Test robustos para superioridad entre dos curvas de regresión

Graciela Boente[1]

y

Juan Carlos Pardo–Fernández[2]

[1]Universidad de Buenos Aires and CONICET, Argentina
[2] Universidad de Vigo, España

El Problema

Dos muestras independientes $\{(X_{ij}, Y_{ij}), i = 1, \ldots, n_j\}$, $j = 1, 2$.

$(X_{ij}, Y_{ij}) \sim (X_j, Y_j)$, $j = 1, 2$, que cumplen **Modelos de regresión nonparametricos**

$$Y_j = m_j(X_j) + \varepsilon_j \qquad j = 1, 2$$

- m_j es una función suave
- ε_j es el error de regresiíon, independiente de las covariables.

Objetivo: testear igualdad de las funciones de regresión versus alternativas unilaterales.

Motivación: Datos Reales

Conjunto de datos del Archivo de Datos del *Journal of Applied Econometrics* usados por Neumeyer and Pardo–Fernández (2009).

- Los datos se relacionan con el gasto total de varios hogares holandeses. En particular, estos autores testearon igualdad de las curvas de regresión que explican la relación entre
  - $X =$ **'log del gasto total'**
  - $Y =$ **'log del gasto en comida'**

  según la cantidad de miembros del hogar.
- La naturaleza de las variables consideradas justifica el uso de un **test unilateral**, ya que se espera que el gasto en comida crezca o al menos no decrezca cuando la cantidad de miembros del hogar crece.

Motivación: Datos Reales

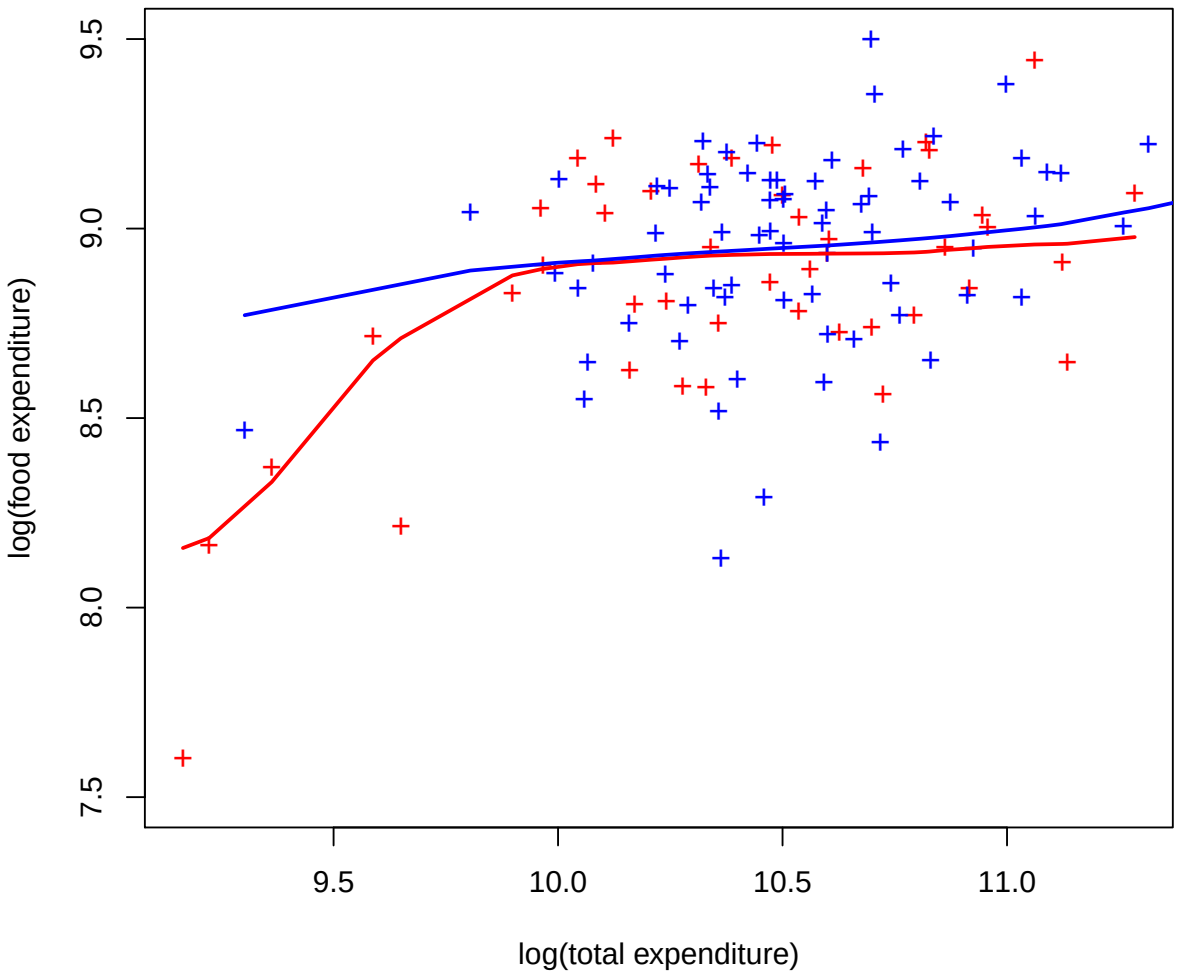


Gráfico de Y='log (gasto en comida)' versus X='log (gasto total)' y estimadores de núcleos de la función de regresión de hogares con 3 integrantes y 4 integrantes.

El Modelo

Sean (X_j, Y_j), $j = 1, 2$, dos vectores aleatorios que siguen el **modelo de regresión no paramétrico**

$$Y_j = m_j(X_j) + \varepsilon_j$$

- m_j es una función suave
- ε_j es el error de regresiíon, independiente de las covariables y tal que

  $$\varepsilon_j = \sigma_j \, U_j,$$

  con $U_j \sim G_j(\cdot)$ con escala 1.

Objetivo: testear igualdad de las funciones de regresión
$H_0 : m_1 = m_2$ versus alternativas unilaterales.

Hipótesis nula y alternativa

La **Hipótesis nula** es

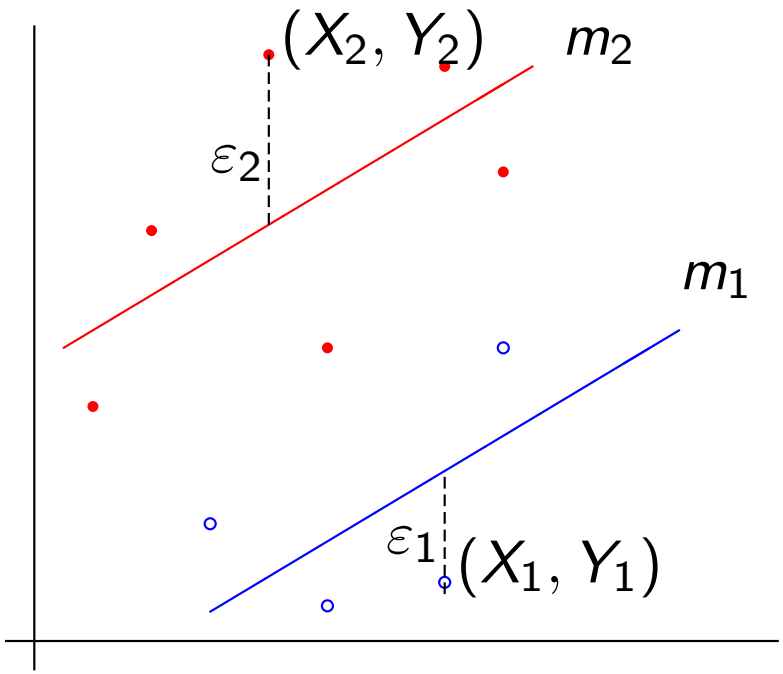
$$H_0 : m_1(x) = m_2(x) \text{ for all } x \in \mathcal{R},$$

con $\mathcal{R}$ la parte común del soporte de la distribución de las covariables X_1 y X_2 donde se realizará la comparación.

