

Regresión Avanzada

Proyecto Final

Mario Becerra Contreras
Edgar López

Otoño 2017

1. Introducción

En este trabajo se presentan distintos modelos para el precio de casas y departamentos en la ciudad de Nueva York, esto con el objetivo de establecer un rango de precios de venta de una casa en una primera instancia, es decir, poder dar una primera cotización acerca de su precio sin conocer demasiados detalles acerca de ella. Un aspecto importante que se debe considerar es que aunque dos casas cuenten con características similares, sus precios pueden variar de manera sustancial dependiendo su ubicación geográfica, por lo que es importante que la cotización de venta de una casa esté en línea con los precios de la zona.

Para generar un modelo que sea útil a una inmobiliaria o agente de ventas, incluyendo los aspectos descritos anteriormente, es necesario utilizar técnicas estadísticas que permitan explicar el precio de venta de una casa a partir de una serie de variables que representen características de esta, considerando la variabilidad de precios que existe por su ubicación geográfica.

Los datos fueron obtenidos de Kaggle¹, una plataforma de concursos de modelado predictivo y aprendizaje estadístico. Los datos de Kaggle, a su vez, provinieron de la página oficial del departamento de finanzas de la ciudad de Nueva York².

Los datos tienen variables relacionadas con las casas y sus ventas, como distrito (*borough*), vecindario (*neighborhood*), código postal (*zip code*), dirección, precio de venta del inmueble, fecha de venta, tipo de inmueble, tamaño del inmueble, etc. No se tiene información muy específica como número de cuartos o de baños.

Se tienen 84,548 observaciones de un periodo de 12 meses (de septiembre de 2016 a agosto de 2017), las cuales no solo incluyen información de inmuebles residenciales, sino todo tipo de bienes raíces, por lo que se filtraron los datos para obtener solamente las observaciones correspondientes a casas. Además, había muchos datos que tenían como precio de venta 0, lo cual puede ser por herencia de padres a hijos o algún otro tipo de traspaso sin dinero³. También existían ventas con valores no creíbles, como unos pocos miles de dólares, por lo que tampoco se tomaron en cuenta para este trabajo; además, también se filtraron las observaciones que no tenían información sobre el código postal. Despues de este filtrado, quedaron 25,299 observaciones.

En las siguientes subsecciones se muestra el análisis exploratorio de los datos.

2. Datos

Antes de pasar al análisis de los datos, hay que mencionar que la ciudad de Nueva York está dividida en 5 distritos (*boroughs*), los cuales a su vez están divididos en 42 vecindarios, los cuales están divididos en 178 códigos postales. Los datos tienen información de venta en 154 códigos postales en todos los vecindarios y, por ende, en todos los distritos.

¹<https://www.kaggle.com/new-york-city/nyc-property-sales>

²<http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>

³http://www1.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf

2.1. Gráficas univariadas

La principal variable de interés es el precio en dólares de venta de las casas en Nueva York. En la figura 1 se muestra una gráfica de frecuencias absolutas con y sin la transformación logarítmica. Como se puede observar, los precios de venta se asemejan a una distribución exponencial o gamma, pero aplicando la transformación logarítmico, los datos se asemejan a una muestra de una distribución normal, por lo que se usará esta transformación en los modelos posteriores.

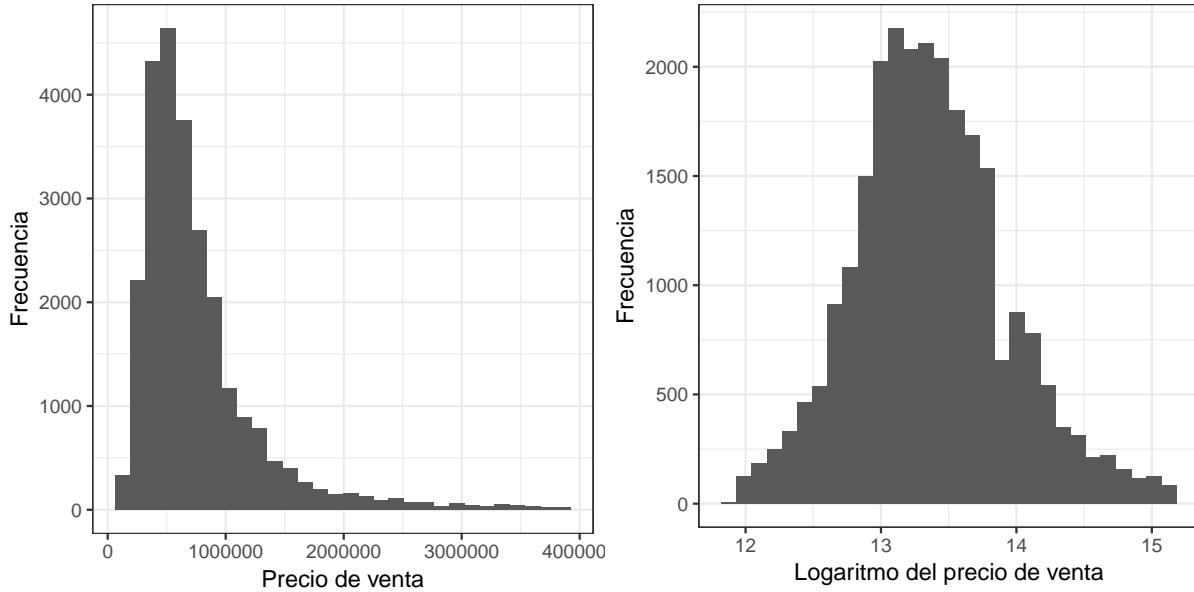


Figura 1: Histogramas del precio de venta en escala original y en escala logarítmica

Otra de las variables de interés es la superficie total que esta medida en pies cuadrados. La figura 2 muestra las gráficas de frecuencias absolutas para esta variable con y sin transformación logarítmico. De igual manera que el precio de ventas, sería más conveniente usar los datos usando la transformación logarítmico pues muestran un comportamiento semejante a una muestra de una distribución normal.

