Capítulo 6

EXPERIMENTOS DE SIMULACIÓN: ESTIMADOR POLINOMIAL LOCAL

Boukichou-Abdelkader, N.[1]; Montero-Alonso, M.Á.[2]; Muñoz-García, A.[3]
Y
Canário, P.N.[4]
[1]Centro de Investigación Ceiis - IdiPAZ. Hospital Universitario La Paz. Madrid. España.
[2]Departamento de Estadística e Investigación Operativa, Universidad de Granada, España.
[3]Departamento de Estadística, Universidad Carlos III de Madrid, España.
[4]C3i, Polytechnic Institute of Portalegre, P -7300 -110,Portalegre, Portugal.

ABSTRACT

This simulation study estimates the trend of a nonparametric dataset using local polynomial estimators. The local regression technique is based on perform various adjustments parametric considering the near data to the point where you want to estimate the regression function. Necessarily, this simulation method is determined by three key parameters: the sample size, the type of domain or grid, where the data are simulated, and the model trend function. For the development of this technique we used four R libraries: kernSmooth, locpol, locfit and sm, to practically implement the algorithms and interpret the results. Therefore, the objective of this simulated experiment is to facilitate a good adjusted model by applying the nonparametric estimate, using another alternative, when the trend model of observed data is unknown.

KEY WORDS: Nonparametric Regression, Local Polynomial Regression, type of Grid, sample size.

RESUMEN

En este estudio de simulación se pretende estimar la tendencia de un conjunto de datos de manera no paramétrica mediante el estimador polinomial local. La técnica de regresión local se basa en realizar varios ajustes paramétricos teniendo únicamente en cuenta los datos cercanos al punto donde se desea estimar la función de regresión. Previamente, para este método simulado es necesario determinar tres parámetros fundamentales: el tamaño muestral, el tipo de dominio o rejilla, donde se simularán los datos, y la función de tendencia para la simulación de los datos y la estimación del modelo. Para el desarrollo de esta técnica se utilizaran las librerías: kernSmooth, locpol, locfit y sm del software estadístico R, que facilitaran la aplicación práctica y la interpretación de los resultados simulados. Por tanto, el objetivo de este experimento simulado es facilitar un buen modelo ajustado aplicando la estimación no paramétrica, mediante otra alternativa, cuando se desconoce el modelo de tendencia de los datos observados.

Palabras claves: Regresión No Paramétrica, Regresión Polinomial Local, tipo de rejilla, tamaño muestral.

1. INTRODUCCIÓN

En la mayoría de las investigaciones científicas, uno de los problemas más importantes, es la formulación de los modelos estadísticos para representar de forma adecuada el fenómeno objeto de estudio.

En muchas ocasiones, no se dispone de toda la información real para realizar un análisis exhaustivo de los datos observados. Y en este momento es cuando se requiere la utilización de otros procedimientos más específicos para poder comprobar desde otra vía las sospechas que se tiene en la investigación objeto de estudio.

Afortunadamente, los contrastes paramétricos no son los únicos disponibles en la rama de la Estadística y existen otros métodos más flexibles para abordar este tipo de planteamiento, como son las técnicas no paramétricas. Las estimaciones no paramétricas permiten construir modelos que se ajustan a los datos de forma local ([1]), cuando no se puede asumir una distribución conocida.

Como es sabido, la teoría y los métodos de suavizamiento o regresión no paramétrica han cobrado un gran auge en las últimas décadas unido al avance en materia computacional, pudiéndose encontrar una revisión de los mismos en los libros [2], [3] y [4].

El creciente interés por estas metodologías ha tenido dos razones principales: la primera, que los planteamientos puramente paramétricos no aportaban la flexibilidad necesaria para la estimación de las curvas que aparecían en la práctica. Y la segunda razón, estaba ligada al avance de la informática y al desarrollo de un hardware que posibilitaba el costoso cálculo de esos estimadores no paramétricos.

Los primeros estimadores de regresión no paramétrica propuestos fueron los sencillos estimadores de tipo núcleo, [5] y [6].

