UNIVERSIDAD DE JAÉN

Material del curso "Análisis de datos procedentes de investigaciones mediante programas informáticos"

Manuel Miguel Ramos Álvarez

MATERIAL IX "EXPLICACIÓN MULTIVARIADA CON REGRESIÓN"

Índice

9. Acercamiento con el fin explicativo: análisis inferencial orientado a Regresión.

9.1. Bases de Regresión Lineal

- Todas las predicciones del modelo, $\widehat{Y}_i$, descansan sobre la línea recta.
- Los errores de predicción ó residuales, $e_i = Y_i - \widehat{Y}_i$, se definen como la distancia vertical entre los puntos de datos y la recta.
- Los parámetros:
  - El parámetro de intersección B_0 corresponde al valor de $\widehat{Y}_i$ cuando X_i es cero ó punto de origen de la recta.
  - La pendiente B_1 cuantifica el cambio en $\widehat{Y}_i$ por cada incremento unitario en X_i. Bien postivo (crecimiento en el criterio conforme aumenta el predictor) bien negativo (decrementos en el criterio correspondiendo a incrementos en el predictor).
- Para adoptar decisiones de significación estadística se compara el valor de F con un valor crítico obtenido a partir del modelo de distribución F según el nivel de significación que imponemos. Si el valor de F asociado a la magnitud RPE supera el valor crítico, entonces nos inclinamos en contra de la Hipótesis Nula, o lo que es equivalente, a favor del modelo Ampliado frente al modelo Compacto y al contrario si el valor es inferior.

A) Análisis global de la regresión lineal

Reducc . Err. AMP
Error AMP
Error COM

Fuente	SC	gl (ν)	MC	F_k	η^2	p
Regres.						
Err. o Residual						
Total						

*p≤α

$$F_k = \frac{MCR}{MC_e} \equiv \frac{RPE/gl}{(1-RPE)/gl}$$

$$\eta^2 = \frac{SCR}{SCE(COM)}$$

- Supuesto M2

Analizar → Regresión → lineal → Dependiente: ExCard; Independientes: Cigarrillos; Estadísticos→ Estimaciones → Continuar → Aceptar

Statistics→ Multiple Regression →Variables: Dependent: ExCard; Independent: Cigarrillos → OK→ Aceptar→*Pestaña* Advanced → Summary→ ANOVA → *Pestaña* Residuals

B) Análisis de los parámetros

- Para B_0 comparar los modelos: $\left\{\begin{matrix} AMP: Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \\ COM1: Y_i = \beta_1 X_i + \varepsilon_i \end{matrix}\right\} \equiv \left\{\begin{matrix} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{matrix}\right\}$
- Para B_1 entonces compararíamos: $\left\{\begin{matrix} AMP: Y_i = \beta_0 + \beta_1 \bullet X_i + \varepsilon_i \\ COM2: Y_i = \beta_0 + \varepsilon_i \end{matrix}\right\} \equiv \left\{\begin{matrix} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{matrix}\right\}$

C) Resumen del Modelo

- **Los Intervalos Cofidenciales, p**ara la intersección y para la pendiente
- Para estimar la **potencia estadística** nos basaremos en RPE como medida del efecto de tratamiento, o mejor la medida ajustada, y a partir del mismo buscaremos en las curvas de potencia o mediante un programa especializado (i.e. Módulo Statistica: Power Calculation).

9.2. Análisis tipo regresión en el contexto Multivariado

9.2.1. Bases de Regresión múltiple

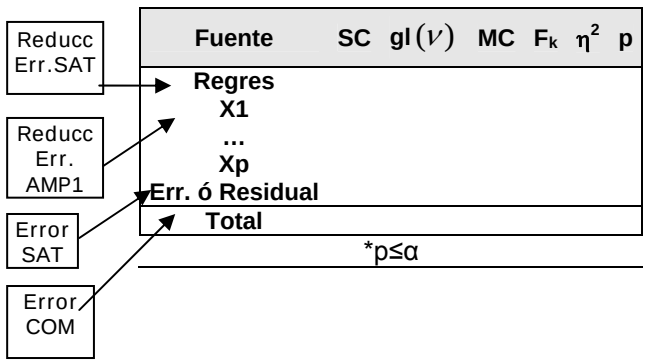
- Evaluar la significación de cada uno de los predictores a través de su pendiente asociada:

