Introducción al método de regresión para Investigación Urbana

Métodos cuantitativos aplicados a estudios urbanos II - MEU UTDT

Ricardo Pasquini - rpasquini@utdt.edu

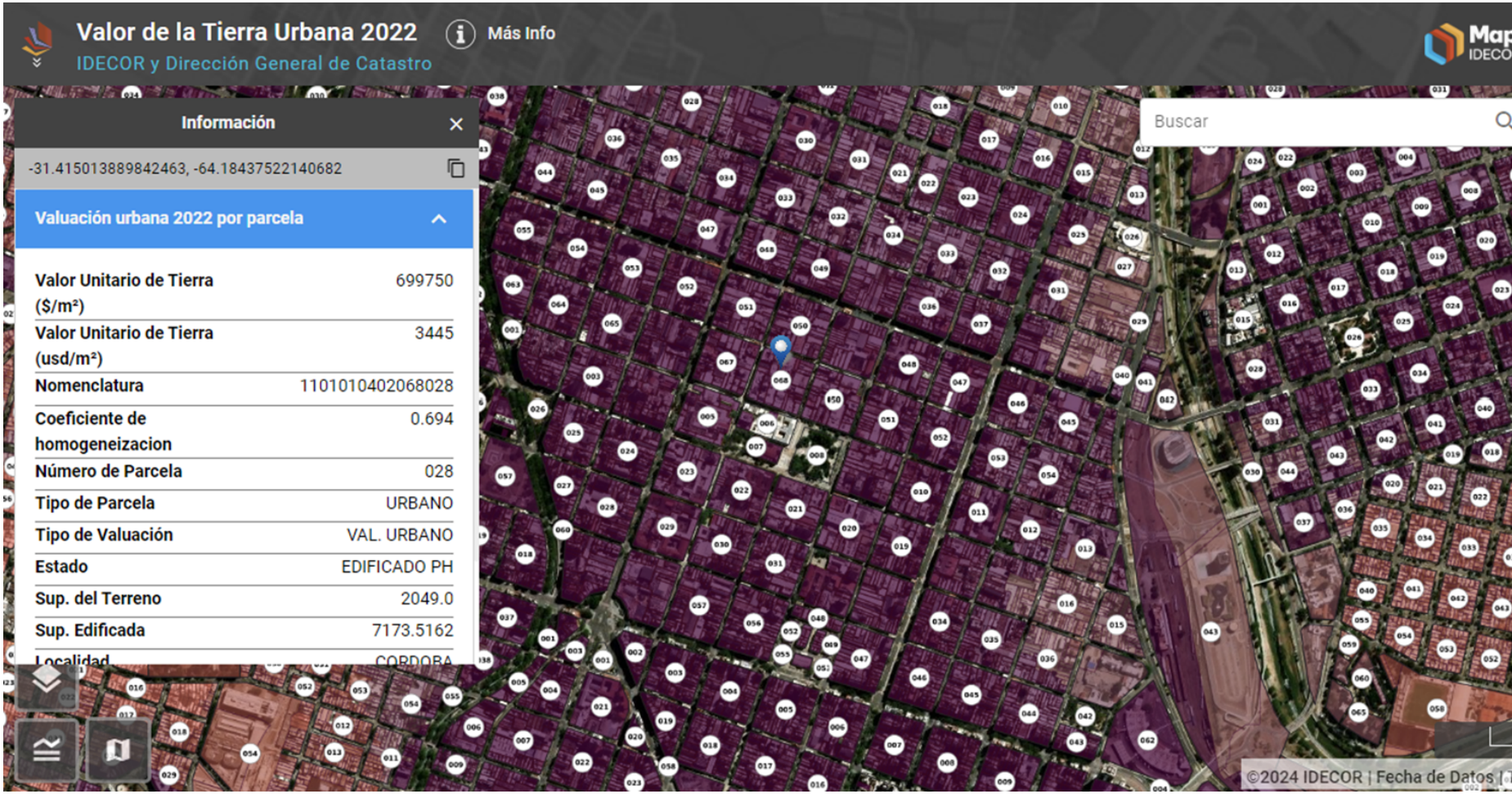
June 8, 2025

Usos de los Modelos de Regresión

1. Predicción
2. Extrapolación (una forma de predicción)
3. Inferencia causal
   - Experimental
   - Cuasi-experimental y observacional