Dichos estimadores se han ido modificando y moldeando dentro de los denominados **métodos de regresión polinomial local**, convirtiéndose en uno de los métodos más empleados por diversos analistas en la actualidad.

No obstante, para estos métodos de regresión no paramétrica se introduce uno de los problemas técnicos y críticos en la práctica, la elección del parámetro de suavizado o ancho de banda, que define la complejidad del modelo. Dada la dificultad que supone dicha selección, se distingue entre los métodos basados en la metodología plug-in y los basados en el criterio de validación cruzada.

En este sentido, el trato realizado de dichos métodos, ha sido dirigido fundamentalmente hacia la práctica, sin profundizar demasiado en aspectos teóricos de complejidad como son los procedimientos de tipo asintótico.

Desde este punto de vista no paramétrico, en concreto, con el estimador polinomial local, se pretende presentar un experimento simulado donde las observaciones analizadas son generadas mediante un modelo previamente definido, que está determinado apriorispor tres parámetros fundamentales, como son el tamaño muestral, el tipo de rejilla (o de dominio) donde se simularán los datosy la función de tendencia (tanto para la simulación de los datos como para la estimación de la tendencia).

Bajo este planteamiento se han explorado los métodos de regresión polinomial local como una de las mejores opciones de análisis, puesto que estas técnicas poseen características de flexibilidad, aplicabilidad e interpretabilidad bastante acertadas para este fin analítico. Para ello se ha realizado todo este proceso en el entorno de análisis y programación estadística R mediante algunas librerías específicas, como son *kernSmooth*, *locpol*, *locfit*, y*sm*, que se pueden descargar directamente através de la web (http://cran.es.r-project.org).

2. MÉTODO

Sea un conjunto de n observaciones, $\{(X_i, Y_i), i = 1,..., n\}$, buscamos un estimador de la función de regresión $m(x) = E[Y|X = x]$ de manera que los datos siguen el modelo,

$$Y_i = m(X_i) + \varepsilon_i \qquad i = 1,..., n,$$

dondelos residuos ε_i son variables aleatorias independientes con media cero y varianza $\sigma^2(X_i)$.

Para alcanzar tales objetivos se puede optar por una regresión paramétrica, y supone que la función de regresión desconocida, *m*, pertenece a alguna familia paramétrica de funciones, $m \in \{m_\theta | \theta \in \Theta\}$, donde θ se estima mediante mínimos cuadrados. La regresión no paramétrica no asume ninguna forma paramétrica para la función *m*([1], [2], [3] y [4]), y la única restricción que se le impone es que sea suave, entendiendo esta suavidad en términos de derivabilidad.

Los primeros estimadores de regresión no paramétrica propuestos fueron los sencillos estimadores de tipo núcleo [5] y [6], estimadores que se han ido refinando y perfeccionando dentro de los denominados métodos de regresión polinomial local, convirtiéndose en uno de los métodos más empleados por diversos analistas en la actualidad, ya que obtiene un estimador sencillo y corrige de forma automática los efectos frontera.

La regresión polinomial local supone que la función de regresión*m*, tiene *p* derivadas en un punto x_0, obteniéndose una aproximación para los valores en un entorno de x_0.

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + ... + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p,$$

es decir, se puede aproximar localmente *m* por funciones polinómicas de grado *p*

$$P_p(x) = \sum_{j=0}^{p} \beta_j (x - x_0)^j,$$

obteniéndose estimaciones de los coeficientes $\hat{\beta}_j$ con $j = 0,..., p.$

Con el fin de estimar *m* localmente mediante polinomios de grado *p* se considerara un problema de mínimos cuadrados ponderados:

$$\min \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 k_h(X_i - x_0)$$

donde*h* es un parámetro denominado ancho de banda o parámetro de suavizado que controla las observaciones que caen en cada entorno, $K_h(u) = h^{-1}K(\frac{u}{h})$, donde la función $K(\cdot)$, se denomina función núcleo. Dicha función define las ponderaciones que se asignan a cada observación en el entorno local considerado. Habitualmente se supone una densidad simétrica y con soporte compacto, y *p* es el grado del ajuste polinomiallocal[1].