$$\left\{\begin{array}{c} AMP: Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_{p-1} XP-1_i + \beta_p XP_i + \varepsilon_i \\ COM1: Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \beta_{p-1} XP-1_i + \varepsilon_i \end{array}\right\} \equiv \left\{\begin{array}{l} \mathrm{H}_0 : \beta_P = 0 \\ \mathrm{H}_1 : \beta_P \neq 0 \end{array}\right\}$$

- La correlación que interviene en la estimación del parámetro es básicamente una **correlación semiparcial** en la que se controla el influjo del resto de predictores secundarios.

> - Informa del influjo de una variable relevante sobre el predictor objetivo de manera **selectiva,** asociación o interrelación entre los predictores en el modelo general. $r_{(X1-X1')\bullet Y}$.
> - Cuánto se incrementa la correlación múltiple al añadir una variable predictora en la ecuación de regresión, o, de otra manera, la correlación semiparcial de esa variable añadida con el criterio, parcializando el influjo sobre dicha variable objetivo de los otros predictores que ya estaban incluidos en el modelo.

- RPE (η^2) equivale directamente al coeficiente R^2



$$F_k = \frac{MCR}{MC_\varepsilon} \equiv \frac{RPE/gl}{(1-RPE)/gl}$$

$$\eta^2 = \frac{SCR}{SCE(COM)}$$

$$\eta_1^2 = \frac{SCR1}{SCE(COM1)}$$

> ➢ Supuesto M2.1
> Analizar → Regresión lineal → Dependiente: Ex.Card; Independientes: Hostilidad, Estres; Estadísticos→ Estimaciones → Continuar → Aceptar
> Statistics→ Multiple Regression →Variables: Dependent: ExCard; Independent: Hostilidad, Estres → OK→ Aceptar→*Pestaña* Advanced → Summary→ ANOVA → *Pestaña* Residuals

Conceptos destacados:

- **Redundancia.** La correlación R_p^2 es la medida RPE obtenida cuando se emplea a todos los predictores p-1 restantes en la predicción del predictor focal p, a modo de asociación entre predictores.

> - A veces su complementaria: medida de **tolerancia**, lo que es único para Xp en la predicción. Si la tolerancia asociada a un predictor Xp es baja entonces Xp será poco útil en la predicción. $1-R_p^2$
> - Incluso la inversa de la tolerancia, exactamente lo que entra en el intervalo de confianza, recibe un nombre: el factor de **inflación** de la varianza (*VIF: Variance Inflation Factor*). $\frac{1}{(1-R_p^2)}$

A) Regresión simultánea (estándar)

- Todas las variables independientes se introducen a la vez en la ecuación de predicción.

B) Regresión múltiple secuencial o jerárquico

- Se introducen en la ecuación **de modo progresivo** (por bloques) de acuerdo con un criterio teórico especificado por el investigador.
- Puesto que en ella ya hay algún predictor presente, lo que se evalúa es la aportación que hace la nueva variable a la predicción.
- Variantes:
  - Introducir (Regresión). Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso.
  - Borrar. Procedimiento de selección de variables en el que todas las variables de un bloque se eliminan en un solo paso.

C) Regresión múltiple o paso a paso

- El orden en que entran las variables en la ecuación de predicción puede establecerse por criterios estadísticos en lugar de teóricos o lógicos.
- El método de regresión interactivo "regresión paso a paso" que va incorporando ("forward" o hacia delante) o eliminando ("backward" o hacia atrás) sucesivamente variables, o una mezcla de ambos eliminando-introduciendo de manera convergente (***stepwise regression o por etapas***).
- El objetivo general es explicar un porcentaje de varianza del criterio similar al explicado por el total de predictores.
  - Se fija un nivel de significación, lo que impone un umbral de inclusión de variables.
  - En el método incremental, se calculan las correlaciones de todos los predictores con el criterio y se selecciona la variable con mayor correlación, siempre que supera el umbral de inclusión (F-entrar).
  - A continuación se elige el siguiente mejor predictor pero según la correlación parcial para controlar la influencia del predictor que ya estaba en el modelo y siempre que vuelva a superar el umbral.
  - Así sigue el procedimiento hasta que el incremento en correlación múltiple deja de ser significativo, es decir no sobrepasa el umbral.
  - La otra variante opera a la inversa (según F-salir).
  - También se puede plantear por bloques o secuencial (variante anterior).

