Contrastes en regresión no paramétrica basados en los momentos de los residuos

Proyecto Fin de Máster - Máster en Técnicas Estadísticas

Alumna: Ana Vanessa Vinseiro Mera

Tutor: Juan Carlos Pardo Fernández

Juan Carlos Pardo Fernández, profesor del Departamento de Estadística e Investigación Operativa de la Universidade de Vigo,

HACE CONSTAR

que el presente trabajo titulado *Contrastes en regresión no paramétrica basados en los momentos de los residuos* ha sido realizado por Ana Vanessa Vinseiro Mera bajo su dirección para su presentación como Trabajo Fin de Máster del *Máster en Técnicas Estadísticas*.

Vigo, 29 de junio del 2012.

Fdo.: Juan Carlos Pardo Fernández

Resumen

Los contrastes de bondad de ajuste en modelos de regresión se ocupan de verificar si una curva de regresión pertenece a una cierta familia paramétrica de funciones. En los últimos años se han publicado diversos trabajos sobre contrastes en modelos de regresión basados en el estudio de diversas características de los errores. Dichos procedimientos se basan en la comparación entre la función de distribución empírica de los residuos paramétricos bajo la hipótesis nula y la función de distribución empírica de los residuos no paramétricos.

En este trabajo se plantean contrastes alternativos basado en los momentos de primer y segundo orden de los errores de regresión. Los métodos de contraste propuestos se validan mediante simulaciones y finalmente se ilustran mediante el análisis de tres conjuntos de datos reales.

Índice general

Capítulo 1

Introducción

1.1. Generalidades sobre los contrastes de hipótesis

Dentro de la inferencia estadística, un **contraste de hipótesis** (también denominado **test de hipótesis** o **prueba de significación**) es un procedimiento para juzgar si una propiedad supuesta en una población estadística es compatible con lo observado en una muestra de dicha población.

Con el fin de realizar un contraste, es necesario establecer dos hipótesis, llamadas hipótesis nula e hipótesis alternativa:

- La **hipótesis nula**, H_0, es la afirmación que inicialmente se supone verdadera.
- La **hipótesis alternativa**, H_1, es la hipótesis que queremos contrastar en contra de la hipótesis nula.

Un contraste de hipótesis es un procedimiento estadístico para decidir si la hipótesis nula debe ser rechazada, en base a la información contenida en una muestra de datos. La hipótesis nula H_0 se acepta si los datos no muestran una fuerte evidencia en contra de ella. Por otro lado, si se observa que los datos sugieren fuerte evidencia en contra de H_0, entonces esta hipótesis será rechazada a favor de la hipótesis alternativa H_1.

Con el fin de realizar un contraste de hipótesis, necesitamos un criterio estadístico para decidir si la hipótesis nula debe ser rechazada o no sobre la base de una muestra

de observaciones. El **estadístico de contraste**, D, es un criterio utilizado para medir la discrepancia entre la hipótesis nula H_0 y las observaciones. Dicho estadístico, es una función de los datos, de tal manera que su distribución se conoce completamente (o puede ser aproximada) cuando la hipótesis nula se supone que es verdadera.

Una vez que el estadístico de contraste se ha elegido, debemos clasificar sus posibles valores en dos regiones:

- La **región de aceptación** es el conjunto de todos los valores del estadístico de contraste para los cuales la hipótesis nula H_0 será aceptada. En otras palabras, es el conjunto de todos los valores del estadístico de contraste que no evidencian gran discrepancia entre H_0 y los datos.

- La **región de rechazo** o **región crítica** es el conjunto de todos los valores del estadístico de contraste que muestran gran discrepancia entre los datos y la hipótesis nula H_0. Por lo tanto, si el valor observado del estadístico de contraste pertenece a la región de rechazo, entonces la hipótesis nula H_0 será rechazada en favor de la hipótesis alternativa H_1.

La decisión a favor o en contra de la hipótesis nula se hará en base a la diferencia observada entre la hipótesis nula y una muestra de observaciones. Obviamente, esa decisión puede ser correcta o incorrecta, lo que nos puede llevar a cometer dos tipos de errores:

- En un contraste de hipótesis, la decisión de rechazar la hipótesis nula H_0 cuando ésta es verdadera se denomina **error de tipo I**. La probabilidad de este error es el **nivel de significación** del contraste, y se denota por α:

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

- En un contraste de hipótesis, la decisión de aceptar la hipótesis nula H_0 cuando H_0 no es cierta se llama **error de tipo II**, y su probabilidad se denota por β:

$$\beta = P(\text{ error de tipo II }) = P(\text{ aceptar } H_0 \mid H_0 \text{ no es cierta }).$$

En una situación ideal, ambas probabilidades de error, α y β, deberían ser pequeñas, pero esto no siempre es posible. Generalmente, cuando tratamos de hacer pequeña una probabilidad de error, entonces la otra aumenta. En la mayor parte de los casos, se considera que el error de tipo I es el más relevante, por lo que solo se trata de controlar la probabilidad de este tipo de error. Así, se fija de antemano una probabilidad máxima para el error de tipo I (por lo general $\alpha = 0.05$) y entonces la hipótesis nula H_0 se rechaza si hay una evidencia muy fuerte contra ella.

Una vez que hemos fijado un nivel de significación, α, los valores del estadístico de contraste se dividen en las regiones de aceptación y de rechazo:

- La región de aceptación, con probabilidad $1 - \alpha$ cuando H_0 es verdadera. Si el valor del estadístico de contrastes D obtenido de la muestra, pertenece a la región de aceptación, entonces no hay suficiente evidencia contra la hipótesis nula, con un nivel de significación α, se dice entonces que el contraste no es estadísticamente significativo.

- La región de rechazo o región crítica, con probabilidad α cuando H_0 es cierta. Si el valor del estadístico de contraste D obtenido en la muestra pertenece a la región de rechazo, entonces la evidencia en contra de la hipótesis nula H_0 es fuerte. En este caso, la hipótesis nula será rechazada en favor de la hipótesis alternativa, y se dirá que el contraste es estadísticamente significativo.

La elección de dichas regiones de aceptación y rechazo se realiza tras un análisis de las hipótesis H_0 y H_1 y la distribución del estadístico de contraste D, de forma que los valores en la región de rechazo sean aquellos que muestran una mayor discrepancia entre H_0 y los datos (generalmente en las colas de la distribución de D bajo H_0).

Si el contraste rechaza la hipótesis nula a un nivel de significación α dado, también se rechaza a cualquier nivel α', tal que $\alpha' > \alpha$. Supongamos que por cada nivel de significación $0 < \alpha < 1$ se obtiene una región de rechazo R_α, entonces se define el **p-valor** como

$$\text{p-valor} = \text{ínf}\left\{\alpha : d \in R_\alpha\right\},$$

donde d es el valor del estadístico de contraste observado en la muestra. Cuanto menor sea el p-valor, mayor es la evidencia en contra de la hipótesis nula H_0. Si el p-

valor$< \alpha$ entonces se rechaza H_0 (contraste significativo) y si el p-valor$\geq \alpha$ entonces se acepta H_0 (contraste no significativo).

De forma general, los contrastes de hipótesis se pueden clasificar en dos grandes grupos: contrastes paramétricos y contrastes no paramétricos. Los **contrastes de hipótesis paramétricos** se emplean cuando la estructura global de los modelos implicados en el contraste (distribución, modelos de regresión, etc.) es conocida, a falta de un conjunto de parámetros, y el contraste se establece sobre esos parámetros. Por otra parte, los **contrastes de hipótesis no paramétricos** son aquellos en los que se plantean hipótesis muy generales sobre la estructura de los datos. En este trabajo nos centraremos en los contrastes de hipótesis no paramétricos.

Dentro de los contrastes no paramétricos, destacan dos grupos: los contrastes sobre modelos de distribución y los contrastes sobre modelos de regresión. En los **contrastes sobre modelos de distribución** nos encontramos con:

- **Contrastes de bondad de ajuste**. Muchos de los procedimientos formales de la inferencia estadística se basan en el supuesto que la distribución de una población es de un tipo específico. El uso de estos procedimientos puede ser inadecuado si la verdadera distribución de probabilidad subyacente es muy diferente de la distribución que se supone. En estos casos, los contrastes estadísticos que comprueban la calidad de un modelo o la forma de una distribución para un conjunto de datos se conocen como contrastes de bondad de ajuste.

- **Comparación de distribuciones**. Al analizar varias muestras, se puede estar interesado en comprobar si provienen de la misma población. Si es así, las muestras pueden ser tratadas conjuntamente con el fin de obtener estimaciones e inferencias más eficientes. Por otro lado, si sucede lo contrario, no debemos utilizar el total de la muestra para realizar cualquier proceso de inferencia. Por lo general, este tipo de problemas se llaman problemas de homogeneidad o de k muestras.

- **Contraste de independencia**. Cuando se observan dos o más características de los elementos de una población, es interesante comprobar si estas características son independientes o no. En el caso de que las características sean inde-

pendientes, las muestras pueden ser analizadas por separado. Por otra parte, si se puede establecer alguna relación entre ellas, entonces pueden ser analizadas conjuntamente, por medio, por ejemplo, de modelos de regresión.

Entre los **contrastes sobre modelos de regresión** podemos citar los siguintes casos:

- **Contrastes de bondad de ajuste**. Actualmente, la expresión bondad de ajuste se utiliza no sólo en problemas de distribuciones, sino también en contextos más generales, tales como la regresión. Muchos de los procedimientos estadísticos se basan en suponer que una familia de regresión pertenece a una cierta familia paramétrica (lineal, polinómica, logística,...). En este caso los contrastes de bondad de ajuste comprueban la forma de la familia de regresión de un conjunto de datos. El presente trabajo se encuadra en este bloque de contrastes.

- **Comparación de curvas**. Uno de los problemas más importantes en la inferencia estadística es la comparación de dos o más grupos de variables. Esta comparación puede realizarse comparando medias, medianas o alguna otra característica de la variable de interés medida para cada grupo. Cuando esta variable es acompañada por covariables, un objetivo más ambicioso consiste en comparar las funciones de regresión correspondientes.

- **Otros contrastes**, entre los cuales nos podemos encontrar contrastes sobre la varianza condicional, contrastes sobre modelos parcialmente lineales, contrastes sobre modelos aditivos, ...

Como hemos dicho, este trabajo se enmarca dentro de los contrastes de bondad de ajuste para modelos de regresión. Para formalizar el planteamiento de estos contrastes, consideremos un vector aleatorio (X, Y) de forma que la respuesta Y y la(s) covariable(s) X satisfacen el modelo de regresión siguiente:

$$Y = m(X) + \sigma(X)\varepsilon \tag{1.1}$$

donde ε es el error de regresión, $m(x) = E(Y|X = x)$ es la función de regresión desconocida y $\sigma^2(x) = Var(Y|X = x)$ es la función de varianza condicional. Bajo este modelo de regresión, queremos contrastar la hipótesis

$$H_0 : m \in \mathcal{M} \tag{1.2}$$

frente a la alternativa general

$$H_1 : m \notin \mathcal{M},$$

donde $\mathcal{M} = \{m_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ es alguna clase paramétrica de funciones de regresión indexadas por el parámetro θ (véase, por ejemplo, Seber (1977) para el caso de regresión lineal y Seber and Wild (1989) para m_θ no lineal).

A continuación se muestran algunos ejemplos de clases de funciones de regresión que podrían constituir clases paramétricas:

- **Regresión polinómica**. Sea X una variable aleatoria unidimensional, entonces
$$m_\theta(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \ldots + \theta_p x^{p-1}$$
- **Regresión lineal multiple**. Si la covariable X es p-dimensional, entonces
$$m_\theta = \theta^t x,$$
siendo $x = (x_1, \ldots, x_p)^t \in \mathbb{R}^p$ y $\theta = (\theta_1, \ldots, \theta_p)^t \in \Theta \subset \mathbb{R}^p$.
- **Algún modelo de regresión no lineal**. Considérese $X = (X_1, X_2)^t$ y sea $\theta = (\theta_1, \theta_2)^t \in \Theta = \{(s,t) \in \mathbb{R}^2, s + t \neq 0\}$. Un ejemplo de modelo de regresión no lineal es
$$m_\theta(x) = \frac{1}{\theta_1 + \theta_2} \exp(\theta_1 x_1 + \theta_2 x_2).$$

1.2. Estimaciones de la función de regresión

Los modelos de regresión tratan de explicar la dependencia que existe entre la variable respuesta Y y la variable dependiente o covariable X a partir de una muestra aleatoria simple $(X_1, Y_1), \ldots, (X_n, Y_n)$ de (X, Y). Esta dependencia se estudia a

través de la función de regresión, m, la cual se puede estimar de forma paramétrica, $m_{\widehat{\theta}}$, o bien de forma no paramétrica, $\widehat{m}$.