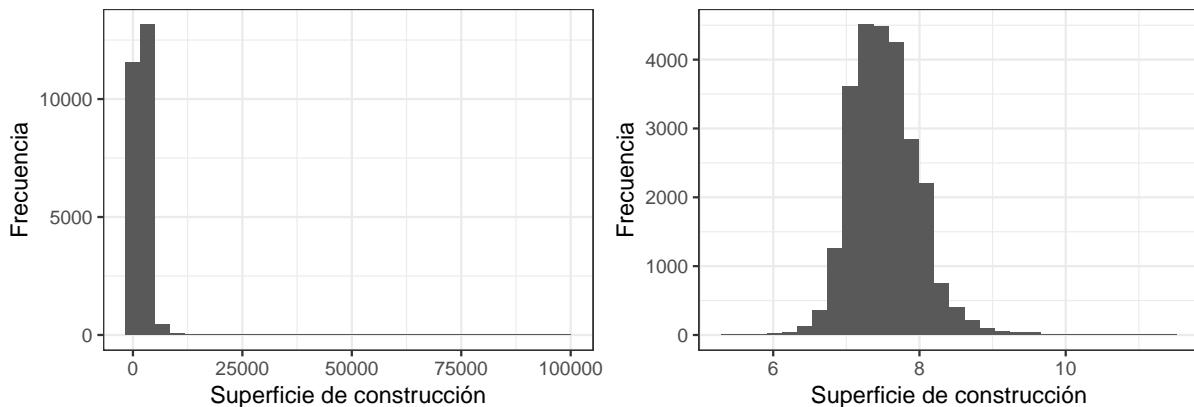


Figura 2: Histogramas de la superficie de construcción en escala original y en escala logarítmica

Finalmente, la variable de superficie del terreno en pies cuadrados se muestra en la figura 3. En este caso también sería conveniente usar la transformación logarítmico en los datos pues mejora la distribución muestral y reescala los datos a una escala más pequeña.

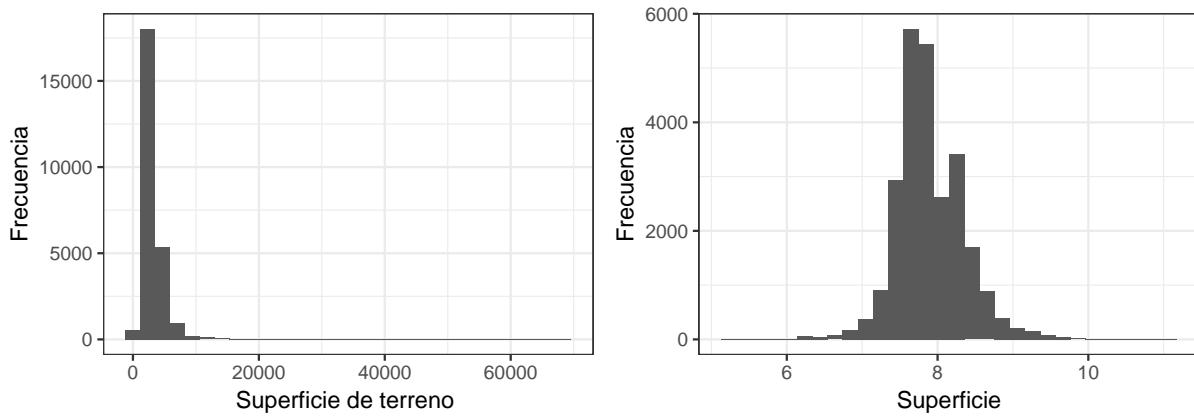


Figura 3: Histogramas de la superficie del terreno en escala original y en escala logarítmica

2.2. Gráficas bivariadas

Primero se analizará la posible relación entre el precio de venta y la superficie total mediante un diagrama de dispersión. En la figura 4 se puede ver la gráfica de dispersión de la superficie de construcción contra el precio. Existe una tendencia lineal creciente entre las dos variables, es decir, a mayor superficie total también se tiene un mayor precio de venta. Por lo que la variable de superficie total puede ser usada como variable explicativa en un modelo de regresión.

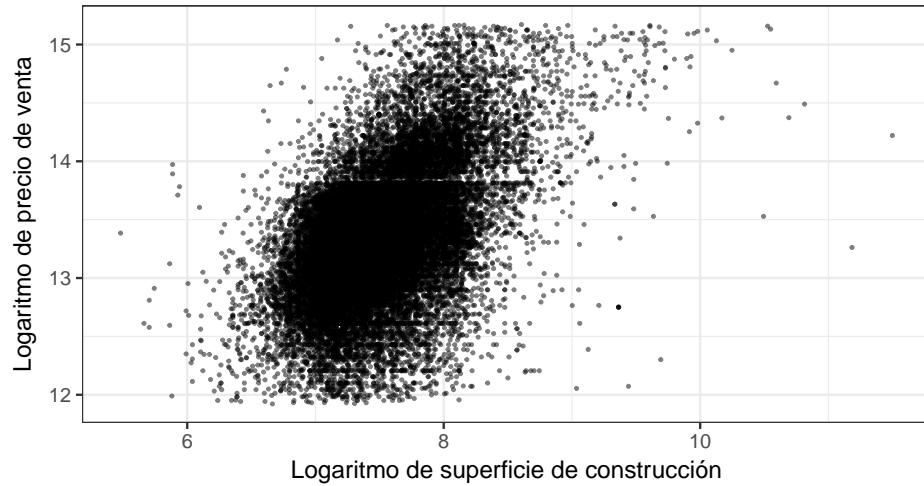


Figura 4: Gráfica de dispersión de superficie de construcción contra precio

La relación entre la superficie de terreno y el precio de venta se muestra en el diagrama de dispersión de la figura 5. Visualmente no existe una relación entre estas dos variables pues no muestra alguna tendencia. En la figura 6 se muestra la posible relación entre las covariables superficie de construcción y superficie de terreno. Como es de esperarse, estas dos variables están relacionadas pues se puede esbozar una relación creciente, es decir, a mayor superficie total se tiene mayor superficie. Dada esta colinealidad, en un modelo de regresión se debería de usar alguna de estas dos variables pues proporcionan la misma información. En el caso de este trabajo, se seleccionó la variable de superficie de construcción debido a su correlación con el precio.

