Proyecto Fin de Máster

Contraste de Bondad de Ajuste de Modelos de Regresión Cuantil

Alumna: Mercedes Conde Amboage

Tutores: César A. Sánchez Sellero
Wenceslao González Manteiga

Máster en Técnicas Estadísticas
Universidad de Santiago de Compostela
Julio 2013

Proyecto Fin de Máster

Contraste de Bondad de Ajuste de Modelos de Regresión Cuantil

Alumna: Mercedes Conde Amboage

Tutores: César A. Sánchez Sellero
Wenceslao González Manteiga

El presente documento recoge el Proyecto Fin de Máster para el Máster en Técnicas Estadísticas realizado por Dª. Mercedes Conde Amboage bajo el título 'Contraste de Bondad de Ajuste de Modelos de Regresión Cuantil'.

Ha sido realizado bajo la dirección de D. Wenceslao González Manteiga y D. César A. Sánchez Sellero, que lo consideran terminado y dan su conformidad para su presentación y defensa.

Santiago de Compostela, a 8 de julio de 2013.

Fdo.: Wenceslao González Manteiga

Fdo.: César A. Sánchez Sellero

Fdo.: Mercedes Conde Amboage

Resumen

La regresión cuantil se utiliza cuando el objetivo de estudio se centra en la estimación de las diferentes posiciones (cuantiles) de una variable de interés en función de ciertas variables explicativas. De esta manera facilita un análisis más completo y robusto de los datos. Por todo esto, la regresión cuantil es una técnica estadística muy útil en diversos ámbitos de aplicación, como por ejemplo, la Ecología, la Economía o la Medicina.

Como consecuencia de la gran difusión de la que ha gozado la regresión cuantil a lo largo de los últimos años, hemos desarrollado un contraste de bondad de ajuste en este contexto basado en proyecciones para evitar de este modo el desastre de la dimensionalidad. A la largo de este trabajo presentaremos el estadístico de contraste asociado a la nueva propuesta de test y comentaremos su comportamiento a la vista de un estudio de simulación. Además, aplicaremos el contraste de bondad de ajuste propuesto a un conjunto de datos reales.

Debo agradecer al Ministerio de Educación, Cultura y Deporte por la concesión de la beca FPU AP2012-5047 que ha servido de apoyo para las investigaciones incluidas en este trabajo.

Índice general

Capítulo 1

Introducción

A día de hoy la regresión cuantil es un tema de máximo interés para los investigadores en Estadística, que están adaptando gran parte de las técnicas de inferencia relacionadas con la tradicional regresión en media estimada por mínimos cuadrados a los modelos de regresión cuantil. La razón es que los modelos de regresión cuantil permiten una descripción más detallada del comportamiento de la variable respuesta, se adaptan a situaciones bajo condiciones más generales de la distribución del error, gozan de propiedades de robustez y permiten abordar problemas de regresión con datos complejos (como por ejemplo, los datos censurados), en muchos casos en mejores condiciones que una regresión en media.

Dentro del contexto de la regresión cuantil, nuestro objetivo será proponer un nuevo contraste de bondad de ajuste que permita contrastar modelos de regresión cuantil con muchas variables explicativas, superando de este modo el conocido desastre de la dimensionalidad.

Empezaremos este primer capítulo estableciendo el concepto de regresión cuantil (Sección 1.1). A continuación presentamos un breve resumen acerca de contrastes de bondad de ajuste en el caso de la regresión en media (Sección 1.2) y de la regresión cuantil (Sección 1.3).

1.1. La regresión cuantil

Aunque la regresión en media, ajustada por el método de mínimos cuadrados, ha alcanzado la mayor difusión en la Estadística del siglo XX, resulta muy llamativo observar que las ideas de regresión cuantil fueron anteriores a los procedimientos basados en los mínimos cuadrados. Así, mientras el inicio de la regresión por mínimos cuadrados se puede datar en el año 1805 por el trabajo de Legendre, a mediados del siglo XVIII Boscovich ya ajustó datos sobre la elipticidad de la Tierra mediante procedimientos de regresión cuantil.

El método de mínimos cuadrados gozó de la ventaja que le proporcionaba la existencia de expresiones cerradas para la estimación, la sencillez de los argumentos de probabilidad y ciertos resultados de optimalidad. Aún así, siempre pesaba la duda sobre las hipótesis del modelo, y sobre la necesidad de una descripción más completa y flexible de la realidad.

Los métodos de regresión cuantil encontraron un gran desarrollo desde el surgimiento de la Estadística Robusta, que alcanzó una gran expansión a principios de los años 80. El libro de Huber (1981) o el de Hampel y otros (1986) son buenas recopilaciones de las aportaciones que hicieron sus autores a la Teoría de la Robustez, cuyos conceptos siguen siendo aplicados hoy en día a los métodos estadísticos modernos.

A lo largo de esta primera sección definiremos que se entiende por cuantil y estableceremos el concepto de modelo de regresión cuantil como solución de un problema de programación lineal, estableciendo previamente el concepto de función de pérdida cuantílica.

1.1.1. Definición del cuantil

Definición 1.1. Dada cualquier variable aleatoria $X : \Omega \to \mathbb{R}$ definida en un espacio muestral Ω asociado a un experimento aleatorio estará caracterizada por su **función de distribución** que viene determinada por la siguiente expresión:

$$F(x) = \mathbb{P}\big(X \leq x\big)$$

Si la variable aleatoria X es discreta (es decir, su recorrido es un conjunto discreto) entonces su función de distribución viene dada por:

$$F(x) = \mathbb{P}\big(X \leq x\big) = \sum_{x_i \leq x} \mathbb{P}\big(X = x_i\big)$$

donde $x_i \leq x$ representa todos aquellos valores que toma la variable X que son inferiores o iguales al valor dado x.

Por otra parte, si la variable aleatoria X es continua (es decir, si su recorrido no es un conjunto numerable) entonces se tiene que:

$$F(x) = \mathbb{P}\big(X \leq x\big) = \int_{-\infty}^{x} f(x)dx$$

para cualquier valor x donde $f : \mathbb{R} \to \mathbb{R}$ es una función no negativa e integrable, denominada **función de densidad**.

Definición 1.2. Dada cualquier variable aleatoria X, para cada $0 < \tau < 1$ se puede definir el **cuantil de orden** τ, que denotaremos por c_τ, como el valor que verifica que:

$$\mathbb{P}(X \leq c_\tau) \geq \tau$$

$$\mathbb{P}(X \geq c_\tau) \geq 1 - \tau$$

De todos modos, debemos tener en cuenta que si la variable aleatoria X que estamos tratando es una variable continua entonces se verifica que :

$$\mathbb{P}\big(X \leq c_\tau\big) = \mathbb{P}\big(X < c_\tau\big) = \tau$$

Aparece así la **función cuantil** de una distribución de probabilidad, que se define como la inversa de la función de distribución. En este momento podemos distinguir los siguientes casos:

1. Dada una función de distribución $F : \mathbb{R} \to (0, 1)$ continua y estrictamente monótona, la función cuantil, F^{-1}, devolvería un valor x tal que:

   $$\mathbb{P}\big(X \leq x\big) = \tau$$

   La situación sería similar a la siguiente:

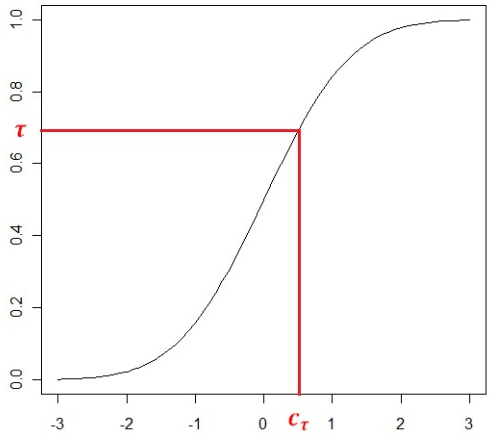


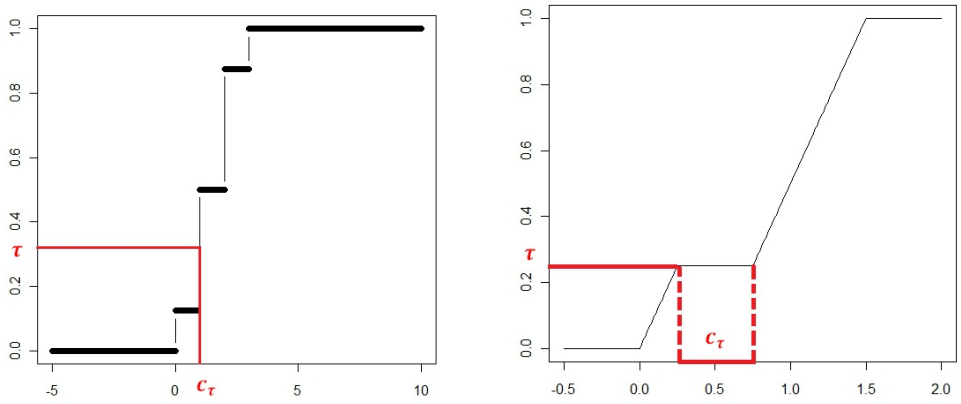
Figura 1.1: Ejemplo de función de distribución continua y estrictamente creciente.

2. Si la función de distribución fuese discreta, entonces puede haber saltos entre los valores del dominio de dicha función. Mientras que si la función de distribución es monótona no estricta, puede haber 'zonas llanas' (es decir, intervalos en los que el valor de la función se mantiene constante) en su rango.

   En cualquiera de los casos anteriores, la función inversa no estaría bien definida, por lo que se establece la siguiente definición alternativa:

   $$F^{-1}(\tau) = \text{ínf}\,\{x \in \mathbb{R} : \tau \leq F(x)\}$$

   Así para una probabilidad $0 < \tau < 1$, la función cuantil nos devolverá el valor mínimo de x para el cual se mantiene la probabilidad anterior. Las situaciones que estamos describiendo podrían ser las siguientes:



(a) (b)

Figura 1.2: Ejemplos de función de distribución discreta (a) y función de distribución no estrictamente monótona (b).

1.1.2. La función de pérdida cuantílica

Si consideramos un problema sencillo de teoría de la decisión, los diferentes cuantiles se pueden calcular como resultado de un problema de optimización. La **función de pérdida cuantílica** vendría determinada por la siguiente función lineal definida a trozos:

$$\rho_\tau(u) = u\big(\tau - \mathbb{I}(u < 0)\big) = \begin{cases} u\,\tau & \text{si } u \geq 0 \\ u\,(\tau - 1) & \text{si } u < 0 \end{cases}$$

cuya representación para diferentes valores del cuantil τ sería:

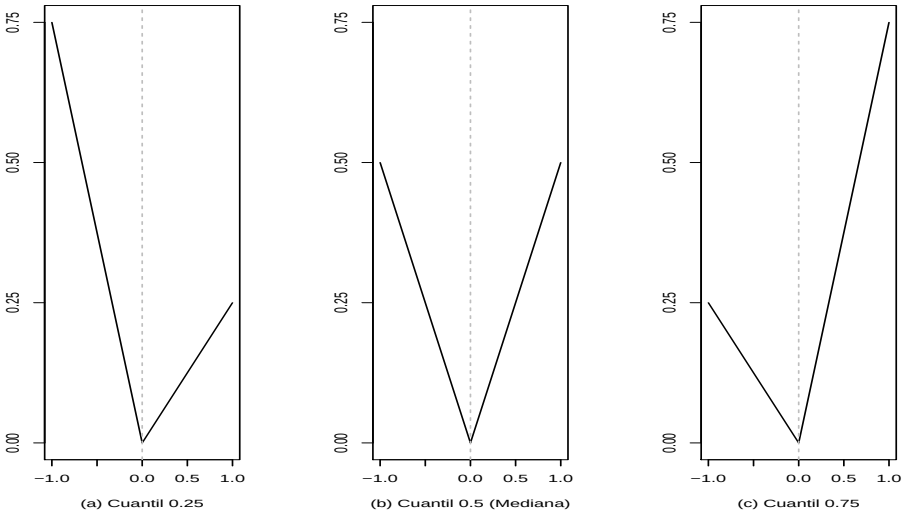


Figura 1.3: Representación de funciones de pérdida cuantílica para diferentes valores del cuantil τ.

Nota 1.1. Nótese que si $\tau = 0.25$ entonces se penalizan más las observaciones inferiores que las superiores. Como consecuencia, se arrastra la estimación hacia posiciones inferiores, por ser 0.25 un valor pequeño, o al menos más pequeño que 0.5. Con el cuantil 0.75 ocurre lo contrario, se penalizan más las desviaciones hacia arriba y la estimación queda más elevada.

Entonces para cada $\tau \in (0,1)$ pretendemos encontrar un elemento c_τ que minimice la pérdida esperada. Con lo cual, hemos reducido nuestro problema a un problema de teoría de la decisión. Es decir, lo que queremos es minimizar:

$$\mathbb{E}\Big[\rho_\tau(X-x)\Big] = (\tau-1)\int_{-\infty}^{x}(y-x)dF(y) + \tau\int_{x}^{\infty}(y-x)dF(y)$$

y si derivamos la expresión anterior con respecto a x tenemos que:

$$0 = (1-\tau)\int_{-\infty}^{x} dF(y) - \tau\int_{x}^{\infty} dF(y) = F(x) - \tau$$

Por lo tanto, si se verifica que F es una función estrictamente creciente entonces existe algún elemento del conjunto $\{x : F(x) \geq \tau\}$ que minimice la pérdida esperada. Cuando la solución es única entonces $c_\tau = F^{-1}(\tau)$ y en otro caso tenemos un intervalo de cuantiles τ de los cuales el más pequeño es el elemento c_τ que estamos buscando.

1.1.3. Optimización mediante programación lineal

Sea X una variable aleatoria que toma los valores $\{X_1, X_2, \cdots, X_n\}$ y de la cual no conoceremos la función de distribución (como ocurre en la mayor parte de los escenarios prácticos). Es decir, solo disponemos de la **función de distribución empírica** que recordemos viene dada por:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leq x)$$

En esta situación, la búsqueda del cuantil también se reduce a encontrar el elemento c_τ que minimice la pérdida esperada. O lo que es lo mismo, el elemento c que haga mínima la siguiente expresión:

$$\int \rho_\tau(x-c)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(X_i - c)$$

y así conseguimos calcular el cuantil de orden τ de la muestra, que denotamos por c_τ. Como ya hemos visto, este problema no siempre tendrá solución única, puesto que podríamos obtener un intervalo de cuantiles del mismo orden.

Lo realmente importante es que hemos logrado expresar el problema de la búsqueda de los cuantiles muestrales como un problema de optimización de la forma:

$$\min_{c\in\mathbb{R}} \sum_{i=1}^{n}\rho_\tau(X_i - c)$$

El problema anterior, se puede reformular como un problema de programación lineal si introducimos $2n$ variables artificiales $\{u_i, v_i \text{ con } i = 1, \cdots, n\}$ que representan las partes positivas y negativas de $X_i - c$, respectivamente. Tenemos entonces el nuevo problema:

$$\min_{(c,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \left\{ \tau 1_n' u + (1-\tau) 1_n' v \ : \ 1_n c + u - v = X \right\}$$

donde 1_n denota el vector n-dimensional de unos, $X = (X_1, \cdots, X_n)$, $u = (u_1, \cdots, u_n)$ y $v = (v_1, \cdots, v_n)$. Por lo tanto, estamos minimizando una función lineal en un conjunto poliédrico de restricciones formado por la intersección de hiperplanos $(2n+1)$-dimensionales determinados por las restricciones lineales impuestas dentro de $\mathbb{R} \times \mathbb{R}_+^{2n}$.

