UNIVERSIDAD DE JAÉN

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

Manuel Miguel Ramos Álvarez

MATERIAL IV "EXPLICACIÓN CON REGRESIÓN"

Índice

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

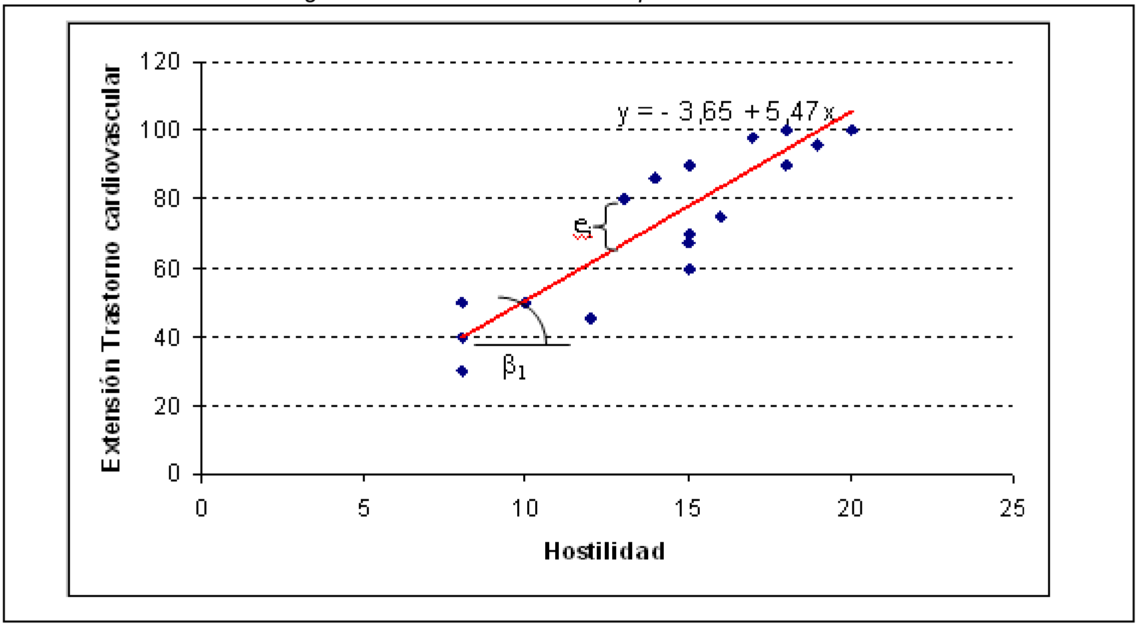
- 2 Aproximaciones:
  - Basada en el contraste de Hipótesis Estadísticas.
  - Basada en la potencia estadística y en los intervalos confidenciales.
- Tener presente el repaso sobre el contraste de Hipótesis y en general el módulo inicial sobre Modelización.

4.1. Análisis de Regresión para investigaciones correlacionales/covariacionales

4.1.1. Introducción a regresión

Interpretación básica a partir del Diagrama de dispersión

Figura adaptada a partir de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.



Bases de la estimación 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.
- El parámetro de intersección B0 corresponde al valor de $\widehat{Y}_i$ cuando X_i es cero ó punto de origen de la recta.
- La pendiente B1 cuantifica el cambio en $\widehat{Y}_i$ por cada incremento unitario en X_i.
- Positiva, lo que expresa un crecimiento en el criterio conforme aumenta el predictor
- negativa, expresando decrementos en el criterio correspondiendo a incrementos en el predictor.

- A partir de la Suma de Cuadrados Error, y teniendo en cuenta el modelo ampliado ó completo en comparación al restringido, podemos reconstruir el proceso de contrastación de Hipótesis tal y como vimos en el Modelo General.
- Proceso de análisis general.
- Se ajustan los modelos correspondientes a los datos con objeto de estimar los parámetros correspondientes
- Se estima la medida de Reducción Proporcional del Error (RPE) del modelo Ampliado en referencia a un modelo Compacto
- Entonces, La medida RPE y su complementaria, 1-RPE, se transforman en Medias Cuadráticas dividiendo por los grados de libertad correspondientes.
- El cociente entre ambas MMCC nos lleva a un estadístico F que nos proporciona información sobre lo que ganamos con el modelo Ampliado por parámetro añadido.
- Finalmente comparamos 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.

4.1.2. Análisis global de la regresión lineal

Tabla Resumen de la perspectiva de Modelización en el contexto de Regresión.

Cuadro 8.1 Adaptado a partir de de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.

