Trabajo Fin de Máster

Estimación de modas condicionales

José Carlos Soage González

Máster en Técnicas Estadísticas

Curso 2018-2019

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación de modas condicionais
Título en español: Estimación de modas condicionales
English title: Conditional mode estimation
Modalidad: Modalidad A
Autor: José Carlos Soage González, Universidad de Vigo
Directores: Daniel Miles Touya, Universidad de Vigo; Javier Roca Pardiñas, Universidad de Vigo
Breve resumen del trabajo: En este trabajo se presentan las ventajas de la moda en el contexto de la regresión en determinados escenarios. Con tal objetivo se realiza una revisión bibliográfica, se estudian con detalle dos procedimientos de estimación y se propone un nuevo método basado en la distribución de los cuantiles condicionales. Mediante un estudio de simulación se compara el comportamiento de los métodos considerados en el estudio para distintos parámetros. Por último, se realiza una aplicación a datos reales en el contexto de la Economía.

Don Daniel Miles Touya, Titular de la Universidad de Vigo y don Javier Roca Pardiñas, Titular de la Universidad de Vigo, informan que el Trabajo Fin de Máster titulado

Estimación de modas condicionales

fue realizado bajo su dirección por don José Carlos Soage González para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Vigo, a 4 de julio de 2019.

El director:

Don Daniel Miles Touya

El director:

Don Javier Roca Pardiñas

El autor:

Don José Carlos Soage González

Agradecimientos

Agradecer en primer lugar a mis tutores Daniel Miles Touya y Javier Roca Pardiñas. A Daniel por su incesante ayuda, valorar mi esfuerzo a lo largo de estos años y por motivarme a sacar la mejor versión de mi. A Javier por su apoyo desde que comencé el máster y ofrecerme tantas oportunidades para poder seguir mejorando. En segundo lugar, agradecer a mis compañeros del Grupo SiDOR, en especial a Natalia Pérez Veiga, Marta Cousido Rocha y Francisco Mora Posada por ayudarme tanto, por estar ahí siempre y haber hecho tan agradable esta etapa. Por último, pero no menos importante, me gustaría agradecer a Jacobo de Uña Álvarez, por la confianza depositada en mi.

Índice general

Resumen

Resumen en español

La regresión modal utiliza la moda para modelizar la dependencia entre variables. Las principales ventajas son su robustez frente a datos atípicos y que permite conocer los distintos patrones que presenten los datos. Para su estimación existen métodos no paramétricos y semiparamétricos. Los no paramétricos son más flexibles pero presentan el problema del desastre de la dimensionalidad al incrementar el número de covariables, que afectan a las tasas de convergencia de las estimaciones. La estimación semiparamétrica es más restrictiva pero evita los problemas de los métodos no paramétricos, permitiendo analizar cómo cambios en las características afectan a la moda condicional. El principal inconveniente es que en la mayoría de los casos es necesario solucionar problemas de optimización multidimensional no convexa.

En este estudio proponemos un método eficiente y escalable de estimación de modas condicionales a través de la distribución empírica de los cuantiles condicionales. El método se compara en un estudio de simulación con un método no paramétrico basado en el algoritmo mean-shift y un estimador semiparamétrico que utiliza la densidad de los cuantiles condicionales para estimar las modas. Por último, la regresión modal se aplica para analizar la toma de decisiones en el mercado laboral así como las diferencias salariales entre hombres y mujeres desde el punto de vista de la moda. Con tal objetivo se han utilizado datos de varias encuestas de estructura salarial española.

English abstract

Modal regression uses the mode to model the dependence between variables. The main advantages are its robustness against outliers and that allows knowing the different patterns that are present on the data. For its estimation there are nonparametric and semi-parametric methods. The nonparametric ones are more flexible but the main problem is the curse of dimensionality when the number of covariables increases, affecting the convergence rates of the estimates. The semi-parametric estimation is more restrictive but avoids the problems of the non-parametric methods, allowing to analyze how changes in the characteristics affect the conditional mode. The main drawback is that in most cases it is necessary to solve non-convex multidimensional optimization problems.

In this study we propose an efficient and scalable conditional mode estimation method through the empirical distribution of conditional quantiles. The method is compared in a simulation study with a non-parametric method based on the mean-shift algorithm and a semi-parametric estimator that uses the density of the conditional quantiles in order to estimate the modes. Finally, modal regression is applied to analyze decision making in the labor market as well as wage differences between men and women from the point of view of the mode. For this purpose, data from several spanish salary structure surveys has been used.

Prefacio

Los *modelos de regresión* son técnicas estadísticas que se utilizan para modelizar la dependencia entre variables. La primera forma de regresión la publicó Legendre en 1805 y los primeros métodos fueron utilizados por Gauss y Laplace en el contexto de la Astrofísica y la Astronomía. Sin embargo, el término regresión proviene de los estudios de Galton a finales del siglo XIX para determinar la dependencia entre la estatura de los padres y los hijos. En el estudio encontró lo que denominó como "regresión a la media". Desde entonces, los métodos de estimación de la media condicional han avanzado desde la regresión lineal simple hasta distintos tipos de métodos no paramétricos. No fue hasta 1978 cuando Koenker y Bassett propusieron la regresión cuantil. El método permitió estimar distintos cuantiles en lugar de la media, resultando de gran interés en distintos escenarios. Al igual que sucedió con la regresión lineal, la literatura en el contexto de la regresión cuantil se extendió y su uso se generalizó. En 1982 Sager y Thisted publicaron la primera propuesta de regresión utilizando la moda. Desde entonces surgieron distintos métodos de estimación modal condicional, pero su uso todavía no se ha generalizado a pesar de sus ventajas en distintas situaciones.

En este Trabajo de Fin de Máster se presentan varios métodos de estimación de modas condicionales, analizando las ventajas de este tipo de regresión frente a las técnicas de regresión ordinarias. Es preciso destacar que en esta memoria se realizan las siguientes aportaciones al contexto de la regresión modal:

- Presentación y análisis de un nuevo método de estimación modal basado en regresión cuantil. La estimación se realiza en los puntos de inflexión de las distribuciones de los cuantiles condicionales.
- Estudio de simulación para analizar el comportamiento de los diferentes métodos presentados a lo largo del trabajo.
- Estudio de simulación del método no paramétrico de estimación basado en el algoritmo mean-shift, ya que hasta donde alcanza nuestro conocimiento no se ha realizado antes.
- Estudio econométrico para analizar las diferencias salariales entre hombres y mujeres y la toma de decisiones en el mercado laboral desde el punto de vista de la moda, así como el impacto de la crisis en los salarios. Para ello se han utilizado varios conjuntos de datos de encuestas cuatrienales de estructuración salarial española.

La memoria se estructura de la siguiente manera. En el Capítulo 1 se realiza una introducción y revisión bibliográfica de los métodos existentes y sus características. En el Capítulo 2 se presentan los conceptos básicos de la terminología empleada a lo largo del trabajo y tres métodos de estimación de modas condicionales. El primero de ellos es no paramétrico y está basado en el algoritmo mean-shift, el segundo se basa en regresión cuantil lineal y por último, presentamos un nuevo procedimiento de estimación a través de la distribución de los cuantiles condicionales. Para finalizar el capítulo se revisa el software disponible. El Capítulo 3 se dedica a comparar los distintos métodos presentados mediante un estudio de simulación donde se consideran distintos modelos, tanto con funciones modales lineales como no lineales. En el Capítulo 4 se realiza un estudio econométrico a partir de los datos de las encuestas de estructura salarial españolas de los años 2002, 2006, 2010 y 2014. Por último, el trabajo se cierra con un capítulo de conclusiones.

Capítulo 1

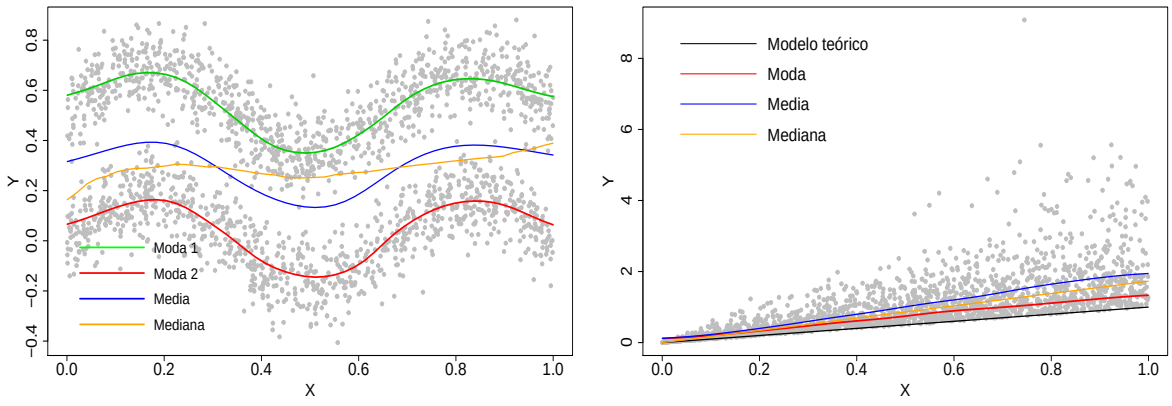
Introducción

En el contexto de la regresión existen distintos procedimientos para establecer la relación entre una variable dependiente Y y una serie de covariables X dado $X = x$. Con tal objetivo se debe determinar una función $m : \mathbb{R} \to \mathbb{R}$ que puede ser expresada de manera genérica como $m(x) = \Omega(Y|X)$, donde $\Omega(\cdot)$ relaciona X e Y. La elección de Ω debe realizarse de manera adecuada en base a las características e información que se necesiten sobre los datos de estudio.

La selección más habitual de Ω en la literatura es la esperanza, esto es, $\Omega(\cdot) = \mathbb{E}(\cdot)$. Otra elección ampliamente estudiada son los cuantiles (Koenker y Bassett, 1978), en especial la mediana $Med(\cdot)$. A pesar de ello, la media y la mediana presentan una serie de desventajas cuando la distribución de los datos no es simétrica, o cuando las funciones de densidad condicionales presentan múltiples modas (varios máximos locales), impidiendo que se capturen todos los patrones que presentan los datos. Para solventar estos inconvenientes puede utilizarse la moda a la hora de modelizar la dependencia entre variables. Las principales ventajas de la moda son su robustez ante datos atípicos y que permite conocer los distintos patrones que presenten los datos. Las estimaciones de regresión modal son aquellas donde se acumula mayor probabilidad. Además, en el caso de que las densidades condicionales sean simétricas, la estimación modal será equivalente a la de la media o la mediana. De estas características nace el interés de trabajar con la moda condicional $\Omega(\cdot) = Moda(\cdot)$.

Como ejemplo para resaltar la importancia de utilizar la moda podemos pensar en los salarios y en la toma de decisiones de los individuos a la hora de seguir estudiando o no (Manski, 2003). Los datos salariales presentan una gran asimetría debido a las rentas más elevadas. Por tanto, el salario más probable estará lejos del salario promedio. En consecuencia, ¿qué información es más importante para tomar la decisión de estudiar más tiempo: un incremento en el salario esperado o un incremento del salario más probable? Otro ejemplo es la implementación de políticas de gestión de recursos para aumentar la satisfacción de los trabajadores. ¿Qué indicador es más preciso para medir su efectividad, el promedio o la moda de los cambios en los niveles de satisfacción inducidos por estas nuevas políticas? Por lo general, la distribución de la satisfacción laboral es asimétrica a la izquierda y centrarse en la media podría ser perjudicial a la hora de tomar decisiones.

En la Figura 1.1 se muestran dos ejemplos con datos simulados donde la estimación de la media o mediana refleja resultados que no se corresponden con el patrón que presentan los datos. En (a) como consecuencia de la existencia de dos patrones. En (b) debido al carácter heterocedástico de los datos. En línea negra se muestra el modelo que genera los datos. Al ser las distribuciones condicionales asimétricas la moda es una mejor aproximación. En los ejemplos la media ha sido calculada mediante regresión Nadaraya-Watson, la mediana a partir de regresión cuantil tipo núcleo y las modas usando el método de estimación propuesto en Einbeck y Tutz (2006).



(a) (b)

Figura 1.1: Ejemplos en los que sería conveniente aplicar regresión modal.

1.1. Revisión bibliográfica

La *moda* es una medida de localización que se define como el valor más probable de una variable aleatoria o como el valor más frecuente de un conjunto de observaciones. Como ya se ha mencionado, es una medida robusta ante datos atípicos que puede tomar varios valores. Además, coincide con la media y la mediana para densidades simétricas.

La estimación de las modas se puede realizar tanto de manera paramétrica como no paramétrica. Generalmente se realiza a través de la solución de un problema de maximización de una función de densidad no paramétrica. A pesar de ello, las densidades pueden tener una moda (unimodalidad) cuando existe un solo máximo local o múltiples modas (multimodalidad) si existen varios máximos locales. Cabe destacar que el supuesto de unimodalidad de las densidades ha sido general en la literatura de estimación modal.

Los distintos procedimientos de regresión modal se pueden agrupar en dos categorías dependiendo del supuesto sobre el número de modas de las densidades. Por una parte, existen métodos que estiman la moda condicional global, que se conocen como *regresión unimodal* (Collomb et al., 1986; Lee, 1989; Manski, 1991; Ohta et al., 2018, entre otros). Como ya se ha mencionado, el supuesto de unimodalidad ha sido una condición general (y restrictiva) en la literatura de estimación de modas. Por otra parte, la estimación de múltiples modas se consideró en un principio en el contexto del reconocimiento de imágenes y patrones aplicando el algoritmo mean-shift (Fukunaga y Hostetler, 1975; Cheng, 1995; Comaniciu y Meer, 2002). No fue hasta el año 2006 cuando Einbeck y Tutz (2006) propusieron el primer procedimiento para estimar las modas condicionales mediante una modificación de dicho algoritmo. Posteriormente, en Chen et al. (2016) se demostró la consistencia de la estimación multimodal. Hasta donde alcanza nuestro conocimiento, este procedimiento de estimación es el único que existe de lo que se conoce como *regresión multimodal*. Una de sus mayores ventajas es que no es necesario hacer supuestos sobre el número de modas que presentan los datos. En el caso de que se desconozca el número de modas puede iniciarse el algoritmo con múltiples puntos de arranque. Las distintas estimaciones deberían converger entre ellas al número de modas.

Los primeros estudios sobre la estimación no paramétrica de modas se remonta a la década de los 60 para el caso univariante. En Parzen (1962) se introdujo la estimación no paramétrica tipo núcleo como primera aproximación para estimar densidades y sus modas. En dicho artículo se estudia consistencia,

normalidad asintótica de los estimadores y el error cuadrático medio de la moda muestral tipo núcleo. Estos resultados se extendieron en varias direcciones. Entre ellas destacan Chernoff (1964), que define la moda como el valor central del intervalo que contiene un mayor número de observaciones de una serie de intervalos equiespaciados. El estimador es idéntico al de Parzen cuando el kernel elegido es el uniforme; Dalenius (1965), que presenta tres estimadores explotando el hecho de que cerca de la moda las observaciones muestrales pueden presentar agrupamientos; Robertson y Cryer (1974), donde la estimación se obtiene a partir de un conjunto de intervalos anidados, cada uno de los cuales es un intervalo más corto que contiene un cierto número de observaciones y Eddy (1980), que investiga las funciones tipo núcleo óptimas para estimar la moda. Recientemente, Dasgupta y Kpotufe (2014) proponen un estimador de la moda basado en el método de los k vecinos más próximos.

No fue hasta 1982 cuando se consideró por primera vez la estimación de modas condicionales. La regresión modal fue propuesta en Sager y Thisted (1982), donde se estudió un estimador no paramétrico de máxima verosimilitud suponiendo la moda global como una función monótona (isotónica) de las covariables. La estimación de la regresión se realiza minimizando una función de pérdida 0-1. Desde entonces se han publicado numerosos artículos relacionados con la regresión modal, entre los que podemos destacar Lee (1989, 1993), Kemp y Santos-Silva (2012), Yao et al. (2012), Yao y Li (2014), Einbeck y Tutz (2006), Chen et al. (2016) y Ohta et al. (2018). Dentro de los distintos procedimientos existen dos tipos diferentes de estimación: métodos no paramétricos y semiparamétricos.

Los *métodos no paramétricos* son los más habituales para la estimación de modas a través de la estimación tipo núcleo de densidades condicionales. Por tanto, en estos métodos no se asume una forma funcional particular de las densidades. Collomb et al. (1986) estudiaron la regresión modal en el caso de dependencia, aplicándose a series de tiempo estrictamente estacionarias. Posteriormente, Samanta y Thavaneswaran (1990) demostraron que los estimadores de modas condicionales que se derivan maximizando densidades tipo núcleo son consistentes y asintóticamente normales. En Quintela-Del-Río y Vieu (1997) se propone un estimador no paramétrico que estima la moda a través de una de la estimación tipo núcleo de la primera derivada de la función de densidad condicional. Como ya se ha comentado, Eibeck y Tutz (2006) desarrollaron un estimador plug-in de las densidades condicionales y estiman las modas locales modificando el algoritmo mean-shif. La extensión del procedimiento al caso multivariante se estudia en Taylor (2012) y el soporte teórico en Chen et al. (2016). En Zhou y Huang (2016) se generaliza dicho procedimiento para el caso en el que existan errores de medida en las variables, pero se asume que se conoce la distribución de los errores de medida. Además, los parámetros ventana requeridos en este procedimiento pueden ser calculados mediante los selectores descritos en Bashtannyk y Hyndman (2001) o en Zhou y Huang (2019). Por otra parte, Yao et al. (2012) proponen un estimador modal local e introducen un parámetro que se selecciona automáticamente utilizando los datos observados para conseguir robustez y eficiencia de las estimaciones. A pesar de la flexibilidad que aportan, los métodos no paramétricos presentan una serie de desventajas. La primera de ellas es el desastre de la dimensionalidad al incrementar el número de covariables, que afectan a las tasas de convergencia de las estimaciones. En segundo lugar, presentan el problema de cómo encontrar los máximos de las funciones de densidad condicionales (Einbeck y Tutz, 2006). Por último, no es posible determinar cómo cambios en las covariables afectan a la moda.

Los métodos de regresión modal *semiparamétrica* presentan una serie de ventajas respecto a la estimación no paramétrica. Por lo general asumen que la moda es una función lineal de x tal que $\text{Moda}(Y|X = x) = x'\beta$, siendo β un vector de pendientes. Aunque la especificación paramétrica es restrictiva, evita el desastre la dimensionalidad que existe en los métodos de estimación no paramétrica. Además, permite analizar cómo cambios en las características afectan a la moda condicional. En Lee (1989) se desarrolla un método de estimación modal para variables dependientes truncadas. Con tal objetivo se utiliza una función de pérdida con kernel rectangular cuya esperanza se minimiza en la moda. El método de estimación se generaliza en Lee (1993) para funciones tipo núcleo cuadráticas, suavizando el kernel rectangular. El principal inconveniente de los dos trabajos anteriores es que se

impone una condición restrictiva sobre la simetría de las distribuciones condicionales. Bajo dicha simetría, media, moda y mediana coinciden, por lo que plantear la estimación modal carece de sentido. Posteriormente, Kemp y Santos-Silva (2012) consideraron regresión modal semiparamétrica para el caso en que la variable dependiente tenga una densidad condicional continua con una moda global bien definida. En este último trabajo se relajó la condición de simetría que se imponía en Lee (1989, 1993) y se propusieron estimadores con distribución asintótica Normal. Recientemente, Yao y Li (2014) proponen la regresión modal lineal para poder tratar con datos de alta dimensionalidad. Para la estimación se utiliza el algoritmo EM (expectation-maximization). Como curiosidad, podemos destacar que dicho algoritmo guarda relación con el mean-shift. En Carreira-Perpiñán (2007) se llega a la conclusión de que el mean-shift gaussiano es equivalente al algoritmo EM. El principal inconveniente de los métodos de estimación semiparamétricos es que es necesario solucionar problemas de optimización multidimensional no convexa. En Khardani (2019) se propone un método para el caso en el que exista censura a la derecha y en Ohta et al. (2018) se propone por primera vez otra aproximación semiparamétrica pero utilizando regresión cuantil lineal para evitar el desastre de la dimensionalidad. Dicho método es computacionalmente escalable ya que las estimaciones de regresión cuantil pueden ser formuladas como problemas de programación lineal convexa. Además, no son necesarios supuestos de simetría. El principal problema que presenta el método es que no existe una regla de selección óptima para las ventanas requeridas.

La regresión modal se ha aplicado en diferentes ámbitos de estudio, como la econometría (Kemp y Santos-Silva, 2012) o el aprendizaje automático (Feng et al., 2017) y ha sido aplicado a distintos tipos de datos como el tráfico (Einbeck y Tutz, 2006), la predicción de la temperatura (Hyndman et al., 1996), consumo eléctrico (Yao y Li, 2014) y dietética (Zhou y Huang, 2016), demostrando ser una mejor opción que las alternativas en numerosos contextos.

