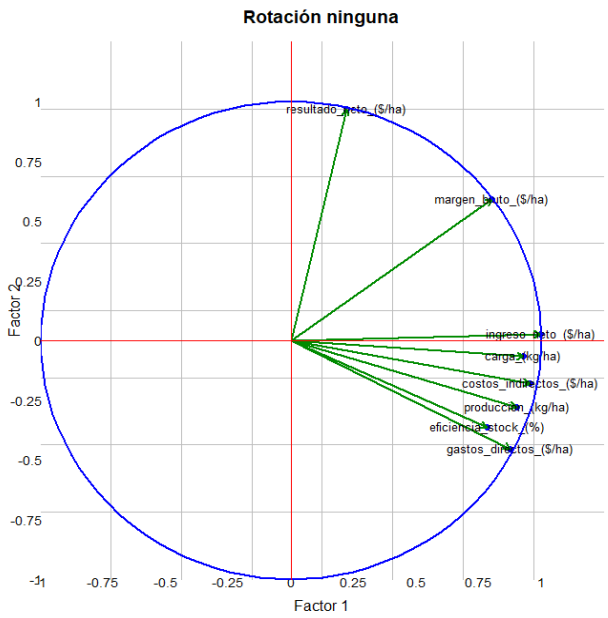


Figura 4.1.11. Matriz de carga sin rotación de factores.

En la Figura 4.1.11, se muestran las cargas de cada variable sin sufrir rotación, los cuales están representadas por vectores indicando: la dirección, el sentido y la magnitud de cada variable. En general, las magnitudes se encuentran cercanas entre si mostrando un grado de correlación positivo con respecto al ingreso_neto. Por otra parte, el resultado_ neto y el margen_bruto se encuentran muy poco correlacionado con el ingreso_ neto, dado que los vectores se hallan más distantes. Si bien, la mayoría de los parámetros están mejor explicados por el Factor 1, el resultado neto está representado por el Factor 2.

Por lo tanto, el comportamiento entre el ingreso_neto y la mayoría de las variables puede ser explicada a través del Factor 1. Luego, con el propósito reducir la pérdida de varianza entre las variables originales y los factores, se realiza la siguiente rotación ortogonal ajustando los vectores a los ejes de coordenadas:

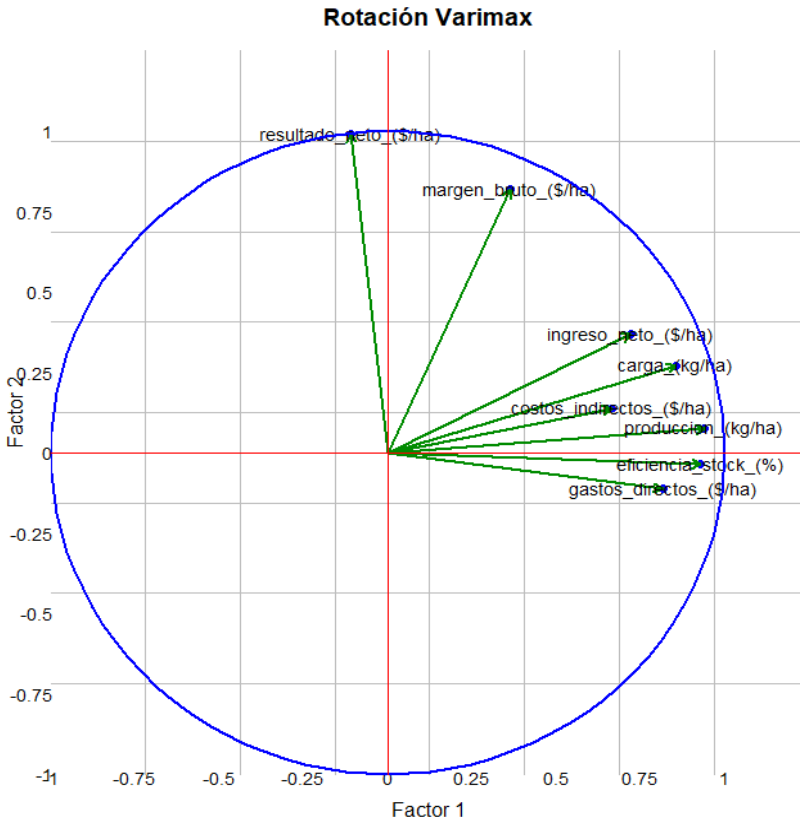


Figura 4.1.12. Matriz de carga con rotación ortogonal.

De acuerdo a la Figura 4.1.12. se observa que las puntuaciones provenientes del ingreso_neto y de la mayoría de las variables, son mejor interpretadas por el Factor 1, mostrando un nivel de correlación positivo fuerte, mientras que el comportamiento de los parámetros: margen_bruto y resultado_neto, es atribuible al Factor 2 representando una correlación negativa y débil con respecto a la variable ingreso_neto.