La **Hipótesis alternativa** es del tipo **unilateral**

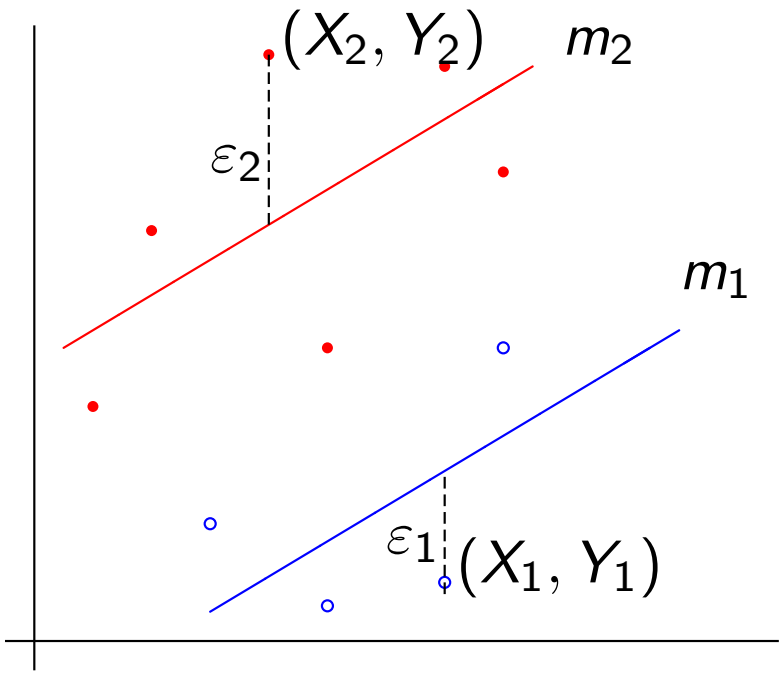
$$\begin{aligned} H_1 : \quad & m_1(x) \leq m_2(x) \text{ para todo } x \in \mathcal{R} \\ & m_1(x) < m_2(x) \text{ para } x \in \mathcal{A} \subset \mathcal{R}, \\ & \text{donde } \mathcal{A} \text{ es tal que } \mathbb{P}(X_j \in \mathcal{A}) > 0, \text{ para } j = 1, 2. \end{aligned}$$

Idea del test

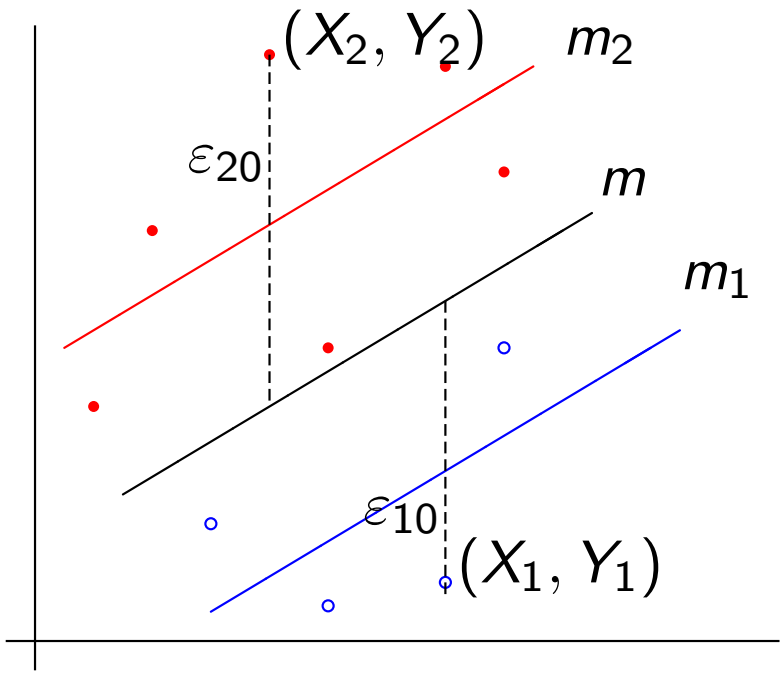


$$\mathbb{E}(\varepsilon_1) = 0 \quad \text{y} \quad \mathbb{E}(\varepsilon_2) = 0$$

Idea del test



$$\mathbb{E}(\varepsilon_1) = 0 \quad \text{y} \quad \mathbb{E}(\varepsilon_2) = 0$$



$$\mathbb{E}(\varepsilon_{10}) < 0 \quad \text{y} \quad \mathbb{E}(\varepsilon_{20}) > 0$$

$$\varepsilon_{j0} = Y_j - m(X_j)$$

Idea del test

- $m(x) = p_1(x)m_1(x) + p_2(x)m_2(x)$, $0 \leq p_1(x) \leq 1$
  $p_2(x) = 1 - p_1(x)$

$$m_1(x) \leq m(x) \leq m_2(x) \quad \text{para todo } x \in \mathcal{R}$$

$$\varepsilon_{j0} = Y_j - m(X_j) \qquad j = 1, 2\,.$$

El test clásico usa que $\mathbb{E}(\varepsilon_{20}) - \mathbb{E}(\varepsilon_{10}) > 0$

Idea del test

- $m(x) = p_1(x)m_1(x) + p_2(x)m_2(x)$, $0 \leq p_1(x) \leq 1$
  $p_2(x) = 1 - p_1(x)$

$$m_1(x) \leq m(x) \leq m_2(x) \quad \text{para todo } x \in \mathcal{R}$$

$$\varepsilon_{j0} = Y_j - m(X_j) \qquad j = 1, 2\,.$$

El test clásico usa que $\mathbb{E}(\varepsilon_{20}) - \mathbb{E}(\varepsilon_{10}) > 0$

- Sea Ψ una **función no decreciente y acotada**
  - $\Psi(t) = \min(k, \max(-k, t))$,
  - $\Psi(t) = t/\sqrt{1 + t^2/k^2}$,
  - $\Psi(t) = k \arctan(t/k)$.

- w_j, $j = 1, 2$, función de peso no–negativa con soporte compacto $\mathcal{S}_j$ tal que $\mathcal{A} \cap \mathcal{S}_1 \cap \mathcal{S}_2 \neq \emptyset$.

Idea del test

Dado $\sigma > 0$,

$$\mathbb{E}\left[\boldsymbol{\Psi}\left(\frac{\boldsymbol{\varepsilon_{10}}}{\boldsymbol{\sigma}}\right)\mathbf{w_1(X_1)}\right] \leq \mathbb{E}\left[\Psi\left(\frac{\varepsilon_1}{\sigma}\right)\right]\mathbb{E}[w_1(X_1)], \tag{1}$$

$$\mathbb{E}\left[\Psi\left(\frac{\varepsilon_2}{\sigma}\right)\right]\mathbb{E}[w_2(X_2)] \leq \mathbb{E}\left[\boldsymbol{\Psi}\left(\frac{\boldsymbol{\varepsilon_{20}}}{\boldsymbol{\sigma}}\right)\mathbf{w_2(X_2)}\right]. \tag{2}$$

- Bajo H_0, las desigualdades en (3) y (4) son igualdades.
- Bajo H_1, **(3) o (4) o ambas desigualdades son estrictas** cuando, por ejemplo, Ψ es nodecreciente, estrictamente creciente en un entorno de 0 y los errores asignan masa positiva a ese entorno.

Idea del test

$$\mathbb{E}\left[\boldsymbol{\Psi}\left(\frac{\boldsymbol{\varepsilon_{10}}}{\boldsymbol{\sigma_1}}\right)\mathbf{w_1(X_1)}\right] \leq \mathbb{E}\left[\Psi\left(\frac{\varepsilon_1}{\sigma_1}\right)\right]\mathbb{E}[w_1(X_1)]\,, \tag{3}$$

$$\mathbb{E}\left[\Psi\left(\frac{\varepsilon_2}{\sigma_2}\right)\right]\mathbb{E}[w_2(X_2)] \leq \mathbb{E}\left[\boldsymbol{\Psi}\left(\frac{\boldsymbol{\varepsilon_{20}}}{\boldsymbol{\sigma_2}}\right)\mathbf{w_2(X_2)}\right]. \tag{4}$$

- Bajo H_0, las desigualdades en (3) y (4) son igualdades.
- Bajo H_1, **(3) o (4) o ambas desigualdades son estrictas** cuando, por ejemplo, Ψ es nodecreciente, estrictamente creciente en un entorno de 0 y los errores asignan masa positiva a ese entorno.

