

Estimación robusta de regiones de referencia condicionales con R. Aplicación en la investigación de diabetes

Sara Rodríguez-Pastoriza¹
Pardiñas^{1,3}

Óscar Lado-Baleato²
Francisco Gude^{4,5,6}

Javier Roca

¹Department of Statistics and Operations Research, University of Vigo, 36310 Vigo, Galicia, Spain.

²Statistics and Operations Research Department, Oviedo University, Asturias, Spain.

³Galician centre for mathematical research and technology, (CITMAGA), Galicia, Spain.

⁴Research Methods Group (RESMET), Health Research Institute of Santiago de Compostela (IDIS), Spain.

⁵Department of Psychiatry, Radiology, Public Health, Nursing and Medicine, University of Santiago de Compostela, Galicia, Spain.

⁶Concepción Arenal Primary Care Center, Santiago de Compostela, Galicia, Spain.

SiDOR
Statistical Inference
Decision & Operations Research Group

Universidade de Vigo



① Introducción

② Objetivos

③ Metodología

④ Estudios de simulación

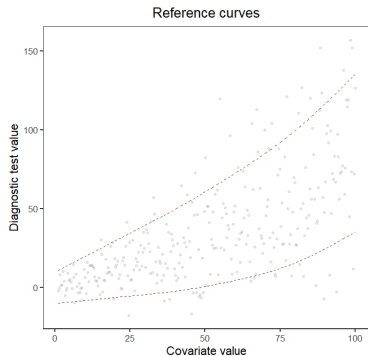
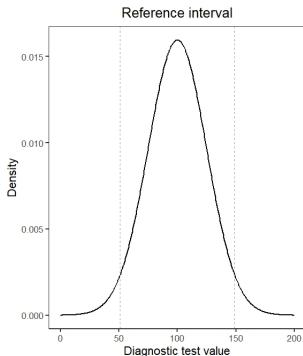
⑤ Aplicación a datos

⑥ Conclusiones y trabajos futuros

INTRODUCCIÓN

Con uno o dos marcadores: intervalos y regiones

- En la **práctica clínica**, la interpretación de un determinado **marcador diagnóstico continuo** requiere un **intervalo** de referencia.
- Para un **único marcador**, los intervalos de referencia se definen como **dos puntos**, entre los cuales se encuentran el **95 %** de los resultados de los **sujetos sanos**.



Con uno o dos marcadores: intervalos y regiones

- La interpretación de dos marcadores continuos dio origen al concepto de **región de referencia** como **extensión de los intervalos**, una **envoltura convexa** que contiene el **95 %** de los valores de la **distribución conjunta** de dos o más **marcadores diagnósticos continuos**.

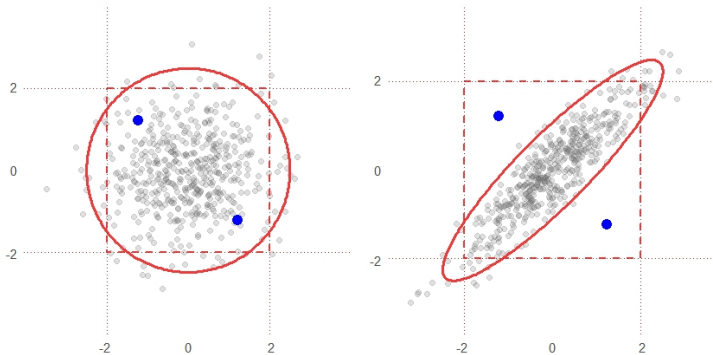
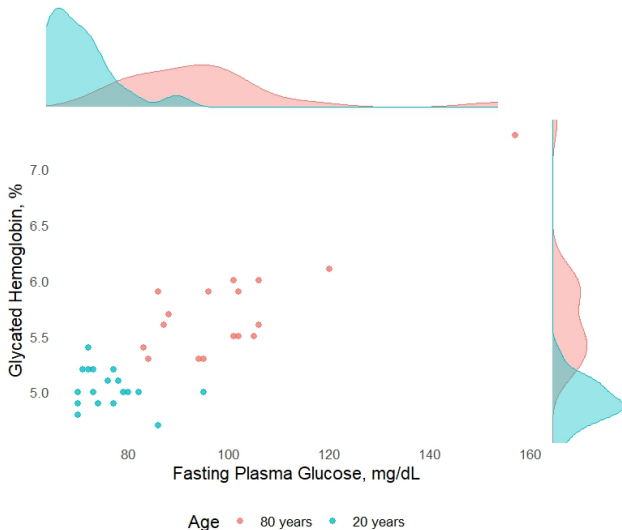


Fig.