Lo extendemos ahora al problema de regresión. Supongamos entonces que nos interesa explicar una variable aleatoria Y escalar en función de d covariables que denotaremos por X de las cuales conocemos una muestra $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ siendo $X_i = (X_{1,i}, \cdots, X_{d,i})$. Entonces si el cuantil muestral de orden τ se define como la solución del problema:

$$\min_{c\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - c)$$

y la función cuantil condicional viene dada por $Q_\tau(x) = \theta(\tau)'x$, consideraremos $\widehat{\theta}(\tau)$ el elemento que resuelva el siguiente problema:

$$\min_{\theta\in\mathbb{R}^{d+1}} \ \sum_{i=1}^{n} \rho_\tau(Y_i - \theta' \mathbf{X_i})$$

donde $\mathbf{X_i} = (1, X_i)$; siendo éste es el punto de partida de la idea desarrollada por Koenker y Bassett (1978). Esto nos permitiría establecer un modelo de la forma:

$$Y_i = \theta' \mathbf{X_i} + \varepsilon_i$$

donde los residuos verificarían que $\mathbb{P}(\varepsilon_i \leq 0 \mid X) = \tau$, es decir, el cuantil de orden τ del error es cero.

Hemos planteado así un problema de optimización sin restricciones. Pero debemos tener en cuenta que la función que pretendemos optimizar no es una función diferenciable. Con lo cual no podemos utilizar los métodos empleados para el caso de la media condicional, en cuya situación si trabajamos con una función diferenciable.

De todas formas, recordemos que hemos formulado el problema de la búsqueda de un cuantil muestral como un problema de programación lineal de la forma:

$$\min_{(c,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \left\{ \tau 1_n' u + (1-\tau) 1_n' v \ : \ 1_n c + u - v = Y \right\}$$

donde 1_n denota el vector n-dimensional de unos y $u = (u_1, \cdots, u_n)$ y $v = (v_1, \cdots, v_n)$ representan las partes positivas y negativas de $Y_i - c$, respectivamente.

Extendiendo este razonamiento, el problema de regresión cuantil se puede formular como un problema de programación lineal de la siguiente manera:

$$\min_{(\theta,u,v)\in\mathbb{R}^{d+1}\times\mathbb{R}_+^{2n}} \left\{ \tau 1_n' u + (1-\tau) 1_n' v \ : \mathbb{X}\theta + u - v = Y \right\}$$

donde $\mathbb{X}$ denota la matriz de diseño que es una matriz $n \times (d+1)$. Además, hemos descompuesto el vector residual $Y - \mathbb{X}\theta$ en sus partes positivas y negativas (u y v respectivamente). Las soluciones $\widehat{\theta}(\tau)$ de este problema son las estimaciones de los coeficientes de regresión cuantil.

Hasta este momento, hemos conseguido expresar la búsqueda de un cuantil como la solución de un problema de programación lineal y extendimos este razonamiento al contexto de la regresión cuantil. Este hecho nos permite proponer métodos para el cálculo de los estimadores de regresión cuantil.

Barrodale y Roberts (1973) proponen una simplificación de la forma estándar del método del Simplex para el resolver el problema del cálculo de los estimadores en el caso de la regresión en mediana, donde la función de pérdida sería el valor absoluto. Posteriormente, Koenker y D'Orey (1987) extendieron este razonamiento a cualquier cuantil $0 < \tau < 1$.

Realmente, la simplificación del método del Simplex es consecuencia de que en un único paso del algoritmo propuesto por Barrodale y Roberts (1973) realizamos varios pasos del método del Simplex. Como consecuencia, este nuevo algoritmo es computacionalmente mucho más eficiente que el método del Simplex clásico.

1.1.4. Propiedades del estimador paramétrico

Debemos tener en cuenta que los estimadores asociados a la regresión cuantil no tienen expresión explícita por lo que sería necesario recurrir a expresiones asintóticas como la representación de Bahadur propuesta por Bahadur (1966). Además, ni siquiera la distribución de estos estimadores es conocida bajo hipótesis de normalidad como en el caso de la regresión en media estimada por mínimos cuadrados.

De todas formas si se verifican resultados sobre la distribución asintótica de los estimadores de regresión cuantil como el siguiente:

Teorema 1.1. *Consideremos un modelo lineal para explicar una variable respuesta escalar Y en función de una variable explicatica X de la forma:*

$$Y_i = \theta' \mathbf{X_i} + \varepsilon_i \qquad con\ i = 1, \cdots, n$$

donde los errores verifican que $\mathbb{P}(\varepsilon_i \leq 0 \mid X) = \tau$. Supongamos además, que se verifican las siguientes condiciones:

Condición A1. *Las funciones de distribución condicionales F_i (de Y_i condicionada a X_i) son absolutamente continuas y con densidades f_i continuas y uniformemente acotadas lejos de 0 e ∞ en los cuantiles condicionales $c_{\tau,i}$.*

Condición A2. *Existen matrices D_0 y $D_1(\tau)$ definidas positivas tales que:*

1. $\lim_{n\to\infty} \frac{1}{n} \sum \mathbf{X_i} \mathbf{X_i}' = D_0$

2. $\lim_{n\to\infty} \frac{1}{n} \sum f_i(c_{\tau,i})\mathbf{X_i}\mathbf{X_i}' = D_1(\tau)$

3. $\max_{i=1,\cdots,n} \|\mathbf{X_i}\|/\sqrt{n} \to 0$

se tiene que:

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}\right)$$

Demostración. No desarrollaremos la demostración de este resultado que puede verse en la página 121 de Koenker (2005). □

Centrémonos ahora en el caso homocedástico en el cual la función de densidad condicional de la variable Y viene dada por una función común f. Entonces si comparamos el resultado anterior con el resultado análogo en el caso de la regresión en media estimada por el método de mínimos cuadrados observamos que $\tau(1-\tau)$ está haciendo en este caso el papel de σ^2 en el caso de la regresión en media. Por otra parte también obtenemos una expresión del tipo $(\mathbb{X}'\mathbb{X})^{-1}$ como ocurre en el caso de la regresión en media donde por $\mathbb{X}$ denotamos la matriz de diseño.Cabe destacar que en el caso de la regresión cuantil adquiere gran importancia el comportamiento la función de densidad condicional f evaluada en el cuantil de orden τ de interés que venimos denotando por c_τ.

Surge entonces la duda de como estimar la matriz de covarianzas asintótica de los estimadores de la regresión cuantil que es de la forma $D_1^{-1}D_0D_1^{-1}$ donde las matrices D_0 y D_1 han sido definidas en la condición A2 del Teorema 1.1. A la vista de dichas definiciones, la precisión de la regresión cuantil depende de la inversa de la función de densidad evaluada en el cuantil que nos interesa, a dicha función Tukey (1965) la denominó **función 'sparsity'** que viene dada por:

$$s(\tau) = [f(F^{-1}(\tau))]^{-1}$$

Teniendo ésto en cuenta, concluimos que las estimaciones del cuantil serán más precisas cuantas más observaciones aparezcan en torno al cuantil que nos interesa. Por el contrario, si en un entorno del cuantil que estamos estudiando no existen muchas observaciones, los resultados que obtengamos no serán muy precisos.

En el caso de que los errores de la regresión sean independientes e idénticamente distribuidos, la función 'sparsity' juega un papel análogo al de la desviación típica en el caso de la regresión por mínimos cuadrados en este mismo escenario independiente e idénticamente distribuido.

Si derivamos la expresión $F(F^{-1}(t)) = t$ nos damos cuenta que la función 'sparsity' es la derivada de la función cuantil:

$$\frac{d}{dt}F^{-1}(t) = s(t)$$

Esto nos proporcionaría un modo de estimar la función 'sparsity' de la siguiente forma:

$$\hat{s}_n(t) = \frac{\hat{F}_n^{-1}(t+h_n) - \hat{F}_n^{-1}(t-h_n)}{2h_n}$$

donde $\hat{F}_n^{-1}$ estima la función F^{-1} y h_n es una sucesión de elementos que tienden a cero. Bofinger (1975) y Hall y Sheather (1988) propusieron diferentes sucesiones h_n para estimar la función 'sparsity'.

1.2. Contrastes de bondad de ajuste en el contexto de la regresión en media

Una vez que hemos formulado un modelo de regresión, éste nos permitirá extraer conclusiones sobre la variable respuesta en función de las diferentes variables explicativas consideradas. Para que las conclusiones de un modelo de regresión sean un reflejo de la realidad debemos estar seguros de que el modelo propuesto se ajusta bien al conjunto de datos con el que estamos trabajando. Surgen de este modo los contrastes de bondad de ajuste en el contexto de la regresión.

Consideremos un modelo de regresión en media de la forma:

$$Y = m(X) + \varepsilon$$

donde $m(x) = \mathbb{E}(Y|X = x)$ es la conocida función de regresión en media de la variable Y sobre X y los errores verifican que $\mathbb{E}(\varepsilon|X) = 0$. Supongamos conocida una muestra aleatoria simple $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ de las variables $(X, Y) \in \mathbb{R}^{d+1}$.

Nuestro objetivo será realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : m \in \mathcal{M}_\theta = \left\{m_\theta \ : \ \theta \in \Theta \subset \mathbb{R}^q\right\} & \text{Hipótesis nula} \\ H_a : m \notin \mathcal{M}_\theta & \text{Hipótesis alternativa} \end{cases}$$

Propondremos a continuación diferentes familias de test que nos permitan realizar el contraste de bondad de ajuste anterior. Tanto para la elaboración de esta sección como para la siguiente hemos tomado como referencia fundamental a González-Manteiga y Crujeiras (2013).

1.2.1. Contrastes basados en métodos de suavizado

Dentro de los contrastes basados en métodos de suavizado nos centraremos en aquellos que se basen en estimadores tipo kernel como el estimador de Nadaraya-Watson (Nadaraya (1964) y Watson (1964)), que viene dado por:

$$m_{nh}(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i = \sum_{i=1}^{n} \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)} Y_i$$

siendo $K_h(x) = h^{-1}K(x/h)$ donde K representa una función de densidad tipo kernel y h es el conocido parámetro de suavizado (ventana). A partir del estimador anterior consideraremos el siguiente proceso empírico:

$$\alpha_n(x) = \sqrt{nh^p} \left(m_{nh}(x) - \mathbb{E}_{\widehat{\theta}}(m_{nh}(x)) \right) = \sqrt{nh^p} \sum_{i=1}^{n} W_{ni}(x) r_i$$

donde $r_i = Y_i - m_{\widehat{\theta}}(X_i)$ y $\mathbb{E}_{\widehat{\theta}}$ representa una estimación de $\mathbb{E}_{\theta_0}$ siendo θ_0 el parámetro real bajo H_0 y $\widehat{\theta}$ un estimador $\sqrt{n}$-consistente de θ_0.

Entonces un test de bondad de ajuste basado en el proceso empírico α_n se puede basar en un estadístico que sea resultado de la aplicación de un funcional continuo a este proceso empírico. Es decir, tendríamos un estadístico de la forma:

$$T_n = \int \alpha_n^2(x) w(x) dx$$

donde w es una función de peso. Además, el estadístico anterior se puede expresar como $T_n = nh^p T_{1n}$ siendo T_{1n} el estadístico propuesto por Härdle y Mammen (1993) que viene dado por:

$$T_{1n} = \int \left(m_{nh}(x) - m_{nh}(x, \widehat{\theta}) \right)^2 w(x) dx$$

donde:

$$m_{nh}(x, \widehat{\theta}) = \sum_{i=1}^{n} W_{ni}(x) m_{\widehat{\theta}}(X_i)$$

Nótese que una versión discreta del estadístico T_{1n} anterior se puede encontrar en González-Manteiga y Cao (1993).

Es claro que el test propuesto por Härdle y Mammen (1993) está basado en una caracterización de la hipótesis nula (H_0). Es decir, se verificará H_0 si y solo si $\mathbb{E}(C_1) = 0$ siendo:

$$C_1 = \mathbb{E}^2(\varepsilon_0 | X) w(X)$$

donde $\varepsilon_0 = Y - m_{\theta_0}(X)$. Alternativamente, es posible definir otras pruebas estadísticas basadas en estimadores consistentes de diferentes características que permitan probar la hipótesis nula. Por ejemplo, Zheng (1996) se basa en la cantidad:

$$C_2 = \varepsilon_0 \, \mathbb{E}(\varepsilon_0 | X) f(X) w(X)$$

siendo f la densidad de la variable explicativa X. Entonces el valor esperado de C_2 será cero si y solo si se verifica la hipótesis nula. Surge así el llamado test de Zheng cuyo estadístico de contraste viene dado por:

$$T_{2n} = \frac{1}{n(n-1)} \sum_{i \neq j} K_h(X_i - X_j)(Y_i - m_{\widehat{\theta}}(X_i))(Y_j - m_{\widehat{\theta}}(X_j)) w(X_i)$$

que es un estimador que no presenta sesgo asintótico.

Dette (1999) caracteriza la hipótesis nula por el cumplimiento de $\mathbb{E}(C_3) = 0$ siendo en este caso:

$$C_3 = \mathbb{E}(\varepsilon_0^2 - (\varepsilon_0 - \mathbb{E}(\varepsilon_0 | X))^2) w(X)$$

En función de ésto, propone un test de bondad de ajuste basado en el estadístico:

$$T_{3n} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{\widehat{\theta}}(X_i))^2 w(X_i) - \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{nh}(X_i))^2 w(X_i)$$

Cualquiera de los estadísticos T_{1n}, T_{2n} o T_{3n} anteriores converge asintóticamente a una distribución gaussiana que nos permitirá calibrar el correspondiente test de bondad de ajuste. En Zhang y Dette (2004) se desarrolla una completa comparativa de los test anteriores.

En la práctica, el uso de las distribuciones asintóticas de cualquiera de los estimadores propuestos implica seleccionar el parámetro h de suavizado. Problema que hoy en día sigue siendo un aspecto aún en estudio dentro del campo de los contrastes de bondad de ajuste. Aunque podemos citar referencias en esta línea como Kulasekera y Wang (1997), Zhang (2004) o Gao y Gijbels (2008).

Además de la selección de la ventana de suavizado, esta familia de test de bondad de ajuste debe enfrentarse a la lenta tasa de convergencia a la distribución gaussiana así como a la necesidad de estimar de manera no paramétrica los modelos involucrados en el estudio.

