Estudio de la relación de Alzheimer con la edad y escolaridad

Martin Alonso Flores Gonzalez

20 de abril de 2023

1. Contexto

Los datos a estudiar son el resultado de aplicar ciertas pruebas cognitivas a dos grupos de personas: portadores del gen (C) y no portadores del gen (NC).

Es necesario mencionar que las pruebas realizadas a cada uno de los participantes son: minimental state examination (MMSE), semantic verbal fluency test using the category animals (SVF), una versión modificada de Boston naming (VBN), word list memory (WLM), word list delayed (WLD), word recognition (WLR), constructional practice copy (CPC) y constructional practice evocation (CPE). Además de considerar los resultados de cada una de las pruebas anteriores, también se consideran las edades y los años de estudios académicos realizados por cada uno de los participantes.

La intención del análisis que se mostrará, es observar los efectos que la edad y años de estudios tienen en cada uno de las grupos de interés (C y NC) mediante los resultados de cada participante en las pruebas.

Como es común es este tipo de investigaciones, es complicado encontrar a personas que sean aptas y además quieran participar en el estudio (especialmente a los portadores del gen); es por ello que los datos originales de cada grupo no son suficientes para obtener resultados concluyentes. Si bien es cierto que es un problema obtener nuevos datos a partir de participantes nuevos, es posible simularlos a partir de los datos que se ya se tienen. Para ello se utilizarán ciertas herramientas estadísticas que se mostrarán un poco más adelante.

1.1. Simulación

Para comenzar, necesitamos definir un vector aleatorio que represente la información que es posible obtener para cada participante del estudio.

Definimos a $X^{\alpha,\beta} \in \mathbb{N}^3$ como un vector aleatoria dado por $X^{\alpha,\beta} = (x_1^{\alpha,\beta}, x_2^{\alpha,\beta}, x_3^{\alpha,\beta})$, que representa la información que es posible obtener de un participante en el estudio, donde $\alpha = 1, 2$ es el grupo al que pertenece el participante (C o NC respectivamente), $\beta = 1, 2, \ldots, 8$ es la prueba aplicada (MMST, SVF, VBN, WLM, WLD, WLR, CPC y CPE respectivamente), $x_1^{\alpha,\beta}$ representa el resultado de la prueba aplicada, $x_2^{\alpha,\beta}$ y $x_3^{\alpha,\beta}$ representa la edad y la escolaridad.

Derivado de lo anterior, denotamos a las realizaciones de la variable aleatoria $X^{\alpha,\beta}$ por $X_i^{\alpha,\beta} = (x_{1,i}^{\alpha,\beta}, x_{2,i}^{\alpha,\beta}, x_{3,i}^{\alpha,\beta})$ que representa la información obtenida del i-ésimo participante del grupo α, con $i = 1, 2, \ldots, n_\alpha$ y donde n_α representa el número de participantes por grupo ($n_1 = 15$ y $n_2 = 24$).

Como los elementos de las realizaciones $X_i^{\alpha,\beta}$ no tienen las mismas unidades de medida, optamos por trabajar con los z-scores de cada uno de ellos y por ende, empleamos las siguientes transformaciones:

$$
\begin{aligned}
z_{1,i}^{\alpha,\beta} &= \frac{x_{1,i}^{\alpha,\beta} - \bar{x}_1^{\alpha,\beta}}{s_1^{\alpha,\beta}}\ , \\
z_{2,i}^{\alpha,\beta} &= \frac{x_{2,i}^{\alpha,\beta} - \bar{x}_2^{\alpha,\beta}}{s_2^{\alpha,\beta}}\ , \\
z_{3,i}^{\alpha,\beta} &= \frac{x_{3,i}^{\alpha,\beta} - \bar{x}_3^{\alpha,\beta}}{s_3^{\alpha,\beta}}\ ,
\end{aligned}
\tag{1}
$$

donde $\bar{x}_k^{\alpha,\beta}$ y $s_k^{\alpha,\beta}$, para $k = 1, 2, 3$, son la media y varianza muestral de los resultados de las pruebas, edad y años de escolaridad respectivamente. De lo anterior obtenemos un nuevo vector aleatorio $Z^{\alpha,\beta} \in \mathbb{R}^3$ cuyas realizaciones están dadas por $Z_i^{\alpha,\beta} = (z_{1,i}^{\alpha,\beta}, z_{2,i}^{\alpha}, z_{3,i}^{\alpha})$. En adelante, utilizamos las realizaciones de este nuevo vector aleatorio para analizar la información obtenida.

