LA REEXPRESIÓN DE LAS VARIABLES

La clave para abordar los problemas creados por la nonormalidad y nolinealidad es la reexpresión. **La reexpresión consiste en el uso de una escala de medida (unidad de medida) diferente a la que en la variable fue originalmente medida.**

La reexpresión o transformación es un importante modo de corregir el sesgo de las distribuciones nonormales. Mediante ellas la dispersión de las variables puede modificarse de modo que casos que anterioremente aparecían como atípicos o extremos, estén ahora, dentro del rango de distribución nromal de la variable. Es decir, es como si la transformación "tirara" de la cola o colas de la distribución dónde aparecen los sesgos, llevando la distribución hasta la normal.

NORMALIDAD Y LINEALIDAD

Lo que no es tan clarao es que la reexpresión también consiga linealizar los Scatter plots correspondientes. Podríamos pensar que linealidad y normalidad son independientes y que la transformación de una variable ahcia su normalización podría no tenner ningún efecto sobre su ajuste lineal con otra, como de hecho ocurre. Sin embargo, nonormalidad y nolinealidad amenudo van de la mano, y **la reexpresión** puede responder a **ambos problemas**. Para comprender porqué, es necesario en primer lugar hacer una distincinción entre **transformaciones lineales** y **no lineales**.

TIPOS DE TRANSFORMACIONES

Suma, resta, multiplicación y división, de una variable por una constante, o cualquier combinación de estas operaciones, no sólo preserva el orden entre los datos sino la distancia entre ellos, por lo que la forma de la relación no se ve afectada. De este modo un Scatter Plot entre los nuevos valores y los originales produce una linea recta, de ahí la denomionación de transformación lineal. **Transparencia**

Por el contrario la trnasformación de una variable por **logaritmos, raices, potencias o exponenciales** cambia la distancia relativa entre los datos, produciendo por tanto distribuciones con diferentes formas, por lo que un Scatter Plot entre los valores originales y los tranformados produce un gráfico no lineal, de ahí el nombre de transformación no lineal. **Transparencia**

Las transformaciones no lineales no lineales no monotónicas, alteran la cambian la información de manera substancial , pues modifican los valores absolutos de los datos, la distancia entre ellos e incluso el orden entre los mismos. Debido a ello es poco recomendable su utilización aplicada a los fenómenos que son nuestro objeto de estudio.

Podemos ver pues, que en estos términos la ecuación de regresión lineal produce un conjunto de valores estimados de Y' que son una **transformación lineal** de los valores observados de **X Y= a+bX**
Simplemente consiste en multiplicar a **X** por una constante y añadirle otra. De este modo la forma de la distribución de los valores de Y y los de X es la misma, o si la relación es – un reflejo de los de X. Para que ambas variables estén relacionadas deben pues tener una forma similar o de espejo. Estono significa que variables sesgadas, en la misma dirección es decir parecidas en su forma, no

estén relacionadas lienalmente, pero si las simetrizaramos la relación lineal entre no se varía alterada y su nueva forma sería beneficiosa para otro tipo de análisis y relación con otras variables.

Por todo ello si la distribución de Y difiere sustancialemnte de la de X la transformación no puede ser lineal ya que no cambiaría la forma de la relación entre las variables. En este caso la transformación ha de ser no lineal. En otras palabras. La relación entre dos distribuciones no iguales ha de ser no lineal, como en el caso de una distribución simétrica con una sesgada, pues es imposible dibujar una linea recta que satisfaga los contrastes impuestos por las dos distribuciones. Esto no quiere decir que la relación entre variables con distribuciones similares distribuciones haya de ser necesariamente lineal sino sólo que las distribuciones con formas diferentes no pueden tener una relación lineal. Debido a ello no todas las situaciones de no linealidad son tratables. Sin embargo la reexpresión sigue siendo una herramienta útil de la aproximación exploratoria ya que a menudo responde simultáneamente a los problemas de no linealidad y no normalidad.

Esta relación entre la forma de la distribución y la forma de la relación entre las varibles, ofrece las bases de una aproximación sistemática a la modelización de las relaciones no lienales. **En primer lugar el analista debiera comenzar el análisis simetrizando las variables y a continuación trabajar, en los futuros análisis, con las variables simetrizadas.** Las ventajas de ello son :

1.- En primer lugar los problemas de no linealidad por distribuciones sesgadas, quedarían eliminados. Por supuesto no se eliminarían todos los casos de no linealidad, pero los más comunes pueden ser abordados de este modo, ya que cualquier resto de no linealidad puede ser más fácilmente modelada si las variables son simétricas y además tienden a la normal.

2.- En segundo lugar, la simetría es una cualidad y propiedad deseable en si misma. En estos casos las relaciones con otras variables en cualquier tipo de análisis se hacen más fáciles, cuando los datos no están concentrados en un área. Además incluso los estadísticos más robustos se ven alterados cuando las distribuciones no son simétricas por tener una desproporcionada concentración de casos en un determinado rango. Y si como es sabido la mayor parte de los análisis se basan en al utilización de estos estadísticos de localización y dispersión, los resultados de los mismos pueden aparecer sesgados.