En el siguiente capítulo se hará una breve introducción sobre conceptos previos para entender la regresión modal, revisaremos en profundidad el método no paramétrico multimodal (Einbeck y Tutz, 2006; Chen et al., 2016), el procedimiento basado en regresión cuantil lineal (Ohta et al., 2018), propondremos un nuevo método semiparamétrico también basado en regresión cuantil para solventar los problemas que presenta el método de Ohta et al. (2018) y por último revisaremos el software disponible para la estimación de modas condicionales, así como las nuevas implementaciones que se han realizado.

Capítulo 2

Métodos de estimación de modas condicionales

2.1. Conceptos previos

Sea $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$ una muestra aleatoria i.i.d. de un vector aleatorio (X, Y) donde Y es la variable respuesta continua y X un vector de variables regresoras d-variantes $X = (X_1, \ldots, X_d)'$. Suponiendo que existe una función de densidad condicional $f(y|x)$ de Y dado $X = x$ donde $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ continua en y, entonces, siguiendo Chen et al. (2016) el conjunto de modas condicionales para un punto x puede ser representado como los puntos en los que la derivada de la función de densidad condicional es cero y la derivada segunda es negativa,

$$\mathcal{M}(x) = \left\{ y : \frac{\partial}{\partial y} f(y|x) = 0, \frac{\partial^2}{\partial y^2} f(y|x) < 0 \right\}. \tag{2.1}$$

Nótese que f tiene que ser una función dos veces diferenciable. Con el objetivo de estimar la función de densidad condicional necesaria en (2.1) desde un punto de vista no paramétrico es habitual utilizar la estimación tipo núcleo (kernel) de la función de densidad. Si X es un vector unidimensional ($d = 1$) el estimador de $f(y|x)$ toma la forma

$$\hat{f}(y|x) = \frac{\hat{f}(x, y)}{\hat{f}(x)} = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) G\left(\frac{Y_i - y}{g}\right)}{g \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, \tag{2.2}$$

donde K y G son funciones kernel univariantes, es decir, funciones de densidad con media cero y h y g son parámetros ventana (también los denotaremos indistintamente como bandwidths) que deben ser seleccionados. Ahora bien, si X es un vector multivariante ($d > 1$) la función de densidad condicional puede ser estimada mediante

$$\hat{f}(y|x) = \frac{\hat{f}(x, y)}{\hat{f}(x)} = \frac{\sum_{i=1}^n G\left(\frac{Y_i - y}{g}\right) \prod_{j=1}^d K_j\left(\frac{X_{ij} - x_j}{h_j}\right)}{g \sum_{i=1}^n \prod_{j=1}^d K_j\left(\frac{X_{ij} - x_j}{h_j}\right)}, \tag{2.3}$$

donde G y K_j, $j = 1, \ldots, d$ son kernels univariantes y los valores g y h_j, $j = 1, \ldots, d$ son parámetros ventana que requieren de selección. Nótese que la diferencia principal respecto a (2.2) radica en el productorio de los kernels de las variables explicativas.

Dadas (2.2) y (2.3), el estimador de $\mathcal{M}(x)$ es

$$\widehat{\mathcal{M}}(x) = \left\{ y : \frac{\partial}{\partial y}\hat{f}(y|x) = 0, \frac{\partial^2}{\partial y^2}\hat{f}(y|x) < 0 \right\}. \tag{2.4}$$

Además, si se conoce que la variable respuesta es unimodal o solo se quiere estimar la moda que acumula mayor probabilidad, estaremos ante un caso particular de (2.1). En este caso, la moda condicional corresponde con el único máximo de la función de densidad condicional. Como ya se comentó en el Capítulo 1, en la literatura es habitual el supuesto de unimodalidad. En consecuencia, el estimador toma la forma

$$m(x) = \text{Moda}(Y|X = x) = \arg\max_{y\in\mathbb{R}} f(y|x). \tag{2.5}$$

Dadas (2.2) y (2.3) la moda condicional (2.5) puede ser estimada como

$$\widehat{m}(x) = \widehat{\text{Moda}}(Y|X = x) = \arg\max_{y\in\mathbb{R}} \hat{f}(y|x). \tag{2.6}$$

Dado que $f(y|x) = \dfrac{f(x,y)}{f(x)}$, cabe destacar que si $f(x) > 0$ las modas de $f(y|x)$ y $f(x,y)$ son iguales para un x dado. En consecuencia (2.1), (2.4), (2.5) y (2.6) se pueden reescribir respectivamente como

$$\mathcal{M}(x) = \left\{ y : \frac{\partial}{\partial y}f(x,y) = 0, \frac{\partial^2}{\partial y^2}f(x,y) < 0 \right\}, \quad \widehat{\mathcal{M}}(x) = \left\{ y : \frac{\partial}{\partial y}\hat{f}(x,y) = 0, \frac{\partial^2}{\partial y^2}\hat{f}(x,y) < 0 \right\},$$

$$m(x) = \arg\max_{y\in\mathbb{R}} f(x,y), \qquad \widehat{m}(x) = \arg\max_{y\in\mathbb{R}} \hat{f}(x,y).$$

Los distintos conjuntos de modas estimadas para cada $x \in X$ formarán las funciones modales $\widehat{m} = \{\widehat{m}(x_1), \ldots, \widehat{m}(x_n)\}$. Sea K el número de modas, entonces las funciones modales estimadas pueden representarse como $\widehat{\mathcal{M}} = \{\widehat{m}_1, \ldots, \widehat{m}_K\}$. Si $K > 1$ la regresión se considera multimodal.

Por otra parte, como ya se introdujo en la revisión bibliográfica, las estimaciones unimodales también se pueden calcular de manera semiparamétrica. Sea K_2 un kernel esférico $K_2(x) = \dfrac{1}{2}\mathbb{I}(|x| \leq 1)$ y sea $\hat{f}(x) = \dfrac{1}{2nh}\mathbb{I}(|x - X_i| \leq h)$ la estimación obtenida con un kernel esférico se tiene que

$$\begin{aligned} \arg\max_x \hat{f}(x) &= \arg\max_x \frac{1}{2nh}\mathbb{I}(|x - X_i| \leq h) \\ &= \arg\max_x \sum_{i=1}^{n} \mathbb{I}(|x - X_i| \leq h) \\ &= \arg\min_x \sum_{i=1}^{n} \mathbb{I}(|x - X_i| > h). \end{aligned} \tag{2.7}$$

Las estimaciones de los parámetros se obtienen de (2.7) o de sus generalizaciones. Por ejemplo, si se quiere ajustar un modelo de la forma $m(x) = \beta_0 + \beta_1 x$, los parámetros se estiman como

$$\begin{aligned} (\widehat{\beta}_0, \widehat{\beta}_1) &= \arg\max_{(\beta_0,\beta_1)} \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{I}(|\beta_0 + \beta_1 X_i - Y_i| \leq h) \\ &= \arg\min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} \mathbb{I}(|\beta_0 + \beta_1 X_i - Y_i| > h), \end{aligned} \tag{2.8}$$

para construir el estimador final $m(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ (Lee, 1989; Yao y Li, 2014).

Aunque no entraremos en más detalles, la estimación de la densidad se ha generalizado en Khardani et al. (2010, 2011) y en Ould-Saïd y Cai (2005) para el caso de variables respuesta censuradas a través del estimador de Kaplan-Meier (Kaplan y Meier, 1958).

La diferencia entre la regresión unimodal y multimodal es evidente y la elección entre un tipo u otro de estimación no es trivial. Aunque los beneficios de utilizar regresión multimodal parecen notables dada la mayor flexibilidad que ofrecen, esta no siempre es la mejor opción, ya que la detección de varias modas puede complicar la interpretación de los datos. Además, las modas locales pueden presentar muy baja probabilidad, por lo que se deberá utilizar un tipo de estimación u otra dependiendo de las características y objetivos del estudio que se realice.

En la Figura 2.1 se muestra un ejemplo de una distribución bimodal. En este ejemplo es evidente que una moda es más relevante que otra, pero dependiendo del objeto de estudio podría resultar más interesante tener en cuenta la existencia de la función modal que acumula mayor probabilidad. En el caso de regresión unimodal el estimador correspondería con el máximo global de las densidades mostradas, perdiéndose información sobre el conjunto de datos. Por último, cabe destacar que en la literatura existen contrastes para determinar la multimodalidad. Como ejemplo práctico podemos nombrar el paquete de `R multimode` (Ameijeiras-Alonso et al., 2018), donde se incluyen distintos procedimientos no paramétricos contrastar multimodalidad. De todas formas, hasta donde alcanza nuestro conocimiento no existen contrastes de multimodalidad adaptados al caso condicional.

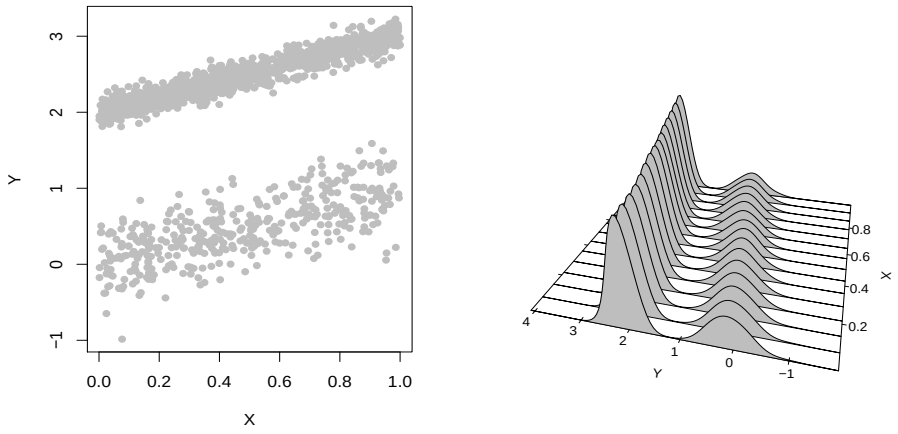


Figura 2.1: Ejemplo de una distribución bimodal. A la izquierda se representa el diagrama de dispersión, mientras que a la derecha se muestran sus densidades condicionales para distintos puntos de diseño.

Como ya se ha mencionado, una cuestión importante consiste en determinar la relevancia de las modas estimadas, con el objetivo de analizar la probabilidad que acumulan. Siguiendo Einbeck y Tutz (2006), la estimación se realiza integrando numéricamente sobre distintas partes de la densidad condicional. Con tal objetivo se ha utilizado la función `integrate.xy` de la librería `sfsmisc` (Maechler et al., 2019). Dada una moda local y en $X = x$, se desciende desde el máximo local $f(y|x)$ en pequeños pasos de longitud δ a la derecha ($k = 0, 1, 2, \dots$) y a la izquierda ($k = -1, -2, \dots$), aumentando la integral en cada paso por $\delta f(y + k\delta|x)$ hasta que se alcanza el mínimo. Cuando éste se alcanza, la secuencia $f(y + k\delta|x)$ deja de decrecer. En la Figura 2.2 se representa la densidad condicional para $x = 0$ de los datos de la Figura 2.1, donde una moda acumula más del doble de probabilidad que la otra. Nótese que según este procedimiento de estimación puede ocurrir el caso en el que dos modas acumulen la misma probabilidad aunque una presente un máximo más alto que la otra.

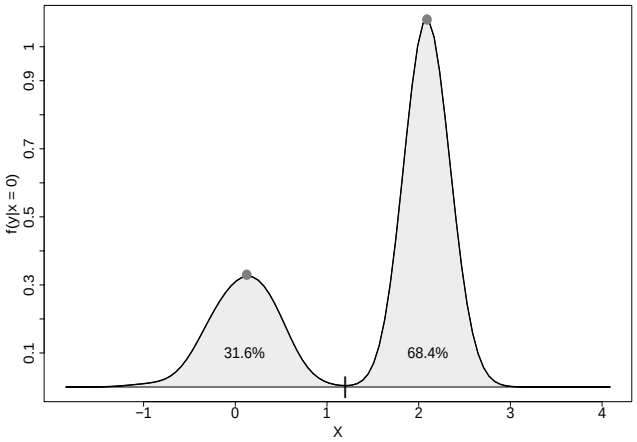


Figura 2.2: Ejemplo de aproximación de las probabilidades asociadas a cada moda local.

2.2. Estimación modal a través del algoritmo mean-shift

En el trabajo de Einbeck y Tutz (2006), se propone un estimador plug-in de las densidades condicionales para cada punto de diseño fijo considerado. Para calcular los máximos locales se utiliza el algoritmo mean-shift. Las propiedades teóricas y la convergencia de esta forma de estimación modal se estudiaron posteriormente en Chen et al. (2016). En esta sección se introducirá el algoritmo mean-shift, su aplicación para la estimación de máximos locales de las densidades kernel condicionales y el cálculo de intervalos de confianza y conjuntos de predicción.

2.2.1. Introducción al algoritmo

El algoritmo *mean-shift* es un método de búsqueda de modas/clusters no paramétrico propuesto en Fukunaga y Hostetler (1975) y adapatado posteriormente en Cheng (1995) y Comaniciu y Meer (2002). Este método no necesita conocer de antemano el número de modas o clusters ni son necesarios supuestos sobre la distribución de los datos.

La idea general del procedimiento es tratar los datos en un espacio d-dimensional a través de una estimación de la función de densidad, donde las zonas en las que se concentran más puntos corresponden con los máximos locales (modas) de la distribución subyacente. Para cada dato se asciende en la dirección del gradiente de la densidad local estimada hasta que se alcanza la convergencia (bajo algún criterio establecido).

Para ilustrar el procedimiento de estimación de los máximos de una densidad tipo núcleo vamos a utilizar un ejemplo unidimensional y un kernel gaussiano. Sea $\{X_i\}_{i=1}^n$ una m.a.s. de X, la estimación tipo núcleo de la densidad se puede calcular para un kernel gaussiano K y una ventana h como

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{2.9}$$

La derivada de (2.9) es

$$\begin{aligned}\frac{\partial}{\partial x}\hat{f}(x) &= \frac{1}{x}\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \\ &= \frac{1}{nh^3}\sum_{i=1}^{n}(x - X_i)K\left(\frac{X_i - x}{h}\right) \\ &= \frac{1}{nh^3}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) - \frac{x}{nh^3}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).\end{aligned} \tag{2.10}$$

Multiplicando ambos lados de (2.10) por nh^3 y dividiéndolos por $\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$ se tiene

$$\frac{nh^3}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}\frac{\partial}{\partial x}\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) X_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} - x. \tag{2.11}$$

Reorganizando la expresión (2.11) se llega a

$$\underbrace{x}_{\text{Localización actual}} + \underbrace{\frac{nh^3}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}\frac{\partial}{\partial x}\hat{f}(x)}_{\text{Gradiente de ascenso } \Delta f(x)} = \underbrace{\frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) X_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}}_{\text{Nueva localización}}.$$

El lado derecho de la expresión (2.11) es lo que se conoce como mean-shift en Fukunaga y Hostetler (1975). Para calcular un máximo local (donde $\frac{\partial}{\partial x}\hat{f}(x) = 0$ y en consecuencia $\Delta f(x) = 0$) el algoritmo necesita partir de una solución inicial $x^{(t)}$ que actualiza como

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

hasta que se alcanza la convergencia, dando lugar al punto estacionario $x^{(\infty)}$ que corresponderá con una de las modas (máximos locales) de la estimación tipo núcleo de la densidad. Para encontrar todos los máximos el algoritmo se puede iniciar desde distintos puntos de arranque que convergerán a los máximos locales de la estimación de la densidad.

2.2.2. Mean-shift condicional

El algoritmo mean-shift clásico permite obtener las modas de la estimación kernel de la densidad pero no permite utilizar covariables para el análisis. En Einbeck y Tutz (2006) se introduce un estimador plug-in de densidades condicionales. Para calcular dichas modas se utiliza el algoritmo mean-shift. Esta aproximación es la primera en la literatura que permite considerar la estimación multimodal en el contexto de la regresión.

Considérese que K_2 pertenece a una clase especial de funciones kernel radialmente simétricas[1] que satisfacen

$$K_2(\cdot) = c_k k[(\cdot)^2],$$

donde c_k es una constante positiva y k es una función que se conoce como perfil kernel (Cheng, 1995). Sea $Y_1, \ldots, Y_n$ una realización muestral de la variable aleatoria Y y X un vector unidimensional, generalizando (2.2) se trabajará con la estimación de la densidad

$$\hat{f}(y|x) = \frac{c_k}{g} \sum_{i=1}^{n} w_i(x) k \left[\left(\frac{Y_i - y}{g} \right)^2 \right], \tag{2.12}$$

donde g es una ventana y $w_i(x)$ es una función de pesos independiente de y. Derivando (2.12) e igualando a cero se tiene

$$\frac{\partial \hat{f}(y|x)}{\partial y} = \frac{2c_k}{g^3} \sum_{i=1}^{n} w_i(x) k' \left[\left(\frac{Y_i - y}{g} \right)^2 \right] (y - Y_i) = 0,$$

de modo que el estimador de la moda condicional $m(x)$ que en esta sección denotaremos como $y(x)$ viene dado por

$$y(x) = \frac{\sum_{i=1}^{n} w_i(x) k' \left[\left(\frac{Y_i - y(x)}{g} \right)^2 \right] Y_i}{\sum_{i=1}^{n} w_i(x) k' \left[\left(\frac{Y_i - y(x)}{g} \right)^2 \right]}.$$

Ahora bien, sea $g(\cdot) = -k'(\cdot)$, donde g es un perfil kernel perteneciente a una función tipo núcleo de la forma $G(\cdot) = c_g g[(\cdot)^2]$ con c_g una constante positiva. Entonces, si K_2 es un kernel gaussiano G también lo será. Reescribiendo la expresión anterior usando G se obtiene

$$y(x) = \frac{\sum_{i=1}^{n} w_i(x) G \left(\frac{Y_i - y(x)}{g} \right) Y_i}{\sum_{i=1}^{n} w_i(x) G \left(\frac{Y_i - y(x)}{g} \right)}.$$

Finalmente, considerando los pesos

$$w_i(x) = \frac{K \left(\frac{X_i - x}{h} \right)}{\sum_{j=1}^{n} K \left(\frac{X_j - x}{h} \right)},$$

se tiene

$$y(x) = \frac{\sum_{i=1}^{n} K \left(\frac{X_i - x}{h} \right) G \left(\frac{Y_i - y(x)}{g} \right) Y_i}{\sum_{i=1}^{n} K \left(\frac{X_i - x}{h} \right) G \left(\frac{Y_i - y(x)}{g} \right)}. \tag{2.13}$$

Denotando el lado derecho de (2.13) como $\mu(y(x))$ podemos calcular $y(x)$ partiendo de una solución inicial $y(x)^{(t)}$ que se actualiza como

$$y(x)^{(t+1)} = \frac{\sum_{i=1}^{n} K \left(\frac{X_i - x}{h} \right) G \left(\frac{Y_i - y(x)^{(t)}}{g} \right) Y_i}{\sum_{i=1}^{n} K \left(\frac{X_i - x}{h} \right) G \left(\frac{Y_i - y(x)^{(t)}}{g} \right)}, \tag{2.14}$$

[1] El procedimiento mean-shift ha sido extendido a kernels anisótropos en Wang et al. (2004).

hasta que se alcance la convergencia, esto es, cuando $\mu(y(x)) - y(x) = 0$. Por lo general se considera que 30 iteraciones son suficientes (Einbeck y Tutz, 2006).

En la Figura 2.4 se muestra una representación gráfica del funcionamiento del algoritmo mean-shift para calcular el máximo de la estimación de una densidad, partiendo de un punto inicial $y_0(x)$.

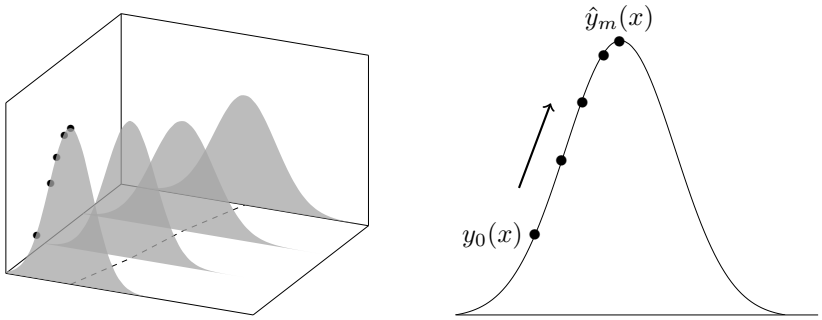


Figura 2.3: Estimación del máximo de una densidad condicional a través del algoritmo mean-shift. A la izquierda, las densidades condicionales y la curva de regresión (línea discontinua). A la derecha, el procedimiento para calcular el máximo a través del algoritmo mean-shift.

