Experiencia de estancia investigativa, financiada por el CONACYT a través del Programa PROCIENCIA con recursos del Fondo para la Excelencia de la Educación e Investigación del FONACIDE

Gustavo Rivas

Marzo-2017

Datos de identificación

Fecha de estancia: del 14 de noviembre de 2016 al 13 de febrero de 2017 (3 meses).

Lugar: Universidad de Sevilla (España).

Departamento: de Estadística e Investigación Operativa (DEIO).

Responsable: Prof. Dra. María Dolores Jiménez Gamero

Objetivos

1. Fortalecer relación en materia de cooperación científica con el DEIO.
2. Realizar difusión científica a través de congresos, charlas, artículos científicos.
3. Recibir capacitación para consolidar competencias investigativas.

Actividades desarrolladas

1) Preparación para disertación en el Congreso Internacional *9th International Conference of the ERCIM WG on Computational and Methodological Statistics* en Sevilla-España. Titulada: *A two-sample test for the error distribution in nonparametric regression*
Fecha de presentación: 11 de diciembre de 2016.
Detalle de las actividades: Resumen del artículo – Preparación de la presentación (diapositivas) – Preparación de la presentación (discurso) – Práctica de la presentación.

Actividades desarrolladas

2) Corrección de un artículo sometido a *Statistical Papers*, titulado *A two-sample test for the error distribution in nonparametric regression based on the characteristic function*
Fecha solicitud de *Major Revision* 10 de noviembre de 2016. Fecha de respuesta: 25 de noviembre de 2016.
Fecha solicitud de *Minor Revision* 15 de diciembre de 2016. Fecha de respuesta: 22 de diciembre de 2016.
Fecha de aceptación: 22 de diciembre de 2016. Fecha de publicación online: 30 de enero de 2017.

Actividades desarrolladas

3) Artículo sometido a *Advances in Statistical Analysis*, titulado *A weighted bootstrap approximation for comparing the error distributions in nonparametric regression*
Fecha: 3 de febrero de 2017.
Detalle de las actividades: Definición del problema a estudiar – Revisión de la literatura – Adopción de una metodología – Demostración de Resultados Teóricos – Verificación de los resultados teóricos mediante simulación – Revisión de la escritura.

Resumen de un artículo

Sean (X_1, Y_1) y (X_2, Y_2) dos vectores aleatorios e idependientes,

$$Y_k = m_k(X_k) + \sigma_k(X_k)\varepsilon_k, \quad k = 1, 2, \tag{1}$$

donde $m_k(x) = E(Y_k \mid X_k = x)$ es la función de regresión,

$\sigma_k^2(x) = Var(Y_k \mid X_k = x)$ es la varianza condicional,

ε_k es el error de regresión, el cual se asume que es independiente de la covariable X_k.

La igualdad de la distribución de los errores es una suposición usual en varios problemas estadísticos

- Young y Bowman 1995
- Hall y Hart 1990
- Kulasekera y Wang 2001

La igualdad de la distribución de los errores es una suposición usual en varios problemas estadísticos

- Young y Bowman 1995
- Hall y Hart 1990
- Kulasekera y Wang 2001

La igualdad de la distribución de los errores es una suposición usual en varios problemas estadísticos

- Young y Bowman 1995
- Hall y Hart 1990
- Kulasekera y Wang 2001

Hipótesis

$$H_0 : F_1 = F_2,$$

contra la alternativa

$$H_1 : F_1 \neq F_2,$$

donde F_1, F_2 son la funciones de distribución de ε_1 y ε_2, respectivamente.

La cantidad de artículos que se ocupan de la misma hipótesis que la nuestra son escasos

- Mora (2005)
- Pardo-Fernández (2007)

La distribución nula de los estadísticos propuestos en estos artículos son desconocidas y para aproximarlas utilizan un bootstrap suavizado.

La cantidad de artículos que se ocupan de la misma hipótesis que la nuestra son escasos

- Mora (2005)
- Pardo-Fernández (2007)

La distribución nula de los estadísticos propuestos en estos artículos son desconocidas y para aproximarlas utilizan un bootstrap suavizado.

Problema

Se han identificado dos principales problemas con los procedimientos anteriores:

- Se asumen condiciones muy fuertes sobre la distribución de los errores, el cual se asume que posee una densidad.
- Además, aunque sea muy fácil de implementar el bootstrap suavizado, se vuelve computacionalmente muy costoso a medida que el tamaño muestral de los datos se incrementa.

Objetivo

Construir un test que no imponga condiciones muy fuertes sobre la distribución de los errores y que desde el punto de vista computacional sea más eficiente que el bootstrap suavizado.