Por último, se realiza una rotación oblicua con el fin de determinar si los vectores se encuentran más próximos a los ejes, como se muestra en el gráfico:

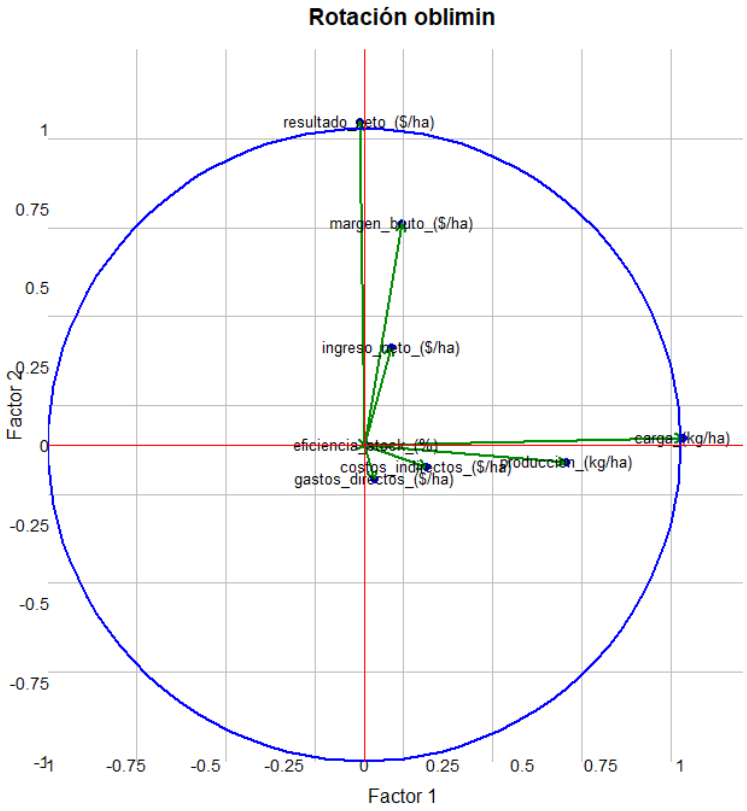


Figura 4.1.13. Matriz de carga con rotación oblicua.

En la Figura 4.1.13. al efectuar una rotación oblicua se observa que, las magnitudes resultado_neto, ingreso_neto y margen_bruto están mejor alineadas al Factor 2, las cuales presentan correlación positiva. Por otra parte, las variables: carga y producción son mejor explicadas por el Factor 1, cuyo grado de correlación con respecto al ingreso_neto es positivo pero muy débil.

En vista a los gráficos utilizados en el análisis factorial exploratorio, se concluye que la rotación oblicua es la que mejor interpreta el comportamiento de las variables, dado que las puntuaciones se encuentran más próximas a los ejes de coordenadas disminuyendo así la pérdida de información entre los parámetros y los factores. Por lo tanto, de acuerdo a la Figura 4.1.13. se observa que el ingreso_neto se ve fuertemente influenciado por el margen_bruto y el resultado_neto, cuya correlación es fuerte y positiva, mientras que el volumen de ingreso está menor afectado por el comportamiento de las variables: gastos _directos,costos_indirectos, eficiencia_stock, producción y carga..

Si bien tanto el método de componentes principales como el de análisis factorial permiten reducir la dimensionalidad de las variables, se observa que el grado de correlación entre el nivel de ventas representado por la variable ingreso neto con el resto de las variables es positiva.

4.5. Discusión de los resultados.

Al analizar el efecto de las variables que repercuten sobre el volumen de ventas de la actividad ganadera en diferentes provincias de la Argentina, se obtuvieron los siguientes resultados: el ingreso_neto representa la magnitud con mayor varianza y con mayor rango por hectárea, mientras que la producción, es la variable con menor varianza y rango, según la Tabla 4.1.1. Por otro lado, de acuerdo al Gráfico 4.1.1. se observa que: Entre Ríos, Buenos Aires y Santa Fe, presentan mayor nivel de ingreso neto por ventas, mientras que, Chubut, Rio Negro y Salta, son las provincias con menor ingreso neto por hectárea. Luego, con el fin de determinar que regiones presentan mayor gastos_directos en relación a la producción bovina, se tiene que: Entre Ríos, Santa Fe y La Pampa son las más representativas. En tanto, Chubut, Rio Negro y Salta tienen el menor gasto directo de producción bovina por hectárea según la Figura 4.1.2.

Después, al analizar el histograma de la Figura 4.1.3. se puede notar que la distribución de los valores no es normal, sino que se encuentran sesgados hacia la derecha.

Según la Tabla 4.1.3. correspondiente a la matriz de varianza y covarianza calculada en el análisis de componentes principales se observa que las covarianzas entre el ingreso_neto y el resto de las variables son positivas, mientras que el resultado_neto y los gastos_directos; la eficiencia_stock y la producción, presentan covarianzas negativas mostrando un nivel de correlación inversa. Luego, al estandarizar los valores del set de datos, se tiene que el coeficiente de correlación de Pearson entre las variables y el ingreso_neto es positivo fuerte. En tanto, el grado de asociación entre el resultado_neto y el ingreso_neto es positivo débil, de acuerdo a la Tabla 4.1.4. Estas últimas observaciones se afirman mediante el correlograma de la Figura 4.1.4., donde las esferas azules de color intenso muestran correlación positiva fuerte, mientras que las de menor intensidad representan correlación débil.

Posteriomente, para determinar cómo se distribuyen los datos entre el ingreso_neto y las demás magnitures, de acuerdo a la Figura 4.1.5., se puede notar que las nubes de puntos están cercanas a la recta de regresión, con excepción de los gráficos pertenecientes al

ingreso_neto y resultado_neto e ingreso_neto y eficiencia_Stock, donde los datos se encuentran más dispersos.