Por otra parte, aunque hasta el momento se ha condicionado a una única variable explicativa, es posible generalizar el procedimiento descrito a más covariables. Esta extensión solo afecta a la definición de los pesos $w_i(x)$. Teniendo en cuenta que la densidad condicional tipo núcleo con varias covariables tiene la forma descrita en (2.3), la extensión de la expresión para el caso X univariante mostrada en (2.13) da lugar a la expresión multivariante (2.15), que se resuelve siguiendo el mismo procedimiento que en (2.14).

$$y(x) = \frac{\sum_{i=1}^{n} G\left(\dfrac{Y_i - y(x)}{g}\right) \prod_{j=1}^{d} K_j\left(\dfrac{X_{ij} - x_j}{h_j}\right) Y_i}{\sum_{i=1}^{n} G\left(\dfrac{Y_i - y(x)}{g}\right) \prod_{j=1}^{d} K_j\left(\dfrac{X_{ij} - x_j}{h_j}\right)}. \tag{2.15}$$

El procedimiento descrito en el presente capítulo se muestra en el Algoritmo 1 para un $x \in X$ dado.

Algoritmo 1 : *Regresión modal mediante mean-shift para un* $x \in X$

1: Elegir un conjunto de soluciones iniciales $y_1(x) < \ldots < y_K(x)$.
2: Para $k = 1, \ldots, K$:
   Fijar $t = 0$ e iterar $y_k^{(t+1)}(x) = \mu(y_k^{(t)}(x))$ hasta que se alcance la convergencia (bajo el criterio establecido), resultando los estimadores $\hat{y}_1(x), \ldots, \hat{y}_K(x)$.
3: El estimador de $\mathcal{M}(x)$ es el conjunto aleatorio $\widehat{\mathcal{M}}(x) = \{\hat{y}_1(x), \ldots, \hat{y}_K(x)\}$.

Es necesario destacar que en el caso en el que se desconozca el número de modas (ramas) que presenten los datos, es recomendable utilizar múltiples puntos de arranque del algoritmo. En este caso, si K excede el número de modas, los valores de las estimaciones convergerán a las modas locales.

Por último, tal y como se muestra en Taylor (2012), resulta de interés mencionar que cuando $g = \infty$, la regresión modal presentada en esta sección es equivalente al estimador de Nadaraya-Watson.

2.2.3. Selección de ventanas

Como es habitual, por ser (2.13) un estimador no paramétrico es necesario un procedimiento de selección de las ventanas h y g. Con este objetivo se pueden utilizar diversos métodos, como los distintos selectores descritos en Bashtannyk y Hyndman (2001) y en Zhou y Huang (2019). Entre ellos destacamos el selector bootstrap y el basado en regresión de Bashtannyk y Hyndman (2001).

En la Figura 2.4 se muestra un ejemplo con datos simulados para resaltar la importancia de una selección adecuada de ventanas. Una mala elección puede llevar al algoritmo a máximos locales que disten del máximo global y por tanto la moda estimada de la densidad kernel se alejará de la moda teórica.

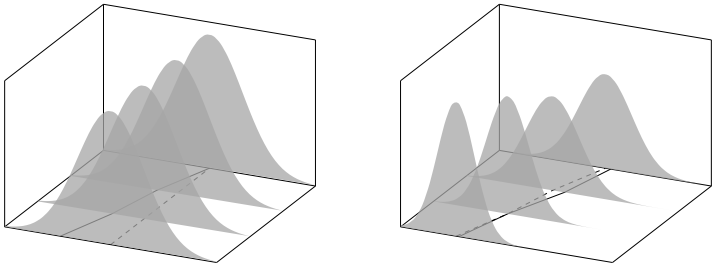


Figura 2.4: Estimación no paramétrica de las densidades condicionales. A la izquierda, estimación con ventanas demasiado grandes. A la derecha, una selección correcta de las ventanas. Las curvas de regresión estimadas se muestran con línea discontinua y la función modal teórica con línea continua.

En el caso en el que existan varias variables explicativas, el problema de selección de ventanas presenta una mayor dificultad, pues se debe seleccionar la ventana g y los valores para cada h_j. Con tal objetivo Taylor (2012) recomienda calcular g como la media de las ventanas calculadas para cada covariable de manera independiente según los procedimientos descritos en Bashtannyk y Hyndman (2001). Para determinar los h_j se recomienda extender al caso multivariante el procedimiento de selección de ventana basado en regresión de Bashtannyk y Hyndman (2001), estandarizando las variables y buscando el bandwidth óptimo que minimiza la media de la raíz del error de predicción penalizado $Q(h)$ respecto a h para un g fijo donde

$$Q(h) = \frac{\Delta}{n}\sum_{k=1}^{N}\sum_{i=1}^{n}\left[\frac{1}{g}G\left(\frac{Y_i - y'_k}{g}\right) - \hat{f}(y'_k|X_i)\right]^2 \left(1 - \frac{(K(0))^d}{\sum_{l=1}^{n}\prod_{j=1}^{d}K_j\left(\frac{X_{ij}-X_{lj}}{h_j}\right)}\right)^{-2},$$

siendo y'_k con $k = 1, \ldots, N$ espaciados equitativamente sobre el soporte de Y tal que $\Delta = y'_{k+1} - y'_k$.

2.2.4. Intervalos de confianza para la función modal

El *bootstrap* es un procedimiento estadístico que sirve para aproximar la distribución en el muestreo. Además, no requiere hipótesis sobre el mecanismo generador de los datos. En Chen et al. (2016) se desarrolla un método bootstrap para la obtención de intervalos de confianza de las funciones modales. Para explicarlo primero debemos definir la distancia de Hausdorff.

Definición 2.2.1. La *distancia de Hausdorff* entre dos conjuntos A y B, tales que $A, B \subset \mathbb{R}^d$ se define como

$$\text{Hausdorff}(A, B) = \{r \geq 0 : A \subset B \oplus r, B \subset A \oplus r\} = \text{máx}\left\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\right\},$$

donde $A \oplus r = \{x \in \mathbb{R}^d : d(x, A) \leq r\}$ siendo $d(x, A) = \text{ínf}_{y\in A}\,||x - y||$.

Así, la distancia de Hausdorff de la estimación puntual entre $\widehat{\mathcal{M}}(x)$ y el valor teórico $\mathcal{M}(x)$ es

$$\Delta(x) = \text{Hausdorff}(\widehat{\mathcal{M}}(x), \mathcal{M}(x)). \tag{2.16}$$

Análogamente, la distancia de Hausdorff de la estimación uniforme entre $\widehat{\mathcal{M}}(x)$ y el valor teórico $\mathcal{M}(x)$ es

$$\Delta = \sup_{x \in \mathcal{D}} \Delta(x) = \sup_{x \in \mathcal{D}} \text{Hausdorff}(\widehat{\mathcal{M}}(x), \mathcal{M}(x)).$$

Dada la muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ y su análogo bootstrap $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ y sea $\mathcal{M}^*(x)$ la estimación bootstrap de la función modal, podemos definir la estimación de la distancia de Hausdorff entre la estimación obtenida mediante la remuestra y la muestra como

$$\hat{\Delta}^*(x) = \text{Hausdorff}(\widehat{\mathcal{M}}^*(x), \widehat{\mathcal{M}}(x)).$$

Repitiendo el proceso un número grande B de veces se obtiene $\hat{\Delta}_1^*(x), \ldots, \hat{\Delta}_B^*(x)$. Definiendo $\hat{\delta}_{1-\alpha}(x)$ como el valor que cumple

$$\frac{1}{B} \sum_{j=1}^{B} I\left(\hat{\Delta}_j^*(x) > \hat{\delta}_{1-\alpha}(x)\right) = \alpha,$$

el intervalo de confianza puntual de $\mathcal{M}(x)$ vendrá dado por

$$\widehat{C}(x) = \widehat{\mathcal{M}}(x) \oplus \hat{\delta}_{1-\alpha}(x),$$

Por otra parte, también es posible calcular los intervalos de confianza uniformes, así como aplicar distintos tipos de bootstrap tales como el paramétrico, suavizado o basado en residuos. Para el caso del bootstrap uniforme o naïve se puede utilizar (2.16) para construir intervalos de confianza uniformes. Definiendo $\delta_{1-\alpha}$ como el valor que verifica

$$P\left(\mathcal{M}(x) \subseteq \widehat{\mathcal{M}}(x) \oplus \delta_{1-\alpha}, \forall x \in \mathcal{D}\right) = 1 - \alpha,$$

podemos estimar vía bootstrap $\hat{\delta}_{1-\alpha}$ a través de los cuantiles de

$$\Delta^* = \sup_{x \in \mathcal{D}} \Delta^*(x) = \sup_{x \in \mathcal{D}} \text{Hausdorff}(\widehat{\mathcal{M}}^*(x), \widehat{\mathcal{M}}(x)).$$

De esta forma se puede definir el intervalo de confianza uniforme como

$$\widehat{C}(x) = \{(x, y) : x \in \mathcal{D}, y \in \widehat{\mathcal{M}}(x) \oplus \hat{\delta}_{1-\alpha}\},$$

cuya cobertura asintótica basada en bootstrap naïve puede encontrarse en Chen et al. (2016).

En la Figura 2.5 se muestran los intervalos de confianza puntuales (izquierda) y uniformes (derecha) de un conjunto de datos simulados de tamaño muestral 500 para un nivel de confianza del 95 % calculados mediante 300 réplicas bootstrap. Puede observarse que los intervalos uniformes son más anchos que los calculados mediante estimación puntual.

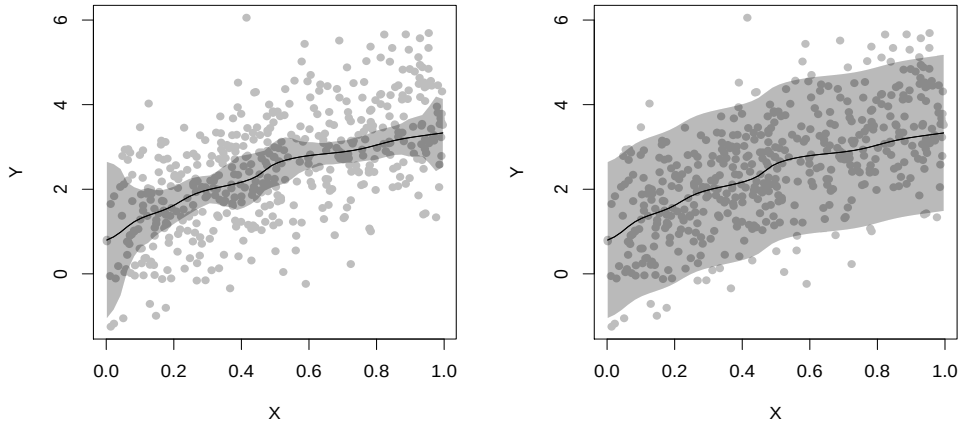


Figura 2.5: Intervalos de confianza para regresión modal (sombreado gris). A la izquierda, los intervalos puntuales. A la derecha, los uniformes.

2.2.5. Conjuntos de predicción para la función modal

En el contexto de la regresión modal también es posible el cálculo de conjuntos de predicción, que pueden utilizarse como método para seleccionar bandwidths óptimos para la estimación kernel de la densidad subyacente (Chen et al., 2016). Además, en el artículo se demuestra que a nivel poblacional los conjuntos de predicción para la regresión modal son más pequeños que para la función de regresión $\mu(x) = \mathbb{E}(Y|X = x)$. Sea

$$\varepsilon_{1-\alpha}(x) = \{\varepsilon \geq 0 : P(d(Y, \mathcal{M}(X)) > \varepsilon | X = x) \leq \alpha\}, \tag{2.17}$$

$$\varepsilon_{1-\alpha} = \{\varepsilon \geq 0 : P(d(Y, \mathcal{M}(X)) > \varepsilon) \leq \alpha\}, \tag{2.18}$$

entonces los conjuntos de predicción puntuales y uniformes a nivel poblacional son, respectivamente,

$$\mathcal{P}_{1-\alpha}(x) = \mathcal{M}(x) \oplus \varepsilon_{1-\alpha}(x),$$

$$\mathcal{P}_{1-\alpha} = \{(x, y) : x \in D, y \in \widehat{\mathcal{M}}(x) \oplus \varepsilon_{1-\alpha}\} \subseteq D \times \mathbb{R}.$$

A nivel muestral se puede estimar (2.17) y (2.18) respectivamente como

$$\widehat{\varepsilon}_{1-\alpha}(x) = \text{ínf}\left\{\varepsilon \geq 0 : \int_{\widehat{\mathcal{M}}(x)\oplus\varepsilon} \hat{f}(y|x)dy \geq 1 - \alpha\right\},$$

$$\widehat{\varepsilon}_{1-\alpha} = Q\left(\left\{d(Y_i, \widehat{\mathcal{M}}(X_i)) : i = 1, \ldots, n\right\}, 1 - \alpha\right).$$

donde $Q(u, 1-\alpha)$ denota el cuantil $1-\alpha$ de u. En consecuencia, los conjuntos de predicción puntuales y uniformes estimados tienen respectivamente la forma

$$\widehat{\mathcal{P}}_{1-\alpha}(x) = \widehat{\mathcal{M}}(x) \oplus \widehat{\varepsilon}_{1-\alpha}(x)$$

$$\widehat{\mathcal{P}}_{1-\alpha} = \{(x, y) : x \in D, y \in \widehat{\mathcal{M}}(x) \oplus \widehat{\varepsilon}_{1-\alpha}\}$$

En la Figura 2.6 se muestran los conjuntos de predicción para los mismos datos de la Figura 2.5 considerando $\alpha = 0.05$.

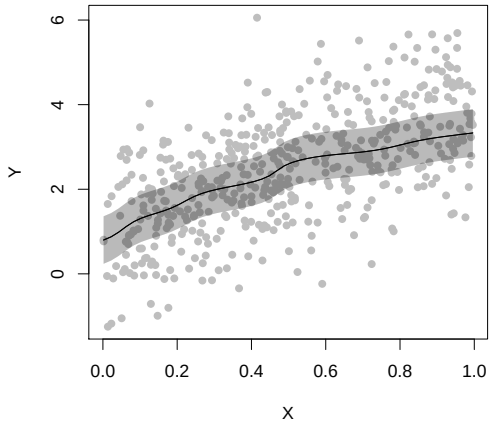


Figura 2.6: Conjuntos de predicción uniformes (sombreado en gris) para regresión modal.

2.3. Estimación modal a través de la densidad condicional

En esta sección se estudia el estimador propuesto en Ohta et al. (2018). La idea principal consiste en estimar los mínimos de la derivadas de las funciones de densidad condicionales, que corresponden con los máximos de la funciones de densidad condicionales y por tanto con la función modal. Este método semiparamétrico se basa en regresión cuantil lineal, permitiendo una estimación unimodal escalable a un número elevado de variables explicativas, a diferencia del método estudiado en el capítulo anterior. En lo que sigue se presentará el estimador, el procedimiento de selección de ventanas y los distintos métodos para calcular intervalos de confianza.

2.3.1. El estimador de Ohta

El procedimiento de estimación consiste en determinar la densidad condicional para cada $X = x$. Este objetivo se logra a través de un modelo de regresión cuantil lineal. En este sentido, se procede calculando y minimizando la derivada de la función cuantil condicional por diferenciación numérica. La moda se encontrará en el mínimo de esa derivada, que corresponde con el máximo de la función de densidad de los cuantiles condicionales.

Con el objetivo de estimar la función modal, en Ohta et al. (2018) se propone invertir un modelo de regresión cuantil (Koenker y Bassett, 1978). Sea $Q(\tau|X)$ el τ-cuantil condicional de Y dado X para $\tau \in (0,1)$ y $Q_x(\tau) = Q(\tau|X = x)$, se puede observar que bajo ciertas condiciones de regularidad

$$s_x(\tau) = Q'_x(\tau) = \frac{\partial Q_x(\tau)}{\partial \tau} = \frac{1}{f(Q_x(\tau)|x)}.$$

Definiendo τ_x como el cuantil que minimiza $s_x(\tau)$, esto es, $\tau_x = \arg\min_{\tau\in(0,1)} s_x(\tau)$, se tiene que la moda estimada en $X = x$ es la función cuantil en ese τ_x, por tanto $m(x) = Q_x(\tau_x)$.

En la práctica $s_x(\tau)$ es desconocida, por lo que debe ser estimada. La estimación se realiza por diferenciación numérica del estimador de la función cuantil condicional $Q_x(\tau)$. En consecuencia también es necesario estimar esta última. Con este objetivo se emplea un modelo de regresión cuantil con la forma

$$Q(\tau|X) = X'\beta(\tau), \quad \tau \in (0,1).$$

donde $\beta(\tau) \in \mathbb{R}^d$ es un vector de pendientes para cada $\tau \in (0,1)$. El vector de pendientes $\beta(\tau)$ puede ser estimado como

$$\hat{\beta}(\tau) = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \rho(Y_i - X_i'\beta), \tag{2.19}$$

donde $\rho_\tau(u) = u\{\tau - \mathbb{I}(u < 0)\}$, siendo $\mathbb{I}$ una función indicadora. La representación de la función de pérdida ρ_τ se puede ver en la Figura 2.7.

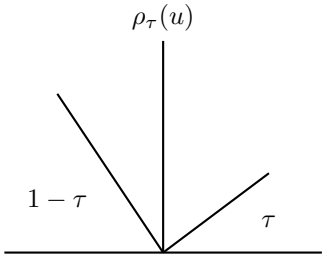


Figura 2.7: Representación de la función de pérdida.

Nótese que la linealidad de la función de regresión cuantil no implica la linealidad de la función modal. En consecuencia, es posible estimar funciones modales no lineales debido a la flexibilidad que aporta el modelo considerado. Este hecho puede observarse en la Figura 2.8, donde en (a) se tiene una función modal teórica no lineal (en rojo) generada en un modelo lineal. En este caso, para cada $x \in X$ considerado se puede observar que la distancia mínima entre cuantiles estimados parece seguir la función teórica. A modo ilustrativo, se puede ver que para $x = 0.1$, la distancia más pequeña corresponde con los cuantiles 0.05 y 0.15. Sin embargo, para $x = 0.7$, esta distancia se encuentra entre los cuantiles 0.65 y 0.75. Esto es lo que permite ajustar funciones modales no lineales mediante regresión cuantil lineal. Por otra parte, para (b), la función modal es lineal. Por tanto, la distancia entre cuantiles estimados más pequeña para distintos x se encuentra siempre entre los cuantiles 0.45 y 0.55 a lo largo del soporte de X.

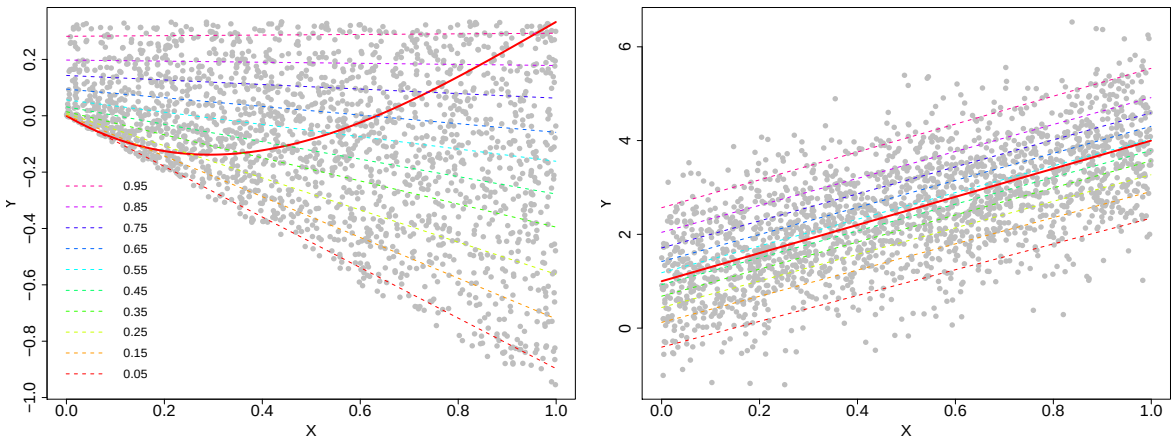


Figura 2.8: Distintos ajustes de regresión cuantil lineal. En (a) un modelo con función modal no lineal. En (b) un modelo con función modal lineal. En rojo (sólido) se representan las funciones modales teóricas.