ELIGIENDO EL TIPO DE TRANSFORMACIÓN

No cualquier tipo de función es igualmente útil para la reexpresión en un análisis exploratorio. Siempre son preferibles las más sencillas a las complejas. La complejidad de una función no viene determinada por el número de operaciones matemáticas para llevarla a cabo sino por los efectos de esas operaciones sobre los valores originales.
Tres son las características a observar en relación a los valores originales: EL VALOR ABSOLUTO DE LOS DATOS ORIGINALES, LA DISTANCIA ENTRE ELLOS Y EL ORDEN DE LOS MISMOS. Cuanto más compleja sea la transformación a más características afectará la transformación.

En términos de esta tipología, las funciones lineales, son las más simples ya que sólo afectan el valor absoluto de los datos. Por el contrario la raíz cuadrada de un conjunto de valores positivos

no sólo cambia sus valores absolutos sino que modifica las distancias relativas entre ellos. Las funciones más complejas afectan también al orden entre los datos, es decir a las tres características.

Estos tres tipos de cambios indican la disntinción principal entre las funciones aritméticas: **entre lineales y no lineales y monotónicas y no monotónicas.** Una función monotónica no cambia el orden entre los casos minetras que una monotónica lo hace y las lineales no alteran la disntnacia entre los casos mientras que las no lineales lo hacen. Asi que todas las funciones lineales son monotónicas y todas las no lineales no monotónicas. Así pues en orden a su complejidad, quedan clasificadas : lineales, no lineales monotónicas, y no lineales no monotónicas.

De este modo, las relaciones no lineales monotónicas pueden ser modeladas con reexpresiones que simplemente ajusten las distnacias de unos valores con otros em ambas variables. Las no monotónicas, sin embargo, requieren reexpresiones que cambien el orden entre los datos en orden a modelar los cambios en la dirección de la relación, en resumen requiren funciones no monotónicas.

ERROR TOTAL DE PREDICCIÓN (TPE) Y MÍNIMOS CUADRADOS

TPE=(4+ 3 - 3)=4 **TPE=(6+3+9)=18** **TPE=(- 4 – 6 +10)=0**

$TPE = \Sigma (Yi – Yr)$

MINIMOS CUADRADOS

MINIMIZAR $\Sigma (Yi - Yr)^2$

TSS = RSS + ESS

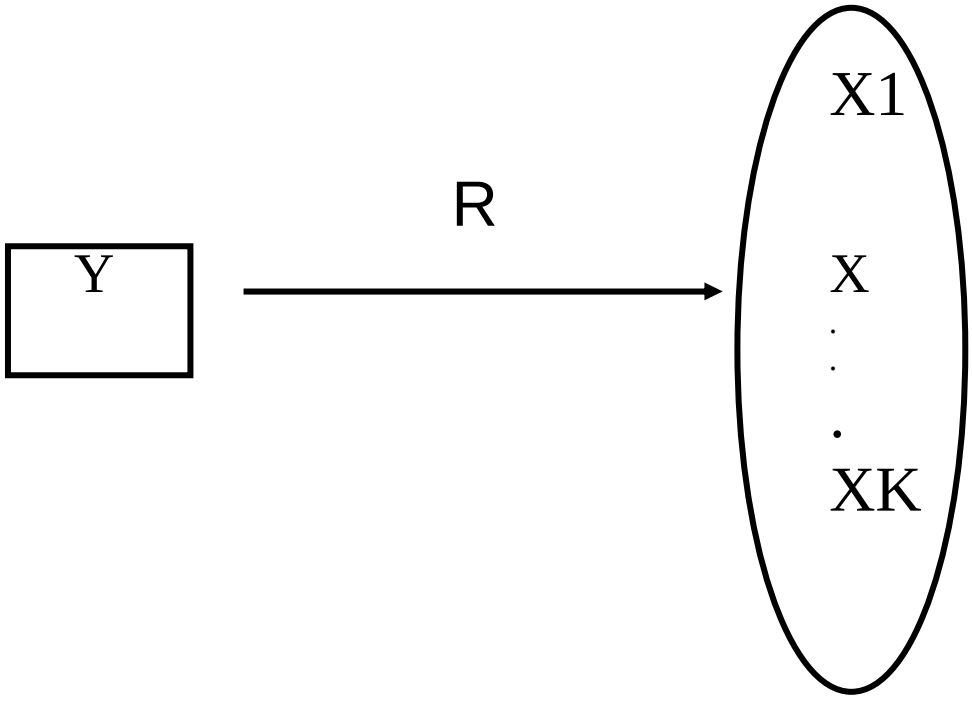
$$\Sigma (Yi - \bar{Y})^2 = \Sigma (Yr - \bar{Y})^2 + \Sigma (Yi - Yr)^2$$

Componentes de la variación de Y como variable dependiente

TSS = sumatorio total de las desviaciones al cuadrado
RSS = sumatorio de las desviaciones explicadas por la regresión al cuadrado.
ESS = sumatorio de las desviaciones no explicadas al cuadrado.