Predicción

Valuación de terrenos



https://mapascordoba.gob.ar/viewer/mapa/401

Predicción

Recomendación de valor de alquiler

Startups

Airbnb Adds A Pricing Recommendation Tool For Renters

Matthew Lynley / 10:34 AM PDT • June 4, 2015

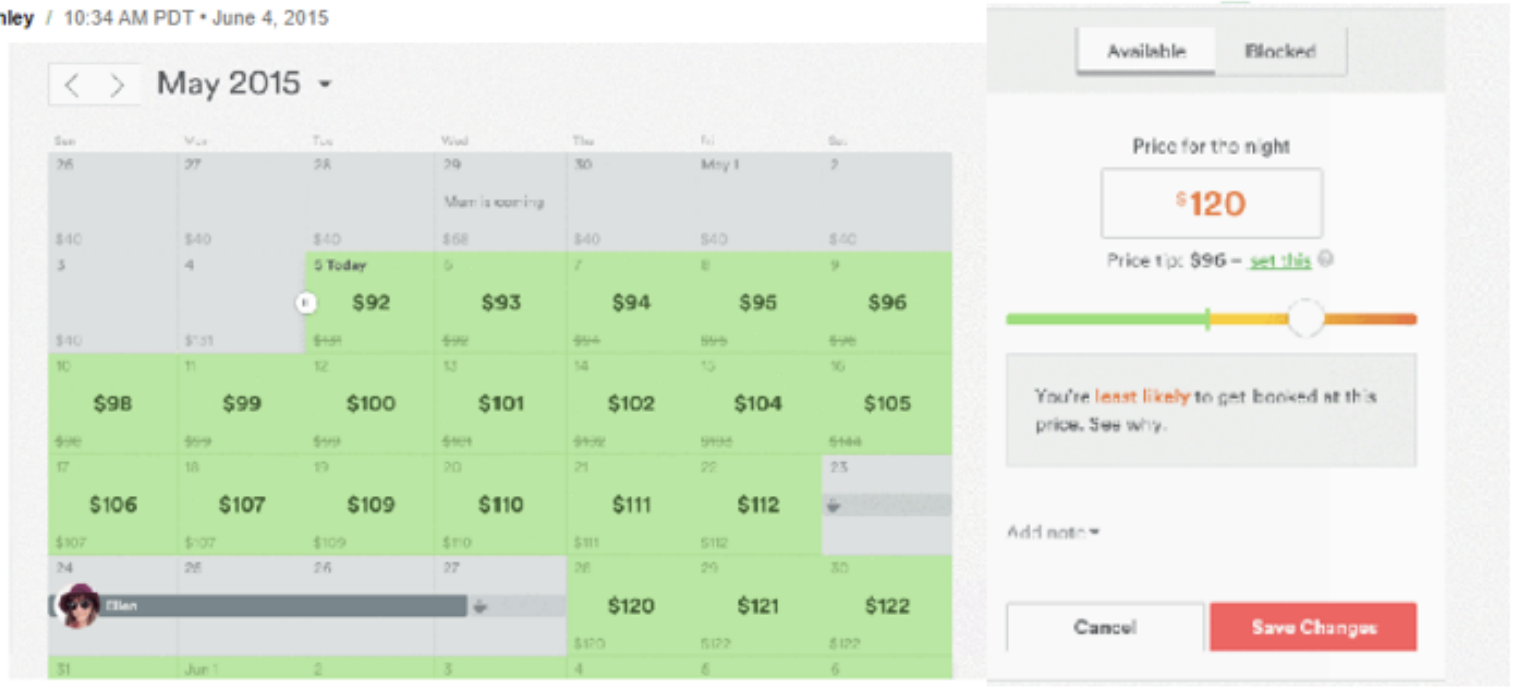


Image Credits: Airbnb

https://techcrunch.com/2015/06/04/airbnb-adds-a-pricing-recommendation-tool-for-renters/

Extrapolación

Extrapolación entre subgrupos

Forecasting elections with non-representative polls

Wei Wang [a,*], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a,c]

A B S T R A C T

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates which are in line with the forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle. We conclude by arguing that non-representative polling shows promise not only for election forecasting, but also for measuring public opinion on a broad range of social, economic and cultural issues.

