Aproximación computacionalmente eficiente para un test de independencia en modelos de regresión no paramétrica

Gustavo Rivas Martínez[1], María Dolores Jiménez Gamero[2]

[1] Laboratorio de Sistemas de Potencia y Control. Universidad Nacional de Asunción-Paraguay

[2] Dpto. de Estadística e Investigación Operativa. Universidad de Sevilla-España

marzo-2019

Modelo

Sean $(X_1, Y_1), \ldots (X_n, Y_n)$ vectores aleatorios independientes e idénticamente distribuidos (iid) de (X, Y) que satisfacen el siguiente modelo de regresión no paramétrica,

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

donde
$m(X) = E(Y|X = x)$ es la función de regresión,
$\sigma^2(X) = Var(Y|X = x)$ es la varianza condicional,
ε es el error aleatorio.

Hipótesis

Cuando la independencia entre X y ε no se cumple, los métodos estadísticos basados en este supuesto no son válidos. Por esa razón, muchos autores han propuesto distintos tests para la siguiente Hipótesis nula

$$H_0 : X \text{ and } \varepsilon \text{ son independientes} \Leftrightarrow F_{X,\varepsilon} = F_X F_\varepsilon.$$

donde $F_{X,\varepsilon}$ es la función de distribución (fdd) conjunta de X y ε, y F_X y F_ε denotan la fdd de X y ε, respectivamente.

Trabajos anteriores

- Einmahl and Van Keilegom (2008) han propuesto tests estadísticos del tipo Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling
- Neumeyer (2009) propuso un test estadístico basado en un estimador tipo kernel para una distancia L_2 entre la función de distribución condicional y la no condicional de X
- Dhar et al. (2018) han propuesto una test basado en una modificación de la medida τ de asociación de Kendall

Trabajos anteriores

La H_0 es equivalente a

$$H_0 : c_{X,\varepsilon} = c_X c_\varepsilon.$$

donde $c_{X,\varepsilon}$ es la función característica (FC) conjunta de X y ε, y c_X y c_ε denotan las FCs marginal de X y ε, respectivamente. Hlávka et al. (2011) han propuesto el estadístico $T_{n,W}$, que está basado en una distancia L_2 entre una estimación de la FC conjunta (X, ε) y una estimación de las FCs marginal de X y ε.

Test

Para contrastar H_0, Hlávka et al. (2011) propusieron el test

$$\Psi = \begin{cases} \text{Rechazar la } H_0 & \text{si} \quad T_{n,W} \geq t_{n,W,\alpha}, \\ \text{No rechazar la } H_0 & \text{si} \quad T_{n,W} < t_{n,W,\alpha} \end{cases}$$

donde $t_{n,W,\alpha}$ es el percentil $1-\alpha$ de la distribución nula de $T_{n,W}$,

$$T_{n,W} = n \int \int |D_n(t,s)|^2 W(t,s) dt ds, \tag{2}$$

donde $D_n(t,s) = \hat{c}(t,s) - \hat{c}_X(t)\hat{c}_{\hat{\varepsilon}}(s)$, con $\hat{c}(t,s) = \frac{1}{n}\sum_{j=1}^{n} \exp(\mathrm{i}tX_j + \mathrm{i}s\hat{\varepsilon}_j)$, corresponde a la función característica empírica (FCE) conjunta de (X, ε), $\hat{c}_X(t) = \hat{c}(t,0)$, $\hat{c}_{\hat{\varepsilon}}(s) = \hat{c}(0,s)$ corresponde a las FCEs X y ε, respectivamente, y $W(t,s)$ es una función de peso.

La distribución nula de $T_{n,W}$ es desconocida. Como una primera aproximación en el Teorema 1 de Hlávka et al. (2011) obtienen su distribución nula asintótica

$$T_{n,W} \xrightarrow{\mathcal{L}} \int \int Z(t,s)^2 \omega(t)\omega(s)dtds,$$

donde $\{Z(t,s),\ (t,s) \in \mathbb{R}^2\}$ es un proceso gaussiano centrado con estructura de covarianza $U(X,\varepsilon;t,s) = \{\cos(s\varepsilon) - R_\varepsilon(s) + s\varepsilon I_\varepsilon(s)$
$-\frac{1}{2}s(\varepsilon^2-1)R'_\varepsilon(s)\}\{\cos(tX)+\sin(tX)-R_X(t)-I_X(t)\}$
$+\{\sin(s\varepsilon) - I_\varepsilon(s) - s\varepsilon R_\varepsilon(s)$
$-\frac{1}{2}s(\varepsilon^2-1)I'_\varepsilon(s)\}\{\cos(tX)-\sin(tX)-R_X(t)+I_X(t)\}$
donde $R_\varepsilon(s)$ y $I_\varepsilon(s)$ denotan la parte real e imaginaria la FC de ε, respectivamente, y $R_X(t)$ y $I_X(t)$ están definidas análogamente.