Por otra parte, la estimación de $s_x(\tau) = Q'_x(\tau)$ se realiza mediante diferenciación numérica. Sea $h = h_n \to 0$ una secuencia de ventanas tales que $nh^2 \to \infty$, se tiene que

$$\hat{s}_x(\tau) = \frac{\widehat{Q}_x(\tau + h) - \widehat{Q}_x(\tau - h)}{2h}. \tag{2.20}$$

A efectos prácticos se utiliza el estimador (2.21), ya que en una muestra finita, $[\tau - h, \tau + h]$ puede no estar incluido para algún $\tau \in [\epsilon, 1 - \epsilon]$.

$$\hat{s}_x(\tau) = \frac{\widehat{Q}_x(\tau + \text{mín}\{h, \tau_{\text{máx}} - \tau\}) - \widehat{Q}_x(\tau - \text{mín}\{h, \tau - \tau_{\text{mín}}\})}{\text{mín}\{h, \tau_{\text{máx}} - \tau\} + \text{mín}\{h, \tau - \tau_{\text{mín}}\}}. \tag{2.21}$$

Finalmente, $m(x)$ se estima como $\widehat{m}(x) = \widehat{Q}_x(\widehat{\tau}_x) = x'\hat{\beta}(\widehat{\tau}_x)$, donde $\widehat{\tau}_x$ es un minimizador aproximado de $\hat{s}_x(\tau)$ en $[\epsilon, 1 - \epsilon]$ con parámetro $\epsilon \in (0, 1/2)$ suficientemente pequeño. Es decir, se busca el τ que minimiza la estimación de la función de densidad calculada mediante diferenciación numérica excluyendo un porcentaje ϵ de los extremos de los datos de la densidad condicional.

2.3.2. Selección de la ventana

Por ser (2.21) un estimador obtenido mediante diferenciación numérica, es necesario el desarrollo de un método para la selección del parámetro ventana h. En Ohta et al. (2018) se propone modificar la regla de selección de ventana de Koenker y Machado (1999) dado que ésta no cumple las condiciones establecidas en el artículo y es τ-dependiente. El bandwidth establecido en Koenker y Machado (1999) se define como

$$h^{KM}(\tau) = n^{-1/3} z_\alpha^{2/3} \left\{ 1.5 \frac{\phi(\Phi^{-1}(\tau))}{2\Phi^{-1}(\tau)^2 + 1} \right\}^{1/3},$$

donde ϕ y Φ son respectivamente la densidad y distribución de la $N(0, 1)$ y $z_\alpha = \Phi^{-1}(1 - \alpha/2)$. Dada la regla de selección para cada punto de diseño x en el soporte de X se procede mediante un algoritmo en dos pasos. El primero de ellos consiste en determinar el bandwidth óptimo τ-independiente. Una vez se calcula esa ventana se vuelven a realizar los mismos pasos empleando mencionada ventana, es decir:

1. Usar la ventana piloto $h^{pilot} = n^{1/6} h^{KM}(0.5) \propto n^{-1/6}$ y construir un estimador preliminar $\widehat{\tau}_x^{prelim}$ de τ_x.

2. Utilizar $h_n = h_{n,x} = n^{1/6} h^{KM}(\widehat{\tau}_x^{prelim})$ para construir un estimador final $\widehat{m}(x)$.

El procedimiento de estimación descrito se muestra en el Algoritmo 2.

Algoritmo 2 : *El estimador de Ohta.*

1: Seleccionar una serie de cuantiles τ equiespaciados, ϵ y α.
2: Calcular la ventana piloto $h^{pilot} = n^{1/6} h^{KM}(0.5) \propto n^{-1/6}$.
3: Para un punto de diseño x en el soporte de X:
  Construir un estimador preliminar $\widehat{\tau}_x^{prelim}$ de τ_x a través de (2.21).
  Calcular $h_n = h_{n,x} = n^{1/6} h^{KM}(\widehat{\tau}_x^{prelim})$.
  Estimar τ_x^{opt} a través de (2.21).
  Calcular los parámetros β para τ_x^{opt}.
  Realizar la predicción como $\widehat{m}(x) = Q_x(\tau_x^{opt})$.
4: La función modal es el conjunto formado por cada moda estimada en cada $x \in X$.

2.3.3. Intervalos de confianza

Para este método es posible construir intervalos de confianza puntuales para $m(x)$ estimando consistentemente los parámetros σ_x^2, v_x y $s_x(\tau_x)$ (la estimación de este último parámetro puede realizarse a través de (2.21)) definidos en el Teorema 1.

Teorema 1. Dado cualquier punto x en el soporte de X, suponiendo que se cumplen ciertas condiciones (pueden consultarse en Ohta et al., 2018) y que $f^{(2)}(m(x)|x) < 0$ siendo $f^{(2)}(y|x) = \partial^2 f(y|x)\ \partial y^2$ y $m(x) \in [Q_x(\varepsilon), Q_x(1-\varepsilon)]$, entonces

$$(nh^2)^{1/3}(\widehat{\tau}_x - \tau_x) \xrightarrow{d} (\sigma_x/v_x)^{2/3} Z$$

cuando $n \to \infty$, donde $\sigma_x = \sqrt{\mathbb{E}[(x'J(\tau_x)^{-1}X)^2]/2}$, con $J(\tau) = \mathbb{E}[f(X'\beta(\tau)|X)XX']$, $v_x = s_x''(\tau_x)/2$ y $Z = \arg\max_{t\in\mathbb{R}}\{B(t) - t^2\}$ siendo $\{B(t) : t \in \mathbb{R}\}$ un movimiento browniano estándar bilateral cuya distribución es la distribución de Chernoff. Además se tiene que

$$(nh^2)^{1/3}(\widehat{m}_x - m_x) \xrightarrow{d} s_x(\tau_x)(\sigma_x/v_x)^{2/3} Z.$$

Por otra parte, sea $\Sigma = \mathbb{E}(XX')$, entonces $\sigma_x^2 = x'J(\tau_x)^{-1}\Sigma J(\tau_x)^{-1}x/2$. Las matrices Σ y $J(\tau)$ pueden ser estimadas como

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i', \quad \widehat{J}(\tau) = \frac{1}{2nh}\sum_{i=1}^{n} I(|Y_i - X_i'\hat{\beta}(\tau)| \le h) X_i X_i',$$

de modo que la estimación de σ_x^2 es $\hat{\sigma}_x^2 = x'\widehat{J}(\tau_x)^{-1}\widehat{\Sigma}\widehat{J}(\tau_x)^{-1}x/2$. Por otra parte, la estimación de $v_x = s_x''(\tau_x)/2 = Q_x'''$ se realiza por diferenciación numérica de $\widehat{Q}_x(\tau)$. Sea Δ_h tal que $\Delta_h g(\tau) = (g(\tau+h) - g(\tau-h))/2h$ y $\Delta_h^j g = \Delta_h(\Delta_h^{j-1} g)$ recursivamente para $j = 2, 3, \ldots$, se puede estimar v_x por

$$\hat{v}_x = \frac{1}{2}\Delta_h^3 \widehat{Q}_x(\widehat{\tau}_x).$$

Cabe destacar que en Ohta et al. (2018) pueden encontrarse estimadores alternativos para v_x.

Finalmente los intervalos de confianza asintóticos para $m(x)$ al nivel $1-\alpha$ vendrán dados por

$$\widehat{m}(x) \pm \frac{\hat{s}_x\ (\widehat{\tau})(\hat{\sigma}_x/\hat{v}_x)^{2/3}}{(nh^2)^{1/3}} q_{1-\alpha/2},$$

siendo $q_{1-\alpha/2}$ el cuantil $1-\alpha/2$ de la distribución de Chernoff.

Además, también es posible calcular intervalos de confianza mediante submuestreo. No se consideran intervalos de confianza bootstrap porque no serían consistentes para este estimador (Ohta et al., 2018). Sea $\widehat{m}(x) = \widehat{m}_n(x) = \widehat{m}_n(x; (Y_1, X_1)\ldots., (Y_n, X_n))$, $h = h_n$ y $W_1, \ldots, W_N$ los N subconjuntos de $(Y_1, X_1)\ldots., (Y_n, X_n)$ de tamaño ℓ tales que $N =$ y $\ell < n$. Considérese la distribución de submuestreo

$$U_{n,l}(x;t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left\{(\ell h_l^2)^{1/3}(\widehat{m}_l(x;W_i) - \widehat{m}_n(x)) \le t\right\}.$$

Cabe destacar que en la práctica el submuestreo es computacionalmente costoso, aunque se puede utilizar un subconjunto aleatorio de $\{1, \ldots, N\}$. Ahora bien, sea $\hat{q}_{n,\ell}(x; 1-\alpha)$ el cuantil $1-\alpha$ de $U_{n,l}(x;\cdot)$, el intervalo de confianza asintótico a nivel $1-\alpha$ de $m(x)$ es

$$\left[\widehat{m}(x) - \frac{\hat{q}_{n,\ell}(x; 1-\alpha/2)}{(nh^2)_n^{1/3}}, \widehat{m}(x) - \frac{\hat{q}_{n,\ell}(x; \alpha/2)}{(nh^2)_n^{1/3}}\right].$$

2.4. Estimación modal a través de la distribución condicional

El principal inconveniente del método de estimación presentado en la Sección 2.3 es que no existe una regla de selección óptima para la selección de ventanas. Además, el selector propuesto tiene un elevado coste computacional. En esta sección proponemos un nuevo estimador semiparamétrico de modas condicionales a través de cuantiles condicionales. La principal diferencia con respecto al método de estimación descrito en Ohta et al. (2018) es que se empleará la distribución condicional empírica en lugar de la densidad. Además de describir el nuevo estimador, se analizarán las ventajas que se obtienen de este cambio de perspectiva y sus posibles extensiones.

2.4.1. Motivación y ventajas del nuevo estimador

El método propuesto es similar a los métodos descritos en Dalenius (1965), Venter (1967) o Robertson y Cryer (1974), que propusieron estimar la moda fijando un número de observaciones y eligiendo el intervalo más corto que contiene dicho número.

La idea es fijar la probabilidad de cada intervalo en lugar del número de observaciones y buscar el intervalo más pequeño que acumula dicha probabilidad, esto es, buscar la distancia más corta entre dos cuantiles consecutivos, que llamaremos q_i y q_{i+1}. Dicho intervalo es en el que se encuentra el punto de inflexión de la función de distribución condicional. La representación de esta idea se muestra en la Figura 2.9, donde la moda se estimará como la media de los cuantiles 0.4 y 0.6, pues son los que representan el intervalo más corto con extremos q_i y q_{i+1}. En este ejemplo es evidente que la estimación de la moda es el cuantil 0.5, pues se está representando una distribución gaussiana.

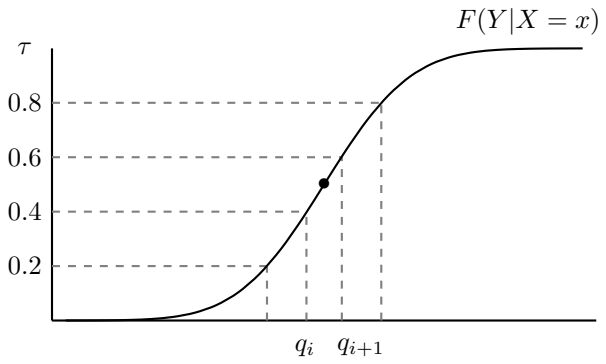


Figura 2.9: Representación de la estimación de la moda en el caso de la distribución Normal.

En otras palabras, la idea consiste en estimar la moda o las modas como el punto medio del intervalo de menor distancia entre dos cuantiles consecutivos definidos de tal manera que acumulen la misma cantidad de probabilidad. Como ejemplo, se podría considerar una rejilla formada por intervalos del 5 % de probabilidad para la distribución condicional. En este caso, dado que dos cuantiles condicionales consecutivos acumulan una cantidad fija de probabilidad, la moda puede ser estimada como el punto medio del intervalo más corto.

Cabe destacar que en distribuciones continuas, las modas se localizan en los puntos de inflexión, donde el ratio de acumulación de probabilidad cambia de grandes acumulaciones a acumulaciones cada vez más pequeñas. En este sentido, el intervalo definido por la secuencia de los cuantiles será cada vez más corto hasta que se alcance la moda. Cuando los cuantiles pasan la moda, los intervalos entre cuantiles serán cada vez más grandes a medida que nos alejamos. Como ya se había adelantado, se pueden estimar las modas en los mínimos de la función que representa la amplitud de estos intervalos.

En la Figura 2.10 se representan las distancias teóricas entre los q_{i+1} y los q_i (para un número suficientemente grande de cuantiles estimados) frente a la distribución condicional teórica. En el caso de existir más de una moda condicional, la distancia entre cuantiles volvería a disminuir hasta alcanzar otro mínimo local.

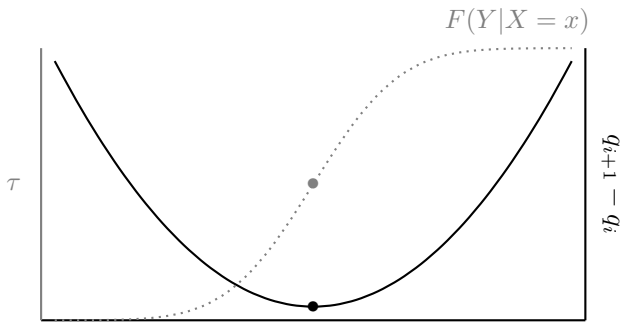


Figura 2.10: Representación de las distancias entre los q_i y q_{i+1} (en negro) frente a la función de distribución teórica (en gris).

Las ventajas[2] de la propuesta de estimación modal planteada son las siguientes:

- Las estimaciones de los cuantiles recuperan la función de distribución condicional.
- Desde un punto de vista teórico se puede aplicar a distribuciones multimodales.
- No es necesario fijar ningún tipo de bandwidth, eliminando los problemas habituales de selección de parámetro ventana.
- Los cuantiles pueden ser estimados tanto de forma paramétrica como no paramétrica.
- Utilizando regresión cuantil paramétrica es posible analizar cómo cambios en las variables afectan a la estimación modal. Además, permite que el método sea escalable a un número elevado de covariables.
- Esta propuesta es computacionalmente mucho más eficiente que las alternativas consideradas, pues no requiere del uso de selectores de ventanas.

2.4.2. Análisis del estimador

La función cuantil condicional puede ser definida como

$$Q_q(y|X=x) = \text{ínf}\{y : F(y|X=x) \geq q\},$$

esto es, el punto en el rango de la función de distribución condicional que acumula al menos probabilidad q. A su vez, el cuantil condicional puede ser definido como el valor de y que verifica

$$F\left(Q_q(y|x)\right) = P\left(Y \leq Q_q(y|x)\right) = q. \tag{2.22}$$

Nótese que (2.22) solo es aplicable si Y es una variable aleatoria continua.

Se puede definir una secuencia de funciones de cuantiles condicionales para probabilidades

$$q_j = q_{j-1} + \alpha,$$

[2]Nótese que varias de estas ventajas son comunes al método de estimación estudiado en la Sección 2.3.

con $j = 1, \ldots, J$, $q_0 = 0$ y $J = 1/\alpha$, esto es, $q_j = j\alpha$. Por tanto, se tiene que la probabilidad entre dos cuantiles sucesivos es el valor α fijado,

$$P\left(Q_{j-1}(y|x) < Y \leq Q_j(y|x)\right) = F\left(Q_j(y|x)\right) - F\left(Q_{j-1}(y|x)\right) = \alpha, \tag{2.23}$$

para $j = 1, \ldots, J$. Variando j, la probabilidad entre dos cuantiles consecutivos se fija en α y la única modificación es la distancia entre los cuantiles definidos por $Q_j(y|x) - Q_{j-1}(y|x)$. Siguiendo Chernoff (1964), las modas se encuentran en el intervalo más corto que suma un número determinado de observaciones del total de las mismas. Con el objetivo de obtener una primera intuición del estimador, dado (2.23), asumiendo que Y es continua y expandiendo por Taylor $F(Q_j(y|x))$ en $Q_{j-1}(y|x)$ se tiene

$$F(Q_j(y|x)) = F\left(Q_{j-1}(y|x)\right) + F'\left(Q^*_{j-1}(y|x)\right)\left(Q_j(y|x) - Q_{j-1}(y|x)\right),$$

por lo que despejando

$$f\left(Q^*_{j-1}(y|x)\right)\left(Q_j(y|x) - Q_{j-1}(y|x)\right) = \alpha, \tag{2.24}$$

donde $Q^*_{j-1}(y|x)$ es un punto entre $Q_j(y|x)$ y $Q_{j-1}(y|x)$ para $j = 1, \ldots, n$. La expresión (2.24) puede ser reescrita como

$$\left(Q_j(y|x) - Q_{j-1}(y|x)\right) = \frac{\alpha}{f\left(Q^*_{j-1}(y|x)\right)} = \alpha s(q^*_{j-1}),$$

donde

$$s(q^*_{j-1}) = \frac{1}{f\left(Q^*_{j-1}(y|x)\right)} = \left[f\left(F^{-1}(q^*_{j-1}|x)\right)\right]^{-1},$$

es la función sparsity evaluada en el punto medio $Q^*_{j-1}(y|x)$.

Si $Q^*_{j-1}(y|x)$ fuese la moda, entonces habría muchas observaciones cerca de ese cuantil forzando a $s(q^*_{j-1})$ a ser pequeña. En consecuencia, el intervalo definido por $(Q_j(y|x) - Q_{j-1}(y|x))$ será pequeño. En otras palabras, dado que la función sparsity mide la tasa del cambio de la función cuantil,

$$s(q) = \frac{\partial}{\partial q} F^{-1}(q|x), \tag{2.25}$$

los cuantiles cambiarán muy lentamente, esto es, el intervalo entre ellos será pequeño en las regiones donde hay datos abundantes y viceversa.

Una vez presentado el procedimiento de estimación, en teoría parece muy sencillo estimar el punto o puntos de inflexión de las distribuciones condicionales para cada $x \in X$ y por tanto las modas condicionales. Sin embargo, en la práctica la distribución que se obtiene no es suave, por lo que es necesario aplicar un algoritmo en dos pasos similar al procedimiento propuesto en Bickel y Frühwirth (2006), pero con carácter condicional, dividiendo primero los cuantiles en una secuencia "pequeña", por ejemplo 4 cuantiles $\tau \in (0.05, 0.95)$, de modo que la primera secuencia sería $\tau = (0.05, 0.35, 0.65, 0.95)$ y de estos se calcula el intervalo de mínima amplitud. Una vez calculado dicho intervalo se vuelve a seleccionar una secuencia mayor de cuantiles, por ejemplo 25 en ese intervalo y se vuelve a calcular el más pequeño. La moda se encontrará en la media de este último intervalo. Así, podemos reducir la variabilidad de las estimaciones y mejorarlas sin necesidad de suavizar las distribuciones empíricas.

Cabe destacar que aunque en este procedimiento no es necesario seleccionar una ventana de suavizado, sigue siendo necesario calcular la secuencia inicial "óptima" de cuantiles, ya que una selección inadecuada podría dar lugar malas estimaciones. Además, el carácter multimodal del procedimiento solo es posible en teoría, pues en la práctica no es sencillo calcular más de un punto de inflexión, por lo que esta metodología no debería aplicarse más que al caso unimodal. Los pasos para llevar a cabo este procedimiento se encuentran en el Algoritmo 3.

Algoritmo 3 : *Estimación modal a través de la distribución condicional*

1: Definir una secuencia con un número pequeño de cuantiles equiespaciados.
2: Para $x \in X$:
Ajustar un modelo de regresión cuantil y realizar la predicción en x.
Calcular el intervalo de cuantiles con extremos q_{i+1} y q_i más pequeño.
Volver a considerar una secuencia de cuantiles con más puntos que la anterior, cuyo mínimo y máximo son los extremos inferior y superior del intervalo calculado en el paso anterior.
Ajustar un modelo de regresión cuantil para los nuevos cuantiles y realizar la predicción en x.
Calcular el τ_{opt} como el punto medio del nuevo intervalo con mínima amplitud.
Ajustar el modelo para τ_{opt} y obtener la estimación $\widehat{m}(x)$ de $m(x)$.
3: Las estimaciones serán aquellas obtenidas en los τ_{opt} en cada $x \in X$ considerado.

2.4.3. Extensión no lineal y futuras líneas de investigación

Aunque no entraremos en más detalles, sería posible extender el procedimiento de estimación al caso no lineal mediante el procedimiento presentado, sustituyendo el modelo de regresión cuantil lineal por una versión no paramétrica (tipo núcleo, lineal local, entre otros). Un modelo de regresión cuantil no paramétrico puede ser representado como

$$Y = f_\tau(X) + \varepsilon,$$

donde Y es la variable respuesta, X la covariable y f_τ es una función de regresión cuantil de orden τ con $\tau \in (0, 1)$. Además, se asume que para un $\tau \in (0, 1)$ fijo, el cuantil τ-ésimo de ε es cero.