3. EXPERIMENTOS DE SIMULACIÓN

Se ha realizado un análisis con datos simulados donde las observaciones analizadas son generadas mediante un modelo previamente definido y constituido por tres parámetros fundamentales: el tamaño muestral, el tipo de rejilla y la función de tendencia. Desde este punto de vista, se pretende ilustrar los métodos de regresión no paramétrica, en concreto, el estimador polinomial local (véase las citas de la [8] a la [16]).

El objetivo de este experimento será cuantificar la bondad de las estimaciones (dado que se conocen los modelos exactos) y además mostrar aspectos interesantes del problema de regresión como será el del efecto del tamaño muestral y la variabilidad de la muestra considerada.

Para estos aspectos, se estudiará el comportamiento de los estimadores con distintos tamaños de muestra (n = 25, 50, 100 y 500) y con distintas desviaciones típicas para los residuos del modelo (0,3, 0,4 y 0,1). Tras esta comparativa, se pretende observar la convergencia de la curva teórica y asimismo ver cómo el problema de estimación se hace más difícil de resolver a medida que se va aumentando el valor de la desviación típica de los residuos del modelo.

Para realizar el planteamiento anterior se considerará el siguiente modelo de regresión:

$$Y = m(x) + \varepsilon \text{ donde } m(x) = sen(2x) + 2\exp(-16x^2)$$

donde x se genera según una distribución uniforme continua en el intervalo (−2, 2) y los residuos se consideran normales con media 0 y desviación típica σ.

En primer lugar, se empezará comparando el estimador polinómico local (**EPL**) con grados *p* = 0, 1, 3. Fijando el parámetro ancho de banda en *h* = 0,15. Y en segundo lugar, se comparará el **EPL** con los distintos métodos de selección para el ancho de banda (plug-in, CV, regla del pulgar), fijando ajustes de grado *p* = 1.

[1]Dicho estimador queda determinado por tres parámetros, fijados o definidos a priori para este experimento: el ancho de banda, la función núcleo y el grado p.

Para cuantificar la precisión de las estimaciones resultantes se utilizará como criterio de error la suma residual de cuadrados sobre una rejilla de puntos de estimación. De este modo se evaluará el estimador sobre una red de puntos $x_l \quad l = 1,..., ngrid,$ equiespaciada en (−2, 2) de tamaño *ngrid* = 500. Una vez calculadas las estimaciones sobre la rejilla se calcularán los errores con la fórmula: $1/500 \times \sum_{i=1}^{500} (m(x_l) - \hat{m}_h(x_l))^2$ y se compararan los resultados tomando la raíz cuadrada.

Bajo este diseñose considerará la estimación con diferentes casos tomando distintos tamaños muestrales (n= 25, 50, 100 y 500) y σ= 0,4. Para estos casos y tras implementar el código correspondiente, los resultados obtenidos se reflejan en conjunto en la siguiente **Figura 1**.

En este sentido, para el **caso 3**, como se puede observar en la **Figura 1**, el gráfico cuando p = 0 y p = 1 son muy parecidos los estimadores salvo en la frontera, debido a que p = 1 permite corregir de forma automática los efectos frontera. Los resultados para p = 3 muestran una mayor irregularidad.

Pero en el **caso 1**, **Figura 1**, cuando se ha disminuido el tamaño muestral a 25 se ve que los estimadores presentan bastantes irregularidades, sobre todo cuando se intenta ajustar un polinomio de grado alto (p = 3). Sin embargo, en el **caso 4**, **Figura 1**, cuando se considera un tamaño de muestra elevado se observa que los tres estimadores ajustados coinciden.

Por otro lado, también se ha ilustrado el comportamiento de los estimadores lineales locales con diferentes métodos de selección del parámetro de suavizado.