Idea del test

Si

- $X_1 \sim X_2$, $w_1 = w_2$ y $U_1 \sim U_2$

  o
- $\mathbb{E}[\Psi(U_j)] = \mathbb{E}[\Psi(\varepsilon_j/\sigma_j)] = 0$, para $j = 1, 2$,

Para distinguir H_1 de H_0 parece razonable considerar.

$$\begin{aligned} \mathcal{T}_0 &= E_{02} - E_{01} \\ &= \mathbb{E}\left[\Psi\left(\frac{\varepsilon_{20}}{\sigma_2}\right) w_2(X_2)\right] - \mathbb{E}\left[\Psi\left(\frac{\varepsilon_{10}}{\sigma_1}\right) w_1(X_1)\right] \geq 0\,, \end{aligned}$$

$$\varepsilon_{j0} = Y_j - m(X_j) \qquad j = 1, 2\,.$$

Estadístico del test

Para realizar el test, necesitamos **estimadores robustos y consistentes** de $m(x) = p_1(x)m_1(x) + p_2(x)m_2(x)$ y σ_j.

- **Muestras:** i.i.d. observaciones $\{(X_{ij}, Y_{ij}), 1 \leq i \leq n_j\}$, $j = 1, 2$.
- $n = n_1 + n_2$.
- $\widehat{m}_j(x)$ $M-$**estimador local robusto** de $m_j(x)$ basado en una función de escores Ψ_j con ventana h_j. ▸ $\widehat{m}_j(x)$
- $\widehat{m}(x) = p_1(x)\widehat{m}_1(x) + p_2(x)\widehat{m}_2(x)$ estimador de m bajo H_0,
- $\widehat{\sigma}_j$ **estimadores robustos consistentes** de σ_j. ▸ $\widehat{\sigma}_j$

Estadístico del test

Para distinguir H_1 de H_0 consideramos el funcional

$$\mathcal{T}_0 = E_{02} - E_{01} = \mathbb{E}\left[\Psi\left(\frac{\varepsilon_{20}}{\sigma_2}\right) w_2(X_2)\right] - \mathbb{E}\left[\Psi\left(\frac{\varepsilon_{10}}{\sigma_1}\right) w_1(X_1)\right] \geq 0\,,$$

El **Estadístico del test** es

$$T = \left(\frac{n_1 n_2}{n}\right)^{1/2} (\widehat{E}_{20} - \widehat{E}_{10})$$

con

$$\widehat{E}_{j0} = \frac{1}{n_j} \sum_{i=1}^{n_j} \Psi\left(\frac{Y_{ij} - \widehat{m}(X_{ij})}{\widehat{\sigma}_j}\right) w_j(X_{ij})\,.$$

H_0 se rechazará para valores grandes positivos de T.

Distribución asintótica del estadístico del test

Bajo supuestos generales, $n_j/n \to \kappa_j$, $0 < \kappa_j < 1$,

(a) Bajo H_0, $T \xrightarrow{D} N(0, \sigma_T^2)$

(b) Bajo H_1, $T \xrightarrow{p} \infty$.

(c) Sea $\Delta : \mathbb{R} \to \mathbb{R}$ tal que $\Delta(x) \geq 0$ para todo $x \in \mathcal{R}$.

Bajo $H_{1n} : m_2(x) = m_1(x) + n^{-1/2}\,\Delta(x)$, $T \xrightarrow{D} N(c, \sigma_T^2)$.

El test rechaza H_0 si $T > z_{1-\alpha}\,\widehat{\sigma}_T$

- $z_{1-\alpha}$ es el percentil $(1-\alpha)$ de una normal estándar.
- $\widehat{\sigma}_T^2$ es un estimador consistente de σ_T^2

Estudio de Monte Carlo

Objetivo: comparar T_{CL} (Neumeyer & Pardo–Fernández, 2009) con T_{R}.

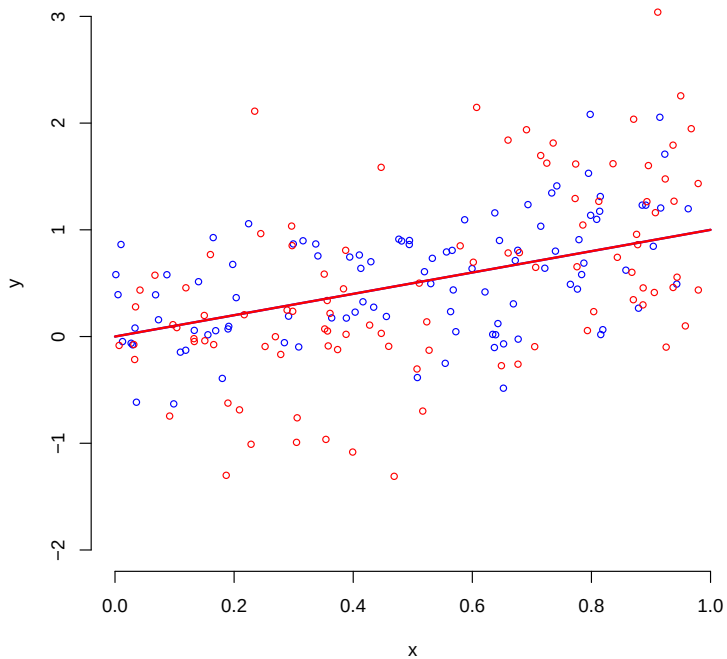
- Covariables: $X_j \sim \mathcal{U}[0,1]$
- Pesos: $w_1 = w_2 = \mathbb{I}_{(0,1)}$
- Funciones de Regressión:

  $m_1(x) = x$

  $m_2(x) = m_1(x) + \Delta n^{-1/2}(\sin(2\pi x)+1)$

  (alternativas locales)

- $p_1(x) = p_2(x) = 0.5$
- $\sigma_1 = 0.5$ y $\sigma_2 = 0.75$



Estudio de Monte Carlo

Objetivo: comparar T_{CL} (Neumeyer & Pardo–Fernández, 2009) con T_{R}.

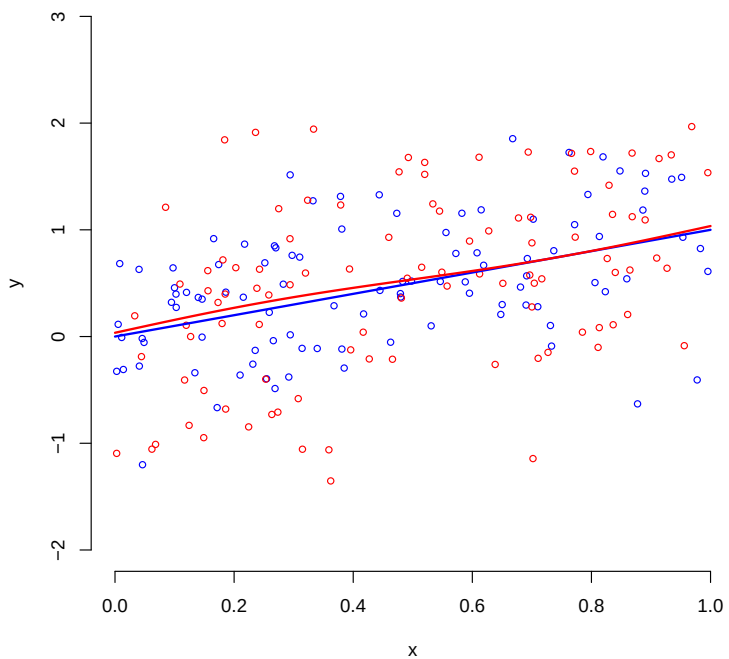
- Covariables: $X_j \sim \mathcal{U}[0,1]$
- Pesos: $w_1 = w_2 = \mathbb{I}_{(0,1)}$
- Funciones de Regressión:

  $m_1(x) = x$

  $m_2(x) = m_1(x) + \Delta n^{-1/2}(\sin(2\pi x)+1)$

  (alternativas locales)

- $p_1(x) = p_2(x) = 0.5$
- $\sigma_1 = 0.5$ y $\sigma_2 = 0.75$



Estudio de Monte Carlo

Objetivo: comparar T_{CL} (Neumeyer & Pardo–Fernández, 2009) con T_{R}.