Aplicando regresión cuantil no paramétrica podríamos ajustar modelos no lineales, pero sería necesario suavizar la estimación obtenida, considerando una nueva ventana. Esto se debe a que las estimaciones obtenidas mediante los métodos que utilizan regresión cuantil no son, por lo general, suaves. Otra alternativa sería suavizar las distribuciones condicionales. En lo que resta de este trabajo no se considerará esta extensión, pero podría ser interesante como objeto de futuras investigaciones. A modo ilustrativo, en la Figura 2.11 se muestra un ejemplo de estimación de unos datos simulados no lineales. La regresión modal se ha realizado mediante regresión cuantil tipo núcleo con pesos normales. En rojo se muestra la función modal teórica, en verde la estimación sin suavizado y en azul la suavizada.

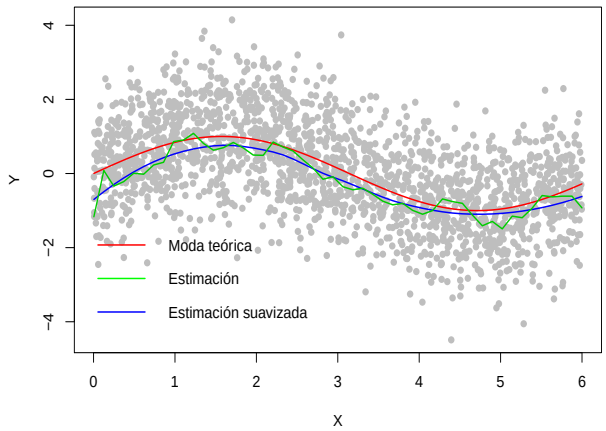


Figura 2.11: Conjuntos de predicción uniformes para regresión modal.

Por otra parte, en futuros trabajos también se espera desarrollar métodos para la obtención de intervalos de confianza del método de estimación, tanto para el propuesto en esta memoria como su posible extensión.

2.5. Software disponible

Existen pocas implementaciones disponibles para aplicar regresión modal. Entre ellas destacamos los siguientes paquetes de `R`, publicados en el repositorio CRAN[3]:

hdrcde (Hyndman et al., 2018)

La función `modalreg` de este paquete permite calcular las modas condicionales mediante el método propuesto por Einbeck y Tutz (2006) utilizando kernels gaussianos. Además, permite el cálculo automático de ventanas óptimas mediante distintos métodos propuestos en Bashtannyk y Hyndman (2001) con la función `cde.bandwidths`.

lpme (Zhou y Huang, 2017)

La función `modereg` permite calcular la regresión modal mediante el método de Einbeck y Tutz (2006) en el caso en el que existan o no errores de medida en las variables explicativas (Zhou y Huang, 2016). Permite el cálculo de ventanas mediante los selectores propuestos en Zhou y Huang (2016) y Zhou y Huang (2019) con las funciones `moderegbwSIMEX` y `moderegbw` respectivamente.

Cabe destacar que se han realizado una serie de implementaciones nuevas en `R`. La función `modalreg` del paquete `hdrcde` se ha mejorado realizando las siguientes modificaciones:

- Cálculo de intervalos de confianza bootstrap propuestos en Chen et al. (2016).
- La opción para elegir entre distintos tipos de kernels.
- Implementación de un criterio de parada del algoritmo mean-shift, mejorando así su velocidad computacional.
- Selección manual de puntos de arranque del algoritmo, junto a las opciones por defecto.

Además, se han creado dos nuevas funciones que permiten realizar las estimaciones de los métodos de estimación modal estudiados en las Secciones 2.3 y 2.4.

En los capítulos siguientes se analizan tres métodos de estimación de modas condicionales que serán comparados en un estudio de simulación y aplicados al datos económicos. Los métodos elegidos para analizar en esta memoria son el no paramétrico (Einbeck y Tutz, 2006), el basado en regresión cuantil (Ohta et al., 2018) y una nueva propuesta de estimación.

[3]CRAN es una red de servidores FTP y web de todo el mundo que almacena versiones de código y documentación actualizadas e idénticas para R.

Capítulo 3

Resultados de simulación

En este capítulo se realiza un estudio de simulación con el objetivo de comparar el comportamiento de los métodos de regresión modal estudiados en el capítulo anterior. Se evaluarán los distintos procedimientos de estimación en escenarios cada vez más complejos, que se irán presentando gradualmente. Se tendrán en cuenta dos posibles escenarios: el primero en el que la función modal es lineal y el segundo en el que ésta no lo es. En este segundo caso se considerará un modelo lineal cuyos residuos forman la no linealidad de la función modal y un modelo no lineal. Además, se han realizado estudios para evaluar el comportamiento de los distintos modelos frente a cambios en las ventanas. Los puntos de diseño fijo (grid) de estimación considerado serán una secuencia de tamaño 50 desde el mínimo hasta el máximo de la covariable.

La medida de error para comparar los distintos procedimientos será la Raíz del Error Cuadrático Medio (también conocida como RMSE por sus siglas en inglés). Para M réplicas de Monte Carlo y una rejilla formada por T puntos de estimación, se pueden estimar la media o la mediana de los RMSE. Para la media se tiene

$$\overline{\text{RMSE}} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{m}_j(x_t) - m(x_t)\right)^2},$$

donde $\widehat{m}_j(x_t)$ representa la estimación de $m(x_t)$ para la simulación j en el punto de estimación x_t. De manera análoga se calculará la mediana de los RMSE.

En el estudio se ha considerado $M = 1000$ y distintos tamaños muestrales ($n = 500, n = 1000$ y $n = 2000$).

Para el método que aplica el mean-shift se han empleado kernels gaussianos para estimar (2.13) y se han calculado las ventanas mediante el procedimiento bootstrap descrito en Bashtannyk y Hyndman (2001).

Para el procedimiento de estimación desarrollado en Ohta et al. (2018) se ha utilizado el parámetro $\epsilon = 0.1$, una secuencia de 100 cuantiles equiespaciados tales que $\tau \in (0.05, 0.95)$ y la regla de selección del parámetro h para cada $X = x$ descrita en el artículo, donde se menciona que "aunque no es óptima, funciona razonablemente bien".

Por último, en el método propuesto se ha utilizado una secuencia inicial de 4 cuantiles para el primer paso y una secuencia de 25 para el segundo.

Por simplicidad se denotará al método analizado en la Sección 2.2 por "MS", al estudiado en la Sección 2.3 por "Densidad" y al propuesto en la Sección 2.4 por "Distribución".

3.1. Modelos con función modal lineal

En esta sección trabajaremos con casos particulares del modelo

$$Y = X'\beta + \sigma(x)v\varepsilon,$$

donde β es un vector de pendientes, X la variable independiente, Y la variable dependiente (univariante), $\sigma(x)$ es un término de error que depende del valor de la covariable, v una constante y ε es un término de error independiente de la covariable. Este modelo genérico permitirá modificar los distintos términos y evaluar los métodos bajo el mismo modelo con distintas generaciones de error. En primer lugar se considerará un modelo homocedástico y posteriormente lo compararemos con una modificación del mismo pero con errores heterocedásticos. De esta forma, si $\sigma(x) = v = 1$ y $\beta = 3$ podemos construir el sencillo modelo

$$Y = 1 + 3X + \varepsilon, \qquad \text{(Modelo 1)}$$

donde $X \sim U(0,1)$ y $\varepsilon \sim N(0,1)$. En este caso $\mathbb{E}(\varepsilon) = 0$ y $\text{Moda}(\varepsilon) = 1$, por lo que la función modal tiene la forma $m(X) = 1 + 3X$. Además, media, moda y mediana condicionales coinciden, ya que los errores son homocedásticos.

Por otra parte, si $\sigma(x) = 1 + 2X$, $v = 1$ y $\varepsilon = 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ se tiene el modelo descrito en el estudio de simulación de Yao y Li (2014), que tiene la forma

$$Y = 1 + 3X + \sigma(X)\varepsilon, \qquad \text{(Modelo 2)}$$

donde $X \sim U(0,1)$. En este caso, $\mathbb{E}(\varepsilon) = 0$, $\text{Moda}(\varepsilon) = 1$ y $\text{Mediana}(\varepsilon) = 0.67$. Las funciones de regresión condicionales son, respectivamente, $\mathbb{E}(Y|X) = 1 + 3X$, $\text{Moda}(Y|X) = 2 + 5X$ y $\text{Mediana}(Y|X) = 1.67 + 4.34X$. Los residuos de la regresión modal son $Y - \text{Moda}(Y|X) = (1 + 2X)(\varepsilon - 1)$, cuya distribución tiene moda cero pero es heterocedástica. Como se puede apreciar, este modelo se ha construido a partir del primero pero modificando la distribución de los errores.

Los resultados de la media y mediana de las raíces de los errores cuadráticos medios para las ventanas óptimas[1] se muestran en las Tablas 3.1 y 3.2, destacando en negrita (en la presente y siguientes tablas de este capítulo) el valor más pequeño para cada tamaño muestral en cada modelo.

n	Densidad		Distribución		MS	
	Media	Mediana	Media	Mediana	Media	Mediana
500	0.386	0.372	**0.233**	**0.226**	0.256	0.243
1000	0.351	0.333	**0.211**	**0.213**	0.219	0.215
2000	0.311	0.299	0.204	0.204	**0.193**	**0.188**

Tabla 3.1: Resultados de simulación para el modelo lineal homocedástico.

[1] Con ventanas óptimas nos referimos a las calculadas mediante las reglas de selección mencionadas.

n	Densidad		Distribución		MS	
	Media	Mediana	Media	Mediana	Media	Mediana
500	3.834	3.636	2.979	2.897	**2.643**	**2.652**
1000	3.665	3.541	2.934	2.874	**2.613**	**2.584**
2000	3.620	3.571	2.948	2.871	**2.561**	**2.535**

Tabla 3.2: Resultados de simulación para el modelo lineal heterocedástico.

En los resultados anteriores se puede apreciar que para el modelo más sencillo, el método de estimación que comete menor error es el que proponemos (excepto para $n = 2000$), con una notable diferencia sobre el otro método basado en regresión cuantil. Sin embargo, para el modelo con heterocedasticidad, el error más pequeño lo comete el método no paramétrico, mientras que el propuesto en Ohta et al. (2018) es el que presenta los peores resultados. Además, se puede apreciar que una ligera modificación en la generación de los residuos provoca que los métodos presenten un comportamiento más errático.

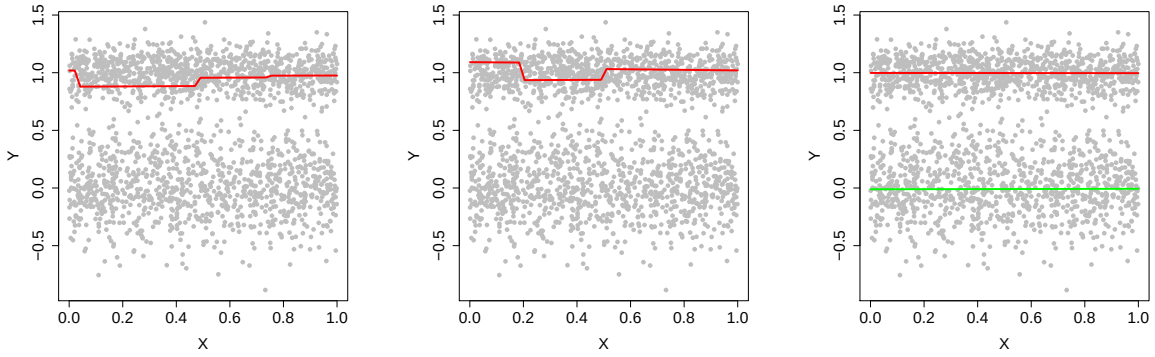
Por otra parte, con el objetivo de estudiar el comportamiento de los procedimientos de estimación bajo cambios en los parámetros se han calculado las medidas de error para un grid de ventanas y tamaño muestral $n = 2000$ para los métodos presentados en las Secciones 2.3 y 2.2. Los resultados se muestran en la Tabla 3.3. En color gris se resaltan los errores con selección de ventanas "óptimas" mostrados en las Tablas 3.1 y 3.2. Nótese que no se incluyen los valores de las ventanas ya que para el primero de los casos se tiene una ventana óptima para cada punto de diseño de cada simulación, mientras que para el segundo caso existen dos ventanas óptimas para cada una de las simulaciones.

	Densidad					MS	
n	h	Modelo 1	Modelo 2	h	g	Modelo 1	Modelo 2
2000	$0.25 \cdot h_{opt}$	0.425	4.028	$0.25 \cdot h_{opt}$	$0.25 \cdot g_{opt}$	0.850	2.799
	$0.90 \cdot h_{opt}$	0.337	3.783	$0.90 \cdot h_{opt}$	$0.90 \cdot g_{opt}$	0.193	2.561
	$0.95 \cdot h_{opt}$	0.320	3.662	$0.95 \cdot h_{opt}$	$0.95 \cdot g_{opt}$	0.207	2.567
	h_{opt}	0.311	3.620	h_{1opt}	h_{2opt}	0.193	2.561
	$1.05 \cdot h_{opt}$	0.316	3.590	$1.05 \cdot h_{opt}$	$1.05 \cdot g_{opt}$	0.181	**2.559**
	$1.10 \cdot h_{opt}$	0.317	3.597	$1.10 \cdot h_{opt}$	$1.10 \cdot g_{opt}$	**0.171**	2.560
	$4 \cdot h_{opt}$	**0.269**	**3.480**	$4 \cdot h_{opt}$	$4 \cdot g_{opt}$	0.301	2.707

Tabla 3.3: Media de los RMSE para distintas ventanas en los modelos con función modal lineal.

Como se puede apreciar en la Tabla 3.3, en el método de Ohta et al. (2018) a medida que incrementamos el tamaño de las ventanas se reduce el error de estimación. Dado el extraño comportamiento que presenta se ha comprobado que el error comienza a aumentar para ventanas aproximadamente 20 veces superior a las calculadas según la regla de selección del artículo. Se hace evidente que la regla de selección del parámetro ventana no es adecuada, como ya se mencionó en el artículo. En lo referente a las ventanas para el otro procedimiento se observa un comportamiento adecuado, pues para valores muy alejados de los calculados según la regla de selección utilizada los errores se incrementan considerablemente.

Por último en esta sección, se considera a modo ilustrativo un modelo bimodal con el objetivo de evaluar el comportamiento de los distintos métodos cuando existe más de una moda. Los datos considerados son una mixtura de dos normales con $X = (X_1, X_2)$ donde $X_1 \sim N_1(1, 0.12)$ y $X_2 \sim N_2(0, 0.15)$. Por tanto, las modas teóricas se encuentran en $Y = 1$ y en $Y = 0$. Como consecuencia de la menor desviación típica de X_1, la moda en $Y = 1$ forma el máximo global. en la Figura 3.1 se muestran los ajustes mediante los distintas formas de estimación.



(a) (b) (c)

Figura 3.1: Ajustes para el mismo conjunto de datos. En (a) para el método de Ohta, en (b) para el que proponemos y en (c) para el no paramétrico.

Si entrar en más detalles, el método de la Sección 2.2 permite el ajuste de las dos modas, mientras que los otros dos métodos ajustan la moda que corresponde con los máximos globales de las funciones de densidad.

3.2. Modelos con función modal no lineal

Para comparar los métodos en escenarios más complejos que los descritos en la sección anterior se considera el modelo utilizado en el estudio de simulación de Ohta et al. (2018)

$$Y = U^3/3 - X(U-1)^2, \qquad \text{(Modelo 3)}$$

donde $X \sim U(0,1)$ y $U \sim U(0,1)$ independiente de X. En este se caso la función cuantil condicional es lineal de la forma $Q_\tau(X) = \tau^3/3 - X(\tau-1)^2$ mientras que la función modal es no lineal con la forma $m(X) = -2X^3/3 + 2X^2 - X$.

Por último se considera el modelo no lineal

$$Y = \sin(X) + \varepsilon, \qquad \text{(Modelo 4)}$$

donde $X \sim U(0,6)$ y $\varepsilon \sim N(0,1)$. En este caso la función modal es no lineal y tiene la forma $m(X) = \sin(X)$.

Los resultados de las simulaciones realizadas para estos dos modelos se muestran en las Tablas 3.4 y 3.5.

n	Densidad		Distribución		MS	
	Media	Mediana	Media	Mediana	Media	Mediana
500	0.208	0.190	**0.141**	**0.125**	0.146	0.135
1000	0.193	0.180	**0.121**	**0.108**	0.139	0.130
2000	0.173	0.163	**0.106**	**0.097**	0.130	0.123

Tabla 3.4: Resultados de simulación para el modelo lineal y función modal no lineal.

n	Densidad		Distribución		MS	
	Media	Mediana	Media	Mediana	Media	Mediana
500	0.605	0.585	0.501	0.494	**0.278**	**0.274**
1000	0.572	0.551	0.489	0.483	**0.235**	**0.231**
2000	0.546	0.534	0.482	0.476	**0.203**	**0.200**

Tabla 3.5: Resultados de simulación para el modelo no lineal y función modal no lineal.

Por una parte, en cuanto a los resultados para el Modelo 3, el mejor comportamiento lo presenta de nuevo el método presentado en la Sección 2.4. El procedimiento de estimación de Einbeck y Tutz (2006) es el que obtiene los segundos mejores resultados, mientras que el estimador de Ohta presenta el peor comportamiento.

Por otra parte, para el Modelo 4, el mejor ajuste es el que se obtiene mediante el procedimiento no paramétrico de Einbeck y Tutz (2006). Este último resultado se debe a que el modelo de regresión cuantil lineal en el que se basan las otros dos métodos de estimación no es capaz de adaptarse adecuadamente a la no linealidad del modelo. De todas formas, nuestro método obtiene menores errores que el de Ohta et al. (2018). En este tipo de modelos no lineales, para obtener resultados competitivos sería necesario comparar el mean-shift la extensión no lineal del método propuesto introducido en la Sección 2.4 del Capítulo 2.

Al igual que se realizó en la sección previa, se han realizado simulaciones para distintas ventanas. En este caso las conclusiones que se derivan de este estudio son equivalentes a los obtenidos en la sección anterior. Los resultados se pueden ver en la Figura 3.6.

	Densidad			MS			
n	h	Modelo 3	Modelo 4	h	g	Modelo 3	Modelo 4
2000	$0.25 \cdot h_{opt}$	0.211	0.631	$0.25 \cdot h_{1opt}$	$0.25 \cdot h_{2opt}$	0.146	0.691
	$0.90 \cdot h_{opt}$	0.177	0.559	$0.90 \cdot h_{1opt}$	$0.90 \cdot h_{2opt}$	0.132	0.230
	$0.95 \cdot h_{opt}$	0.175	0.549	$0.95 \cdot h_{1opt}$	$0.95 \cdot h_{2opt}$	0.131	0.216
	h_{opt}	0.173	0.546	h_{1opt}	h_{2opt}	0.130	0.203
	$1.05 \cdot h_{opt}$	0.169	0.547	$1.05 \cdot h_{1opt}$	$1.05 \cdot h_{2opt}$	0.129	0.192
	$1.10 \cdot h_{opt}$	0.166	0.546	$1.10 \cdot h_{1opt}$	$1.10 \cdot h_{2opt}$	**0.128**	**0.182**
	$4 \cdot h_{opt}$	**0.143**	**0.521**	$4 \cdot h_{1opt}$	$4 \cdot h_{2opt}$	0.169	0.258

Tabla 3.6: Media de los RMSE para distintas ventanas en los modelos con función modal no lineal.

Por otra parte, dado el comportamiento que presentan las ventanas del método de Ohta et al. (2018), se ha estudiado su comportamiento para el caso en el que se utilizase la ventana de Koenker y Machado (1999) con el objetivo de estudiar si la modificación de la regla de selección presentada en el artículo carece o no de sentido. En la Tabla 3.7 se presentan los resultados obtenidos.

n	Modelo 1		Modelo 2		Modelo 3		Modelo 4	
	Media	Mediana	Media	Mediana	Media	Mediana	Media	Mediana
500	0.501	0.490	4.688	4.573	0.232	0.211	0.706	0.697
1000	0.449	0.446	4.538	4.365	0.221	0.204	0.674	0.661
2000	0.414	0.405	4.285	4.178	0.202	0.187	0.621	0.607

Tabla 3.7: Medias de los RMSE en el método de Ohta et al. (2018) para los distintos modelos y tamaños muestrales considerando como ventana la propuesta en Koenker y Machado (1999)

Como se puede apreciar, los errores cometidos con esta ventana son mayores que con la modificación de la regla de selección original. De todas formas, sería necesario determinar una regla de selección óptima para el método de Ohta, o una que presente un mejor comportamiento ante cambios en las ventanas y evaluar de nuevo su rendimiento frente a las alternativas.

Además, aunque nuestro método no necesite de una selección de ventana como tal, es necesario hacer una selección inicial del tamaño de la secuencia de cuantiles iniciales que actúan en cierta medida como un parámetro ventana. Nótese que una secuencia inicial de tamaño dos es equivalente a aplicar el algoritmo en un único paso. Los resultados se muestran en la Tabla 3.8.