Y = 45.720 + 10.000 X

CONCEPTO DE CORRELACIÓN MÚLTIPLE



REGRESIÓN MÚLTIPLE PROCEDIMIENTO PASOS SUCESIVOS

DEPENDIENTE : SALARIO

INDEPENDIENTES:

- SEXO
- NIVEL EDUCATIVO ALCANZADO
- EXPERIENCIA LABORAL
- EDAD
- CLASE SOCIAL

1.- INCREMENTO DE R MÚLTIPLE

2.- TOLERANCIA = $1 - R^2_{ix}$

R2ix : correlación múltiple al cuadrado entre cada independiente y el resto

3.- COEFICIENTE t

$$t = \frac{B}{SEB}$$

SEB : Error típico de B

$$t^2 = F$$

4.- ECUACIÓN DE PREDICCIÓN:

$$Y = 86585,31 + 4062,98\ X1 - 12454,77\ X2 + 624,73\ X3 - 4328,84\ X4$$

LA REGRESIÓN MÚLTIPLE COMO EJEMPLO PARADIGMÁTICO DE LOS ANÁLISIS PARAMÉTRICOS

El modelo de regresión múltiple es un buen ejemplo para desentrañar el sentido último de modelos paramétricos en el análisis de lo social, y los supuestos que acompañan a éstos. Con el objeto de aplicar determinados modelos, que se muestran potentes desde el punto de vista matemático estadístico, se asumen ciertos procedimientos analíticos que requieren para su aplicación unas condiciones concretas. La exposición de los supuestos paramétricos necesarios para la aplicación de dichos modelos, tiene aquí una función ilustrativa respecxto a las restricciones a las que se somete la información básica que sirve de punto de partida. Si estos supuestos no están presentes en los datos que maneja el investigador, éste se ve impelido a realizar determinadas transformaciones, en orden a conseguir su cumplimiento o por el contrario a sacrificar la validez y fiabilidad de sus resultados.

A pesar de ello y , como anteriormente comentaba, parece admitido por muchos investigadores que el hecho de que los supuestos paramétricos no se cumplan estrictamente no afecta a los resultados de un modo tan determinante como para anularlos. Es decir, se suponen suficientemente robustos ante ciertas desviaciones de los supuestos implícitos en ellas. Pero la cuestión de fondo no es tanto si los modelos son robustos o no, como si son válidos en su aplicación para el análisis social. Pues lo que si parece evidente es, que definida la realidad social desde el punto de vista de la ***complejidad***, la naturaleza de ésta está lejos de encerrar un cumplimiento formal de dichos supuestos, haciendo necesaria una mayor adecuación de los modelos para su estudio.

Sea como fuere, el punto central es, aquí, que la aplicación de modelos matemático-estadísticos para el análisis social implican o presuponen el cumplimiento de ciertas características por parte de los datos, cuando éstas generalmente no se dan. En este sentido, la mera observación de las funciones matemáticas elegidas en cada caso para representar la relación funcional entre las diferentes partes -variables- da una idea de la naturaleza y características de estos supuestos. Es decir, que adoptar un tipo u otro de función matemática supone la asunción, desde un punto de vista

teórico, de determinados supuestos respecto al funcionamiento y características de los fenómenos que se intentan representar y analizar. La aplicación de un modelo lineal implica por tanto, que la definición de la relación entre las variables y procesos que se investiga se corresponde con un comportamiento tal. Del mismo modo, asume la consideración de una realidad social lineal fruto de la adición de la variabilidad de estos mismos procesos de modo que el resultado es función de la suma de las partes. Una concepción semejante, contradice la visión de una realidad social *multidimensional*, *compleja* en la que las interacciones de variables y procesos se superponen y dan lugar a fenómenos que van más allá de una mera acumulación de efectos. Así pues la asimilación de ciertos modelos matemáticos para representar y analizar la realidad social debe de hacerse en función de una consideración teórica previa fruto de una concepción concreta de la realidad social respecto a una epistemología precisa de lo social y no al contrario.

El análisis de regresión múltiple es una extensión de la regresión lineal simple a más de una variable independiente. Mediante un modelo semejante, se trata de analizar las relaciones de interdependencia de un conjunto de variables considerando una de ellas como dependiente del influjo de resto de independientes.

> " La regresión múltiple no es más que un método para especificar, e interpretar un modelo explicativo en el que una variable dependiente se estudia en función de una serie de una o más variables explicativas independientes. El objetivo consiste en cuantificar la relación entre la variable dependiente y las independientes; y en establecer con qué grado de confianza podemos afirmar que la cuantificación realizada se ajusta a la realidad observada...En realidad la regresión múltiple descansa sobre dos pilares de orígenes distintos: el cálculo diferencial y la teoría de la probabilidad. El cálculo diferencial (estudiado por Ferrita, Leibniz y Newton en el siglo XVI, y por los Bernoulli y D'alambert en el XVII) nos permite cuantificar la relación entre las variables. En concreto el principio que se utiliza en la mayoría de los casos se conoce con el nombre de 'mínimos cuadrados' enunciado por Gauss en 1794. La teoría de la probabilidad (desarrollada en los siglos XVII y XIX por Moivre, Bayes, Laplace y Pearson) nos permite determinar

> con qué confianza podemos afirmar que nuestras estimaciones cuantitativas se ajustan a la realidad de la población que estamos estudiando" (F. Mauro Guillén:1992, pag. 21)

Así la esencia y naturaleza de este análisis queda reflejada en la ecuación de regresión múltiple que se concreta en la siguiente expresión matemática:

$$\mathbf{Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k}$$

Como se ha dicho, su simple observación nos remite a algunos de los supuestos paramétricos básicos para la aplicación de este tipo de modelos. En primer lugar, en cuanto a la notación funcional que se desprende de la fórmula anterior

$$\mathbf{Y = f (X_1 + X_2 + X_3 + ... + X_k)}$$

-**Y** es función de $\mathbf{X_1 , X_2 , X_3 , ...}$ y $\mathbf{X_k}$ - el supuesto que se desprende es el de linealidad