Al efectuar las pruebas de adecuación muestral, tanto el test de esfericidad de Bartlett como la prueba de Kaiser-Meyer-Olkin (KMO) verifican que la adecuación de la muestra es buena, por lo tanto los datos pueden ser utilizados para el análisis de componentes principales. En particular, la medida de adecuación muestral del ingreso_neto presenta un valor medio alto en relación a los demás parámetros, de acuerdo a la Tabla 4.1.5.

Luego al obtener los componentes principales se observa que las dos primeras componentes, denotadas por Dim_1 y Dim_2 , son las que capturan mayor proporción de varianza, según la Tabla 4.1.7. y la Figura 4.1.6. Posteriormente, al sumar los autovalores de la matriz de correlación, se obtiene la traza de la matriz, cuyo resultado representa la cantidad total de variables utilizadas, la cual es este caso es igual a 8. Entonces, al obtener la proporción de varianza representada por cada componente principal, a partir de la Tabla 4.1.8 se tiene que: la primera componente captura 72,43% de varianza, mientras que la segunda componente 21,34 %, cuyo porcentaje de varianza acumulada es de 93,78%. Por lo tanto, las dos primeras componentes principales son las que mejor explican el comportamiento de las variables observables, dado que reducen la pérdida de información entre las magnitudes y los componentes.

Luego, al representar gráficamente la relación entre las variables y los componentes principales Dim_1 y Dim_2, de acuerdo a la Figura 4.1.8, se observa que *resultado_ neto* está representado por el componente Dim_2, cuya proporción de varianza es 21,3%, mientras que: el ingreso_neto ,la carga, los costos_indirectos, la producción, la eficiencia_stock, los gastos_directos y el margen_bruto son explicados por la primer componente principal Dim_1 , donde la proporción de varianza es mayor correspondiente a 72,4%.

A partir de una representación vectorial de las cargas, según la Figura 4.1.9. se tiene que el ingreso_neto correlaciona positivamente con la mayoría de los parámetros, pero con menor intensidad con el resultado_neto. De lo cual se concluye que el ingreso neto de la producción bovina en la República Argentina estaría relacionada al comportamiento de las siguientes variables: carga, costos_indirectos, producción, eficiencia_stock, gastos_directos y margen_bruto, cuyo nivel de correlación es positivo fuerte.

Para contrastar los resultados obtenidos en el análisis de componentes principales se realiza un análisis factorial. Al aplicar el test de normalidad de Shapiro-wilk, de acuerdo a la Tabla 4.10., se tiene que las variables se distribuyen normalmente, es decir, los residuos presentan un comportamiento homocedástico, dado que el coeficiente de Shapiro-Wilk es superior a 0,70, con excepción de los gastos_directos cuyo coeficiente es medio, correspondiente a 0,60.

De la misma manera que en el análisis de componentes principales, se observa que la adecuación de los datos a la muestra es buena, según la prueba de esfericidad de Bartlett y del coeficiente Kaiser-Meyer-Olkin. Por otra parte, se efectúan análisis de confianza de los datos utilizando los coeficientes Alfa de Crombach y Lambda de Guttman, cuyos resultados son 95% y 99 % correlativamente, lo cual indica que los datos utilizados presentan un alto grado de confianza. Luego, se aplica el estimador Omega para determinar la cantidad de factores a utilizar en el análisis factorial, del cual se constata que la cantidad de factores que capturan mayor varianza es igual a cinco, correspondientes a: F_1, F_2, F_3, F_4 y F_5, como lo muestra la Figura 4.1.10. Posteriormente, se realiza un análisis factorial exploratorio, para determinar las cargas entre las variables observables y los factores y se determinan la comunalidad y la unicidad. Según la Tabla 4.1.12., los parámetros que presentan mayor comunalidad son: margen_bruto, resultado_neto, ingreso_neto, gastos_directos y costos_indirectos, lo cual indica que la proporción de varianza explicada por los factores es elevada. Por otro lado, las magnitudes con menor comunalidad son: eficiencia_stock, producción y carga. Además, las variables con valores altos de unicidad son: ingreso_neto, el margen_bruto, el resultado_neto, los gastos_directos y los costos_indirectos, lo cual comprueba que el grado de varianza no explicada por los factores es alta.

Finalmente, para poder interpretar la relación entre el ingreso_neto y las demás magnitudes se comparan tres instancias. En la primera etapa, según el Grafico 4.1.11., donde los factores no son rotados, se observa que las variables presentan correlación positiva con el ingreso_neto dado que los vectores se encuentran próximos entre si y se dirigen en la misma dirección. No obstante, el ingreso_neto y el margen_bruto poseen poco grado de asociación, ya que los vectores están más alejados. Luego para ajustar los vectores con los ejes de coordenadas, se efectúa una rotación ortogonal, donde se observa que el ingreso_neto y gran parte de las variables, son mejor interpretadas por el Factor 1, cuyo coeficiente de correlación es positivo fuerte, mientras que el margen_bruto y el

resultado_neto son mejor representados por el Factor 2, cuya corelación es negativa y débil con respecto a la variable ingreso_neto. Por último, al realizar una rotación oblicua se concluye que: el resultado_neto, el ingreso_neto y el margen_bruto presentan correlación positiva y corresponden al Factor 2, mientras que la carga y producción, tienen correlación positiva débil y están asociadas al Factor 1, cuyo porcentaje de varianza es menor. De acuerdo a las dos últimas rotaciones de factores realizadas, se observa que la rotación oblicua es la que mejor describe el comportamiento de los parámetros dado que los vectores están cercanos a los ejes, donde se representan los factores con mayor varianza acumulada. Entonces, según la Figura 4.1.13., se tiene que el ingreso_neto está fuertemente relacionado con el margen_bruto y el resultado_neto, siendo su correlación positiva fuerte. En tanto que, el ingreso_neto y las siguientes magnitudes: gastos _directos, costos_indirectos, eficiencia_stock, producción y carga, tienen un nivel de correlación menor.

