Duodécimas Jornadas *"Investigaciones en la Facultad"* de Ciencias Económicas y Estadística, noviembre de 2007

Marí, Gonzalo
Cuesta, Cristina
Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística

COMPARACIÓN DE LA ESTIMACIÓN DE LA TASA DE OCUPACIÓN DE PLAZAS POR REGRESIÓN LOESS Y POR MUESTREO

1. INTRODUCCIÓN

Una de las utilidades de los modelos de regresión es reflejar el comportamiento de una variable respuesta a partir de una función que modele la relación existente entre la misma y variables explicativas. Debido a la rigidez de estos modelos, en muchas ocaciones los mismos no logran captar en forma correcta la relación existente. Sumado a esta dificultad, existen además los supuestos de distribución que deben realizarse sobre los modelos para que las inferencias sean válidas, los cuales no siempre se cumplen.

Una de las soluciones viene de la mano de herramientas que permitan cierta flexibilidad de los supuestos, además de representar en forma más fidedigna las relaciones existentes. Entre las mismas, y en el caso univariado, se pueden mencionar la construcción de curvas "suaves" que no hacen supuestos de antemano de la forma de la relación existente entre las variables explicativas y respuesta, ni tampoco sobre los supuestos distribucionales.

Las curvas "loess" (local regression) (Cleveland, 1979) se basan en ajustar modelos de regresión polinómicos locales para ajustar cada punto y unir luego las estimaciones. El resultado será una curva más o menos suave de acuerdo a la definición de cercanía de los puntos utilizados para estimar el dato en cuestión y a la función de ponderación que se aplique en dicha vecindad. El método de estimación de la curva es el de mínimos cuadrados ponderados, repitiendose la estimación en un número grande de puntos dentro del campo de variación de la variable explicativa.

El objetivo de esta técnica es hallar una curva que refleje el comportamiento de la relación, que no sea demasiado "suave" ni demasiado "rugosa". Uno de los factores que influyen en el grado de suavidad de la misma es la elección del ancho de ventana a considerar en la estimación de los puntos. Existen diferentes elecciones, pudiendo ser prefijada de antemano (bandwidth), fijando proporciones de datos a considerar en cada estimación (span), optando por anchos de ventana calculados en forma óptima, etc.

Este trabajo tiene por objetivo evaluar la utilización de ancho de ventana calculados en forma óptima para el ajuste de curvas loess en datos medidos a través del tiempo. Se analizan datos provenientes de la Encuesta de Ocupación Hotelera (EOH) que el Instituto Nacional de Estadística y Censos realiza mensualmente en distintas localidades de nuestro país. Se ajustará el comportamiento de la tasa de ocupación de plazas durante un periodo de tiempo y se evaluará si las curvas de distintos anchos caen dentro de intervalos de confianza estimados a partir de la teoría general de muestreo.

2. METODOLOGÍA

La relación entre una variable explicativa y una respuesta puede expresarse a través del modelo: $y_i = m(x_i) + \varepsilon_i$, donde la curva de regresión $m(x)$ es la esperanza condicional $m(x) = E(Y / X = x)$. Cuando un modelo de regresión paramétrico no es apropiado (porque no llega a "captar" adecuadamente la estructura que relaciona ambas variables), una alternativa es utilizar un modelo de regresión no paramétrica (que elimina la restricción paramétrica sobre $m(x)$)

Estos modelos sugieren ajustar polinomios locales de mayor grado ya que una constante local sólo tendría sentido sobre un pequeño vecindario (ancho de ventana muy pequeño). Así, el estimador de la regresión polinómica local es el que minimiza:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - ... - \beta_p (x - x_i)^p \right)^2 K\left(\frac{x - x_i}{h}\right) \qquad (2.1)$$

Como puede advertirse en la expresión (2.1), la estimación dependerá del grado del polinomio, de la función K de ponderación y del ancho de ventana *h*.

Una vez realizada la estimación para todos los valores de x (e incluso para otros valores que pertenezcan al campo de variación de la variable explicativa, para lograr una curva más suave), estas se unen formando la curva suavizada.