Como los elementos de $Z_i^{\alpha,\beta}$ no son independientes, utilizamos las correlaciones que existen entre ellos para simular nuevas muestras. Para ello empleamos el teorema de Sklar [4] y una cúpula Gaussiana [6, 1] para obtener vectores aleatorios uniformes con elementos correlacionados entre sí. Por último, utilizamos muestreo por transformada inversa (inverse transform sampling) [5, 2] para crear las nuevas muestras.

La metodología que utilizamos para simular las nuevas variables aleatorias se resume en los siguientes pasos:

1. Estimación de la matriz de correlación $\rho^{\alpha,\beta}$ del vector aleatorio $Z^{\alpha,\beta}$ utilizando las realizaciones $Z_i^{\alpha,\beta}$. Los elementos de la matriz estimada están dados por

$$
\hat{\rho}_{p,q}^{\alpha,\beta} = \frac{1}{n_\alpha} \sum_{j=1}^{n_\alpha} z_{p,j}^{\alpha,\beta} z_{q,j}^{\alpha,\beta}\,, \tag{2}
$$

donde $p, q = 1, 2, 3$. La anterior forma de estimar la correlación se obtiene al tomar en cuenta que los z-scores tienen media cero y varianza unitaria.

2. Simulación de un vector aleatorio que se distribuye como una normal multivariada, con vector de medias igual a cero y covarianza $\hat{\rho}^{\alpha,\beta}$, es decir:

$$W^{\alpha,\beta} = (w_1^{\alpha,\beta}, w_2^{\alpha,\beta}, w_3^{\alpha,\beta}) \sim N_3(\mu, \hat{\rho}^{\alpha,\beta}), \quad (3)$$

donde $\mu^T = (0, 0, 0)$.

3. Obtención de las variables aleatorias uniformes ya correlacionadas al aplicar la función de distribución acumulada Φ de una normal estándar:

$$\begin{aligned} U^{\alpha,\beta} &= (u_1^{\alpha,\beta}, u_2^{\alpha,\beta}, u_3^{\alpha,\beta}) \\ &= (\Phi(w_1^{\alpha,\beta}), \Phi(w_2^{\alpha,\beta}), \Phi(w_3^{\alpha,\beta})) \end{aligned} \quad (4)$$

4. Transformación de las variables uniformes en las nuevas muestras utilizando inverse transform sampling

$$\begin{aligned} \hat{Z}^{\alpha,\beta} &= (\hat{z}_1^{\alpha,\beta}, \hat{z}_2^{\alpha,\beta}, \hat{z}_3^{\alpha,\beta}) \\ &= (F_{\alpha,\beta}^{-1}(u_1^{\alpha,\beta}), G_{\alpha,\beta}^{-1}(u_2^{\alpha,\beta}), H_{\alpha,\beta}^{-1}(u_3^{\alpha,\beta}), \end{aligned} \quad (5)$$

donde $F_{\alpha,\beta}, G_{\alpha,\beta}, H_{\alpha,\beta}$ son las interpolaciones lineales de las funciones de distribución acumulada empíricas [3] obtenidas con las realizaciones $z_{1,i}^{\alpha,\beta}, z_{2,i}^{\alpha,\beta}, z_{3,i}^{\alpha,\beta}$ respectivamente.

Este procedimiento nos permitió simular tantos vectores aleatorios $\hat{Z}^{\alpha,\beta}$ como se necesitaron para realizar un análisis más robusto.

1.2. Rendimiento cognitivo como función de la edad y la escolaridad

Para analizar el desempeño de las personas en las tareas con relación a su edad y escolaridad, utilizamos las simulaciones $\hat{Z}^{\alpha,\beta}$ cuya obtención se detallamos en la sección anterior. En particular, modelamos a $\hat{z}_1^{\alpha,\beta}$ como una función dependiente de $\hat{z}_2^{\alpha,\beta}$ y $\hat{z}_3^{\alpha,\beta}$ (recordemos que $\hat{z}_1^{\alpha,\beta}, \hat{z}_2^{\alpha,\beta}$ y $\hat{z}_3^{\alpha,\beta}$ son simulaciones de los z-scores de $x_1^{\alpha,\beta}, x_2^{\alpha,\beta}$ y $x_3^{\alpha,\beta}$) mediante una regresión lineal de acuerdo con el siguiente modelo:

$$\hat{z}_1^{\alpha,\beta} = b_1 + b_2\hat{z}_2^{\alpha,\beta} + b_3\hat{z}_3^{\alpha,\beta} + \epsilon, \quad (6)$$

donde b_1, b_2 y b_3 representan los parámetros de la regresión y ϵ representa el error.