Diabetes mellitus

- La **diabetes mellitus** es un **trastorno metabólico** crónico que se caracteriza por **niveles elevados de glucosa** en sangre.
- Existen **dos biomarcadores** que, en conjunto, ofrecen una **evaluación** de la **homeostasis de la glucosa**:
 - **Glucosa plasmática (FPG)**: **glucosa en sangre** después de un período de **ayuno de 8 a 12 horas**.
 - **Hemoglobina glucosilada (HbA1c)**: niveles promedio de **glucosa en sangre** durante los **dos a tres meses** midiendo la proporción de hemoglobina glucosilada.

Diabetes mellitus



OBJETIVOS

Objetivos planteados

- **Explorar** los **avances recientes** en la estimación de **regiones de referencia condicionales**.
- Proponer un **nuevo modelo de regresión bivariada** que estima **directamente** estas regiones condicionales.
- Estudiar la **robustez del método** frente a la presencia de **valores atípicos**.
- **Comparar las estimaciones** del modelo propuesto **con los publicados** en estudios anteriores (modelo de localización y escala, paquete refreg)

METODOLOGÍA

Método propuesto

En esta sección, presentamos el procedimiento de estimación de la región bivariada condicionada dada en la ecuación

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \mu_1(\mathbf{X}) \\ \mu_2(\mathbf{X}) + \mu_{12}(\mathbf{X})Y_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}. \quad (1)$$

PASO 1. Estimación de las medianas condicionales de Y_1 e Y_2 :

Construcción de modelos de regresión para la mediana bivariada planteado de la siguiente forma

$$Y_r = \mu_j(X) + \varepsilon_r = \alpha_r + f^r(X) + \varepsilon_r \quad \text{para } r = 1, 2 \quad (2)$$

donde α_r son los interceptos de regresión, f^r son los efectos parciales asociados a la covariable X , y ε_r representa el error, que cumple que la mediana de ε_r dado X es cero.

Método propuesto

PASO 2. Modelado de la dependencia condicional entre respuestas centradas.

Obtenemos una muestra centrada $\{X_i, Y_{i1}^C, Y_{i2}^C\}_{i=1}^n$ a partir de la fórmula

$$Y_{ir}^C = Y_{i1} - \hat{\mu}_r(X_i), r = 1, 2.$$

Modelamos cómo depende Y_2^C de Y_1^C a partir de la mediana, condicionando esta relación a X :

$$Y_2^C = \mu_{12}(X) \cdot Y_1^C + \varepsilon_{12} = (\alpha_{12} + f^{12}(X)) \cdot Y_1^C + \varepsilon_{12} \quad (3)$$

donde la mediana del error ε_{12} es cero.

Método propuesto

PASO 3. Estandarización de residuos y transformación a coordenadas polares.

- Calcular las variables respuesta re-escaladas $\tilde{Y}_{i1}^C = Y_{i1}^C / Rl_1$ y $Y_{i2}^C = \left(\tilde{Y}_{i2}^C - \hat{\mu}_{12}(X_i) \cdot Y_{i1}^C \right) / Rl_2$ con $Rl_r = Q_{0,75}^{Y_r} - Q_{0,25}^{Y_r}$ ($r = 1, 2$) siendo $Q_{\tau}^{Y_1}$ y $Q_{\tau}^{Y_2}$ los τ -cuantiles empíricos de los valores $\{Y_{i1}^C\}_{i=1}^n$ y $\{Y_{i2}^C - \hat{\mu}_{12}(X_i) \cdot Y_{i1}^C\}_{i=1}^n$ respectivamente.
- Para $i = 1, \dots, n$, calcular la distancia euclídea R_i y el ángulo α_i como sigue:

$$R_i = \sqrt{\left(\tilde{Y}_{i1}^C\right)^2 + \left(\tilde{Y}_{i2}^C\right)^2} \quad \text{y} \quad \alpha_i = \arctan \left(Y_{i2}^C / Y_{i1}^C \right)$$

Método propuesto

PASO 4. Estimación del contorno condicional mediante regresión cuantil polar.

Usando la muestra $\{X_i, R_i, \alpha_i\}_{i=1}^n$, ajustar el modelo de regresión cuantílica

$$\log R_i = f(X, \alpha) + \varepsilon_R = \alpha_r + g(X, \alpha) + \varepsilon_r, \quad (4)$$

donde el τ -ésimo cuantil de ε_r es cero.