Fuente	SC	gl (ν)	MC	F_k	η^2	p
Regres.	SCR= SCe(COM)- SCe(SAT)	1	$MCR = \frac{SCR}{1}$	$\frac{MCR}{MC\varepsilon}$*	$\frac{SCR}{SCE(COM)}$	$p(F_k)$
Err. o Residual	SCe(SAT)	N-2	$MC\varepsilon = \frac{SC(SAT)}{N-2}$			
Total	SCe(COM)	N-1				

*p≤α

Supuesto 1
Analizar → Regresión → lineal → Dependiente: FreqY; Independientes: X1; Estadísticos→ Estimaciones → Continuar → Aceptar

Statistics→ Multiple Regression →Variables: Dependent: FreqY; Independent: X1→ OK→ Aceptar→*Pestaña* Advanced → Summary→ ANOVA → *Pestaña* Residuals

Interpretación:

Análisis de los parámetros

- Si deseamos probar la significación del parámetro B_0 entonces, según la perspectiva de modelización tendríamos que comparar los modelos:

$$\left\{\begin{matrix} AMP: Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \\ COM1: Y_i = \beta_1 X_i + \varepsilon_i \end{matrix}\right\} \equiv \left\{\begin{matrix} \mathrm{H}_0: \beta_0 = 0 \\ \mathrm{H}_1: \beta_0 \neq 0 \end{matrix}\right\}$$

- Para probar la significación del parámetro B_1 entonces compararíamos los modelos:

$$\left\{\begin{matrix} AMP: Y_i = \beta_0 + \beta_1 \bullet X_i + \varepsilon_i \\ COM2: Y_i = \beta_0 + \varepsilon_i \end{matrix}\right\} \equiv \left\{\begin{matrix} \mathrm{H}_0: \beta_1 = 0 \\ \mathrm{H}_1: \beta_1 \neq 0 \end{matrix}\right\}$$

4.1.3. Resumen del Modelo

Los Intervalos Cofidenciales:

- Intersección. $\beta_0 \pm \sqrt{_\alpha F_{1;n-2}}\sqrt{\frac{MC_\varepsilon}{SC_X}}\sqrt{\frac{\sum X^2}{n}}$ en escala directa ó $\beta_0 \pm \sqrt{_\alpha F_{1;n-2}}\sqrt{\frac{MC_\varepsilon}{n}}$ en escala diferencial respecto a la media del predictor.
- Pendiente. $\beta_1 \pm \sqrt{_\alpha F_{1;n-2}}\sqrt{\frac{MC_\varepsilon}{SC_X}}$

Donde $SC_X = \sum(X-\overline{X})^2$

- 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.

Mejorar la interpretación de las predicciones

- En ocasiones interesa cambiar la escala de la ecuación de regresión, básicamente refiriendo todos los puntos con respecto al promedio de las variables. $\widehat{Y} = \beta_0^* + \beta_1(X_i - \overline{X})$.
- Donde $\beta_0^* = \overline{Y}$. En otras palabras, el parámetro de origen recoge ahora la coordenada $\left(\overline{X},\overline{Y}\right)$.

INTERVALOS CONFIDENCIALES
- ➢ Supuesto 1

Analizar → Regresión → lineal → Dependiente: FreqY; Independientes: X1; Estadísticos→ Estimaciones, Intervalos de confianza → Continuar → Aceptar

Statistics→ Advanced Linear/NonLinear→General Linear Models→ General Linear Models →OK→ Variables: Dependent: FreqY; Continuous pred: X1→ OK→ Aceptar→*Pestaña* Summary→ Coefficients

POTENCIA
- ➢ Supuesto 1

Statistics→ Power Analysis→ Power Calculation → One Correlation, t-Test→ OK → Rho: 0,44 *(es* R^2_{Adj} *)*; N: 18 Alpha: 0,05→ OK→ Calculate Power → start N: 10; End N: 100 → Power vs. N; Power vs. Rho; Power vs. Alpha

Opciones del programa Statistica:
- Power Calculation. Cálculo de la potencia y Funciones de Potencia para estimar la potencia a partir del tamaño del efecto, alfa y tamaño muestral.
- Sample Size Calculation. Cálculo del tamaño muestral a requerido para lograr un determinado nivel de potencia y también en function del resto de parámetros.
- Interval Estimation. Estimación por Intervalos a partir de variantes analíticas especializadas que no suelen aparecer en los programas de análisis convencionales.
- Probability Distributions. Modelos No centralizados que están implicados en las estimaciones de Potencia y del tamaño muestral.

Interpretación:

Interpretación:

4.2. El análisis de múltiples variables predictoras cuantitativas o perspectiva de Regresión Múltiple

- Evaluar la significación de cada uno de los predictores a través de su pendiente asociada y se corresponde con la vertiente condicional de modelización.

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

- 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. En definitiva, para un modelo de regresión múltiple con por ejemplo dos predictores,

Extensión de la Tabla Resumen de la perspectiva de Modelización a Regresión Múltiple

Cuadro 8.6 Adaptado a partir de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.