Es importante notar que los parámetros de la regresión lineal varían cuando se utilizan arreglos de simulaciones diferentes, por lo tanto, utilizamos la siguiente metodología para obtener una estimación robusta.

Sean α, β fijos, entonces:

1. Se simula una muestra M que consiste de diez mil vectores $\hat{Z}^{\alpha,\beta} = (\hat{z}_1^{\alpha,\beta}, \hat{z}_2^{\alpha,\beta}, \hat{z}_3^{\alpha,\beta})$.

2. Con la muestra M se obtienen las estimaciones $\hat{b}_1, \hat{b}_2$ y $\hat{b}_3$ de los parámetros del modelo de regresión (6).

3. Se calculan los p-valores p_1, p_2 y p_3 correspondientes a las estimaciones $\hat{b}_1, \hat{b}_2$ y $\hat{b}_3$ respectivamente.

4. Se define un vector de resultados dado por

$$R = (\hat{b}_1, \hat{b}_2, \hat{b}_3, p_1, p_2, p_3) \quad (7)$$

5. Se realizan los pasos 1,2,3 y 4, mil veces, esto nos permite obtener las muestras R_i con $i = 1, 2, \ldots, 1000$.

6. Se descartan los vectores de resultados R_i cuando uno de los p-valores es menor de 0.05 (nivel de significación estadística elegida).

7. Se calcula la media de las estimaciones de los parámetros con los vectores de resultados R_i no descartados por el paso 6. Así obtuvimos los estimadores promedio de la regresión $\bar{b}_1, \bar{b}_2$ y $\bar{b}_3$

8. Se calcula el porcentaje de aceptación A para conocer la cantidad de vectores de resultados R que no se rechazan, tal y como se muestra a continuación.

$$A = \frac{n_a}{n_s} 100\,\%, \quad (8)$$

donde n_a es el número de vectores de resultados no rechazados y n_s es el número de muestras M simuladas.

Realizamos los pasos anteriores para todas las combinación de indices α y β.

Notemos que no hay manera de obtener los p-valores relacionados a los parámetros $\bar{b}_1^{\alpha,\beta}, \bar{b}_2^{\alpha,\beta}$ y $\bar{b}_3^{\alpha,\beta}$, ya que estas estimaciones no provienen de una muestra de simulaciones M en particular. Sin embargo, estimamos los respectivos p-valores con la siguiente metodología.

Con α, β, $\bar{b}_1^{\alpha,\beta}$, $\bar{b}_2^{\alpha,\beta}$ y $\bar{b}_3^{\alpha,\beta}$ fijos, entonces:

1. Se simula una muestra M que consiste de diez mil vectores $\hat{Z}^{\alpha,\beta} = (\hat{z}_1^{\alpha,\beta}, \hat{z}_2^{\alpha,\beta}, \hat{z}_3^{\alpha,\beta})$.

2. Se calcula los p-valores $\hat{p}_1^{\alpha,\beta}, \hat{p}_2^{\alpha,\beta}$ y $\hat{p}_3^{\alpha,\beta}$ correspondientes a las estimaciones $\bar{b}_1^{\alpha,\beta}$, $\bar{b}_2^{\alpha,\beta}$ y $\bar{b}_3^{\alpha,\beta}$ con los datos de la muestra M. Además definimos el vector $P^{\alpha,\beta} = (\hat{p}_1^{\alpha,\beta}, \hat{p}_2^{\alpha,\beta}, \hat{p}_3^{\alpha,\beta})$.

3. Se realizan los pasos 1 y 2, mil veces, así se obtienen los vectores $P_i^{\alpha,\beta}$ con $i = 1, 2, \ldots, 1000$.

4. Se calculan las medias ($\bar{p}_1^{\alpha,\beta}, \bar{p}_2^{\alpha,\beta}$ y $\bar{p}_3^{\alpha,\beta}$) de los p-valres con los vectores $P_i^{\alpha,\beta}$.