El ancho de ventana *h* puede ser obtenido por un procedimiento plug-in. Uno de los más conocidos que será utilizado en este trabajo es el desarrollado por Ruppert, Sheather y Wand (1995). El mismo es una función de propiedades desconocidas de *m*, las cuales son estimadas dividiendo el rango de la variable explicativa *x* en bloques y estimando *m* en cada bloque a través de un polinomio.

3. ENCUESTA DE OCUPACIÓN HOTELERA

Hacia fines del año 2003 la Secretaría de Turismo de la Nación (SECTUR) y el Instituto Nacional de Estadística y Censos (INDEC) firmaron un convenio para medir el impacto y la participación del turismo en el conjunto de la economía de la Argentina. Entre los operativos diseñados para tal fin se encuentra la Encuesta de Ocupación Hotelera (EOH) cuyo objetivo es medir el impacto del turismo internacional e interno sobre la actividad de los establecimientos hoteleros y para-hoteleros, la oferta y utilización de infraestructura, la evolución de tarifas, etc.. Se define como establecimientos "hoteleros" a aquellos categorizados como 1,2,3,4 y 5 estrellas y apart-hoteles, mientras que el grupo de establecimientos "para-hoteleros" lo conforman los hoteles sindicales, albergues, cabañas, bungalows, hospedajes, bed & breakfast, hosterías, residenciales, etc.

En el año 2004, la EOH se desarrolló con periodicidad mensual en 17 localidades, mientras que en el año 2005 se redefinió en 39 localidades de 6 regiones turísticas del país (determinadas por SECTUR).

Para la encuesta desarrollada durante el año 2004 se seleccionó, en cada una de las 17 localidades en forma independiente, una muestra estratificada considerando como estratos las categorías hotelero y para-hotelero. En el primero de los estratos, se incluyó con probabilidad 1 a aquellos establecimientos 4 y 5 estrellas, mientras que para las restantes categorías se contempló la selección de hoteles con probabilidad proporcional al total de plazas del mismo. Ésta estrategia de selección fue utilizada también para seleccionar establecimientos en el estrato para-hotelero.

Para el año 2005 se planteó la ampliación de los dominios de estimación, agregándose a las 17 localidades consideradas durante el 2004, la posibilidad de dar estimaciones en distintas regiones. Para ello, se seleccionaron 22 localidades que, sumadas a las 17 del año 2004, permiten dar estimaciones para regiones. En cada una de las 22 localidades, se consideró para la selección de establecimientos un diseño similar al utilizado en las 17 primeras localidades.

Los principales índices estudiados son: la tasa de ocupación de plazas (TOP), la tasa de ocupación de habitaciones (TOH) y las plazas por personal ocupado (PPO) que se definen como:

$$TOP = \frac{plazas\ ocupadas}{plazas\ disponibles \times días\ abiertos} \times 100$$

$$TOH = \frac{(unidades + habitaciones\ ocupadas)}{(unidades + habitaciones\ disponibles) \times días\ abiertos} \times 100$$

$$PPO = \frac{plazas\ disponibles}{personal\ ocupado}$$

4. APLICACIÓN

En el presente trabajo se va a evaluar la utilización de anchos de ventana óptimos obtenidos a partir del método plug-in en datos medidos a través del tiempo.

Se consideran las estimaciones de la TOP de la EOH en algunas de las 17 localidades medidas durante el año 2004 y 2005. Se considera también las estimaciones por intervalo de las mismas siendo el objetivo que las curvas que se obtengan caigan dentro de las estimaciones por intervalo obtenidas a partir de considerar el diseño muestral utilizado.

La estimación de la TOP para una localidad y un mes dado viene dada por

$$\hat{TOP} = \frac{\sum_h \sum_s w_{hk} y_{hk}}{\sum_h \sum_s w_{hk} x_{hk}}$$

donde w_{hk} es el ponderador, y_{hk} son las plazas ocupadas, y x_{hk} son las plazas disponibles correspondientes al establecimiento *k* del estrato *h*.

