Cook County, IL: Índices de precios de viviendas

Luis Alejandro Rubiano Guerrero
202013482
la.rubiano@uniandes.edu.co

Andres Felipe Rosas Castillo
202013471
a.rosasc@uniandes.edu.co

Carlos Andrés Castillo Cabrera
202116837
ca.castilloc1@uniandes.edu.co

I. Introducción

A través de este informe se construyen y se comparan **cuatro índices de precios de vivienda** para *Cook County, Illinois* para el periodo comprendido entre 2000 y 2020. Las metodologías correspondientes a cada uno de los índices son las siguientes: (i) y (ii) índices de precios hedónicos, (iii) Índice de precios de la vivienda usada (IPVU), y (iv) estimador de efectos fijos con errores agrupados a nivel de propiedad. El objetivo de este informe es comparar y contrastar los resultados obtenidos a través de cada una de las metodologías, así como discutir las ventajas y desventajas de cada una de ellas.

En la sección II. se describe el conjunto de datos utilizado, así como el proceso de limpieza y tratamiento de los datos. La sección III. detalla la metodología utilizada para cada uno de los índices. En la sección IV. se presentan los resultados obtenidos, así como un análisis comparativo entre los diferentes índices. Finalmente, en la sección V. se presentan las conclusiones del informe y en la sección VI. se proporciona información adicional para la reproducibilidad del análisis.

II. Datos y preparación

1. Descripción del conjunto de datos

El conjunto de datos utilizado en este informe es `dataTaller01_PriceIndeces.Rds`, el cual contiene información sobre ventas de viviendas en *Cook County, Illinois* entre los años 2000 y 2020. En total, el conjunto de datos contiene 427.649 observaciones y 31 variables. Estas variables incluyen el `pin` (identificador único de la propiedad), `year` (fecha de venta de la vivienda), `sale_price` (precio de venta de la vivienda), `township_code` (código local correspondiente al *township* donde se ubica la propiedad), así como 27 covariables continuas y categóricas de características estructurales y de ubicación de las viviendas.

2. Limpieza y tratamiento

Se encontraron valores faltantes para 1060 de las observaciones, en las variables `building_sqft`, `num_bedrooms`, `num_rooms`, `num_full_baths`, `num_half_baths` y `num_fireplaces`. Estas observaciones fueron eliminadas del conjunto de datos.

Adicionalmente, después de eliminar las anteriores observaciones, se encontraron 151.929 observaciones con valores faltantes en la variable `land_sqft`, lo cual representa aproximadamente el 36% del total de las observaciones. Estas fueron tratadas de manera diferente dependiendo de la metodología utilizada para construir los índices de precios de vivienda.

Adicionalmente se creó la variable `log_sale_price`, la cual corresponde al logaritmo natural del precio de venta de la vivienda, la cual es la variable dependiente de interés en los respectivos índices de precios.

III. Metodología

En todas las metodologías se normaliza el índice tomando como base el año 2000, es decir, el índice toma valor 100 en el año 2000.

1. Índice hedónico

Un índice hedónico se encuentra definido por la siguiente regresión.

$$\log(P)_{it} = \sum_{t=t_0}^{T} \delta_t D_{it} + \sum_{j=1}^{h} \beta_j H_{ij} + \sum_{k=1}^{n} \beta_k N_{ik} + \mu_{it}$$

En donde:

- $\log P_{it}$ es el logaritmo natural del precio de venta de la vivienda i en el tiempo t, $(t = t_0, \ldots, T)$.
- D_{it} es una variable indicadora de la venta de la vivienda i en el tiempo t.
- H representa características estructurales de la vivienda.
- N representa características de la ubicación de la vivienda.
- μ_{it} es el término de error, el cual se asume que es idénticamente distribuido, $\mu_{it} \sim N(0, \sigma^2)$.

En este caso estimamos dos índices hedónicos diferentes, en donde $t_0 = 2000$ y $T = 2020$.

En el primer modelo las características estructurales H son: `class`, `year_built`, `building_sqft`, `land_sqft`, `num_bedrooms`, `num_rooms`, `num_full_baths`, `num_half_baths`, `num_fireplaces`, `type_of_residence`, `construction_quality`, `attic_finish`, `garage_attached`, `garage_area_included`, `garage_size`, `garage_ext_wall_material`, `attic_type`, `basement_type`, `ext_wall_material`, `central_heating`, `basement_finish`, `roof_material`, `renovation`, `recent_renovation`, `porch`, `central_air`

y en el segundo modelo son las mismas características estructurales pero sin incluir `land_sqft`.

en ambos modelos las características de ubicación N son: `township_code, site_desirability`.

La diferencia entre los modelos es que en el primero se incluye la covariable `land_sqft`, pero se eliminan las observaciones con valores faltantes en esta variable, mientras que en el segundo modelo no se incluye esta covariable, pero se utilizan todas las observaciones.