Extrapolación

Extrapolación entre subgrupos

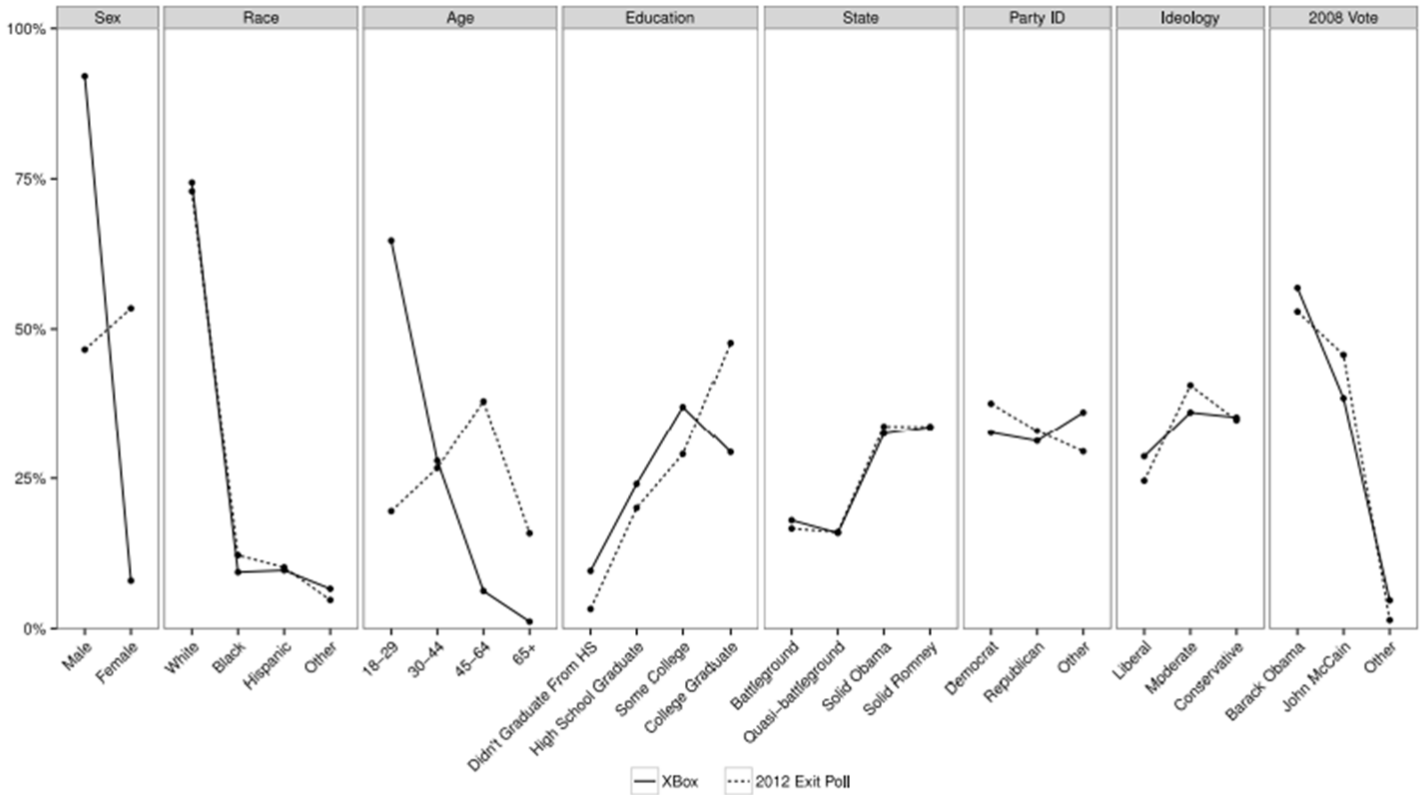
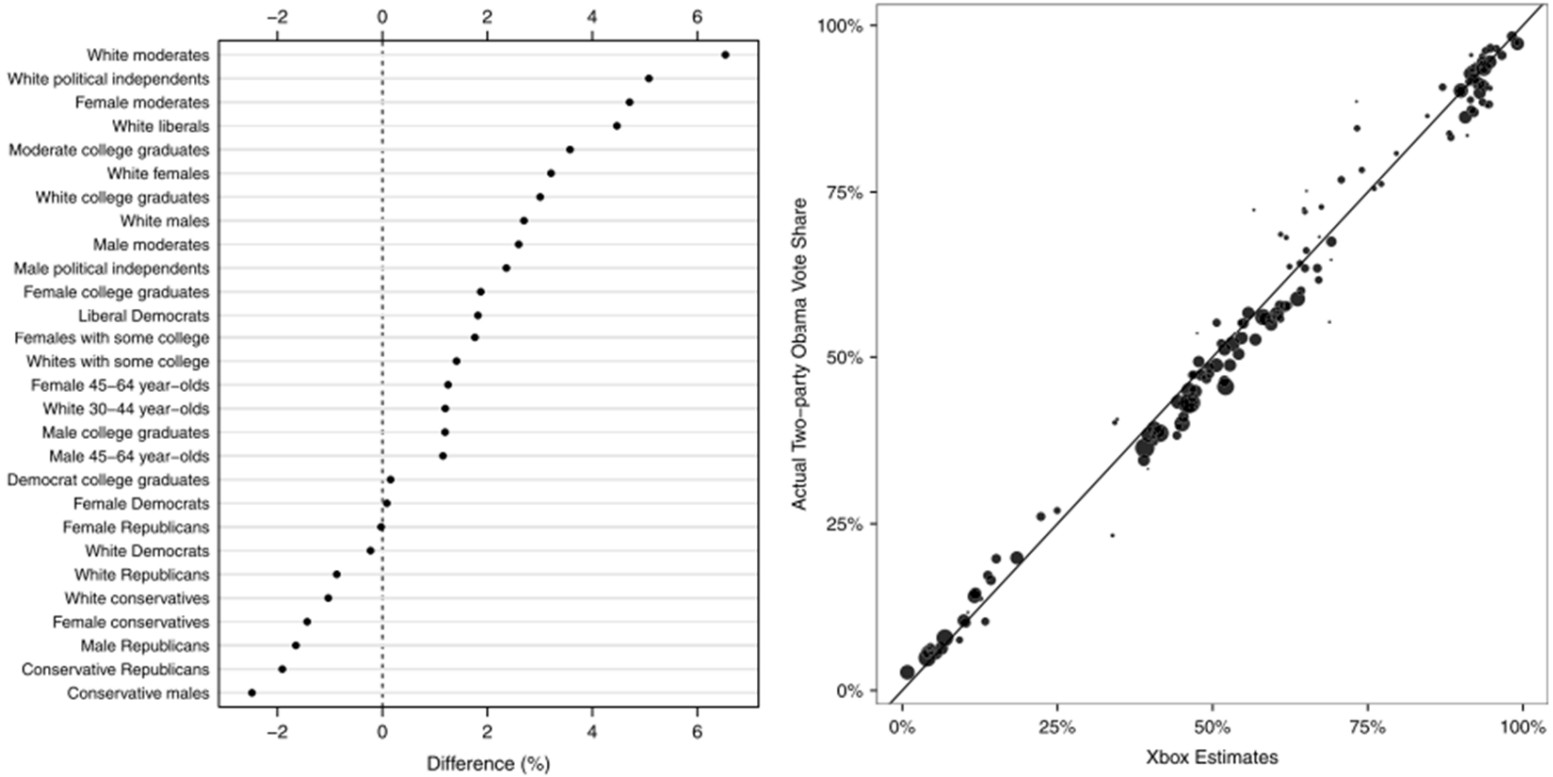