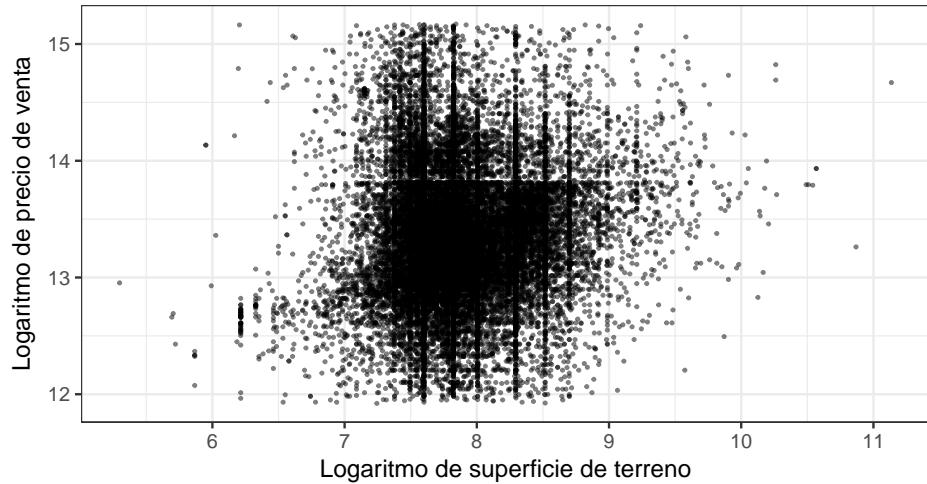


Figura 5: Gráfica de dispersión de superficie de terreno contra precio

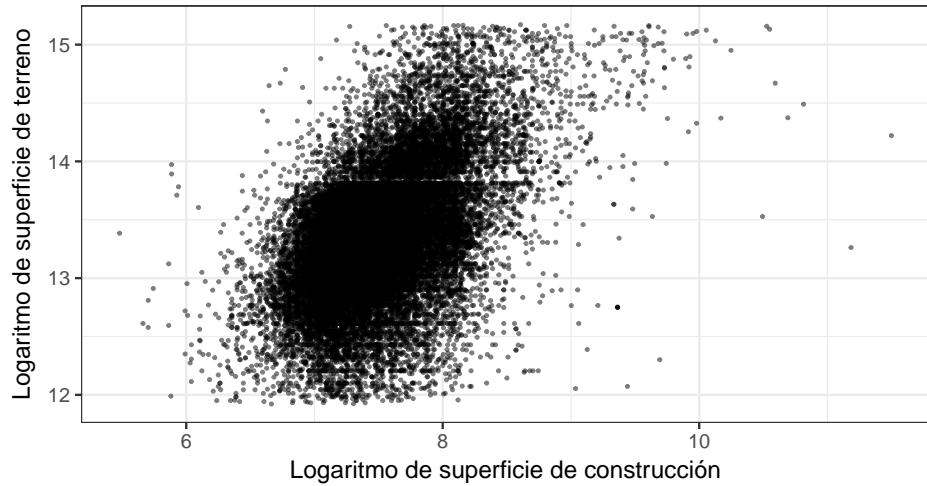


Figura 6: Gráfica de dispersión de superficie de construcción contra superficie del terreno

Es de esperar que el precio de venta cambie dependiendo si la casa esta ubicada en cierto distrito (*borough*). Para corroborar esta hipótesis se graficaron los precios en cada uno de los distritos, los cuales se pueden ver en las figuras 7 y 8. Se puede ver que, en efecto, cambian las distribuciones muestrales dependiendo el distrito en el que se encuentran las casas. También se puede ver que Manhattan muestra una media más alta que el resto de los distritos (línea punteada en figura 8), y también presenta más variación en los precios de venta. El siguiente distrito con una media más alta es Brooklyn con una variación más grande que el resto de los distritos (sin considerar Manhattan).