Asimismo, de manera preventiva se agrupan los errores a nivel de `township_code`, es decir, se asume que los errores pueden estar correlacionados dentro de cada *township*, pero son independientes entre diferentes *townships*.

Finalmente, para estimar los índices se calcula $I_t = 100 \cdot \exp(\beta_t - \beta_{2000}) = 100 \cdot \exp(\beta_t)$, y se calculan sus errores estándar utilizando el método del delta, es decir $\text{var}(I_t) \approx (100 \cdot \exp(\beta_t)) \cdot \text{var}(\beta_t)$.

2. Índice de ventas repetidas (IPVU)

En este índice requiere identificar viviendas que hayan sido vendidas por lo menos dos veces dentro del periodo de estudio, basado en la metodología de *Case y Shiller* (1989).

Por lo tanto, filtramos el conjunto de datos para incluir únicamente aquellas viviendas que han sido vendidas más de una vez durante el periodo de estudio y que no hayan presentado modificaciones significativas en su estructura física. Para lo último, descartamos las ventas tales que la variable `recent_renovation` sea `TRUE`. La base de datos resultante tiene 233.233 observaciones.

En este modelo se asume que el comportamiento del precio de la misma vivienda P_t es un proceso estocástico dado por:

$$\ln(P_{i,t}) = \beta_t + H_{i,t} + N_{i,t}$$

En donde β_t corresponde al índice del precio, $H_{i,t}$ es una caminata aleatoria gaussiana que describe como el cambio del precio de una vivienda individual se desvía en el tiempo respecto a la variación del índice de mercado, y $N_{i,t}$ son errores que se asumen normales y representa las diferencias idiosincrásicas de las propiedades en un momento del tiempo.

Para este índice se crea una variable que representa el cambio porcentual total en el precio de una vivienda entre dos transacciones. Esta variable se define como:

$$\begin{aligned}\Delta V_i &= \ln(P_{i,t}) - \ln(P_{i,s}) \\ &= \beta_t - \beta_s + H_{i,t} - H_{i,s} + N_{i,t} - N_{i,s}\end{aligned}$$

Sobre los términos de perturbación se asume que

$$\begin{aligned}\mathbb{E}[H_{i,t} - H_{i,s}] &= 0(\text{media cero}) \\ \text{Var}(H_{i,t} - H_{i,s}) &= A(t-s) + B(t-s)^2 (\text{var cuadrática en el tiempo}) \\ \mathbb{E}[N_{i,t}] &= 0(\text{media cero}) \\ \text{Var}(N_{i,t}) &= c(\text{varianza constante}) \\ \mathbb{E}[H_{i,t} N_{i,t}] &= 0(\text{independencia})\end{aligned}$$

Partiendo de lo anterior, los índices de precios se estiman en tres etapas:

1) Primera etapa: Se estiman los β iniciales y los errores. Para una venta repetida de una vivienda i se estima el cambio porcentual de la siguiente forma:

$$\begin{aligned}\Delta V_i &= \sum_{t=t_0}^{T} \ln(P_{i,t}) D_{it} \\ &= \sum_{t=t_0}^{T} \beta_t D_{it} + \varepsilon_i\end{aligned}$$

En donde D_{it} es una variable indicadora que toma valor 1 cuando el precio de la vivienda i es observado por segunda vez en t, -1 si el precio de la vivienda i fue observado por primera vez en t, y cero de lo contrario. Aquí β_t se estima a través de mínimos cuadrados ordinarios.

2) Segunda etapa: Se estima la varianza de la caminata aleatoria.
En esta etapa se estiman los coeficientes A, B y c a través de mínimos cuadrados ordinarios, mediante la expresión:

$$\mathbb{E}[\varepsilon_i^2] = A(t-s) + B(t-s)^2 + c$$

Es decir, se estima la varianza de los errores al cuadrado como función cuadrática del tiempo entre ventas. En este caso $\mathbb{E}[\varepsilon_i^2]$ es la varianza muestral de los errores estimados en la primera etapa. Si los coeficientes estimados A y B son positivos y significativos, se debe continuar con la tercera etapa, en tal caso la raíz cuadrada de los valores estimados por la ecuación se utiliza para ponderar por mínimos cuadrados generalizados.

3) Tercera etapa: Se re-estiman los β utilizando mínimos cuadrados generalizados.
Es decir, se estima

$$\frac{\Delta V_i}{\sqrt{\hat{\varepsilon_i^2}}} = \sum_{t=t_0}^{T} \frac{\beta_t}{\sqrt{\hat{\varepsilon_i^2}}} D_{it} + \frac{\varepsilon_i}{\sqrt{\hat{\varepsilon_i^2}}}$$

Posteriormente, los indices y los errores estándar se calculan de forma similar a la metodología de índices hedónicos.

3. Estimador de efectos fijos con errores agrupados