Esta aplicación comparativa se centra en el **caso 3,** que se simularon n = 100 datos con σ = 0,4, donde los métodos considerados son el selector basado en validación cruzada calculado usando la función **regCVBwSelC**(**cv**), el de tipo plug-in calculado usando **pluginBw** (**pi**) y el calculado según la simple regla del pulgar, ofrecido por la función **thumbBw**(**th**), todas ellas contenidas en la librería **locpol**. Y además, esta comparación de los métodos se realizará vía la raíz cuadrada del error definido anteriormente.

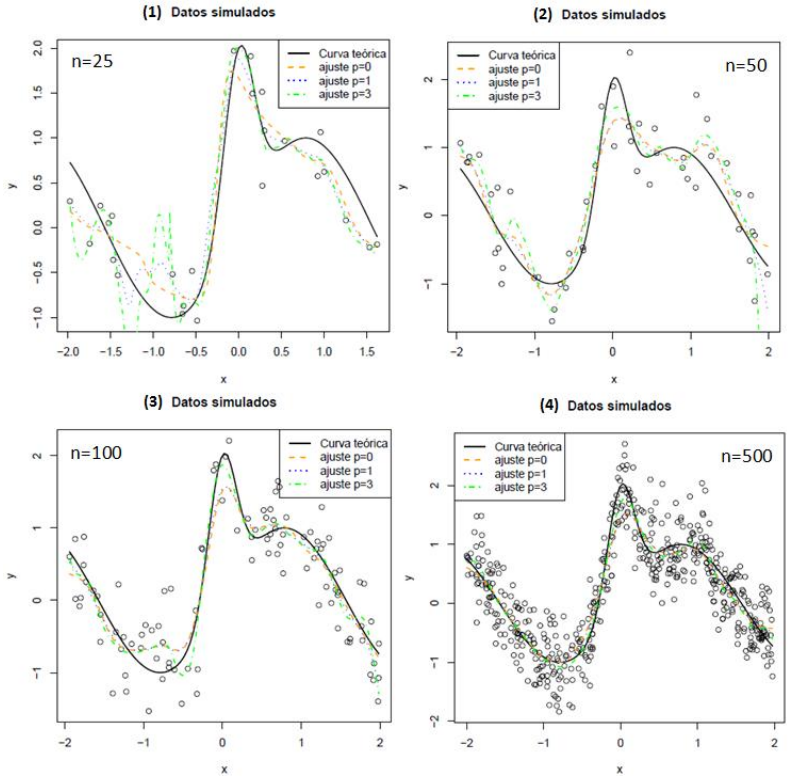

Figura 1: Estimación polinomial local a partir de datos simulados. Tamaños muestrales de 25, 50, 100 y 500 y desviación típica residual 0,4.

De esta manera, implementado el código correspondiente, los resultados obtenidos comparando con los tamaños muestrales n = 25, 50, 100, 500, según los diferentes selectores, son los siguientes:

	cv	th	pi
n= 25	0.3238807	0.2492679	NA
n= 50	0.2768001	0.2395432	0.2738346
n=100	0.2140991	0.2080944	0.2093302
n=500	0.07953835	0.09508265	0.08881742

De dichos resultados se puede observar que el comportamiento de los métodos plug-in es ligeramente superior a validación cruzada. No obstante la diferencia se hace menos patente en tamaños de muestra límite. También es de destacar que cuando se consideran pocos datos n = 25 no es posible el cálculo del selector de tipo plug-in. Esto es debido a que dichos métodos requieren estimaciones de las derivadas que no son posibles en este caso.

Y finalmente, se repite la comparación de los selectores variando la dificultad del problema de estimación. Esto se hará variando la desviación típica de los residuos del modelo, considerando $\sigma = 0.001, 0.1, 0.5$. Y el tamaño muestral se mantiene en n = 100. De este modo, los resultados obtenidos para los diferentes selectores son los siguientes:

	cv	th	pi

sigma= 0.001	0.01265973	0.08732214	0.03904589
sigma= 0.1	0.06306383	0.0912005	0.06364838
sigma= 0.5	0.2278124	0.1755693	0.1888333

Por tanto, la relación de los selectores cuando se muestran **distintos tamaños muestrales** (n=25, 50, 100, 500) fijando la desviación típica (σ = 0,4) son muy diferentes, permitiendo reflejar que los métodos de **tipo *plug-in*** se comportan ligeramente superior a los de validación cruzada, aunque estas diferencias son menos patente en tamaños de muestra límite.