Conclusión general

De acuerdo a los temas abordados en esta tesis, tanto el análisis de componentes principales como el análisis factorial pueden ser utilizados para describir el comportamiento de un conjunto de variables mediante una reducción de sus dimensiones ya sea, a través de componentes principales o de factores. Al utilizar ambos métodos se observa que el ingreso_neto presenta correlación positiva con la mayoría de los parámetros, es decir, al incrementar el valor de las variables descriptas en este caso de estudio, aumenta el valor del volumen de ventas.

Finalmente, se observa que, el análisis de componentes principales es el método que mejor describe la relación entre el ingreso_neto y las demás magnitudes, dado que presenta mayor proporción de varianza explicada correspondiente a 72,4% en comparación con el análisis factorial, cuyo resultado es 21,35%. Es decir, el ingreso_neto de la producción de carne bovina en la República Argentina se incrementa al aumentar el valor de los siguientes parámetros: carga, costos_indirectos, producción, ingreso_neto, eficiencia_stock, gastos_directos y margen_bruto.

Bibliografía

[1] Blomhøj, M. (2008). ***Modelización matemática - Una teoría para la práctica*.** Vol.23, No.2, pp.20-35. https://revistas.unc.edu.ar/index.php/REM/article/view/10419

[2] Bocco, M. (2010). ***Funciones elementales para construir modelos matemáticos*.** Buenos Aires: Ministerio de Educación de la Nación. Instituto Nacional de Educación Tecnológica. Cáp.1-2. https://www.educ.ar/sitios/educar/resources/151449/funciones-elementales-para-construir-modelos-matematicos-parte-1/download

[3] Brito Vallina, M.L. Et al. (2011). ***Papel de la modelación matemática en la formación de los ingenieros***. Ingeniería mecánica. Vol.14, No.2, pp.129-139. http://scielo.sld.cu/pdf/im/v14n2/im05211.pdf

[4] Romo-Vázquez, A. (2014). ***La modelización matemática en la formación de ingenieros.*** Educación Matemática. Vol.1, pp.314-338. https://www.redalyc.org/pdf/405/40540854016.pdf

[5] Chow J.W. and Knudson D. V. (2011). ***Use of deterministic models in sports and exercise biomechanics research.*** Sports Biomechanics .Vol.10, No.3, pp 219-233. https://doi.org/10.1080/14763141.2011.592212

[6] Todorov A. Et al. (2002). ***The persuasion handbook: developments in theory and practice*.** Cap. 11, pp. 195-212. https://dx.doi.org/10.4135/9781412976046.n11

[7] Macciota, N.P.P. Et al. (2005). ***Detection of different shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models***. Journal of diary science. Vol.88, No. 3, pp. 1178-1191. https://doi.org/10.3168/jds.s0022-0302(05)72784-3

[8] Cai, Z., Hong, H., and Wang, S. (2018). ***Econometric modeling and economic forecasting***. Journal of management science and engineering, Vol.3,No.4, pp.179-182. https://doi.org/10.3724/SP.J.1383.304010

[9] McFadden, D. L. (1994).***Econometric analysis of qualitative response models***. Handbook of econometrics. Vol.2, Cap. 24, pp. 1355-1457 https://doi.org/10.1016/S1573-4412(84)02016-X

[10] Smith, W.R. (1983). ***Qualitative mathematical models of endocrine systems***. American journal of physiology. Vol. 245, No. 4, pp. R473-R477. https://doi.org/10.1152/ajpregu.1983.245.4.r473

[11] Yao, D. D. Et al. (2012). ***Stochastic modeling and analysis of manufacturing systems***. Springer science and business media. https://books.google.com.ar/books?hl=es&lr=&id=OiXUBwAAQBAJ&oi=fnd&pg=PR14&dq=Stochastic+models+of+manufacturing+systems&ots=UEf38rrzyO&sig=TSAj-dbEPB6FjFk8PPZipVajKCg&redir_esc=y#v=onepage&q=Stochastic%20models%20of%20manufacturing%20systems&f=false

[12] Yoram, R. (2003). ***Applied stochastic hydrogeology***. Oxford Academic. https://doi.org/10.1093/oso/9780195138047.001.0001

[13] Farmer, W. H. and Vogel, R.M. (2016). ***On the deterministic and stochastic use of hydrologic models.*** Water Resources Research. Vol.52, No.7, pp.5619-5633 https://doi.org/10.1002/2016WR019129

[14] Lahrouz, A. Et al. (2011). ***Deterministic and stochastic stability of a mathematical model of smoking***. Statistics and probability letters. Vol.81, No.8, pp. 1276-1284. https://doi.org/10.1016/j.spl.2011.03.029

[15] Majda, A. (1981). ***A qualitative model for dynamic combustion***. SIAM Jounal on Applied Mathematics. Vol.41, No.1 . https://doi.org/10.1137/0141006

[16] Sargent, R.G. (2013) ***Verification and validation of simulation models***. Journal of Simulation. Vol.7, No.1, pp.12-24. https://doi.org/10.1057/jos.2012.20