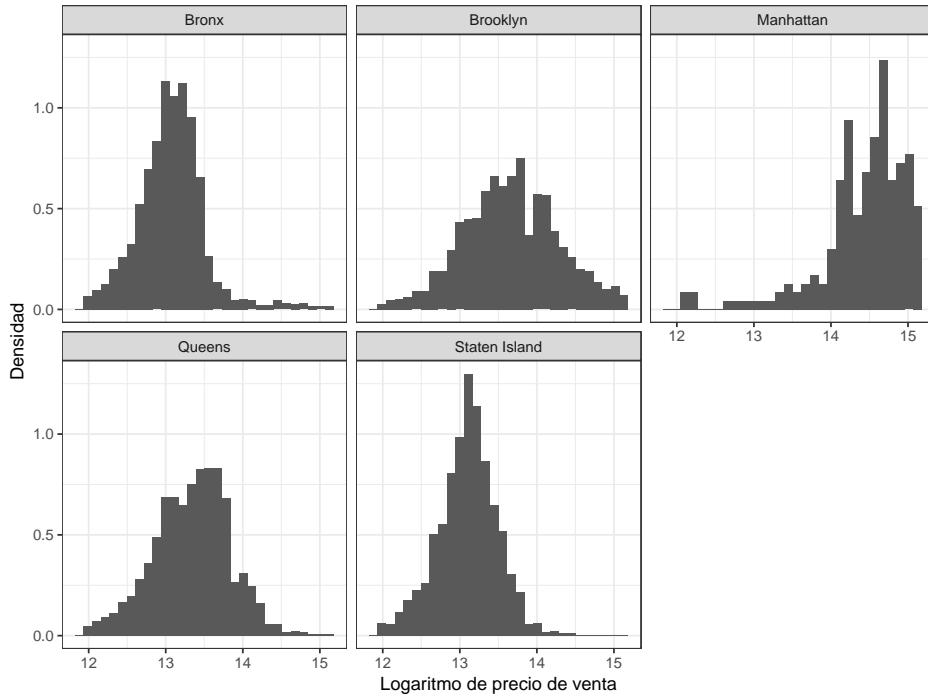


Figura 7: Histogramas de precio de venta por distrito

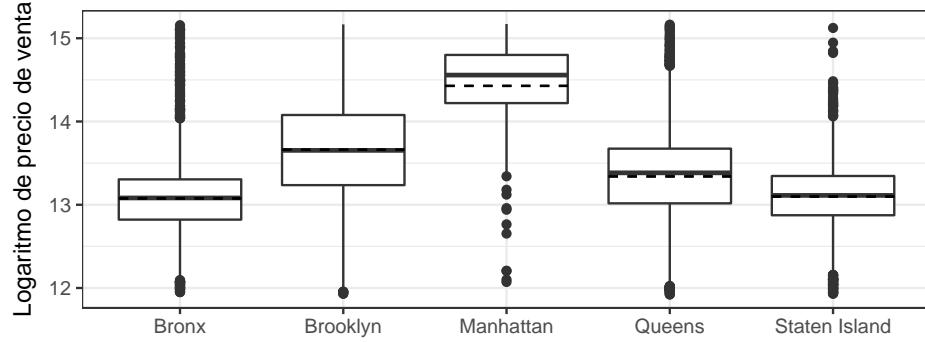


Figura 8: Diagrama de caja y brazos de precio de venta por distrito

Hasta este momento se ha omitido la variable vecindario en el análisis exploratorio. La figura 9 muestra la dispersión de los datos en cada vecindario. Se puede observar que considerando los vecindarios dentro de cada distrito, en algunos la relación creciente no es tan clara, o incluso llega a verse decreciente, como en el Upper West Side de Manhattan. Hay que tomar en cuenta que el número de observaciones en este vecindario es más pequeño.

El propósito de esta gráfica es mostrar que las relaciones cambian de vecindario a vecindario, por lo que hay que tomar esto en cuenta al momento de hacer los modelos. Aún cuando existe otro nivel geográfico (los códigos postales), no es factible graficar la relación a este nivel pues son demasiados para que se pueda ver en una hoja, sin embargo, se hizo un análisis y también se ve que las relaciones cambian.

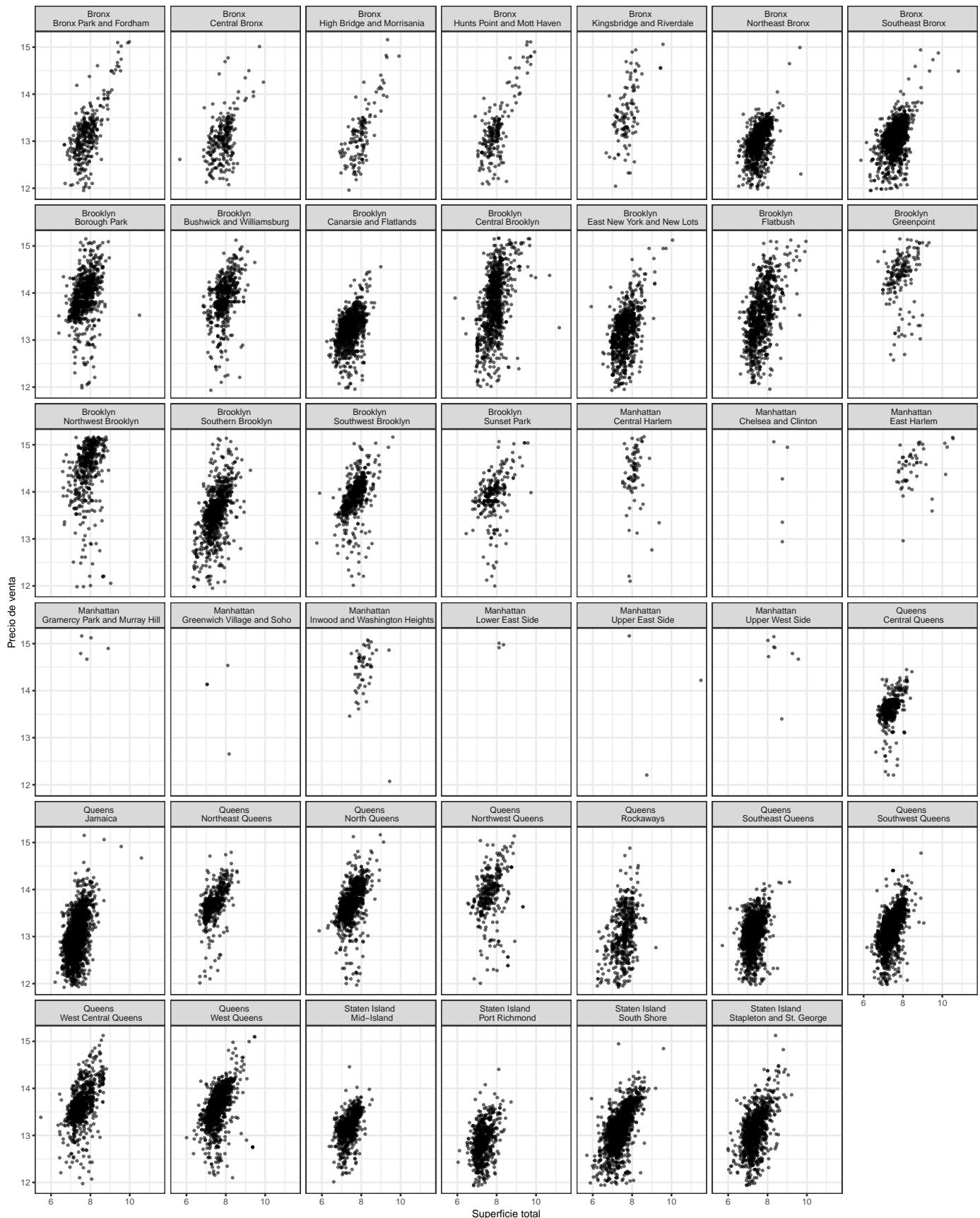


Figura 9: Diagramas de dispersión del logaritmo del precio contra el logaritmo del tamaño en cada vecindario

3. Modelos estadísticos

El objetivo es obtener estimaciones del precio de las casas a partir del tamaño en pies cuadrados. Para probar la capacidad predictiva, se dividieron los datos en dos: un conjunto de entrenamiento con el 90 % de los datos, y un conjunto de prueba con el 10 % restante. Para tener información de todos los códigos postales, se hizo un muestreo estratificado, tomando el 90 % de observaciones de cada código postal para el conjunto de entrenamiento. En el conjunto de entrenamiento quedaron 22,769 observaciones y en el de prueba 2,530.