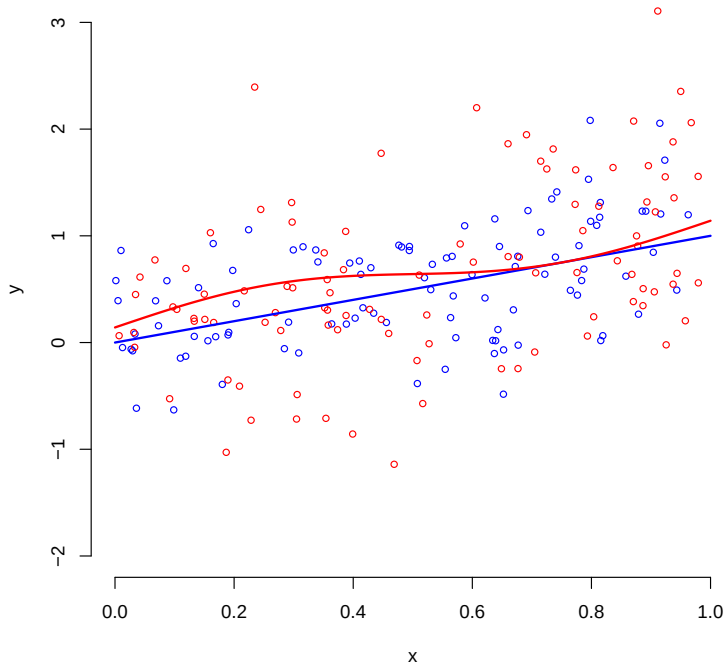
- Covariables: $X_j \sim \mathcal{U}[0,1]$
- Pesos: $w_1 = w_2 = \mathbb{I}_{(0,1)}$
- Funciones de Regressión:

  $m_1(x) = x$

  $m_2(x) = m_1(x) + \Delta n^{-1/2}(\sin(2\pi x)+1)$

  (alternativas locales)

- $p_1(x) = p_2(x) = 0.5$
- $\sigma_1 = 0.5$ y $\sigma_2 = 0.75$



Estudio de Monte Carlo

Objetivo: comparar T_{CL} (Neumeyer & Pardo–Fernández, 2009) con T_{R}.

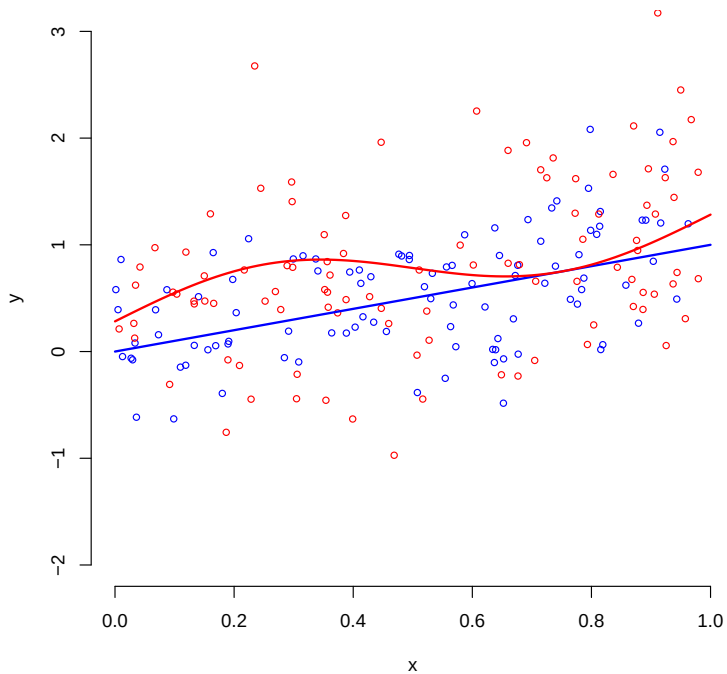
- Covariables: $X_j \sim \mathcal{U}[0,1]$
- Pesos: $w_1 = w_2 = \mathbb{I}_{(0,1)}$
- Funciones de Regressión:

  $m_1(x) = x$

  $m_2(x) = m_1(x) + \Delta n^{-1/2}(\sin(2\pi x)+1)$

  (alternativas locales)

- $p_1(x) = p_2(x) = 0.5$
- $\sigma_1 = 0.5$ y $\sigma_2 = 0.75$



Estudio de Monte Carlo

Objetivo: comparar T_{CL} (Neumeyer & Pardo–Fernández, 2009) con T_{R}.

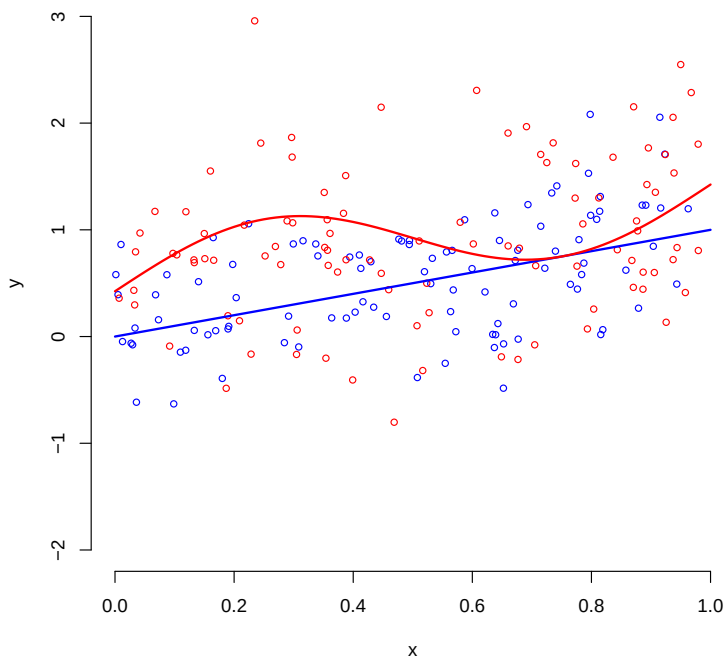
- Covariables: $X_j \sim \mathcal{U}[0,1]$
- Pesos: $w_1 = w_2 = \mathbb{I}_{(0,1)}$
- Funciones de Regressión:

  $m_1(x) = x$

  $m_2(x) = m_1(x) + \Delta n^{-1/2}(\sin(2\pi x)+1)$

  (alternativas locales)

- $p_1(x) = p_2(x) = 0.5$
- $\sigma_1 = 0.5$ y $\sigma_2 = 0.75$



Estudio de Monte Carlo

Objetivo: comparar T_{CL} (Neumeyer & Pardo–Fernández, 2009) con T_{R}.

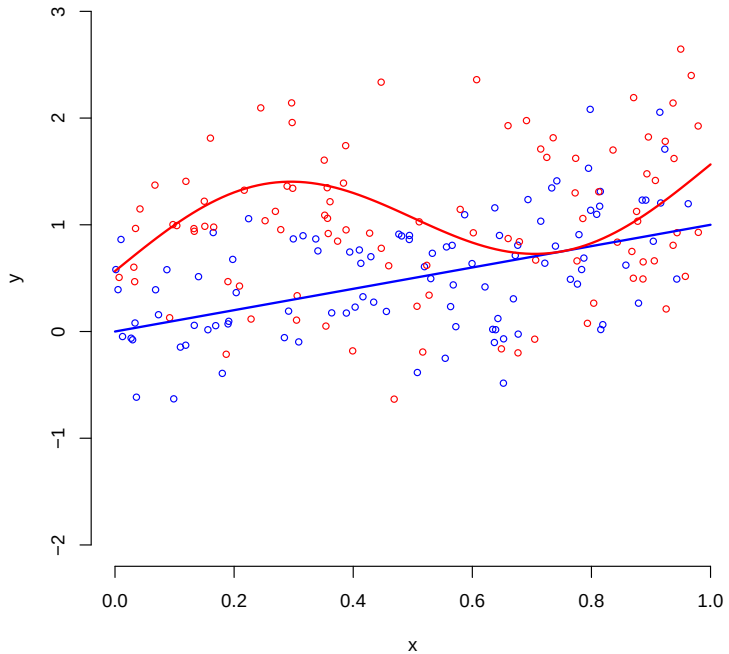
- Covariables: $X_j \sim \mathcal{U}[0,1]$
- Pesos: $w_1 = w_2 = \mathbb{I}_{(0,1)}$
- Funciones de Regressión:

  $m_1(x) = x$

  $m_2(x) = m_1(x) + \Delta n^{-1/2}(\sin(2\pi x)+1)$

  (alternativas locales)

- $p_1(x) = p_2(x) = 0.5$
- $\sigma_1 = 0.5$ y $\sigma_2 = 0.75$



Estudio de Monte Carlo

Errores:

$\mathcal{N}_0$ (sin outliers): $\varepsilon_j \sim N(0, \sigma_j^2)$.