Fuente	SC	gl (ν)	MC	F_k	η^2	p
Regres	SCR= SCe(COM)-SCe(SAT)	p	$MC = \frac{SC}{gl}$	$\frac{MCR}{MC\varepsilon}$*	$\frac{SCR}{SCE(COM)}$	$p(F_k)$
X1	SCR1= SCe(COM1)-SCe(SAT)	1		$\frac{MCR1}{MC\varepsilon}$*	$\frac{SCR1}{SCE(COM1)}$	$p(F_k)$
...	...	...		...	...	...
Xp	SCRp= SCe(COMp)-SCe(SAT)	1		$\frac{MCRp}{MC\varepsilon}$*	$\frac{SCRp}{SCE(COMp)}$	$p(F_k)$
Err. ó Residual	SCe(SAT)	N-(p+1)				
Total	SCe(COM)	N-1				

*p≤α

- RPE equivale directamente al coeficiente R^2

Supuesto 1
Analizar → Regresión→ lineal → Dependiente: Y; Independientes: X1, X2, X3, X4; Estadísticos→ Estimaciones → Continuar → Aceptar

Statistics→ Multiple Regression →Variables: Dependent: FreqY; Independent: X1-X4 → OK→ Aceptar→*Pestaña* Advanced → Summary→ ANOVA → *Pestaña* Residuals.

Interpretación: (evitar comparaciones entre variables a partir de los parámetros B, mejor a partir de RPE).

4.2.1. Resumen del Modelo

- La correlación global ó **múltiple** ahora expresa un índice general de relación entre el conjunto de predictores y el criterio, que por ejemplo para dos predictores se puede calcular mediante la fórmula:

$$R_{Y\bullet 12} = \sqrt{\frac{r_{Y1}^2 + r_{Y2}^2 - 2 \cdot r_{Y1} \cdot r_{Y2} \cdot r_{12}}{1 - r_{12}^2}}$$

- Nuevamente la regresión general llevará asociado un **Error Típico** de Estimación, pero basado ahora en la correlación múltiple:

$$S_{Y\bullet X} = S_Y \cdot \sqrt{1 - R^2}$$

- La Correlación **parcial** que controla el influjo de una variable relevante sobre el predictor focal y sobre el criterio de manera simultánea:
- Similar al anterior, la Correlación **Semiparcial** controla el influjo de una variable relevante sobre el predictor objetivo de manera selectiva.

Estimación de intervalos confidenciales, también son válidas las fórmulas de regresión simple pero incluyendo una medida de redundancia. En general, para la pendiente de cada predictor p, la ecuación es la siguiente:

$$\beta_p \pm \sqrt{{}_{\alpha}F_{1;n-2}}\sqrt{MC_{\varepsilon}}\sqrt{\frac{1}{SC_{Xp}}}\sqrt{\frac{1}{(1 - R_{p.1\ldots p-1}^2)}}$$

- La correlación $R_{p.1\ldots p-1}^2$ que abreviaremos en adelante como R_p^2 y 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. Una medida de **redundancia**
- A veces se expresa, su complementaria: medida de **tolerancia**, lo que es único para Xp en la predicción.
- Incluso la inversa de la tolerancia, exactamente lo que entra en el intervalo, recibe un nombre: el factor de **inflación** de la varianza (*VIF: Variance Inflation Factor*).

Respecto a la **estimación de la potencia**, bastaría intercambiar las estimaciones del efecto de tratamiento propias de regresión múltiple con las que aparecían dentro del planteamiento de regresión simple

Analizar … Estadísticos→ Estimaciones, Intervalos de confianza, Correlaciones parcial y semiparcial, Diagnósticos Colinealidad …

Statistics … *Pestaña* Advanced→Partial Correlations & Redundancy & Current sweep matrix.

Para los Intervalos Confidenciales y la Potencia seguir las indicaciones de Regresión Simple

Interpretación:

4.2.2. La especificación de la interacción en el Modelo

Supongamos una investigación con dos predictores (X1 y X2) de un criterio.

- Para evaluar el efecto principal/aditivo de la variable X1 de manera independiente:

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

- Para el efecto principal de la segunda variable predictora, X2:

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

- Y finalmente, para evaluar la interacción o efecto conjunto de los dos predictores:

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

➢ **Como consecuencia, volveríamos a replantear el análisis de regresión de manera que el modelo final incluyera exclusivamente los parámetros que son significativos.**

➢ Primero Creamos nosotros la interacción: Transformar → Calcular → Interacc = X1 * X2; Aceptar. • Entonces Análisis regresión: Analizar → Regresión lineal → Dependiente: FreqY; Independientes: X1, X2, 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: FreqY; Predictor: X1, X2→ OK→ Aceptar→*Pestaña* Summary→ Coefficients →*Pestaña* Advanced → Summary→ ANOVA → *Pestaña* Residuals → *Pestaña* Advanced→Partial Correlations & Redundancy & Current sweep matrix. ➢ Alternativamente se puede hacer a través del módulo convencional Multiple Regression si creamos nosotros la variable interacción (Añadir nueva variable “Interacc“ y Pulsar sobre su Nombre e incluir la fórmula “=X1*X2” en la ventana Functions).	Interpretación: ¿Se gana en ajuste R^2 al cambiar a un modelo más complejo?