1.2.2. Contrastes basados en la función de regresión integrada

Con el objetivo de evitar la problemática asociada a la selección del parámetro de suavizado surge una nueva metodología basada en los contrastes de bondad de ajuste para la función de distribución de una variable aleatoria. Esta nueva familia de contrastes estadísticos se basa en la **función de regresión integrada** que viene dada por:

$$I(x) = \mathbb{E}[Y\ \mathbb{I}(X \leq x)] = \int_{-\infty}^{x} m(z)dF(z)$$

donde la función $\mathbb{I}$ representa la función indicadora. Teniendo en cuenta la expresión anterior, la función de regresión integrada se puede estimar empíricamente de la siguiente manera:

$$I_n(x) = \frac{1}{n}\ \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)\ Y_i$$

Asociado a la función de regresión integrada podemos definir el proceso empírico:

$$R_n(x) = \sqrt{n}\ \left(I_n(x) - \mathbb{E}_{\widehat{\theta}}(I_n(x)) \right) = \sqrt{n}\ \sum_{i=1}^{n} r_i\ \mathbb{I}(X_i \leq x)$$

donde recordemos que $r_i = Y_i - m_{\widehat{\theta}}(X_i)$. Stute (1997) demostró que el proceso empírico anterior marcado por los residuos converge a un proceso gaussiano de media cero y con una complicada función de covarianzas, lo cual hacía complicado el calibrado del test. Posteriormemente, Stute, González-Manteiga y Presedo-Quindimil (1998) propusieron un procedimiento wild bootstrap para el calibrado del test de Stute (1997).

1.2.3. Contrastes construidos para evitar el desastre de la dimensionalidad

Hemos considerado hasta el momento contrastes de bondad de ajuste construidos a partir de la comparación de un estimador no paramétrico del modelo de regresión y un estimador bajo la hipótesis nula o en la comparación análoga con estimadores de la función de regresión integrada. En ambos casos, se aprecia el desastre de la dimensionalidad a medida que aumenta la dimensión de la variable explicativa.

Para la primera familia de contrastes estadísticos el efecto del aumento de la dimensión conlleva una pérdida de potencia asintótica. Por otra parte, dentro de la segunda familia, el desastre de la dimensionalidad también se observa en muestras pequeñas, como puede extraerse de diferentes estudios de simulación.

Con el objetivo de intentar resolver este problema surgen diferentes modificaciones de los contrastes propuestos con anterioridad. Por ejemplo, dentro de los test basados en métodos de suavizado podemos citar el trabajo de Lavergne y Patilea (2008) que proponen la siguiente modificación del test de Zheng (1996) utilizando proyecciones:

$$T_n = \sup_{\beta,||\beta||=1} \sum_{i<j} K_h(\beta'(X_i - X_j))(Y_i - m_{\hat{\theta}}(X_i))(Y_j - m_{\hat{\theta}}(X_j))$$

Otra opción en esta misma línea sería proyectar la covariable X en una cierta dirección β_0 que minimice:

$$\mathbb{E}^2(\varepsilon - \mathbb{E}(\varepsilon \mid \beta'X)) = \mathbb{E}^2(\varepsilon - m_\beta(X))$$

siendo éste el punto de partida de Xia (2009).

Dentro del contexto de los test basado en la función de regresión integrada, Stute y otros (2008) proponen reemplazar el proceso empírico R_n por:

$$R_n^g(t) = n^{-1/2} \sum_{i=1}^{n} (g(X_i) - \bar{g})\mathbb{I}(r_i \leq t)$$

con $t \in \mathbb{R}$ siendo $\bar{g} = n^{-1}\sum_{i=1}^n g(X_i)$ y r_i denotan los residuos del modelo.

En esta misma línea debemos mencionar a Escanciano (2006) que establece la siguiente caracterización casi segura de la hipótesis nula:

$$H_0 \text{ cierta} \Longleftrightarrow \mathbb{E}[\varepsilon(\theta_0)\, \mathbb{I}(\beta'X \leq u)] = 0$$

siendo $\varepsilon(\theta_0) = Y - m_{\theta_0}(X)$ y β el vector de proyección. Ésto da lugar a un nuevo proceso empírico marcado por los residuos de la forma:

$$R_n^1(\beta, u) = n^{-1/2} \sum_{i=1}^{n} r_i\, \mathbb{I}(\beta'X \leq u)$$

donde $r_i = Y_i - m_{\widehat{\theta}}(X_i)$ representan los residuos del modelo.

Nota 1.2. Debemos mencionar que el campo de los contrastes de bondad de ajuste es muy amplio y ha gozado de un gran desarrollo en los últimos años. Como consecuencia de ello, nos podemos encontrar con una gran variedad de contrastes de bondad de ajuste que no hemos tratado en esta sección. Citamos a modo de ejemplo, test basados en la distribución empírica de los residuos (Van Keilegom y otros (2008)) o test basados en la función de verosimilitud (Fan y otros (2001) o Fan y Jiang(2007)).

1.3. Contrastes de bondad de ajuste en el contexto de la regresión cuantil

En la sección anterior hemos citado diferentes contrastes de bondad de ajuste en el contexto de la regresión en media ajustada por el método de mínimos cuadrados. De igual modo, surge la necesidad de contrastar el ajuste de cualquier modelo de regresión cuantil a un conjunto de datos, pudiendo entonces extraer conclusiones válidas a partir del modelo de regresión cuantil propuesto. En esta sección, hablaremos, por lo tanto, de contrastes de bondad de ajuste en el contexto de la regresión cuantil.

Consideraremos un modelo de regresión asociado a un cuantil $\tau \in (0,1)$ de interés de la forma:

$$Y = Q_\tau(X) + \varepsilon$$

siendo ε el error desconocido que debe verificar que $\mathbb{P}(\varepsilon \leq 0|X) = \tau$.

En este caso, nuestro objetivo será realizar el siguiente contraste de hipótesis:

$$H_0 \ : \ Y = Q_\tau(X) + \varepsilon = g(X, \theta_0) + \varepsilon$$

donde $\theta_0 \in \mathbb{R}^q$, que es equivalente a:

$$H_0 \ : \ \mathbb{E}\big[\mathbb{I}(Y \leq g(X, \theta_0)) \mid X\big] = \tau$$

Para intentar llevar a cabo el anterior contraste de bondad de ajuste, al igual que hemos descrito en la sección anterior, nos encontramos con diferentes enfoques a la hora de plantear un contraste aunque en este caso, el tema está notablemente menos desarrollado. Para ello supongamos conocida una muestra aleatoria simple $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ de las variables $(X, Y) \in \mathbb{R}^{d+1}$.

1.3.1. Contrastes basados en métodos de suavizado

Dentro del contexto de los test basados en métodos de suavizado destacamos la metodología de Zheng (1998) que extiende el conocido test de Zheng (1996), para la regresión en media, al caso de la regresión cuantil. En el caso de la regresión cuantil, el estadístico de contraste vendría dado por:

$$T_n = \frac{nh^{d/2}}{\widehat{\sigma}} \ \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{h^p} K\left(\frac{X_i - X_j}{h}\right) \left[\mathbb{I}\big(Y_i \leq g(X_i, \widehat{\theta})\big) - \tau\right] \left[\mathbb{I}\big(Y_j \leq g(X_j, \widehat{\theta})\big) - \tau\right]$$

donde K es una función de densidad tipo kernel, h es el parámetro de suavizado y:

$$\widehat{\sigma}^2 = 2\tau^2(1-\tau)^2 \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{h^d} K^2\left(\frac{X_i - X_j}{h}\right)$$

El estadístico anterior T_n converge en distribución a una normal estándar. Cabe destacar el conocido problema de la elección del parámetro de suavizado h, al igual que ocurría en el caso de los contrastes de bondad de ajuste para la regresión en media.

En la línea marcada por Zheng (1998), Dette y otros (2012) proponen un contraste de bondad de ajuste basado en métodos de suavizado para el contexto de modelo aditivos cuantiles. En el contexto de los modelos aditivos asociados a la regresión cuantil podríamos plantearnos el siguiente contraste:

$$H_0 : Q_\tau(x) = Q_\tau(x_1, \cdots, x_p) = \sum_{i=1}^{d} Q_{\tau,i}(x_i) + c(\tau)$$

siendo $X \in \mathbb{R}^d$ la variable explicativa en función de las cual intentamos explicar la variable respuesta que venimos denotando por Y.

Suponiendo conocida una muestra $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ de las variables $(X, Y) \in \mathbb{R}^{d+1}$, Dette y otros (2012) proponen un test de bondad de ajuste basándose en el estadístico:

$$T_n = \frac{1}{n(n-1)h^d} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) \widehat{R}_i \widehat{R}_j$$

donde K representa una función de densidad tipo kernel, h es el parámetro de suavizado y los elementos $\widehat{R}_i$ vienen dados por:

$$\widehat{R}_i = \mathbb{I}(Y_i \leq \widehat{Q}^{-i}_{\tau,add}(X_i)) - \tau$$

donde $\widehat{Q}^{-i}_{\tau,add}(X_i)$ denota la estimación de la función de regresión cuantil sin incluir la $i-$ésima observación. Estandarizando el estadístico T_n descrito anteriormente, obtendríamos una convergencia asintótica gaussiana, de todas formas, es aconsejable utilizar un calibrado bootstrap para este test.

1.3.2. Contrastes basados en la función de regresión integrada

En la línea de estudiar la función de regresión integrada debemos citar el trabajo de He y Zhu (2003) en el contexto de la regresión cuantil. He y Zhu (2003) proponen un test de bondad de ajuste asumiendo como hipótesis nula que el modelo de regresión cuantil tiene una forma paramétrica frente a alternativas no paramétricas. Es decir, suponen que la relación entre la variable explicativa $X \in \mathbb{R}^d$ y la respuesta Y se puede modelar como:

$$Y_i = g(X_i, \theta) + \varepsilon_i \qquad \text{con } i = 1, \cdots, n$$

Entonces, proponen estudiar el siguiente proceso empírico:

$$R_n^\tau(t) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i) \dot{g}(X_i, \widehat{\theta}) \mathbb{I}(X_i \leq t)$$

donde $\psi(r) = \tau \mathbb{I}(r > 0) + (\tau - 1)\mathbb{I}(r > 0)$ denota la derivada de la función de pérdida cuantílica $\rho_\tau = u\big(\tau - \mathbb{I}(u < 0)\big)$, $\dot{g}$ representa la derivada de la función $g(x, \theta)$ con respecto al parámetro θ, $\widehat{\theta}$ es el parámetro estimado que recordemos es la solución del siguiente problema de optimización:

$$\min_{\theta} \sum_{i=1}^{n} \rho_\tau(Y_i - g(X_i, \theta))$$

y $r_i = Y_i - g(X_i, \widehat{\theta})$ denotan los residuos del modelo.

Bajo las condiciones anteriores, el proceso empírico $R_n^\tau(t)$ converge a un proceso gaussiano de media cero y con función de covarianzas:

$$W(t_1, t_2) = \tau(1 - \tau)\mathbb{E}\Big[\dot{g}(X, \theta)\dot{g}'(X, \theta)\mathbb{I}(X \leq \min(t_1, t_2)) - S(t_1)S^{-1}S(t_2)\Big]$$

siendo:

$$S = \mathbb{E}[\dot{g}(X, \widehat{\theta})\dot{g}'(X, \widehat{\theta})]$$
$$S(t) = \mathbb{E}[\dot{g}(X, \widehat{\theta})\dot{g}'(X, \widehat{\theta})\mathbb{I}(X \leq t)]$$

1.3.3. El desastre de la dimensionalidad en los contrastes de bondad de ajuste

Dentro del contexto de la regresión cuantil, no nos consta un contraste de bondad de ajuste para evitar el temido desastre de la dimensionalidad. De todas formas, debemos citar a Wilcox (2008) que propone una modificación del estadístico de He y Zhu (2003) en el caso de que la hipótesis nula refleje un modelo cuantil lineal. Así, bajo las hipótesis de la sección anterior, el proceso empírico asociado al contraste de He y Zhu (2003) en el caso particular de la regresión cuantil lineal vendría dado por:

$$R_n(X_i) = n^{-1/2} \sum_{k=1}^{n} \psi(r_k) \mathbf{X_k} \mathbb{I}(X_k \leq X_i)$$

donde $\mathbf{X_i} = (1, X_i)$ representa la derivada de la función cuantil en el punto X_i siendo X_i la i-ésima observación de la variable explicativa $X \in \mathbb{R}^d$.

La propuesta de Wilcox (2008) para simplificar el problema de la dimensión consiste en lo siguiente. Fijado j, sean U_{ij} los rangos de los n valores en la j-ésima columna de la matriz de diseño con $j = 2, \cdots, d + 1$ (se omite la columna $j = 1$ de la matriz de diseño porque en el caso de la regresión lineal consiste en una columna de unos). Denotemos por $F_i = \max U_{ij}$ tomado sobre $j = 2, \cdots, d + 1$. De este modo, Wilcox (2008) pretende

reducir la dimensión del problema original considerando los valores $F_i \in \mathbb{R}$ en lugar de las correspondientes $X_i \in \mathbb{R}^p$. En esta línea, define el proceso empírico:

$$R_n^w(t) = n^{-1/2} \sum_{k=1}^{n} \psi(r_k)\mathbf{X_k}\mathbb{I}(F_k \leq t)$$

a partir del cual se establece el estadístico de contraste como:

$$C_n = \text{mayor autovalor de } \int R_n^w(t)[R_n^w(t)]' dF_{n,w}(t)$$

siendo $F_{n,w}(t)$ la función de distribución empírica de las variables F_i.

El propio Wilcox propone este nuevo estadístico de contraste como una simplificación del test propuesto por He y Zhu (2003) y es claro que ésta no es una solución que garantice la consistencia universal del problema. Entonces, nuestro objetivo será proponer un nuevo test de bondad de ajuste en el contexto de la regresión cuantil basado en proyecciones para evitar el desastre de la dimensionalidad en la línea del propuesto por Escanciano (2006).

El presente documento se estructura de la siguiente manera: en el Capítulo 2 se presenta la nueva propuesta de contraste de bondad de ajuste, en el Capítulo 3 se comentan los resultados obtenidos gracias a un estudio de simulación y en en Capítulo 4 se aplica la nueva propuesta de contraste de bondad de ajuste a un conjunto de datos reales. Finalmente, se presenta parte del código utilizado para la implementación informática del estadístico de contraste.

Capítulo 2

Propuesta del nuevo test de bondad de ajuste

A lo largo de este segundo capítulo vamos a presentar una nueva propuesta de contraste de bondad de ajuste en el contexto de la regresión cuantil con el objetivo de evitar el desastre de la dimensionalidad.

Empezaremos este capítulo presentando el nuevo test de bondad de ajuste (Sección 2.1) para luego tratar propiedades asintóticas (Sección 2.2) del mismo y una aproximación bootstrap (Sección 2.3) para el calibrado del test. Para terminar, se presentan ciertos aspectos computacionales (Sección 2.4) sobre la implementación del contraste.

2.1. Estadístico de contraste

Supongamos que queremos explicar una variable aleatoria escalar Y en función de un vector aleatorio X de dimensión d mediante un modelo de regresión cuantil. Nuestro objetivo será contrastar si la relación entre X e Y se puede modelar como:

$$Y = g(X, \theta) + \varepsilon \tag{2.1}$$

donde ε representa el error desconocido del modelo que recordemos debe verificar que $\mathbb{P}(\varepsilon \leq 0 \mid X) = \tau$, es decir, su cuantil de orden τ es 0. Además, $\theta \in \Theta \subset \mathbb{R}^q$ es un parámetro desconocido y $g \in \mathcal{M}_\theta = \{g(\cdot, \theta) : \theta \in \Theta \subset \mathbb{R}^q\}$ es una función conocida excepto por el parámetro θ.