Dado un modelo paramétrico de regresión, la **estimación paramétrica de la función de regresión** tiene como objetivo buscar los parámetros que mejor ajustan la nube de puntos al modelo de regresión. La diferencia entre el valor real de la respuesta, Y_i, y el valor de predicción dado por el modelo paramétrico, $m_\theta(X_i)$, se denomina residuo: $e_i = Y_i - m_\theta(X_i)$. Un método muy utilizado para estimar el vector de paramétros θ es el **método de mínimos cuadrados**, que consiste en estimar θ mediante el valor que minimiza

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - m_\theta(X_i))^2.$$

En algunos casos tenemos que introducir pesos en el proceso de minimización para obtener una estimación adecuada, entonces se utilizará el **método de mínimos cuadrados ponderados** para estimar los parámetros, es decir, se buscan θ que minimiza la siguiente suma de cuadrados ponderada

$$\sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} w_i (Y_i - m_\theta(X_i))^2,$$

donde w_i es el vector de pesos positivos.

La **estimación no paramétrica de la función de regresión** se basa en no suponer ninguna hipótesis paramétrica sobre la función de regresión. Simplemente se asumen condiciones generales de suavidad (continuidad o derivabilidad). Los métodos empleados en esta situación se llaman métodos de suavizado.

Consideremos el modelo de regresión (1.1), donde m es la función de regresión. Como estimador no paramétrico de m, se tiene el estimador tipo núcleo propuesto por Nadaraya (1964) y Watson (1964), que viene dado por la siguiente expresión:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)},$$

donde K es la función núcleo (típicamente, una función de densidad simétrica) y h es el llamado parámetro de suavizado o ventana. Algunos de los núcleos más utilizados son los siguintes:

- Uniforme: $K(x) = \frac{1}{2}\mathbb{I}(\mid x \mid < 1)$.
- Gaussiano: $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$.
- Epanechnikov: $K(x) = \frac{3}{4}(1-x^2)\mathbb{I}(\mid x \mid < 1)$.

El estimador tipo núcleo hereda las propiedades de suavidad del núcleo, sin embargo, la forma del mismo no tiene un papel tan determinante como la ventana h. Para un valor de h fijo, el nivel de suavizado que proporcionan los diferentes núcleos puede ser muy distinto. Sin embargo, una vez que se elimina este efecto, el papel del núcleo no es demasiado determinante; véase por ejemplo Wand and Jones (1995). Nótese que el estimador de Nadaraya-Watson se puede reescribir como

$$\widehat{m}(x) = \sum_{j=1}^{n} W_{nj}(x) Y_j,$$

donde

$$W_{nj}(x) = \frac{K(x - X_j)}{\sum_{k=1}^{n} K(x - X_k)}.$$

Por tanto el estimador tipo núcleo de la función de regresión no es más que una media (local) ponderada de los valores observados de la variable Y, ya que $\sum_{j=1}^{n} W_{nj}(x) = 1$. La cantidad de datos que entra en esta media, así como sus ponderaciones dependen fundamentalmente del parámetro de suavizado.

El comportamiento del estimador de Nadaraya-Watson depende de los valores del parámetro ventana, h. A medida que h se hace pequeño, es decir, cuando se aproxima a 0, el estimador converge a las observaciones de la variable aleatoria Y, Y_i, por tanto, el estimador tiende a interpolar los datos si h es pequeño, lo que nos lleva al infrasuavizado. Por otra parte, si h es grande, el estimador converge a la media de Y, es decir, el estimador es una función constante, lo que conlleva al sobresuavizado.

Veamos un ejemplo con datos simulados. Los datos provienen de $X \sim U[0,1]$ y $\varepsilon \sim N(0,1)$, con tamaño muestral $n = 500$. Se considera el modelo de regresión

$$Y = -5X + \cos(2\pi X) + \varepsilon.$$

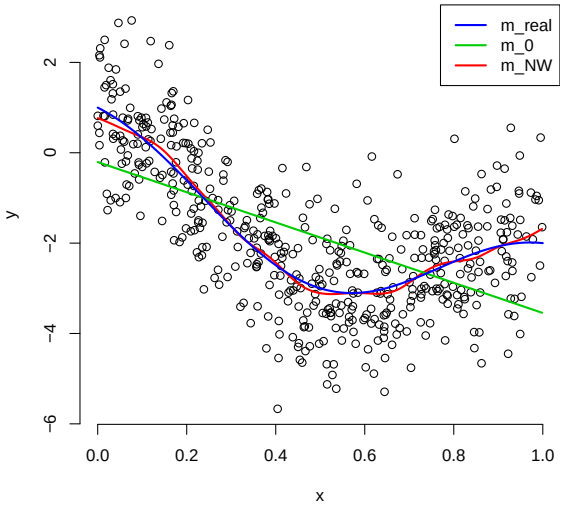

Figura 1.1: Ejemplo simulado

Gráficamente, tenemos la situación dada en la Figura 1.1, en la que se muestran los datos, la verdadera función de regresión, el estimador de Nadaraya-Watson y el estimador paramétrico si se asume un modelo lineal. Evidentemente, al no hacer suposiciones sobre el modelo con el estimador no paramétrico, éste aproxima mejor la función de regresión que el estimador paramétrico.

Otro estimador no paramétrico también muy popular es el de la regresión polinómica local (Fan and Gijbels, 1996), que ajusta localmente un polinomio. Se trata de una generalización del estimador de Nadaraya-Watson y posee buenas propiedades a la hora de su aplicación práctica.

1.3. Breve revisión de métodos de bondad de ajuste en regresión

En los últimos veinte años la literatura relacionada con el problema de bondad de ajuste en el contexto de la regresión ha crecido rápidamente. De forma general, los procedimientos de contraste se pueden clasificar en tres grandes bloques:

1. Contrastes basados en la estimación de la función de regresión.

2. Contrastes basados en la estimación de la función de regresión integrada.

3. Contrastes basados en los residuos.

1.3.1. Contrastes basados en la estimación de la función de regresión

Partiendo del modelo de regresión (1.1), se considera el problema básico de contrastar (1.2). En muchos casos no existe ninguna teoría que diga como debe ser m. El análisis de la información empírica disponible nos debería proporcionar dicha información.

Básicamente, los métodos de contraste basados en la estimación de la función de regresión consisten en comparar una estimación paramétrica con una estimación no paramétrica a través de algún tipo de distancia funcional. Es muy natural comparar ambas estimaciones con el fin de decidir si el modelo paramétrico está en consonancia con los datos. Esta comparación se puede realizar de dos formas:

- Suavizando los datos: se considera la distancia entre $\widehat{m}$ y $m_{\widehat{\theta}}$.
- Suavizando los datos y la hipótesis nula: se considera la distancia entre $\widehat{m}$ y $\widehat{m}_{\widehat{\theta}}$, siendo $\widehat{m}_{\widehat{\theta}}(x) = \sum_{j=1}^{n} W_{nj}(x) m_{\widehat{\theta}}(X_j)$.

Durante los años noventa, las propuestas de diferentes alternativas se introducen en la literatura estadística:

- Härdle and Mammen (1993) introducen un contraste basado en la distancia L_2:
$$d_1(m, H_0) = \int (\widehat{m}(x) - \widehat{m}_{\widehat{\theta}}(x))^2 \pi_1(x) dx$$
donde $\widehat{m}_{\widehat{\theta}}(x)$ es el estimador polinómico-local basado en la muestra $(X_i, m_{\widehat{\theta}}(X_i))$, con $i = 1, \ldots, n$.
La idea de este contraste se basa en el valor esperado de
$$C_1 = E^2(\varepsilon_0 \mid X)\pi_1(X),$$

siendo $\varepsilon_0 = Y - m_{\theta_0}(X)$ y π_1 una función de peso, ya que $E(C_1) = 0$ si y sólo si la hipótesis nula H_0 es cierta.

- El contraste de Zheng (1996) considera:

$$d_2(m, H_0) = \frac{1}{n} \sum_{i \neq j} K_h(X_i - X_j)(Y_i - m_{\widehat{\theta}}(X_i))(Y_j - m_{\widehat{\theta}}(X_j))\pi_2(X_i).$$

Este contraste se basa en que el valor esperado de

$$C_2 = \varepsilon_0 E(\varepsilon_0 \mid X) f(X) \pi_2(X),$$

con π_2 una función de peso, es cero si y sólo si la hipótesis nula es cierta. La versión muestral de $E(C_2)$ está dada por

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{\widehat{\theta}}(X_i))(\widehat{m}_h(X_i) - \widehat{m}_{\widehat{\theta}}(X_i))\widehat{f}_h(X_i)\pi_2(X_i),$$

que coincide con $d_2(m, H_0)/n$, salvo por un término constante aditivo.

- El contraste propuesto en Dette (1999) (véase también Azzalini and Bowman (1993)) se basa en procesos del tipo:

$$d_3(m, H_0) = \sum_{i=1}^{n} (Y_i - m_{\widehat{\theta}}(X_i))^2 \pi_3(X_i) - \sum_{i=1}^{n} (Y_i - \widehat{m}_h(X_i))^2 \pi_3(X_i).$$

Este contraste se basa en que el valor esperado de

$$C_3 = E(\varepsilon_0^2 - (\varepsilon_0 - E(\varepsilon_0 \mid X))^2)\pi_3(X),$$

donde π_3 es una función de peso, es igual a cero si y sólo si H_0 es cierta. La versión muestral de $E(C_3)$ está dada por

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} ((Y_i - m_{\widehat{\theta}}(X_i))^2 - ((Y_i - m_{\widehat{\theta}}(X_i)) - (\widehat{m}_h(X_i) - \widehat{m}_{\widehat{\theta}}(X_i)))^2)\pi_3(X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} ((Y_i - m_{\widehat{\theta}}(X_i))^2 - (Y_i - \widehat{m}_h(X_i) - m_{\widehat{\theta}}(X_i) + \widehat{m}_{\widehat{\theta}}(X_i))^2)\pi_3(X_i), \end{aligned}$$

que coincide con $d_3(m, H_0)/n$ cuando $m_{\widehat{\theta}}(X_i) = \widehat{m}_{\widehat{\theta}}(X_i)$.

Estas tres propuestas de contraste utilizan la metodología bootstrap para la obtención de los valores críticos. En general, la hipótesis nula se rechaza para valores grandes de los estadísticos de contraste.

Nótese que si $\pi_1 \equiv \pi_2 f \equiv \pi_3 \equiv c$, para $c > 0$, entonces

$$E(C_1) = E(C_2) = E(C_3) = c(E(E^2(\varepsilon_0 \mid X))),$$

lo cual permite establecer una cierta analogía entre los tres contrastes descritos. Zhang and Dette (2004) ofrecen un estudio exhaustivo de la comparación de estos procedimientos bajo la hipótesis nula y alternativas fijas.

1.3.2. Contrastes basados en la estimación de la función de regresión integrada

Considérese el vector aleatorio (X, Y) que satisface el modelo (1.1). Se define la **función de regresión integrada** como

$$I(x) = E(Y 1_{\{X \leq x\}}) = \int_{-\infty}^{x} m(y) dF(y),$$

donde F es la función de distribución de X. Esta función caracteriza a la función de regresión (véase Stute (1997)) y además puede ser estimada empíricamente sin necesidad de recurrir a técnicas de suavizado mediante

$$I_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}} Y_i. \tag{1.3}$$

La idea principal del contraste es comparar un estimador no paramétrico de la regresión integrada con algún estimador basado en los supuestos de la hipótesis nula. Para ello se distinguen dos casos:

- Caso de la hipótesis nula simple (θ especificado completamente)
- Caso de la hipótesis nula compuesta (θ en un espacio paramétrico)

En el caso de la hipótesis nula simple se considera la hipótesis nula $H_0 : m = m_0$. Entonces, se toma como estimador no paramétrico el dado por la expresión (1.3) y como estimador paramétrico bajo la hipótesis nula H_0

$$I_0(X) = \int_{-\infty}^{x} m_0(y) dF_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}} m_0(X_i).$$

El proceso empírico marcado por los errores de regresión

$$R_n(x) = n^{1/2}(I_n(x) - I_0(x)) = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}(Y_i - m_0(X_i))$$

ha sido estudiado por Stute (1997).Para contrastar H_0 sólo se tiene que elegir algún funcional continuo sobre $R_n(x)$ (por ejemplo, el supremo que lleva a un estadístico de tipo Kolmogorov-Smirnov). El valor crítico puede obtenerse a partir de la distribución límite de tales funcionales.