Lo anterior se realizó para todas las combinaciones de índices α y β.

1.3. Resultados

Los resultados de los anteriores procedimientos se muestran en los cuadros (1) y (2), en donde se encuentran los parámetros estimados $\bar{b}_1$, $\bar{b}_2$ y $\bar{b}_3$ para el modelo lineal ajustado a cada prueba cognitiva. En los cuadros también se observan los p-valores estimados $\bar{p}_1$, $\bar{p}_2$ y $\bar{p}_3$ para los parámetros de cada regresión y además mostramos el porcentajes A de vectores R no rechazados durante las simulaciones.

Lo primero a notar es que los porcentajes A son altos a excepción de tres casos: para la prueba VBN (con 13 %) realizada al grupo de portadores, para las pruebas WLR y CPE (con 8 % y 3.4 % respectivamente) realizadas en el grupo de no portadores.

Lo segundo a notar es que los p-valores $\bar{p}_1$, $\bar{p}_2$ y $\bar{p}_3$ son muy pequeños para todas las pruebas con alto porcentaje A. Obteniendo evidencia de que las variables $\hat{z}_2^{\alpha,\beta}$ y $\hat{z}_3^{\alpha,\beta}$, que representan la edad y escolaridad respectivamente, si contribuyen de manera significativa a $\hat{z}_1^{\alpha,\beta}$ que representa los resultados obtenidos en cada prueba.

También es importante resaltar que en el grupo de portadores, los resultados obtenidos en las pruebas MMSE, SVF, WLM, WLD, WLR y CPE son mayormente influenciados por los años de escolaridad que por la edad del participante, esto debido a que $|\bar{b}_2| < |\bar{b}_3|$. Mientras que el resultado de la prueba CPC aplicada al grupo de portadores es mayormente influenciado por la edad, ya que $|\bar{b}_2| > |\bar{b}_3|$.

Por otro lado, vemos que en el grupo de no portadores los resultados obtenidos en las pruebas MMSE, VBN, WLM, WLD y CPC también son mayormente influenciados por los años de escolaridad que por la edad del participante, esto debido a que $|\bar{b}_2| < |\bar{b}_3|$. Mientras que el resultado de la prueba SVF aplicada al grupo de no portadores es mayormente influenciado por la edad, ya que $|\bar{b}_2| > |\bar{b}_3|$.

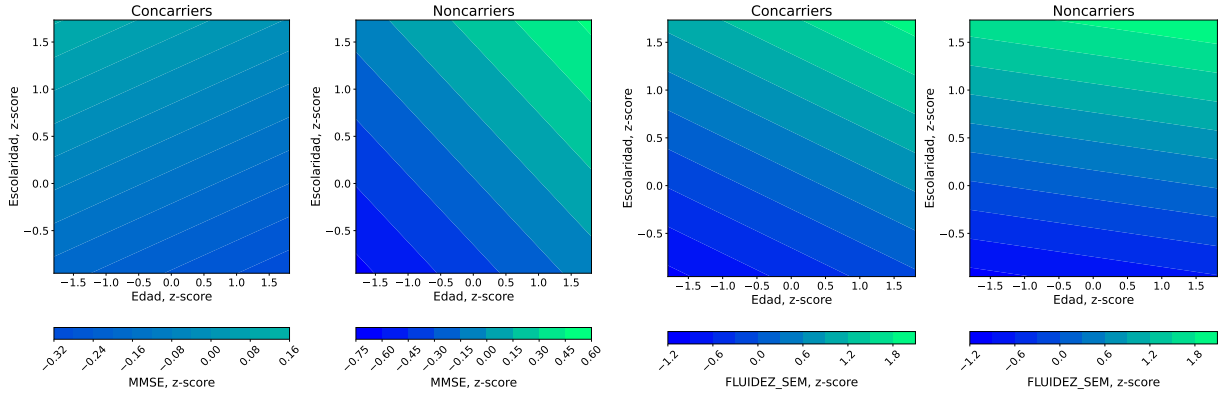
Se puede observar los modelos obtenidos al graficar la ecuación (6) (sin la parte del error) utilizando los parámetros $\bar{b}_1$, $\bar{b}_2$ y $\bar{b}_3$ según lo que se muestra en los cuadros (1) y (2). Las graficas se pueden observar en la figura (1) de la siguiente página.