Fig. 1. A comparison of the demographic, partisan, and 2008 vote distributions in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). As one might expect, the sex and age distributions exhibit considerable differences.

Extrapolación

Extrapolación entre subgrupos

Fig. 5. Comparison of the two-party Obama vote share for various demographic subgroups, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election.



Inferencia Causal

Análisis de experimentos

Journal of Urban Economics
Volume 142, July 2024, 103612

JUE insight: The unintended effect of Argentina's subsidized homeownership lottery program on intimate partner violence

Bruno Cardinale Lagomarsino [a], Martin A. Rossi [b]

Abstract

We study a natural experiment in Argentina, where low-income women were selected through a lottery system to receive a house and a heavily subsidized long-term mort- gage. We exploit the random assignment to estimate the causal link between sub- sidized homeownership programs and intimate partner violence (IPV). Our analysis utilizes administrative records of the population of women applicants to assess the im- pact of homeownership on IPV, differentiating between women under joint-ownership contracts with their partners and those under single-ownership contracts. We find that the program causes an increase in IPV for women under joint-ownership contracts and a decrease in IPV for women under single-ownership contracts. Our results highlight the importance of considering the design of subsidized homeownership programs and explicitly incorporating measures to facilitate exit from conflicting relationships.

https://www.sciencedirect.com/science/article/abs/pii/S0094119023000827

Inferencia Causal

Análisis de experimentos

	Dependent variable: IPV							
	Joint ownership		Women ownership		Joint ownership		Women ownership	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beneficiary	0.173***	0.196***	-0.113	-0.116*	0.072	0.069	-0.608***	-0.501***
	(0.049)	(0.047)	(0.070)	(0.069)	(0.122)	(0.127)	(0.172)	(0.146)
Control group mean	0.114	0.114	0.239	0.239	0.386	0.386	0.887	0.887
Control variables	No	Yes	No	Yes	No	Yes	No	Yes
Sample	Admin	Admin	Admin	Admin	Survey	Survey	Survey	Survey
Observations	290	290	147	147	85	85	37	37

Notes: Heteroskedastic-robust standard errors in parentheses. The 437 observations correspond to the sample of eligible non-attritor women. Dependent variable equals 1 if the women reported IPV in the administrative data and zero otherwise in columns 1 to 4. Dependent variable is equal to 1 if the woman reported IPV in the survey and zero otherwise in columns 5 to 8. The main independent variable equals 1 if the woman is a beneficiary of the program and zero otherwise. The instrument equals 1 if the woman was assigned by the lottery as a beneficiary of the program and zero otherwise. The set of controls is the same as in Table 3. *** p<0.01, ** p<0.05, * p<0.1.

Inferencia Causal

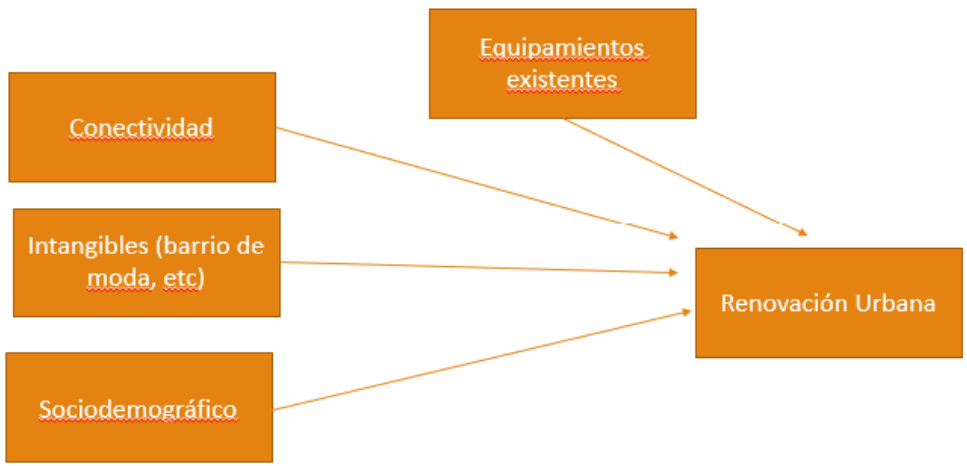
Análisis de experimentos

	Dependent variable: IPV							
	Joint ownership		Women ownership		Joint ownership		Women ownership	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beneficiary	0.173***	0.196***	-0.113	-0.116*	0.072	0.069	-0.608***	-0.501***
	(0.049)	(0.047)	(0.070)	(0.069)	(0.122)	(0.127)	(0.172)	(0.146)
Control group mean	0.114	0.114	0.239	0.239	0.386	0.386	0.887	0.887
Control variables	No	Yes	No	Yes	No	Yes	No	Yes
Sample	Admin	Admin	Admin	Admin	Survey	Survey	Survey	Survey
Observations	290	290	147	147	85	85	37	37

Notes: Heteroskedastic-robust standard errors in parentheses. The 437 observations correspond to the sample of eligible non-attritor women. Dependent variable equals 1 if the women reported IPV in the administrative data and zero otherwise in columns 1 to 4. Dependent variable is equal to 1 if the woman reported IPV in the survey and zero otherwise in columns 5 to 8. The main independent variable equals 1 if the woman is a beneficiary of the program and zero otherwise. The instrument equals 1 if the woman was assigned by the lottery as a beneficiary of the program and zero otherwise. The set of controls is the same as in Table 3. *** p<0.01, ** p<0.05, * p<0.1.