Secuencia inicial	n	Modelo 1	Modelo 2	Modelo 3	Modelo 4
2	500	0.390	3.680	0.172	0.625
	1000	0.319	3.347	0.150	0.569
	2000	0.273	3.170	0.123	0.530
4	500	**0.233**	**2.979**	**0.141**	**0.501**
	1000	**0.211**	**2.934**	**0.121**	**0.489**
	2000	**0.204**	**2.948**	**0.106**	**0.482**
6	500	0.311	3.371	0.148	0.555
	1000	0.266	3.193	0.126	0.528
	2000	0.215	3.005	0.108	0.496
8	500	0.313	3.291	0.149	0.558
	1000	0.270	3.126	0.128	0.526
	2000	0.234	3.037	0.111	0.503
10	500	0.327	3.352	0.151	0.568
	1000	0.281	3.172	0.133	0.533
	2000	0.236	3.039	0.112	0.503
12	500	0.333	3.377	0.156	0.573
	1000	0.285	3.188	0.135	0.536
	2000	0.234	3.062	0.114	0.506

Tabla 3.8: Medias de los RMSE de los modelos para distintos tamaños de secuencia de cuantiles iniciales tales que $\tau \in (0.05, 0.95)$ en el método de estimación propuesto.

En este último estudio se observa que los errores más pequeños son aquellos en los que se ha utilizado la secuencia de tamaño 4, que corresponde con los cuantiles iniciales $\tau = (0.05, 0.35, 0.65, 0.95)$. A su vez se puede apreciar que a medida que incrementamos el tamaño de la secuencia inicial se incrementan los RMSE medios para los distintos modelos. A falta de determinar una posible secuencia de cuantiles óptima o una regla de selección que funcione razonablemente bien, en base a los resultados se recomienda utilizar una secuencia inicial de tamaño 4 para calcular las estimaciones.

Capítulo 4

Diferencias salariales entre hombres y mujeres en España

En este capítulo se analiza la evolución de las brechas salariales de género usando los datos de distintas encuestas de estructura salarial española a través de las diferencias en moda, media y mediana.

La *discriminación laboral* ocurre cuando dos trabajadores con capacidades similares en términos de productividad son tratados de forma diferente por tener características personales distintas, como pueden ser el género, la raza, la edad, la nacionalidad o la orientación sexual (Arrow, 1971). La discriminación no solo genera efectos sociales negativos, sino que produce pérdidas de eficiencia económica y una mala asignación de los recursos.

En este capítulo nos centraremos en la *brecha salarial*, que se define como la diferencia relativa en el ingreso bruto de mujeres y hombres en una economía. En las últimas décadas se ha producido un incremento del interés de los economistas por analizar la creciente desigualdad salarial en los mercados laborales. Dicho interés se debe a que las diferencias salariales entre hombres y mujeres fueron disminuyendo hasta que se produjo un cambio de tendencia a partir de 1980.

Es necesario destacar que las diferencias salariales son comparaciones entre colectivos de distintas características. Como tal, su comparación no debe entenderse como una interpretación causal. No podemos asumir de antemano que las diferencias salariales tengan un carácter discriminatorio. Es por ello que se suele considerar la brecha salarial no ajustada y ajustada. La primera indica de manera descriptiva las diferencias entre dos grupos sin tener en cuenta las disimilitudes subyacentes entre ellos, mientras que la segunda busca determinar las diferencias tendiendo en cuenta dichos factores. De todas formas, nosotros nos centraremos en la brecha salarial no ajustada.

La determinación de los salarios se realiza en la teoría económica neoclásica a través de la teoría de la productividad marginal. El empresario busca maximizar sus beneficios, resolviendo un problema de optimización del que se deriva que el salario real que percibe el trabajador es igual a su productividad marginal. Nótese que para determinar los salarios de una economía sería preciso conocer los precios para cada empresa. Para obtener los salarios reales lo más habitual es deflactar los precios mediante un índice. Como los salarios son precios, pueden estar expresados en términos nominales (a precios corrientes), si llevan incorporado el efecto de la inflación, o en términos reales (a precios constantes), si vienen corregidos por el efecto de la inflación. El proceso de convertir precios corrientes a constantes se denomina deflactar o actualizar, dependiendo de la dirección de la conversión. Para evitar que los resultados del análisis vengan influenciados por el efecto de la inflación, los datos pueden ser actualizados o deflactados para estar expresados en moneda de igual poder adquisitivo. En este capítulo

realizaremos este procedimiento porque agregaremos distintos datos y los estudiaremos, pero ello no quiere decir que esta conversión sea la más adecuada desde un punto de vista económico.

El objetivo de este capítulo es analizar las diferencias salariales y la toma de decisiones en el mercado laboral desde el punto de vista de la moda. Con tal objetivo, en la Sección 4.1 se revisan los distintos procedimientos que existen en la literatura para determinar las diferencias salariales, en la Sección 4.2 se describen con detalle los datos que se utilizarán para el análisis, en la Sección 4.3 se muestran los resultados sobre las diferencias salariales y se derivan conclusiones. Además, se analiza el impacto de la crisis de 2008 en el salario bruto medio y modal para distintas horas semanales trabajadas diferenciando por sexo. Para ello estimaremos las medias, modas y medianas condicionales para los datos considerados.

Para estimar las modas se utilizará el método no paramétrico presentado en la Sección 2.2, ya que es el modelo más flexible y trabajaremos con una única variable explicativa. Además, permite compararlo con media y medianas no paramétricas. Las ventanas necesarias para la estimación modal han sido calculadas mediante el procedimiento bootstrap de Bashtannyk y Hyndman (2001). En cuanto a la estimación de la media, se ha utilizado la regresión Nadaraya-Watson. En este caso la ventana ha sido calculada utilizando la función `h.select` de la librería `sm` (Bowman y Azzalini, 2019). Por último, los cuantiles se han estimado mediante regresión cuantil tipo núcleo. Todos los kernels utilizados para los distintos procedimientos han sido gaussianos.

4.1. Procedimientos para determinar las diferencias salariales

Para analizar si existen diferencias salariales se debe determinar si las funciones de distribución de los grupos considerados dada una serie de características son iguales o no. En el caso de que no existiesen diferencias las distribuciones serían iguales, esto es

$$F(W_0|X) = F(W_1|X), \tag{4.1}$$

donde $i = 0, 1$ indica el género o grupo. A su vez, las características pueden ser observables (como el sexo, el nivel de estudios o la edad) o inobservables (como la motivación, el esfuerzo o la competitividad). Sean X_o las características observables y X_v las inobservables, (4.1) puede reescribirse como

$$F(W_0|X_o, X_v) = F(W_1|X_o, X_v).$$

En la Figura 4.1 se muestra una representación de las funciones de distribución para cada género en la que existen diferencias salariales.

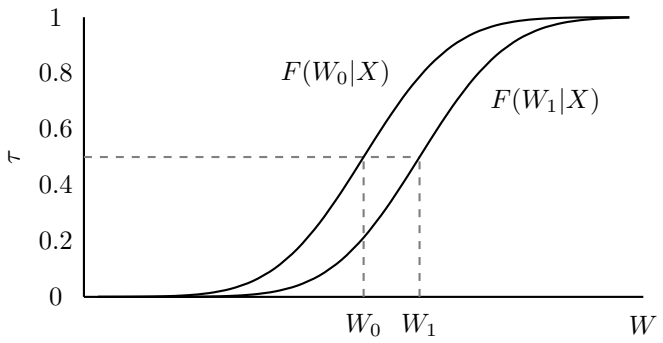


Figura 4.1: Representación de dos funciones de distribución salarial.

La aproximación más habitual para analizar las diferencias salariales es la descomposición de Oaxaca-Blinder (Oaxaca, 1973; Blinder, 1973), que descompone las diferencias en medias. El modelo se asume lineal y separable entre características observables e inobservables con la forma

$$W_g = X'\beta_g + \varepsilon_g \quad \text{para} \quad g = 0, 1,$$

donde W son los salarios, X un vector de características y $\mathbb{E}(\varepsilon|X) = 0$. Si denotamos por G a una variable indicadora de género, las diferencias salariales en media Δ^μ pueden representarse como

$$\begin{aligned}\Delta^\mu &= \mathbb{E}(W|G=0) - \mathbb{E}(W|G=1) \\ &= \mathbb{E}(\mathbb{E}(W|X, G=0)|G=0) - \mathbb{E}(\mathbb{E}(W|X, G=1)|G=1) \\ &= \mathbb{E}(X|G=1)'\beta_1 + \mathbb{E}(\varepsilon_1|G=1) - \left[\mathbb{E}(X|G=0)'\beta_0 + \mathbb{E}(\varepsilon_0|G=1)\right],\end{aligned}$$

donde $\mathbb{E}(\varepsilon_g|G=g) = 0$ para $g = 0, 1$. Por tanto la descomposición salarial se reduce a

$$\begin{aligned}\Delta^\mu &= \mathbb{E}(X|G=1)'\beta_1 - \mathbb{E}(X|G=0)'\beta_0 \\ &= \mathbb{E}(X|G=1)'(\beta_1 - \beta_0) + \left[\mathbb{E}(X|G=1) - \mathbb{E}(X|G=0)\right]'\beta_0.\end{aligned}$$

La idea consiste en determinar en qué medida el efecto de la discriminación se debe a las características o a las diferencias entre los precios salariales, esto es, los coeficientes asociados a cada grupo, $(\beta_1 - \beta_0)$. Su mayor inconveniente es que los coeficientes vienen determinados por las variables, que pueden tener errores de medida o haber sido seleccionadas de manera incorrecta.

Otra aproximación habitual consiste en determinar las diferencias cuantílicas Δ^Q (Machado y Mata, 2005) que toman la forma

$$\Delta^Q = Q(W_0|X) - Q(W_1|X).$$

En este caso la descomposición de las diferencias no resulta tan simple como en el caso de la media, por lo que no entraremos en más detalles sobre ella.

En este estudio, en lugar de descomponer los salarios se considerarán individuos con características homogéneas y se analizarán las diferencias en modas Δ^m calculadas de manera no paramétrica para distintas horas trabajadas. Sean X las horas de trabajo semanal la diferencia en modas toma la forma

$$\Delta^m = \text{Moda}(W_0|X) - \text{Moda}(W_1|X).$$

4.2. Datos

Para el estudio se han utilizado los datos de distintas *encuestas cuatrienales de estructura salarial*[1], ya que es la encuesta salarial con más relevancia a nivel nacional y europeo. Se realiza desde 1995 en el marco de la Unión Europea empleando criterios comunes de metodología y contenidos. El fin es obtener unos resultados comparables sobre la estructura y distribución de los salarios entre los Estados miembros, considerando distintas variables como el sexo, la ocupación, o el tamaño de la empresa.

En dichas encuestas se computan los devengos brutos, es decir, los rendimientos dinerarios antes de haber practicado las deducciones de las aportaciones a la Seguridad Social por cuenta del trabajador o las retenciones a cuenta del Impuesto sobre la Renta de las Personas Físicas (IRPF). A continuación se muestran de manera esquemática las características de las encuestas:

- **Ámbito poblacional**: está formado por todos los trabajadores por cuenta ajena que presten sus servicios en centros de cotización, independientemente del tamaño de los mismos, y hayan estado de alta en la Seguridad Social durante todo el mes de octubre del año de referencia.

[1] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&secc=1254736195110&idp=1254735976596

- **Ámbito geográfico**: todo el territorio nacional y desagregación por Comunidades Autónomas.

- **Cobertura sectorial**: se investigan los centros de cotización cuya actividad económica esté encuadrada en los tres grandes sectores: la Industria, la Construcción y los Servicios.

- **Período de referencia de los resultados**: se divide en dos periodos. El primero de ellos es mes de octubre del año de referencia, que se considera un mes "normal" dentro de la Unión Europea ya que no hay variaciones estacionales ni pagas extras. El segundo es el año en su conjunto.

- **Tamaño muestral**: aproximadamente 28.500 establecimientos y 220.000 trabajadores por encuesta.

- **Tipo de muestreo**: el procedimiento de selección de unidades corresponde con un muestreo bietápico estratificado. La unidad estadística de primera etapa son las cuentas de cotización y se utiliza un muestreo aleatorio estratificado con afijación óptima. La segunda etapa son los trabajadores de las cuentas de cotización. El número de trabajadores seleccionados en cada cuenta depende del tamaño de la misma.

Los microdatos considerados han sido los de la encuesta española de los años 2002, 2006, 2010 y 2014, al igual que en Anghel et al. (2018). Se han tenido en cuenta trabajadores de empresas públicas y privadas del sector de la hostelería. Las distintas encuestas se han organizado en dos grupos. El primero de ellos está formado por las encuestas de 2002 y 2006 y el segundo por las de los años 2010 y 2014. Una vez se han seleccionado los individuos se tiene una muestra de 9774 observaciones de hombres y 13131 de mujeres para el grupo previo a la crisis y 5591 hombres y 8197 mujeres para el grupo posterior. Además, se ha eliminado un 2 % de los individuos de las colas inferior y superior de los datos debido a la existencia de datos atípicos. En la Tabla 4.1 se muestra un resumen descriptivo sobre las distintas encuestas mientras que en la Tabla 4.2 se hace un resumen para los datos agregados. Todos los datos están expresados en euros de 2014.

		2002			2006			2010			2014		
	Horas	Media	Moda	Mediana	Media	Moda	Mediana	Media	Moda	Mediana	Media	Moda	Mediana
Hombres	15	3574	2964	3283	5323	3637	3965	5714	5120	5567	5463	5592	5576
	25	5430	4697	5188	6210	5320	6360	9814	8420	8864	8645	8520	8427
	40	15087	12327	13059	15170	12802	13309	20195	16888	17768	20599	17880	19312
Mujeres	15	3274	2885	3234	4201	3419	3630	4516	3636	4423	4695	4266	4917
	25	5734	6078	6094	6248	6259	6291	8181	7602	8347	8871	8324	9060
	40	11502	11110	11105	11338	11729	11251	15597	14647	14852	16818	15801	16874
Diferencias	15	300	79	49	1122	218	335	1198	1484	1144	768	1326	659
	25	-304	-1381	-906	-38	-939	69	1633	818	517	-226	4254	-633
	40	3585	1217	1954	3832	1073	2058	4598	2241	2916	3781	2079	2438

Tabla 4.1: Salario bruto medio, modal y mediano en términos reales para los datos desagregados.

		2002-2006			2010-2014		
	Horas	Media	Moda	Mediana	Media	Moda	Mediana
Hombres	15	4641	3085	3615	5511	5038	5552
	25	5838	4649	5838	9116	7912	8619
	40	15133	12332	13135	20379	17011	18429
Mujeres	15	3804	2923	3476	4597	4552	4616
	25	6027	6006	6216	8555	8594	8573
	40	11403	11314	11146	16176	15147	15595
Diferencia	15	837	162	139	914	486	936
	25	-189	-1357	-378	561	-682	4003
	40	3730	1018	6919	4203	1864	2834

Tabla 4.2: Salario bruto medio, modal y mediano en términos reales para los datos agregados.

Dado que se han tenido en cuenta datos de empresas privadas y públicas, en la Tabla 4.3 se muestra el resumen descriptivo de los datos diferenciando entre sector público y privado. Para los datos previos a la crisis hay 184 hombres y 282 mujeres. Para el grupo posterior hay datos de 125 hombres y 122 mujeres. Los datos faltantes se representan con una línea.

		2002-2006						2010-2014					
		Sector público			Sector privado			Sector público			Sector privado		
	Horas	Media	Moda	Mediana	Media	Moda	Mediana	Media	Moda	Mediana	Media	Moda	Mediana
Hombres	15	-	-	-	4641	2923	3604	7429	6500	5552	5510	5034	5552
	25	-	-	-	5880	6006	5904	7373	15115	8643	9123	7909	8643
	40	18648	16276	13092	15062	11314	13092	23544	23060	18354	20294	16940	18354
Mujeres	15	-	-	-	3793	3204	3457	-	-	-	4587	4556	4613
	25	3452	6627	2381	6029	6221	6210	13088	13688	13498	8527	8590	8550
	40	14287	13143	13402	11341	11278	11095	17444	15404	16194	16145	15124	15581
Diferencias	15	-	-	-	848	-281	147	-	-	-	923	478	939
	25	-	-	-	-149	-215	-306	-5715	1427	-4855	596	-681	93
	40	4361	3133	-310	3721	36	1997	6100	7656	2160	4149	1816	2773

Tabla 4.3: Salario bruto medio, modal y mediano en términos reales diferenciando entre sector público y privado para los datos agregados.

En las tablas anteriores se puede observar que los hombres tienen salarios mayores que las mujeres para los trabajos de más horas semanales. En los trabajos con pocas horas semanales se produce todo lo contrario. Además, tras la crisis se ha producido un incremento en las rentas reales de los trabajadores pertenecientes al sector de la hostelería, que ha beneficiado en mayor medida a hombres que a mujeres.

También se puede observar que los salarios más elevados siempre son los medios y los medianos y que para pocas horas semanales las mujeres ganan más que los hombres. Aunque no entraremos en más detalles, esto puede deberse a la productividad marginal de los hombres en ese sector para los trabajos de horario parcial.

En lo que se refiere a las diferencias salariales entre sector público y privado, observamos que en el sector privado las diferencias en moda son las más pequeñas, sin embargo, para el sector público sucede lo contrario, con diferencias salariales más elevadas en la mayoría de los casos.

4.3. Análisis de las diferencias salariales

Una vez se han presentado los datos y sus características, en esta sección se realizará un análisis de las diferencias salariales por grupos. Considerando las distintas estimaciones no paramétricas, en las Figuras 4.2 y 4.3 se muestran las diferencias entre medias y modas condicionales diferenciando por sexo. En la Figura 4.4 se muestran las estimaciones de las moda condicionales para ambos grupos. En las figuras se muestran los intervalos de confianza tanto de las medias como de las modas para un nivel de significación $\alpha = 0.05$. Para el cálculo de los intervalos de confianza de las modas se ha utilizado realizado 500 réplicas bootstrap.

Además, en la Figura 4.5 se muestran las diferencias salariales en media, moda y mediana. Las diferencias se han calculado como los salarios de los hombres menos los de las mujeres. En consecuencia, una diferencia negativa significa que el salario de las mujeres es mayor que el de los hombres para las horas semanales en las que ocurre.

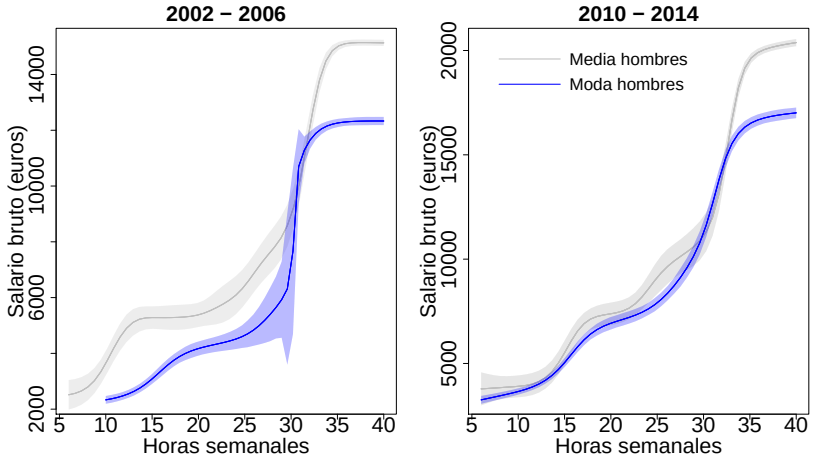


Figura 4.2: Modas y medias condicionales para hombres.

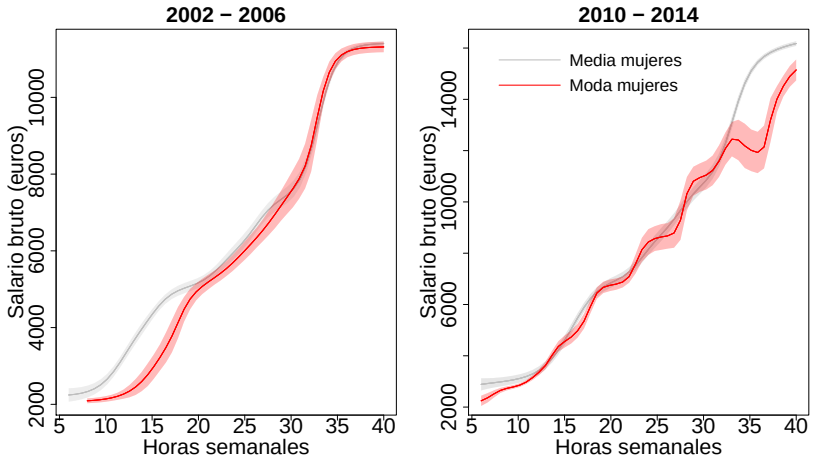


Figura 4.3: Modas y medias condicionales para mujeres.