- **La linealidad**

Todo lo anterior pone de manifiesto el supuesto básico de la regresión múltiple: la relación de interdependencia entre las variables debe de responder a un modelo **lineal**. Al igual que muchas otras técnicas estadísticas derivadas del modelo lineal general, su formulación presupone que los fenómenos sociales analizados a través de ellas pueden explicarse en términos de linealidad. Para averiguar si ésta está presente en los datos -pues se desconoce, ya que es un presupuesto del que se parte, pero no una constatación del mismo- es conveniente analizar los gráficos de dispersión -scatter plot- para observar su cumplimiento. En este caso, el análisis de los valores residuales también es fundamental. Generalmente, éstos se representan en un gráfico de dispersión de residuales, respecto a los valores estimados a partir de la ecuación de regresión o respecto a los de cada variable independiente. De este modo y mediante la observación de la dispersión de estos valores puede realizarse un diagnóstico sobre el cumplimiento del supuesto de linealidad. Pero un modo más preciso de llevar a cabo este diagnóstico es el que puede realizarse a través de los gráficos de regresión parcial. En ellos, una vez eliminados los efectos lineales del resto de las variables independientes, se aprecia la regresión parcial entre la dependiente y la independiente seleccionada. Si se confirma el no cumplimiento de este primer supuesto se presentan fundamentalmente dos alternativas. Una de ellas, pasaría por la realización de un profundo análisis exploratorio para proponer un modelo alternativo. Otra opción, la más comúnmente practicada, consiste en realizar las transformaciones o re-expresiones oportunas sobre la variable/s que provoca/n la desviación del supuesto de linealidad. Además de ello las transformaciones generalmente utilizadas, -las no lineales monotónicas como logaritmos y raíces- no sólo pueden hacer que el modelo recobre la linealidad sino que pueden solucionar problemas de normalidad y variabilidad, aspectos relacionados con otros tantos supuestos de la regresión lineal.

Una vez que se ha optado por esta segunda alternativa, la elección concreta de la transformación a realizar depende de diversas consideraciones. Generalmente, como se ha indicado,

las transformaciones más frecuentes son la no lineales monotónicas[23] . Para seleccionar la más adecuada el criterio más frecuente suele ser la asimetría positiva o negativa de la distribución de residuales. Cuando ésta aparece positivamente sesgada la transformación más adecuada sobre la variable dependiente es la logarítmica o la extracción de raices; por el contrario si la distribución aparece negativamente sesgada es más recomendable la utilización de potencias. Una vez realizadas estas estrategias para preservar el cumplimiento de los supuestos paramétricos, el análisis se realiza sustituyendo las variables originales por las transformadas, de modo que si realizáramos de nuevo una análisis de residuales obtendríamos que éstos se distribuyen aleatoriamente, tanto para los "plots" relativos, a la regresión como en los correspondientes a la regresión parcial.

El objetivo de todas estas estrategias es capacitar los datos para su tratamiento bajo determinados modelos, de modo que la cuestión aquí no es tanto si las transformaciones son las adecuadas -preservan la información original- sino si los supuestos se dan o no realmente. Es decir, las re-expresiones que podemos utilizar suponen expresar los datos en una escala diferente manteniendo su originalidad a la vez que se favorece el cumplimiento de determinados supuestos. Por otra parte, el hecho fundamental, es que definida la realidad social desde la perspectiva de la complejidad tales procesos lineales no se dan realmente salvo en periodos de tiempo y bajo circunstancias concretas, en general los procesos sociales tienen una configuración no lineal en la que se combinan multitud de aspectos de *indeterminación* y *caos*. Por ello, a pesar de que desde un punto de vista matemático estas transformaciones sean adecuadas al fin que pretenden, desde un punto de vista sustantivo, su funcionalidad y sentido es limitado. En definitiva, *"estrangula"'* la complejidad connatural a los procesos sociales no resulta una aproximación metodológica adecuada si lo que se busca es una análisis de los mismos y no una conformidad a ciertos modelos dados por válidos para su estudio.

[23] Ello es debido, fundamentalmente, a que las lineales monotónicas no afectan a la distancia ni al valor de los datos, por lo que no resuelven problemas de linealidad ni normalidad. Por otra parte, las no lineales no monotónicas no sólo alteran el valor de los datos originales y su distancia sino que además modifican el orden entre ellos, lo cual supone una transformación total de la información original y provoca dificultades para su tratamiento.