Dada una malla $\theta_1 < \dots < \theta_N$ de N ángulos equidistantes en $[0, 2\pi)$, la región incondicional puede estimarse mediante el polígono definido por los puntos $\{U_j, V_j\}_{j=1}^N$ dados por

$$U_j = \exp(\hat{g}(\theta_j, \mathbf{X})) \cos(\theta_j) \cdot Rl_1, V_j = \exp(\hat{g}(\theta_j, \mathbf{X})) \sin(\theta_j) \cdot Rl_2$$

.

Método propuesto

PASO 5. Reconstrucción de la región bivariada condicional en el espacio original: Finalmente, se obtiene la estimación $\hat{R}_\tau(\mathbf{X})$ a partir del polígono definido por los puntos $\{\tilde{U}_j, \tilde{V}_j\}_{j=1}^N$ donde

$$\begin{pmatrix} \tilde{U}_j(\mathbf{X}) \\ \tilde{V}_j(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1(\mathbf{X}) \\ \hat{\mu}_2(\mathbf{X}) \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \hat{\mu}_{12}(\mathbf{X}) & 1 \end{pmatrix} \begin{pmatrix} U_j \\ V_j \end{pmatrix} \quad j = 1, \dots, n$$

ESTUDIOS DE SIMULACIÓN

Scenario 1: Gaussian data

Scenario 2: Gaussian data with outliers

Scenario 3: Non gaussian data

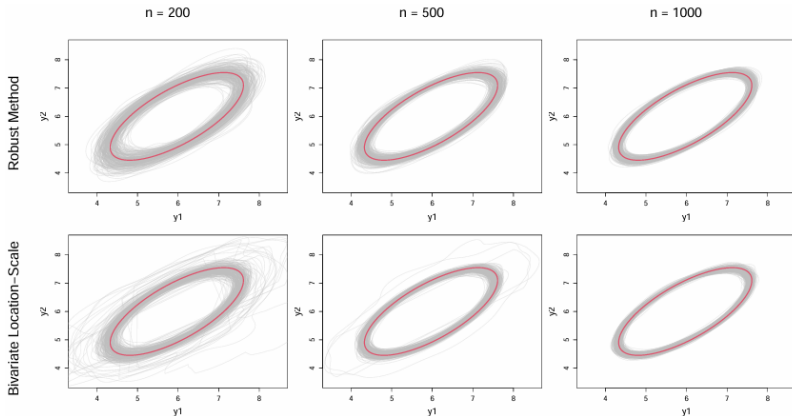


Fig.. Representación de las regiones bivariadas estimadas $\hat{R}_\tau(X)$ para $\tau = 0.95$ y $X = 0.5$ con el nuevo modelo y el previo (datos gaussianos).

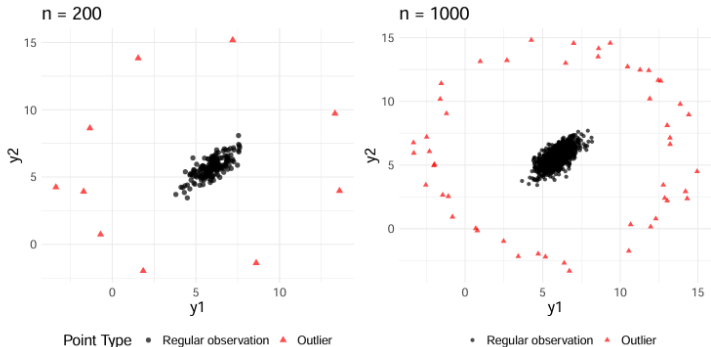
Scenario 1: Gaussian
dataScenario 2: Gaussian
data with outliersScenario 3: Non
gaussian data

Fig.. Ejemplo de dos conjuntos de datos generados para el escenario de simulación 2, con los valores atípicos resaltados en rojo y las observaciones regulares representadas en negro.

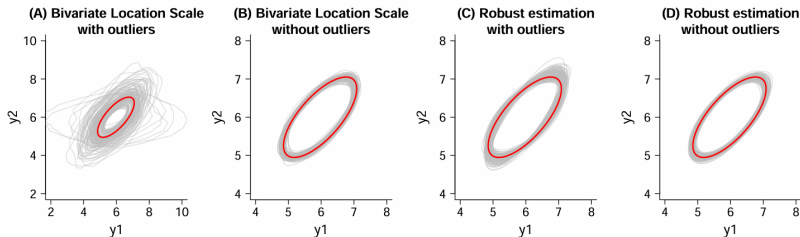
Scenario 1: Gaussian
dataScenario 2: Gaussian
data with outliersScenario 3: Non
gaussian data

Fig.. Comparación de las regiones estimadas utilizando el modelo de localización y escala y el nuevo método propuesto. Los resultados se presentan para $n = 1000$ y $\tau = 0,75$, con 500 iteraciones y diferentes valores de $X1$. Cada línea gris representa una región condicional estimada, mientras que la curva roja corresponde a la región teórica.