4.2.3. Análisis detallado mediante regresión. Las tendencias curvilíneas

- Supongamos que nos interesase evaluar una tendencia más compleja, como por ejemplo de orden-3 ó cúbica. El análisis mediante modelización implicaría, entonces los siguientes pasos:

Perspectiva global: $\left\{\begin{array}{l} AMP: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ COM: \widehat{Y}_i = \beta_0 \end{array}\right\} \equiv \left\{\begin{array}{l} \mathrm{H}_0: R^2 = 0 \\ \mathrm{H}_1: R^2 \neq 0 \end{array}\right\}$

Perspectiva Condicional:

Lineal u orden-1 $\left\{\begin{array}{l} AMP: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ COM: \widehat{Y}_i = \beta_0 + \beta_2 X_i^2 + \beta_3 X_i^3 \end{array}\right\} \equiv \left\{\begin{array}{l} \mathrm{H}_0: \beta_1 = 0 \\ \mathrm{H}_1: \beta_1 \neq 0 \end{array}\right\}$

Cuadrática u orden-2 $\left\{\begin{array}{l} AMP: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ COM: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_3 X_i^3 \end{array}\right\} \equiv \left\{\begin{array}{l} \mathrm{H}_0: \beta_2 = 0 \\ \mathrm{H}_1: \beta_2 \neq 0 \end{array}\right\}$

Cúbica u orden-3: $\left\{\begin{array}{l} AMP: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \\ COM: \widehat{Y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \end{array}\right\} \equiv \left\{\begin{array}{l} \mathrm{H}_0: \beta_3 = 0 \\ \mathrm{H}_1: \beta_3 \neq 0 \end{array}\right\}$

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.*

Interpretación: ¿Se gana en ajuste R² al cambiar a un modelo más complejo? Tener presentes indicaciones sobre especificidad de la Hipótesis (crear los modelos mediante opción de Datos, Calcular).

Conclusión: Probar con los diferentes Modelos, omitiendo predictores no significativos o problemáticos hasta dejar un Modelo Final.

Interpretación: Especifique el modelo final:

4.2.4. El caso general: análisis de regresión de modelos complejos

- En concreto, se usa un método de regresión interactivo “regresión paso a paso” (***stepwise regression***) que va incorporando (“forward”) o eliminando (“backward”) sucesivamente variables.
- 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.
- A continuación se elige el siguiente mejor predictor pero según la correlación semiparcial 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.

- El problema es que 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.
- Si la investigación incluye muchos predictores estará claramente enfocada desde el punto de vista correlacional y será preferible realizar los análisis dentro de la perspectiva especializada de “**análisis causal**”, en la que se corrige el problema de “colinealidad”.

- Además de lo anterior, existe la posibilidad de plantear modelos complejos con interacciones y polinomios, lo que se analiza mediante Modelos de regresión de superficie (i.e. Response Surface Regression dentro del módulo GLM de Statistica).

Supuesto 1
Analizar → Regresión→ lineal → Dependiente: Y; Independientes: X1, X2, X3, X4; Método: Hacia adelante → Aceptar

Statistics→ Multiple Regression →Variables: Dependent: FreqY; Independent: X1-X4 → *Pestaña* Advanced → Marcar Advanced options (stepwise …) → OK→ Method: Forward stepwise → OK.