- **Aditividad y *multicolinealidad*[24]**

Del mismo modo que la ecuación de regresión pone de manifiesto el supuesto de la linealidad, otras características del mismo quedan patentes a partir de esta función matemática. Una de ellas, es el requisito de aditividad y no *multicolinealidad* entre las variables. Es decir: los efectos de las variables deben ser aditivos, por lo que no es conveniente que las variables independientes estén correlacionadas. El significado último de la asunción y utilización de modelos semejantes supone que los fenómenos o procesos sociales que se pretenden explicar son fruto de otros sucesos o variables independientes entre si . De nuevo nos topamos aquí con otro de los supuestos que constituyen una construcción artificiosa de la realidad social para hacerla susceptible de ser tratada mediante determinados modelos. Son, precisamente, estos modelos los que conllevan unos determinados requisitos para su aplicación, supuestos que tanto en este caso como en otros no son la característica esencial de lo social. Desde el punto de vista de la complejidad, es difícil concebir una realidad social no compleja no multidimensional en la que el todo es simplemente la suma de las partes. En cierto sentido, el supuesto de no *multicolinealidad* lleva implícita la idea de aditividad de elementos independientes que configuran un todo; por el contrario, el funcionamiento de un sistema complejo es inconcebible sin la interacción de sus múltiples dimensiones, que engendran nuevos procesos distintos e independientes de ellas. Por tanto, asumir un modelo aditivo y de interacción restrictiva entre las variables supone renunciar a una visión sistémica y estructural de la realidad social, al tiempo que nos traslada a un ámbito reduccionista y empobrecido, tanto desde el punto de vista teórico, como metodológico.

Para comprobar el cumplimiento de este supuesto, ya que al igual que en el caso anterior se parte de él y no de su constatación, existen una serie de procedimientos y técnicas precisas. Uno de los modos más frecuentes suele ser la simple observación de la matriz de correlaciones inicial. La existencia de altos coeficientes de correlación entre pares de variables independientes puede ser un

[24] La *multicolinealidad* hace referencia a ciertos niveles de correlación o interacción entre las variables independientes, es decir al hecho de que unas variables se aproximan a ser combinación lineal de otras. Por el contrario el modelo de regresión parte del supuesto de ausencia de correlación entre las variables independientes de modo que puedan calcularse, con suficiente confianza, los efectos de cada una de las independientes sobre la dependiente controlándose el efectos del las restantes. Si la relación entre ellas es alta se dice que hay una alta '*multicolinealidad*' lo que produce sesgos que restan fiabilidad, no sólo a los coeficientes de regresión parcial, sino al resto de los estadísticos y parámetros hallados en la regresión.

buen indicador de la existencia de *multicolinealidad*. Sin embargo, el problema en torno a esta cuestión es que puede existir *multicolinealidad* a pesar de que dichos coeficientes no sean apreciablemente altos. Debido a ello, y para una detección más precisa de la existencia de *multicolinealidad*, suele recurrirse a otro criterio: el de la 'tolerancia'. Este criterio está íntimamente relacionado con la *multicolinealidad*, ya que indica en qué medida una variable independiente dada está relacionada con el resto de las independientes. En un modelo de regresión lineal, la tolerancia de una determinada variable es la proporción de varianza no explicada por el resto de las independientes y es:

$$\textbf{Tolerancia} = 1 - R^2_{ix}$$

dónde R^2_{ix} es el coeficiente de determinación entre cada independiente considerada como dependiente, y el resto de las independientes. Cuanto mayor sea la tolerancia, mayor es la independencia de la variable respecto al resto -no *multicolinealidad*-, por lo que puede contribuir en mayor medida a explicar la dependiente. Por ello, además de ser un indicador de *multicolinealidad*, la tolerancia también es un criterio adecuado para determinar la inclusión o no de una determinada variable en el análisis, según su capacidad explicativa en función de su relación con el resto de las independientes y la dependiente. Así el procedimiento **'stepwise'** de regresión múltiple, requiere en cada paso del análisis un nivel mínimo de tolerancia para la inclusión de una determinada variable en el modelo. Todos estos procedimientos, se llevan a cabo con el objeto de detectar la *multicolinealidad* y evitar las consecuencias de ésta. Su existencia sesga y resta fiabilidad a los estadísticos obtenidos en la regresión. Así no podrá obtenerse una cuantificación precisa del la relación de cada una de las independientes y la dependiente, a partir de los coeficientes de correlación parcial, al no poderse controlar adecuadamente los efectos del resto de las independientes, si existe correlación entre ellas. Del mismo modo, si la correlación entre las independientes es alta el coeficiente de correlación múltiple incluye información redundante. Estos y otros sesgos influyen indudablemente en la interpretación global de resultados por lo que es necesario conocer en qué medida nuestros datos se apartan del cumplimiento de los supuestos paramétricos. Esta consideración orientará una interpretación más adecuada. No obstante, la mayor parte de los autores coinciden en señalar que el modelo de regresión es una técnica suficientemente

robusta como para soportar el no cumplimiento estricto de los supuestos, especialmente de los de normalidad y *homocedasticidad*. No ocurre lo mismo en el caso de la *multicolinealidad*, por lo que este aspecto debe de considerarse con especial atención. Por ello, si se detecta una alta *multicolinealidad* alguna/s de las variables habrán de ser excluidas del análisis, pero en este caso, como en otras tantas ocasiones, lo que es adecuado matemáticamente puede no serlo desde un punto de vista teórico o sustantivo. Eliminar ciertas variables del análisis, puede restar sentido al análisis mismo, desvirtuando un planteamiento metodológico-conceptual, en aras de la utilización de determinadas técnicas e instrumentos de análisis que dejan de ser guía para convertirse en un fin en si mismo.