La hipótesis nula se establece de manera equivalente basados en la función característica de los errores

$$H_0 : C_1 = C_2,$$

contra la alternativa

$$H_1 : C_1 \neq C_2,$$

donde C_1 y C_2 denotan la función característica correspondiente a F_1 y F_2, respectivamente.

Test

El estadístico de contraste propuesto tiene la siguiente forma

$$T_{n_1,n_2} = \int [\hat{C}_{n_1}(t) - \hat{C}_{n_2}(t)]^2 \omega(t) dt = \| \hat{C}_{n_1} - \hat{C}_{n_2} \|^2_{\omega},$$

donde

$$\hat{C}_{n_k}(t) = \hat{R}_{n_k}(t) + \mathrm{i}\hat{I}_{n_k}(t) = \frac{1}{n_k} \sum_{j=1}^{n_k} \cos(t\hat{\varepsilon}_{kj}) + \mathrm{i}\frac{1}{n_k} \sum_{j=1}^{n_k} \sin(t\hat{\varepsilon}_{kj}),\ k = 1, 2,$$

$\omega(t)$ es una función de peso.

Para aproximar los valores críticos se propone un bootstrap ponderado. Este estimador fue utilizado antes, por ejemplo, en

- Kojadinovic y Yan (2012)
- Jiménez-Gamero y Kim (2015)

Para aproximar los valores críticos se propone un bootstrap ponderado. Este estimador fue utilizado antes, por ejemplo, en

- Kojadinovic y Yan (2012)
- Jiménez-Gamero y Kim (2015)

Para la obtención de los resultados teóricos se establecen condiciones regulares no muy fuertes (A1)-(A3). De modo a dar una justificación al estadístico T_{n_1,n_2} para contrastar la H_0, hallamos su límite

TEOREMA 1

Theorem

Si la condiciones (A1)–(A3) se cumplen, entonces $T_{n_1,n_2} \xrightarrow{P} \kappa = \|C_1 - C_2\|_\omega^2$.

TEOREMA 2

Theorem

Si las condiciones (A1)–(A3) se cumplen y H_0 es verdadera

$$\frac{n_1 n_2}{N} T_{n_1,n_2} \xrightarrow{\mathcal{L}} \|Z\|^2_\omega,$$

donde $\{Z(t),\ t \in \mathbb{R}\}$ es un proceso Gaussiano centrado en $L_2(\omega)$ con estructura de covarianza $\varrho_0(t,s) = Cov_0\{Z_0(\varepsilon;t), Z_0(\varepsilon;s)\}$ y

$$\begin{aligned} Z_0(\varepsilon;t) = \quad & \cos(t\varepsilon) + t\varepsilon I(t) - t\tfrac{\varepsilon^2-1}{2} R'(t) - R(t) \\ & + \sin(t\varepsilon) - t\varepsilon R(t) - t\tfrac{\varepsilon^2-1}{2} I'(t) - I(t). \end{aligned}$$

Sean $\xi_{11}, \ldots, \xi_{1n_1}, \xi_{21}, \ldots, \xi_{2n_2}$ variables aleatorias IID con media 0 y varianza 1, independientes de $(X_{11}, Y_{11}), \ldots, (X_{1n_1}, Y_{1n_1}), (X_{21}, Y_{21}), \ldots, (X_{2n_2}, Y_{2n_2})$.

Definimos la versión Bootstrap Ponderado de T_{n_1,n_2}, como

$$T^*_{1,n_1,n_2} = \| C^*_{n_1} - C^*_{n_2} \|^2_\omega,$$

donde

$$C^*_{n_k}(t) = \frac{1}{n_k} \sum_{j=1}^{n_k} \xi_{kj} Z_{k,\tau}(\varepsilon_{kj}; t),$$

$$\begin{aligned} Z_{k,\tau}(\varepsilon; t) = \quad & \cos(t\varepsilon) + t\varepsilon I_k(t) - t\frac{\varepsilon^2-1}{2} R'_k(t) - R_\tau(t) + \\ & \sin(t\varepsilon) - t\varepsilon R_k(t) - t\frac{\varepsilon^2-1}{2} I'_k(t) - I_\tau(t). \end{aligned}$$

$k = 1, 2.$

TEOREMA 3

Theorem

Si las condiciones (A1)–(A3) se cumplen, entonces

$$\sup_x \left| P_* \left\{ \frac{n_1 n_2}{N} T^*_{1,n_1,n_2} \leq x \right\} - P\left\{ T_\tau \leq x \right\} \right| \xrightarrow{P} 0,$$

where $T_\tau = \|Z_\tau\|^2_\omega$, $\{Z_\tau(t),\ t \in \mathbb{R}\}$ *is a centered Gaussian process on* $L_2(\omega)$ *with covariance kernel*
$\varrho_\tau(t,s) = (1-\tau)\varrho_{1,\tau}(t,s) + \tau\varrho_{2,\tau}(t,s)$ *and*
$\varrho_{k,\tau}(t,s) = E\{Z_{k,\tau}(\varepsilon_k; t) Z_{k,\tau}(\varepsilon_k; s)\}$, $k = 1, 2$.