En contraposición, se detecta aun más la dificultad de la estimación cuando se modifican las **desviaciones típicas** de los residuos del modelo (σ = 0.001, 0.1, 0.5) y se fija el tamaño muestral en un valor en concreto (n=100), arrojando en los resultados obtenidos que cuando existe una **altavariabilidadmuestral** (σ = 0.5) validación cruzada ofrece resultados poco representativos siendo el mejor ajuste el que viene mostrado por los selectores de **tipo *plug-in*** y en concreto, el de la sencilla regla del pulgar.

En resumen, con estos resultados obtenidos se ve un paralelismo entre aumentar el tamaño muestral y reducir la variabilidad residual. Si se observan estos datos cuando hay **alta variabilidad muestral** σ = 0.5 **validación cruzada ofrece resultados pobres**. Por eso, en este caso el **mejor ajuste** sería con los selectores de tipo plug-in y en particular con la sencilla regla del pulgar.

No obstante, nótese que a la vista de estos resultados aprioris se sabe que la **elección del ancho de banda** es uno de los aspectos cruciales del procedimiento de estimación para intentar buscar una adecuada **compensación entre sesgo y varianza**. Y además, se sabe que esta elección supone la complejidad del modelo.

También, esto mismo, ocurre con la **elección del grado de los ajustes polinomiales**. Es decir, la utilización de **ajustes de grado cero ó uno** nos darán estimaciones con **poca variabilidad**, muy suaves,pero con sesgos muy elevados. Por el contrario, cuando los ajustes son con **grados mayores** (dos ó tres) nos permitirán mayor adaptabilidad, o sea, menores sesgos pero, obtendremos **mayor varianza**.

Por eso, para este experimento se han fijado ajustes de grado p = 1, obteniendo de este modo varianzas pequeñas para poder obtener anchos de bandas aceptables que abarquen observaciones muy próximas al punto de estimación describiendo muy bien comportamientos locales y de esta manera poder comparar los resultados de los diferentes tipos de selectores calculados.

4. CONCLUSIONES

Con este experimento de simulación se pretende generar observaciones que puedan ser analizadas mediante el modelo previamente definido en base a los tres parámetros esenciales (el tamaño muestral, el tipo de rejilla y la función de tendencia).

El objetivo del experimento es cuantificar la bondad de las estimaciones, conociendo los modelos exactos, y además, mostrar aspectos interesantes del problema de regresión como son el del efecto del tamaño muestral y la variabilidad de la muestra considerada, permitiendo de esta forma visualizar los distintos comportamientos de los estimadores.

Además, desde esta comparativa, se pretende observar la convergencia de la curva teórica y asimismo ver cómo el problema de estimación se hace más difícil de resolver a medida que se va aumentando el valor de la desviación típica de los residuos del modelo.

En esta aplicación se ha utilizado el software R como entorno de análisis y programación estadística y en concreto, algunas de las librerías específicas del mismo.

Los resultados logrados mediante la aplicación de estas técnicas no paramétricas sobre los datos simulados son:

- ✓ Los estimadores polinomiales con grados menores (p=0 y p=1) para una muestra de **tamaño 100** son muy parecidos salvo en la frontera, debido a que cuando $p = 1$ se corrigen de forma automática los efectos frontera. En cambio, con grados mayores (p=3) los resultados del estimador muestran demasiadas irregularidades en la curva.
- ✓ Cuando se disminuye el **tamaño muestral a 25,** los estimadorespolinomiales presentan bastantes irregularidades sobre todo cuando se ajusta un polinomio de grado alto (p = 3).
- ✓ Los tres estimadores ajustados coinciden suficientemente bien sobre las observaciones cuando el tamaño muestral es bastante grande (**n=500**).
- ✓ El estimador polinomial local se comporta de forma distinta cuando se comparan varios tamaños muestrales y desviaciones típicas mediante la utilización de **diferentes métodos deselección** para el parámetro de suavizado reflejando resultados bastantes sorprendentes a los esperados.