Se utiliza el método de expansion por series de Taylor para estimar la variancia de la estimación de la tasa. El estimador es

$$\hat{V}(\hat{TOP}) = \sum_h \hat{V}(\hat{TOP})$$

$$= \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (g_{hi} - \bar{g}_h)^2$$

donde

$$g_{hi} = \frac{\sum_{i=1}^{n_h} w_{hi}(y_{hi} - x_{hi}\hat{TOP})}{\sum_{h}\sum_{i=1}^{n_h} w_{hi}x_{hi}}$$

$$\bar{g}_h = \left(\sum_{i=1}^{n_h} g_{hi}\right) \Big/ n_h$$

Para cada una de las localidades consideradas y en cada mes, se utilizó una estimación de la tasa y de un intervalo de confianza del 95%.

Se ajusto cada una de las curvas considerando un ajuste polinomico local, utilizando una función de núcleo gaussiana, y un grado de polinomio igual a 2. El ancho de ventana fue obtenido a partir del procedimiento plug-in, lo que a priori permite obtener un ancho considerado óptimo. En muchas situaciones, el mismo se debio disminuir debido a que el requerimiento de caer dentro de los intervalos de confianza de las estimaciones puntuales no fue satisfecho.

Ciudad Autónoma de Buenos Aires

En los siguientes gráficos se muestran las curvas ajustadas utilizando un ancho de ventana óptima y tomando la mitad del mismo como ancho de ventana

Figura 4.1. Tasa de Ocupación de Plazas para Buenos Aires. Años 2004 y 2005. (h óptimo)

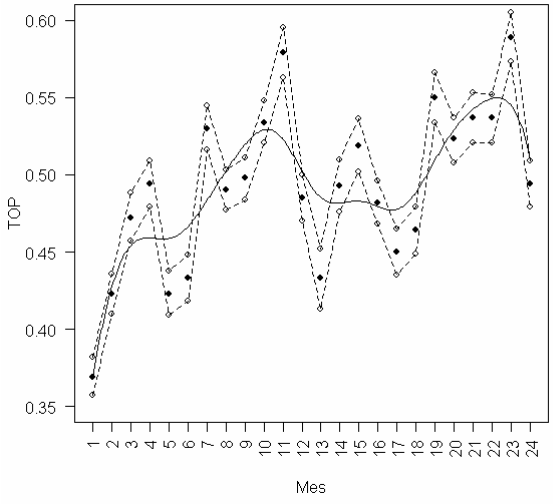
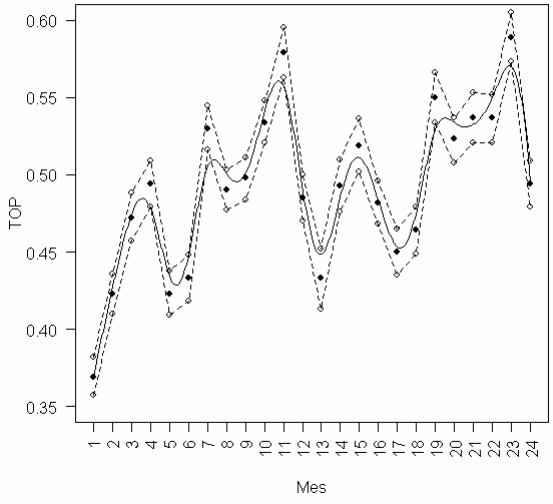


Figura 4.2 Tasa de Ocupación de Plazas para Buenos Aires. Años 2004 y 2005. (h óptimo/2)



Como puede observarse, al utilizar el h óptimo, la curva no logra captar los picos que presenta la TOP a través del tiempo. Esto es solucionado al tomar la mitad del h óptimo, quedando además la curva contenida en la mayor parte del recorrido dentro de las bandas de confianza

Puerto Iguazú

De igual forma se presentan los gráficos para la estimación de la TOP para los establecimientos de esta localidad durante los dos años considerados.