$\mathcal{T}_1$ (sin momentos): $\varepsilon_j \sim \mathcal{C}(0, 25\sigma_j^2)$, con $\mathcal{C}(\mu, \sigma^2)$ la distribución Cauchy con posición μ y dispersión σ^2.

C_{1,π_1,π_2} (errores con contaminación):

$$\varepsilon_j \sim (1 - \pi_j)\, N(0, \sigma_j^2) + \pi_j\, N(0, 25\sigma_j^2).$$

$C_{2,c}$ (Agregamos 1 outlier a $\mathcal{N}_0$): Genere como en $\mathcal{N}_0$.

Sea $X_{(1),1} \leq \cdots \leq X_{(n_1),1}$

- $(X_{(1),1}, Y_{D_{1,1},1})^{\mathrm{T}}, \ldots, (X_{(n_1),1}, Y_{D_{n_1,1},1})^{\mathrm{T}}$,
- fije $X_{(\frac{n_1}{2}),1} = 0.5$ y $Y_{D_{\frac{n_1}{2}},1} = c$.

Estudio de Monte Carlo

De ahora en más

- **Funciones de escores** para estimación robusta y para test:

  Ψ, Ψ_1 y Ψ_2

  Función de Huber $\psi_{k,\mathrm{H}}(t) = \min(k, \max(-k, t))$ con constante $k = 1.345$

$\longrightarrow \quad T_{\mathrm{R}}$.

- **Ventana(s)**:

  Para T_{CL}: L^2 convalidación cruzada
  Para T_{R}: convalidación cruzada robusta

Estudio de Monte Carlo . Frecuencias de rechazo $n_1 = n_2 = 100$

T_{CL}: Test clásico

Línea continua: $\mathcal{N}_0$

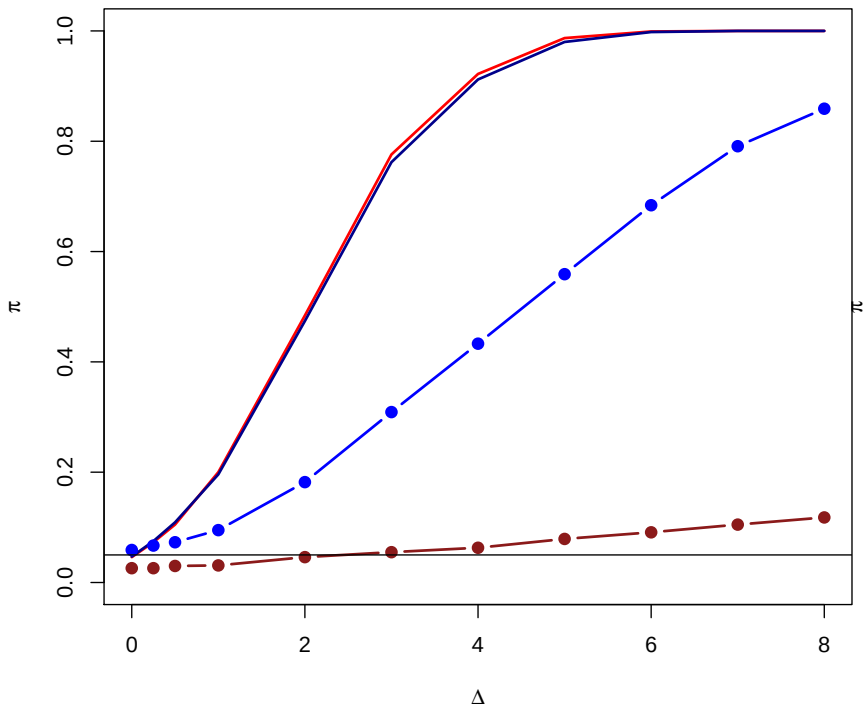
•: $\mathcal{T}_1$

T_{R}: Nuestra propuesta

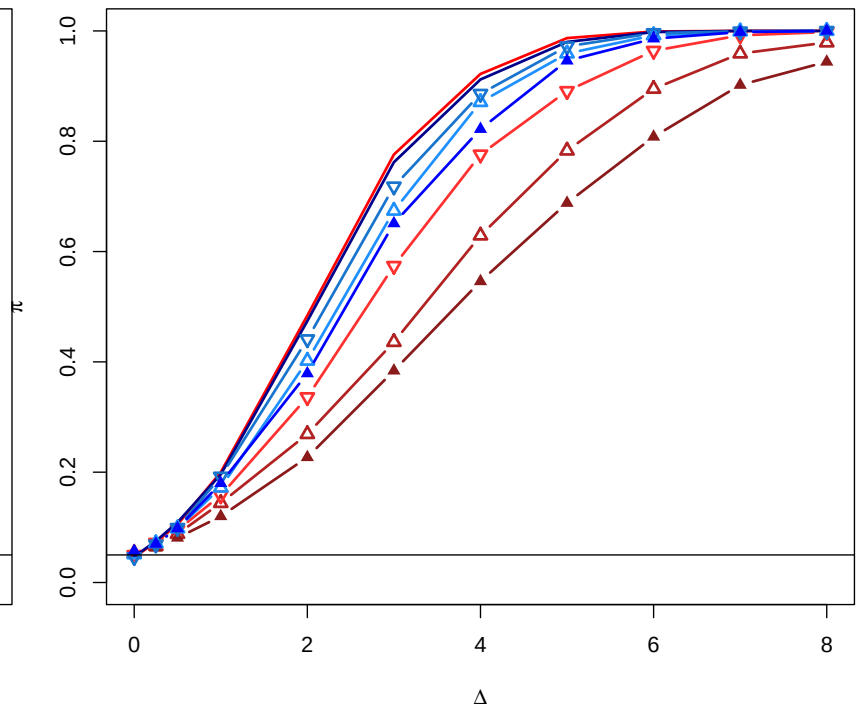
△: $C_{1,0,0.1}$

▽ $C_{1,0.1,0}$, ▲ $C_{1,0.1,0.1}$

$\mathcal{T}_1$



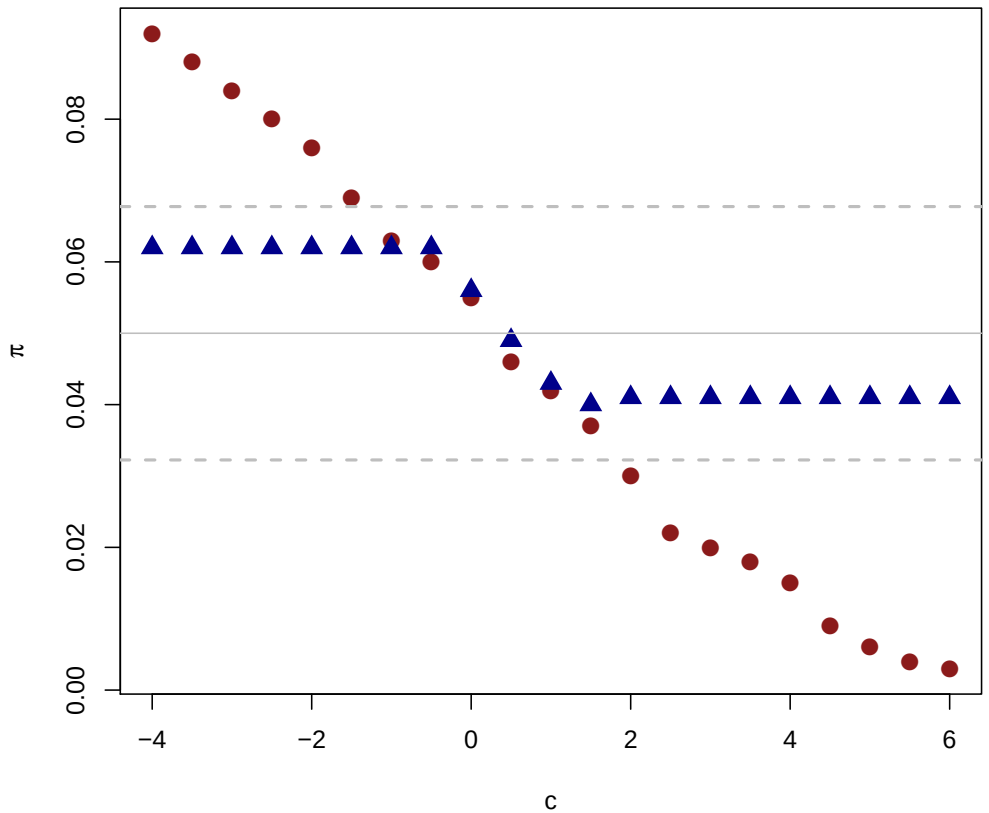
C_{1,π_1,π_2}



La línea horizontal indica el nivel nominal $\alpha = 0.05$.