Problema

- Porque la distribución nula exacta y la asintótica son desconocidas, Hlávka et al. (2011) proponen un bootstrap para estimar el p-valor.

- El Bootstrap es, en general, de fácil implementación pero el coste computacional se incrementa notablemente a medida que el tamaño muestral aumenta.

Objetivo

Encontrar una aproximación a la distribución nula de $T_{n,W}$ que sea, desde un punto de vista computacional, más eficiente.

Metodología

Desarrollar un bootstrap ponderado (BP) en el contexto del problema abordado que sea consistente
Para variables observables,

- Los artículos de Fan et al. (2017) y Jiménez-Gamero et al. (2018) proponen un BP para aproximar la distribución nula de test estadísticos similares a $T_{n,W}$.

En nuestro caso, una de las variables de interés (ε) es no observable. Esto hace que el tratamiento teórico del desarrollo del BP sea distinto al caso de variables observables.

Para la estimación de $m(\cdot)$ y $\sigma^2(\cdot)$, se usan los siguientes estimadores núcleo de Nadaraya-Watson

$$\hat{m}(x) = \sum_{j=1}^{n} R_j(x;h)\, Y_j, \quad x \in S,$$

$$\hat{\sigma}^2(x) = \sum_{j=1}^{n} R_j(x;h) \left\{ Y_j - \hat{m}(x) \right\}^2, \quad x \in S,$$

donde

$$R_j(x;h) = \frac{K_h\left(X_j - x\right)}{\sum_{s=1}^{n} K_h\left(X_s - x\right)}, \quad x \in S,$$

Condiciones

(A.1*) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con $E(\varepsilon_j/X_j) = 0$, $E(\varepsilon_j^2/X_j) = 1$, $E(\varepsilon_j^4) < \infty$ y función característica $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.1) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con mediad cero y varianza unitaria, $E(\varepsilon_j^4) < \infty$ y función características $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.2) $X_1, \ldots, X_n$ son variables aleatorias iid que toman valores en un compacto S, con densidad común positiva y contínuamente diferenciable f_X y función característica $c_X(t)$, $t \in \mathbb{R}$.

(A.3) $\varepsilon_1, \ldots, \varepsilon_n$ y $X_1, \ldots, X_n$ son independientes.

Condiciones

(A.1*) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con $E(\varepsilon_j/X_j) = 0$, $E(\varepsilon_j^2/X_j) = 1$, $E(\varepsilon_j^4) < \infty$ y función característica $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.1) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con mediad cero y varianza unitaria, $E(\varepsilon_j^4) < \infty$ y función características $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.2) $X_1, \ldots, X_n$ son variables aleatorias iid que toman valores en un compacto S, con densidad común positiva y contínuamente diferenciable f_X y función característica $c_X(t)$, $t \in \mathbb{R}$.

(A.3) $\varepsilon_1, \ldots, \varepsilon_n$ y $X_1, \ldots, X_n$ son independientes.

Condiciones

(A.1*) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con $E(\varepsilon_j/X_j) = 0$, $E(\varepsilon_j^2/X_j) = 1$, $E(\varepsilon_j^4) < \infty$ y función característica $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.1) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con mediad cero y varianza unitaria, $E(\varepsilon_j^4) < \infty$ y función características $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.2) $X_1, \ldots, X_n$ son variables aleatorias iid que toman valores en un compacto S, con densidad común positiva y contínuamente diferenciable f_X y función característica $c_X(t)$, $t \in \mathbb{R}$.

(A.3) $\varepsilon_1, \ldots, \varepsilon_n$ y $X_1, \ldots, X_n$ son independientes.

Condiciones

(A.1*) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con $E(\varepsilon_j/X_j) = 0$, $E(\varepsilon_j^2/X_j) = 1$, $E(\varepsilon_j^4) < \infty$ y función característica $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.1) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias iid con mediad cero y varianza unitaria, $E(\varepsilon_j^4) < \infty$ y función características $c_\varepsilon(t)$, $t \in \mathbb{R}$.