	b_1	b_2	b_3	A_b	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$
MMSE	-0.136	-0.036	0.114	99.9 %	2.20×10^{-86}	8.703×10^{-7}	1.099×10^{-46}
SVF	0.181	0.258	0.738	100 %	3.655×10^{-174}	4.451×10^{-285}	0
VBN	-0.23	-0.018	0.285	13 %	8.211×10^{-195}	0.021	4.078×10^{-223}
WLM	-0.072	-0.285	0.382	100 %	4.864×10^{-20}	6.15×10^{-231}	0
WLD	-0.059	-0.365	0.51	100 %	3.541×10^{-22}	0	0
WLR	-0.217	-0.12	0.166	100 %	7.726×10^{-241}	2.597×10^{-67}	9.22×10^{-108}
CPC	0.171	0.175	0.125	100 %	2.31×10^{-154}	2.603×10^{-134}	1.11×10^{-62}
CPE	-0.149	-0.066	0.267	100 %	1.393×10^{-67}	1.16×10^{-12}	1.893×10^{-150}

Cuadro 1: Resultados de las simulaciones para el grupo de portadores

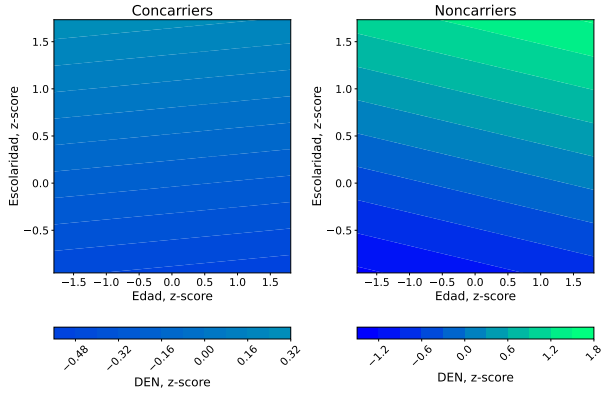
	b_1	b_2	b_3	A_b	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$
MMSE	-0.163	0.156	0.211	100 %	9.532×10^{-79}	4.649×10^{-69}	5.077×10^{-12}
SVF	0.139	0.104	0.99	100 %	5.234×10^{-44}	1.839×10^{-23}	8.988×10^{-154}
VBN	-0.194	0.142	0.85	100 %	3.14×10^{-77}	2.766×10^{-40}	2.286×10^{-110}
WLM	-0.041	-0.192	0.526	98.8 %	5.519×10^{-6}	4.55×10^{-88}	2.262×10^{-56}
WLD	-0.333	-0.276	-0.314	100 %	7.941×10^{-209}	4.823×10^{-140}	2.823×10^{-17}
WLR	-0.491	-0.085	-0.075	8 %	0	4.434×10^{-17}	0.031
CPC	-0.0542	-0.042	1.295	95.6 %	4.481×10^{-7}	1.526×10^{-4}	2.612×10^{-221}
CPE	-0.017	-0.040	1.345	3.4 %	0.086	6.612×10^{-5}	6.296×10^{-284}