Estudio de Monte Carlo. Nivel empírico bajo $C_{2,c}$

T_{CL}: Test clásico T_{R}: Nuestra propuesta



La línea horizontal indica el nivel nominal $\alpha = 0.05$.

Las líneas cortadas representan la región de aceptación para testear si el nivel empírico es significativamente distinto del nominal, con nivel 0.05.

Análisis de datos reales

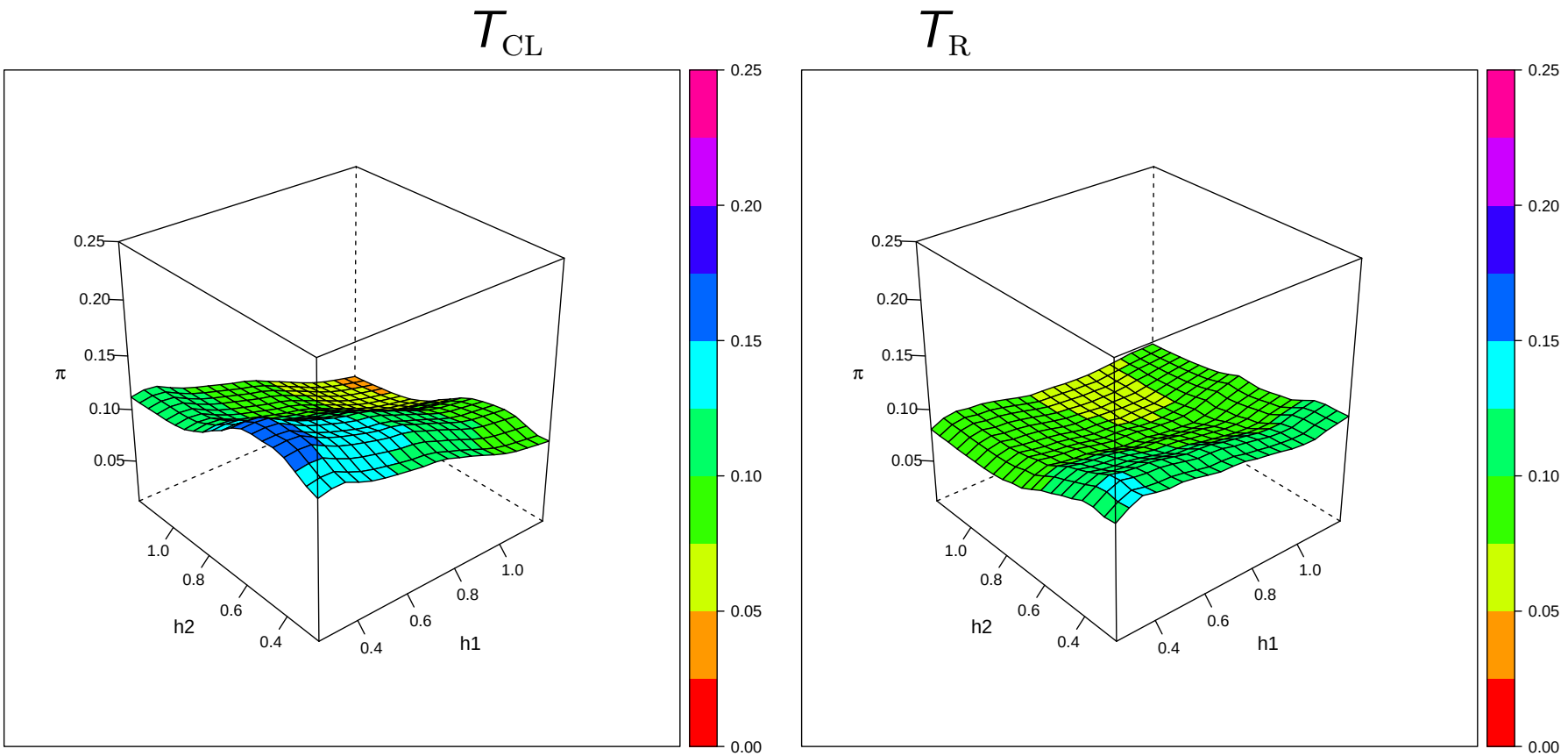
Neumeyer and Pardo–Fernández (2009) usaron un conjunto de datos del Archivo de Datos del *Journal of Applied Econometrics* para ilustrar su procedimiento.

- Los datos se relacionan con el gasto total de varios hogares holandeses.
- Igualdad de las curvas de regresión que explican la relación entre
  - $X =$ **'log del gasto total'**
  - $Y =$ **'log del gasto en comida'**

  según la cantidad de miembros del hogar.
- **Test Unilateral**
- Los $p-$valores son 0.125 para T_{CL} y 0.102 para T_{R}.

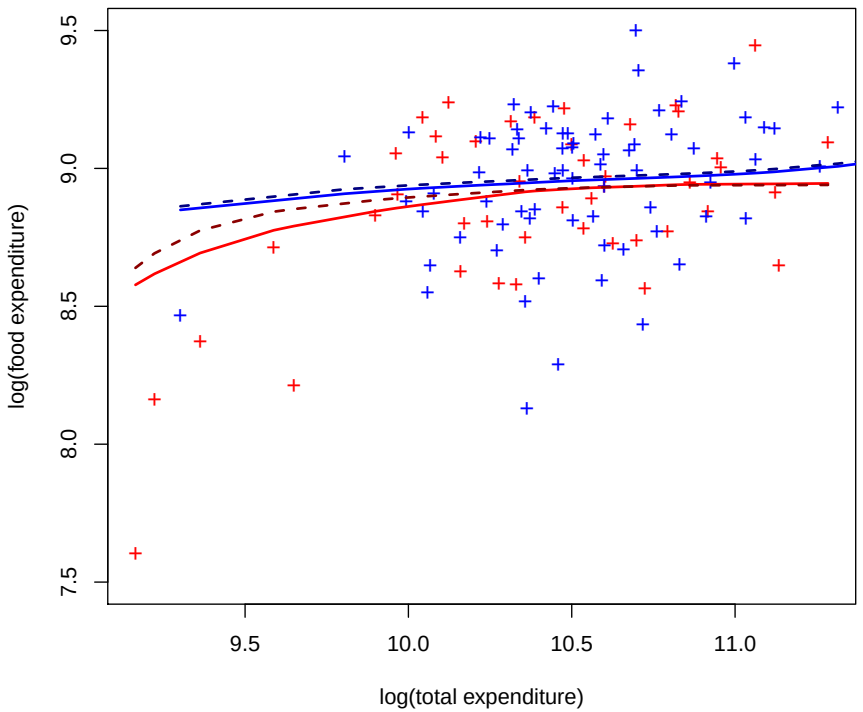
Análisis de datos reales: p-valores para distintos valores de h_1 y h_2,