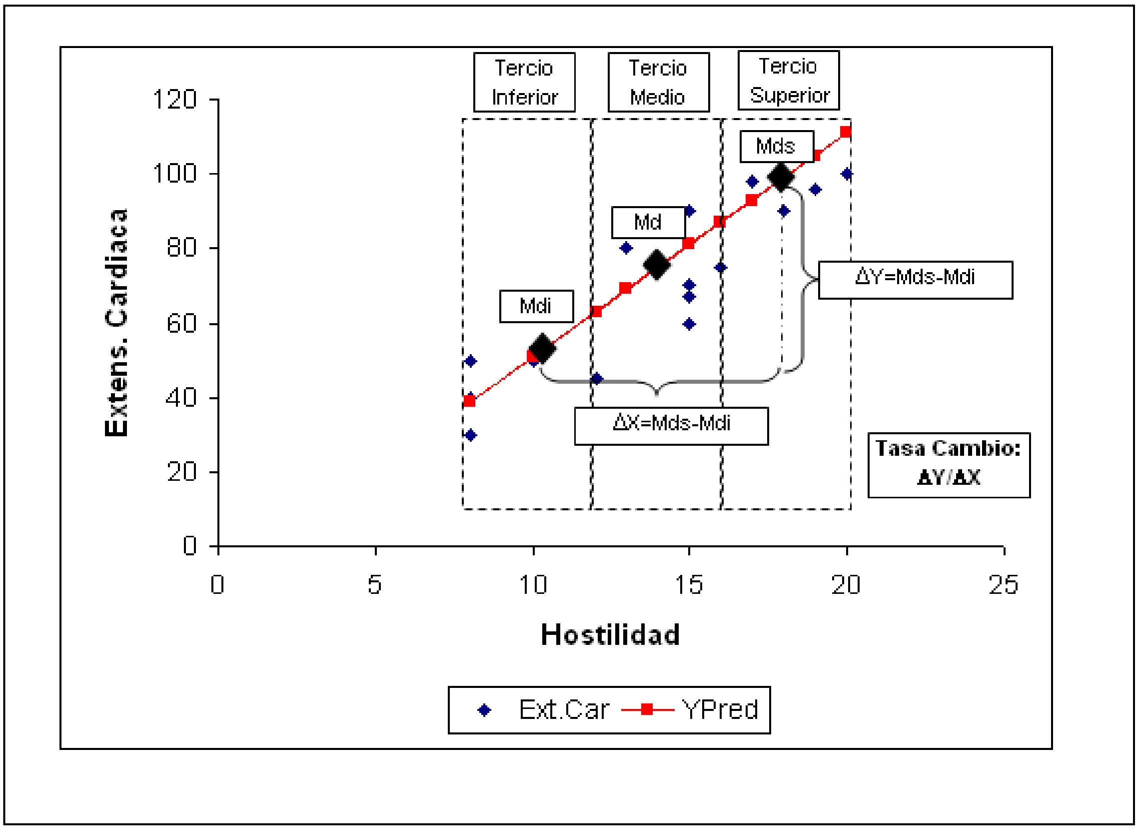
- Probar con otras variantes buscando convergencia, i.e. Hacia atrás (backward).

Interpretación:
Hacia delante y hacia atrás consideran las variables una a una.
En los métodos introducir y borrar se consideran bloques de variables.
Pasos sucesivos realiza convergentemente la introducción y la eliminación.

4.3. Alternativas robustas y No paramétricas de regresión

- ✓ Linea resistente de Tukey y Reajuste de los parámetros mediante un método iterativo de Emerson y Hoaglin (1985)

Figura 4-1: Interpretación gráfica del parámetro de tasa de cambio en regresión robusta



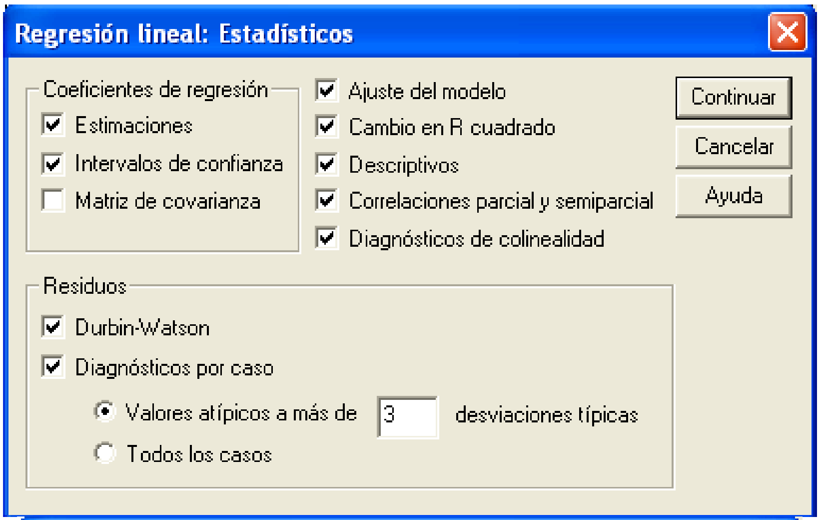
- ✓ Alternativa basada en los MM-Estimadores de regression
- ✓ Alternativa No paramétrica basada en la prueba de Brown-Mood

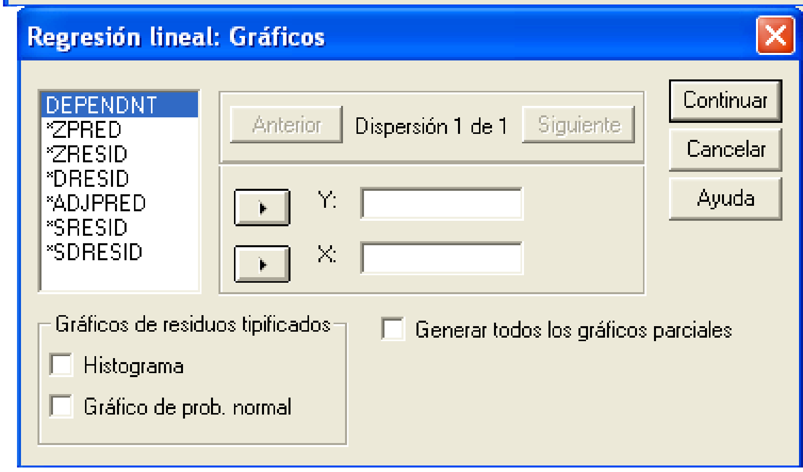
Ver para todas ellas el manual general recomendado y se ilustrarán en la última sesión.