Cuadro 2: Resultados de las simulaciones para el grupo de no portadores

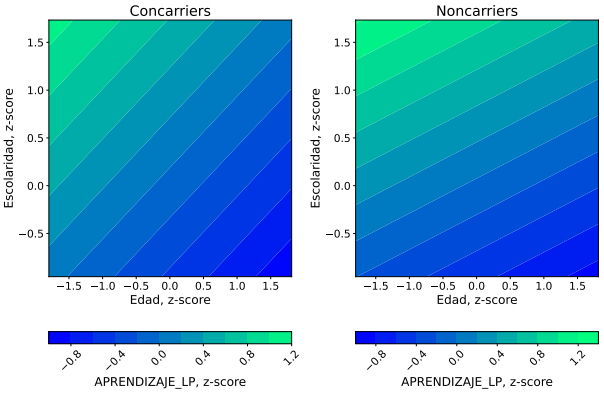


(a) MMSE

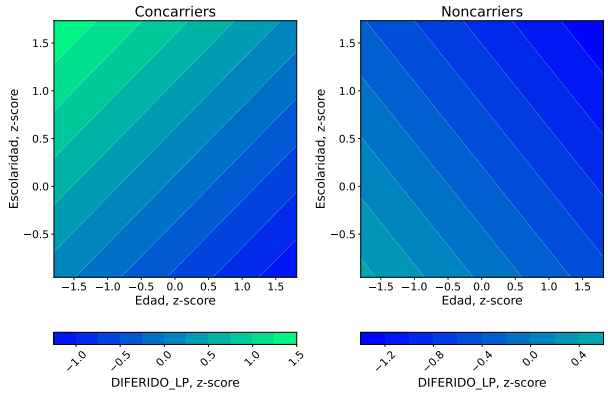
(b) SVF



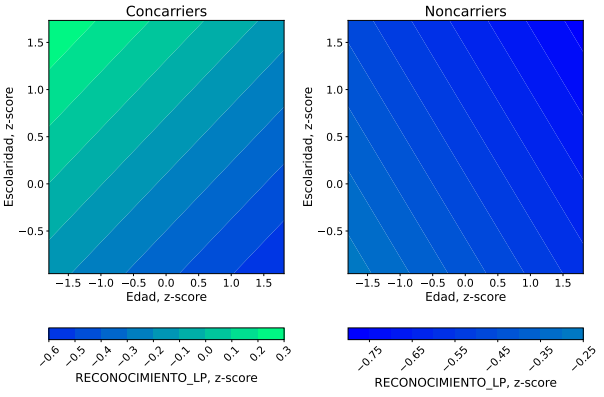
(c) VBN



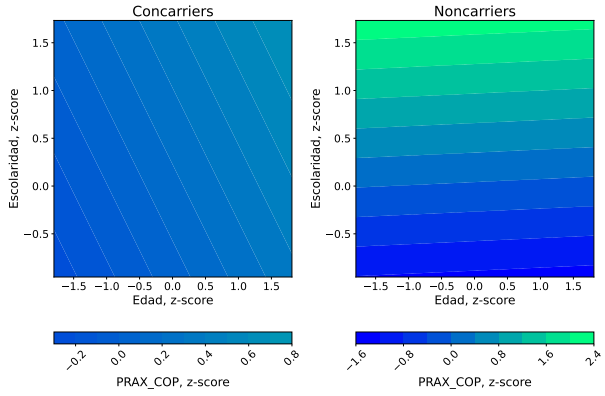
(d) WLM



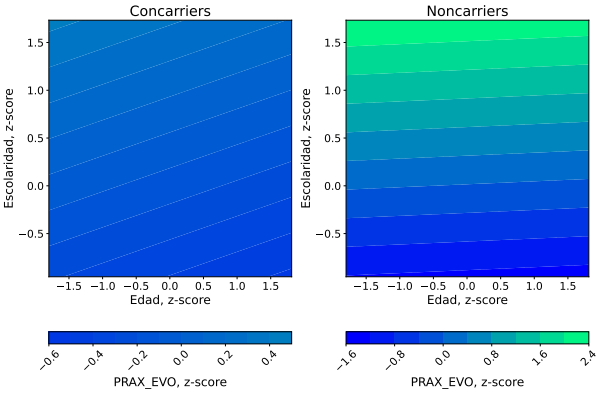
(e) WLD



(f) WLR



(g) CPC



(h) CPE

Figura 1: Graficas de los modelos lineales obtenidos para cada uno de los grupos de estudio según la prueba aplicada.

Referencias

[1] Alireza Bayestehtashk e Izhak Shafran. "Parsimonious multivariate copula model for density estimation". En: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), págs. 5750-5754.

[2] George Casella Christian P. Robert. *Monte Carlo Statistical Methods*. Springer New York, NY, 2004.

[3] O Fercheluc. "Some asymptotics and applications of the interpolated bootstrap method". En: *Statistica Applicata* 11 (), págs. 7-22.

[4] Roger B. Nelsen. *An Introduction to Copulas*. Springer New York, NY, 2006.

[5] Sheehan Olver y Alex Townsend. "Fast inverse transform sampling in one and two dimensions". En: (jul. de 2013).

[6] Pravin K. Trivedi y David M. Zimmer. "Copula Modeling: An Introduction for Practitioners". En: *Foundations and Trends® in Econometrics* 1.1 (2007), págs. 1-111. ISSN: 1551-3076. DOI: 10.1561/0800000005. URL: http://dx.doi.org/10.1561/0800000005.