- ***Homocedasticidad***

Este concepto está en la base de otro de los supuestos básicos de la mayor parte de los análisis paramétricos. Con él, se hace referencia a la situación en la que la varianza de las variables implicadas en el análisis no es significativamente diferente; en caso contrario, se habla de *heterodasticidad*. De este modo, las variables incluidas en un análisis de regresión múltiple -tanto la dependiente como las independientes- deben de presentar una varianza similar. Como se señalaba más arriba, la resistencia del análisis de regresión múltiple a la desviación del cumplimiento estricto de ciertos supuestos, permite que los resultados sigan siendo válidos. En este caso, por tanto, se tolera cierta divergencia de las varianzas, a condición de que la diferencia entre ellas no llegue a ser significativa. Si, por el contrario, a partir del análisis de residuales observamos que la dispersión de éstos aumenta o decrece respecto a los valores de la variable independiente o de los valores estimados de la regresión, podemos sospechar que el supuesto de *homocedasticidad* ha sido violado

Además del análisis de residuales, y con el objeto de conseguir una mayor precisión que permita determinar si la diferencia entre las varianzas es significativa o no, pueden realizarse diferentes pruebas. Entre ellas -prueba de Bartlett, prueba de C. Cohran, prueba de Levene, prueba de la Fmax. de Hartley, y otras- la más comúnmente utilizada es la F de Fisher. Esta prueba permite comparar las varianzas de las variables dos a dos, de modo que pueda evaluarse si la diferencia entre sus varianzas es significativa o no y comprobar así el cumplimiento del supuesto de *homocedasticidad*.

$$F = \frac{S^2_1}{S^2_2}$$

Siendo $S^2_1 > S^2_2$

- **Normalidad univariable y normalidad multivariable**

Bajo este supuesto, todas las variables incluidas en el análisis deben ajustarse a la distribución normal. Un requisito para la mayor parte de los análisis paramétricos multivariables es la normalidad multivariable, que a su vez incluye como condición la normalidad univariable. Este requisito es una condición necesaria pero no suficiente, ya que el hecho de que todas las variables se ajusten a la ley normal no implica que conjuntamente sigan una distribución normal multivariable. Es necesario, por tanto, que todas las variables, incluida la dependiente, se distribuyan normalmente para los valores dados de las otras variables.

> "Para que los datos procedan de esta distribución es necesario que cada variable sea normal y que el conjunto de las variables tengan una buena unidad experimental, de modo que sea apropiado estudiarlas conjuntamente...La normalidad univariante de cada una de las variables X_i, no es condición suficiente de normalidad multivariante. Se puede construir diferentes familias de distribuciones multivariantes no normales, cuyas marginales sin embargo lo sean." (C. M. Cuadras: 1991, pp.38-43)

Bajo estas condiciones, una primera aproximación para comprobar el cumplimiento o no de la normalidad multivariable, supone la constatación de la normalidad univariable. Respecto a ello, cabe destacar una vez más la conveniencia de realizar un análisis exploratorio previo a cualquier análisis multivariable, con el objeto de poder detectar desviaciones a los supuestos paramétricos, así como la estructura latente a los datos. De este modo, es fácil corregir y detectar desviaciones en torno a la normalidad, a la igualdad de varianzas, a la linealidad, etc. En este caso, si se observa que

algunas de las variables no se distribuyen normalmente, se realizarán las transformaciones oportunas para aproximarlas a la normal, con lo que es más probable que el supuesto de normalidad multivariable tienda a cumplirse.

Como en casos anteriores, los gráficos de residuales -histograma de residuales estandarizados, "normal probability (P-P) plot" y "casewise plot of residuals outliers"- pueden ser de gran utilidad, pues permiten diagnosticar la existencia de casos extremos en alguna de las variables -lo que permite concluir que no se distribuye normalmente- y ciertas asimetrías y sesgos.

Pero, como se ha indicado más arriba, comprobar el supuesto de normalidad univariable es condición necesaria pero no suficiente para la normalidad multivariable. Por ello, son necesarias algunas pruebas adicionales que permitan determinar si ésta se encuentra presente en los datos o no. En este caso, Cuadras recomienda una prueba de normalidad de las *componentes principales*

> " Si efectuamos un análisis de componentes principales a las n variables y aplicamos una prueba de normalidad a cada una de las componentes principales, tenemos una prueba de normalidad conjunta que además es suficiente. En efecto, la distribución $x_1,..., x_n$ es normal multivariante si y sólo si es normal univariante de la distribución de las n componentes. Es una consecuencia de la propia definición de la normal multivariante." (C. M. Cuadras: op. cit., p. 71)

- **Nivel de medida interval**

Otro de los supuestos del análisis de regresión múltiple obliga a que todas las variables que forman parte de la ecuación de regresión tengan, como mínimo, un nivel de medición de intervalo. De nuevo, en este caso, los intereses teóricos y metodológicos de la investigación pueden topar con una dificultad que no es inherente al análisis mismo, sino que se deriva de la utilización de ciertos modelos para el análisis social. Dado que la mayor parte de los fenómenos que podemos abordar tienen una naturaleza nominal -cualitativa, hacen referencia a cualidades más que a cantidades- u ordinal, es necesario realizar ciertas transformaciones para poder incluir estas variables en la investigación. En estos casos, la solución es la creación de variables 'dummy' -variables ficticias-