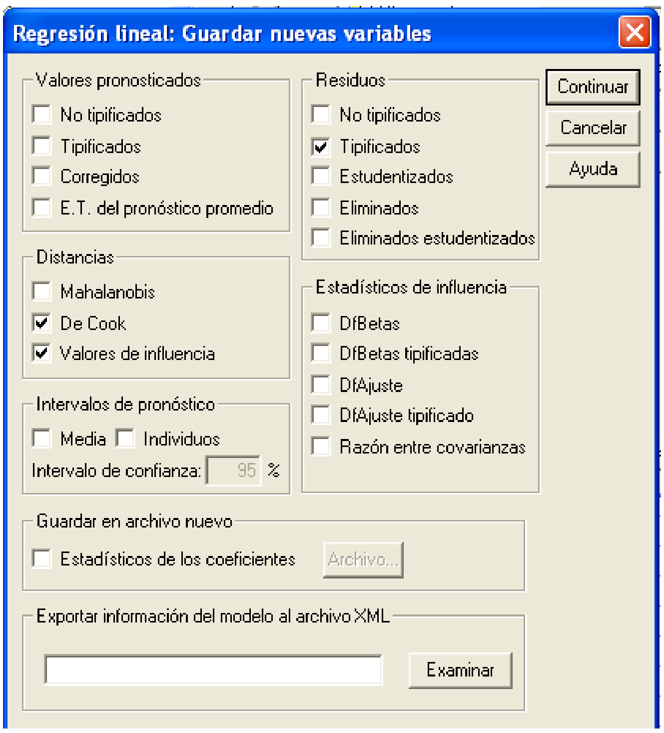
4.4. Opciones de Regresión Lineal en los paquetes de Análisis

- Probablemente, paquetes como SPSS tienen ventajas en cuanto al análisis de Modelos complejos a partir de la opción de regresión por pasos.
- En cambio, paquetes como Statistica son preferibles desde el punto de vista de la Modelización.
- Otros paquetes como S-Plus son la opción para Regresión Robusta.

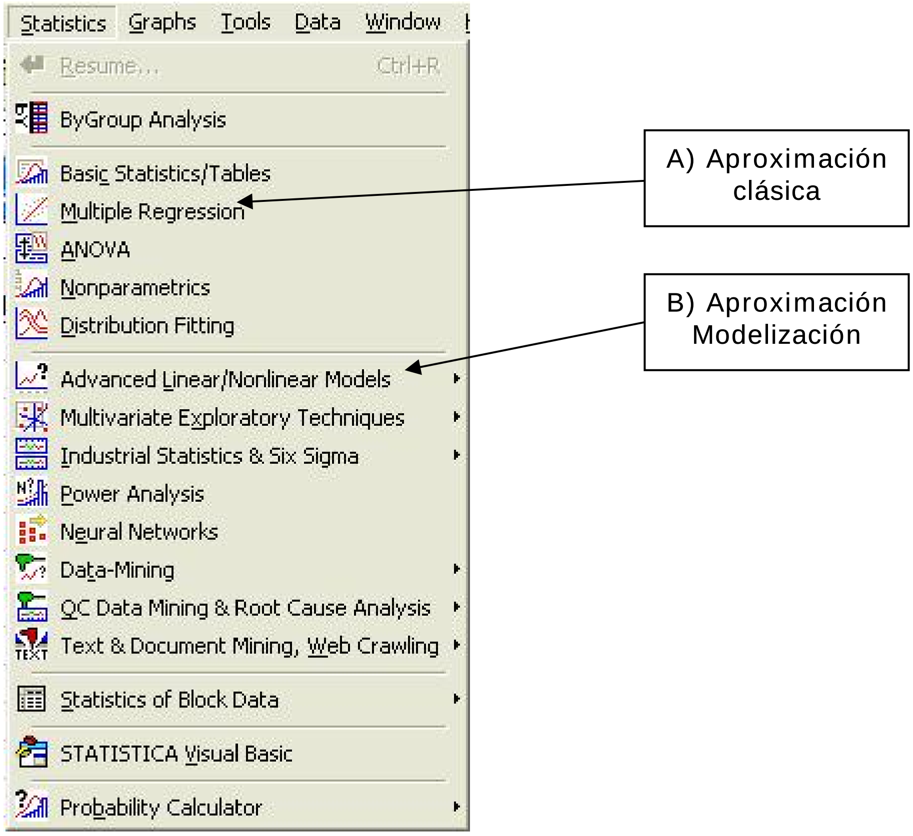
4.4.1. Opciones de Regresión Lineal en SPSS 12.0/15.0