[17] Calafiore, G. C. and El Ghaoui, L. (2014). ***Optimization models***. https://openlibrary.org/works/OL21076406W/Optimization_Models

[18] Ziemba, W. T. and Vickson, R. G. (Eds.). (2006). ***Stochastic optimization models in finance***. World scientific handbook in financial economics series. Vol.1. https://doi.org/10.1142/6101

[19] Holland, J. H. (1992). ***Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence***. https://doi.org/10.7551/mitpress/1090.001.0001

[20] Miroslav, F. and Mikleš J. (2007). ***Process modelling, identification and control***. . https://doi.org/10.1007/978-3-540-71970-0

[21] Sontag, E. D. (2013). ***Mathematical control theory: deterministic finite dimensional systems***.Vol.2 Springer Science and Business. https://books.google.com.ar/books?hl=es&lr=&id=f9XiBwAAQBAJ&oi=fnd&pg=PR5&dq=Mathematical+control+theory:+deterministic+finite+dimensional+sytems&ots=bLuVQep0LX&sig=EnKABEHFk-TAL-jwjtFOYhB7HiQ&redir_esc=y#v=onepage&q=Mathematical%20control%20theory%3A%20deterministic%20finite%20dimensional%20sytems&f=false

[22] Kaznessis, Y. N. (2011). ***Mathematical models in biology: from molecules to life*.** Vol.3, No.1, pp. 314-322. https://doi.org/10.1002/wsbm.142

[23] Fowler, A. C. (1997). ***Mathematical models in the applied sciences***. Cambridge University Press. https://books.google.com.ar/books?hl=es&lr=&id=2KeYPU78AsMC&oi=fnd&pg=PR7&dq=Mathematical+models+&ots=lTWNauEWaJ&sig=PZVjgV9TfP0Nke62IH_oFycc7tE&redir_esc=y#v=onepage&q&f=false

[24] Astorga Gómez, J. M. (2014). ***Aplicación de modelos de regresión lineal para determinar las armónicas de tensión y corriente***. Ingeniería Energética. Vol.35, No.3, pp.234-241. https://www.redalyc.org/pdf/3291/329132445008.pdf

[25] Ramajo Hernández, J. (1994). ***Estimación de respuestas renta de los consumidores: una aplicación del modelo de regresión lineal discontinua a tramos***. Estudios de Economía Aplicada. Vol.1, No.1, pp.61-86. https://dialnet.unirioja.es/servlet/articulo?codigo=175950

[26] Rodríguez, J. (2011). ***Aplicación de un modelo de crecimiento en biología***. Revista de Ciencias. Vol.3. https://doi.org/10.25100/rc.v3i0.555

[27] Triola, Mario. (2009). ***Estadística***. Pearson Educación. Cap.10, 11, 12.

[28] Spiegel, M. R. (1970). ***Estadística.*** Mc Graw-Hill. Cap.10, 13, 14,15.

[29] Lind, D. A. Et al. (2012). ***Estadística aplicada a los negocios y a la economía***. Mc Graw-Hill. Cap.12, 13, 14.

[30] Dagnino, J. S. (2014). ***Regresión lineal.*** Revista chilena de anestesia. Vol.43, Nro.2. pp.143-149. https://doi.org/10.25237/revchilanestv43n02.14

[31] Chan, D. Et al. (2019). ***Análisis inteligente de datos con lenguaje R: con aplicaciones e imágenes.*** Ciudad de Buenos Aires.UTN-edUTecNe. https://ria.utn.edu.ar/xmlui/handle/20.500.12272/4371?show=full

[32] Rigalli, A. Et al. (2019***). Uso de herramientas informáticas para la recopilación, análisis e interpretación de datos de interés en las ciencias biomédicas: estadística básica con R***. Facultad de Ciencias médicas. Universidad Nacional de Rosario. No.1, pp 76-105. https://rephip.unr.edu.ar/bitstream/handle/2133/15296/libroRmodulo3.pdf?sequence=3&isAllowed=y

[33] Reyes Martin, G. (2012) ***La función de probabilidad normal: Características y aplicaciones***. No.6, pp.107-110. https://dialnet.unirioja.es/servlet/articulo?codigo=5582675

[34] Brunetzarza, K. (2019). ***Distribución normal y algunas aplicaciones***. Universidad Autónoma del Estado de México. http://ri.uaemex.mx/handle/20.500.11799/106113

[35] Piol, R. (2014). ***Validación de la regresión mediante el análisis de homocedasticidad***. SOITAVE 260 / UPAV 94. pp. 1-28. https://es.scribd.com/document/408911767/Validacion-de-La-Regresion-Mediante-El-Analisis-de-Homocedasticidad

[36] Pértega Díaz, S. y Pita Fernández, S. (2001). ***Técnicas de regresión: regresión lineal simple***. Vol.7, No.2, pp.91-94. https://dialnet.unirioja.es/servlet/articulo?codigo=2331559

[37] Devore, J. L. (2018). ***Fundamentos de probabilidad y estadística***. Cengage. Vol.1, Cap. 7, 8, 9.