Figura 4.3. Tasa de Ocupación de Plazas para Puerto Iguazú. Años 2004 y 2005. (h óptimo)

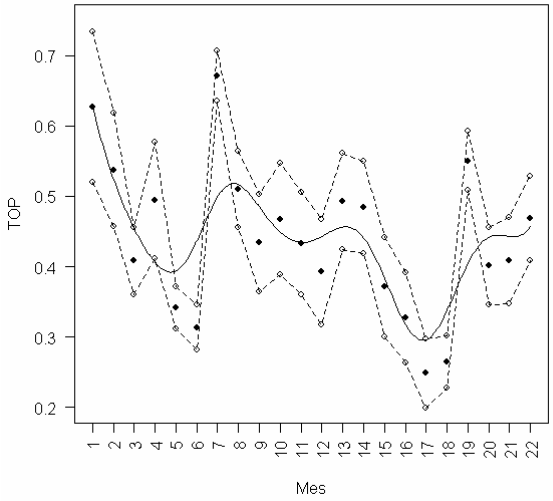
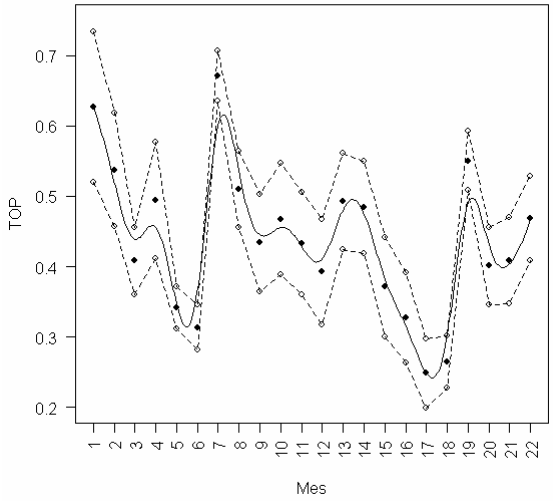


Figura 4.4 Tasa de Ocupación de Plazas para Puerto Iguazú. Años 2004 y 2005. (h óptimo/2)



En este caso al utilizar el h óptimo, la curva está contenida en gran parte de su recorrido dentro de las bandas de confianza, pero no logra captar las pequeñas variaciones que se presentan por ejemplo entre los meses 9 y 12. Esto es solucionado al tomar la mitad del h óptimo como ancho de ventana como se aprecia en la figura 4.4.

Salta

El ejemplo donde quizás sea más visible se presenta a continuación, donde para la ciudad de Salta se tuvo que considerar un cuarto del h óptimo para obtener una curva que reflejara la relación existente.

Figura 4.5. Tasa de Ocupación de Plazas para Salta. Años 2004 y 2005. (h óptimo)

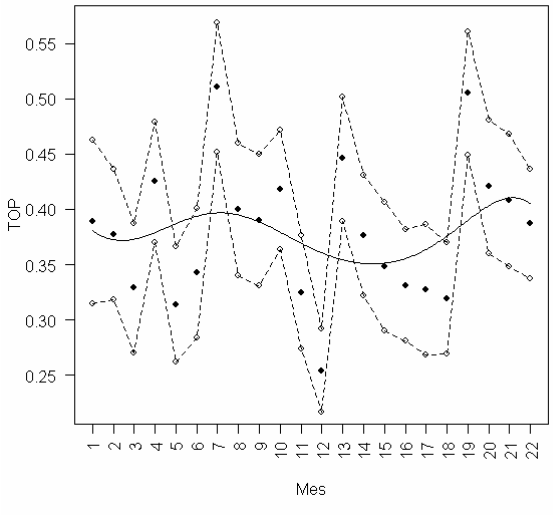


Figura 4.6. Tasa de Ocupación de Plazas para Salta. Años 2004 y 2005. (h óptimo/2)