Para el caso de la hipótesis nula compuesta se necesita un estimador de θ bajo H_0, digamos $\widehat{\theta}$, y el estadístico de contraste será un funcional del proceso

$$R_n^1 = n^{-1/2} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}(Y_i - m_{\widehat{\theta}}.(X_i))$$

El análisis teórico es considerablemente más complejo que para el caso de la hipótesis nula simple. Como consecuencia, esto provoca un problema a la hora de obtener los valores críticos del contraste, que se puede solventar utilizando técnicas bootstrap (Stute et al., 1998).

1.3.3. Contrastes basados en los residuos

Dada la hipótesis nula (1.2) acerca de las características individuales del modelo de regresión, Van Keilegom et al. (2008) proponen contrastes basados en los residuos partiendo del hecho que H_0 es verdadera si y sólo si los residuos del modelo de regresión, ε, y los residuos construidos bajo H_0, ε_0, tienen la misma distribución. Akritas and Keilegom (2001) propusieron la estimación de la distribución de los errores de regresión a través de la distribución empírica de los residuos estimados, mediante la siguiente expresión:

$$\widehat{F}_\varepsilon(y) = \frac{1}{n} \sum_{i=1}^{n} I(\widehat{\varepsilon}_i \leq y),$$

donde $\widehat{\varepsilon}_i$ son los residuos estimados

$$\widehat{\varepsilon}_i = \frac{Y_i - \widehat{m}(X_i)}{\widehat{\sigma}(X_i)},$$

para $i = 1, \ldots, n$, y donde $\widehat{m}$ y $\widehat{\sigma}^2$ son estimadores no paramétricos de la regresión y de la función de varianza, respectivamente. Para dichas estimaciones se emplean los estimadores de Nadaraya-Watson.

En Van Keilegom et al. (2008) el procedimiento de contraste está basado en la comparación entre la función de distribución empírica de los residuos paramétricos (bajo la hipótesis nula), $\widehat{F}_{\varepsilon 0}$, y la de los residuos no paramétricos (en la situación general), $\widehat{F}_{\varepsilon}$. Se estudia el proceso empírico $n^{1/2}(\widehat{F}_{\varepsilon 0}(y) - \widehat{F}_{\varepsilon}(y))$, y se consideran estadísticos de contrastes de tipo Kolmogorov-Smirnov,

$$T_{KS} = \sup_{y} \mid \widehat{F}_{\varepsilon 0}(y) - \widehat{F}_{\varepsilon}(y) \mid \tag{1.4}$$

y Cramér-von Mises,

$$T_{CM} = \int (\widehat{F}_{\varepsilon 0}(y) - \widehat{F}_{\varepsilon}(y))^2(y) d\widehat{F}_{\varepsilon}(y), \tag{1.5}$$

para medir la distancia entre las dos distribuciones empíricas. Finalmente, se usan técnicas bootstrap para aproximar los valores críticos del contraste.

Capítulo 2

Contrastes basados en los momentos de los residuos

Considerando el modelo de regresión $Y = m(X)+\sigma(X)\varepsilon$, donde $m(X) = E(Y|X)$ y $\sigma^2(X) = Var(Y|X)$ son desconocidas, y el error ε es independiente de la covariable X, Van Keilegom et al. (2008) proponen un procedimiento estadístico para contrastar si la función de regresión m pertenece a alguna familia parámetrica. Dicho contraste se basa en el estudio de la distancia entre las funciones de distribución empíricas de los residuos calculados de forma paramétrica y no paramétrica. En este capítulo, propondremos un contraste alternativo, basado en el estudio de los momentos de primer y segundo orden (media y varianza) de los residuos bajo la hipótesis nula.

2.1. Motivación y justificación teórica

Sea (X, Y) un vector aleatorio bidimensional satisfaciendo el siguiente modelo de regresión

$$Y = m(X) + \sigma(X)\varepsilon, \tag{2.1}$$

donde ε es la variable de error independiente de X con media $E(\varepsilon) = 0$ y varianza $Var(\varepsilon) = 1$, $m(x) = E(Y|X = x)$ es la función de regresión y $\sigma^2(x) = Var(Y|X = x)$ es la función de varianza condicional.

Se pretende contrastar la hipótesis nula

$$H_0 : m \in \mathcal{M}, \tag{2.2}$$

frente a la hipótesis alternativa general

$$H_1 : m \notin \mathcal{M},$$

donde $\mathcal{M} = \{m_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ es una familia paramétrica de funciones de regresión indexada por el parámetro θ. La hipótesis nula se puede reescribir como,

$$H_0 : \text{ existe } \theta_0 \in \Theta \text{ tal que } m = m_{\theta_0}. \tag{2.3}$$

Como ya mencionamos anteriormente, en el planteamiento de este contraste juegan un papel importante los errores. Despejando del modelo de regresión (2.1), se tiene la expresión de los verdaderos errores

$$\varepsilon = \frac{Y - m(X)}{\sigma(X)}.$$

Por otra parte, dado un valor del parámetro, $\tilde{\theta}$, también podemos construir los errores bajo la hipótesis nula:

$$\varepsilon_0 = \frac{Y - m_{\tilde{\theta}}(X)}{\sigma(X)} = \frac{Y - m(X)}{\sigma(X)} + \frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)} = \varepsilon + \frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}.$$

Van Keilegom et al. (2008) demuestran que la hipótesis nula (2.2) es cierta si y solo si las variables aleatorias ε y ε_0 tienen una misma distribución. El contraste propuesto en el presente trabajo se basa en la siguiente caracterización alternativa de la hipótesis nula H_0.

TEOREMA 2.1

$$H_0 \text{ es cierta} \iff \begin{cases} E(\varepsilon_0) = E(\varepsilon) = 0 \\ \text{y} \\ Var(\varepsilon_0) = Var(\varepsilon) = 1 \end{cases}$$

Demostración. Como hemos dicho, Van Keilegom et al. (2008) demuestran que la hipótesis nula (2.2) es cierta si y solo si las variables aleatorias ε y ε_0 tienen una

misma distribución. Esto nos basta para justificar que si la hipótesis H_0 es cierta, por tener los residuos la misma distribución, en particular se tiene que los momentos de primer y segundo orden coinciden.

La justificación del resultado inverso no es tan evidente. Considérese la expresión de los errores construidos bajo la hipótesis nula:

$$\varepsilon_0 = \varepsilon + \frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)},$$

para los cuales se asume $E(\varepsilon_0) = 0$ y $Var(\varepsilon_0) = 1$. Calculando la esperanza de ε_0 se tiene que

$$E\left(\varepsilon_0\right) = E\left(\varepsilon + \frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}\right) = E\left(\varepsilon\right) + E\left(\frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}\right),$$

de donde se sigue, por ser $E(\varepsilon_0) = E(\varepsilon) = 0$, que

$$E\left(\frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}\right) = 0. \tag{2.4}$$

Por otra parte, se calcula la varianza de los errores ε_0 (téngase en cuenta que ε y X son independientes):

$$Var\left(\varepsilon_0\right) = Var\left(\varepsilon + \frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}\right) = Var\left(\varepsilon\right) + Var\left(\frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}\right). \tag{2.5}$$

De modo análogo a la esperanza, puesto que $Var(\varepsilon_0) = Var(\varepsilon) = 1$, se sigue que

$$Var\left(\frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)}\right) = 0,$$

por lo tanto

$$P\left(\frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)} = \text{ constante }\right) = 1.$$

Además teniendo en cuenta (2.4) se obtiene que

$$P\left(\frac{m(X) - m_{\tilde{\theta}}(X)}{\sigma(X)} = 0\right) = 1.$$

Dado que la función σ es siempre positiva podemos concluir que

$$P\left(m(X) = m_{\tilde{\theta}}(X)\right) = 1,$$

o dicho de otra forma, existe $\tilde{\theta}$ tal que $m \in \mathcal{M}$. Se verifica por tanto la hipótesis nula H_0, como queríamos demostrar.

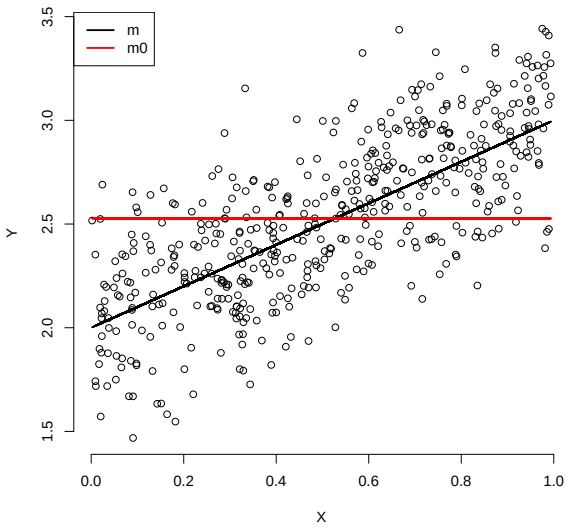

Figura 2.1: Datos simulados.

Teniendo en cuenta el Teorema 2.1, podemos plantear la hipótesis nula H_0 de forma equivalente como:

$$H_0^{'} : E(\varepsilon_0) = 0 \qquad \text{y} \qquad Var(\varepsilon_0) = 1. \tag{2.6}$$

Veamos un ejemplo. Consideramos un modelo de regresión lineal y contrastaremos si la función de regresión proviene de una familia parámetrica formada por las funciones constantes. En caso de provenir de dicha familia, la media y la varianza de los residuos deberían coincidir con la media y la varianza de los residuos bajo la hipótesis nula.

Sea $X \sim U\,[0,1]$ y $\varepsilon \sim N(0,1)$. Se define el modelo de regresión $Y = 2 + X + 0.25\varepsilon$, y se quiere contrastar la hipótesis nula $H_0 : m$ es constante, es decir, $H_0 : m = \theta_0$. En la Figura 2.1 se muestra el gráfico de dispersión de un conjunto de datos simulado según el modelo propuesto.

Trabajando a nivel poblacional, el valor de θ que minimiza $E[(Y-\theta)^2]$ es $\tilde{\theta} = E(Y)$. Entonces los residuos bajo la hipótesis nula vienen dados por

$$\varepsilon_0 = \frac{Y - m_{\tilde{\theta}}(X)}{\sigma(X)} = \frac{Y - E(Y)}{0.25}.$$

Seguidamente teniendo en cuenta que $E(Y) = 2 + E(X)$ se obtiene que la media de

ε_0 es

$$\begin{aligned} E(\varepsilon_0) &= E(\varepsilon) + E\left(\frac{X-\overline{X}}{0.25}\right) = E(\varepsilon) + E\left(4X - 4\overline{X}\right) = E(\varepsilon) + 4E(X) - 2 \\ &= E(\varepsilon) = 0, \end{aligned}$$

lo cual está en consonancia con H_0. Sin embargo, la varianza de ε_0 es (téngase en cuenta que ε y X son independientes)

$$\begin{aligned} Var(\varepsilon_0) &= Var(\varepsilon) + Var\left(\frac{2+X-E(Y)}{0.25}\right) = 1 + Var\left(\frac{X-0.5}{0.25}\right) \\ &= 1 + \frac{1}{0.25^2}\frac{1}{12} = 1 + \frac{16}{12} = \frac{7}{3}, \end{aligned}$$

lo cual contradice la hipótesis nula. Como se puede observar, a pesar que la media de ε_0 es cero, su varianza no es uno, por lo tanto la función de regresión no proviene de un modelo de regresión constante, como ya sabíamos. En situaciones generales también puede producirse un cambio en $E(\varepsilon_0)$.