➢ Supuesto M2.3

Analizar → Regresión → lineal → Dependiente: Y; Independientes: X1, X2, X3, X4; → Método: Introducir → Estadísticos: Matriz de covarianzas; Correlaciones Parcial y Semiparcial → Continuar → Aceptar

---

Statistics→ Multiple Regression→ *Pestaña* Advanced → Variables: Dependent: Y; Independent: X1, X2, X3, X4; Advanced Options & Review descriptive → OK→ OK Method: Standard

➢ Supuesto M2.3

Analizar → Regresión → lineal → Dependiente: Y; Independientes: X3, X4; → Bloque Siguiente → X1, X2 → Aceptar

---

Statistics→ Multiple Regression→ ... Method: Standard (forward stepwise/ backward stepwise)

➢ Supuesto M2.3

Analizar → Regresión → Lineal → Independiente: Y → Dependientes: X1, X2, X3, X4 → Método: Hacia delante (Hacia atrás/Pasos sucesivos) → Opciones → Probabilidad de F: Entrada: 0.05; Salida: 0.10 → Continuar → Aceptar

---

Statistics→ Multiple Regression→ ... Method: Standard (forward stepwise/ backward stepwise)

Interpretación:

Problemas

- Si los predictores son redundantes (recordar los conceptos asociados como tolerancia o tasa de Inflación), entonces el algoritmo implementado por algunos programas especializados no lleva a modelos realmente óptimos.
- Además, la interpretación del modelo resultante puede ser difícil. Siempre es preferible realizar un análisis guiado por hipótesis de investigación que doten de sentido a los resultados del análisis estadístico. Los criterios estadísticos pueden promover que diferencias mínimas entre variables pueden llevar a excluir a unas a favor de otras.

➢ F-entrar < F-salir, para que la variable que acaba de entrar no sea inmediatamente eliminada de la ecuación.
  - Para introducir muchas variables en la ecuación: incrementar la probabilidad (i.e. 0.10 mejor que 0.05).
  - Para ser exigente con las variables que queden en la ecuación: probabilidades de salida bajas (i.e. 0.05 en lugar de 0.10).

➢ Comparación métodos:
  - Simultáneo. para realizar predicciones/ comprobar correlaciones.
  - Por etapas. Si el objetivo es la construcción de un modelo de regresión, tanto más interesante cuanto mayor número de variables y si no tenemos modelos teóricos que guíen el análisis.
  - Secuencial. Si hemos medido variables independientes por razones teóricas, que no está orientado a la construcción de modelos, sino a su contraste.

➢ Si la investigación incluye muchos predictores estará claramente enfocada desde el punto de vista correlacional/covariacional y será preferible realizar los análisis dentro de la perspectiva especializada de "**análisis causal**", en la que se corrige el problema de "colinealidad".

➢ Recordar la posibilidad de incluir la interacción en el modelo o bien Modelos Polinómicos para Relaciones curvilíneas.

Para evaluar la **interacción o efecto conjunto** de los dos predictores:

$$\left\{\begin{array}{c} SAT: Y_i = \beta_0 + \beta_1 \bullet X1_i + \beta_2 \bullet X2_i + \beta_3 \bullet X1 \bullet X2_i + \varepsilon_i \\ COM: Y_i = \beta_0 + \beta_1 \bullet X1_i + \beta_2 \bullet X2_i + \varepsilon_i \end{array}\right\}$$

