CAPITULO 2

**2.MARCO TEORICO DE COMPONENTES PRINCIPALES Y ANALISIS DE REGRESION MULTIPLE**

**2.1 Análisis de Componentes Principales**

El Análisis de Componentes Principales trata de explicar la estructura de la varianza y covarianza del ajuste de las variables .

Estos Objetivos generales son:

1.-Reducción de datos

2.-Interpretación

Aunque p componentes son requeridas y producidas en la variabilidad del sistema total, frecuentemente mucha de está variabilidad puede ser contada por un pequeño número de k componentes principales . Si aún así, existe mucha información en la k componentes es aquí donde se utiliza p variables.

Las k componentes principales pueden entonces reemplazar en su inicio p-variables, y los datos originales, consisten de n medidas sobre p variables, esta reduce los datos consistentes de n medidas sobre k componentes principales.

Un análisis de componentes principales frecuentemente revela la relación de que no fueron previamente supuesto y con relación a eso permiten interpretaciones que no deberían resultar ordinarias.

El análisis de las componentes principales es más, que un medio para llegar al fin de algo, es un fin en ellos mismos, porque

frecuentemente sirven como intermediarios paso a paso en largas investigaciones.

El núcleo fundamental del Análisis de Componentes Principales (ACP), es el problema de la obtención de los vectores y valores propios (principales) de un operador vectorial, que en el campo del cálculo matricial se da bajo el problema de la diagonalización de una matriz cuadrada. Este problema algebraico, que inicialmente impulsó el desarrollo del Análisis Factorial en el estudio de la regresión lineal entre múltiples variables, se ha convertido, a lo largo de nuestro siglo, en uno de los instrumentos más extendidos en todas las ramas científicas. No sólo es una técnica de análisis empírico de la varianza, sino que puede jugar un papel decisivo en la formulación teórica.

**2.1.1 Componentes Principales de la Población**

Algebraicamente, las componentes principales son combinaciones lineales particulares de las p-variables aleatorios X1,X2,...,Xp. Geométricamente, estás combinaciones lineales representan un

nuevo sistema coordinado obtenido por la rotación de el sistema original con X1,X2,...,Xp en los ejes coordenados.

Los nuevos ejes representan las direcciones con la variabilidad máxima y provista de una simple y parsimoniosa descripción de la estructura de la covarianza.

Como deberíamos ver, los componentes principales dependen solamente de la matriz de covarianza  (o la matriz de correlación ) de X1,X2,...,Xp. Su desarrollo no requiere que se asuma una normal multivariada. Otra forma de transmitir, componentes principales derivadas de poblaciones normales multivariadas tienen interpretaciones útiles en términos de la constante de densidad elipsoidal. Más allá se pueden hacer inferencias sobre componentes muestreados cuando la población es una normal multivariada.

Introduciendo el vector aleatorio X'=[ X1,X2,...,Xp] se tiene la matriz de covarianza con los valores propios .

Considere las combinaciones lineales



Entonces, usando Z=CX tenemos:



Nosotros Obtenemos



Las componentes principales son combinaciones lineales incorrectas Y1,Y2,...,Yp cuya varianza en (2.2) son las más largas posibles.

La primera componente principal es la combinación lineal con máxima varianza. Esto es maximizando . Está claro que  puede incrementarse multiplicándolo por algún a1 que es una constante. Eliminando está indeterminación, es conveniente que se restrinja la atención al coeficiente de vectores de la unidad de longitud.

Por eso se define de la siguiente manera:

**Primera Componente Principal** = Combinación Lineal a1'X es aquel que maximiza la **Var(a1'X)** sujeta a **a1' a1=1**

**Segunda Componente Principal** = Combinación Lineal a1'X es aquel que maximiza la **Var(a2'X)** sujeta a **a2' a2=1** y **Cov(a1'X, a2'X)=0**

La etapa de i se define de la siguiente manera:

i Componente Principal = Combinación Lineal **ai'X** es aquel que maximiza la **Var(ai'X)** sujeta a **ai' ai=1** y **Cov(ai'X, ak'X)=0** para k<i.

**2.1.2 Matriz de Covarianza asociada con Vectores Aleatorios**

Después que  matriz de covarianza está asociada con el vector aleatorio X'=[ X1,X2,...,Xp] . La  tiene pares de

vectores y valores propios como  donde  .

Entonces las i componentes principales está dada por:



Con está seleccionamos



Si algunos de los  son iguales, se selecciona los correspondientes coeficientes de vectores ei, y Yi, no son únicos.

Nosotros conocemos que :



Con B=, esto es que:



Pero e1’e1=1 iniciando con vectores propios que son normalizados. De está manera :



## Similarmente usando



Obtenemos lo siguiente:



Para la selección de a=ek+1, con e’k+1ei=0, para i=1,2,...,k y k=1,2,...,p-1.



## Pero



De está manera Var(Yk+1 ) =. Si se puede mostrar , que ei es perpendicular a ek(esto es, eiek=0,i) conociendo Cov(Yi,Yk)=0.