2.2. Contraste de un modelo paramétrico basado en el momento de segundo orden

No siempre es necesario contrastar ambas condiciones sobre la media y la varianza de los errores dadas en (2.6). En el caso de estar ante un modelo de regresión paramétrico homocedástico con término aditivo constante, la media de los residuos bajo la hipótesis nula, ε_0, es cero. Entonces nos basta con analizar su varianza.

Probémoslo para el siguiente caso particular. Consideramos un modelo lineal para m de forma que la hipótesis nula es

$$H_0 : m(x) = a + bx.$$

El parámetro $\theta = (a, b)^t$ es estimado mediante mínimos cuadrados, que en su versión poblacional viene dado por

$$\tilde{a} = E(Y) - \frac{Cov(X,Y)}{Var(X)}E(X) \qquad \text{y} \qquad \tilde{b} = \frac{Cov(X,Y)}{Var(X)}.$$

Seguidamente se obtienen los residuos bajo la hipótesis nula:

$$\varepsilon_0 = \frac{Y - m_{\tilde{\theta}}(X)}{\sigma(X)} = \frac{Y - \tilde{a} - \tilde{b}X}{\sigma(X)}.$$

Calculando $E(\varepsilon_0)$ se tiene

$$\begin{aligned} E(\varepsilon_0) &= E\left(\frac{Y - \tilde{a} - \tilde{b}X}{\sigma(X)}\right) \\ &= E\left\{E\left(\frac{Y - E(Y) + \frac{Cov(X,Y)}{Var(X)}E(X) - \frac{Cov(X,Y)}{Var(X)}X}{\sigma(X)} \mid X = x\right)\right\} \\ &= E\left\{\frac{1}{\sigma(X)}\left(E\left(Y \mid X = x\right) - E\left(Y\right) + \frac{Cov(X,Y)}{Var(X)}E(X) - \frac{Cov(X,Y)}{Var(X)}X\right)\right\} \\ &= E\left[\frac{m(X)}{\sigma(X)}\right] - \frac{E(Y)}{E(\sigma(X))} + \frac{Cov(X,Y)}{Var(X)}\frac{E(X)}{E(\sigma(X))} - \frac{Cov(X,Y)}{Var(X)}E\left(\frac{X}{\sigma(X)}\right). \end{aligned}$$

Entonces si $\sigma(X) = \sigma$, entonces $E(\varepsilon_0) = 0$ ya que $E(Y) = E[m(X)]$.

Bajo estas condiciones (homocedasticidad y término aditivo constante en el modelo paramétrico), el contraste se reduciría a comprobar la hipótesis nula siguiente:

$$H_0 : Var(\varepsilon_0) = 1,$$

frente a la alternativa

$$H_1 : Var(\varepsilon_0) > 1.$$

Nótese que según la expresión (2.5) la varianza de ε_0 es siempre mayor o igual que la varianza de ε, es decir, mayor o igual que 1, lo cual justifica la elección de la alternativa unilateral.

El procedimiento a seguir para realizar este contraste consta de los siguientes pasos:

1. Se considera una muestra aleatoria, $(X_1, Y_1), \ldots, (X_n, Y_n)$, del modelo de regresión (1.1) con $\sigma(X) = \sigma$.

2. Se estima el paramétro θ por mínimos cuadrados, obteniendo $\tilde{\theta}$.

3. Se estima la varianza (constante), σ^2. Dado el modelo de regresión homocedástico, en Rice (1984) se propone un método para la estimación de la varianza

utilizando diferencias, que tienen como objetivo eliminar la tendencia en la función de regresión m, una idea originaria del análisis de series temporales. Tal método no requiere un estimador de la función de regresión. Asumiendo que X es univariante y $X_1 \leq \ldots \leq X_n$, el estimador propuesto para la varianza viene dado por:

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2. \tag{2.7}$$

4. Se construyen los residuos bajo la hipótesis nula H_0:

$$\widehat{\varepsilon}_{0i} = \frac{Y_i - m_{\tilde{\theta}}(X_i)}{\widehat{\sigma}}. \tag{2.8}$$

5. Se construye el estadístico de contraste para la varianza:

$$D = (n-1)S^2_{\varepsilon_0},$$

siendo, $S^2_{\varepsilon_0} = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (\widehat{\varepsilon}_{0i} - \overline{\varepsilon}_0)^2$.

Para llevar a la práctica el contraste necesitamos estudiar la distribución del estadístico de contraste bajo la hipótesis nula, lo cual puede hacerse mediante el estudio teórico de su distribución exacta o asintótica o bien mediante técnicas de remuestreo. El análisis teórico sobrepasa los objetivos de este trabajo. Como alternativa, procederemos mediante el uso de técnicas bootstrap. El proceso finaliza por lo tanto con los siguientes pasos:

6. Se calcula mediante bootstrap el punto crítico para el estadístico D a un nivel de significación α.

7. Finalmente se compara el valor del estadístico de contraste obtenido en la muestra con el punto crítico. Si el estadístico es mayor que el punto crítico se rechaza H_0, mientras que si es menor se acepta.

En la sección 2.4. se especificarán los detalles del algoritmo bootstrap empleado en el punto 6 para la obtención de los valores críticos.

2.3. Contraste de un modelo paramétrico basado en los momentos de primer y segundo orden

Cuando nos encontramos con un modelo de regresión parámetrico general no tenemos garantía que la media de los residuos construidos bajo la hipótesis nula sea cero. En este caso debemos contrastar las dos condiciones establecidas en (2.6).

Consideremos el caso particular del siguiente modelo lineal $Y = a + bX + \varepsilon$, del cual se obtiene

$$E(Y) = a + bE(X) \qquad \text{y} \qquad Var(Y) = b^2 Var(X).$$

Supongamos que se pretende contrastar la hipótesis nula $H_0 : m(x) = \theta x$. Para ello se busca el estimador de θ a través del problema de minimización

$$\tilde{\theta} \;=\; \arg\min_{\theta} E\left[(Y - m_\theta(X))^2\right].$$

Si desarrollamos la función que tenemos que minimizar obtenemos

$$\begin{aligned} E\left[(Y - m_\theta(X))^2\right] &= E\left[(Y - \theta X)^2\right] \\ &= E\left[Y^2\right] - 2\theta E\left[YX\right] + \theta^2 E\left(X^2\right). \end{aligned}$$

Esta expresión es una parábola en θ, cuyo mínimo está en

$$\tilde{\theta} = \frac{E((YX)}{E(X^2)}.$$

Teniendo en cuenta que $Y = a + bX + \varepsilon$, entonces

$$E(YX) = aE(X) + bE(X^2) + E(\varepsilon X) = aE(X) + bE(X^2),$$

dado que ε y X son independientes. Por lo tanto,

$$\tilde{\theta} = a\frac{E(X)}{E(X^2)} + b.$$

Entonces

$$\varepsilon_0 = Y - m_{\tilde{\theta}}(X) = Y - \tilde{\theta}X = Y - \frac{E((a + bX)X)}{E(X^2)}X,$$

y por lo tanto si calculamos su esperanza, se tiene

$$\begin{aligned} E(\varepsilon_0) &= E(Y) - \frac{E((a+bX)X)}{E(X^2)}E(X) = a + bE(X) - \frac{aE(X) + bE(X^2)}{E(X^2)}E(X) \\ &= a - a\frac{[E(X)]^2}{E(X^2)} = a\left(1 - \frac{[E(X)]^2}{E(X^2)}\right). \end{aligned}$$

En conclusión, se sigue que $E(\varepsilon_0) \neq 0$ siempre que a sea distinto de cero, ya que $E(X^2) > [E(X)]^2$ y por lo tanto $[E(X)]^2 / E(X^2) < 1$.

También se puede calcular la varianza de ε_0:

$$\begin{aligned} Var(\varepsilon_0) &= Var(Y - \tilde{\theta}X) = Var(a + bX + \varepsilon - \tilde{\theta}X) = Var((b - \tilde{\theta})X) + Var(\varepsilon) \\ &= \left(b - \tilde{\theta}\right)^2 Var(X) + 1 = \left(b - b - a\frac{E(X)}{E(X^2)}\right) Var(X) + 1 \\ &= \left(a\frac{E(X)}{E(X^2)}\right)^2 Var(X) + 1, \end{aligned}$$

que resulta ser mayor que 1, excepto si $a = 0$, lo cual está excluido del modelo propuesto.

En esta situación el contraste consiste en probar las dos condiciones descritas en la hipótesis nula (2.6). Con un nivel de significación α, el contraste se realiza en **dos etapas**:

1. **Etapa 1**. Se contrasta $H_0^{'(1)} : E(\varepsilon_0) = 0$ frente a $H_1^{'(1)} : E(\varepsilon_0) \neq 0$ con un nivel de significación α_1. Si se rechaza $H_0^{'(1)}$ esto implica que rechazamos H_0'. En caso contrario pasamos a la etapa siguiente.

2. **Etapa 2**. Se contrasta $H_0^{'(2)} : Var(\varepsilon_0) = 1$ frente a $H_1^{'(2)} : Var(\varepsilon_0) > 1$ con un nivel de significación α_2. Si se rechaza $H_0^{'(2)}$ entonces rechazamos H_0'. En cambio, si se acepta $H_0^{'(2)}$, aceptamos H_0'.

Los niveles de significación α_1 y α_2 se obtienen una vez fijado un nivel de significación global α para el contraste en dos etapas. En Qiu and Sheng (2008) se propone un procedimiento en dos etapas para contrastar la igualdad de dos funciones de riesgo en un contexto de análisis de supervivencia. Se empleará en este trabajo un proceso análogo para la obtención de los niveles de significación.

Por definición, bajo la hipótesis nula, se tiene

$$\begin{aligned}\alpha &= P(\textit{rechazar } H_0' \mid H_0' \textit{ es cierta}) \\ &= P(\textit{rechazar } H_0'^{(1)} \mid H_0' \textit{ cierta}) + P(\textit{rechazar } H_0'^{(2)} \cap \textit{no rechazar } H_0'^{(1)} \mid H_0' \textit{ cierta}),\end{aligned}$$

de donde, suponiendo independencia entre las dos etapas del contraste, se sigue que

$$\alpha = \alpha_1 + P(\textit{rechazar } H_0'^{(2)} \mid H_0' \textit{ es cierta})(1-\alpha_1) = \alpha_1 + \alpha_2(1-\alpha_1). \tag{2.9}$$

Por tanto, una vez controlado este punto, para que el contraste en dos etapas tenga un nivel de significación global α, α_1 y α_2 deben satisfacer la ecuación (2.9), es decir, para un α_1 dado, α_2 se debería tomar como $\alpha_2 = \dfrac{\alpha - \alpha_1}{1-\alpha_1}$.

Si no se tiene información previa sobre el modelo, entonces se puede dejar $\alpha_1 = \alpha_2$, y optar por

$$\alpha_1 = \alpha_2 = 1 - \sqrt{1-\alpha}.$$

Así, α_1 y α_2 determinan la región de rechazo del procedimiento en dos etapas. Sin embargo, el p-valor del contraste no está definido todavía. Generalmente, el p-valor de un contraste en dos etapas se puede construir de varias maneras. Para el procedimiento propuesto, ya que los contrastes en sus dos etapas individuales son independientes, una definición adecuada de p-valor es

$$\text{p-valor} = \begin{cases} p_1 & \text{si } p_1 \leq \alpha_1 \\ \\ \alpha_1 + p_2(1-\alpha_1) & \text{en otro caso} \end{cases}$$

donde p_1 y p_2 denotan los p-valores en la etapa 1 y etapa 2, respectivamente, y α_1 es el nivel de significación de la primera etapa. Obsérvese que el contraste rechazará la hipótesis nula H_0' cuando el p-valor global sea menor que α, ya que

(a) o bien el contraste en la primera etapa rechaza la hipótesis nula $H_0'^{(1)}$, es decir, $p_1 \leq \alpha_1$;

(b) o bien el contraste en la primera etapa no rechaza la hipótesis nula $H_0'^{(1)}$, pero el contraste en la segunda etapa sí rechaza $H_0'^{(2)}$, es decir, $p_1 > \alpha_1$ y $p_2 \leq \alpha_2$.