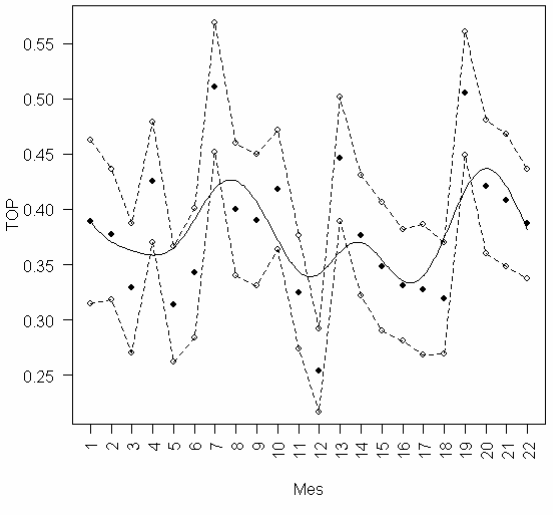
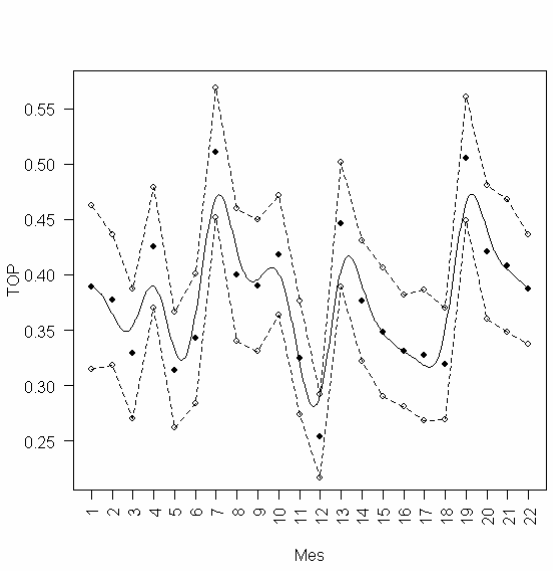


Figura 4.7. Tasa de Ocupación de Plazas para Salta. Años 2004 y 2005. (h óptimo/4)



Como puede observarse, la curva obtenida con el h óptimo no logra captar nada del comportamiento de la tasa a través de los dos años. Tampoco tomar como ancho de ventana la mitad del h óptimo soluciona este problema, ya que si bien logra mejorar la anterior, no logra quedar dentro de las bandas de confianza. El mejor ajuste se logra tomando la cuarta parte del h óptimo, obteniendo una curva que refleja el comportamiento de la tasa y que logra quedar dentro

5. DISCUSIÓN

En el presente trabajo se evaluó la utilización de utilizar regresión loess para estimar tasas de ocupación de plazas de la Encuesta de Ocupación Hotelera y su comparación con estimaciones surgidas a partir de la teoría del muestreo. Se utilizó un procedimiento plug-in para la obtención de anchos de ventana óptimo para la estimación de curvas de regresión polinómicas locales en datos obtenidos a través del tiempo.

Se presentaron ajustes en tres localidades donde el ancho óptimo de ventana daba como resultado una curva suave que no lograba captar la relación existente, además de no quedar contenida en las bandas de confianza definidas por las estimaciones obtenidas a partir de la muestra.

Esta dificultad se vio subsanada en parte al considerar la mitad del ancho óptimo, si bien en algunas localidades se tuvo que tomar una fracción aun mayor del mismo.

En estudios futuros se contemplará la adaptación de los métodos de estimación de las curvas para considerar la estructura de correlación subyacente por tratarse de datos medidos a través del tiempo, y verificar si la debilidad observada en este trabajo respecto a la utilización del h óptimo puede ser corregida de esa forma.

REFERENCIAS BIBLIOGRÁFICAS

Browman, A.W., Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-plus illustrations*. Oxford University Press.

Cleveland W. (1979) *Robust Locally Weighted Regression and Smoothing Scatterplots*. Journal of the American Statistical Association. Vol 74 pp 829-836

Fox, J. 2000. *Nonparametric Simple Regression: Smoothing Scatterplots*. Sage University Paper

Fox, J. 2000. *Multiple and Generalized Nonparametric Regression*. Sage University Paper

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Simonoff, J. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Venables, W.N., Ripley, B.D. (1999). *Modern Applied Statistics with S-plus*. New York: Springer-Verlag.

FUENTE

Encuesta de Ocupación Hotelera (EOH), 2005. Instituto Nacional de Estadística y Censos (INDEC).