Se ajustaron tres modelos lineales a los datos: un modelo de unidades iguales, un modelo de unidades independientes, y un modelo jerárquico. El modelo de unidades iguales asume que todas las realizaciones provienen de la misma distribución; mientras que el de unidades independientes asume que los precios varían de acuerdo a diferentes sectores (en este caso son los códigos postales); y finalmente, el modelo jerárquico es un compromiso entre ambos modelos que toma fuerza de los demás sectores, esto es particularmente útil cuando hay sectores con pocas observaciones.

3.1. Modelo de unidades iguales

El modelo de unidades iguales es simplemente un modelo de regresión lineal con un parámetro fijo para el intercepto y un parámetro fijo para cada uno de los regresores. Sean y_i el logaritmo del precio de la casa i y x_i el logaritmo del número de pies cuadrados en la casa i , para $i \in \{1, \dots, n\}$, con $n = 22,769$. El modelo de unidades independientes es $y_i \sim N(\alpha + \beta x_i, \tau_y)$, con distribuciones previas $\alpha \sim N(0, 0.001)$, $\beta \sim N(0, 0.001)$ y $\tau_y \sim Ga(0.001, 0.001)$.

De antemano se tiene conocimiento como para pensar que este modelo no es el más adecuado para los datos, pues se vio en el análisis exploratorio de datos que los precios varían por código postal, por lo que no es muy sensato suponer que no existen relaciones entre las observaciones. De hecho, en la figura 10 se puede ver este efecto. Más adelante se ahonda en este resultado.

3.2. Modelo de unidades independientes

Este es un modelo de interceptos y pendientes cambiantes de acuerdo al código postal, es decir, es de la forma

$$y_i \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \tau_y),$$

donde $j[i]$ se refiere al código postal correspondiente a la i -ésima observación. Las distribuciones previas son

- $\alpha[j] \sim N(0, 0.001)$
- $\beta[j] \sim N(0, 0.001)$
- $\tau_y \sim Ga(0.001, 0.001)$

para $j = 1, \dots, J$, con $J = 154$ el número de códigos postales en la ciudad.

3.3. Modelo multinivel

En el modelo multinivel o jerárquico, también ajustamos distintos interceptos y pendientes de acuerdo a cada código postal, pero en lugar de considerar cada código postal como una unidad independiente, se le agregaron dos niveles más de hiperparámetros, correspondientes a los vecindarios de la ciudad, y a cada distrito (*borough*). Además, en este modelo no se asume varianza constante en las observaciones, sino que cambian de acuerdo al código postal.

El modelo ajustado fue de la forma $y_i \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \tau_{j[i]})$, donde nuevamente $j[i]$ se refiere al código postal correspondiente a la i -ésima observación. Las distribuciones previas son $\alpha_j \sim N(\mu_{\alpha,k[j]}, \tau_{\alpha,k[j]})$, $\beta_j \sim N(\mu_{\beta,k[j]}, \tau_{\beta,k[j]})$ y $\tau_{j[i]} \sim Ga(\alpha_{y,l[k]}, \beta_{y,l[k]})$, donde $k[j]$ se refiere al vecindario correspondiente al j -ésimo código postal, y $l[k]$ se refiere al distrito correspondiente al k -ésimo vecindario. Las distribuciones previas de estos hiperparámetros son

- $\mu_{\alpha,k[j]} \sim N(\mu_{\alpha,l[k]}, \tau_{\alpha,l[k]})$
- $\mu_{\beta,k[j]} \sim N(\mu_{\beta,l[k]}, \tau_{\beta,l[k]})$
- $\alpha_{y,l[k]} \sim \exp(\lambda_{\alpha_y,l[k]})$
- $\tau_{\alpha,k[j]} \sim \exp(\lambda_{\alpha,l[k]})$
- $\tau_{\beta,k[j]} \sim \exp(\lambda_{\beta,l[k]})$
- $\beta_{y,l[k]} \sim \exp(\lambda_{\beta_y,l[k]})$

Y sus correspondientes hiperparámetros se distribuyen:

- $\mu_{\alpha,l[k]} \sim N(\mu_{\alpha_0})$
- $\mu_{\beta,l[k]} \sim N(\mu_{\beta_0})$
- $\lambda_{\alpha_y,l[k]} \sim \exp(\lambda_{\alpha_y})$
- $\lambda_{\beta_y,l[k]} \sim \exp(\lambda_{\beta_y})$
- $\tau_{\alpha,l[k]} \sim \exp(\lambda_{\tau_{\alpha_0}})$
- $\tau_{\beta,l[k]} \sim \exp(\lambda_{\tau_{\beta_0}})$
- $\lambda_{\alpha_0} \sim \exp(\lambda_{\lambda_{\alpha_0}})$
- $\lambda_{\beta_0} \sim \exp(\lambda_{\lambda_{\beta_0}})$

Con sus correspondientes distribuciones previas:

- $\mu_{\alpha_0} \sim N(0, 0.0001)$
- $\mu_{\beta_0} \sim N(0, 0.0001)$
- $\lambda_{\lambda_{\alpha_y}} \sim \exp(0.01)$
- $\lambda_{\lambda_{\beta_y}} \sim \exp(0.01)$
- $\lambda_{\tau_{\alpha_0}} \sim \exp(0.01)$
- $\lambda_{\tau_{\beta_0}} \sim \exp(0.01)$
- $\lambda_{\lambda_{\alpha_0}} \sim \exp(0.01)$
- $\lambda_{\lambda_{\beta_0}} \sim \exp(0.01)$

En la siguiente sección se muestran los resultados de los tres modelos presentados aquí.

4. Resultados

En las figuras 10, 11 y 12 se muestran para cada uno de los modelos, los residuales en el eje y y en el eje x se muestra el índice de la observación, donde las observaciones están ordenadas de acuerdo a código postal. En el modelo de unidades iguales es evidente un patrón, que viene de la correlación entre las observaciones que existe dentro de cada código postal. En el modelo de unidades independientes los patrones del código postal ya no son tan evidentes como en el modelo de unidades iguales. Lo mismo pasa con el modelo multinivel, ya no hay un patrón evidente.