$$w_j = \mathbb{I}_{9 \leq x \leq 12}$$



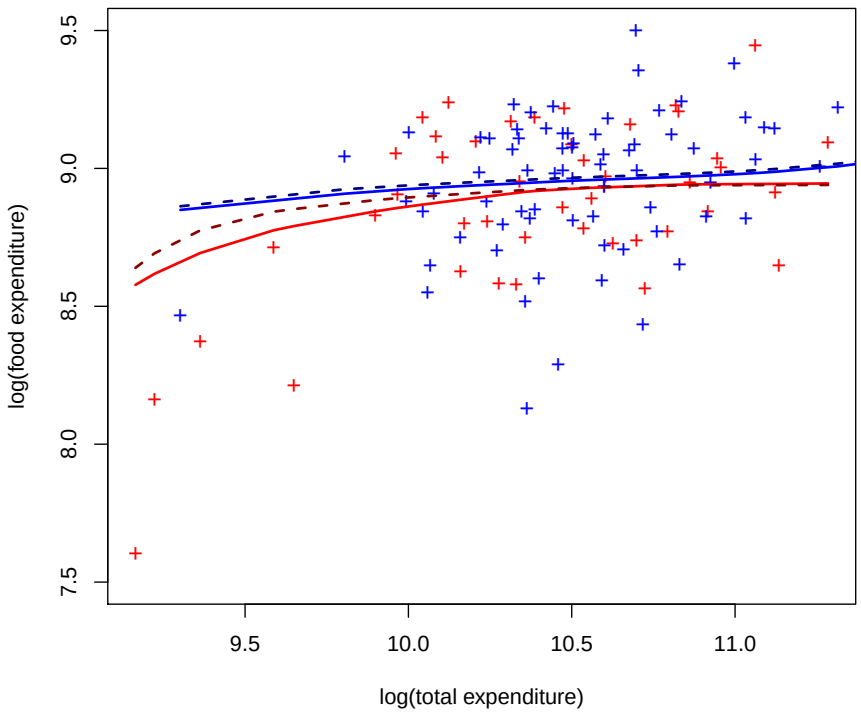
Análisis de datos reales: Estimadores de regresión $h_1 = h_2 = 1.2$

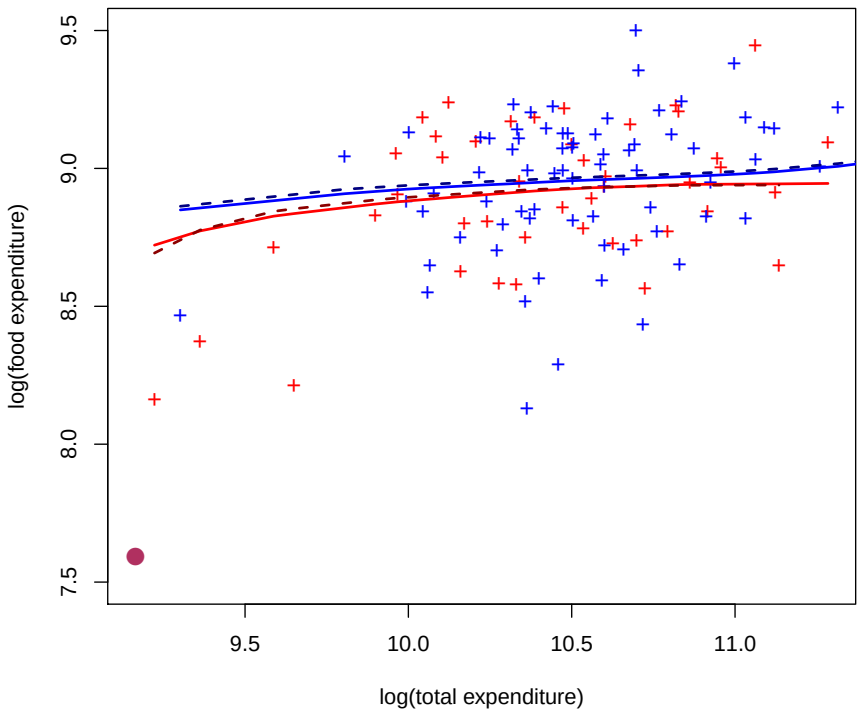
- Líneas continuas: estimadores de Nadaraya Watson
- Líneas cortadas: $M-$estimadores locales



Análisis de datos reales: Estimadores de regresión $h_1 = h_2 = 1.2$

- Líneas continuas: estimadores de Nadaraya Watson
- Líneas cortadas: $M-$estimadores locales