[38] Flores Tapia, C y Flores Cevallos, K.L. (2021). ***Pruebas para comprobar a normalidad de datos en procesos productivos: Anderson-darling, Ryan-Joiner, Shapiro-Wilk y Kolmogórov -Smirnov***. Vol.23, No.2. , pp. 1-15. http://portal.amelica.org/ameli/jatsRepo/341/3412237018/index.html

[39] Hanusz, Z. Et al. (2016). ***Shapiro–Wilk Test with known mean***. REVSTAT-statistical journal. Vol.14, No.1, pp. 89–100. https://doi.org/10.57805/revstat.v14i1.180

[40] Szretter Noste, M. E. (2017). ***Apunte de regresión lineal***. Carrera de Especialización en Estadística para Ciencias de la Salud Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. http://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf

[41] Dos Santos Freitas, M.M. Et al. (2022). ***KNN algorithm and multivariate analysis to select and classify starch films.*** Science direct. Vol. 34. https://doi.org/10.1016/j.fpsl.2022.100976.

[42] Montero Granados, R. (2016). ***Modelo de regresión lineal múltiple***. Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España. pp.1-61. https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf

[43] Ordaz Sanz, J. A. Et al. (2010). ***Introducción a las técnicas de análisis multivariante en el ámbito de la economía y la empresa.*** Métodos Estadísticos y Econométricos en la Empresa y para Finanzas. https://libros.metabiblioteca.org/jspui/bitstream/001/362/5/978-84-694-7251-4.pdf

[44] Da Costa Pereira, N. Et al. (2006). ***La matriz de correlación***: ***una dicotomía entre soporte estadístico y herramienta agenciada.*** I Encuentro Latinoamericano de metodología en Ciencias Sociales. Mesa E-2.- Estado actual de los métodos/técnicas cuantitativas y cualitativas y de la triangulación metodológica. Ponencia: Universidad Nacional de Lujan. Buenos Aires, Argentina. http://sedici.unlp.edu.ar/handle/10915/109909

[45] Montgomery, D. C. Et al. (2021). ***Introduction to linear regression analysis*****.** Vol.6. https://books.google.com.ar/books?hl=es&lr=&id=tCIgEAAAQBAJ&oi=fnd&pg=PR13&dq=linear+regression&ots=lfxfYqd_Jt&sig=RZizcWArdsjb5noFNsoTF1xvj4k&redir_esc=y#v=onepage&q=linear%20regression&f=false

[46] Garibaldi, L. A. Et al. (2019). ***Modelos estadísticos en lenguaje R.*** Editorial Universidad Nacional de Rio Negro. http://rid.unrn.edu.ar/handle/20.500.12049/5789

[47] Raykov, T. and Marcoulides, G. A. (2008). ***An Introduction to applied multivariate analysis***. Routledge. Vol.1. https://doi.org/10.4324/9780203809532

[48] Timm, N. H. (2002). ***Applied multivariate analysis***. Cap.3. https://doi.org/10.1007/978-0-387-22771-9_4

[49] Rencher, A. C. (2002). ***Methods of multivariate analysis***.Vol.2. https://epdf.tips/methods-of-multivariate-analysisa2631e582da20c7b5ec9939d8ea6b0bb19356.html

[50] Blanca Mena, M. J. (2004). ***Alternativas de análisis estadístico en los diseños de medidas repetidas***. Psicothema. Vol.16, No.3, pp.509-518.. https://reunido.uniovi.es/index.php/PST/article/view/8229

[51] Warne, R. (2014). ***A primer on multivariate analysis of variance (MANOVA) for behavioral scientists.*** Vol. 19. Practical Assessment, Research, and Evaluation. https://doi.org/10.7275/sm63-7h70

[52] Bray, J. H., Et al. (1985). ***Multivariate analysis of variance***. No. 54. https://books.google.com.ar/books?hl=es&lr=&id=QGsKyIixFjAC&oi=fnd&pg=IA1&dq=multivariate+analysis+of+variance+(MANOVA)+&ots=97xCialGEh&sig=g1DYA_qtq_SPUKxHjpKggFgMDH8&redir_esc=y#v=onepage&q=multivariate%20analysis%20of%20variance%20(MANOVA)&f=false

.

[53] Mora Catalá, R. y Rodríguez-Jaume, M. J. (2001). ***Estadística informática: casos y ejemplos con el SPSS.*** Cap. 4. https://rua.ua.es/dspace/handle/10045/8143

[54] Rodríguez Gómez, G. (1992). ***El análisis multivariante de la varianza (MANOVA): claves para su interpretación***. Revista investigación educativa.Vol.10, No.19, pp.69-79. https://digitum.um.es/digitum/handle/10201/94959

[55] Fang, K. T., Et al. (2018). ***Symmetric multivariate and related distributions***. Chapman and Hall/CRC. Vol.1. https://doi.org/10.1201/9781351077040

[56] Haase, R. F. and Ellis, M. V. (1987). ***Multivariate analysis of variance***. Journal of counseling psychology. Vol. 34, No.4, pp.404–413. https://doi.org/10.1037/0022-0167.34.4.404

[57] Muller, K. E. (2012). ***A new F approximation for the Pillai—Bartlett trace under H 0.*** Journal of computational and graphical statistics. Vol.7, No.1, pp.131-137. https://doi.org/10.1080/10618600.1998.10474765

[58] Serlin, R. C. (1982). ***A multivariate measure of association based on the Pillai-Bartlett procedure.*** Psychological bulletin. Vol. 91, No. 2, pp. 413–417. https://doi.org/10.1037/0033-2909.91.2.413

[59] Zwick, R. (1985). ***Nonparametric one-way multivariate analysis of variance: A computational approach based on the Pillai-Bartlett trace***. Psychological bulletin. Vol. 97, No.1, pp. 148–152. https://doi.org/10.1037/0033-2909.97.1.148

[60] Vargas, J. J. Et al. (2020). ***Aplicación de la técnica multivariada Manova a dos variables de control provenientes de tres modelos de simulación estocásticos de un proceso productivo***. Entre ciencia e ingeniería. Vol. 14, No. 28, pp. 66-75. https://doi.org/10.31908/19098367.2056

[61] Niño, M. F. y Simonetti, E. F. (2005). ***El análisis de datos desde una perspectiva integradora. Una introducción al análisis multivariado: las Componentes Principales***. Facultad de Humanidades y Ciencias Sociales Universidad Nacional de Misiones Posadas Misiones.