Para llevar a cabo este contraste supongamos conocida una muestra aleatoria simple $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ siendo $X_i = (X_{1,i}, \cdots, X_{d,i})'$, Debemos recordar en este punto que el parámetro desconocido $\theta \in \Theta \subset \mathbb{R}^q$ se puede estimar resolviendo el siguiente problema de optimización:

$$\min_{\theta} \sum_{i=1}^{n} \rho_\tau(Y_i - g(X_i, \theta))$$

siendo $\rho_\tau(u) = u\big(\tau - \mathbb{I}(u < 0)\big)$ la conocida función de pérdida cuantílica introducida por Koenker y Bassett (1978). A partir de este momento vamos a denotar por $\widehat{\theta}$ a esta estimación.

A lo largo de esta sección vamos a proponer un nuevo test de bondad de ajuste para contrastar si la relación entre la variable respuesta Y y el vector de variables explicativas X es de la forma descrita en (2.1). Es decir, nuestro objetivo será realizar el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : g \in \mathcal{M}_\theta = \{g(\cdot, \theta) \ : \ \theta \in \Theta \subset \mathbb{R}^q\} \\ H_a : g \notin \mathcal{M}_\theta \end{cases}$$

Es claro entonces que $g \in \mathcal{M}_\theta$ es equivalente a que:

$$\mathbb{P}(\varepsilon(\theta_0) \leq 0 \mid X) = \tau$$

para cierto $\theta_0 \in \Theta \subset \mathbb{R}^q$ donde por $\varepsilon(\theta_0) = Y - g(X, \theta_0)$ estamos representando los errores del correspondiente modelo de la forma (2.1) con parámetro θ_0.

Como ya hemos mencionado en la Sección 2 del Capítulo 1, He y Zhu (2003) estudiaron este mismo problema para el cual definieron el siguiente proceso empírico marcado por los residuos:

$$S_n(t) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i) \dot{g}(X_i, \widehat{\theta}) \mathbb{I}(X_i \leq t) \tag{2.2}$$

donde $\psi(r) = \tau \mathbb{I}(r > 0) + (\tau - 1)\mathbb{I}(r < 0)$ es la derivada de la función ρ_τ, $\dot{g}$ es la derivada de la función $g(\cdot, \theta)$ con respecto al parámetro θ, $\widehat{\theta}$ es la estimación del parámetro θ y $r_i = Y_i - g(X_i, \widehat{\theta})$ representan los residuos del modelo. Debemos notar además que $x \leq t$ si y solo si cada componente de x es menor o igual que la correspondiente componente de t.

En función del proceso empírico (2.2) se define el estadístico de contraste que viene dado por:

$$V_n = \underset{||a||=1}{\text{máx}} \ n^{-1} \sum_{i=1}^{n} \big(a' S_n(X_i)\big)^2$$

que coincide con el mayor autovalor de $n^{-1} \sum_{i=1}^{n} S_n(X_i) S_n(X_i)'$.

A la vista del proceso empírico anterior, debemos mencionar que $S_n(t)$ representaría un vector del espacio $\mathbb{R}^q$ si el parámetro θ se mueve en $\mathbb{R}^q$, es decir, $S_n(t)$ es un proceso empírico multidimensional que está ponderado por la función $\psi(\cdot)$ aplicada a los residuos y por los valores observados de la variable explicativa a través de la derivada de la función g. Estas dos características del proceso empírico $S_n(t)$ marcan una clara diferencia frente al proceso estudiado por Stute (1997) que toma valores en $\mathbb{R}$ puesto que los residuos del modelos son unidimensionales. La modificación propuesta por He y Zhu (2003) surge con la idea de evitar la baja potencia de los test basados en el proceso empírico unidimensional frente a desviaciones en torno al 0 (problema estudiado por Eubank y Hart (1993)).

Recordemos que el test de bondad de ajuste propuesto por He y Zhu (2003) comparte con el propuesto por Stute (1997) la no necesidad de construir un modelo no paramétrico mediante suavizadores así como la conveniencia de un procedimiento bootstrap para el calibrado del test.

En el Capítulo 1 ya hemos hablado de cómo afecta a los contrastes de bondad de ajuste el aumento de la dimensión de la variable explicativa $X \in \mathbb{R}^d$ que es lo que se conoce como **desastre de la dimensionalidad**. El nuevo test que vamos a presentar surge con el objetivo de evitar el desastre de la dimensionalidad en la línea del test propuesto por Escanciano (2006). Es decir, vamos a proponer un contraste de bondad de ajuste para el contexto de la regresión cuantil basado en el uso de proyecciones sobre la variable explicativa X.

En primer lugar debemos presentar el siguiente lema que nos proporciona una caracterización de la hipótesis nula que justifica la utilización de proyecciones en la nueva propuesta de test de bondad de ajuste. Entonces si denotamos por $\|a\|$ a la norma euclídea del elemento $a \in \mathbb{R}^q$ se tiene el siguiente resultado:

Lema 2.1. *Una condición necesaria y suficiente para que $H_0 : g \in \mathcal{M}_\theta$ se verifique es que para cualquier vector $\beta \in \mathbb{R}^d$ con $\|\beta\| = 1$,*

$$\mathbb{P}[\varepsilon(\theta_0) \leq 0 \mid \beta' X] = \tau$$

casi seguro para algún $\theta_0 \in \Theta \subset \mathbb{R}^p$ donde $\varepsilon(\theta_0) = Y - g(X, \theta_0)$.

Demostración. Es claro que la afirmación anterior será cierta si conseguimos probar la siguiente cadena de equivalencias:

$$\begin{aligned} H_0 : \quad g(\cdot, \theta_0) \in M &\iff \mathbb{P}[\varepsilon(\theta_0) \leq 0 \mid X] = \tau \\ &\overset{(i)}{\iff} \mathbb{E}[\psi(\varepsilon(\theta_0)) \mid X] = 0 \\ &\overset{(ii)}{\iff} \mathbb{E}[\psi(\varepsilon(\theta_0))\dot{g}(X, \theta_0) \mid X] = 0 \\ &\overset{(iii)}{\iff} \mathbb{E}[\psi(\varepsilon(\theta_0))\dot{g}(X, \theta_0) \mid \beta' X] = 0 \end{aligned}$$

En primer lugar, en la equivalencia (i) debemos tener en cuenta que:

$$\begin{aligned} \mathbb{E}\big[\psi(r)\big] &= \mathbb{E}\big[\tau\, \mathbb{I}(r > 0) + (\tau - 1)\, \mathbb{I}(r < 0)\big] = \tau\, \mathbb{P}(r > 0) + (\tau - 1)\, \mathbb{P}(r < 0) \\ &= \tau\, \mathbb{P}(r > 0) + (\tau - 1)\, (1 - \mathbb{P}(r > 0)) = \mathbb{P}(r > 0) + \tau - 1 \end{aligned}$$

Por otra parte, la equivalencia (ii) es clara teniendo en cuenta las propiedades de la esperanza condicional así como el hecho de que $\dot{g}$ es una función de X. Finalmente, en (iii) la implicación es inmediata. De todas formas, tenemos que probar la suficiencia, si denotamos por $Z = \psi(\varepsilon(\theta_0))\dot{g}(X, \theta_0)$ se tiene que para cada $\beta \neq 0$, la σ-álgebra generada por $X'\beta$ coincide con la σ-álgebra generada por $X'\beta/||\beta||$. Como consecuencia de propiedades elementales de la esperanza condicional, se tiene que para cualquier β, incluyendo $\beta = 0$,

$$0 = \mathbb{E}[e^{iX'\beta}\, \mathbb{E}[Z \mid X'\beta]] = \mathbb{E}[e^{iX'\beta} Z] = \mathbb{E}[e^{iX'\beta}\, \mathbb{E}[Z \mid X]]$$

por lo tanto, como consecuencia del teorema 3.1 (Página 75) de Parthasarathy (1967), concluimos que $\mathbb{E}\left[Z|X\right] = 0$ casi seguro como queríamos demostrar. □

El lema anterior nos permite proponer un nuevo test de bondad de ajuste basado en proyecciones unidimensionales puesto que hemos llegado a la siguiente caracterización de la hipótesis nula:

$$H_0 \text{ cierta} \Longleftrightarrow \mathbb{P}\Big(\varepsilon(\theta_0)\, \mathbb{I}(\beta' X \leq u) \leq 0\Big) = \tau$$

casi seguro para todo $(\beta, u) \in \Pi$ siendo $\Pi = \mathbb{S}_d \times [-\infty, +\infty]$ donde $\mathbb{S}_d$ representa la esfera unidad en $\mathbb{R}^d$, es decir:

$$\mathbb{S}_d = \left\{\beta \in \mathbb{R}^d \ : \ \|\beta\| = 1\right\}$$

Teniendo todo esto en cuenta consideraremos el siguiente proceso empírico:

$$R_n(\beta, u) = n^{-1/2} \sum_{i=1}^{n} \psi(\varepsilon(\theta_0)_i) \dot{g}(X_i, \theta_0) \mathbb{I}(\beta' X_i \leq u) \tag{2.3}$$

donde recordemos que θ_0 representa el parámetro verdadero del modelo, que en la práctica no conocemos y será por tanto necesario estimarlo. Entonces el proceso empírico que utilizaremos en la práctica vendría dado por:

$$R_n^1(\beta, u) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i) \dot{g}(X_i, \widehat{\theta}) \mathbb{I}(\beta' X_i \leq u) \tag{2.4}$$

donde $r_i = Y_i - g(X_i, \widehat{\theta})$ representan los residuos del modelo y $\widehat{\theta}$ es una estimación de θ.

Llegados a este punto debemos elegir una norma en función de la cual definir el estadístico de contraste. Consideraremos la norma de Cramér-von Mises, al igual que propone Escanciano (2006) como consecuencia de sus claras ventajas computacionales. Con lo cual, siguiendo en este caso la línea marcada por He y Zhu (2003), el estadístico en función del cual se define el nuevo contraste de bondad de ajuste vendría dado por:

$$T_n = \text{mayor autovalor de } \int_{\Pi} R_n^1(\beta, u)[R_n^1(\beta, u)]' F_{n,\beta}(du) d\beta \tag{2.5}$$

donde $F_{n,\beta}(u)$ representa la función de distribución empírica de las variables explicativas proyectadas $\{\beta' X_1, \cdots, \beta' X_n\}$ y $d\beta$ representa la densidad uniforme en la esfera unidad en el espacio $\mathbb{R}^d$.

2.2. Propiedades asintóticas

En esta sección se estudia el comportamiento asintótico del proceso de contraste R_n^1 y del estadístico T_n, bajo la hipótesis nula y bajo la alternativa. Bajo la hipótesis

nula sirve de orientación sobre la calibración del test, y bajo la alternativa justifica la consistencia del mismo.

Vamos a introducir una matriz que juega un papel importante en la distribución del proceso de contraste,

$$S = E\left[\dot{g}(X,\theta_0)\dot{g}'(X,\theta_0)\right]$$

que suponemos que es una matriz finita y no singular. Se trata por tanto de una matriz simétrica y definida positiva.

Empezamos con un teorema que prueba la convergencia del proceso R_n en el cual el parámetro θ_0 se supone conocido. Este teorema es un análogo del Teorema 1 de Escanciano (2006). Como diferencia respecto de aquel resultado, en este caso no es necesario imponer suposiciones sobre los momentos de la variable respuesta o los errores de regresión, pues éstos han sido sustituidos por la función ψ, que está acotada por 1. Sin embargo, aparece la condición de que S sea finita y no singular, que es consecuencia de incluir el vector gradiente $\dot{g}(X,\theta_0)$ en la expresión del proceso de contraste, en la línea de lo propuesto por He y Zhu (2003).

Teorema 2.2. *Bajo la hipótesis nula $H_0 : g \in \mathcal{M}_\theta$, se tiene que:*

$$R_n \xrightarrow{d} R_\infty$$

siendo R_∞ un proceso gaussiano de media cero y matriz de covarianzas dada por:

$$K(x_1,x_2) = E\left[\dot{g}(X,\theta_0)\dot{g}'(X,\theta_0)\mathbb{I}(\beta_1'X \le u_1)\mathbb{I}(\beta_2'X \le u_2)\right]$$

donde $x_1 = (\beta_1', u_1)'$ y $x_2 = (\beta_2', u_2)'$.

Demostración. La convergencia de las distribuciones finito-dimensionales de R_n se obtiene aplicando el teorema central del límite clásico. Para la equicontinuidad asintótica, se pueden emplear los mismos argumentos dados en la demostración del Teorema 1 de Escanciano (2006). □

El resultado anterior establece la distribución límite del proceso cuando no se requiere estimar el parámetro θ_0. En la práctica será necesario estimar θ_0, lo cual conduce al proceso R_n^1. La distribución asintótica del proceso R_n^1 se verá afectada por el problema de estimación. De hecho, necesitamos disponer de una representación asintótica del estimador.

He y Shao (1996) establecieron la siguiente representación del estimador bajo diseño fijo y con ciertas suposiciones de regularidad para g y f, siendo f la función de densidad del error,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \left(\sqrt{n}f(0)S\right)^{-1}\sum_{i=1}^{n}\psi(\epsilon_i)\dot{g}(X,\theta_0) + o_P(1) \tag{2.6}$$

He y Zhu (2003) extendieron este resultado en su Lema A.1 al contexto de diseño aleatorio e incluyendo la estimación bajo la hipótesis alternativa.

El interés de una representación como la anterior es que simplifica el estudio de las propiedades del estimador, pues lo convierte en una suma de variables independientes e idénticamente distribuidas (salvo un término despreciable). De este modo, se le pueden aplicar resultados asintóticos clásicos, como leyes de los grandes números, para demostrar la consistencia del estimador, o teorema central del límite, para obtener su distribución asintótica.

En la representación del estimador dada en (2.6) destaca el papel que juega la matriz S, que sería el análogo de la matriz del diseño de un modelo lineal de regresión, así como el valor $f(0)$, que indica la densidad de la respuesta en torno al cuantil condicional. Su inversa sería la función 'sparsity', que ya fue comentada en el Capítulo 1 al tratar sobre la estimación de modelos paramétricos de regresión cuantil. Lo ideal es que haya poca 'sparsity', o lo que es lo mismo, mucha densidad de observaciones en las proximidades de la curva de regresión cuantil.

El teorema siguiente proporciona una representación del proceso de contraste R_n^1.