Modelos de regresión en el marco de la investigación

- Teoría $\rightarrow$ Proposiciones $\rightarrow$ Hipótesis
- Proposición: "Áreas con mayor accesibilidad experimentarán mayor renovación"
- Hipótesis: "Radios censales con menor distancia a estaciones de subte/tren tendrán mayor porcentaje de parcelas renovadas"
- Representación como un gráfico causal



Pérsico 2019

Representación del modelo de renovación como una función

$$renovación_i = f(conectividad_i, nivel\ socioeconómico_i, intangibles_i, equipamientos_i)$$

- i (la unidad de análisis) refiere a un radio censal.
- $f()$ es una función cuya forma desconocemos.
- *renovación* se aproximará con el % de parcelas renovadas (cambiaron uso, altura, etc.)
- *conectividad* se aproximará con distancia a estaciones de subte/tren etc.
- *nivel socioeconómico* se aproximará por alguna variable del censo (% población universitaria, % NBI, etc)
- *equipamientos*, aproximada, por ejemplo, con la distancia promedio a un equipamiento (escuelas, centros de salud, etc.)

Representación del modelo de renovación como una función lineal

$$\begin{aligned}\text{porcentaje\_parcelas\_renovadas}_i = \beta_0 + \beta_1 \text{distancia\_transporte}_i \\ + \beta_2 \text{NBI}_i + \beta_3 \text{clima\_barrial}_i \\ + \beta_4 \text{distancia\_equipamientos}_i + \varepsilon_i\end{aligned}$$

- Proponemos una forma lineal para la función en cuestión que podremos estimar.
  - El supuesto de linealidad es restrictivo, pero veremos que admite igualmente relaciones no lineales.
- Cada coeficiente β_j representa la *contribución* de la variable X_j a la variable dependiente Y.
- Estos coeficientes podrán ser estimados a partir de la muestra de datos.
- ε_i es el error aleatorio de la observación i.

Base de datos necesaria para la estimación

- En este caso, la base de datos tendría una observación por cada radio censal.
- Se corresponde con una base de tipo "corte transversal", esto es, una observación es una instancia de la unidad de análisis.

id_radio_censal	porcentaje_parcelas_renovadas	distancia_transporte	NBI	clima_barrial	distancia_equipamientos
1	0.911746	8.069655	0.892558	0.695538	1.995177
2	0.564947	4.596649	0.419448	0.849102	4.199008
3	0.710200	5.465771	0.780366	-0.293967	0.942703
4	0.382642	4.328160	0.476336	-0.071599	3.362302
5	0.200962	0.438999	0.497540	-1.517874	4.885035

Ejemplo: Modelización hedónica de precios de terrenos

- La modelización hedónica parte de la idea teórica que el valor de un terreno o inmueble es una función de sus características observables.

$$precio_i = f(\text{superficie}_i, \text{regulación}_i, \text{distanciaCDN}_i, \text{Amenidades}_i)$$

$$\begin{aligned} precio_i = \beta_0 + \beta_1 \text{superficie}_i + \beta_2 \text{regulación}_i \\ + \beta_3 \text{distanciaCDN}_i + \beta_4 \text{Amenidades}_i + \varepsilon_i \end{aligned}$$

Ejemplo modelo hedónico

Base de datos

FECHA	CALLE	NUMERO	M2	DOLARES	U_S_M2	FOT	COTIZ	BARRIOS	COMUNA	DIRECCIOI	CODIGO_P	CODIGO_POSTAL_ARGENTINO
13/06/201	15 DE NOV	1600	330	340000	1030,3	3	17,5	CONSTITU(	COMUNA 1	15 DE NOV	1130	C1130ABJ
8/6/2017	ANTARTIDA	1400	6885	35000000	5083,5	0	17,5	RETIRO	COMUNA 1	ANTARTIDA	1104	C1104ACN
9/6/2017	BELGRANC	800	1720	7500000	4360,5	0	17,5	MONSERR/	COMUNA 1	BELGRANC	1092	C1092AAU
12/6/2017	BRASIL	1500	800	600000	750	0	17,5	CONSTITU(	COMUNA 1	BRASIL 15(	1154	C1154AAZ
7/6/2017	CALVO, CA	614	254	800000	3149,6	0	17,5	SAN TELMC	COMUNA 1	CALVO, CA	1102	C1102AAN
12/6/2017	CALVO, CA	900	135	280000	2074,1	5	17,5	CONSTITU(	COMUNA 1	CALVO, CA	1102	C1102AAR
12/6/2017	CALVO, CA	1100	458	650000	1419,2	5	17,5	CONSTITU(	COMUNA 1	CALVO, CA	1102	C1102AAV
12/6/2017	CALVO, CA	1500	1356	2200000	1622,4	0	17,5	CONSTITU(	COMUNA 1	CALVO, CA	1102	C1102ABD
9/6/2017	CALVO, CA	1600	275	390000	1418,2	0	17,5	CONSTITU(	COMUNA 1	CALVO, CA	1102	C1102ABF
12/6/2017	CALVO, CA	1600	520	460000	884,6	3,4	17,5	CONSTITU(	COMUNA 1	CALVO, CA	1102	C1102ABF
12/6/2017	CASEROS /	1701	1200	2700000	2250	0	17,5	CONSTITU(	COMUNA 1	CASEROS /	1152	C1152ABA
12/6/2017	CEVALLOS	200	620	450000	725,8	4	17,5	MONSERR/	COMUNA 1	CEVALLOS	1077	C1077AAD
8/6/2017	CEVALLOS	300	1400	2100000	1500	0	17,5	MONSERR/	COMUNA 1	CEVALLOS	1077	C1077AAF
9/6/2017	CEVALLOS	345	1400	2100000	1500	0	17,5	MONSERR/	COMUNA 1	CEVALLOS	1077	C1077AAG