Ahora los vectores de son ortogonales si todos los valores propios son distintos. Si los valores propios no son distintos, los vectores propios corresponden a valores propios comunes quizás se diga que son ortogonales. Por consiguiente , para los dos vectores propios ei y ek , ei'ek=0, .

A partir que ek=kek, multiplicamos ei' conociendo



* + 1. **Componentes Principales de la Población con la Matriz de Covarianza**

# Decimos que X'=[ X1,X2,...,Xp] teniendo una matriz de covarianza con los pares de valores y vectores propios donde

Si  son componentes principales.

Entonces



Por definición . Para  con , podemos escribir donde es la matriz diagonal de los valores propios y  dado que PP’=P’P=I. Usando tr(AB)=tr(BA).

### Tenemos que



De está manera:



### Se dice que

Varianza Total de la Población = (2.7)

=+...+

y consecuentemente, la proporción total de la varianza, explicada por k componentes principales es:



Por decir, del 80 al 90% de la varianza total poblacional, para un p largo, podemos atribuir al Primero uno, dos o tres componentes entonces estas componentes pueden “reemplazar” las variables originales p, en el exterior perderíamos más información.

Cada componente del vector de coeficiente  también tiene mérito de inspección. La magnitud de que son medidas de importancia de k variables a i componentes principales, sin consideración a otras variables . En particular es proporcional al coeficiente de correlación entre Yi y Xk .

**2.1.4 Coeficientes de correlación entre las componentes Yi y la**

**variable Xk**

Si son componentes principales obtenidas de la matriz de covarianza , entonces:



Son los coeficientes de correlación entre los componentes Yi y la variable Xk.

Aquí **** son los pares de valores y vectores propios para .

### Demostración

Si colocamos  tal que Xk = ak’X y la  de acuerdo a las combinaciones

lineales:



Donde y son el vector media y la matriz de covarianza de X respectivamente.

A partir que . Entonces  y las  con lo cual obtenemos lo siguiente:



Aunque la correlación de las variables con las componentes principales muchas veces nos ayudan a interpretar las componentes, ellos miden solamente univariado de manera individual a X del componente Y.

Esto es , ellos no nos indican la importancia de una X con una componente Y en la presencia de las otras X’s. Por está razón, algunos estadísticos recomiendan solamente que los coeficientes eik se use para la interpretación de los componentes.

Aunque los coeficientes y las correlaciones pueden guiar a diferenciar el rango como medida importante de las variables conociendo cada componente, esto es, que los rangos no son apreciablemente diferentes.

En la práctica, las variables con coeficientes relativamente largos, tienden a ser correlaciones relativamente largas, por lo tanto dos medidas de importancia, la primera multivariada y la segunda univariada, frecuentemente se obtienen resultados similares. Se recomienda que ambos coeficientes y correlaciones examinen la ayuda de interpretar componentes principales.

* + 1. **Componentes Principales con Variables Estandarizadas**

Las componentes principales que se obtienen a través de variables estandarizadas son las siguientes:



La Notación de la Matriz es :



Donde la matriz de desviación estandar V1/2 está definida de la siguiente manera:



E[Z] = 0 y Cov(Z)=(V1/2)-1∑ (V1/2)-1=ρ

#### Donde



#### Y





**2.2 Análisis de Regresión Múltiple**

**2.2.1 Introducción**

En la mayor parte de las aplicaciones del análisis de regresión se usan modelos que son más complejos que el de la simple línea recta.

Los modelos probabilísticos en los que intervienen x2 , x3 (o términos con potencias más elevadas ) o más que una variable independiente se llaman modelos de regresión múltiple. La forma general de esos modelos es:



Ahora la variable dependiente Y se expresa como función de k variables independientes Se suma el término de error aleatorio para tener en cuenta la desviación entre la parte deterministica del modelo,  , y el valor de la variable dependiente Y. El componente aleatorio hace que el modelo sean probabilístico y no determinista. El valor del coeficiente determina la contribución de la variable independiente xi y  es la ordenada al origen Y. En general, se desconocen los coeficientes porque representan parámetros de la población.

En el caso del modelo de regresión múltiple son aplicables los pasos que se siguieron para desarrollar un modelo de línea recta.

Paso1. Proponer la forma del modelo. Esto significa seleccionar las variables independientes que debe incluir el modelo.

Paso2. Estimar los parámetros desconocidos  .

Paso3. Especificar la distribución de probabilidad del componente aleatorio de error  y estimar su varianza .

Paso4. Comprobar lo adecuado del modelo .

Paso5. Emplear el modelo ajustado para estimar el valor promedio de Y o para predecir un valor determinado de Y para valores dados de las variables independientes.

**2.2.2 Ajuste del Modelo: Método de los Mínimos Cuadrados**

El método de ajustar modelos de regresión múltiple es idéntico al modelo simple de línea recta: es el método de los cuadrados mínimos. Es decir, se selecciona el modelo estimado



que hace mínimo a



La diferencia principal entre los modelos de regresión simple y múltiple es la dificultad de cálculo. Las (k+1) ecuaciones lineales simultáneas que deben resolverse para calcular los (k+1) coeficientes estimados  son difíciles de resolver.