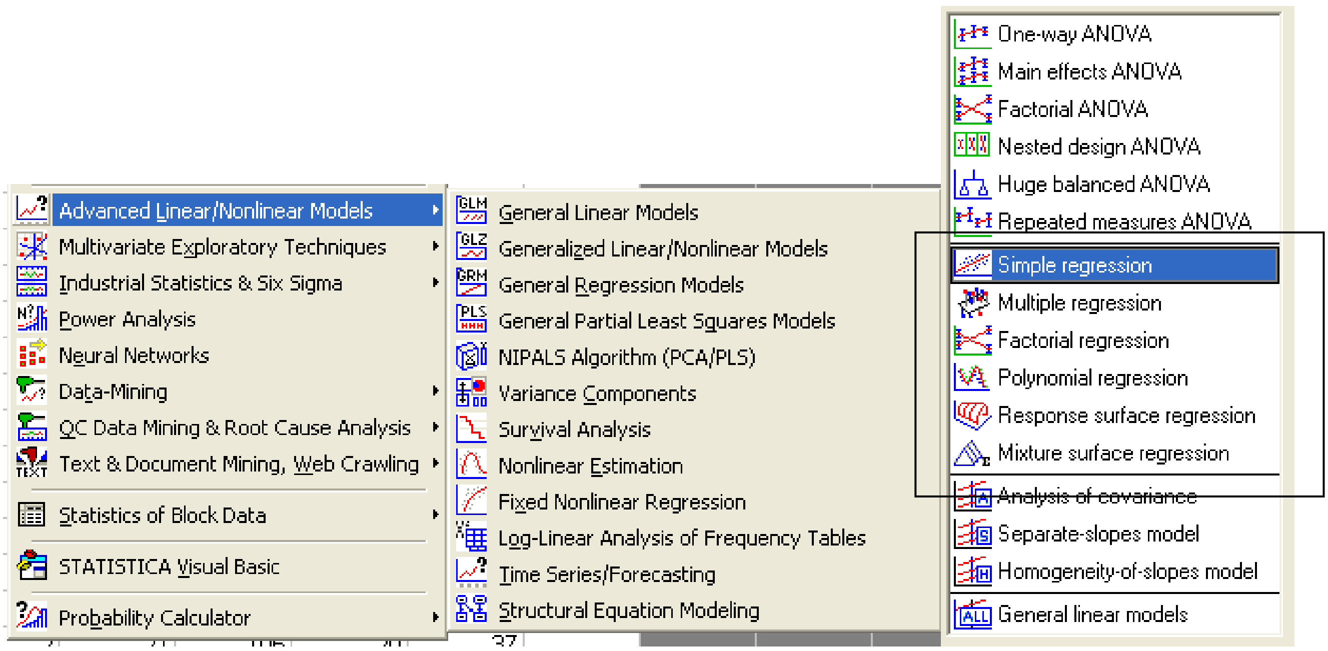




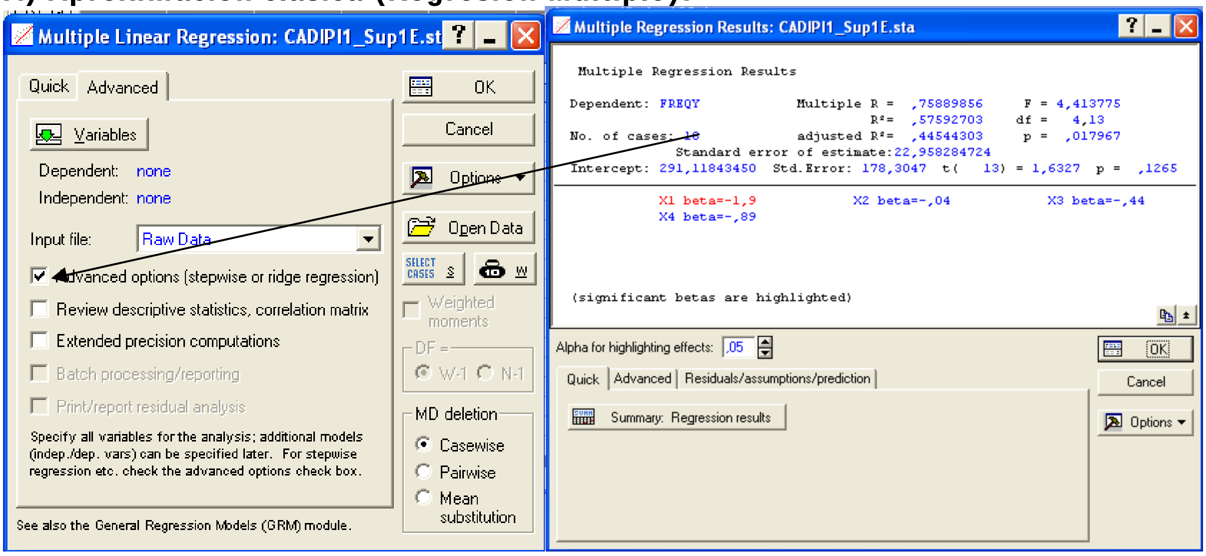


4.4.2. Opciones de Regresión Lineal en Statistica

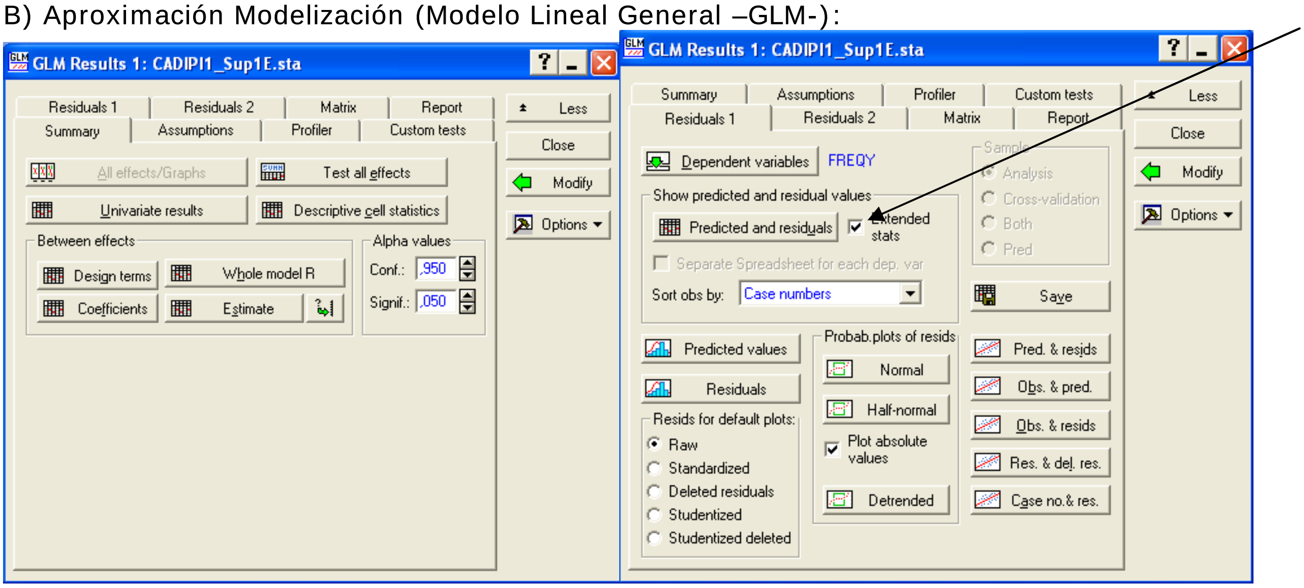




A) Aproximación clásica (Regresión Múltiple):



B) Aproximación Modelización (Modelo Lineal General –GLM-):



4.5. Realización de los supuestos de prácticas

Guía del análisis en Regresión

- Análisis exploratorio de los diagramas de dispersión
- Análisis del modelo SATURADO de referencia
- Análisis de los residuales (estandarizados) y de las distancias de influencia indebida para decidir sobre los posibles puntos extremos.
- Confirmar conclusiones de significación para cada predictor mediante regresión por pasos.
- Si gráficos de residuales sugieren funciones curvilíneas entonces pasar al análisis polinómico (regresión curvilínea).
- Incluir interacciones si están justificadas desde el punto de vista teórico.

4.5.1. Ejemplificación del análisis de Regresión mediante el Supuesto 1

[Gráficos|Dispersión…] o mejor [Gráficos|Interactivos|Dispersión …] para ajustar la linea de regresión lineal.
Los diagramas de dispersión de cada predictor frente al criterio

Análisis del modelo SATURADO de referencia, es decir el que incluye todos los predictores (X1 a X4) y de manera lineal.
[Analizar|Regresión lineal] Seleccionamos las opciones de Estadísticos, Gráficos y Guardar que aparecen en las ventanas precedentes…

Análisis de los residuales
Observar al menos los gráficos de RRESID frente a ADJPRED, incluyendo los gráficos parciales y el de probabilidad normal
Confirmarlo a partir de las Distancias de Cook.
El caso 8 parece tener un error estandarizado muy elevado pero puesto que está justificado, entonces no se omitiría

Confirmar las conclusiones de los parámetros significativos (únicamente la inversa de la variable predoctora X1) mediante regresión por pasos, por ejemplo introduciendo variables en pasos sucesivos.

Probar Modelo polinómico
[Analizar|Regresión|Estimación curvilínea …]
La pregunta clave es ¿Se gana en ajuste R^2 al cambiar a un modelo más complejo que el lineal?

Centrar el modelo significativo y resumirlo adecuadamente.

Volver Principio