➢ Supuesto M2.1

- Primero Creamos la interacción: Transformar → Calcular → Interacc = Hostilidad * Estres; Aceptar.
- Entonces Análisis regresión: Analizar → Regresión lineal → Dependiente: Ex.Card; Independientes: Hostilidad, Estres, Interacc,; Estadísticos→ Estimaciones → Continuar → Aceptar
- Analizar … Estadísticos→ Estimaciones, Intervalos de confianza, Correlaciones parcial y semiparcial, Diagnósticos Colinealidad …

- Statistics→ Advanced Linear/NonLinear→General Linear Models→ Factorial Regression →OK→ Variables: Dependent: Ex.Card; Predictor: Hostilidad, Estres → OK→ Aceptar→*Pestaña* Summary→ Coefficients →*Pestaña* Advanced → Summary→ ANOVA → *Pestaña* Residuals → *Pestaña* Advanced→Partial Correlations & Redundancy & Current sweep matrix.

Modelos **Polinómicos**:

Cuadrática u orden-2:

$$\left\{\begin{array}{l} AMP: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ COM: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_3 X_i^3 \end{array}\right\} \equiv \left\{\begin{array}{l} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{array}\right\}$$

Cúbica u orden-3:

$$\left\{\begin{array}{l} AMP: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ COM: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \end{array}\right\} \equiv \left\{\begin{array}{l} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{array}\right\}$$

➢ Supuesto M2.2

Analizar → Regresión → Estimación curvilínea → Dependientes: Y; Independiente: X1; Modelos: Lineal, Cuadrático, Cúbico → Aceptar.

Statistics→ Advanced Linear/NonLinear→General Linear Models→ Polynomial Regression→ *Una vez definidas las variables, el Botón* Between Effects *permite ampliar la complejidad del modelo polinómico.*

9.2.3. Análisis de correlación canónica

1) ¿Cuándo se puede emplear la técnica?

- Para ajustar modelos de regresión lineal pero entre dos conjuntos de variables más que entre variables aisladamente.
- Usualmente servirá para predecir un conjunto a partir del otro aunque en origen los dos conjuntos de variables tienen el mismo estatus. Cuando se hayan medido a varios sujetos dos conjuntos de variables (métricas todas ellas) diferentes, e interese determinar si los conjuntos correlacionan.

2) Bases conceptuales de la técnica

- Se logra computando un coeficiente de correlación entre sendos conjuntos.
- Se combinan linealmente (se suman) las variables de cada conjunto por separado según una suma ponderada y normalizándolas para evitar las diferencias dadas por la escala. Cada una de las sumas es el variado canónico del conjunto. Los pesos se estiman para maximizar la correlación entre los dos conjuntos.
- Por tanto, la correlación canónica es **la correlación lineal entre dos variados canónicos**.
- Como vimos en las bases de análisis Multivariado, habrá **más de un coeficiente** de correlación canónico pues hay más de un variado normalmente. Pero únicamente nos quedaremos con los que resultan significativos.

- Podremos construir, todas las matrices de correlación y combinarlas en la supermatriz **M**

$M = \begin{pmatrix} R_{YY} & R_{YX} \\ R_{XY} & R_{XX} \end{pmatrix}$	**R_{YY}** entre las variables de uno de los dos conjuntos.
	R_{XX} entre las variables del otro de los dos conjuntos.
	R_{YX}, y su complementaria **R_{XY}**, que es entre sendos conjuntos.

- A partir de M, se invierten R_{YY} y R_{XX} para obtener una matriz R:

$$R = R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} ,$$

- Por lo demás se aplican las bases de ajuste multivariado, para obtener los valores propios de R, y entonces los vectores propios a partir de la ecuación determinantal:

$$\left| R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY} - \rho^2 I \right| = 0 .$$

➢ Supuesto M2.4 En SPSS, en el editor de sintaxis ejecutamos la orden: manova E D with P S /discrim all alpha (1) /print=sig(eigen dim)
Statistics → Multivariate Exploratory Technics → Canonical Analysis → Variables: ALL → Ok → Variables for canonical analysis → First variable list: E, D; Second variable list: P, S –> Ok →Ok→ Canonical factors → Eigenvalues → Canonical Scores → Left & right set canonical weights.