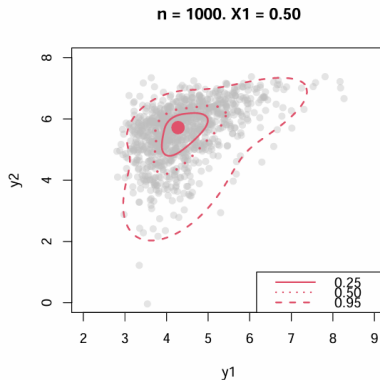
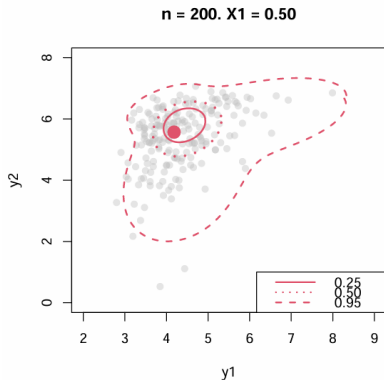
Scenario 1: Gaussian
dataScenario 2: Gaussian
data with outliersScenario 3: Non
gaussian data

Fig.. Ejemplos de dos conjuntos de datos generados en el escenario 3 para tamaños de muestra $n = 200, 1000$, junto con la estimación mediante los métodos propuestos para diferentes τ_s .

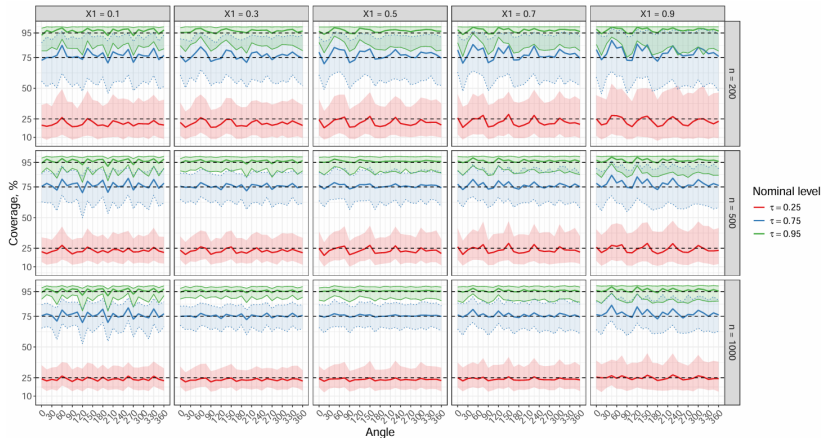
Scenario 1: Gaussian
dataScenario 2: Gaussian
data with outliersScenario 3: Non
gaussian data

Fig.. Cobertura angular de las regiones de referencia condicionadas a diferentes valores de X_1 y τ , para $n = 200, 500, 1000$.

APLICACIÓN A DATOS

Aplicación a la investigación de diabetes

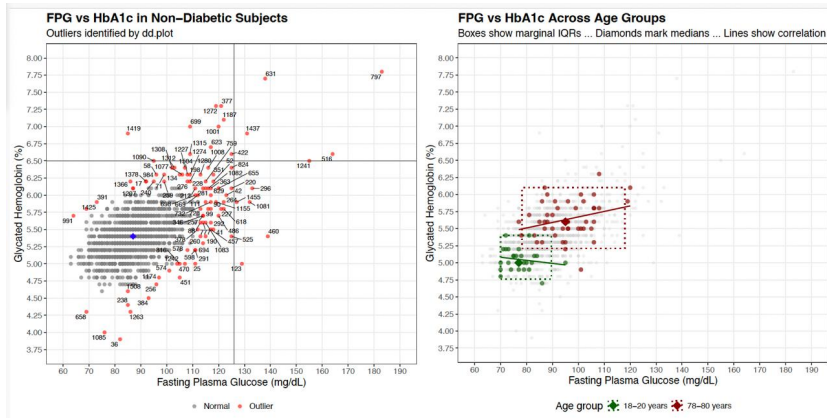


Fig.. Relación entre FPG y HbA1c en sujetos no diabéticos. (A) Detección de valores atípicos mediante dd.plot (paquete mvoutliers). (B) Comparación de la relación FPG-HbA1c entre grupos etarios (18-20 y 78-80 años).