Teorema 2.3. *Bajo la hipótesis nula $H_0 : g \in \mathcal{M}_\theta$, suponiendo que*

$$Y_i = g(X_i, \theta_0) + \varepsilon_i$$

donde $\varepsilon_1, \ldots, \varepsilon_n$ son errores independientes e idénticamente distribuidos con cuantil τ igual a cero, y función de densidad f, y bajo ciertas condiciones de regularidad para g y f, se tiene que:

$$\begin{aligned} R_n^1(\beta, u) \; &= n^{-1/2} \sum_{i=1}^{n} \psi(\varepsilon_i) \dot{g}(X_i, \theta_0) \mathbb{I}(\beta' X \leq u) \\ &\quad - S(\beta, u) S^{-1} n^{-1/2} \sum_{i=1}^{n} \psi(\varepsilon_i) \dot{g}(X_i, \theta_0) + o_P(1) \end{aligned} \tag{2.7}$$

uniformemente en (β, u), siendo:

$$S(\beta, u) = E\left[\dot{g}(X, \theta_0) \dot{g}'(X, \theta_0) \mathbb{I}(\beta' X \leq u)\right]$$

Demostración. La demostración de este teorema se obtiene agregando los argumentos de la demostración del Teorema 1 de He y Zhu (2003) y del Teorema 2 de Escanciano (2006). □

En la representación dada en el teorema anterior, destacamos que el primer sumando coincide con el proceso R_n, en el cual no se estimaba el parámetro θ_0. El segundo sumando se debe precisamente a la estimación del mismo, y da lugar a un proceso que resulta de multiplicar una función no aleatoria $S(\beta, u)S^{-1}$ por una suma de vectores aleatorios independientes e idénticamente distribuidos de dimensión q. A esta suma se le puede aplicar un teorema central del límite clásico, que multiplicado por la función no aleatoria $S(\beta, u)S^{-1}$ da lugar de manera inmediata a la convergencia del segundo sumando de la representación (2.7)

A partir de la representación anterior es inmediato obtener la convergencia del proceso a un proceso gaussiano de media cero y con la matriz de covarianzas resultante de las distribuciones bidimensionales.

Además, por el teorema de la aplicación continua, resulta la convergencia del estadístico de contraste, como se indica en el corolario siguiente. Además del teorema de la aplicación continua, es precido añadir que la integración respecto de $F_{n\beta}$, que es una distribución empírica, no afecta al límite en distribución, como se indica en Escanciano (2006).

Corolario 2.4. *Bajo la hipótesis nula $H_0 : g \in \mathcal{M}_\theta$, y asumiendo las mismas suposiciones del teorema anterior, el estadístico T_n definido en (2.5) converge en distribución al mayor autovalor de $\int_\Pi R^1(\beta,u)[R^1(\beta,u)]' F_{\beta' X}(du)d\beta$, siendo R^1 el proceso límite de R_n^1 y $F_{\beta' X}$ la función de distribución de $\beta' X$.*

A continuación estudiamos las propiedades del proceso de contraste R_n^1 y el estadístico T_n, frente a alternativas locales. El propósito es mostrar la consistencia del test, esto es, que si los datos proceden de un modelo que no cumple la hipótesis nula, entonces la potencia del test tiende a infinito al aumentar el tamaño muestral. Es más, se conseguirá esta convergencia incluso cuando la sucesión de alternativas se aproximen a la hipótesis nula a una tasa de convergencia próxima a $\sqrt{n}$. En concreto, supondremos que:

$$Y_i = g(X_i,\theta_0) + n^{-1/2}h(X_i) + \varepsilon_i \qquad i \in \{1,\dots,n\} \tag{2.8}$$

donde la función h representa la desviación respecto de la hipótesis nula.

Teorema 2.5. *Si los datos proceden del modelo (2.8), donde $\varepsilon_1,\dots,\varepsilon_n$ son errores independientes e idénticamente distribuidos con cuantil τ igual a cero, y función de densidad f, y bajo ciertas condiciones de regularidad para g, f y h, se tiene que:*

$$\begin{aligned} R_n^1(\beta,u) &= n^{-1/2}\sum_{i=1}^n \psi(\varepsilon_i)\dot{g}(X_i,\theta_0)\mathbb{I}(\beta' X \le u) \\ &\quad -S(\beta,u)S^{-1}n^{-1/2}\sum_{i=1}^n \psi(\varepsilon_i)\dot{g}(X_i,\theta_0) + q(\beta,u) + o_P(1) \end{aligned}$$

uniformemente en (β,u), siendo:

$$\begin{aligned} q(\beta,u) &= f(0)\left\{S(\beta,u)S^{-1}E\left[h(X)\dot{g}(X,\theta_0)\right]\right. \\ &\qquad \left. - E\left[h(X)\dot{g}(X,\theta_0)\mathbb{I}(\beta' X \le u)\right]\right\} \end{aligned}$$

Demostración. La demostración de este teorema sigue los mismos argumentos de la demostración del Teorema 1 de He y Zhu (2003). □

La función q es la que va a marcar la diferencia en la distribución del proceso de contraste, y en consecuencia del estadístico de contraste, entre la hipótesis nula y la alternativa. Así, mientras bajo la hipótesis nula el estadístico de contraste estandarizado converge a una distribución constante, bajo la alternativa hay que sumarle a esa distribución la función q. Si h va precedida por una tasa de convergencia a cero más lenta

que $n^{-1/2}$, entonces el desplazamiento de la distribución del estadístico se irá a infinito. Esto queda expresado formalmente en el corolario siguiente, que es una adaptación del Corolario 1 de He y Zhu (2003).

Corolario 2.6. *Bajo las condiciones del teorema anterior, suponiendo que los datos proceden del modelo:*

$$Y_i = g(X_i, \theta_0) + c_n n^{-1/2} h(X_i) + \varepsilon_i$$

siendo c_n una sucesión que converge a infinito (a cualquier tasa), el estadístico de contraste T_n converge a infinito.

Para obtener este resultado se requiere que la sucesión $g(x\theta_0) + c_n n^{-1/2} h(x)$ no coincida con ningún elemento del modelo paramétrico, esto es, no se pueda expresar como $g(x, b_n)$ para ninguna sucesión de parámetros b_n, y que $Var(h(X)\dot{g}(X, b) > 0$ para cualquier valor del parámetro b.

Este corolario garantiza que la potencia del test converge a 1, bajo alternativas locales que se aproximan al modelo paramétrico a una tasa arbitrariamente próxima a $n^{-1/2}$.

2.3. Aproximación bootstrap

Una vez que hemos propuesto un nuevo contraste de bondad de ajuste, será necesario llevar a cabo el calibrado del estadístico (2.5). Para ello aplicaremos un procedimiento wild bootstrap sobre los residuos. Debemos recordar en este punto que estamos trabajando en el contexto de la regresión cuantil por lo que el procedimiento bootstrap debe ser el adecuado para este contexto. He y Zhu (2003) ya proponen un procedimiento bootstrap para llevar a cabo el calibrado de su propuesta de contraste de bondad de ajuste. Posteriormente, Feng, He y Hu (2011) demuestran la consistencia del wild bootstrap en el contexto de la regresión cuantil paramétrica, siendo ésta nuestra principal referencia.

Consideremos un modelo de regresión cuantil de la forma:

$$Y_i = g(X_i, \theta_0) + \varepsilon_i \quad \text{con } i = 1, \cdots, n \tag{2.9}$$

donde Y_i denota la i-ésima observación de una variable aleatoria Y, X_i denota el vector de observaciones de las variables explicativas $X \in \mathbb{R}^d$ y ε_i representan los errores independientes. Para que el modelo anterior resulte ser identificable supondremos que para cualquier cuantil de interés $\tau \in (0, 1)$ se tiene que el cuantil condicional τ de ε_i dado X_i es cero. Entonces podemos calcular un estimador del parámetro θ_0 como el elemento que minimiza:

$$\sum_{i=1}^{n} \rho_\tau\big(Y_i - g(X_i, \theta)\big)$$

donde recordemos que $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ denota la función de pérdida cuantílica.

Feng, He, y Hu (2011) proponen una simple modificación del wild bootstrap (para la regresión en media) para adaptarse a la pérdida asimétrica en la función de regresión

cuantil, e identificar una clase de distribuciones de peso en las cuales el método propuesto es asintóticamente válido. El procedimiento bootstrap para llevar a cabo el calibrado del test es el siguiente:

Paso 1: Ajustar el modelo (2.9), y denotar por $\widehat{\theta}$ a la estimación del parámetro θ_0 y por r_i a los residuos con $i = 1, \cdots, n$. Además, calculamos:

$$T_n = \text{mayor autovalor de } \int_{\Pi} R_n^1(\beta, u)[R_n^1(\beta, u)]' F_{n,\beta}(du) d\beta$$

donde $F_{n,\beta}(u)$ representa la función de distribución empírica de las variables explicativas proyectadas $\{\beta' X_1, \cdots, \beta' X_n\}$, $d\beta$ representa la densidad uniforme en la esfera unidad en el espacio $\mathbb{R}^d$ y además:

$$R_n^1(\beta, u) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i) \dot{g}(X_i, \widehat{\theta}) \mathbb{I}(\beta' X_i \leq u)$$

donde $r_i = Y_i - g(X_i, \widehat{\theta})$ representan los residuos del modelo y $\widehat{\theta}$ es una estimación de θ.

Nótese que la función de distribución empírica $F_{n,\beta}(u)$ no varía para el cálculo de las remuestras bootstrap del estadístico de contraste puesto que no estamos remuestreando la variable explicativa $X \in \mathbb{R}^d$.

Paso 2: Generar los pesos w_i procedentes de una distribución que satisfaga ciertas propiedades que proponen Feng, He, y Hu (2011). A partir de estos pesos calculamos $\epsilon_i^\star = w_i |r_i|$ donde $|\cdot|$ denota el valor absoluto.

En nuestro caso, utilizaremos como función de pesos una variable Bernoulli que toma los valores $2(1-\tau)$ y -2τ con probabilidad $1-\tau$ y τ.

Paso 3: Calcular las remuestras bootstrap de la forma $Y_i^\star = g(X_i, \widehat{\theta}) + \epsilon_i^\star$.

Paso 4: Ajustar el modelo (2.9) a las remuestras bootstrap y denotar por $\widehat{\theta^\star}$ a las estimaciones del parámetro $\widehat{\theta}$ obtenidas con la muestra bootstrap del paso 3. Además, calcularemos el valor del estadístico de contraste para las muestras bootstrap:

$$T_{n,b}^\star = \text{mayor autovalor de } \int_{\Pi} R_n^{1\star}(\beta, u)[R_n^{1\star}(\beta, u)]' F_{n,\beta}(du) d\beta$$

siendo en este caso:

$$R_n^{1\star}(\beta, u) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i^\star) \dot{g}(X_i, \widehat{\theta^\star}) \mathbb{I}(\beta' X_i \leq u)$$

donde $r_i^\star = Y_i^\star - g(X_i, \widehat{\theta^\star})$.

Paso 5: Repetir los pasos 2, 3 y 4 B veces.

Finalmente calculamos el p-valor estimado por Monte Carlo como la proporción de valores del estadístico bootstrap que superan el estadístico original, esto es:

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(T_n < T_{n,b}^{\star})$$

siendo B el número de remuestras bootstrap.

2.4. Aspectos computacionales del estadístico de contraste

En este sección se proporcionan ciertas expresiones que facilitan la computación del estadístico de contraste definido en (2.5). De hecho, hemos elegido la norma de Cramér-von Mises debido a la computación sencilla del estadístico de contraste en este caso. En efecto, se tiene que:

$$\begin{aligned}
\alpha_n &= \int_{\Pi} R_n^1(\beta,u)[R_n^1(\beta,u)]' F_{n,\beta}(du)d\beta \\
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta})'\int_{\Pi}\mathbb{I}(\beta'X_i\le u)\mathbb{I}(\beta'X_j\le u)F_{n,\beta}(du)d\beta \\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta})\int_{\mathbb{S}_d}\mathbb{I}(\beta'X_i\le\beta'X_r)\mathbb{I}(\beta'X_j\le\beta'X_r)d\beta \\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta})A_{ijr}
\end{aligned}$$

donde $F_{n,\beta}(u)$ representa la función de distribución empírica de las variables explicativas proyectadas $\{\beta'X_1,\cdots,\beta'X_n\}$ y $d\beta$ representa la densidad uniforme en la esfera unidad $\mathbb{S}_d$.

Debemos tener en cuenta que dado $d>1$, Escanciano (2006) llegó a que la integral A_{ijr} es proporcional al volumen de una sección de la esfera $\mathbb{S}_d$ que se puede calcular de la siguiente manera:

$$A_{ijr} = A_{ijr}^{(0)}\,\frac{\pi^{d/2-1}}{\Gamma\left(\frac{d}{2}+1\right)}$$

donde $A_{ijr}^{(0)}$ es el ángulo complementario entre los vectores (X_i-X_r) y (X_j-X_r) medido en radianes y $\Gamma(\cdot)$ representa la función gamma. Por lo tanto, $A_{ijr}^{(0)}$ viene dado por:

$$A_{ijr}^{(0)} = \left|\pi - arcos\left(\frac{(X_i-X_r)'(X_j-X_r)}{\|X_i-X_r\|\,\|X_j-X_r\|}\right)\right|$$

A la vista de las expresiones anteriores queda claro el simple cálculo computacional asociado a la elección de la norma de Cramér-von Mises. Además para el cálculo de los

elementos A_{ijr} debemos tener en cuenta que:

$$A_{ijr}^{(0)} = \begin{cases} \pi & X_i = X_j \text{ y } X_i \neq X_r \\ 2\pi & \text{si } X_i = X_j = X_r \\ \pi & X_i \neq X_j \text{ y } X_i = X_r \text{ ó } X_j = X_r \end{cases}$$

y que $A_{ijr} = A_{jir}$, lo cual nos permite una considerable reducción del tiempo de computación necesario para calcular el estadístico de contraste.

Debemos notar que:

$$\alpha_n = n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n} \psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta})A_{ijr}$$

representa un vector tridimensional (dependiendo de los parámetros $i, j, r \in \{1, \cdots, n\}$) lo cual representa a priori un problema de memoria a la hora de almacenar dichos valores. Entonces, debemos tener en cuenta lo siguiente:

$$\begin{aligned} \alpha_n &= n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n} \psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta})A_{ijr} \\ &= n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} \psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta}) \sum_{r=1}^{n} A_{ijr} \\ &= n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} \psi(r_i)\psi(r_j)\dot{g}(X_i,\widehat{\theta})\dot{g}'(X_j,\widehat{\theta})A_{ij\bullet} \end{aligned}$$

donde $A_{ij\bullet}$ representa una matriz de dimensión $n \times n$ donde el elemento que ocupa la fila i-ésima y la columna j-ésima viene dado por A_{ijr}.

Es decir, $A_{ij\bullet} \in \mathcal{M}_{n\times n}$ siendo n el número de observaciones conocidas de las variables $(X, Y) \in \mathbb{R}^{d+1}$. Por ejemplo, si estuviéramos trabajando con una muestra de tamaño $n = 100$ entonces $A_{ij\bullet} \in \mathcal{M}_{100\times 100}$ con el alto coste de almacenamiento que ello conlleva. Ya hemos dicho anteriormente que $A_{ij\bullet}$ es una matriz simétrica con lo cual el número de elementos que debemos calcular en memoria pasa de n^2 a $n(n+1)/2$.

Debemos mencionar que el estadístico de contraste (2.5) ha sido programado utilizando el software libre R (http://cran.r-project.org/). De todos modos, el cálculo de las matrices $A_{ij\bullet}$ y α_n ha sido programado en Fortran consiguiendo de esta forma que la implementación del estadístico de contraste sea mucho más rápida.