Interpretación:

Aunque las correlaciones canónicas constituyen el cálculo más destacado, hay otras estimaciones de interés.

A) Las correlaciones canónicas se obtienen a partir de la raíz cuadrada de los valores propios y el porcentaje de varianza compartida a partir de la traducción en porcentajes de los mismos.

- La significación de cualquier coeficiente de correlación canónica puede contrastarse mediante el test de Bartlett,

$$\Lambda = \prod_{i=1}^{s}(1-\lambda_i)$$

- que se aproxima a una chi-cuadrado según:

$$\chi^2_{ab} = -\left[N-1-\left(\frac{a+b+1}{2}\right)\right]\ln\Lambda;\ {}_{\alpha}\chi^2_{ab}$$

"a" y "b" son el número de variables en los conjuntos X e Y
N es el número total de sujetos.

B) Los pesos para combinar linealmente las variables de cada conjunto nos informarán de la importancia relativa de cada una. Para ello tendremos que operar en los dos sistemas siguientes:

$$\left(R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}-\rho^2 I\right)b=0\ ;\ \text{para los pesos del conjunto Y}$$
$$\left(R_{XX}^{-1}R_{XY}R_{YY}^{-1}R_{YX}-\rho^2 I\right)a=0\ ;\ \text{para los pesos del conjunto X}$$

Mejor las cargas canónicas (correlación entre cada variable y sus variados) para muestras pequeñas.

C) La **Varianza** de las variables observadas en cada conjunto por cada variado canónico.

$$p_Y=\sum_{i=1}^{b}\frac{c_{Li}^2}{b}\ \text{ y }\ p_X=\sum_{i=1}^{a}\frac{c_{Ri}^2}{a}\ ;\ \text{donde "c" son las cargas canónicas}$$

D) **La redundancia** viene dada por:

$$R_{Y/X}=\sum R_{Vi/Wi}$$
$$R_{Vi/Wi}=p_W\lambda_i$$

Es decir la suma de las redundancias del conjunto Y. Y cada redundancia se obtiene con el producto de la varianza por la raíz asociada al variado.

- Interpretación Supuesto M2.4 (tomado de Catena, Ramos y Trujillo (2003):

Valores propios

Raíces	Raíz 1	Raíz 2
Valor	0,833791	0,070008

Vectores de pesos

Variables	Raíz 1	Raíz 2	Variables	Raíz 1	Raíz 2
E	-0,9058	-0,7457	P	0,33479	1,0975
D	-0,1616	1,1621	S	-0,7923	0,8299

Cargas canónicas

Variables	Raíz 1	Raíz 2	Variables	Raíz 1	Raíz 2
E	-0,990	-0,138	P	0,723	0,690
D	-0,636	0,772	S	-0,956	0,292

3) Variantes principales de la técnica.

- En esta técnica sencilla no hay variantes pues las estimaciones son siempre de este tipo. Lo único que podría suceder es que se complicara el análisis por mayor número de variables en cada uno de los conjuntos.

4) Alternativas No Paramétricas y Robustas

- Análisis de regresión canónica no lineal ("OVERALS") mediante Escalamiento Óptimo en el menú:

Analizar → Reducción de datos → Escalamiento Óptimo

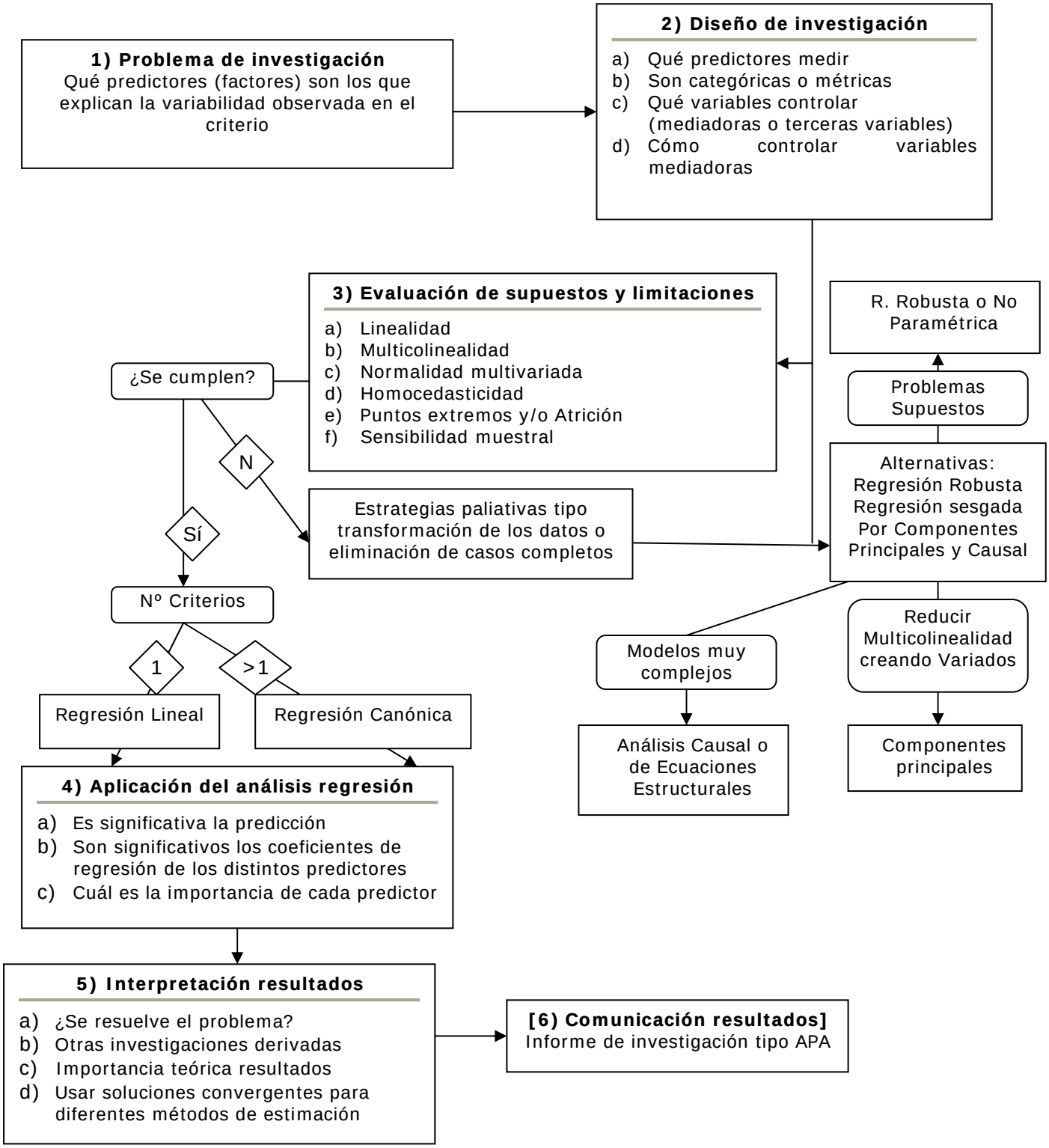
5) Limitaciones, supuestos y condiciones de aplicación

- Linealidad. Gráfico Predichos (abcisas) vs Errores de predicción (ordenadas).
- Multicolinealidad y Singularidad. El diagnóstico de la multicolinealidad se realiza a través de la tolerancia. Alternativas: regresión sesgada (ridge regresión) o regresión por componentes principales (los Variados son los predictores).
- Normalidad Multivariada. Gráficos de de probabilidad normal y Prueba de Kolmogorov-Smirnov o de Shapiro-Wilks. Alternativa No paramétrica basada en la prueba de Brown-Mood
- Problema de puntos extremos. Estadísticos de influencia indebida. Alternativa robusta basada en los MM-Estimadores de regresión
- Tamaño muestral: $n>50+8p$.

Para detalles ver el capítulo 6 (Apdo.4) del manual de Catena, Ramos y Trujillo (2003).

9.3. Secuencia de investigación en el Análisis de Regresión Multivariante

Para detalles ver el capítulo 6 (Apdo.5) del manual de Catena, Ramos y Trujillo (2003).



Volver Principio