Aplicación a la investigación de diabetes

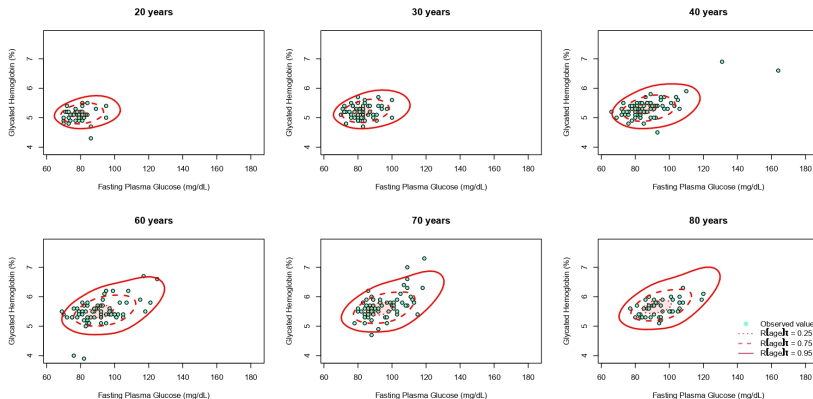


Fig.. Región estimada para los valores de glucosa plasmática en ayunas y hemoglobina glicosilada en función de la edad para $\tau = 0,25, 0,50, 0,95$.

CONCLUSIONES

Conclusiones

- **Estimación más precisa** en comparación con el modelo clásico de localización y escala.
- **Mayor flexibilidad** en los **efectos de la covariable**, permitiendo que su **influencia varíe** a lo largo de diferentes **cuantiles** (por ejemplo, la edad puede impactar más en el percentil 95 que en la mediana).
- **Robustez de valores atípicos**: es **menos susceptible** a **distorsiones** causadas por **valores atípicos**.
- Capacidad para modelar **efectos no lineales** de covariables, como la edad, mediante el uso de funciones spline.

Trabajos futuros

- Analizar y evaluar el método propuesto cuando el **efecto de la co-variable no es constante** a lo largo de la distribución, es decir, la dependencia entre ellos cambia **según el cuantil** τ .
- Estudiar la **selección de variables** en este tipo de modelos de regresión cuantil.
- **Estimar regiones** de referencia condicionales en función de una **variable factor** de interés. ¿Se estima la misma región bivariada para hombres que para mujeres?
- **Actualizar el paquete refreg** de R, de manera que se pueda implementar el nuevo método propuesto.

Bibliografía

- 1 Lado-Baleato Ó, Roca-Pardiñas J, Cadarso-Suárez C, Gude F. Modeling conditional reference regions: Application to glycemic markers. *Statistics in Medicine*. 2021;40(26):5926–5946.
- 2 Roca-Pardiñas J, Ordoñez C, Lado-Baleato O. Nonparametric location–scale model for the joint forecasting of SO₂ and NO_x pollution episodes. *Stochastic Environmental Research and Risk Assessment*. 2021;35(2):231–244.
- 3 Lado-Baleato Ó, Roca-Pardiñas J, Cadarso-Suárez C, Gude F. Refreg: An R package for estimating conditional reference regions. *The R Journal*. 2022;14(2):253–269.
- 4 Young DS, Mathew T. Nonparametric hyperrectangular tolerance and prediction regions for setting multivariate reference regions in laboratory medicine. *Statistical Methods in Medical Research*. 2020;29(12):3569–3585.

Estimación robusta de regiones de referencia condicionales con R. Aplicación en la investigación de diabetes

Sara Rodríguez-Pastoriza¹
Pardiñas^{1,3}

Óscar Lado-Baleato²
Francisco Gude^{4,5,6}

Javier Roca

¹Department of Statistics and Operations Research, University of Vigo, 36310 Vigo, Galicia, Spain.

²Statistics and Operations Research Department, Oviedo University, Asturias, Spain.

³Galician centre for mathematical research and technology, (CITMAGA), Galicia, Spain.

⁴Research Methods Group (RESMET), Health Research Institute of Santiago de Compostela (IDIS), Spain.

⁵Department of Psychiatry, Radiology, Public Health, Nursing and Medicine, University of Santiago de Compostela, Galicia, Spain.

⁶Concepción Arenal Primary Care Center, Santiago de Compostela, Galicia, Spain.

SiDOR
Statistical Inference
Decision & Operations Research Group

Universidade de Vigo