(A.2) $X_1, \ldots, X_n$ son variables aleatorias iid que toman valores en un compacto S, con densidad común positiva y contínuamente diferenciable f_X y función característica $c_X(t)$, $t \in \mathbb{R}$.

(A.3) $\varepsilon_1, \ldots, \varepsilon_n$ y $X_1, \ldots, X_n$ son independientes.

Condiciones

(A.4) $m(x)$, $x \in S$, tiene primera derivada.

(A.5) $\sigma(x)$, $x \in S$, es positiva con primera dirivada.

(A.6) La función de peso $\omega(t)$, $t \in \mathbb{R}$, es simétrica, no negativa y tal que $\int t^4 \omega(t) dt < \infty$.

(A.7) K es simétrica y dos veces contínuamente diferenciable.

(A.8) $\{h_n\}$ es una secuencia de parámetros de ventana tal que $\lim_{n\to\infty} nh_n^2 = \infty$ y $\lim_{n\to\infty} nh_n^{3+\delta} = 0$, para algún $\delta > 0$.

Condiciones

(A.4) $m(x)$, $x \in S$, tiene primera derivada.

(A.5) $\sigma(x)$, $x \in S$, es positiva con primera dirivada.

(A.6) La función de peso $\omega(t)$, $t \in \mathbb{R}$, es simétrica, no negativa y tal que $\int t^4 \omega(t) dt < \infty$.

(A.7) K es simétrica y dos veces contínuamente diferenciable.

(A.8) $\{h_n\}$ es una secuencia de parámetros de ventana tal que $\lim_{n\to\infty} nh_n^2 = \infty$ y $\lim_{n\to\infty} nh_n^{3+\delta} = 0$, para algún $\delta > 0$.

Condiciones

(A.4) $m(x)$, $x \in S$, tiene primera derivada.

(A.5) $\sigma(x)$, $x \in S$, es positiva con primera dirivada.

(A.6) La función de peso $\omega(t)$, $t \in \mathbb{R}$, es simétrica, no negativa y tal que $\int t^4 \omega(t) dt < \infty$.

(A.7) K es simétrica y dos veces contínuamente diferenciable.

(A.8) $\{h_n\}$ es una secuencia de parámetros de ventana tal que $\lim_{n\to\infty} nh_n^2 = \infty$ y $\lim_{n\to\infty} nh_n^{3+\delta} = 0$, para algún $\delta > 0$.

Condiciones

(A.4) $m(x)$, $x \in S$, tiene primera derivada.

(A.5) $\sigma(x)$, $x \in S$, es positiva con primera dirivada.

(A.6) La función de peso $\omega(t)$, $t \in \mathbb{R}$, es simétrica, no negativa y tal que $\int t^4 \omega(t) dt < \infty$.

(A.7) K es simétrica y dos veces contínuamente diferenciable.

(A.8) $\{h_n\}$ es una secuencia de parámetros de ventana tal que $\lim_{n\to\infty} nh_n^2 = \infty$ y $\lim_{n\to\infty} nh_n^{3+\delta} = 0$, para algún $\delta > 0$.

Condiciones

(A.4) $m(x)$, $x \in S$, tiene primera derivada.

(A.5) $\sigma(x)$, $x \in S$, es positiva con primera dirivada.

(A.6) La función de peso $\omega(t)$, $t \in \mathbb{R}$, es simétrica, no negativa y tal que $\int t^4 \omega(t) dt < \infty$.

(A.7) K es simétrica y dos veces contínuamente diferenciable.

(A.8) $\{h_n\}$ es una secuencia de parámetros de ventana tal que $\lim_{n\to\infty} nh_n^2 = \infty$ y $\lim_{n\to\infty} nh_n^{3+\delta} = 0$, para algún $\delta > 0$.

Resultados previos

Si las condiciones (A.1)-(A.8) se cumplen, de la demostración del Teorema 1 de Hlávka et al. (2011), se deduce que

$$T_{n,W} = T_{1,n,W} + o_P(1), \tag{3}$$

con

$$T_{1,n,W} = \int \int Z_1(t,s)^2 \omega(t)\omega(s) dtds,$$

donde $Z_1(t,s) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} U(X_j, \varepsilon_j; t, s)$.