Debemos mencionar que la matriz $A_{ij\bullet}$ no será necesario replicarla para cada una de las remuestras bootstrap puesto que los elementos de dicha matriz solo dependen de las variables explicativas que no se modifican a lo largo del procedimiento bootstrap. Este hecho implica un importante ahorro computacional para calibrar la nueva propuesta de test de bondad de ajuste dado que el cálculo de las remuestras bootstrap realizadas en el paso cuatro del calibrado bootstrap quederían como sigue:

$$T_{n,b}^{\star} = \text{mayor autovalor de } n^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} \psi(r_i^{\star})\psi(r_j^{\star})\dot{g}(X_i,\widehat{\theta^{\star}})\dot{g}'(X_j,\widehat{\theta^{\star}})A_{ij\bullet}$$

Capítulo 3

Estudio de simulación

En este tercer capítulo se presenta un estudio de simulación con el objetivo de evaluar la nueva propuesta de contraste bondad de ajuste y compararla con el test propuesto por He y Zhu (2003).

El capítulo se organiza de la siguiente manera: en la primera sección se presenta el estudio de simulación para luego mostrar los resultados asociados a modelos bajo la hipótesis nula (Sección 3.2) y bajo la hipótesis alternativa (Sección 3.3).

3.1. Presentación del estudio de simulación

A lo largo de todo este capítulo vamos a estudiar si la relación entre las variables $X \in \mathbb{R}^d$ e $Y \in \mathbb{R}$ se puede expresar gracias a un modelo de regresión cuantil de la siguiente manera:

$$Y = g(X, \theta) + \varepsilon$$

donde $\mathbb{P}(\varepsilon \leq 0 \mid X) = \tau$ siendo ε el error desconocido del modelo y τ el cuantil de interés que en nuestro caso será 0.5 (la mediana). Nuestro objetivo será realizar el siguiente contaste de hipótesis:

$$\begin{cases} H_0 : g \in \mathcal{M}_\theta = \{g(X, \theta) \; con \; \theta \in \Theta \subset \mathbb{R}^q\} \\ H_a : g \notin \mathcal{M}_\theta \end{cases} \tag{3.1}$$

Durante el Capítulo 2, hemos propuesto un test de bondad de ajuste para realizar el contraste (3.1) suponiendo que la función de regresión cuantil bajo la hipótesis nula se puede expresar a través de un modelo paramétrico. En este tercer capítulo vamos a estudiar un caso particular, supondremos que la relación entre las variables X e Y se puede formular gracias a un modelo de regresión cuantil lineal. Es decir, contrastamos que el cuantil τ condicional de la variable respuesta Y se puede expresar como combinación lineal de las variables explicativas.

Recordemos que el proceso empírico asociado a la nueva propuesta de contraste de bondad de ajuste viene dado por:

$$R_n^1(\beta, u) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i)\ \dot{g}(X_i, \widehat{\theta})\ \mathbb{I}(\beta' X_i \leq u) \tag{3.2}$$

siendo en este caso $\dot{g}(X_i, \widehat{\theta}) = (1, X_i) = (1, X_{i,1}, \cdots, X_{i,d})$ que denotaremos por $\mathbf{X_i}$. Además, $r_i = Y_i - \widehat{\theta}' \mathbf{X_i}$ representan los residuos del modelo siendo en este caso $\widehat{\theta} \in \mathbb{R}^{p+1}$.

Por otra parte, He y Zhu (2003) resuelven el mismo problema en función del siguiente proceso empírico:

$$S_n(t) = n^{-1/2} \sum_{i=1}^{n} \psi(r_i)\ \mathbf{X_i}\ \mathbb{I}(X_i \leq t) \tag{3.3}$$

Presentamos a continuación los resultados asociados a un estudio de simulación que nos permitirá comparar el nuevo contrastes de bondad de ajuste basado en proyecciones frente al test propuesto por He y Zhu (2003), de los que hemos hablado a lo largo del Capítulo 2.

Para comparar ambos contrastes de bondad de ajuste debemos establecer las dos definiciones siguientes:

Definición 3.1. Dado un contraste de hipótesis, la decisión de rechazar la hipótesis nula H_0 siendo verdadera se llama error de tipo I. La probabilidad de este error es el **nivel de significión** del test y se denota por:

$$\alpha = \mathbb{P}(\text{error de tipo I}) = \mathbb{P}(\text{rechazar } H_0 \mid H_0 \text{ es cierta})$$

Definición 3.2. La **potencia** de un contraste se define como:

$$\beta = \mathbb{P}(\text{aceptar } H_1 \mid H_1 \text{ cierta})$$

es decir, la probabilidad complementaria al error de tipo II.

Nótese que mostraremos la proporción de rechazos asociados a ambos contrastes de bondad de ajuste obtenidos como consecuencia de simular 1000 muestras por Monte Carlo de cada uno de los modelos que mostramos a continuación. Además, como ya hemos mencionado en la Sección 3 del Capítulo 2 será necesario aplicar un procedimiento bootstrap considerando en este caso $B = 500$ réplicas.

3.2. Modelos bajo la hipótesis nula

En esta sección aplicaremos la nueva propuesta de contraste de bondad de ajuste así como el test de He y Zhu (2003) a dos modelos bajo la hipótesis nula. Vamos a empezar por generar valores del siguiente modelo:

$$Y = 1 + X_1 + X_2 + \varepsilon \tag{3.4}$$

donde $X_i \sim U(0, 1)$ con $i = 1, 2$ y $\varepsilon \sim N(0, 1)$ es el error desconocido del modelo. Es de-

cir, estaríamos generando comprobando el **ajuste del nivel de significación** de ambos contrastes de bondad de ajuste que ya generamos valores de un modelo bajo la hipótesis nula.

Antes de mostrar los resultados obtenidos, debemos mencionar que hemos utilizado la función *rq* disponible dentro de la librería *quantreg* de R que nos permite ajustar los correspondientes modelos lineales de regresión cuantil asociados al cuantil $\tau = 0.5$ (la mediana). Todos los detalles de esta librería se pueden ver en la siguiente dirección http://cran.r-project.org/web/packages/quantreg/quantreg.pdf.

En la siguiente tabla mostramos la proporción de rechazos asociada a diferentes tamaños muestrales, que denotaremos por n, y para diferentes niveles de significación, que denotaremos por α. Los resultados obtenidos son los siguientes:

	Test basado en proyecciones			Test de He y Zhu (2003)		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
n=25	0.096	0.049	0.002	0.099	0.057	0.011
n=50	0.112	0.047	0.008	0.108	0.061	0.012
n=100	0.102	0.058	0.016	0.108	0.045	0.007
n=150	0.089	0.048	0.007	0.091	0.049	0.010
n=200	0.100	0.048	0.010	0.106	0.050	0.011

Tabla 3.1: Proporción de rechazos asociada al nuevo test de bondad de ajuste basado en proyecciones y al test de He y Zhu (2003) para el modelo (3.4).

A la vista de los resultados anteriores, ambos contrastes de bondad de ajuste ajustan bien el nivel de significación, incluso para tamaños muestrales no muy grandes (como por ejemplo, $n = 25$). En este sentido no se aprecian diferencias significativas entre el test propuesto por He y Zhu (2003) y el nuevo test basado en proyecciones que nosotros proponemos. Es decir, estamos trabajando con dos contraste de bondad de ajuste que ajustan bien el nivel de significación con lo cual tiene sentido comparar la potencia de cada uno de ellos.

Como nuestro objetivo era proponer un nuevo test de bondad de ajuste para evitar el desastre de la dimensionalidad, vamos a simular entonces valores del siguiente modelo con más variables explicativas:

$$Y = 1 + X_1 + X_2 + X_3 + X_4 + X_5 + \varepsilon \tag{3.5}$$

donde $X_i \sim U(0,1)$ con $i = 1, \cdots, 5$ y por $\varepsilon \sim N(0,1)$ hemos denotado el error desconocido del modelo.

La proporción de rechazos asociada a diferentes tamaños muestrales, n, y para diferentes niveles de significación, α, se muestra en la siguiente tabla:

	Test basado en proyecciones			Test de He y Zhu (2003)		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 25$	0.119	0.066	0.017	0.106	0.053	0.015
$n = 50$	0.112	0.053	0.014	0.125	0.076	0.019
$n = 100$	0.094	0.047	0.011	0.108	0.058	0.016
$n = 150$	0.104	0.056	0.014	0.092	0.042	0.007
$n = 200$	0.106	0.049	0.010	0.108	0.053	0.015

Tabla 3.2: Proporción de rechazos asociada al nuevo test de bondad de ajuste basado en proyecciones y al test de He y Zhu (2003) para el modelo (3.5).

Vemos entonces que los resultados obtenidos para ambos contrastes de bondad de ajuste son muy similares a los obtenidos para el modelo (3.4) en el que se consideraban dos variables explicativas, ya que ambos ajustan bien el nivel de significación del contraste de hipótesis que hemos propuesto. Por lo tanto, el aumento de la dimensión de las variables explicativas no conlleva diferencias significativas en el ajuste del nivel de ninguno de los dos contraste de bondad de ajuste propuestos.

Una vez que hemos estudiado el ajuste del nivel de significación de ambos contrastes de bondad de ajuste, pasaremos a estudiar sus potencias simulando para ello modelos bajo la hipótesis alternativa. En este caso, deberíamos observar que bajo la hipótesis alternativa la proporción de rechazos tiende a 1 a medida que aumentamos el tamaño muestral lo cual pondría de manifiesto la **consistencia** de ambos contrastes de bondad de ajuste.

3.3. Modelos bajo la hipótesis alternativa

Para comparar la **potencia** de ambos contrastes de bondad de ajuste vamos a simular valores de diferentes modelos de regresión cuantil bajo la hipótesis alternativa. Empezaremos por generar valores del modelo:

$$Y = 1 + X_1 + X_2 + \frac{1}{3}\left(X_1^2 + X_1X_2 + X_2^2\right) + \varepsilon \tag{3.6}$$

donde $X_1 \sim U(0,1)$, $X_2 \sim N(0,1)$ y $e^{\varepsilon+1} \sim N(0,1)$ siendo ε el error desconocido del modelo (modelo muy similar a uno de los considerados por He y Zhu (2003)). Es decir, estaríamos generando valores de un modelo bajo la hipótesis alternativa ya que hemos añadido una desviación cuadrática al modelo lineal. Por lo tanto, nos interesará que la proporción de rechazos sea elevada.

En la siguiente tabla recogemos la proporción de rechazos asociada a diferentes

niveles de significación (α) y diferentes tamaños muestrales (n). Los resultados obtenidos son los siguientes:

	Test basado en proyecciones			Test de He y Zhu (2003)		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 25$	0.252	0.154	0.057	0.225	0.135	0.035
$n = 50$	0.675	0.564	0.361	0.487	0.357	0.163
$n = 100$	0.961	0.918	0.776	0.822	0.725	0.460
$n = 150$	0.993	0.983	0.943	0.949	0.903	0.751
$n = 200$	0.999	0.998	0.990	0.982	0.965	0.897

Tabla 3.3: Proporción de rechazos asociada al nuevo test de bondad de ajuste basado en proyecciones y al test de He y Zhu (2003) para el modelo (3.6).

A la vista de los resultados anteriores vemos que la proporción de rechazos del nuevo test basado en proyecciones es mayor que la correspondiente proporción de rechazos para el test propuesto por He y Zhu (2003), sean cuales sean el nivel de significación y el tamaño muestral considerados. Así, por ejemplo, para un tamaño muestral $n = 100$ y un nivel de significación del 5 %, la proporción de rechazos asociada al nuevo test basado en proyecciones es casi 0.20 más que la correspondiente al test de He y Zhu (2003). Es decir, a la vista de este modelo de regresión en mediana, la potencia de nuestra propuesta de test es mayor que la potencia asociada al test de He y Zhu (2003).

Además, observamos que la proporción de rechazos aumenta a medida que aumenta el tamaño muestral tendiendo hacia 1. Teniendo en cuenta la sección anterior, se pone de manifiesto de esta forma la consistencia de ambos contrastes de bondad de ajuste ya que en ambos casos la proporción de rechazos tiende hacia el nivel de significación α bajo la hipótesis nula y hacia 1 bajo la hipótesis alternativa.

Dado que el objetivo de nuestra propuesta de contraste de bondad de ajuste es evitar el desastre de la dimensionalidad, deberíamos observar que las diferencias entre las proporciones de rechazos para ambos contrastes deberían ser mayores a medida que aumentamos el número de variables explicativas involucradas en el modelo.

Por lo tanto, vamos a aumentar la dimensión de la variable explicativa $X = (X_1, \cdots, X_d)$ en el modelo (3.6), considerando en este caso 4 variables explicativas y manteniendo la misma desviación sobre la hipótesis nula. Es decir, vamos a generar entonces valores del siguiente modelo de regresión cuantil:

$$Y = 1 + X_1 + X_2 + X_3 + X_4 + \frac{1}{3}\left(X_1^2 + X_1X_2 + X_2^2\right) + \varepsilon \tag{3.7}$$

donde $X_i \sim U[0,1]$ con $i = 1, 4$, $X_j \sim N(0,1)$ con $j = 2, 3$ y $e^{\varepsilon+1} \sim N(0,1)$ siendo ε el error desconocido del modelo.

En este nuevo escenario la proporción de rechazos asociada a diferentes niveles de significación (α) y diferentes tamaños muestrales (n) es la siguiente:

	Test basado en proyecciones			Test de He y Zhu (2003)		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 25$	0.108	0.052	0.012	0.103	0.051	0.007
$n = 50$	0.263	0.149	0.037	0.165	0.094	0.031
$n = 100$	0.746	0.641	0.448	0.281	0.170	0.060
$n = 150$	0.965	0.915	0.780	0.424	0.305	0.118
$n = 200$	0.994	0.990	0.946	0.565	0.410	0.216

Tabla 3.4: Proporción de rechazos asociada al nuevo test de bondad de ajuste basado en proyecciones y al test de He y Zhu (2003) para el modelo (3.7).

A la vista de los resultados anteriores, podemos decir que al aumentar la dimensión de la variable explicativa X la proporción de rechazos disminuye en ambos contrastes de bondad de ajuste aunque la pérdida de potencia es mucho más significativa en el caso del test propuesto por He y Zhu (2003). Por el contrario, la nueva propuesta de contraste de bondad de ajuste resiste mucho mejor el temido desastre de la dimensionalidad como consecuencia de la utilización de proyecciones sobre las variables explicativas, cumpliéndose de este modo el propósito de la propuesta de este nuevo contraste de bondad de ajuste.

Así, por ejemplo, considerando un tamaño muestral $n = 150$ y un nivel de significación del 10 %, la diferencia entre la proporción de rechazos asociada a la nueva propuesta de contraste de bondad de ajuste y el test de He y Zhu (2003) es de 0.044 para el modelo (3.6) con dos variables explicativas mientras que pasa a ser 0.541 para el modelo (3.7) con cuatro variables explicativas.