El procedimiento de contraste consta de los siguientes pasos:

1. Se considera una muestra aleatoria, $(X_1, Y_1), \ldots, (X_n, Y_n)$, del modelo de regresión (2.1).

2. Se estima el paramétro θ por mínimos cuadrados, obteniendo $\tilde{\theta}$.

3. Se estima la varianza condicional. Si estamos ante un modelo heterocedástico, dicha estimación se hace mediante el estimador no paramétrico de Nadaraya-Watson:
$$\widehat{\sigma}^2(x) = \sum_{i=1}^{n} W_{nj}(x) Y_i^2 - \widehat{m}^2(x),$$
donde $\widehat{m}$ es el estimador de Nadaraya-Watson de m. Obviamente, también se podrían emplear otros estimadores no paramétricos. Ahora bien, si nuestro modelo es homocedástico, la estimación de σ^2 se puede realizar empleando la propuesta de Rice (1984) dada en (2.7).

4. Se construyen los residuos bajo la hipótesis nula H_0, dados por la expresión (2.8).

5. Se construyen los estadísticos de contraste para la media y la varianza, respectivamente:
$$D_1 = \sqrt{n}\frac{\overline{\varepsilon}_0}{S_{\varepsilon_0}} \qquad \text{y} \qquad D_2 = (n-1)S_{\varepsilon_0}^2,$$
siendo, $\overline{\varepsilon}_0 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \widehat{\varepsilon}_{0i}$ y $S_{\varepsilon_0}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} (\widehat{\varepsilon}_{0i} - \overline{\varepsilon}_0)^2$.

Como ya mencionamos en la sección anterior, la distribución de los estadísticos de contraste se obtendrá aplicando la metodología bootstrap, ya que la teoría asintótica sobrepasa los objetivos de este trabajo. Se procede entonces del siguiente modo:

6. Se emplea la metodología bootstrap para calcular los puntos críticos para D_1 y D_2 a niveles de significación α_1 y α_2, respectivamente.

7. Finalmente se comparan los valores de los estadísticos de contraste con los puntos críticos obtenidos para cada estadístico. Si se rechaza alguno de los dos contrastes para la media y la varianza, se rechazará entonces la hipótesis nula H_0' del contraste en dos etapas (2.6). Si por el contrario se aceptan ambos contrastes, se acepta H_0'.

2.4. Aproximaciones bootstrap

Para aplicar estos procedimientos de contraste utilizamos metodología bootstrap para aproximar las distribuciones de los estadísticos de contraste bajo la hipótesis nula.

En primer lugar, para $i = 1, \ldots, n$ se estiman los residuos de forma no paramétrica, obteniendo

$$\widehat{\varepsilon}_{1i} = \frac{Y_i - \widehat{m}(X_i)}{\widehat{\sigma}(X_i)}, \tag{2.10}$$

siendo $\widehat{m}$ y $\widehat{\sigma}$ los estimadores de Nadaraya-Watson de m y σ, respectivamente. Estos nuevos residuos son estandarizados a media cero y varianza uno.

En este trabajo, se proponen dos tipos de bootstrap: un bootstrap uniforme de residuos y un bootstrap suavizado de residuos.

El algoritmo de remuestreo se describe como sigue. Sea b el índice de las ejecuciones bootstrap, $b = 1, \ldots, B$. Entonces:

1. Sean $\left\{\varepsilon^*_{i,b} = (1-a^2)^{1/2} V_i + aZ_i,\ i = 1, \ldots, n\right\}$ los residuos bootstrap, donde V_i es una muestra i.i.d. de $\tilde{F}_\varepsilon$ (función de distribución empírica de los residuos (2.10) estandarizados) y Z_i es una muestra i.i.d. de una variable de media cero y desviación típica uno (por ejemplo, una normal estándar).

   Con $a = 0$ se obtiene el bootstrap de residuos **uniforme**, mientras que con $a > 0$ se obtiene el bootstrap de residuos **suavizado**. La constante a, determina la cantidad de suavizado en el bootstrap.

2. Para cada $i = 1, \ldots, n$, se contruyen las variables respuesta bajo la hipótesis nula

   $$Y^*_{i,b} = m_{\tilde{\theta}}(X_i) + \widehat{\sigma}(X_i)\varepsilon^*_{i,b}.$$

3. Sea D^*_b en el contraste basado en el momento de orden dos, o bien $D^*_{1,b}$ y $D^*_{2,b}$ en el contraste basado en los momentos de orden uno y dos, los estadísticos de contraste obtenidos a partir de las muestras bootstrap $\left\{(X_i, Y^*_{i,b}),\ i = 1, \ldots, n\right\}$.

Dado que en el paso 2 las remuestras bootstrap son construidas bajo la hipótesis nula de que la función de regresión pertenece a la familia paramétrica de funciones

especificada en H_0, este procedimiento aproxima la distribución del estadístico de contraste bajo la hipótesis nula. Si $D^*_{(b)}$ es el estadístico ordenado de los valores $D^*_1, \ldots, D^*_B$ obtenidos en el paso 3, y análogamente para $D^*_{1,(b)}$ y $D^*_{2,(b)}$, entonces $D^*_{([\rho B])}$, $D^*_{1,([\rho B])}$ y $D^*_{2,([\rho B])}$ aproximan los cuantiles de orden ρ de la distribución de D, D_1 y D_2 bajo la hipótesis nula, respectivamente.

Tras la aproximación de los cuantiles, los procedimientos de contraste son los siguientes:

- Si el contraste está basado únicamente en el momento de segundo orden (varianza), se rechazará $H_0^{'}$ si $d > D^*_{([(1-\alpha)B])}$, donde d es el valor observado del estadístico D en la muestra original.

- Si el contraste esta basado en los momentos de primer y segundo orden (media y varianza), se rechazará $H_0^{'(1)}$ si $d_1 < D^*_{1,([(\alpha/2)B])}$ o bien $d_1 > D^*_{1,([(1-\alpha/2)B])}$, donde d_1 es el valor observado del estadístico D_1 en la muestra original. Por otra parte, se rechazará $H_0^{'(2)}$ si $d_2 > D^*_{2,([(1-\alpha)B])}$, donde d_2 es el valor observado del estadístco D_2.

2.5. Estudio de simulación

En esta sección se llevan a cabo varios estudios de simulación para mostrar el comportamiento práctico de los métodos de contraste propuestos en las secciones anteriores. En uno de los estudios se harán comparaciones con los resultados obtenidos mediante otros estadísticos de contraste propuestos en la literatura. En particular, se compara con el procedimiento desarrollado por Van Keilegom et al. (2008) que se basa en la comparación de las distribuciones de los residuos.

Las simulaciones se realizan mediante la selección de las siguientes funciones de regresión:

$$
\begin{array}{llll}
(i) & m(x) = 1 + x & (iv) & m(x) = x + 0.5xe^x \\
(ii) & m(x) = x & (v) & m(x) = x + 0.3sen(4\pi x) \\
(iii) & m(x) = x + x^2 & &
\end{array}
$$

Los modelo (i) y (ii) se corresponden con las hipótesis nulas que se presentarán más tarde, mientras que los modelos (iii), (iv) y (v) son los utilizados como hipótesis alternativas.

En las distintas simulaciones, se consideran situaciones homocedásticas y heterocedásticas. En el caso homocedástico, las funciones de escala son:

$$(a) \quad \sigma(x) = 0.2 \qquad \text{y} \qquad (b) \quad \sigma(x) = 0.3,$$

y el caso heterocedástico se toma

$$(c) \quad \sigma(x) = 0.2 + 0.2x.$$

La distribución de ε es una distribución normal estándar, y la covariable X está distribuida uniformemente en el intervalo $[0, 1]$. Para la estimación no paramétrica de las curvas de regresión y de la varianza condicional se empleará el estimador de Nadaraya-Watson con el núcleo de Epanechnikov.

Las tablas recogen la proporción de rechazos en 1000 realizaciones del contraste para tamaños muestrales $n = 100$ y $n = 200$ basados en $B = 200$ repeticiones bootstrap. Los niveles de significación son $\alpha = 0.05$ y $\alpha = 0.10$. En el caso del contraste en dos etapas se opta por $\alpha_1 = \alpha_2 = 1 - \sqrt{1 - \alpha}$.

Como ancho de banda h requerido para la estimación no paramétrica, se considera un conjunto de valores fijos que se especificarán en cada tabla. Sobre la cantidad de suavizado que se aplica en el bootstrap suavizado, se eligen para a los valores 0.05 y 0.10, aunque en las tablas solo se mostrarán los resultados obtenidos para $a = 0.05$ ya que los resultados obtenidos con $a = 0.10$ son muy similares. Recordemos que el caso $a = 0$ se corresponde con el bootstrap uniforme.

En primer lugar tomamos como hipótesis nula

$$H_0 : m(x) = a + bx.$$

En la Tabla 2.1 se muestran los resultados del contraste basado en el momento de segundo orden para los modelos (i), (iii) y (iv) con las funciones de escala homocedásticas (a) y (b). Por otra parte, en la Tabla 2.2 se muestran los resultados del contraste basado en los momentos de primer y segundo orden para para los mismos modelos con función de escala homocedastica (b) y heterocedástica (c).

Tabla 2.1: Probabilidades de rechazo bajo los modelos (i), (iii) y (iv) del contraste basado en el momento de segundo orden. Los modelos son homodedásticos, con varianzas (a) y (b).

				$(a)\ \sigma = 0.2$				$(b)\ \sigma = 0.3$			
				$a = 0$		$a = 0.05$		$a = 0$		$a = 0.05$	
m	n	h	$\alpha:$	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(i)	100	0.10		0.048	0.087	0.050	0.089	0.044	0.100	0.043	0.098
		0.15		0.043	0.088	0.042	0.086	0.045	0.096	0.045	0.097
		0.20		0.046	0.090	0.044	0.089	0.042	0.091	0.043	0.092
	200	0.10		0.063	0.121	0.066	0.122	0.066	0.121	0.063	0.116
		0.15		0.058	0.127	0.058	0.130	0.062	0.121	0.061	0.118
		0.20		0.067	0.117	0.064	0.120	0.064	0.116	0.062	0.116
(iii)	100	0.10		0.341	0.461	0.340	0.462	0.163	0.256	0.164	0.256
		0.15		0.350	0.467	0.345	0.463	0.170	0.255	0.170	0.252
		0.20		0.345	0.469	0.347	0.469	0.169	0.266	0.168	0.263
	200	0.10		0.562	0.681	0.563	0.677	0.232	0.348	0.232	0.348
		0.15		0.559	0.687	0.557	0.686	0.234	0.359	0.234	0.352
		0.20		0.563	0.687	0.563	0.684	0.240	0.348	0.244	0.346
(iv)	100	0.10		0.411	0.553	0.409	0.550	0.174	0.282	0.176	0.277
		0.15		0.405	0.552	0.405	0.551	0.183	0.271	0.182	0.267
		0.20		0.426	0.542	0.421	0.538	0.176	0.267	0.175	0.264
	200	0.10		0.647	0.760	0.647	0.758	0.252	0.375	0.258	0.371
		0.15		0.657	0.755	0.656	0.753	0.255	0.373	0.252	0.374
		0.20		0.651	0.762	0.653	0.764	0.257	0.373	0.254	0.367

La Tabla 2.1 muestra una buena aproximación de los niveles de significación en todos los casos (modelo (i)), sin embargo la potencia no es demasiado elevada. Los mayores valores se alcanzan en el modelo (iv) para el caso de menor varianza. El parámetro ventana h no representa un gran impacto en la aproximación del nivel o en la potencia.

En la Tabla 2.2 se observan que la aproximación del nivel depende en gran medida del parámetro h. El contraste aproxima mejor dichos niveles de significación para valores pequeños del parámetro ventana. Con respecto a la potencia, se obtienen valores aceptables de la misma.

Tabla 2.2: Probabilidades de rechazo bajo los modelos (i), (iii) y (iv) del contraste basado en los momentos de primer y segundo orden. Se considera para la varianza el modelo homocedástico (b) y el modelo heterocedástico (c).