Entonces, en vez de T^*_{1,n_1,n_2}, ahora consideramos

$$T^*_{2,n_1,n_2} = \|\hat{U}^*_1 - \hat{U}^*_2\|^2_\omega.$$

donde

$$\begin{aligned}\hat{U}^*_k(t) = \quad & \frac{1}{n_k}\sum_{j=1}^{n_k}\Big\{\cos(t\hat{\varepsilon}_{kj}) + t\hat{\varepsilon}_{kj}\hat{I}_{n_k}(t) - t\frac{\hat{\varepsilon}^2_{kj}-1}{2}\hat{R}'_{n_k}(t) - \hat{R}_\tau(t) \\ & + \sin(t\hat{\varepsilon}_{kj}) - t\hat{\varepsilon}_{kj}\hat{R}_{n_k}(t) - t\frac{\hat{\varepsilon}^2_{kj}-1}{2}\hat{I}'_{n_k}(t) - \hat{I}_\tau(t)\Big\}\xi_{kj}, \quad k = 1, 2\end{aligned}$$

TEOREMA 4

Theorem

Si las condiciones (A1)–(A3) se cumplen, entonces

$$\sup_x \left| P_* \left\{ \frac{n_1 n_2}{N} T^*_{2,n_1,n_2} \leq x \right\} - P\left\{ T_\tau \leq x \right\} \right| \xrightarrow{P} 0,$$

donde T_τ se define como en el Teorema 3.

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

$$\Psi_* = \begin{cases} 1, & \text{si} \quad T_{n_1,n_2} \geq t^*_{2,n_1,n_2,\alpha}, \\ 0, & \text{para otros casos,} \end{cases}$$

donde $t^*_{2,n_1,n_2,\alpha}$ es el $1-\alpha$ percentil de la distribución condicional de T^*_{2,n_1,n_2}

Corolario 1

Corollary

Si H_0 es verdadera y las condiciones del Teorema 4 se cumplen, entonces

$$\sup_x \left| P_* \left\{ \frac{n_1 n_2}{N} T^*_{2,n_1,n_2} \leq x \right\} - P \left\{ \frac{n_1 n_2}{N} T_{n_1,n_2} \leq x \right\} \right| \xrightarrow{P} 0.$$

Corolario 2

Corollary

Si H_0 no es verdadera, las condiciones del Teorema 4 se cumplen y ω es tal que

$$\kappa = \|C_1 - C_2\|_\omega^2 > 0, \tag{2}$$

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

Objetivo del experimento

- Comparar la aproximación propuesta con el bootstrap suavizado en términos del tiempo de cómputo requerido.

n_1, n_2	Boot/WB	WB
50,50	13.12	1.20
50,100	19.43	1.35
100,100	25.05	1.45
100,150	36.13	1.70
150,150	36.63	1.85

Table : Tiempo de consumo en segundos para el cálculo de un p-valor.

Gracias por la atención

Referencias

- Alba-Fernández V, Jiménez-Gamero MD, Muñoz-García J (2008) A test for the two-sample problem based on empirical characteristics function. Comput Statist Data Anal 52:3730–3748.
- Baringhaus L, Kolbe D (2015) Two-sample tests based on empirical Hankel transforms. Stat Papers 56:597-617.
- Burke MD (2000) Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap. Statist Probab Lett 46:13–20.
- Cleveland W S (1993) Visualizing Data. Summit, New Jersey
- Feller W (1971). An Introduction to Probability Theory and its Applications. Vol 2. Wiley, New York.

References

- Ghoudi K, Rémillard B (2014) Comparison of specification tests for GARCH models. Comput Statist Data Anal 76:291–300.
- Jiménez-Gamero MD, Kim H-M (2015) Fast goodness-of-fit test based on the characteristic function. Comput Statist Data Anal 89:172–191.
- 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.
- Mora J, (2005) Comparing distribution functions of errors in linear models: A nonparametric approach. Statist Probab Lett 73:425–432

References

- Pardo-Fernández JC (2007) Comparison of error distributions in nonparametric regression. Statist Probab Lett 77:350–356.
- Pardo-Fernández JC, Jiménez-Gamero MD, El Ghouch A (2015a) A nonparametric ANOVA-type test for regression curves based on characteristic functions. Scand J Stat 42:197–213.
- Pardo-Fernández JC, Jiménez-Gamero MD, El Ghouch A (2015b) Tests for the equality of conditional variance functions in nonparametric regression. Electron J Stat 9:1826–1851.
- R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015.URL http://www.R-project.org/