[62] Hair, J. F. Et al. (1999). ***Análisis Multivariante.*** Vol.5. https://books.google.com.ar/books/about/An%C3%A1lisis_multivariante.html?id=QV4INQAACAAJ&redir_esc=y

[63] Palacio, F. Et al. (2020). ***Análisis Multivariados para datos biológicos. Teoría y su aplicación utilizando el lenguaje R***. Fundación de Historia Natural Félix de Azara. Departamento de Ciencias Naturales y Antropológicas. Universidad Maimónides. Buenos Aires. No.1, Cap.6. https://www.researchgate.net/publication/341446029_ANALISIS_MULTIVARIADO_PARA_DATOS_BIOLOGICOS_Teoria_y_su_aplicacion_utilizando_el_lenguaje_R

[64] Córdoba, M., B. Et al. (2012). ***Análisis de componentes principales con datos georreferenciados: Una aplicación en agricultura de precisión***. Revista de la Facultad de Ciencias Agrarias. Universidad Nacional de Cuyo. Vol.44, No.1, pp. 27-39. http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1853-86652012000100003&lng=es&nrm=iso

[65] Pizarro Romero, K. y Martínez Mora, O. (2020). ***Análisis factorial exploratorio mediante el uso de las medidas de adecuación muestral kmo y esfericidad de Bartlett para***

determinar factores principales. Journal of science and research. Vol.5, pp.903 - 924. https://revistas.utb.edu.ec/index.php/sr/article/view/1046

[66] Garmendia, M. L. (2007). ***Análisis factorial: una aplicación en el cuestionario de salud general de Goldberg, versión de 12 preguntas.*** Revista chilena de salud pública. Vol 11, No.2, pp. 57-65.https://revistaatemus.uchile.cl/index.php/RCSP/article/view/3095

[67] Ruiz, M. A. Et al. (2010). ***Modelo de ecuaciones estructurales.*** Papeles del psicólogo. Vol. 31, No. 1, pp. 34-45. https://www.redalyc.org/pdf/778/77812441004.pdf

[68] Peña, D. (2002). ***Análisis de datos multivariantes***. ResearchGate. https://www.researchgate.net/publication/40944325_Analisis_de_Datos_Multivariantes

[69] López González, E. (1998). ***Tratamiento de la colinealidad en regresión múltiple***. Psicothema. Vol. 10, No. 2, pp. 491-507. https://www.redalyc.org/pdf/727/72710221.pdf

[70] Ramirez, G. Et al. (2005). ***Detección gráfica de la multicolinealidad mediante el H-Plot de la inversa de la matriz de correlaciones***. Revista colombiana de estadística. Vol. 28, No.2, pp 207-219. www.redalyc.org/pdf/899/89928207.pdf

[71] Lozares Colina, C. y López-Roldán, P. (1991). ***El análisis de componentes principales: aplicación al análisis de datos secundarios.*** *Revista de sociologia*, No.37, pp. 31-63. https://ddd.uab.cat/record/49950?ln=es

[72] Ferrero, S.B. Et al. (2002). ***Análisis de componentes principales en teledetección. Consideraciones estadísticas para optimizar su interpretación***. Revista de teledetección. No.17, pp.43-54. http://www.aet.org.es/revistas/revista17/AET17-05.pdf

[73] Chávez Chong, C. O. Et al. (2015). ***Análisis de componentes principales funcionales en series de tiempo económicas (Analysis of principal functional components in economic time series).*** GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología. Vol.3, No.2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737561

[74] Pernice, S. A. (2020). ***Serie de machine learning: Análisis de Componentes Principales (PCA)*** .Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires. Serie Documentos de Trabajo, No. 770. https://www.econstor.eu/bitstream/10419/238395/1/770.pdf

[75] Watkins, M. W. (2018). ***Exploratory Factor Analysis***: ***A guide to best practice.*** Journal of Black Psychology. Vol. 44, No. 3, pp.219-246. https://doi.org/10.1177/0095798418771807

[76] Martínez Mora, O. and Pizarro Romero, K. (2020). ***Análisis factorial exploratorio mediante el uso de las medidas de adecuación muestral KMO y esfericidad de Bartlett para determinar factores principales.*** Journal of science and research.Vol.5, No.1, pp.1-22. https://doi.org/10.5281/zenodo.4453224

[77] Cudeck, R. (2000***). Handbook of Applied Multivariate Statistics and Mathematical Modeling.*** Science direct. Cap.10. https://doi.org/10.1016/B978-012691360-6/50011-2

[78] Bosten, J. M. Et al. (2017). ***An exploratory factor analysis of visual performance in a large population.*** Vision Research. Vol.141, pp.303-316. https://doi.org/10.1016/j.visres.2017.02.005