				$(b)\ \sigma = 0.3$				$(c)\ \sigma = 0.2 + 0.2x$			
				$a = 0$		$a = 0.05$		$a = 0$		$a = 0.05$	
m	n	h	$\alpha:$	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(i)	100	0.10		0.042	0.086	0.042	0.088	0.044	0.085	0.042	0.084
		0.15		0.032	0.074	0.032	0.074	0.044	0.080	0.043	0.081
		0.20		0.029	0.066	0.025	0.068	0.041	0.076	0.043	0.080
	200	0.10		0.055	0.097	0.051	0.096	0.042	0.081	0.045	0.082
		0.15		0.049	0.086	0.047	0.086	0.037	0.074	0.037	0.074
		0.20		0.035	0.064	0.036	0.067	0.026	0.051	0.027	0.048
(iii)	100	0.10		0.335	0.468	0.333	0.470	0.431	0.546	0.434	0.544
		0.15		0.317	0.417	0.314	0.421	0.448	0.553	0.452	0.554
		0.20		0.230	0.323	0.230	0.322	0.349	0.446	0.347	0.450
	200	0.10		0.644	0.740	0.646	0.741	0.784	0.855	0.785	0.852
		0.15		0.478	0.568	0.481	0.568	0.716	0.792	0.719	0.790
		0.20		0.336	0.434	0.331	0.429	0.618	0.721	0.619	0.714
(iv)	100	0.10		0.370	0.479	0.374	0.479	0.472	0.595	0.472	0.601
		0.15		0.264	0.375	0.266	0.378	0.408	0.513	0.410	0.514
		0.20		0.178	0.249	0.176	0.247	0.285	0.395	0.287	0.393
	200	0.10		0.646	0.738	0.645	0.739	0.772	0.850	0.776	0.849
		0.15		0.393	0.484	0.393	0.486	0.623	0.721	0.620	0.724
		0.20		0.297	0.402	0.291	0.409	0.543	0.663	0.548	0.657

En la Tabla 2.1 y en la Tabla 2.2 los modelos coinciden en el caso de la varianza (b). En los resultados de la Tabla 2.1 la varianza se obtiene mediante el estimador de Rice, mientras que en los resultados de la Tabla 2.2 se obtiene mediante el estimador de Nadaraya-Watson. La diferencia está en el método de contraste. Comparando ambos métodos observamos mejores aproximaciones del nivel de significación con el primero de ellos, mientras que la potencia es mayor para el segundo. No obstante, para el segundo contraste la influencia del parámetro ventana se muestra muy relevante.

Finalmente se realiza una comparativa entre el método basado en los momentos de primer y segundo orden con el método propuesto por Van Keilegom et al. (2008),

cuyos estadísticos de contraste T_{KS} y T_{CM} están dados en las ecuaciones (1.4) y (1.5). En este caso se contrasta la hipótesis nula

$$H_0 : m(x) = \theta x.$$

Las tablas 2.3, 2.4, 2.5 y 2.6 muestran los resultados para los modelos (ii), (iii), (iv) y (v), respectivamente. El modelo (ii) se corresponde con la hipótesis nula y los modelos (iii), (iv) y (v) con la hipótesis alternativa.

Los resultados de la Tabla 2.3 muestran una buena aproximación al nivel de significación para el contraste basado en los momentos y el contraste de tipo Cramér-von Mises, mientras que el contraste de tipo Kolmogorov-Smirnov proporciona aproximaciones regulares, principalmente para el caso $n = 100$. Con respecto a la potencia (Tablas 2.4, 2.5 y 2.6) se observa que el contraste de tipo Kolmogorov-Smirnov obtiene en general peores resultados. En los modelos (iii) y (iv) el contraste basado en los momentos y el contraste de tipo Cramér-von Mises son similares, mientras que en el modelo (v) los resultados del contraste basado en los momentos son considerablemente mejores. Nótese además que el contraste basado en los momentos es más robusto a h, especialmente en el modelo (v) y en el caso heterocedástico para los modelos (iii) y (iv).

Tabla 2.3: Probabilidades de rechazo bajo el modelo (ii) del contraste basado en los momentos de primer y segundo orden (mom.) y en los estadísticos T_{KS} y T_{CM}. Se considera para la varianza el modelo homocedástico (b) y el modelo heterocedástico (c).

					mom. $(a=0)$		mom. $(a=0.05)$		T_{KS} $(a=0)$		T_{KS} $(a=0.05)$		T_{CM} $(a=0)$		T_{CM} $(a=0.05)$	
m	σ	n	h	α:	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(ii)	(b)	100	0.10		0.061	0.111	0.060	0.112	0.020	0.069	0.021	0.077	0.036	0.077	0.038	0.077
			0.15		0.059	0.102	0.057	0.104	0.021	0.074	0.030	0.078	0.033	0.079	0.035	0.085
			0.20		0.051	0.098	0.051	0.098	0.019	0.066	0.022	0.075	0.032	0.066	0.035	0.073
			0.25		0.059	0.096	0.058	0.095	0.020	0.068	0.024	0.079	0.032	0.074	0.033	0.081
			0.30		0.048	0.089	0.051	0.091	0.023	0.059	0.028	0.073	0.032	0.062	0.035	0.071
			0.35		0.049	0.090	0.050	0.088	0.017	0.059	0.028	0.074	0.030	0.069	0.033	0.071
		200	0.10		0.073	0.144	0.074	0.141	0.041	0.084	0.045	0.090	0.049	0.087	0.052	0.089
			0.15		0.074	0.127	0.076	0.124	0.037	0.076	0.046	0.097	0.045	0.083	0.047	0.086
			0.20		0.068	0.111	0.066	0.111	0.031	0.082	0.047	0.104	0.044	0.085	0.047	0.093
			0.25		0.051	0.099	0.051	0.097	0.031	0.073	0.041	0.088	0.041	0.088	0.043	0.098
			0.30		0.063	0.094	0.060	0.094	0.035	0.078	0.041	0.088	0.044	0.083	0.046	0.091
			0.35		0.050	0.083	0.048	0.083	0.035	0.080	0.046	0.097	0.040	0.079	0.047	0.089
	(c)	100	0.10		0.066	0.126	0.066	0.128	0.037	0.081	0.040	0.084	0.038	0.086	0.040	0.084
			0.15		0.068	0.115	0.064	0.113	0.022	0.081	0.025	0.094	0.039	0.084	0.044	0.085
			0.20		0.054	0.112	0.056	0.112	0.017	0.085	0.022	0.098	0.039	0.083	0.039	0.086
			0.25		0.060	0.098	0.058	0.100	0.025	0.090	0.032	0.102	0.038	0.076	0.042	0.082
			0.30		0.057	0.100	0.059	0.098	0.031	0.094	0.037	0.108	0.047	0.088	0.052	0.089
			0.35		0.040	0.082	0.039	0.082	0.035	0.101	0.045	0.115	0.052	0.091	0.056	0.103
		200	0.10		0.074	0.125	0.071	0.125	0.037	0.075	0.042	0.082	0.040	0.083	0.040	0.088
			0.15		0.062	0.119	0.057	0.118	0.026	0.058	0.032	0.072	0.040	0.084	0.045	0.088
			0.20		0.056	0.093	0.057	0.097	0.036	0.072	0.044	0.099	0.044	0.090	0.048	0.095
			0.25		0.044	0.078	0.043	0.080	0.040	0.089	0.049	0.107	0.054	0.087	0.052	0.090
			0.30		0.048	0.081	0.048	0.080	0.049	0.095	0.064	0.111	0.054	0.097	0.059	0.102
			0.35		0.043	0.067	0.044	0.066	0.057	0.109	0.068	0.129	0.064	0.122	0.070	0.127

Tabla 2.4: Probabilidades de rechazo bajo el modelo (iii) del contraste basado en los momentos de primer y segundo orden (mom.) y en los estadísticos T_{KS} y T_{CM}. Se considera para la varianza el modelo homocedástico (b) y el modelo heterocedástico (c).

					mom. $(a=0)$		mom. $(a=0.05)$		T_{KS} $(a=0)$		T_{KS} $(a=0.05)$		T_{CM} $(a=0)$		T_{CM} $(a=0.05)$	
m	σ	n	h	α:	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(iii)	(b)	100	0.10		0.762	0.843	0.757	0.845	0.545	0.719	0.559	0.723	0.739	0.830	0.747	0.835
			0.15		0.728	0.824	0.733	0.824	0.544	0.719	0.550	0.741	0.768	0.850	0.774	0.861
			0.20		0.695	0.786	0.700	0.784	0.529	0.727	0.559	0.756	0.781	0.860	0.786	0.865
			0.25		0.688	0.769	0.686	0.765	0.503	0.695	0.521	0.726	0.783	0.867	0.786	0.869
			0.30		0.673	0.759	0.677	0.763	0.442	0.686	0.471	0.710	0.779	0.862	0.790	0.868
			0.35		0.680	0.756	0.671	0.758	0.405	0.634	0.429	0.674	0.754	0.847	0.765	0.859
		200	0.10		0.976	0.988	0.977	0.987	0.907	0.960	0.922	0.965	0.973	0.988	0.977	0.991
			0.15		0.957	0.978	0.955	0.977	0.919	0.960	0.932	0.971	0.980	0.993	0.982	0.993
			0.20		0.946	0.970	0.947	0.969	0.919	0.965	0.937	0.969	0.977	0.993	0.979	0.995
			0.25		0.938	0.957	0.941	0.959	0.888	0.950	0.910	0.959	0.972	0.987	0.976	0.988
			0.30		0.919	0.951	0.918	0.951	0.826	0.920	0.856	0.937	0.969	0.983	0.968	0.984
			0.35		0.922	0.953	0.921	0.948	0.778	0.891	0.809	0.904	0.953	0.973	0.955	0.976
	(c)	100	0.10		0.836	0.904	0.831	0.907	0.663	0.809	0.670	0.814	0.851	0.910	0.849	0.914
			0.15		0.845	0.915	0.850	0.914	0.637	0.808	0.655	0.820	0.853	0.921	0.853	0.923
			0.20		0.828	0.887	0.828	0.887	0.576	0.764	0.591	0.783	0.819	0.902	0.822	0.909
			0.25		0.840	0.887	0.836	0.885	0.497	0.693	0.516	0.719	0.779	0.867	0.785	0.874
			0.30		0.821	0.885	0.817	0.888	0.396	0.593	0.410	0.633	0.737	0.814	0.742	0.817
			0.35		0.807	0.874	0.808	0.880	0.312	0.488	0.331	0.513	0.629	0.755	0.637	0.764
		200	0.10		0.993	0.999	0.994	0.999	0.963	0.979	0.967	0.985	0.986	0.995	0.987	0.995
			0.15		0.989	0.992	0.989	0.992	0.943	0.976	0.953	0.979	0.985	0.992	0.985	0.995
			0.20		0.983	0.993	0.982	0.992	0.916	0.964	0.928	0.970	0.982	0.989	0.981	0.989
			0.25		0.979	0.992	0.979	0.992	0.844	0.932	0.870	0.940	0.966	0.980	0.971	0.980
			0.30		0.983	0.990	0.984	0.990	0.735	0.854	0.763	0.871	0.931	0.964	0.937	0.961
			0.35		0.979	0.986	0.978	0.987	0.585	0.734	0.620	0.754	0.840	0.920	0.847	0.919

Tabla 2.5: Probabilidades de rechazo bajo el modelo (iv) del contraste basado en los momentos de primer y segundo orden (mom.) y en los estadísticos T_{KS} y T_{CM}. Se considera para la varianza el modelo homocedástico (b) y el modelo heterocedástico (c).