Sean $\xi_1, \ldots, \xi_n$ (multiplicadores) variables aleatorias iid con media 0 y varianza 1, independientes de $(X_1, Y_1), \ldots, (X_n, Y_n)$. La versión BP de $T_{n,W}$ se define así,

$$T^*_{1,n,W} = \int \int Z^*_1(t,s)^2 \omega(t)\omega(s) dtds,$$

con $Z^*_1(t,s) = \frac{1}{\sqrt{n}} \sum_{j=1}^n U(X_j, \varepsilon_j; t, s)\xi_j$.

Resultados previos

De (3) y de los resultados en Dehling and Mikosch (1994), se sigue que

$$\sup_x \left| P_* \left\{ T^*_{1,n,W} \leq x \right\} - P_0 \left\{ T_{n,W} \leq x \right\} \right| \xrightarrow{P} 0, \qquad (4)$$

Es decir, la distribución condicional de $T^*_{1,n,W}$, dado $(X_1, Y_1), \ldots, (X_n, Y_n)$, es un estimador consistente de la distribución nula de $T_{n,W}$.

$Z_1^*(t,s)$ depende de $\varepsilon_1, \ldots, \varepsilon_n$ (no observables), de $R_\varepsilon(t)$, $I_\varepsilon(t)$, $R_X(t)$ y $I_X(t)$, (desconocidos). Para solucionar este problema se reemplaza ε_j por $\hat{\varepsilon}_j$, $R_\varepsilon(s)$ por $\hat{R}_{\hat{\varepsilon}}(s)$, $I_\varepsilon(s)$ por $\hat{I}_{\hat{\varepsilon}}(s)$, $R'_\varepsilon(s)$ por $\hat{R}'_{\hat{\varepsilon}}(s)$, $I'_\varepsilon(s)$ por $\hat{I}'_{\hat{\varepsilon}}(s)$, $R_X(t)$ por $\hat{R}_X(t)$, $I_X(t)$ por $\hat{I}_X(t)$

Estimadores utilizados

$$\hat{\varepsilon}_j = \frac{Y_j - \hat{m}(X_j)}{\hat{\sigma}(X_j)},$$

$$\hat{R}_{\hat{\varepsilon}}(s) = \frac{1}{n}\sum_{j=1}^{n}\cos(s\hat{\varepsilon}_j), \quad \hat{I}_{\hat{\varepsilon}}(s) = \frac{1}{n}\sum_{j=1}^{n}\sin(s\hat{\varepsilon}_j)$$

$$\hat{R}_X(t) = \frac{1}{n}\sum_{j=1}^{n}\cos(tx_j), \quad \hat{I}_X(t) = \frac{1}{n}\sum_{j=1}^{n}\sin(tx_j)$$

$$\hat{R}'_{\hat{\varepsilon}}(s) = -\frac{1}{n}\sum_{j=1}^{n}\hat{\varepsilon}_j\sin(s\hat{\varepsilon}_j), \quad \hat{I}'_{\hat{\varepsilon}}(s) = \frac{1}{n}\sum_{j=1}^{n}\hat{\varepsilon}_j\cos(s\hat{\varepsilon}_j).$$

Anora consideramos

$$T^*_{n,W} = \int \int Z^*_{n,1}(t,s)^2 \omega(t)\omega(s)dtds,$$

donde $Z^*_{n,1}(t,s) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{U}(X_j, \hat{\varepsilon}_j; t, s)\xi_j$, con
$\hat{U}(X, \varepsilon; t, s) = \{\cos(s\varepsilon) - \hat{R}_{\hat{\varepsilon}}(s)s\varepsilon\hat{I}_{\hat{\varepsilon}}(s)$
$-\frac{1}{2}s(\varepsilon^2 - 1)\hat{R}'_{\hat{\varepsilon}}(s)\}\{\cos(tX) + \sin(tX) - \hat{R}_X(t) - \hat{I}_X(t)\}$
$+\{\sin(s\varepsilon) - \hat{I}_{\hat{\varepsilon}}(s) - s\varepsilon\hat{R}_{\hat{\varepsilon}}(s)$
$-\frac{1}{2}s(\varepsilon^2 - 1)\hat{I}'_{\hat{\varepsilon}}(s)\}\{\cos(tX) - \sin(tX) - \hat{R}_X(t) + \hat{I}_X(t)\},$

Teorema 1

Si las condiciones (A.1), (A.2) y (A.4)–(A.8) se cumplen, entonces

$$\sup_x \left| P_* \left\{ T^*_{n,W} \leq x \right\} - P_* \left\{ T^*_{1,n,W} \leq x \right\} \right| \xrightarrow{P} 0,$$

El resultado del Teorema 1 se cumple ya sea H_0 verdadera o no, de este hecho se desprenden dos consecuencias inmediatas.