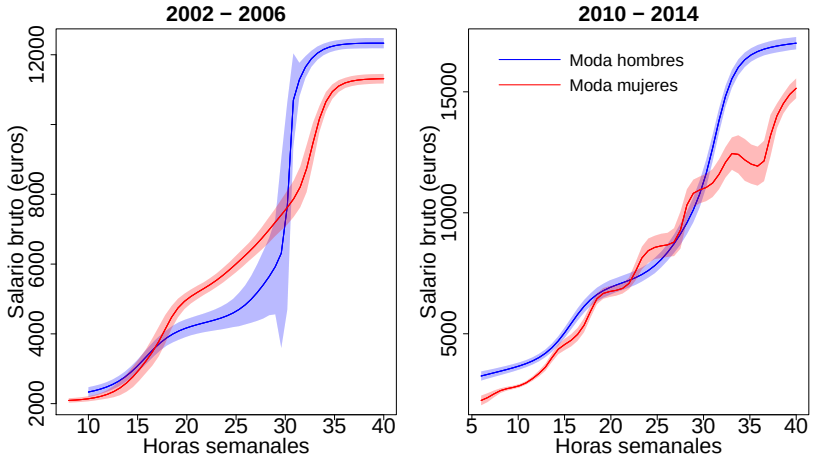


Figura 4.4: Modas condicionales para hombres y mujeres.

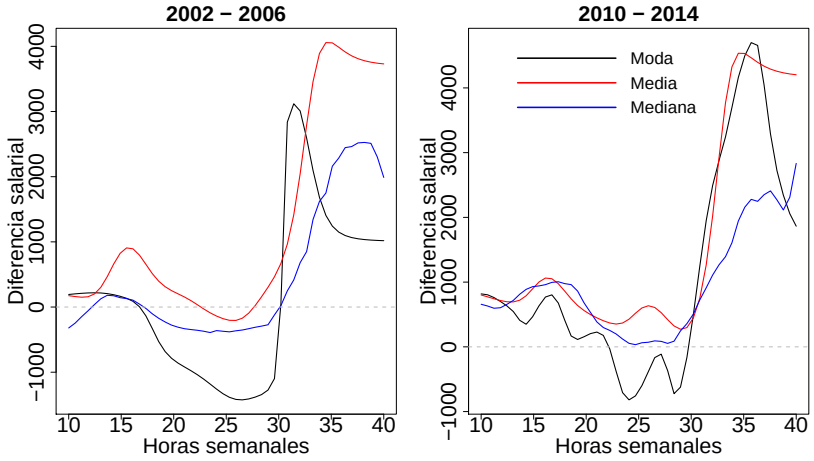


Figura 4.5: Diferencias salariales para media, moda y mediana entre hombres y mujeres.

De las Figuras 4.2, 4.3, 4.4 y 4.5 se concluye que existen diferencias significativas entre las modas y las medias condicionales para distintas horas. Para el conjunto de datos previos a la crisis (años 2002 y 2006) se observa que la media de los hombres difiere de la moda para casi todos los puntos, excepto para los trabajos de aproximadamente 30 horas. A partir de las 35 horas semanales las diferencias son notables. Para las mujeres se observa que la media y moda son iguales excepto para los datos entre 10 y 17 horas semanales, donde la media es mayor que la moda. Este hecho puede deberse a los pocos datos que existen en esas horas.

En cuanto al conjunto de datos posterior a la crisis (años 2010 y 2014), las modas y medias de hombres y mujeres son iguales para menos de 33 y 35 horas semanales respectivamente, donde la media es mayor que la moda. Por último, comparando las modas en los distintos grupos se observa que para el conjunto de datos previo a la crisis existen diferencias estadísticamente significativas para el conjunto entre 17 y 27 horas semanales, donde la moda de las mujeres es mayor que la de los hombres y entre las 31 y 40 horas semanales. En estas horas la moda de los hombres es superior a la de las mujeres. En el segundo conjunto se observan diferencias en pocas horas semanales, que nuevamente pueden ser debidas a los pocos datos que se tienen para esas horas y diferencias a partir de las 30 horas.

Debemos destacar que para analizar si existen diferencias significativas entre las curvas simplemente se han utilizado los distintos intervalos de confianza. En este punto es cuando sería conveniente utilizar un contraste para comparar si las curvas modales estimadas son iguales o si las funciones modales y medias son iguales. Hasta donde alcanza nuestro conocimiento no existe ningún procedimiento para realizar este tipo de contrastes. Por tanto, podría ser objeto de futuras líneas de investigación. Como primera aproximación podría plantearse un contraste bootstrap.

Por otra parte, en la Figura 4.6 se dibujan las densidades condicionadas a 40 horas semanales para los distintos grupos[2]. En línea continua se destacan las modas, en discontinua las medias y en punteada las medianas. Además, en la Figura 4.7 se muestra la evolución de las densidades para ambos sexos.

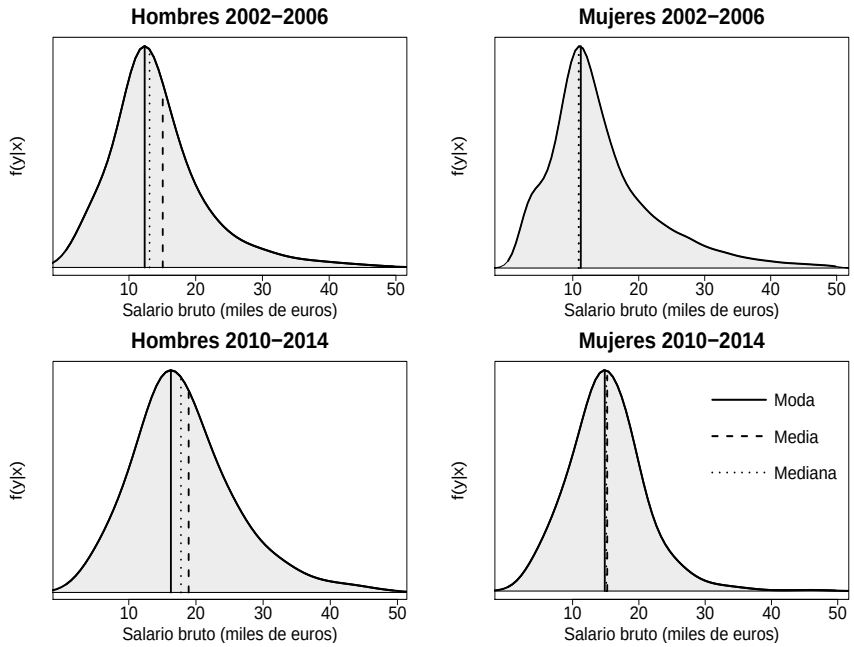


Figura 4.6: Densidades condicionadas a 40 horas semanales para los distintos grupos.

[2]No se representan la colas a partir de 50000 euros brutos para poder ver con mayor claridad los gráficos.

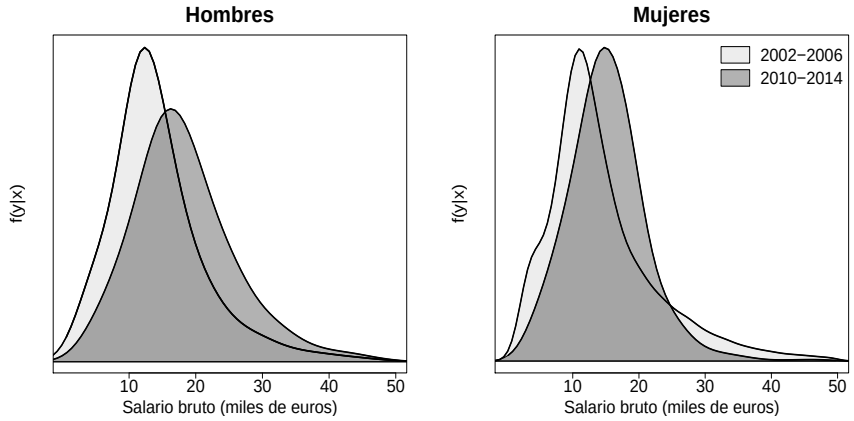


Figura 4.7: Comparación de la evolución de las densidades.

De las Figuras 4.6 y 4.7 podemos concluir que se han producido cambios en las densidades condicionales para la hora considerada y que tanto las modas como las medias se han incrementado para ambos grupos.

Otra cuestión importante es la probabilidad que se acumula en el entorno de las distintas estimaciones. Las probabilidades que acumulan media, moda y mediana para las horas consideradas en la Tabla 4.2 en los dos grupos para un intervalo de ± 1000 euros de las estimaciones se muestran en la Tabla 4.4. En negrita se resalta el valor más alto para cada grupo, que como es lógico, ocurre en las modas.

		2002-2006			2010-2014		
	Horas	Media	Moda	Mediana	Media	Moda	Mediana
Hombres	15	44.42	**48.28**	48.26	29.65	**29.75**	29.63
	25	39.24	**40.77**	39.24	27.05	**27.69**	27.47
	40	23.59	**27.64**	27.30	16.92	**18.60**	18.29
Mujeres	15	59.74	**62.28**	61.58	53.50	**53.51**	53.49
	25	50.47	**50.49**	50.18	40.91	**40.92**	40.92
	40	39.96	**37.98**	37.94	27.56	**28.05**	27.98

Tabla 4.4: Probabilidades acumuladas (en porcentaje) en el entorno de las estimaciones.

De los resultados anteriores podemos concluir que existen diferencias en la probabilidad que acumulan las modas, medias y medianas condicionales. Además, como las modas representan la estimación más probable de la variable aleatoria de los salarios, un individuo que entre en el mercado laboral debería tener más interés en el salario modal que en el medio o el mediano. Conocer el salario modal le permitirá saber el salario más probable que obtendrá dadas sus características.

Análisis del impacto de la crisis en el salario

Por último analizaremos el impacto de la crisis y el sexo sobre los salarios. Con tal objetivo realizaremos dos regresiones lineales múltiples. Para la primera de ellas se considera el modelo

$$W = \alpha + \beta S + \gamma C + \lambda SC + \varepsilon,$$

donde W son los salarios brutos a precios constantes de 2014, S una variable binaria indicativa del sexo, siendo 1 mujer y 0 hombre, C otra variable binaria que es 0 si el dato es de los cuestionarios previos a la crisis o 1 si es de los posteriores, SC es la interacción entre C y S y ε un término de error. Los resultados del ajuste se muestra en la Tabla 4.5.

Parámetro	Estimación	Error estándar	t valor	Pr(>|t|)
α	13726.96	91.82	149.498	0
β	-4488.78	121.27	-37.015	0
γ	3554.53	152.22	23.352	0
λ	-1123.59	198.74	-5.654	0

Tabla 4.5: Estimaciones de los parámetros para el primer modelo.

Los coeficientes estimados son todos estadísticamente significativos para todos los niveles de significación. Sin entrar en más detalles, de los resultados concluimos que las mujeres ganan -4489 euros menos que los hombres. Además, tras la crisis se han incrementado los salarios pero la diferencia de este incremento respecto a los hombres es de -1124 euros. Lo último se puede observar en el gráfico de interacción entre las variables consideradas en la Figura 4.8, donde la pendiente para las mujeres es menor que para los hombres.

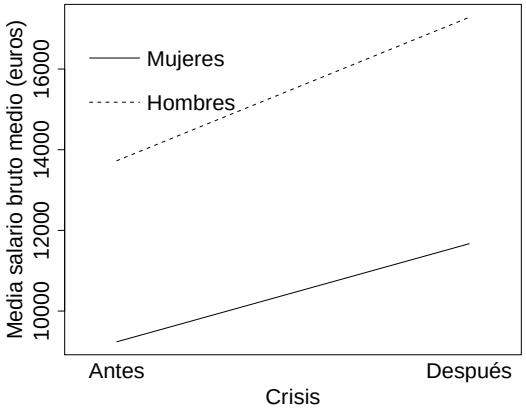


Figura 4.8: Gráfico de interacción entre los factores. Primer modelo.

En segundo lugar consideraremos la misma regresión que antes pero en lugar de trabajar con los salarios trabajaremos con los salarios modales correspondientes para cada individuo. En consecuencia consideraremos el modelo

$$W_{\text{Moda}} = \alpha + \beta S + \gamma C + \lambda SC + \varepsilon,$$

donde la única diferencia respecto al modelo anterior es que la variable dependiente W_{Moda} son los salarios modales para cada individuo. Los resultados del ajuste se muestra en la Tabla 4.6.

Parámetro	Estimación	Error estándar	t valor	Pr(>|t|)
α	11194.57	39.46	283.68	0
β	-2057.26	52.12	-39.47	0
γ	3412.46	65.42	52.16	0
λ	-1577.48	85.41	-18.47	0

Tabla 4.6: Estimaciones de los parámetros para el segundo modelo.

Los parámetros estimados vuelven a ser estadísticamente significativos para los niveles de significación usuales. En este caso, el efecto de ser mujer es una reducción del salario modal de 2057 euros, cantidad inferior a la estimación de la media. En cuanto al efecto de la interacción, se observa que el impacto de la crisis y ser mujer es una renta modal 1577 euros menor que en el caso de los hombres. El gráfico de interacción entre las variables consideradas se muestra en la Figura 4.9.

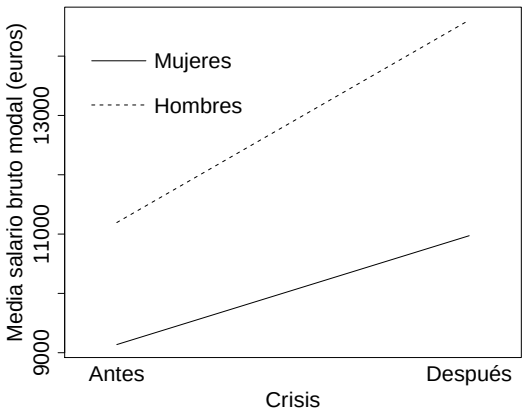


Figura 4.9: Gráfico de interacción entre los factores. Segundo modelo.

Como se ha visto, el impacto de la crisis ha afectado de distinta manera a hombres que a mujeres. Además, tener en cuenta los salarios modales en lugar de los medios ha permitido entender este impacto de una forma distinta que considerando la media. Por tanto, se puede afirmar que las conclusiones a las que podemos llegar si se utiliza la regresión modal en lugar de emplear las técnicas habituales son diferentes.

Capítulo 5

Conclusiones

En esta memoria se ha destacado la importancia de emplear regresión modal frente a las alternativas clásicas cuando los datos presentan distintos patrones, esto es, cuando las densidades condicionales tienen varios máximos locales o cuando éstas son asimétricas.

Con tal objetivo se ha realizado una revisión bibliográfica y se han estudiado con detalle tres procedimientos de estimación. El primero de ellos es el método no paramétrico que emplea el algoritmo mean-shift (Sección 2.2). El carácter no paramétrico lo convierte en el procedimiento más flexible, pero el cálculo de las estimaciones se complica y las tasas de convergencia empeoran a medida que se incrementa el número de covariables. Además, no se puede analizar cómo cambios en las características afectan a las modas. El segundo método estudiado es el primero en la literatura de regresión modal que se basa en regresión cuantil. La estimación del vector de parámetros de un modelo de regresión cuantil puede ser expresado como un problema de programación lineal, por tanto, permite obtener un método computacionalmente escalable a un número elevado de covariables, evitando el desastre de la dimensionalidad. Para la estimación de las modas utiliza la densidad de los cuantiles condicionales (Sección 2.3). El principal problema que presenta es que no existe una regla de selección óptima bajo algún criterio para seleccionar una serie de parámetros ventana. Además, el selector es costoso desde el punto de vista computacional. El tercer método presentado es el que proponemos. Al igual que el método de Ohta está basado un un modelo de regresión cuantil. Con tal objetivo se emplea la distribución empírica de los cuantiles condicionales (Sección 2.4). El método que proponemos intenta solventar la necesidad de calcular parámetros ventana a la vez que se conserva el carácter escalable del procedimiento de la Sección 2.3. Para ello se cambia la perspectiva del segundo método y en lugar de emplear las densidades de los cuantiles condicionales se propone utilizar las distribuciones empíricas. Las modas se encuentran en este caso en sus puntos de inflexión que se calculan mediante un proceso en dos pasos.

Una vez presentados los distintos procedimientos, en el Capítulo 3 se ha realizado un estudio de simulación con distintos tipos de modelos, con funciones modales lineales y no lineales para analizar los métodos bajo distintos escenarios. De los estudios realizados se concluye que nuestra propuesta de estimación presentada en la Sección 2.4 obtiene mejores resultados que el estimador de Ohta, estando ambos métodos basados en regresión cuantil lineal. Además, nuestro procedimiento es más eficiente desde el punto de vista computacional. Por otra parte, el método no paramétrico tiene un buen rendimiento en todo tipo de modelos, además de poder aplicarse al caso multimodal. Es aquí donde nace la necesidad de considerar la versión no paramétrica del método propuesto y realizar un nuevo estudio comparativo. De todas formas, considerar la versión no paramétrica supondría volver al problema de selección de ventanas. En lo que se refiere a los parámetros ventana en el método de Ohta y en el no paramétrico, cabe destacar que las estimaciones son sensibles ante cambios en estos parámetros. Para el estudio se han utilizado los selectores recomendados en ambos artículos. El estimador de Ohta, al no tener una regla de selección óptima presenta unos resultados ante cambios en la ventana no

satisfactorios. Para el otro método el selector utilizado parece presentar un funcionamiento adecuado. En lo que se refiere a nuestra propuesta, es necesario seleccionar un tamaño de secuencia inicial en el primer paso del algoritmo, pero la variabilidad de las estimaciones no se ve gravemente afectada por cambios en los tamaños de dichas secuencias. Además, no es necesario utilizar ningún tipo de ventana para cada punto a estimar.

Por último se han analizado varios conjuntos de datos reales del ámbito de la Economía que se corresponden con varias encuestas de estructura salarial española. El objetivo ha sido evaluar la toma de decisiones en el mercado laboral así como las diferencias salariales entre hombres y mujeres desde el punto de vista de la moda. A raíz de este estudio ha surgido la necesidad de comparar curvas modales mediante algún tipo de procedimiento, que hasta donde alcanza nuestro conocimiento, todavía no se ha realizado en la literatura. Además se ha evaluado el impacto de la crisis en los salarios medios y modales. Del estudio se concluye que existen diferencias salariales para hombres y mujeres de las mismas características y que las expectativas de un individuo en el mercado laboral desde el punto de vista de la moda son distintas que desde el punto de vista de la media o la mediana. Un individuo que entre en el mercado laboral debería tener la expectativa de que el salario que obtendrá dadas sus características es el salario modal, en lugar del salario medio o el mediano, ya que éste es el más probable. Por último, del análisis del impacto se concluye que el mayor impacto lo han sufrido las mujeres y que éste es distinto si se considera la media o la moda.

Bibliografía

[1] Ameijeiras-Alonso J, Crujeiras RM, Rodríguez-Casal A (2018) multimode: Mode Testing and Exploring. R package version 1.4. https://cran.r-project.org/package=multimode.

[2] Anghel B, Conde-Ruiz JI, Marra de Artíñano I (2018). Brechas Salariales de Género en España. FEDEA Estudios de Economía Española, Madrid.

[3] Arrow KJ (1971) The Theory of Discrimination. En Aschenfelter O y Rees A (eds) Discrimination in Labor Markets. Princeton University Press, Princeton.

[4] Bashtannyk DM, Hyndman RJ (2001) Bandwidth selection for kernel conditional density estimation. Computational Statistics and Data Analysis 36:279-298.

[5] Bickel DR, Frühwirth R (2006) On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. Computational Statistics and Data Analysis 50:3500-3530.

[6] Blinder AS (1973) Wage Discrimination: Reduced Form and Structural Estimates. Journal of Human Resources 8:436-455.

[7] Bowman A, Azzalini A (2019) sm: Smoothing Methods for Nonparametric Regression and Density Estimation. R package version 2.2-5.6. https://cran.r-project.org/package=sm. Accedido 2 de marzo de 2019.

[8] Carreira-Perpiñán MA (2007) Gaussian mean-shift is an EM algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 29:767-776.

[9] Chen YC, Genovese CR, Tibshirani RJ, Wasserman L (2016) Nonparametric modal regression. The Annals of Statistics 44:489-514.

[10] Cheng Y (1995) Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 17:790-799.

[11] Chernoff H (1964) Estimation of the mode. Annals of the Institute of Statistical Mathematics 16:31-41.

[12] Collomb G, Härdle W, Hassani S (1986) A note on prediction via estimation of the conditional mode function. Journal of Statistical Planning and Inference 15:227-236.

[13] Comaniciu D, Meer P (2002) Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:603-619.

[14] Dalenius T (1965) The Mode - A Neglected Statistical Parameter. Journal of the Royal Statistical Society: Series A (General) 12:110-117.