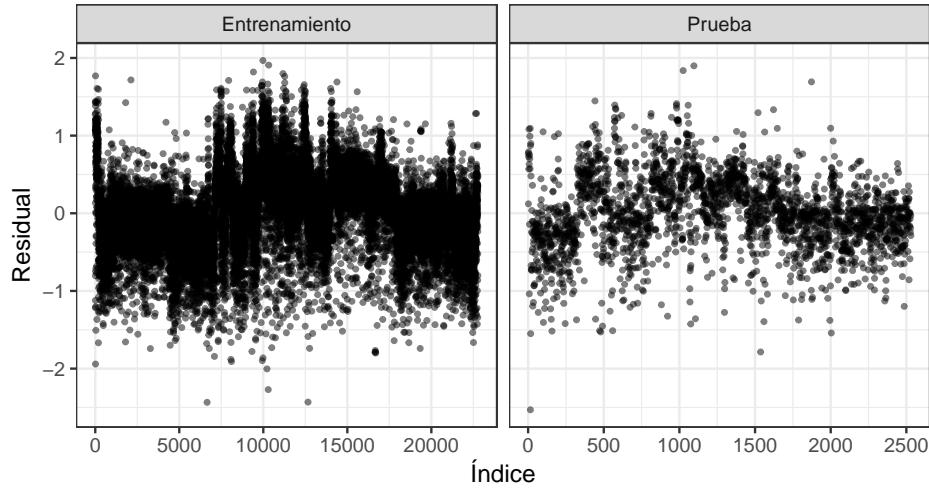


Figura 10: Residuales de modelo de unidades iguales

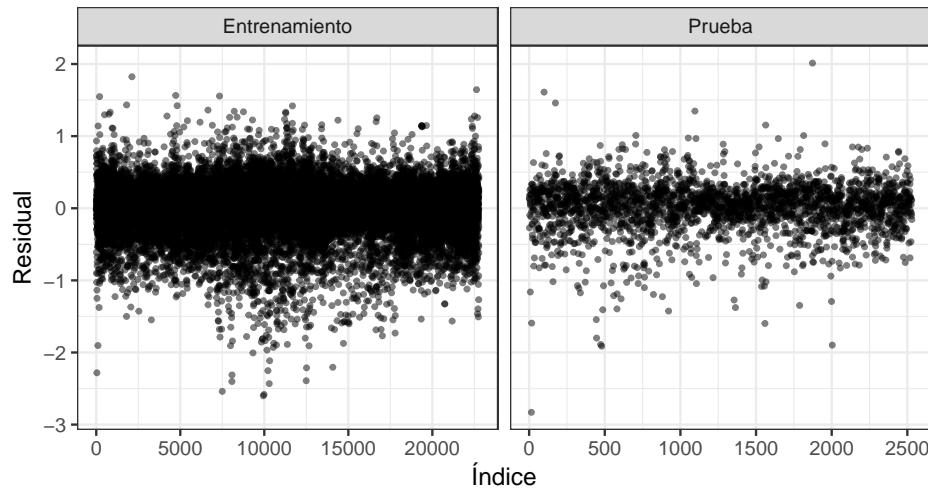


Figura 11: Residuales de modelo de unidades independientes

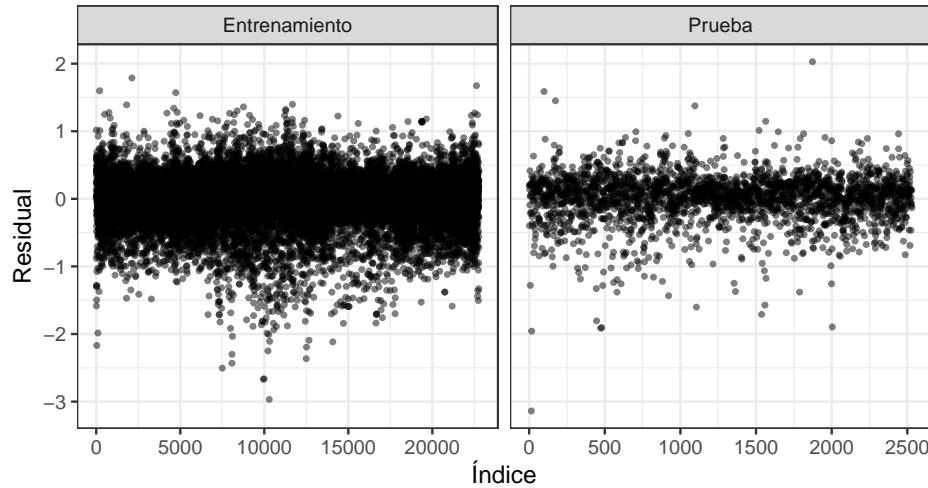


Figura 12: Residuales de modelo multinivel

En las figuras 13, 14 y 15 se puede ver para cada observación el valor observado contra el valor ajustado de cada modelo. En todos se aprecia una varianza considerablemente grande; y en el modelo de unidades iguales, el modelo tiende a sobreestimar los valores pequeños, mientras que en valores grandes pasa lo contrario. Este efecto persiste, pero en mucho menor medida, en los otros dos modelos.