que sustituyan a las originales, de modo que éstas puedan ser tratadas matemáticamente como si fueran continuas. De este modo, a partir de una variable categórica, obtenemos K-1 variables ficticias -siendo K el número de categorías de la variable original-. Dichas variables presentan una estructura dicotómica y pueden así pasar a formar parte del análisis. Esto es debido a que las variables codificadas como 0 y 1 son equivalentes a las variables continuas dicotomizadas -y relacionadas linealmente- que se analizan a partir de la correlación tetracórica. De este modo, al expresar así la variable original ésta adquiere ciertas características que permiten incluirla en la ecuación de regresión. En la construcción de este tipo de variables, el nivel que queda excluido, (K-1), sirve de base de comparación del resto de las categorías, pero no existe ningún criterio para determinar cual de las categorías ha de eliminarse; esta consideración sólo depende de los intereses particulares de cada investigación. La necesidad de eliminar una de las categorías responde al hecho de que, en caso contrario, cualquiera de ellas sería una combinación lineal de las demás, lo que redundaría en problemas de *multicolinealidad*.

En definitiva, de todos estos supuestos, como de otros muchos de la estadística confirmatoria tradicional, podemos decir que obtienen "buenos ajustes", "medidas precisas" y resultados estadísticos “muy significativos", pero quizá no podamos decir que la significación obtenida corresponda a una significación real, sustantiva, que las medidas sean tales y que los ajustes sean fruto de la variabilidad real de los fenómenos.

En virtud de la necesidad de ofrecer una alternativa viable a esta compleja problemática, surgen en la actualidad técnicas no paramétricas, cada vez más potentes, cuya utilización se generaliza. La proliferación y generalización del uso de estas técnicas aparamétricas demuestra la tendencia creciente a considerar las limitaciones de las técnicas tradicionales, en su aplicación a realidades a las que no se adaptan tan adecuadamente como cabría esperar.

> “...más recientemente se han desarrollado otras técnicas que no exigen tantas restricciones sobre la naturaleza de la población. Tales técnicas aparamétricas, o de ‘libre distribución’, permiten obtener conclusiones con menos condiciones. El

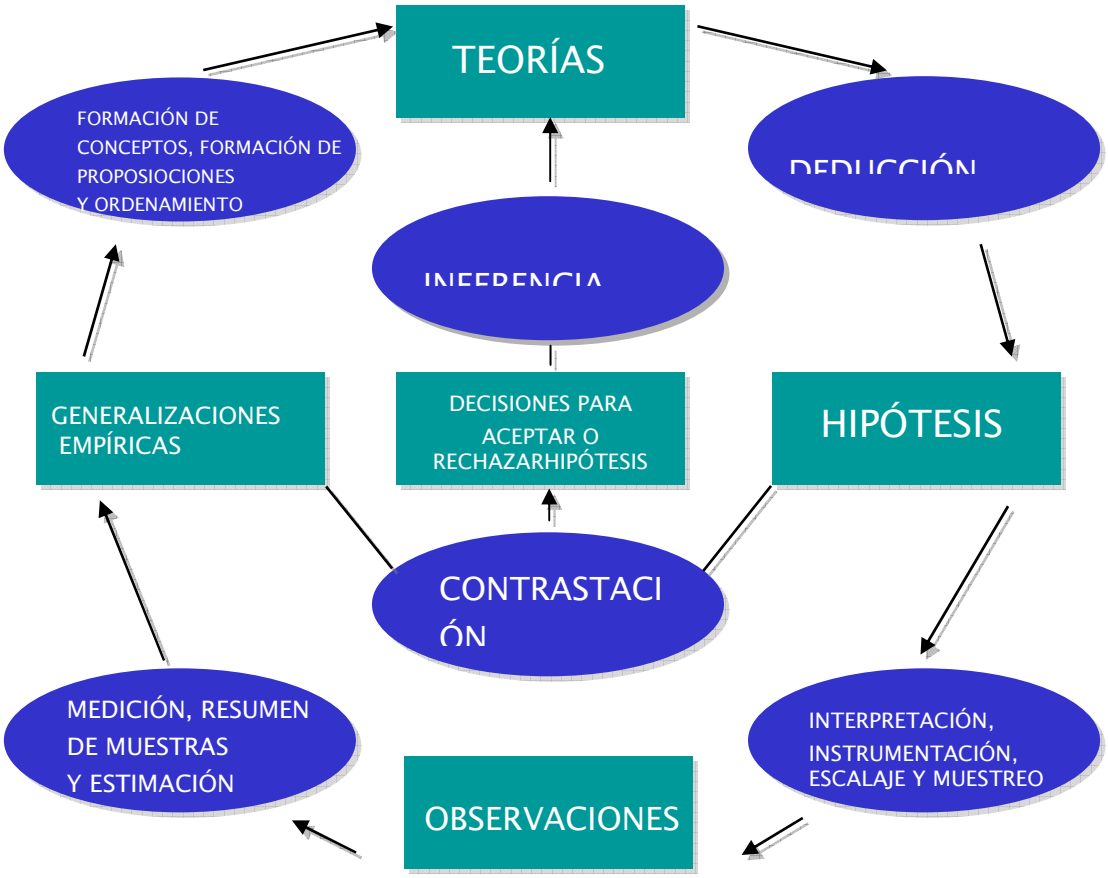
> tipo de conclusión que se puede obtener con el uso de tales técnicas será de la forma: Con independencia de la forma de la población se puede concluir que...'
> Algunos autores denominan también a las técnicas aparamétricas como 'pruebas de ordenación', lo que sugiere la existencia de otro factor diferencial entre las técnicas paramétricas y las aparamétricas. En efecto, en el cálculo de las pruebas paramétricas se pueden realizar todas las operaciones aritméticas con los valores obtenidos de las muestras. Si tales procedimientos aritméticos se aplicaran a valores que no son realmente numéricos se introducirían distorsiones en estos datos y las conclusiones que se obtuvieran vendrían sesgadas. Así, pues, sólo se pueden emplear técnicas paramétricas cuando los valores son verdaderamente numéricos. Sin embargo las pruebas aparamétricas atienden a la ordenación de los 'datos', no a su valor 'numérico' e incluso algunas técnicas pueden utilizarse con datos meramente clasificatorios que no pueden siquiera ser ordenados." (Manuel García Ferrando: 1986, p.156)