De todas formas, parte de la pérdida de potencia asociada al modelo (3.7) podría ser consecuencia de que hemos incluido dos nuevas variables explicativas manteniendo la misma desviación sobre la hipótesis nula, por lo tanto, el porcentaje de varianza asociada a dicha desviación es menor que la correspondiente para el modelo (3.6).

En la línea de conocer como afecta a ambos contrastes de bondad de ajuste el desastre de la dimensionalidad, vamos a simular valores del siguiente modelo de regresión cuantil:

$$Y = 1 + X_1 + X_2 + X_3 + X_4 + X_5 + \left(X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 \right) + \varepsilon \qquad (3.8)$$

donde $X_i \sim N(0,1)$ con $i = 1, \cdots, 5$ y $e^{\varepsilon+1} \sim N(0,1)$ siendo ε el error desconocido del modelo.

La siguiente tabla recoge la proporción de rechazos asociada a diferentes niveles

de significación (α) y diferentes tamaños muestrales (n). Los resultados obtenidos se muestran a continuación:

	Test basado en proyecciones			Test de He y Zhu (2003)		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 25$	0.131	0.069	0.016	0.097	0.052	0.008
$n = 50$	0.245	0.130	0.030	0.145	0.080	0.022
$n = 100$	0.896	0.770	0.484	0.217	0.118	0.040
$n = 150$	0.996	0.985	0.906	0.350	0.212	0.070
$n = 200$	1.000	1.000	0.997	0.466	0.299	0.108

Tabla 3.5: Proporción de rechazos asociada al nuevo test de bondad de ajuste basado en proyecciones y al test de He y Zhu (2003) para el modelo (3.8).

Los resultados anteriores siguen la línea de los obtenidos para el modelo (3.7) aunque en este caso como la dimensión de las variables explicativas es mayor, las diferencias entre las proporciones de rechazos también aumentan. A la vista de la tabla anterior, considerando un tamaño muestral $n = 100$ y un nivel de significación del 5 %, la proporción de rechazos asociada al test basado en proyecciones es 0.652 mayor que la correspondiente proporción para el test propuesto por He y Zhu (2003).

Nótese que la proporción de rechazos asociada a un tamaño muestral $n = 25$ es baja para ambos contrastes de bondad de ajuste como consecuencia de la consideración de cinco variables explicativas en el modelo de regresión cuantil siendo $n = 25$ un tamaño muestral pequeño.

Para terminar con el análisis de la potencia de los dos contrastes de bondad de ajuste simularemos valores del siguiente modelo, que también ha sido estudiado por Escanciano (2006):

$$Y = X'\theta_0 + g\ cos(0.6\ \pi\ X'\theta_0) + \varepsilon \tag{3.9}$$

donde W, W_1 y $W_2 \sim U(0, 2\pi)$ y $\varepsilon \sim N(0, 1)$ representa el error desconocido del modelo, y siendo:

$$X'\theta_0 = 1 + X_1 + X_2$$

$$X_i = \frac{W + W_i}{2} \quad \text{con } i = 1, 2$$

Nótese que en este modelo hemos incluido una cierta dependencia entre las variables explicativas X_i gracias a la consideración de las variables auxiliares W, W_1 y W_2.

Mostramos a continuación la proporción de rechazos obtenida para diferentes valores del nivel de significación (α), diferentes tamaños muestrales (n) y diferentes valores

del parámetro g que representa la desviación del modelo sobre la hipótesis nula. Estos resultados se recogen en la siguiente tabla:

		Test basado en proyecciones			Test de He y Zhu (2003)		
		$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$n = 25$	$g = 3$	0.125	0.058	0.012	0.089	0.046	0.010
	$g = 6$	0.125	0.066	0.015	0.090	0.048	0.010
	$g = 10$	0.117	0.062	0.011	0.084	0.037	0.012
	$g = 12$	0.117	0.056	0.013	0.081	0.041	0.010
$n = 50$	$g = 3$	0.189	0.107	0.034	0.109	0.053	0.015
	$g = 6$	0.209	0.113	0.028	0.096	0.051	0.011
	$g = 10$	0.217	0.122	0.032	0.112	0.056	0.015
	$g = 12$	0.211	0.118	0.028	0.096	0.058	0.012
$n = 100$	$g = 3$	0.587	0.381	0.128	0.152	0.084	0.027
	$g = 6$	0.666	0.430	0.154	0.171	0.089	0.026
	$g = 10$	0.674	0.469	0.152	0.196	0.096	0.027
	$g = 12$	0.686	0.479	0.161	0.165	0.083	0.020
$n = 150$	$g = 3$	0.953	0.840	0.419	0.298	0.165	0.042
	$g = 6$	0.987	0.904	0.490	0.297	0.171	0.041
	$g = 10$	0.984	0.915	0.539	0.312	0.182	0.051
	$g = 12$	0.992	0.929	0.539	0.336	0.197	0.054
$n = 200$	$g = 3$	1.000	0.995	0.794	0.461	0.252	0.052
	$g = 6$	1.000	0.999	0.907	0.542	0.340	0.082
	$g = 10$	1.000	0.999	0.899	0.505	0.312	0.086
	$g = 12$	1.000	1.000	0.907	0.526	0.313	0.072

Tabla 3.6: Proporción de rechazos asociada al nuevo test de bondad de ajuste basado en proyecciones y al test de He y Zhu (2003) para el modelo (3.9).

A la vista de los resultados anteriores vemos que la proporción de rechazos es claramente mayor en el caso del nuevo test de bondad de ajuste basado en proyecciones que

para el test propuesto por He y Zhu (2003). Como ya hemos mencionado anteriormente, el parámetro g marca la desviación del modelo sobre la hipótesis nula, con lo cual es razonable que la proporción de rechazos aumente a medida que aumenta el valor de dicho parámetro g para ambos contrastes de bondad de ajuste, como podemos observar en la tabla anterior.

Al igual que hemos mencionado anteriormente, la proporción de rechazos aumenta a medida que aumenta el tamaño muestral y tiende hacia 1, aunque en este caso la convergencia sea más lenta que para otros modelos estudiados. Nótese además que en este caso obtenemos las diferencias más significativas entre la proporción de rechazos asociada al nuevo contraste de bondad de ajuste frente al test propuesto por He y Zhu (2003), quizá motivado por la forma de las variables explicativas.

A lo largo de todo este capítulo hemos presentado un estudio de simulación que pone de manifiesto que la nueva propuesta de contraste de bondad de ajuste ajusta bien el nivel de significación y la proporción de rechazos tiende hacia 1 bajo la hipótesis alternativa a medida que aumenta el tamaño muestral. Es decir, la propuesta de test basado en proyecciones en el contexto de la regresión cuantil es consistente. Además, este contraste muestra una potencia superior a la del test de He y Zhu (2003).

Capítulo 4

Aplicación a datos reales

4.1. Presentación del conjunto de datos reales

La Unidad de Producción Térmica (U.P.T.) de As Pontes, que está situada en el municipio de As Pontes de García Rodríguez (en el noreste de la provincia de A Coruña), constituye uno de los centros productivos propiedad de Endesa Generación S.A. en la Península Ibérica. La instalación inició su actividad en 1976 con el objetivo de hacer uso racional de los lignitos pardos extraídos de la Mina a cielo abierto situada en sus proximidades.

En el año 1993 se inició un proceso de transformación de la Central Térmica que culminaría en 1996, con objeto de utilizar mezclas de lignito local con carbones de importación caracterizados por sus bajos contenidos en azufre y cenizas. De este modo ha sido posible una reducción global en las emisiones de dióxido de azufre (SO_2) superior al 40 %.

En el emplazamiento de la Central Térmica de As Pontes, entre los años 2007 y 2008 se ha construido una nueva Central de Ciclo Combinado de Gas Natural, también propiedad de Endesa Generación S.A.. La Central de Ciclo Combinado consiste en un grupo generador de electricidad de tecnología de Ciclo Combinado, formado por dos turbinas de gas y una turbina de vapor. El Ciclo Combinado solo operará con gasóleo cuando se produzca un fallo en el suministro de gas natural y el Operador Nacional del Sistema considere que la Planta debe aportar energía a la red en esa situación. Los combustibles que van a ser utilizados en este caso, hacen que el principal interés recaiga en predecir los valores de los óxidos de nitrógeno (NO_x), para así, evitar superar los niveles límite fijados por la legislación.

Ambas Centrales tienen implantado un Sistema de Control Suplementario de la Contaminación Atmosférica que incluye la adquisición de datos de calidad de aire en tiempo real, su tratamiento y la realización de operaciones específicas que nos ayuden a la reducción de emisiones.

La Red de Vigilancia de la Calidad Atmosférica está formada por 7 estaciones automáticas, distribuidas en un radio de 30 km y comunicadas en tiempo real con la U.P.T. de As Pontes. Dichas estaciones automáticas proporcionan una medición continua de las concentraciones de dióxido de azufre, óxidos de nitrógeno, partículas en suspensión, parámetros meteorológicos y, en algún caso, ozono. La distribución geográfica de las mismas se presenta a continuación:

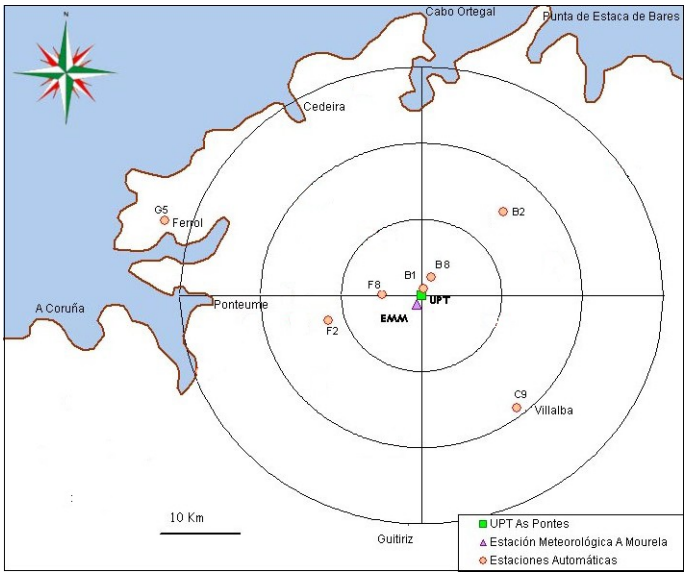


Figura 4.1: Red de Vigilancia de la Calidad Atmosférica de la U.P.T. de As Pontes.

La Estación Meteorológica, denominada E. M. de A Mourela, situada aproximadamente a un kilómetro de la instalación, está dotada de un mástil de 80 metros de altura con medidas de temperatura, velocidad y dirección de viento a distintos niveles, así como de sensores de humedad relativa, precipitación, radiación solar y presión atmosférica a nivel del suelo.

Debido a la legislación vigente y la localización de la U.P.T. de As Pontes, así como el compromiso de Endesa Generación S.A. con el medio ambiente, se estableció una fructífera relación entre la Sección de Medio Ambiente de la U.P.T. As Pontes y el Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela.

De esta relación nace el Sistema de Predicción Estadística de Inmisión (SIPEI) que permite obtener predicciones de los valores de dióxido de azufre y de óxidos de nitrógeno, con media hora de antelación, usando para ello modelos aditivos. Además, este sistema también predice cuál es el origen del episodio de alteración de calidad de aire, ya que este puede ser causado por la Central Térmica, el Ciclo Combinado u otros posibles focos como por ejemplo el tráfico o las actividades agrícolas de la zona.

En resumen, uno de los problemas que se plantea es poder predecir los niveles de

NO_x, a partir de la información que se recibe en continuo de las estaciones de muestreo y la información pasada de dichas medidas. Disponemos entonces de bases de datos minutales a partir de las medidas de emisión, meteorología y calidad de aire medidas. Debemos en este punto agradecer a la Sección de Medio Ambiente de la U.P.T. As Pontes su amabilidad por proporcionarnos los datos y toda la información necesaria para la elaboración de este trabajo.

4.2. Estimación del modelo de regresión cuantil

En esta sección vamos a ajustar un modelo de regresión cuantil asociado al cuantil $\tau = 0.5$ (la mediana) a los datos medioambientales que hemos presentado en la sección anterior. Debemos recordar en este punto que trabajaremos con mediciones minutales de óxidos de nitrógeno (NO_x) tomadas en las inmediaciones de la Unidad de Producción Térmica de As Pontes.

Para construir la muestra de trabajo emplearemos los datos correspondientes a un día completo (1440 observaciones minutales), que dividiremos en bloques de seis observaciones. Las cinco primeras mediciones de cada bloque están asociadas a cinco variables explicativas que denotaremos por X_i, mientras que el último valor de cada bloque representará la variable respuesta que denotaremos por Y. Evitamos de esta forma una cierta correlación entre las variables de distintos bloques.

Trabajaremos entonces con una muestra de la forma $\{(x_i, y_i) \text{ con } i = 1, \cdots, 240\}$ donde $x_i = (x_{1,i}, \cdots, x_{5,i})$, es decir, la variable respuesta representará la medición de NO_x en un cierto minuto y las variables explicativas representarán las mediciones de NO_x en los cinco minutos anteriores. A partir de este momento nos referiremos a $\{(x_i, y_i)\}_{i=1}^{240}$ como **muestra base** y denotaremos por $n = 240$ al tamaño muestral.

Ayudándonos de la función *rq* disponible dentro de la librería *quantreg* (toda la información a cerca de esta función puede encontrarse en http://cran.r-project.org/web/packages/quantreg/quantreg.pdf), ajustaremos un modelo lineal de la forma:

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 + \theta_5 X_5 + \varepsilon \tag{4.1}$$

siendo ε el error desconocido del modelo. La sintaxis, junto con los resultados obtenidos, figuran a continuación:

```
> modelo<-rq(Y~X1+X2+X3+X4+X5,tau=0.5)
> modelo

Call:  rq(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Coefficients:

(Intercept)          X1          X2          X3          X4          X5
 0.09678477  0.06889496 -0.18451555  0.30284451 -0.60023966  1.40393969
```

```
Degrees of freedom: 231 total; 225 residual
```

Con lo cual acabamos de ajustar el siguiente modelo de regresión cuantil a nuestro conjunto de datos:

$$\hat{Y} = 0.0968 + 0.0689X_1 - 0.1845X_2 + 0.3028X_3 - 0.6002X_4 + 1.4039X_5 \qquad (4.2)$$

que nos permite hacer predicciones gracias a la función *predict* disponible dentro de la mencionada librería *quantreg*. A continuación representamos la evolución diaria de los valores reales junto con las predicciones asociadas a la variable respuesta Y que estamos estudiando, calculadas gracias al modelo (4.2). La representación obtenida sería:

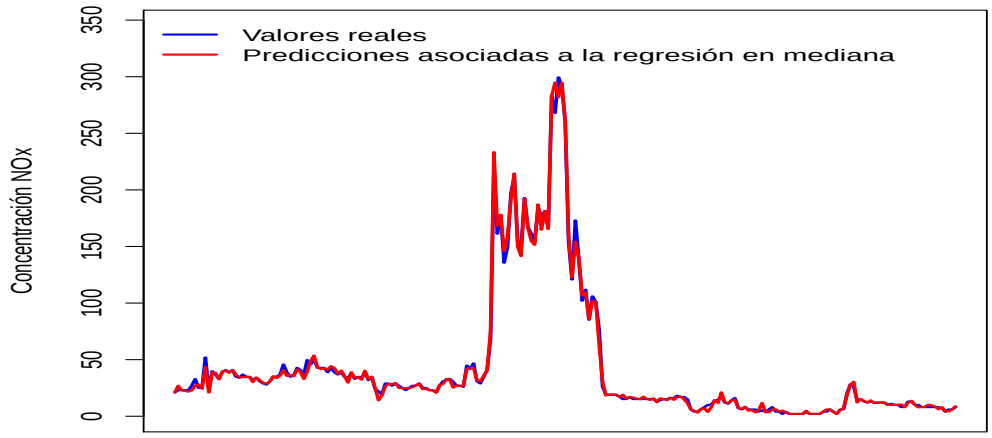


Figura 4.2: Representación de la variable respuesta junto con las correspondientes predicciones calculadas gracias al modelo (4.2) de regresión en mediana.