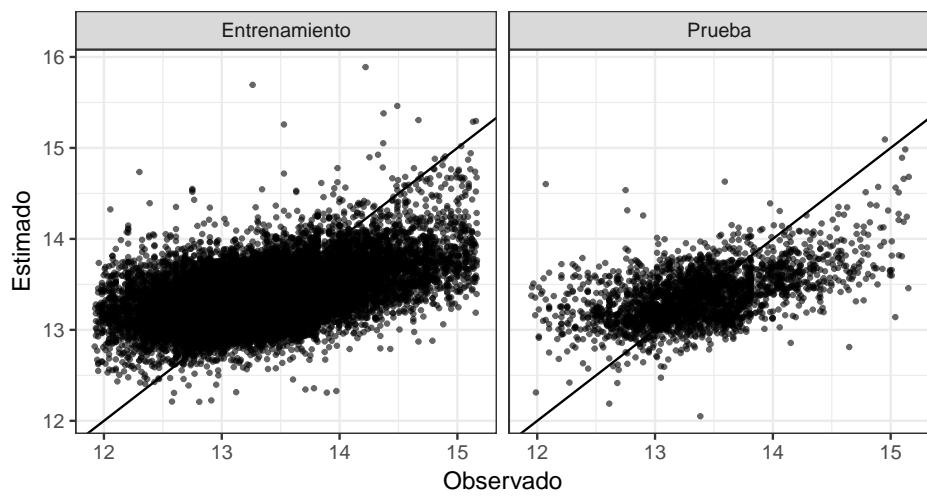


Figura 13: Ajustado contra observado en modelo de unidades iguales

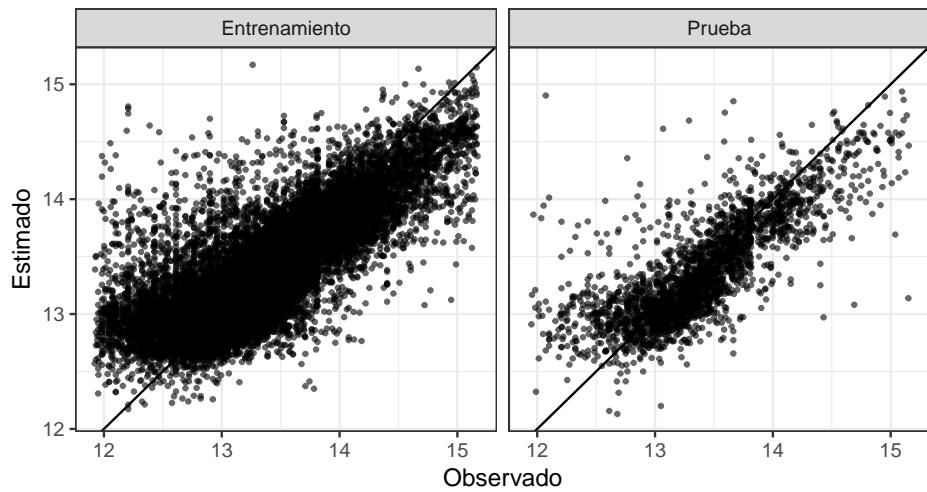


Figura 14: Valor ajustado contra observado en modelo de unidades independientes

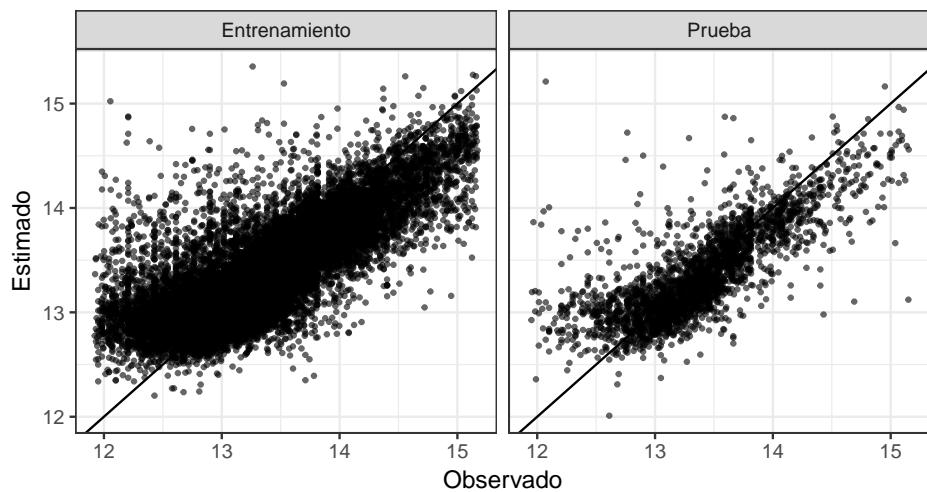


Figura 15: Valor ajustado contra observado en modelo multinivel

En la figura 16 también se muestran los valores observados contra los ajustados con el modelo multinivel y de unidades independientes en el conjunto de prueba. La diferencia en esta gráfica es que se muestran los intervalos al 95 % de predicción de cada observación. Se puede ver por el tamaño de estos intervalos que se está tomando en cuenta la incertidumbre que hay en la predicción. Esta incertidumbre viene de la varianza original de los datos.