Corolario 1

Si H_0 es verdadera y las condiciones del Teorema 1 se cumplen, entonces

$$\sup_x \left| P_* \left\{ T^*_{n,W} \leq x \right\} - P_0 \left\{ T_{n,W} \leq x \right\} \right| \xrightarrow{P} 0.$$

Sea $\alpha \in (0,1)$ y

$$\Psi_* = \begin{cases} 1, & \text{si} \quad T_{n,W} \geq t^*_{n,W,\alpha}, \\ 0, & \text{caso contrario,} \end{cases}$$

donde $t^*_{n,W,\alpha}$ es el percentil $1-\alpha$ de la distribución condicional de $T^*_{n,W}$.

Corolario 2

Si H_0 no es verdadera, las condiciones (A.1*),(A.2) y (A.4)–(A.8) se cumplen y ω es tal que

$$\kappa = \int \int \{C_{X,\varepsilon}(t,s) - C_X(t)C_\varepsilon(s)\}^2 \omega(t)\omega(s) > 0, \qquad (5)$$

entonces $P(\Psi_* = 1) \to 1$.

Observación: Los resultados del Teorema 1 y los Corolarios (1 y 2) siguen siendo válidos si en vez de los multiplicadores en bruto $\xi_1, ..., \xi_n$, se utiliza los multiplicadores centrados $\xi_1 - \overline{\xi}, ..., \xi_n - \overline{\xi}$.

Objetivos del experimento de simulación

- Estudiar la bondad del ajuste de la aproximación propuesta (PB) a la distribución nula del estadístico.
- Analizar la potencia de la aproximación propuesta con el Bootstrap.
- Comparar el tiempo de computo para el BP y el Bootstrap.

Objetivos del experimento de simulación

- Estudiar la bondad del ajuste de la aproximación propuesta (PB) a la distribución nula del estadístico.
- Analizar la potencia de la aproximación propuesta con el Bootstrap.
- Comparar el tiempo de computo para el BP y el Bootstrap.

Objetivos del experimento de simulación

- Estudiar la bondad del ajuste de la aproximación propuesta (PB) a la distribución nula del estadístico.
- Analizar la potencia de la aproximación propuesta con el Bootstrap.
- Comparar el tiempo de computo para el BP y el Bootstrap.

- H_0 : ϵ y X son independientes
- Kernel (Epanechnikov) $K(u) = 0.75 \times (1 - u^2)$
- Parámetro de ventana mediante validación cruzada
- $W_1(t,s) = \exp(-\lambda_{1,1}|t| - \lambda_{1,2}|s|)$ y
  $W_2(t,s) = \exp(-\lambda_{2,1}t^2 - \lambda_{2,2}s^2)$.
- $\xi \sim Normal(0,1)$ (Se muestran resultados para los centrados)
- El modelo que genera los datos es
  $Y_j = X_j - \frac{X_j^2}{2} + 0{,}1\varepsilon_j\sqrt{1+2X_j}$, $1 \leq j \leq n$,

- H_0 : ϵ y X son independientes
- Kernel (Epanechnikov) $K(u) = 0.75 \times (1 - u^2)$
- Parámetro de ventana mediante validación cruzada
- $W_1(t,s) = \exp(-\lambda_{1,1}|t| - \lambda_{1,2}|s|)$ y
  $W_2(t,s) = \exp(-\lambda_{2,1}t^2 - \lambda_{2,2}s^2)$.
- $\xi \sim Normal(0,1)$ (Se muestran resultados para los centrados)
- El modelo que genera los datos es
  $Y_j = X_j - \frac{X_j^2}{2} + 0{,}1\varepsilon_j\sqrt{1+2X_j}$, $1 \leq j \leq n$,