					mom. $(a=0)$		mom. $(a=0.05)$		T_{KS} $(a=0)$		T_{KS} $(a=0.05)$		T_{CM} $(a=0)$		T_{CM} $(a=0.05)$	
m	σ	n	h	α:	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(iv)	(b)	100	0.10		0.697	0.789	0.688	0.785	0.474	0.644	0.479	0.658	0.685	0.793	0.687	0.800
			0.15		0.660	0.757	0.658	0.761	0.449	0.682	0.471	0.714	0.712	0.803	0.714	0.806
			0.20		0.616	0.714	0.611	0.710	0.421	0.650	0.450	0.678	0.703	0.808	0.704	0.815
			0.25		0.580	0.677	0.586	0.678	0.376	0.616	0.397	0.649	0.684	0.798	0.692	0.803
			0.30		0.572	0.677	0.572	0.670	0.345	0.561	0.361	0.595	0.666	0.786	0.677	0.796
			0.35		0.584	0.674	0.581	0.672	0.313	0.528	0.340	0.564	0.642	0.771	0.651	0.780
		200	0.10		0.964	0.978	0.961	0.977	0.868	0.928	0.876	0.938	0.957	0.977	0.956	0.978
			0.15		0.940	0.965	0.938	0.966	0.877	0.939	0.893	0.941	0.966	0.980	0.970	0.983
			0.20		0.907	0.945	0.910	0.946	0.844	0.926	0.867	0.943	0.958	0.977	0.955	0.980
			0.25		0.895	0.943	0.900	0.944	0.802	0.891	0.822	0.904	0.949	0.975	0.952	0.978
			0.30		0.888	0.929	0.887	0.935	0.720	0.856	0.747	0.866	0.930	0.971	0.936	0.971
			0.35		0.874	0.925	0.872	0.923	0.648	0.786	0.676	0.822	0.895	0.944	0.898	0.946
	(c)	100	0.10		0.768	0.844	0.771	0.844	0.560	0.718	0.570	0.731	0.758	0.839	0.756	0.845
			0.15		0.759	0.834	0.763	0.834	0.533	0.712	0.542	0.728	0.749	0.840	0.760	0.843
			0.20		0.731	0.820	0.729	0.820	0.460	0.656	0.479	0.679	0.718	0.808	0.728	0.817
			0.25		0.723	0.808	0.720	0.808	0.375	0.560	0.383	0.589	0.650	0.755	0.656	0.765
			0.30		0.714	0.800	0.711	0.796	0.268	0.452	0.295	0.493	0.552	0.674	0.562	0.689
			0.35		0.702	0.795	0.705	0.791	0.221	0.363	0.236	0.382	0.453	0.577	0.462	0.588
		200	0.10		0.986	0.992	0.985	0.992	0.922	0.961	0.935	0.965	0.978	0.991	0.978	0.992
			0.15		0.974	0.988	0.975	0.989	0.903	0.959	0.915	0.968	0.971	0.990	0.971	0.987
			0.20		0.965	0.988	0.966	0.989	0.828	0.915	0.850	0.931	0.951	0.974	0.950	0.976
			0.25		0.964	0.984	0.961	0.983	0.706	0.829	0.733	0.850	0.904	0.951	0.911	0.953
			0.30		0.958	0.977	0.956	0.975	0.531	0.698	0.562	0.733	0.799	0.879	0.799	0.881
			0.35		0.954	0.978	0.952	0.976	0.359	0.538	0.389	0.564	0.645	0.765	0.657	0.773

Tabla 2.6: Probabilidades de rechazo bajo el modelo (v) del contraste basado en los momentos de primer y segundo orden (mom.) y en los estadísticos T_{KS} y T_{CM}. Se considera para la varianza el modelo homocedástico (b) y el modelo heterocedástico (c).

					mom. $(a=0)$		mom. $(a=0.05)$		T_{KS} $(a=0)$		T_{KS} $(a=0.05)$		T_{CM} $(a=0)$		T_{CM} $(a=0.05)$	
m	σ	n	h	α:	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(v)	(b)	100	0.10		0.994	0.999	0.995	0.999	0.713	0.865	0.723	0.872	0.801	0.897	0.811	0.904
			0.15		0.992	0.996	0.992	0.996	0.675	0.856	0.692	0.864	0.770	0.869	0.782	0.879
			0.20		0.988	0.993	0.985	0.992	0.620	0.810	0.648	0.832	0.679	0.813	0.685	0.829
			0.25		0.974	0.986	0.974	0.985	0.514	0.729	0.562	0.767	0.564	0.709	0.581	0.727
			0.30		0.959	0.971	0.959	0.971	0.426	0.645	0.474	0.680	0.453	0.601	0.468	0.608
			0.35		0.941	0.966	0.941	0.966	0.321	0.527	0.359	0.581	0.380	0.514	0.401	0.519
		200	0.10		1.000	1.000	1.000	1.000	0.975	0.988	0.976	0.989	0.991	0.998	0.991	0.998
			0.15		1.000	1.000	1.000	1.000	0.972	0.986	0.981	0.991	0.983	0.995	0.982	0.995
			0.20		1.000	1.000	1.000	1.000	0.947	0.980	0.960	0.982	0.966	0.986	0.967	0.986
			0.25		1.000	1.000	1.000	1.000	0.889	0.960	0.910	0.974	0.910	0.965	0.920	0.967
			0.30		0.999	1.000	0.999	1.000	0.803	0.890	0.834	0.910	0.820	0.902	0.831	0.915
			0.35		0.998	0.999	0.998	0.999	0.690	0.823	0.723	0.846	0.756	0.839	0.767	0.855
	(c)	100	0.10		0.999	0.999	0.999	0.999	0.817	0.933	0.831	0.938	0.873	0.933	0.873	0.937
			0.15		0.996	1.000	0.996	1.000	0.774	0.914	0.784	0.922	0.855	0.928	0.866	0.935
			0.20		0.988	0.997	0.988	0.996	0.717	0.872	0.741	0.888	0.806	0.891	0.815	0.901
			0.25		0.975	0.989	0.979	0.990	0.628	0.833	0.652	0.855	0.719	0.832	0.727	0.839
			0.30		0.966	0.979	0.966	0.977	0.528	0.743	0.570	0.762	0.631	0.755	0.638	0.761
			0.35		0.954	0.973	0.953	0.974	0.473	0.657	0.498	0.679	0.572	0.676	0.589	0.681
		200	0.10		1.000	1.000	1.000	1.000	0.990	0.999	0.991	0.998	0.997	0.999	0.997	1.000
			0.15		1.000	1.000	1.000	1.000	0.989	0.996	0.991	0.997	0.995	0.999	0.995	0.999
			0.20		1.000	1.000	1.000	1.000	0.975	0.989	0.977	0.992	0.990	0.995	0.991	0.997
			0.25		1.000	1.000	1.000	1.000	0.939	0.979	0.949	0.981	0.964	0.986	0.967	0.986
			0.30		1.000	1.000	1.000	1.000	0.873	0.947	0.895	0.955	0.923	0.968	0.925	0.972
			0.35		0.998	1.000	0.998	1.000	0.795	0.898	0.821	0.922	0.891	0.941	0.896	0.944

Capítulo 3

Aplicaciones a datos reales

En este capítulo se considerarán tres aplicaciones a datos reales para ilustrar el funcionamiento del método de contraste en dos etapas propuesto en este trabajo. Aunque hemos descrito dos procedimientos de contraste, uno basado solo en el momento de segundo orden y otro basado en los momentos de primer y segundo orden, en estas aplicaciones a datos reales empleamos este último, ya que no se hace ninguna suposición sobre los datos.

Con respecto a la utilización de los métodos bootstrap, se muestran los resultados relativos al bootstrap suavizado con parámetro $a = 0.05$. También se realizaron los contrastes para $a = 0$ (bootstrap uniforme) y $a = 0.10$, obteniéndose resultados casi idénticos en los tres ejemplos.

3.1. Aplicación 1: Datos de diabetes

Consideramos un conjunto de datos sobre diabetes analizados en Faraggi (2003) (véase también González-Manteiga et al. (2011)). Los datos provienen de una encuesta sobre la incidencia de la diabetes en El Cairo (Egipto). Consisten en mediciones de la concentración de glucosa en sangre de 286 sujetos obtenidas a partir de una punción en el dedo. De acuerdo con los criterios de la Organización Mundial de la Salud (1985), se clasificaron 88 sujetos como enfermos y 198 como sanos. En esta sección trabajamos con el grupo de sujetos sanos.

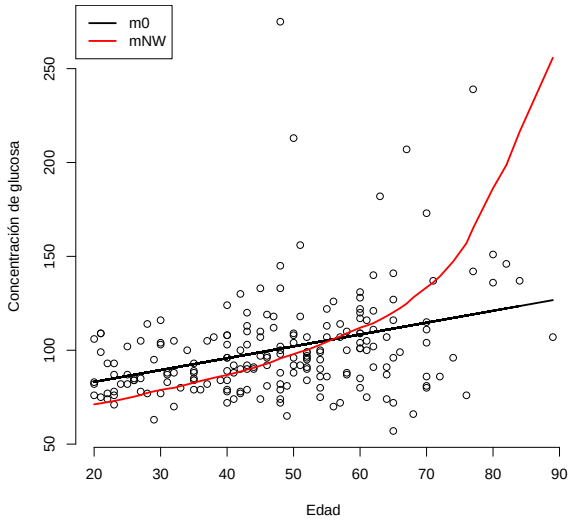

Figura 3.1: Diagrama de dispersión y curvas de regresión estimadas de Y =“concentración de glucosa” frente a X =“edad”.

La edad del sujeto resulta ser una covariable relevante en este ejemplo, ya que debido a razones médicas los niveles de glucosa tienden a ser mayores en las personas de más edad, incluso cuando no sufren diabetes. Los métodos desarrollados en Faraggi (2003) requieren que la relación entre la covariable “edad” y la variable respuesta “concentración de glucosa” sea lineal, es decir, de la forma $Y = \beta_0 + \beta_1 X + \varepsilon$. Se pretende entonces contrastar la hipótesis nula:

$$H_0 : m(x) = a + bx. \tag{3.1}$$

La Figura 3.1 muestra el diagrama de dispersión de los datos, así como la estimación no paramétrica de la función de regresión mediante Nadaraya-Watson, que denotamos por mNW, y la estimación de la función de regresión bajo la hipótesis nula, $m0$. En la estimación no paramétrica de m, se tomó como parámetro ventana $h = 7.89$, el óptimo obtenido por el método de validación cruzada.

La bondad de ajuste de la regresión lineal se comprobó mediante el contraste en dos etapas propuesto en este trabajo. Se utilizaron para ello diez valores del parámetro ventana seleccionados en torno al h óptimo obtenido por validación cruzada. Para cada valor de h se calcularon los estadísticos de contraste y los p-valores según

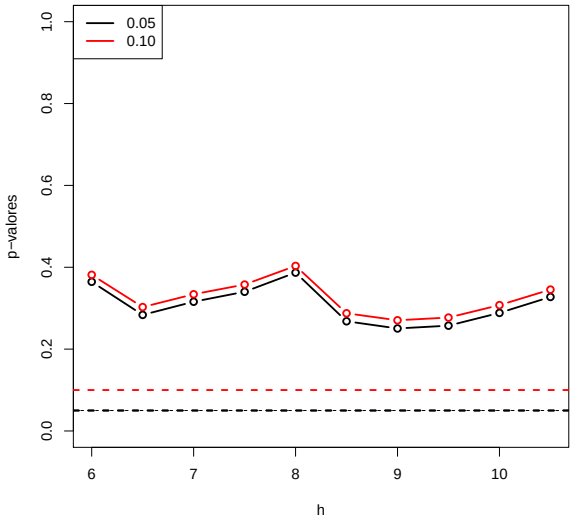

Figura 3.2: Gráfico de los p-valores en función de h para los datos sobre la concentración de glucosa (en negro para $\alpha = 0.05$ y en rojo para $\alpha = 0.10$). Las líneas horizontales discontinuas se corresponden a los niveles 0.05 y 0.10.

lo descrito en la sección 2.3 mediante aproximación bootstrap con 1000 repeticiones. Nótese que el p-valor para el contraste en dos etapas depende del nivel de significación α. Los p-valores para todos estos contrastes se muestran en la Figura 3.2.

Los valores obtenidos indican que un modelo lineal es apropiado para los datos, lo cual concuerda con la suposición de Faraggi (2003).

En la Figura 3.1 la estimación no paramétrica parece desviarse notablemente del modelo lineal en la parte derecha del soporte de la covariable. No obstante la escasa presencia de observaciones que parecen estar alterando dicha estimación no son relevantes en el modelo.