En definitiva, se puede determinar que con estas confrontaciones realmente existe una semejanza entre aumentar el tamaño muestral y disminuir la variabilidad residual del modelo, y según los resultados recogidos, los métodos de tipo *plug-in* funcionan bastante mejor respecto a los de validación cruzada. También, se ha observado que la convergencia a la curva se alcanza cuando el tamaño muestral es bastante elevado (n=500) ya que, con tamaños menores se visualizan demasiadas irregularidades en la curva palpándose estas diferencias cuando se intentan ajustar polinomios de grado alto (p=3).

En esta línea exploratoria, una buena propuesta de investigación sería la aplicación de observaciones en **modelos aditivos** [17], cuya finalidad es la búsqueda de nuevos modelos ajustados que mejoren los resultados del estudio de una forma asequible para la rama de las ciencias sociales y de la biomedicina.

No obstante, para este experimento de simulación se puede concluir que los

métodos de regresión no paramétrica, en concreto, el estimador polimonial local ofrecen una buena vía de solución e interpretación como fuente de análisis primaria, corroborando los resultados que aparecen en [7] y [18].

BIBLIOGRAFÍA

[1] Boukichou-Abdelkader, N.; Montero-Alonso M. Á.; Muñoz-García, A. y Canário, P.N. (2014).**Regresión no paramétrica: estimador polinomial local**, En: Modelación matemática de fenómenos del medio ambiente y la salud (III), 46-52, Granada (España). ISBN: 84-616-7997-0.

[2] Wand, M. P. and Jones, M. C. (1995). **Kernel Smoothing**. Chapman and Hall, London.

[3] Fan, J. and Gijbels, I. (1996). **Local polynomial modelling and its applications**.Chapman and Hall, London.

[4] Loader, C. (1999). **Local Regression and Likelihood**.Springer, New York.

[5] Nadaraya, E.A (1964). On estimating regression.**Theory Probab.Appl**, 9, 141-142.

[6] Watson, G. S. (1964). Smooth regression analysis. **SankhyaSerie A**, 26, 101-116.

[7] Heckman, N.; Ramsay, J.O. (1996). **Spline smoothing with model based penalties**. McGillUniversity, unpublished manuscript.

[8] Ruppert, D.; Sheather, S. J.; Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. **Journal of the American Statistical Association**, 90, 1257-1270.

[9] DeBrabanter, K.; De Brabanter, J.; De Moor, B. and Gijbels, I. (2013).**Derivative Estimation withLocal Polynomial Fitting.**Journal of Machine Learning Research , 14, 281-301.

[10] Tsybakov, A.B. (2009). **Introduction to Nonparametric Estimation**.Springer.

[11] Zhou, S. and Wolfe, D.A. (2000). **On derivative estimation in spline regression**.Statist. Sinica, 10(1), 93-108.

[12] Vilar-Fernández, J.A. and Vilar-Fernández, J.M. (1998). **Recursive estimation of regressionfunctions by local polynomial fitting**. Ann. Ins. Stat. Math., 50(4), 729-754.

[13] Francisco-Fernández, M. and Vilar-Fernández, J. M. (2001). Local polynomial regression estimationwith correlated errors. Communications in Statistics: **Theory and Methods**, 30(7), 1271-1293.

[14] Su, L. and Ullah, A. (2008). **Local polynomial estimation of nonparametric simultaneous equationsmodels**. Journal of Econometrics, 144, 193-218.

[15] Hall, P. and Yatchew, A. (2007). Nonparametric estimation when data on derivatives are available.**Annals of Statistics**, 35, 300-323.

[16] Ichimura, H. and Todd, P. E. (2007). **Implementing Nonparametric and Semiparametric estimators**.Handbooks in Economics, 2, 5369-5468.

[17] Hastie, T. J. and Tibshirani, R. (1990). **Generalized additive models**.Chapman and Hall.Washington, D.C.

[18] Schimek, M. G. (2000). **Smoothing and regression:approaches, computation and application**. New York, Wiley.