Figure: Base de datos de terrenos

- En este caso, la base de datos tendría una observación por cada radio censal.
- Se corresponde con una base de tipo "corte transversal", esto es, una observación es una instancia de la unidad de análisis.

Notación para modelización usando regresión

$$\underbrace{Y_i}_{\text{Variable a explicar o Dependiente}} = \underbrace{\beta_0}_{\text{Constante o intercepto}} + \beta_1 \cdot \underbrace{X_{1,i}}_{\text{Variable explicativa 1 (independiente)}} + \beta_2 \cdot \underbrace{X_{2,i}}_{\text{Variable explicativa 2 (independiente)}} + \cdots + \underbrace{\varepsilon_i}_{\text{Error}}$$

- $\beta_0, \beta_2, \ldots, \beta_k$ son los coeficientes a ser estimados.
- Una vez estimados los denotamos con $\hat{\beta}_0, \hat{\beta}_2, \ldots, \hat{\beta}_k$.
- Los coeficientes medirán la contribución a la variable explicada por cada unidad de la variable explicativa.

Explicación y Predicción

- **Explicación**
- Los coeficientes estimados $\hat{\beta}_0, \hat{\beta}_2, \ldots, \hat{\beta}_k$ son la base para la explicación.
- Si bien estrictamente hablando, los coeficientes solo capturan variaciones, y por lo tanto no representan necesariamente causalidad, son la base bajo las cuales buscaremos identificar efectos causales.
- **Predicción**
- La predicción se obtiene reemplazando los coeficientes estimados en la ecuación, y utilizando estos componentes para calcular el valor de que predice el modelo.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \cdots + \hat{\beta}_k X_{k,i}$$

Medidas de Bondad de Ajuste

R^2

- El R^2 es una medida de la bondad de ajuste del modelo en su conjunto.
- Se define como el cociente entre la varianza explicada por el modelo y la varianza total de la variable dependiente.
- Se interpreta como la proporción de la varianza de la variable dependiente que es explicada por el modelo.

Bondad de Ajuste: R^2

- Puesto que vale que

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i$$

- También vale que la varianza de Y_i es la suma de la varianza de la predicción y la varianza de los residuos.
- Esta igualdad nos permite definir una medida de la bondad de ajuste del modelo: el R^2 es la proporción de la varianza de la variable dependiente que es explicada por el modelo.

$$1 = \frac{Var(\hat{Y}_i)}{Var(Y_i)} + \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)}$$

$$R^2 = \frac{Var(\hat{Y}_i)}{Var(Y_i)} = 1 - \frac{Var(\hat{\varepsilon}_i)}{Var(Y_i)}$$

- Notar que es un valor entre 0 y 1.

Medidas de Bondad de Ajuste

Error Cuadrático Medio

- El error cuadrático medio (MSE) es otra medida de la bondad de ajuste del modelo en su conjunto.
- Es una medida que busca cuantificar la magnitud de los errores.

$$MSE = \frac{\sum \hat{\varepsilon}_i^2}{n - (k+1)}$$

- El $-(k+1)$ en el denominador es una corrección estadística al promedio, dada por el número de coeficientes estimados.
- La raiz cuadrada del MSE (o RMSE) es una medida que puede compararse en magnitud con la variable dependiente.