3.2. Aplicación 2: Datos de talla y peso de mejillones

Se considera el conjunto de datos de talla y peso del mejillón que se produce en bateas en un espacio marítimo delimitado de las Rías Gallegas. La especie que se cultiva en Galicia es el mejillón *Mytilus Galloprovinciales*. Concretamente los datos aquí utilizados se corresponden a la Ría de Arousa y fueron proporcionados para

este trabajo por el Instituto de Investigaciones Marinas de Vigo (IIM), perteneciente al Consejo Superior de Investigación Científica (CSIC). La muestra contiene 287 observaciones.

Generalmente la relación entre la talla y el peso de diversas especies de animales y plantas se suele ajustar por un modelo alométrico del tipo

$$Y = aX^b + \sigma(X)\varepsilon. \tag{3.2}$$

Cuando se relacionan talla (X) y peso (Y), si b (exponente alométrico) es 1, los individuos poseen similitud geométrica (o se dice que son isométricos). En este caso la relación entre las variables es lineal. Cualquier exponente mayor al esperado según la isometría se considera alometría positiva, es decir, hay un crecimiento desproporcionadamente alto del peso. Por otro lado si el exponente alométrico es menor que 1 decimos que hay una alometría negativa, es decir, hay un crecimiento desproporcionadamente bajo del peso. Para el caso del mejillón en la literatura se suele suponer cierto este modelo alométrico.

Comenzamos por contrastar la isometría ($b = 1$). Esto es equivalente a contrastar la siguiente hipótesis nula

$$H_0 : m(x) = ax.$$

En la Figura 3.3 se muestra el diagrama de dispersión, así como la estimación no paramétrica de la función de regresión mediante Nadaraya-Watson. En la estimación no paramétrica de m, se tomó como parámetro ventana $h = 4$, el óptimo obtenido por el método de validación cruzada.

Para la aplicación del contraste en dos etapas se utilizaron diez parámetros ventana seleccionados en torno al h óptimo obtenido por validación cruzada. Los p-valores obtenidos para todos estos contrastes con niveles de significación 0.05 y 0.10 fueron siempre cero. Por tanto, el contraste indica que el modelo isométrico dista mucho de ser apropiado para estos datos, es decir, en caso de ajustarse a un modelo alométrico el paramétro b no es 1. Observando la Figura 3.3 es evidente que el conjunto de datos considerado no puede ser ajustado por un modelo lineal sin constante aditiva.

Por otra parte, si se verifica el modelo alométrico, aplicando logaritmos se puede

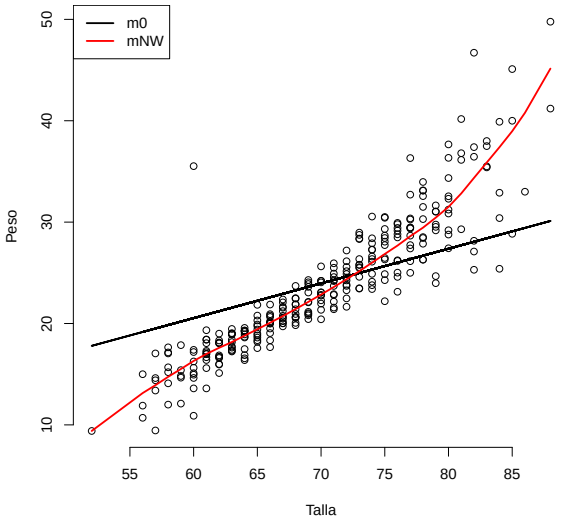

Figura 3.3: Diagrama de dispersión y curvas de regresión estimadas de Y =“peso” frente a X =“talla”.

transformar la expresión (3.2) en la expresión lineal siguiente:

$$\log(Y) = \log(a) + b\log(X) + \sigma^{'}(X)\varepsilon^{'}.$$

Entonces contrastando dicho modelo lineal entre las variables en escala logarítmica se estaría contrastando el modelo alométrico (3.2) entre las variables originales. Consideremos para ello la hipótesis nula de linealidad (3.1).

En la Figura 3.4 se muestra el diagrama de dispersión de las variables $\log(Y)$ y $\log(X)$, así como la estimación no paramétrica de la función de regresión mediante Nadaraya-Watson. En la estimación no paramétrica de m, se tomó como parámetro ventana $h = 0.07$, el óptimo obtenido por el método de validación cruzada. Los p-valores obtenidos para todos estos contrastes se muestran en la Figura 3.5. El contraste indica que el modelo lineal no es apropiado para los datos transformados mediante logaritmos, por lo tanto el modelo alométrico no es apropiado para los datos originales de talla y peso.

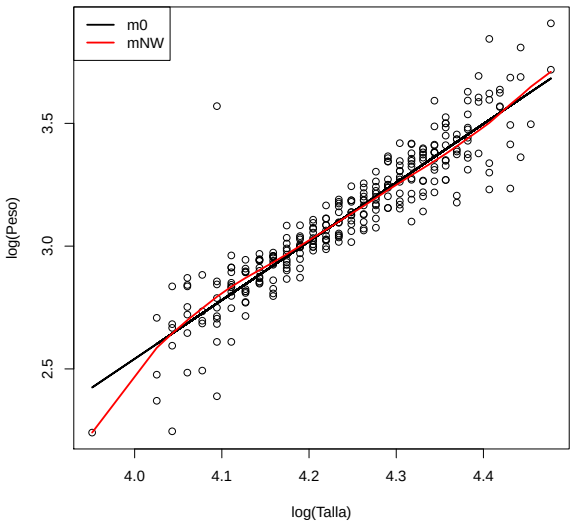

Figura 3.4: Diagrama de dispersión y curvas de regresión estimadas de $Y = \log(\text{“peso”})$ frente a $X = \log(\text{“talla”})$.

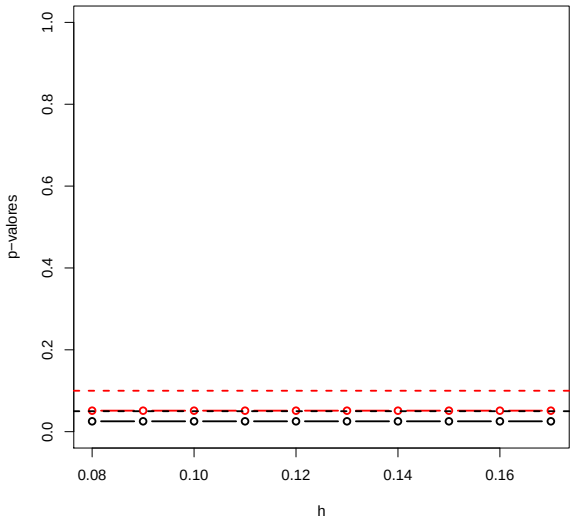

Figura 3.5: Gráfico de los p-valores en función del parámetro ventana h para los datos de talla y peso de mejillón (en negro para $\alpha = 0.05$ y en rojo para $\alpha = 0.10$). Las líneas horizontales discontinuas se corresponden a los niveles 0.05 y 0.10.

3.3. Aplicación 3: Datos sobre el gasto en familias holandesas

Finalmente, en esta sección se ilustra el procedimiento del contraste en dos etapas por medio de un conjunto de datos obtenido del *Journal of Applied Econometrics* introducidos en Adang and Melenberg (1995) (véase también, Neumeyer and Pardo-Fernández (2009) y Pardo-Fernández et al. (2007)). Las observaciones consisten en el gasto anual de los hogares holandeses (desde octubre de 1986 hasta septiembre de 1987). Los datos son registrados en florines holandeses, antigua moneda de los Países Bajos antes de la introducción del euro. Se considera la población formada por los hogares de dos miembros, con 159 observaciones. Einmahl and Van Keilegom (2008) verifican que el modelo

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

con ε independiente de X se cumple cuando X =“logaritmo de los gastos totales” se considera como una covariable e Y =“logaritmo de los gastos en alimento” es la variable respuesta. Ahora contrastaremos la linealidad entre estas variables.

La Figura 3.6 muestra los gráficos de dispersión y la estimación de la curva de regresión mediante Nadaraya-Watson y bajo la hipótesis nula de linealidad. La estimación de la curva no paramétrica se basa en el parámetro ventana $h = 0.47$ obtenido por validación cruzada.

Como en los ejemplos anteriores, para la aplicación del contraste en dos etapas se utilizaron diez parámetros ventana seleccionados en torno al h óptimo obtenido por validación cruzada. Los p-valores obtenidos para todos estos contrastes se muestran en la Figura 3.7. En conclusión, el contraste indica que el modelo lineal es apropiado para estos datos.

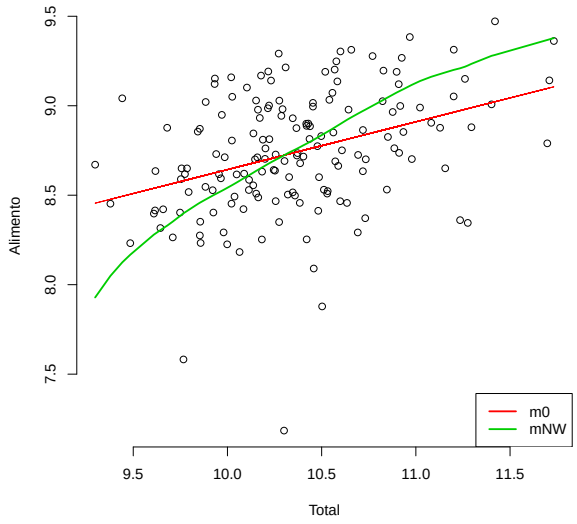

Figura 3.6: Diagrama de dispersión y curvas de regresión estimadas de $Y = \log(\text{“gastos en alimento”})$ frente a $X = \log(\text{“gastos totales”})$.

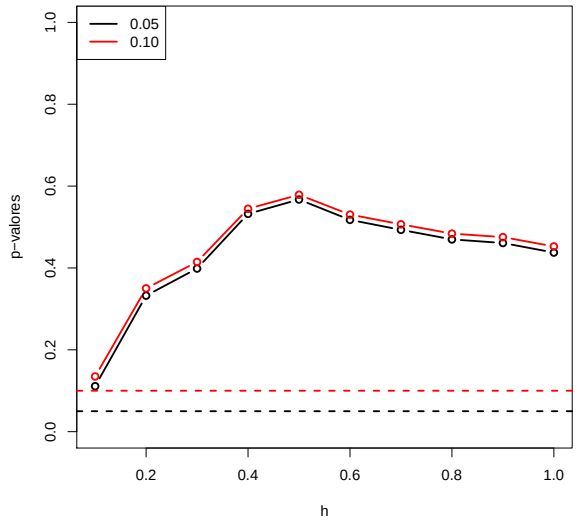

Figura 3.7: Gráfico de los p-valores en función del parámetro ventana h para los datos sobre gasto en familias holandesas (en negro para $\alpha = 0.05$ y en rojo para $\alpha = 0.10$). Las líneas horizontales discontinuas se corresponden a los niveles 0.05 y 0.10.

Bibliografía

Adang, P. and Melenberg, B. (1995). Nonnegativity constraints and intratemporal uncertainty in multi-good life-cycle models. *Journal of Applied Econometrics*, 10:1–15.

Akritas, M. G. and Keilegom, I. V. (2001). Nonparametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, 28:549–568.

Azzalini, A. and Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society. Series B*, 55:549–557.

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

Einmahl, J. and Van Keilegom, I. (2008). Tests for independence in nonparametric regression. *Statistica Sinica*, 18:601–616.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician*, 52:179–192.

González-Manteiga, W., Pardo-Fernández, J. C., and Van Keilegom, I. (2011). Roc curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics*, 38:169–184.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21:1926–1947.

Nadaraya, A. E. (1964). On estimating regression. *Theory of probability and its applications*, 10:186–196.

Neumeyer, N. and Pardo-Fernández, J. C. (2009). A simple test for comparing regression curves versus one-sided alternatives. *Journal of Statistical Planning and Inference*, 139:4006–4016.

Pardo-Fernández, J. C., Van Keilegom, I., and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, 17:1115–1137.

Qiu, P. and Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society. Series B*, 70:191–208.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12:1215–1230.

Seber, G. (1977). *Linear regression analysis.* Wiley.

Seber, G. and Wild, C. (1989). *Nonlinear regression.* Wiley.

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

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

Van Keilegom, I., Manteiga, W. G., and Sellero, C. S. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *TEST*, 17:401–415.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman & Hall.

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

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

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