**2.2.3 Estimación de , La varianza de **

La especificación de la distribución de probabilidad del componente de error  del modelo de regresión múltiple sigue las mismas características generales que la del modelo de línea recta. Se supone que  tiene distribución normal con promedio cero y varianza constante, para cualquier conjunto de valores de las variables independientes .

Para el modelo general de regresión múltiple



se debe calcular los (k+1) parámetro . Así, el estimador de es la SEC(suma de errores al cuadrado) dividida entre la cantidad n-(número de parámetros estimados ) .

Es decir,



El estimador de se usa tanto para comprobar lo adecuado del modelo, como para dar una medida de confiabilidad de los predictores y estimaciones .

* + 1. **Prueba de la Adecuación del modelo: El Coeficiente de**

**Determinación**

Para determinar una medida estadística que cuantifique lo bien que ajusta el modelo de regresión múltiple a un conjunto de datos, se usa el equivalente de r2 , para el caso de regresión múltiple .

Se define el coeficiente de determinación múltiple R2 de la siguiente manera:



donde  es el valor predicho de Yi para el modelo. De la misma manera que en el caso del modelo lineal sencillo, R2 representa la fracción de la variación en la muestra de los valores de y (medida por SSyy ) que explica la ecuación de predicción de los cuadrados mínimos. Así, si R2 =0, quiere decir que falta por completo el ajuste del modelo a los datos, y si R2 =1 quiere decir que se trata de un ajuste perfecto, y que la gráfica del modelo pasa por todos los puntos del diagrama de dispersión .

En general, mientras mayor sea el valor de R2 , mejor será el ajuste del modelo a los datos.

***Prueba de la Utilidad de un Modelo de Regresión Múltiple : Prueba F Global***



Medida estadística :



Región de Rechazo:



Donde

n = número de datos

k = número de parámetros  en el modelo , con excepción de 

* 1. **Prueba de Durbin-Watson en Modelos de Regresión**

Una de las razones de la existencia de autocorrelación es que podrían no haberse tomado en cuenta en el modelo variables importante de predicción . La autocorrelación puede eliminarse mediante la inclusión de las variables omitidas en el modelo de regresión.

La autocorrelación también puede presentarse debido a que los residuos sucesivos tienden a estar positivamente correlacionado, es decir, los grandes residuos negativos siguen a grandes residuos negativos y los grandes residuos positivos siguen a grandes residuos positivos.

Cuando se encuentra presente la autocorrelación, el análisis de regresión es afectado en tres formas:

1. Los estimadores de la Media Cuadrática, aunque son no sesgados ya no tienen varianza mínima.
2. Los estimados s2(Bi) pueden subestimar, en forma seria, las varianzas de los estimadores de la Media Cuadrática de Bi.
3. Los intervalos de confianza y las pruebas de hipótesis que incluyen, ya sea la distribución t de Student o la distribución F, nos son teóricamente válidas.

Al emplear la estadística de Durbin-Watson se prueba la hipótesis nula:

Ho:ρ = 0

contra la alternativa

H1:ρ >0

Nótese que H1 es una hipótesis alternativa unilateral superior exhiben muchas veces una autocorrelación positiva. La estadística de Durbin-Watson se basa en los residuos que resultan después de obtener la ecuación de regresión estimada . Se calcula un valor de esta estadística a partir de la expresión.



donde el residuo es ei= yi – y\*i.

Si los errores se encuentran positivamente autocorrelacionados, es probable que los errores adyacentes tengan la misma longitud.

De está forma, pequeñas diferencias entre los residuos adyacentes sugieren que ρ es mayor que cero; pero cuando las diferencias son pequeñas. Se rechaza la hipótesis nula de autocorrelación cero siempre que d tiene un valor relativamente pequeño.

Durbin-Watson tabularon los límites inferior y superior di y ds, respectivamente, para probar Ho. Dado los límites di y ds, la decisión para Ho se toma de la siguiente forma:

1. Si d< di , rechazar Ho
2. Si d> ds , no puede rechazarse Ho,
3. Si di<d< ds , la prueba no es concluyente.

* 1. **Prueba de Kolmogorov-Smirnov**

Al definir la medida estadística K-S, se definirá la Función de Distribución empírica. Supongamos que Y es una variable aleatoria continua que tiene una función de distribución F(y). Una muestra aleatoria de n realizaciones de Y produce las observaciones y1, y2, …, yn. Es conveniente reordenar esos valores observados de menor a mayor, y las yi ordenadas se representarán mediante y(1)< y(2)<…< y(n).

Supóngase que se toma una variable aleatoria continua Y, como hipótesis nula que tiene una función representada por F(y). La hipótesis alterna es que F(y) no es la función verdadera de distribución de Y.

La medida estadística D de K-S se basa en la distancia máxima entre F(y) y Fn(y), es decir,

**D = max ⏐F(y) - Fn(y) ⏐**

**y**