4.3. Contraste de bondad de ajuste

A la vista de los resultados anteriores, somos capaces de predecir nuevos valores de la variable respuesta Y en función de las cinco variables explicativas que venimos denotando por X_i con $i = 1, \cdots, 5$. En función de estas predicciones calculadas, podríamos extraer determinadas conclusiones acerca de la evolución de las concentraciones de NO_x en el aire.

Surge entonces la necesidad de contrastar si el modelo propuesto se ajusta a nuestro conjunto de datos, puesto que en caso contrario todas las conclusiones que podríamos extraer no serían correctas.

Nos disponemos entonces a aplicar la nueva propuesta de contraste de bondad de ajuste así como el test de He y Zhu (2003) a la muestra base que venimos utilizando.

Entonces el contraste a realizar es el siguiente:

$$\begin{cases} H_0 : g \in \mathcal{M}_\theta = \left\{ g(X,\theta) = \theta_0 + \theta_1 X_1 + \cdots + \theta_d X_d \ : \ \theta = (\theta_0, \cdots, \theta_d) \in \Theta \subset \mathbb{R}^{d+1} \right\} \\ H_a : g \notin \mathcal{M}_\theta \end{cases}$$

siendo en este caso $d = 5$.

Para este caso concreto, el valor de los estadísticos de contraste asociados a la nueva propuesta basada en proyecciones y al test de He y Zhu (2003) vendrían dados por:

$$T_n = \text{mayor autovalor de } \int_\Pi R_n^1(\beta,u)[R_n^1(\beta,u)]' F_{n,\beta}(du)d\beta = 7374.364$$

$$V_n = \text{mayor autovalor de } \ n^{-1}\sum_{i=1}^n S_n(x_i)S_n(x_i)' = 409.021$$

respectivamente, siendo:

$$R_n^1 = n^{-1/2}\sum_{i=1}^n \psi(r_i)\mathbf{x_i}\mathbb{I}(\beta' x_i \leq u)$$

$$S_n(t) = n^{-1/2}\sum_{i=1}^n \psi(r_i)\mathbf{x_i}\mathbb{I}(x_i \leq t)$$

donde $\mathbf{x_i} = (1, x_i)$, r_i denotan los residuos del modelo, $F_{n,\beta}(u)$ representa la función de distribución empírica de las variables explicativas proyectadas y $d\beta$ representa la densidad uniforme en la esfera unidad $\mathbb{S}_5$.

En la siguiente tabla mostramos los valores críticos así como los p-valores de ambos contrastes aproximados gracias a un procedimiento wild bootstrap (detallado en la Sección 3 del Capítulo 2) con diferentes números de réplicas (que denotamos por B) y diferentes niveles de significación (que denotaremos por α). Los resultados obtenidos son los siguientes:

	Test basado en proyecciones				Test de He y Zhu (2003)			
	Nivel de significación			p-valor	Nivel de significación			p-valor
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$		$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	
$B = 500$	45765.93	58848.39	81651.27	0.868	551.01	676.63	1015.56	0.212
$B = 1000$	41439.23	55584.39	81673.10	0.868	573.83	684.19	933.12	0.245
$B = 10000$	43088.33	55538.94	83302.99	0.8663	568.57	698.96	981.25	0.2359

Tabla 4.1: Valores críticos y p-valores asociados a la nueva propuesta de test de bondad de ajuste y al test de He y Zhu (2003) para diferentes números de remuestras bootstrap.

En ninguno de los escenarios anteriores rechazaríamos la hipótesis nula con ninguno de los dos contrastes de bondad de ajuste propuestos. En primer lugar, considerando la

propuesta del nuevo test basado en proyecciones, aceptaríamos la hipótesis nula puesto que el valor T_n del estadístico de contraste es menor que los valores críticos asociados a diferentes niveles de significación. De hecho, el p-valor toma valores próximo a 0.86, claramente superior a cualquier nivel de significación razonable.

Por otra parte, teniendo en cuenta el test propuesto por He y Zhu (2003), tampoco rechazaríamos la hipótesis nula puesto que el valor del estadístico de contraste es menor que cualquiera de los valores críticos que mostramos en la tabla anterior. Además, a la vista de los resultados anteriores el p-valor en este caso del orden de 0.24.

Con lo cual el modelo (4.1) propuesto se ajusta bien a la muestra base de datos medioambientales que estamos estudiando. Es decir, las conclusiones que podríamos extraer de dicho modelo son correctas ya que el modelo propuesto no presenta desviaciones significativas. Nótese además que estamos trabajando con un tamaño muestral $n = 240$ que debería ser suficiente para detectar desviaciones importantes. Por lo tanto, el modelo parece adaptarse a la realidad.

4.4. Conclusiones

A lo largo de todo este capítulo hemos trabajado con un conjunto de datos medioambientales que recogen las concentraciones de NO_x en cada minuto a lo largo de todo un día. La evolución minutal de estas concentraciones se muestra en la siguiente figura:

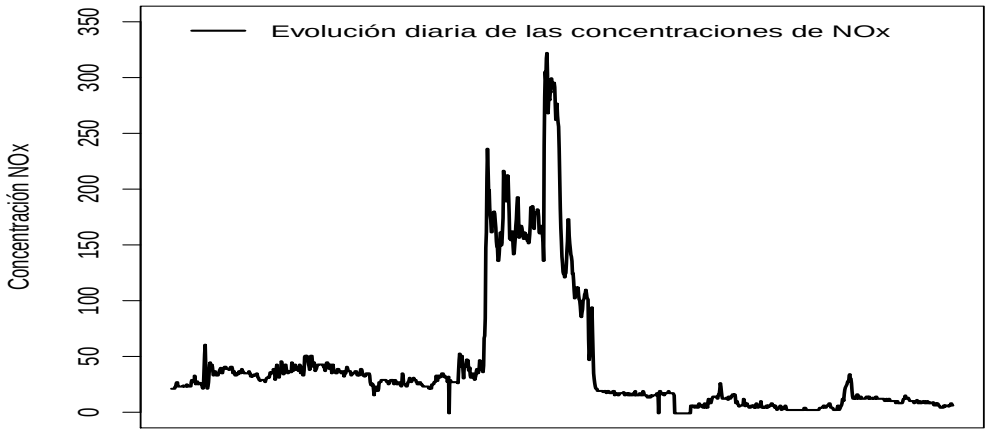


Figura 4.3: Representación de la evolución diaria de la concentración de NO_x.

En función de los datos anteriores, hemos intentado explicar la concentración de NO_x en un cierto minuto en función de las mediciones obtenidas los cinco minutos anteriores, empleando para ello un modelo lineal de regresión en mediana.

Gracias a la nueva propuesta de contraste de bondad de ajuste hemos visto que el

modelo anterior resulta satisfactorio para explicar la evolución minutal de las concentraciones de NO_x para cualquier nivel de significación razonable a la vista de los p-valores obtenidos. Nótese que el test propuesto por He y Zhu (2003) tampoco rechazaría la veracidad del ajuste propuesto.

Por lo tanto, gracias a la aplicación de ambos contrastes de bondad de ajuste nos hemos dado cuenta de que podemos extraer conclusiones en función del modelo propuesto ya que dichas conclusiones no presentarían desviaciones significativas sobre la evolución real de las concentraciones de NO_x.

Implementación informática

Para la realización de este trabajo se ha utilizado mayoritariamente el software libre R (http://www.r-project.org/) aunque ciertas partes han sido programados en Fortran como consecuencia del importante ahorro de tiempo computacional obtenido. En concreto, el cálculo de la matriz $A_{ij\bullet}$(definida en la Sección 4 del Capítulo 2) así como los estadístico de contraste asociados al test de He y Zhu (2003) y a la nueva propuesta de contraste basado en proyecciones han sido programados en lenguaje Fortran. En esta línea, ha sido necesaria la utilización del programa Rtools para poder importar los programas de Fortran al lenguaje R.

Mostramos a continuación, a modo de ejemplo, el código programado para el cálculo de la matriz $A_{ij\bullet}$, que ha sido programada dentro del entorno Fortran:

```
subroutine aij(n, d, x, a)

implicit none
integer n, d, i, j, r, h
double precision x(n,d), a(n,n)
double precision aux1(d), aux2(d)
double precision aux3, norma1, norma2, pi, prod, cte

pi=3.14159265359
cte=(pi**(d/2.0-1))/gamma(d/2.0+1)

a(n,n)=pi*(n+1)*cte

do i=1,(n-1)
       a(i,i)=pi*(n+1)*cte  ! si i=j y i!=r entonces  A_ijr^(0)=pi
                            ! si i=j=r entonces A_ijr^(0)=2*pi
       do j=(i+1),n
               a(i,j)=2*pi ! si r=i o r=j A_ijr^(0)=pi
               do r=1,n
                      if ((r.NE.i).AND.(r.NE.j)) then
                            do h=1,d
                                  aux1(h)=x(i,h)-x(r,h)
                                  aux2(h)=x(j,h)-x(r,h)
                            end do
```

```
                              norma1=dot_product(aux1,aux1)
                              norma2=dot_product(aux2,aux2)
                              prod=dot_product(aux1,aux2)
                              aux3=acos(prod/sqrt(norma1*norma2))
                              a(i,j)=a(i,j)+ abs(pi-aux3)
                        end if
                end do

                a(i,j)=a(i,j)*cte
                a(j,i)=a(i,j) ! Completamos la parte inferior de la matriz
        end do
end do

 return
 end
```

Dentro del software R, hemos utilizado la librería *quantreg* que ha sido desarrollada por R. Koenker y está dedicada íntegramente a la regresión cuantil. Dentro de dicha librería tenemos disponible la función *rq* que nos permite ajustar modelos lineales de regresión cuantil. Gracias a esta función, hemos podido ajustar los diferentes modelos de regresión cuantil que hemos propuesto a lo largo de los Capítulos 3 y 4.

En la sección 2 del Capítulo 4 ya hemos mostrado la sintaxis asociada a la función *rq* que hemos mencionado anteriormente. Por lo tanto, como ejemplo del código desarrollado dentro del lenguaje R, vamos a mostrar la generación de muestras aleatorias asociadas al modelo (3.9) que hemos simulado en el Capítulo 3. El código utilizado en este caso es el siguiente:

```
muestraA4=function(n,theta,g,sigma){

  # Covariables
  w=runif(n,min=0,max=2*pi)
  w1=runif(n,min=0,max=2*pi)
  w2=runif(n,min=0,max=2*pi)
  x1=(w+w1)/2
  x2=(w+w2)/2

  a=theta[1]+theta[2]*x1+theta[3]*x2
  y=a+g*cos(0.6*pi*a)

  # Error del modelo
  error=rnorm(n,sd=sigma)

  # Variable repuesta
  y=as.numeric(y+error)
```

```
    return(list(x1=x1,x2=x2,y=y,error=error))
  }
```

donde por n denotamos el tamaño muestral, *theta* es el vector de parámetros y *sigma* representa la desviación estándar del error. Finalmente, g marca la desviación del modelo simulado sobre la hipótesis nula asociada al contraste de bondad de ajuste que hemos presentado en la Sección 1 del Capítulo 3.

Referencias

Bahadur, R. R. (1966). A note on quantile in large samples. *The Annals of Mathematical Statistics*, *37*, 577-580.

Barrodale, I. y Roberts, F. D. K. (1973). An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, *10*, 839-848.

Bofinger, E. (1975). Estimation of a density function using order statistics. *Austrian Journal of Statistics*, *17*, 1-7.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *The Annals of Statistics*, *27*, 1012-1040.

Dette, H., Guhlich, M. y Neumeyer, N. (2012). Testing for additivity in nonparametric quantile regression. *http://www.ruhr-uni-bochum.de/mathematik3/-research/index.html.*.

Escanciano, J. C. (2006). A consistent disnostic test for regression models using projections. *Econometric Theory*, *22*, 1030-1051.

Eubank, R. L. y Hart, J. D. (1993). Commonality of cusum, von Neumann and smoothing-based goodness-of-fit test. *Biometrika*, *80*, 89-98.

Fan, J. y Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *Test*, *16*, 409-444.

Fan, J., Zhang, C. y Zhang, J. (2001). Generalised likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, *29*, 153-193.

Feng, X., He, X. y Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika*, *98*(4), 995-999.

Gao, J. y Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association*, *103*, 1584-1594.

González-Manteiga, W. y Cao, R. (1993). Testing for hypothesis of a general linear model using nonparametric regression estimation. *Test*, *2*, 161-188.

González-Manteiga, W. y Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test. Pendiente de publicación.*

Hall, P. y Sheather, S. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society, Series B (Methodological)*, *50*, 381-391.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. y Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* Wiley.

Härdle, W. y Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, *21*, 1926-1947.

He, X. y Shao, Q. M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, *24*, 2608-2630.

He, X. y Zhu, L. X. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, *98*, 1013-1022.

Huber, P. J. (1981). *Robust statistics.* Wiley.

Koenker, R. (2005). *Quantile regression.* Cambridge.

Koenker, R. y Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33-50.

Koenker, R. y D'Orey, V. (1987). Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *36*, 383-393.

Kulasekera, K. B. y Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association*, *92*, 500-511.

Lavergne, P. y Patilea, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics*, *143*, 103-122.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, *10*, 186-196.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces.* Academic Press.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, *25*, 613-641.

Stute, W., González-Manteiga, W. y Presedo-Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, *93*, 141-149.

Stute, W., Xu, W. L. y Zhu, X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, *95*, 451-467.

Tukey, J. (1965). Which part of the sample contains the information. *Proceedings of the National Academy of Sciences*, *53*, 127-134.

Van Keilegom, I., González-Manteiga, W. y Sánchez-Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test*, *17*, 401-415.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, *26*, 359-372.

Wilcox, R. R. (2008). Quantile regression: A simplified approach to a goodness-of-fit test. *Journal of Data Science*, *6*, 547-556.

Xia, Y. (2009). Model checking in regression via dimension reduction. *Biometrika*, *96*, 133-148.

Zhang, C. (2004). Assessing the equivalence of nonparametric regression tests based on spline and local polynomial smoothers. *Journal of Statistical Planning and Inference*, *126*, 73-95.

Zhang, C. y Dette, H. (2004). A power comparison between nonparametric regression tests. *Statistics and Probability Letters*, *66*, 289-301.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*, 263-289.

Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, *14*, 123-138.