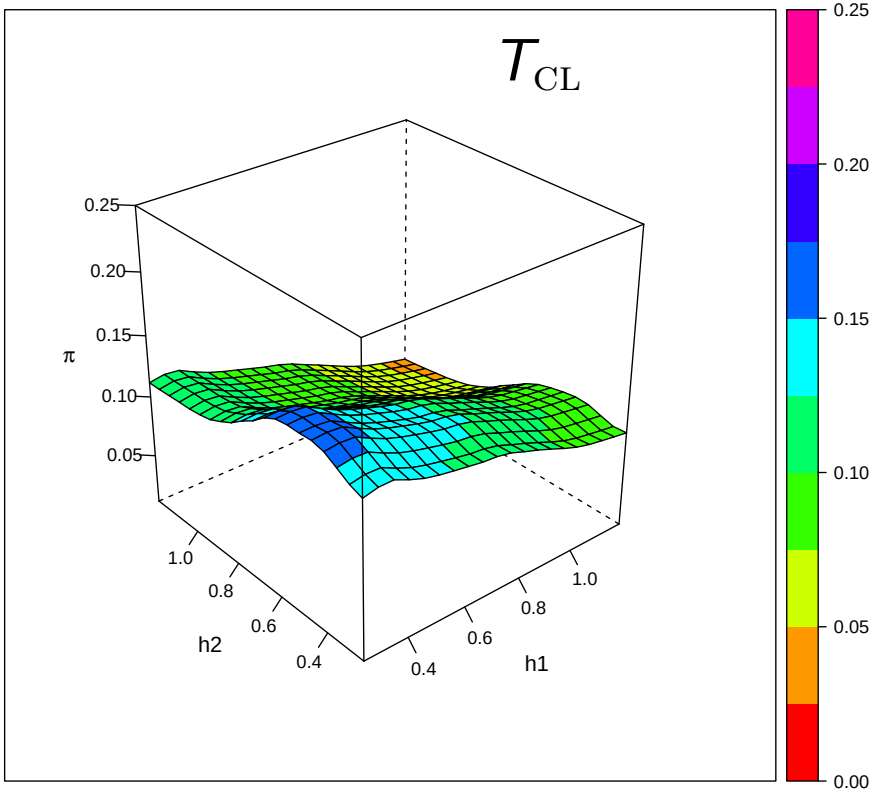


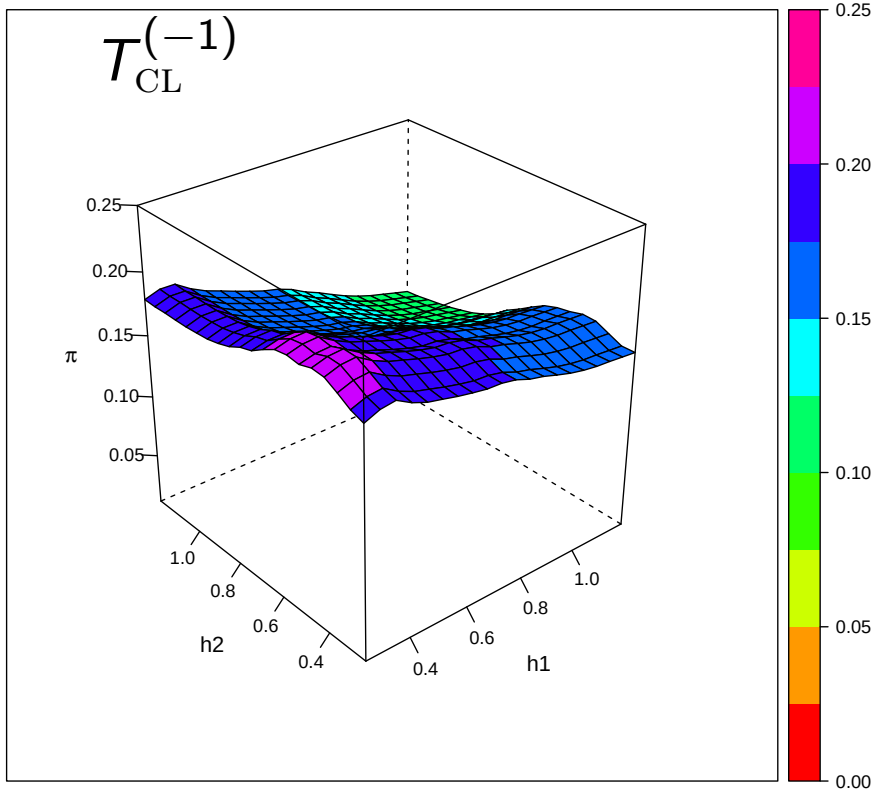


$\widehat{m}_{1,\mathrm{CL}}^{(-1)}(x)$

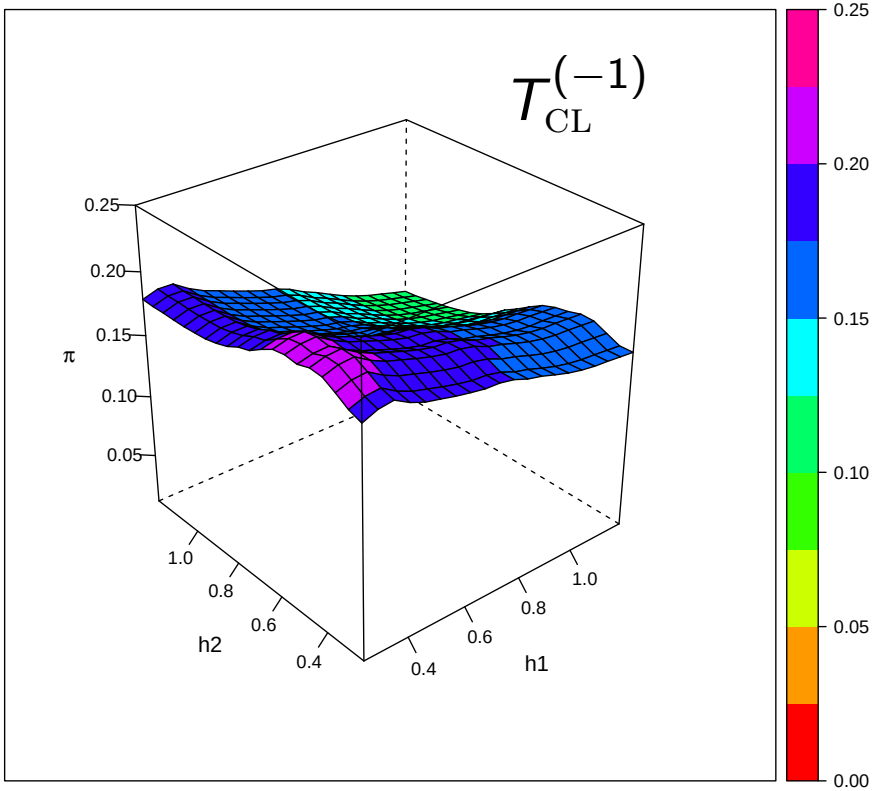
Análisis de datos reales: p-valores para distintos valores de h_1 y h_2,

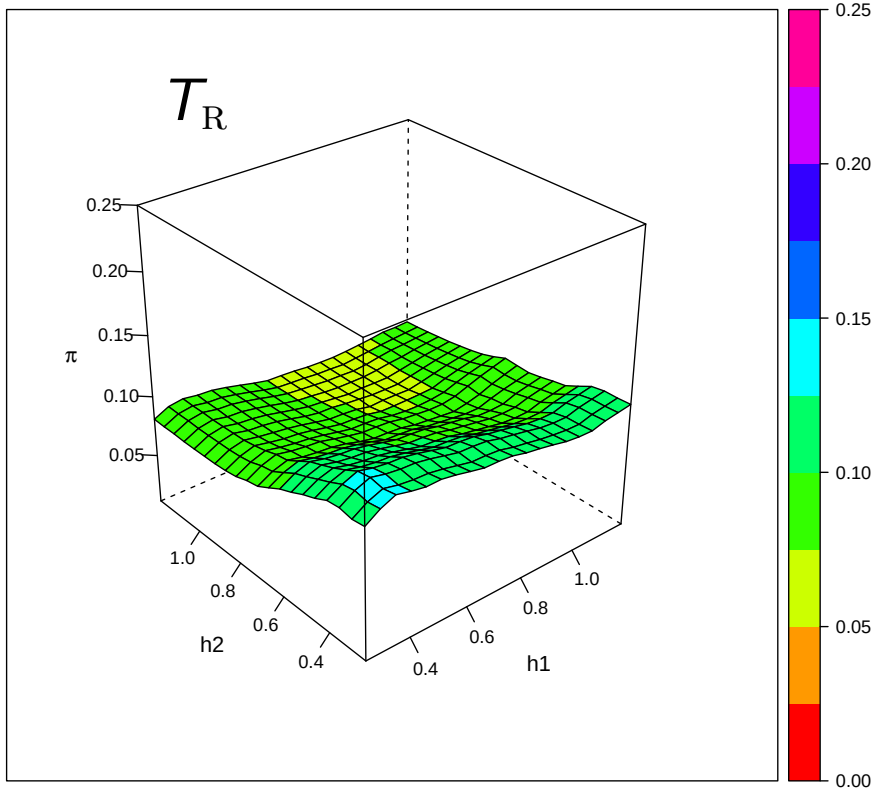
$$w_j = \mathbb{I}_{9 \leq x \leq 12}$$





Análisis de datos reales: *p*-valores





Conclusiones

- Hemos propuesto y estudiado un nuevo método **robusto** para **testear la igualdad de las dos curvas de regresión** versus una alternativa unilateral en un contexto no paramétrico

- El nuevo procedimiento adapta las ideas en Neumeyer and Pardo-Fernández (2009) a la situación en que los errores no tienen **no momentos**.

- Los procedimientos robustos permiten tener
  - puntos de diseño con **densidades diferentes**
  - errores con **distinta distribución** si son simétricas

- SI $\Psi = \Psi_1 = \Psi_2$, el procedimiento da origen a un test consistente aún bajo errores con distribución asimétrica si tienen la misma distribución.

Conclusiones

- El anaálisis de la **Distribución asintótica del Estadístico del test** revela que el procedimiento es consistente versus alternativas locales que convergen a la hipótesis nula a tasa paramétrica $n^{-1/2}$.
- **Valores críticos** pueden obtenerse de la distribución asintótica del Estadístico del test bajo H_0.
- Buen comportamiento en las simulaciones.

Muchas Gracias!!

Estimador robusto de la regresión

El **estimador robusto de** $m_j(x)$ está dado por

$$\text{la solución } \widehat{m}_j(x) \text{ de } \widehat{\lambda}_j(x, \widehat{m}_j(x), \widehat{\sigma}_j) = 0,$$

con

$$\widehat{\lambda}_j(x, a, \sigma) = \sum_{i=1}^{n_j} K_h(x - X_{ij}) \Psi_j\left(\frac{Y_{ij} - a}{\sigma}\right).$$

- K es un núcleo (usualmente una densidad positiva)
- $h = h_n$ una sucesión de ventanas $h_n > 0$
- $K_h(u) = h^{-1}K(u/h)$
- Ψ_j es una función **impar**

$$\widehat{m}(x) = p_1(x)\widehat{m}_1(x) + p_2(x)\widehat{m}_2(x)$$

Volver

Estimador de escala robusto

Bajo el modelo homoscedastico, se pueden obtener **estimadores robustos con tasa $\sqrt{n}$ para la escala σ**.

Sean $X_{(1),j} \leq \cdots \leq X_{(n_j),j}$ los estadísticos de orden de las variables explicativas de la población $j-$ésima. Indiquemos as $(X_{(1),j}, Y_{D_{1,j},j})^{\mathrm{T}}, \ldots, (X_{(n_j),j}, Y_{D_{n_j,j},j})^{\mathrm{T}}$ la muestra de las observaciones ordenadas según los valores de las covariables, o sea, $X_{(\ell),j} = X_{D_{\ell,j},j}$.

Un estimador robusto consistente con tasa $\sqrt{n}$ está dado por

$$\widehat{\sigma}_j = \frac{1}{\sqrt{2}\Phi^{-1}(3/4)} \underset{1\leq\ell\leq n_j-1}{\text{median}} \left| Y_{D_{\ell+1,j},j} - Y_{D_{\ell,j},j} \right|,$$

donde el coeficiente $\sqrt{2}\Phi^{-1}(3/4)$ asegura consistencia Fisher para errores normales (Φ indica la función de distribución de una variable $N(0,1)$). ▸ Volver