- H_0 : ϵ y X son independientes
- Kernel (Epanechnikov) $K(u) = 0.75 \times (1 - u^2)$
- Parámetro de ventana mediante validación cruzada
- $W_1(t,s) = \exp(-\lambda_{1,1}|t| - \lambda_{1,2}|s|)$ y
  $W_2(t,s) = \exp(-\lambda_{2,1}t^2 - \lambda_{2,2}s^2)$.
- $\xi \sim Normal(0,1)$ (Se muestran resultados para los centrados)
- El modelo que genera los datos es
  $Y_j = X_j - \frac{X_j^2}{2} + 0{,}1\varepsilon_j\sqrt{1+2X_j}$, $1 \leq j \leq n$,

- H_0 : ϵ y X son independientes
- Kernel (Epanechnikov) $K(u) = 0.75 \times (1 - u^2)$
- Parámetro de ventana mediante validación cruzada
- $W_1(t,s) = \exp(-\lambda_{1,1}|t| - \lambda_{1,2}|s|)$ y
  $W_2(t,s) = \exp(-\lambda_{2,1}t^2 - \lambda_{2,2}s^2)$.
- $\xi \sim Normal(0,1)$ (Se muestran resultados para los centrados)
- El modelo que genera los datos es
  $Y_j = X_j - \frac{X_j^2}{2} + 0{,}1\varepsilon_j\sqrt{1+2X_j}$, $1 \leq j \leq n$,

- H_0 : ϵ y X son independientes
- Kernel (Epanechnikov) $K(u) = 0.75 \times (1 - u^2)$
- Parámetro de ventana mediante validación cruzada
- $W_1(t,s) = \exp(-\lambda_{1,1}|t| - \lambda_{1,2}|s|)$ y
  $W_2(t,s) = \exp(-\lambda_{2,1}t^2 - \lambda_{2,2}s^2)$.
- $\xi \sim Normal(0,1)$ (Se muestran resultados para los centrados)
- El modelo que genera los datos es

$$Y_j = X_j - \frac{X_j^2}{2} + 0{,}1\varepsilon_j\sqrt{1+2X_j},\ 1 \leq j \leq n,$$

- H_0 : ϵ y X son independientes
- Kernel (Epanechnikov) $K(u) = 0.75 \times (1 - u^2)$
- Parámetro de ventana mediante validación cruzada
- $W_1(t,s) = \exp(-\lambda_{1,1}|t| - \lambda_{1,2}|s|)$ y
  $W_2(t,s) = \exp(-\lambda_{2,1}t^2 - \lambda_{2,2}s^2)$.
- $\xi \sim Normal(0,1)$ (Se muestran resultados para los centrados)
- El modelo que genera los datos es
  $Y_j = X_j - \frac{X_j^2}{2} + 0{,}1\varepsilon_j\sqrt{1+2X_j}$, $1 \leq j \leq n$,

Para la estimación del nivel (α) se consideraron las siguientes distribuciones

$$H_{0,0}\colon \varepsilon|X = x \sim N(0,1),$$
$$H_{0,4}\colon \varepsilon|X = x \sim \frac{\chi^2_4 - 4}{\sqrt{8}},$$

donde χ^2_4 representa una distribución χ-cuadrado con 4 grados de libertad.

Estimación del nivel al 5 % y al 10 %

		W_1			W_2		
	n	B	P	WB	B	P	WB
$H_{0,0}$	80	5.5	6.0	3.0	3.5	5.5	3.6
		9.5	12.0	8.8	7.0	11.0	7.8
	160	5.0	5.5	4.8	4.5	6.8	5.1
		9.0	11.0	9.4	9.5	11.5	9.3
	240	4.0	5.8	5.1	4.5	6.0	5.6
		10.0	11.0	10.2	10.0	11.0	10.7
$H_{0,4}$	80	7.0	7.5	3.5	7.0	8.5	3.8
		15.5	14.5	7.5	11.5	15.5	7.9
	160	4.5	6.0	4.1	4.0	6.5	4.6
		10.5	11.5	8.5	8.0	11.5	9.6
	240	4.0	5.0	4.5	5.0	6.0	4.6
		11.0	10.0	9.0	10.0	11.5	10.1

Para la estimación de la potencia (1-β) se consideraron las siguientes distribuciones

$$H_{1,a}\colon \varepsilon|X = x \sim N(0, 1 + ax), \text{ con } a > 0,$$
$$H_{1,b}\colon \varepsilon|X = x \sim \frac{\chi^2_{r_x} - r_x}{\sqrt{2r_x}}, \text{ donde } r_x = \frac{1}{bx} \text{ con } b > 0,$$
$$H_{1,c}\colon \varepsilon|X = x \sim \sqrt{1 - (cx)^{1/4}}\, t_{2/(cx)^{1/4}}, \text{ con } 0 < c \leq 1.$$

Los parámetros $a > 0$, $b > 0$ y $0 < c \leq 1$ controlan la dependencia de la varianza, la asimetría y la curtosis; respectivamente.