[15] Dasgupta S, Kpotufe S (2014) Optimal rates for k-NN density and mode estimation. Advances in Neural Information Processing Systems 3:2555-2563.

[16] Eddy WF (1980) Optimum Kernel Estimators of the Mode. The Annals of Statistics 8:870-882.

[17] Einbeck J, Tutz G (2006) Modelling beyond regression functions: an application of multimodal regression to speed-flow data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 55:461-475.

[18] Feng Y, Fan J, Suykens JAK (2017) A Statistical Learning Approach to Modal Regression. arXiv:1702.05960

[19] Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 21:32-40.

[20] Hyndman R, Einbeck J, Wand M (2018) hdrcde: Highest Density Regions and Conditional Density Estimation. R package version 3.3. https://cran.r-project.org/package=hdrcde. Accedido 14 de febrero de 2019.

[21] Hyndman R, Bashtannyk DM, Grundwald GK (1996) Estimating and visualizing conditional densities. Journal of Computational and Graphical Statistics 5:315-336.

[22] Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457-481.

[23] Kemp GCR, Santos-Silva JMC (2012) Regression towards the mode. Journal of Econometrics 170:92-101.

[24] Khardani S, Lemdani M, Ould-Saïd E (2010) Some asymptotic properties for a smooth kernel estimator of the conditional mode under random censorship. Journal of the Korean Statistical Society 39:455-469.

[25] Khardani S, Lemdani M, Ould-Saïd E (2011) Uniform rate of strong consistency for a smooth kernel estimator of the conditional mode for censored time series. Journal of Statistical Planning and Inference 141:3426-3436.

[26] Khardani S (2019) A Semi-Parametric Mode Regression with Censored Data. Mathematical Methods of Statistics 28:39-56.

[27] Koenker R, Bassett (1978) Regression quantiles. Econometrica 46:33-50.

[28] Koenker R, Machado JAF (1999) Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94:1296-1310.

[29] Lee M-J (1989) Mode regression. Journal of Econometrics 42:337-349.

[30] Lee M-J (1993) Quadratic mode regression. Journal of Econometrics 57:1-19.

[31] Machado JAF, Mata J (2005) Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of Applied Econometrics 20:445-465.

[32] Maechler M (2019) sfsmisc: Utilities from "Seminar fuer Statistik" ETH Zurich. R package version 1.1-4. https://cran.r-project.org/package=sfsmisc. Accedido 26 de mayo de 2019.

[33] Manski CF (1991) Regression. Journal of Economic Literature 29:34-50.

[34] Manski CF (2003) Identification Problems in the Social Sciences and Everyday Life. Southern Economic Journal 70:11-21.

[35] Oaxaca R (1973) Male-Female Wage Differentials in Urban Labor Markets. International Economic Review 14:693-709.

[36] Ohta H, Kato K, Haram S (2018) Quantile regression approach to conditional mode estimation. arXiv:1811.05379.

[37] Ould-Saïd E, Cai Z (2005) Strong uniform consistency of nonparametric estimation of the censored conditional mode function. Nonparametric Statistics 17:797-806.

[38] Página web del Instituto Nacional de Estadística, https://www.ine.es. Accedido el 5 de Mayo de 2019.

[39] Parzen E (1964) On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics 33:1065-1076.

[40] Quintela-Del-Río A, Vieu P (1997) A nonparametric conditional mode estimate. Journal of Nonparametric Statistics 8:253-266.

[41] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[42] Robertson T, Cryer JD (1974) An Iterative Procedure for Estimating the Mode. Journal of the American Statistical Association 69:1012-1016.

[43] Sager TW, Thisted RA (1982) Maximum likelihood estimation of isotonic modal regression. The Annals of Statistics 10:690-707.

[44] Samanta M, Thavaneswaran A (1990) Non-parametric estimation of the conditional mode. Communication in Statistics - Theory and Methods 19:4515-4524.

[45] Sasaki H, Ono Y, Sugiyama M (2016) Modal regression via direct log-density derivative estimation. En: International Conference on Neural Information Processing, pp 108-116.

[46] Taylor J (2012) Strategies for mean and modal multivariate local regression. Tesis doctoral, Universidad de Durham. http://etheses.dur.ac.uk/3514/.

[47] Venter JH (1967) On estimation of the mode. The Annals of Mathematical Statistics 38:1446-1455.

[48] Wang J, Thiesson B, Xu Y, Cohen M (2004) Image and video segmentationby anisotropic kernel mean shift. En: European Conference on Computer Vision, pp 238-249.

[49] Yao W, Lindsay BG, Li R (2012) Local modal regression. Journal of Nonparametric Statistics 24:647-663.

[50] Yao W, Li, R (2014) A new regression model: modal linear regression. Scandinavian Journal of Statistics 41:656-671.

[51] Zhou H, Huang X (2016) Nonparametric modal regression in the presence of measurement error. Electronic Journal of Statistics 10:3579-3620.

[52] Zhou H, Huang X (2017) lpme: Nonparametric Estimation of Measurement Error Models. R package version 1.1.1. https://cran.r-project.org/package=lpme.

[53] Zhou H, Huang X (2019) Bandwidth selection for nonparametric modal regression. Communications in Statistics - Simulation and Computation 48:968-984.

Índice de figuras

Índice de tablas

Apéndice: Códigos R

En el presente apéndice se muestran los códigos utilizados en el estudio de simulación realizado en el Capítulo 3. Como consecuencia de haber fijado una semilla los resultados obtenidos son completamente reproducibles ejecutado las líneas correspondientes a cada método y comentando o descomentando las líneas que correspondan a los distintos parámetros y modelos.

```
#=============================================================================
# Método de Ohta
#=============================================================================

library(quantreg)

#-----------------------------------------------------------------------------
# Parámetros iniciales
#-----------------------------------------------------------------------------

epsilon <- 0.1
alpha <- 0.05
n <- 500
# n <- 1000
# n <- 2000
n.iter <- 1000

error <- vector("list", n.iter)

for (i in 1:length(error)){
 error[[i]] <- matrix(NA, nrow = 9, ncol = 50)
}

RMSE <- vector("list", n.iter)

for (i in 1:length(RMSE)){
 RMSE[[i]] <- matrix(NA, nrow = 9, ncol = 1)
}

mxhat <- error

pb <- txtProgressBar(min = 0, max = n.iter, style = 3)
set.seed(1)
X2 <- runif(n)
# X3 <- runif(n)
# X4 <- rnorm(n)
```

```r
# X2 <- runif(n, 0, 6)

taus <- seq(0.05, 0.95, length.out = 100)

for(i in 1:n.iter) {

 setTxtProgressBar(pb, i)


 #---------
 # Modelo 1
 #---------

 Y <- 1 + 3 * X2 + rnorm(n)
 Y2.teo <- 1 + 3 * X2
 plot(X2, Y)
 lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


 #---------
 # Modelo 2
 #---------

 # sigma <- 1 + 2 * X2
 # eps <- 0.5 * rnorm(n, -1, 2.5 ^ 2) + 0.5 * rnorm(n, 1, 0.5 ^ 2)
 # Y <- 1 + 3 * X2 + sigma * eps
 # Y2.teo <- 1 + 2 * X2
 # plot(X2, Y)
 # lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)

 #---------
 # Modelo 3
 #---------

 # U <- runif(n, 0, 1)
 # Y <- U ^ 3 / 3 - X2 * (U - 1) ^ 2
 # Y2.teo <- -2 * X2 ^ 3 / 3 + 2 * X2 ^ 2 - X2
 # plot(X2, Y)
 # lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


 #---------
 # Modelo 4
 #---------

 # Y <- sin(X2) + rnorm(n)
 # Y2.teo <- sin(X2)
 # plot(X2, Y)
 # lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)

 for(mc in 1:50) {
```

```
88
89  #-------------------------------------------------------------------------
90  # Cálculo de la ventana piloto h^pilot
91  #-------------------------------------------------------------------------
92
93  tau <- 0.5
94  h_KM <- n ^ (-1 / 3) * pnorm(1 - alpha / 2) ^ (2 / 3) *
95         (1.5 * ((dnorm(pnorm(tau))) / (2 * pnorm(tau) ^ 2 + 1))) ^ 1 / 3
96  h_pilot <- n ^ (1 / 6) * h_KM
97
98  #-------------------------------------------------------------------------
99  # Construímos tau_x_prelim con la ventana piloto
100 #-------------------------------------------------------------------------
101
102 tau_U <- seq(0.05, 0.95, length.out = 100) + sapply(1:length(taus),
103          function(i) min(h_pilot, max(taus) - taus[i]))
104          tau_L <- seq(0.05, 0.95, length.out = 100) - sapply(1:length(taus),
105          function(i) min(h_pilot, taus[i] - min(taus)))
106
107 beta_U <- coef(rq(formula = Y ~ X2, tau = tau_U))
108 beta_L <- coef(rq(formula = Y ~ X2, tau = tau_L))
109
110 # Elegimos un punto de diseño
111
112 XX2 <- seq(min(X2), max(X2), length.out = 50)
113
114 x <- c(1, XX2[mc])
115
116 pred_U <- x %*% beta_U
117 pred_L <- x %*% beta_L
118
119 # Calculamos ŝ_x(τ)  y  τ̂_x^prelim
120
121 sxhat <- (pred_U - pred_L) / (sapply(1:length(taus),
122          function(i) min(h_pilot, max(taus) - taus[i])) +
123          sapply(1:length(taus),
124          function(i) min(h_pilot, taus[i] - min(taus))))
125
126 tau_x_prelim <- tau_U[which.min(sxhat[, (epsilon * ncol(sxhat)):
127                 (ncol(sxhat) - epsilon * ncol(sxhat))])] - h_pilot
128
129
130 #-------------------------------------------------------------------------
131 # Volvemos a calcular el h con τ̂_x^prelim
132 #-------------------------------------------------------------------------
133
134 h_KM <- n ^ (-1 / 3) * pnorm(1 - alpha / 2) ^ (2 / 3) *
135        (1.5 * ((dnorm(pnorm(tau_x_prelim))) /
136        (2 * pnorm(tau_x_prelim) ^ 2 + 1))) ^ 1 / 3
137 h_n <- n ^ (1 / 6) * h_KM
138
```

```
  h_grid <- c(0.10 * h_n, 0.25 * h_n, 0.90 * h_n, 0.95 * h_n, h_n,
              1.05 * h_n, 1.10 * h_n, 4 * h_n, 10 * h_n)

  for(w in 1:length(h_grid)) {

    #-------------------------------------------------------------------------
    # Construímos el estimador final m(x)
    #-------------------------------------------------------------------------

    tau_U <- taus + sapply(1:length(taus),
                           function(i) min(h_grid[w], max(taus) - taus[i]))

    tau_L <- taus - sapply(1:length(taus),
                           function(i) min(h_grid[w], taus[i] - min(taus)))

    beta_U <- coef(rq(formula = Y ~ X2, tau = tau_U))
    beta_L <- coef(rq(formula = Y ~ X2, tau = tau_L))

    pred_U <- x %*% beta_U
    pred_L <- x %*% beta_L

    sxhat <- (pred_U - pred_L) / (sapply(1:length(taus),
                           function(i) min(h_grid[w], max(taus) - taus[i])) +
                           sapply(1:length(taus), function(i) min(h_grid[w],
                           taus[i] - min(taus))))

    tau_opt <- tau_U[which.min(sxhat[,(epsilon * ncol(sxhat)):(ncol(sxhat) -
                             epsilon * ncol(sxhat))])] - h_grid[w]

    beta_opt <- coef(rq(formula = Y ~ X2, tau = tau_opt))

    mxhat[[i]][w, mc] <- x %*% beta_opt

  # mx <- 1 + 3 * x[2]
  # mx <- 2 + 5 * x[2]
  # mx <- -2 * x[2] ^ 3 / 3 + 2 * x[2] ^ 2 - x[2]
  # mx <- sin(x[2])

    error[[i]][w, mc] <- mxhat[[i]][w, mc] - mx
    RMSE[[i]][w, ] <- sqrt(mean(unlist(error[[i]][w, ]) ^ 2))
  # lines(seq(min(X2), max(X2), length.out = 50),
          mxhat[[i]][w,], lwd = 2, col = w)
   }
 }

# lines(seq(min(X2), max(X2), length.out = 50), mxhat[[i]][3,], lwd = 2, col = 3)

 print(paste0("RMSE_", i," = ", RMSE[i]))
}
```

```r
#=============================================================================
# Método basado en la distribución de los cuantiles condicionales
#=============================================================================

#-----------------------------------------------------------------------------
# Parámetros iniciales
#-----------------------------------------------------------------------------

n <- 500
# n <- 1000
# n <- 2000
n.iter <- 1000

RMSE <- numeric(n.iter)
error <- vector("list", n.iter)
mxhat <- error
pb <- txtProgressBar(min = 0, max = n.iter, style = 3)
set.seed(1)
X2 <- runif(n)
# X2 <- runif(n, 0, 6) # Para el Modelo 4

for(i in 1:n.iter){

 setTxtProgressBar(pb, i)


 #---------
 # Modelo 1
 #---------

 Y <- 1 + 3 * X2 + rnorm(n)
 Y2.teo <- 1 + 3 * X2
 plot(X2, Y)
 lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


 #---------
 # Modelo 2
 #---------

 # sigma <- 1 + 2 * X2
 # eps <- 0.5 * rnorm(n, -1, 2.5 ^ 2) + 0.5 * rnorm(n, 1, 0.5 ^ 2)
 # Y <- 1 + 3 * X2 + sigma * eps
 # Y2.teo <- 1 + 2 * X2
 # plot(X2, Y)
 # lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


 #---------
 # Modelo 3
 #---------
```

```r

# U <- runif(n, 0, 1)
# Y <- U ^ 3 / 3 - X2 * (U - 1) ^ 2
# Y2.teo <- -2 * X2 ^ 3 / 3 + 2 * X2 ^ 2 - X2
# plot(X2, Y)
# lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


#---------
# Modelo 4
#---------

# Y <- sin(X2) + rnorm(n)
# Y2.teo <- sin(X2)
# plot(X2, Y)
# lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


for(mc in 1:50) {

 tau <- seq(0.05, 0.95, length.out = 4)

 beta <- coef(rq(formula = Y ~ X2, tau = tau))

 # Elegimos un punto de diseño
 XX <- seq(min(X2), max(X2), length.out = 50)

 x <- c(1, as.numeric(XX[[mc]]))

 pred <- x %*% beta

 ii <- numeric(length(pred)-1)

 for(w in 1:length(pred)){

  ii[w] <- (pred[w + 1] - pred[w])
  # ii[w] <- (abs(pred[w + 1]) - abs(pred[w])) - (abs(pred[w]) - abs(pred[w - 1]))

 }

 min <- which.min(abs(na.omit(ii)))

 ii[min] <- 1000

 min2 <- min + 1 #which.min(abs(na.omit(ii)))

 mins <- sort(c(min, min2))

 tau2 <- seq(tau[mins[1]], tau[mins[2]], length.out = 25)

 beta2 <- coef(rq(formula = Y ~ X2, tau = tau2))
```

```
  pred <- x %*% beta2

  ii <- numeric(length(pred)-1)

  for(w in 1:length(pred)){

   ii[w] <- (pred[w + 1] - pred[w])
   # ii[w] <- (abs(pred[w + 1]) - abs(pred[w])) - (abs(pred[w]) - abs(pred[w - 1]))
  }

  min <- which.min(abs(na.omit(ii)))

  ii[min] <- 1000

  min2 <- min + 1 #which.min(abs(na.omit(ii)))

  mins <- sort(c(min, min2))

  tau_opt <- mean(c(tau2[mins[1]], tau2[mins[2]]))

  beta_opt <- coef(rq(formula = Y ~ X2, tau = tau_opt))

  mx <- 1 + 3 * x[2]
# mx <- 2 + 5 * x[2]
# mx <- -2 * x[2] ^ 3 / 3 + 2 * x[2] ^ 2 - x[2]
# mx <- sin(x[2])

  mxhat[[i]][[mc]] <- x %*% beta_opt
  error[[i]][[mc]] <- mxhat[[i]][[mc]] - mx
# points(XX[order(XX)][[mc]], mxhat[[i]][[mc]], lwd = 2, col = 3, pch = 16)
 }

# lines(XX, mxhat[[i]][order(XX)], lwd = 2, col = 4)

 RMSE[i] <- sqrt(mean(unlist(error[[i]]) ^ 2))

 print(paste0("RMSE_", i," = ", RMSE[i]))
}


mean(RMSE)
```

```r
#=============================================================================
# Método basado en el mean-shift
#=============================================================================

library(hdrcde)

#-----------------------------------------------------------------------------
# Parámetros iniciales
#-----------------------------------------------------------------------------

n <- 500
# n <- 1000
# n <- 2000
n.iter <- 1000

error <- vector("list", n.iter)

for (i in 1:length(error)){
 error[[i]] <- matrix(NA, nrow = 9, ncol = 50)
}

RMSE <- vector("list", n.iter)

for (i in 1:length(RMSE)){
 RMSE[[i]] <- matrix(NA, nrow = 9, ncol = 1)
}

mxhat <- error

pb <- txtProgressBar(min = 0, max = n.iter, style = 3)
set.seed(1)
X2 <- runif(n)
# X3 <- runif(n)
# X4 <- rnorm(n)
# X2 <- runif(n, 0, 6) # Para el Modelo 4

for(i in 1:n.iter){

 setTxtProgressBar(pb, i)

 #---------
 # Modelo 1
 #---------

 Y <- 1 + 3 * X2 + rnorm(n)
 Y2.teo <- 1 + 3 * X2
 plot(X2, Y)
 lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)

 #---------
 # Modelo 2
```

```r
#---------

# sigma <- 1 + 2 * X2
# eps <- 0.5 * rnorm(n, -1, 2.5 ^ 2) + 0.5 * rnorm(n, 1, 0.5 ^ 2)
# Y <- 1 + 3 * X2 + sigma * eps
# Y2.teo <- 1 + 2 * X2
# plot(X2, Y)
# lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)

#---------
# Modelo 3
#---------

# U <- runif(n, 0, 1)
# Y <- U ^ 3 / 3 - X2 * (U - 1) ^ 2
# Y2.teo <- -2 * X2 ^ 3 / 3 + 2 * X2 ^ 2 - X2
# plot(X2, Y)
# lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)

#---------
# Modelo 4
#---------

# Y <- sin(X2) + rnorm(n)
# Y2.teo <- sin(X2)
# plot(X2, Y)
# lines(X2[order(X2)], Y2.teo[order(X2)], col = "red", lwd = 3)


fit <- modalreg(X2, Y, P = 1,
                xfix = seq(min(X2), max(X2), length.out = 50),
                start = "e")

h_grid <- vector("list", 9)

h_grid[[1]] <- c(0.10 * fit$bandwidths[1], 0.10 * fit$bandwidths[2])
h_grid[[2]] <- c(0.25 * fit$bandwidths[1], 0.25 * fit$bandwidths[2])
h_grid[[3]] <- c(0.90 * fit$bandwidths[1], 0.90 * fit$bandwidths[2])
h_grid[[4]] <- c(0.95 * fit$bandwidths[1], 0.95 * fit$bandwidths[2])
h_grid[[5]] <- fit$bandwidths
h_grid[[6]] <- c(1.05 * fit$bandwidths[1], 1.05 * fit$bandwidths[2])
h_grid[[7]] <- c(1.10 * fit$bandwidths[1], 1.10 * fit$bandwidths[2])
h_grid[[8]] <- c(4 * fit$bandwidths[1], 4 * fit$bandwidths[2])
h_grid[[9]] <- c(10 * fit$bandwidths[1], 10 * fit$bandwidths[2])

for(w in 1:length(h_grid)){

 if(w == 3){
  mxhat[[i]][w, ] <- as.numeric(fit$fitted.values)
 } else {
  mxhat[[i]][w, ] <- modalreg(X2, Y, P = 1,
```

```
                                xfix = seq(min(X2), max(X2), length.out = 50),
                                a = unlist(h_grid[[w]][1]),
                                b = unlist(h_grid[[w]][2]),
                                start = "e")$fitted.values
  }

  mx <- 1 + 3 * seq(min(X2), max(X2), length.out = 50)
  # mx <- 2 + 5 * seq(min(X2), max(X2), length.out = 50)
  # mx <- -2 * seq(min(X2), max(X2), length.out = 50) ^ 3 / 3 +
  #         2 * seq(min(X2), max(X2), length.out = 50) ^ 2 -
  #         seq(min(X2), max(X2), length.out = 50)
  # mx <- sin(seq(min(X2), max(X2), length.out = 50))
  error[[i]][w, ] <- mxhat[[i]][w, ] - mx
  RMSE[[i]][w, ] <- sqrt(mean(unlist(error[[i]][w, ]) ^ 2))
 }

 print(paste0("RMSE_", i," = ", RMSE[i]))
}
```