[79] Morata Ramírez, M. A. Et al. (2015). ***Confirmatory factor analysis. Recommendations for unweighted least squares method related to Chi-Square and RMSEA***. Acción psicológica. Vol.12, No.1, pp. 79-90. http://dx.doi.org/10.5944/ap.12.1.14362

[80] Ondé Pérez, D. (2020). ***Revisión del Concepto de Causalidad en el Marco del Análisis Factorial Confirmatorio***. Revista Iberoamericana de Diagnóstico y Evaluación. Vol.1, No.54, pp.103-117. https://doi.org/10.21865/ridep54.1.09

[81] Martínez Ávila, M. (2021). ***Análisis factorial confirmatorio: un modelo de gestión del conocimiento en la universidad pública.*** Revista Iberoamericana para la investigación y el desarrollo educativo. Vo.12, No.23, https://doi.org/10.23913/ride.v12i23.1103

[82] Flora, D. B. (2020). ***Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates***. *Advances in Methods and Practices in Psychological Science*. Vol.3, No.4, pp. 484–501. https://doi.org/10.1177/2515245920951747

[83] Green, S. and Yang, Y. (2009). ***Reliability of summed item scores using Structural Equation Modeling: An alternative to coefficient alpha.*** Psychometrika. Vol. 74, pp.155–167. https://doi.org/10.1007/s11336-008-9099-3

[84] Loken, E. and Gelman, A. (2017). ***Measurement error and the replication crisis: The assumption that measurement error always reduces effect sizes is false.*** Science. Vol. 355, No.6325, pp. 584–585. https://doi.org/10.1126/science.aal3618

[85] Ledesma, R. D. Et al. (2019). ***Uso del análisis factorial exploratorio en RIDEP. recomendaciones para autores y revisores.*** Revista Iberoamericana de Diagnóstico y Evaluación. Vol. 3, No. 52, pp. 173-180. https://doi.org/10.21865/RIDEP52.3.13

[86] Lloret-Segura, S. Et al. (2014). ***El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada***. Anales de psicología. Vol.30, No.3, pp. 1151-1169 http://dx.doi.org/10.6018/analesps.30.3.199361

[87] Campo-Arias, A. and Oviedo, H. C. (2005). ***Aproximación al uso del coeficiente alfa de Cronbach.*** Revista Colombiana de Psiquiatría. Vol. 34, No. 4, pp. 572-580. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0034-74502005000400009&lng=en&tlng=es

[88] Cervantes Botero, V. H. (2005). ***Interpretaciones del coeficiente alpha de Cronbach***. Avances en medición. No.3, pp.9-28. http://www.academia.edu/33218497/Interpretaciones_del_coeficiente_alpha_de_Cronbach

[89] Momirović, K. (1996).***An alternative to Guttman λ6: A measure of true lower bound to reliability of the first principal component***. Psihologija. No.1, pp.99-102. http://scindeks-clanci.ceon.rs/data/pdf/0048-5705/1996/0048-57059601099M.pdf

[90] Callender, J.C. and Osburn, H. G. (1979). ***An empirical comparison of coefficient alpha, Guttman's lambda - 2, and MSPLIT maximized split-half reliability estimates***. Journal of educational measurement. Vol.16, No.2, pp. 89–99. http://www.jstor.org/stable/1434452

[91] Costello, A. B. and Osborne, J. (2005). ***Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis.*** Practical Assessment, Research, and Evaluation. Vol. 10. https://doi.org/10.7275/jyj1-4868

[92] Ferrando, P. J. y Anguiano-Carrasco, C. (2010). ***El análisis factorial como técnica de investigación en psicología.*** Papeles del psicólogo. Vol.31, No.1, pp.18-33. https://www.redalyc.org/articulo.oa?id=77812441003

[93] Manly, B.F.J. and Navarro, A.J.A. (2016). Multivariate Statistical Methods. Chapman and Hall/CRC. https://doi.org/10.1201/9781315382135

[94] Guerra, G. (1992). ***Manual de administración de empresas agropecuarias***. Agroamérica. Vol.30. https://books.google.com.ar/books?hl=es&lr=&id=dlLYWCOylZUC&oi=fnd&pg=PR17&dq=gestion+de+empresas+agropecuarias&ots=VdJCl_xID3&sig=pOD16aiwTsW4H9N9fEGaoI9D88&redir_esc=y#v=onepage&q=gestion%20de%20empresas%20agropecuarias&f=false

[95] Nastasi, A. Et al. (2004) .***Claves para Costos***. Editorial La Ley. https://openlibrary.org/books/OL25632441M/Claves_para_costos

[96] Castaldo, A. O. (2003). ***Caracterización de los sistemas de producción bovina (invernada) en el nordeste de la provincia de La Pampa (Argentina): modelos de gestión***. Tesis presentada por D. Ariel Osvaldo Castaldo para optar al grado de Doctor en Veterinaria Año 2003. https://repo.unlpam.edu.ar/handle/unlpam/740

[97] Schildknecht, F. (2013). ***La tecnología de procesos como un medio para limitar el proceso de substitución de la producción bovina por la producción agrícola: estudio exploratorio en empresas agropecuarias del Grupo CREA Rauch-Udaquiola***. Trabajos de Licenciatura en Administración de Empresas. Universidad de San Andrés. Escuela de Administración y Negocios. https://repositorio.udesa.edu.ar/jspui/handle/10908/2590