Estimación de la potencia al 5 %

		$\alpha = 5\,\%$					
		W_1			W_2		
	n	B	P	WB	B	P	WB
$H_{1,a=5}$	80	6.0	14.0	7.7	16.0	21.5	13.5
	160	7.5	11.5	13.4	14.5	25.0	20.8
	240	11.0	16.0	13.6	21.0	29.5	23.7
$H_{1,b=10}$	80	82.0	84.5	16.2	88.0	89.5	21.9
	160	88.5	90.0	45.7	95.4	96.1	62.6
	240	97.0	99.0	78.5	99.0	99.0	90.4
$H_{1,c=1}$	80	19.5	21.9	24.0	27.0	28.5	27.9
	160	26.0	29.5	26.4	32.0	35.4	32.4
	240	27.5	35.2	30.2	33.5	40.4	47.8

Tiempo empleado en segundos para obtener un p-valor con 1000 repeticiones

n	B_1/WB_1	B_1	WB_1
100	3.79	7.61	2.01
200	11.57	24.41	2.11
300	24.44	54.50	2.23
400	38.46	92.68	2.41
500	52.90	149.17	2.82
1000	93.33	528.27	5.66
1500	78.41	851.58	10.86

Aplicación datos reales (aeronáutica)

La Agencia Espacial Canadiense en el marco de un desarrollo experimental de materiales que disminuyan el sonido en cabina de naves espaciales cuenta con datos sobre el ruido del perfil aerodinámico. Proponen para el estudio la estimación de un modelo de regresión no paramétrica teniendo en cuenta 1503 datos. La variable dependiente (Y) es la presión sonora (dB) y la frecuencia motora (Hz) es la variable independiente (X). Para que las conclusiones del modelo estimado sean válidas se necesita conocer la distribución del error dentro del modelo (se espera que sea ruido blanco). Pero antes es necesaria la aplicación de un test de independencia entre la X y el error ε.

p-valores obtenidos con la aproximación propuesta

		W_1			
		λ_1			
		3.5	2.5	1.5	0.5
λ_2	3.5	0.996	0.981	0.996	0.999
	2.5	0.972	0.980	0.995	0.998
	1.5	0.968	0.979	0.996	0.999
	0.5	0.962	0.972	0.995	0.991

Conclusión: No existe evidencia estadística suficiente que indique que la covariable (X) y el error (ε) sean dependientes.

Gracias por la atención !!

Referencias

- Delhing, H., Mikosch, T. (1994) Random quadratic forms and the bootstrap for *U*-statistics. J. Multivariate Anal. 51, 392-413.
- Hušková, M., Janssen, P. (1993) Consistency of the generalized bootstrap for degenerate *U*-statistics. Test. 19, 92-112.
- Hušková, M., Meintanis, S.G. (2009) Goodness-of-fit tests for parametric regression models based on empirical characteristics functions. Kibernetika. 45, 960-971.
- Hušková, M., Meintanis, S.G. (2010) Test for the error distribution in nonparametric possiby heteroscedatic regression models. Test. 19, 92-112.
- Jiménez-Gamero, M.D., Kim, H-M. (2015) Fast goodness-of-fit test based on the characteristic function. Comput. Statist. Data Anal. 89, 172-191.

Referencias

- Jiménez-Gamero, M.D., Muñoz-García, J., Pino-Mejías, R. (2005) Testing goodness of fit for the distribution of errors in multivariate linear models. J. Multivariate Anal. 95, 301-322.
- Kojadinovic, I., Yan, J., (2012) Goodness-of-fit testing based on a weighted bootstrap: A fast sample alternative to the parametric bootstrap. Can. J. Statist. 40, 480-500.
- Pardo-Fernández, J.C., Jiménez-Gamero, M.D., El Ghouch, A., (2015a) A nonparametric ANOVA-type test for regression curves based on characteristic functions. Scand. J. Stat. 42, 197-213.
- Pardo-Fernández, J.C., Jiménez-Gamero, M.D., El Ghouch, A., (2015b) Tests for the equality of conditional variance functions in nonparametric regression. Electron. J. Statist. 9, 1826-1851.