Además de ello, también cabe resaltar la importancia creciente que adquiere la perspectiva exploratoria en el análisis de datos, cuya utilización facilita la detección de irregularidades en las distribuciones respecto a los supuestos paramétricos de partida, ofreciendo la posibilidad de aplicación de estadísticos que se adecuen a su fiel representación. Se ponen así de manifiesto, en los primeros pasos del análisis, aspectos relevantes de los datos que, de otro modo, hubieran pasado inadvertidos y que alertan en cuanto a la idoneidad de aplicación de análisis con serias restricciones paramétricas.

En definitiva, todas estas consideraciones anteriores, vuelven a remitirnos al problema central que se aludía anteriormente y Coombs expresa como **"el dilema del sociólogo"**. Y es, precisamente, a este dilema al que los diferentes enfoques tratan de dar respuesta. Por una parte, la asunción de los supuestos paramétricos implícitos, por parte de la estadística tradicional (normalidad linealidad, parsimonia...) tiene como resultado la referencia al primer aspecto de este dilema, "escoger entre poner sus datos en un orden sencillo...". Este orden sencillo encaja

perfectamente con el planteamiento tradicional sintetizado en la lógica de la investigación científica expuesta por Walter Wallace,

Gráfico 2 Walter Wallace(1980, p.22)



Este gráfico pretende relacionar los diferentes elementos del proceso de investigación a la vez que reflejar la conexión entre teoría e investigación empírica. De este modo la perspectiva que se ha dado en denominar "confirmatoria" se centra en la zona deductiva del gráfico, la zona derecha del mismo, que partiendo de la teoría, supuestos e hipótesis derivadas de ella, pretenden una contrastación empírica volviendo de nuevo a la teoría.

> "El análisis estadístico tradicional -regresión, factorial, de ecuaciones estructurales, etc.- da por supuesto un determinado modelo de realidad y a él adapta/ajusta los datos obtenidos en un intento estrictamente confirmatorio y de contrastación. Dentro de esta perspectiva tradicional, el analista puede llegar a comprobar los supuestos en los que se basan las técnicas, pero normalmente no se plantea un proceso inductivo de conocimiento previo y detallado de su matriz de datos para llegar a un modelo partiendo del análisis univariado, bivariado después y, por último multivariado. Por el contrario, el camino recorrido por el análisis estadístico tradicional es justo el inverso, es un proceso deductivo de contrastación de hipótesis utilizando modelos de comportamiento de la realidad preestablecidos" (F. Alvira: op. cit., p.332)

Por otra parte, el dilema de Coobms plantea una segunda estrategia "...o preguntarse si sus datos responden a un orden sencillo". De ahí surge otra perspectiva que no pretende imponer a los datos ningún modelo, sino explorar en su variabilidad latente y averiguar los modos y procesos de relación que emanan de los propios datos, en lugar de partir de supuestos a priori sobre sus relaciones y/o comportamiento. Esta estrategia "exploratoria" se centra en la parte izquierda del gráfico anterior, partiendo entonces de los datos hacia la *teorización* de las conclusiones derivadas de la exploración de los mismos, que producen generalizaciones empíricas, formando proposiciones y conceptos que derivan en la teoría. La lógica subyacente con este planteamiento se relacionarla con la perspectiva exploratoria fundamentada y desarrollada por Tukey (1977).

Para completar la información sobre el tema :

Sánchez Carrión , Juan Javier Manural de Análisis de datos Alianza Universidad textos Madrid 1996

Ferrán Aranaz Magdalena SPSS para Windows Programación y Análisis Estadístico McGraw-Hill Madrid 1996

Bisquerra Alzina Inchausti Introducción conceptual al Análisis Multivariable. Un enfoque informático con los paquetes SPSS-X; BMDP, LISREL, y SPAD (vol. I y Vol. II) Edit. PPU, Barcelona 1990

FASES en el desarrollo del ANÁLISIS MULTIVARIABLE

- Elaboración de la matriz de datos
- Depuración de datos
- Análisis univariable: descriptivo y exploratorio
- Análisis bivariable
- Análisis Multivariable

CRITERIOS de clasificación de las Técnicas de Análisis Multivariable

- Objetivo del análisis: **Exploratorias y descriptivas/Explicativas y confirmatorias**
- Tipo de relación entre las variables y número: **relaciones de dependencia / interdependencia**
- Nivel de medición de las variables: **técnicas Paramétricas/no Paramétricas**