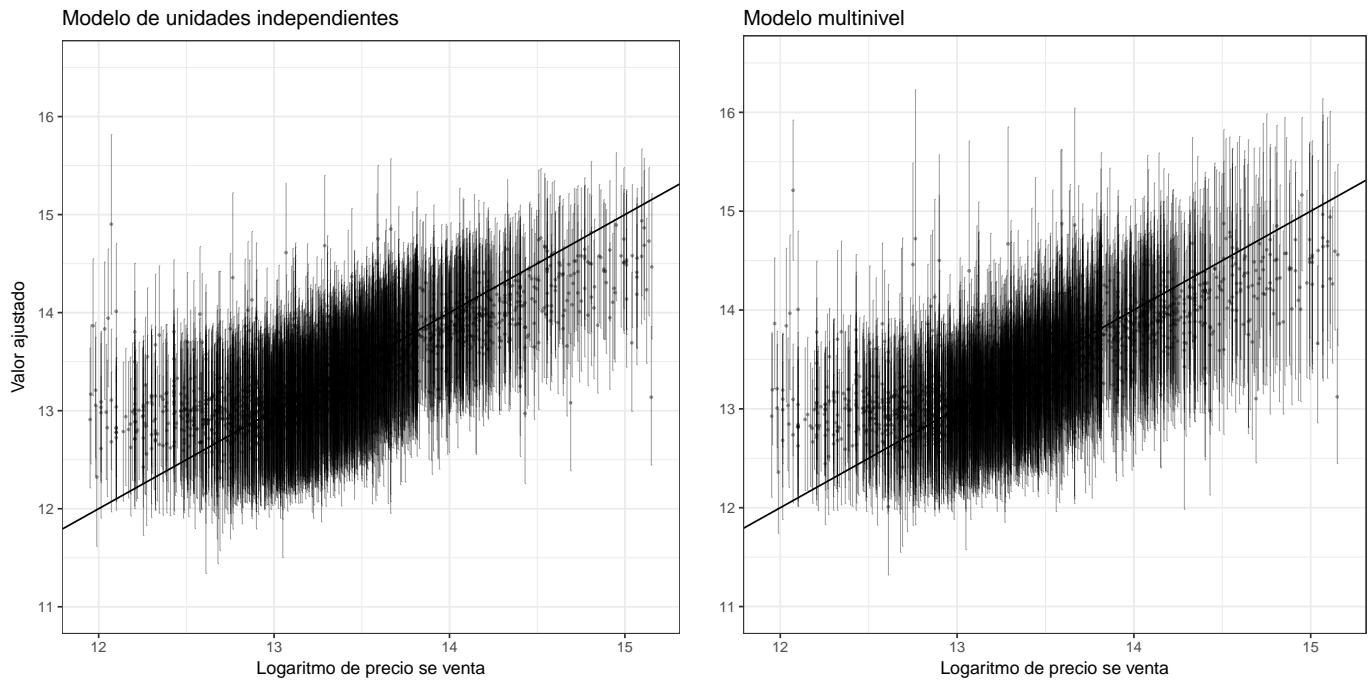


Figura 16: Valor ajustado contra observado del conjunto de prueba en modelo de unidades independientes y modelo multinivel

En las figuras 17 y 18 se muestran los valores de los parámetros junto con sus intervalos al 95 % de probabilidad del modelo de unidades independientes y del modelo multinivel. En el modelo de unidades independientes hay varios parámetros que tienen una varianza muy grande, y hay incluso algunos códigos postales que tienen un parámetro de pendiente negativo, lo cual intuitivamente no hace mucho sentido, pues eso significaría que a menor tamaño, la casa es más grande; pero recordando el análisis exploratorio, estas estimaciones negativas corresponden a los códigos postales con pocas observaciones en las cuales se podía apreciar una tendencia negativa; pero esto no es nada más que ruido de la muestra pequeña que se tiene. En el modelo multinivel, al tomar fuerzas de los hiperparámetros, no se tienen estimaciones puntuales negativas, y los intervalos de probabilidad son mucho más pequeños; sin embargo, sí se aprecia cambio entre los parámetros; es decir, el modelo está captando las diferencias de precio que existen entre los códigos postales.

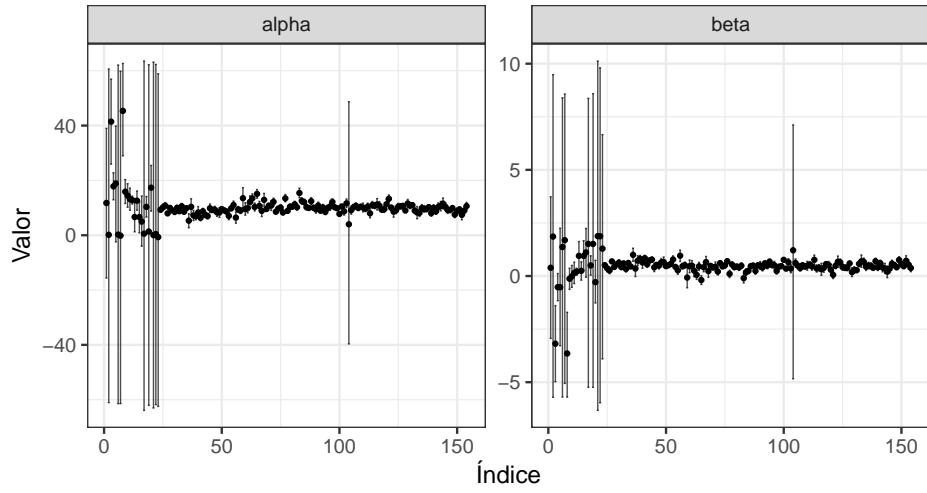


Figura 17: Valor e intervalos de probabilidad de parámetros de modelo de unidades independientes

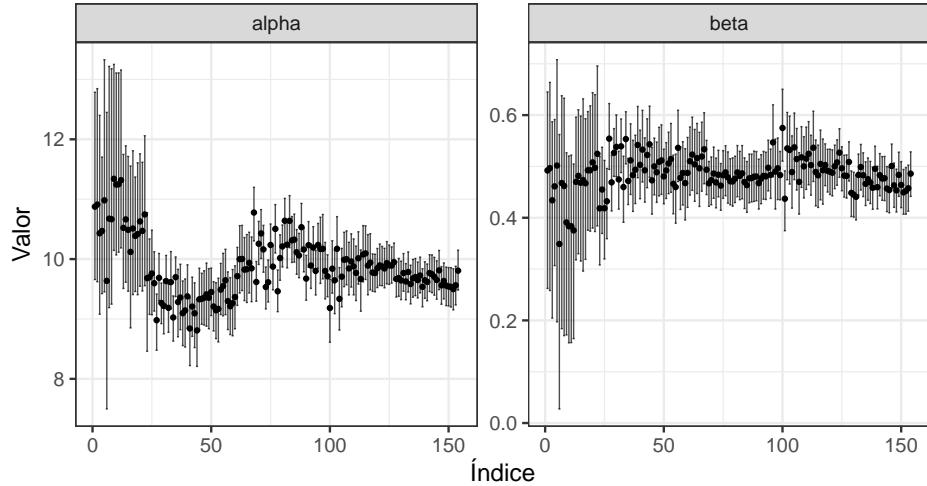


Figura 18: Valor e intervalos de probabilidad de parámetros de modelo multinivel

Para observar el efecto de las diferencias en los parámetros, en las figuras 19 y 20 se muestran las pendientes de regresión de código postal de cada modelo. Cada figura tiene muchas subgráficas, cada una representando un vecindario. Dentro de cada subgráfica se muestran los precios de ventas y además las líneas de regresión ajustadas a cada código postal dentro de cada vecindario. Se puede ver que en el modelo de unidades independientes cambian mucho las líneas, sobre todo en los vecindarios en los que hay pocos datos, llegando a haber líneas con pendientes negativas, lo cual no tiene mucho sentido dado el contexto del problema. El modelo nivel muestra pendientes más estables, y sobre todo, en los vecindarios y códigos postales, se mantiene siempre la pendiente positiva.

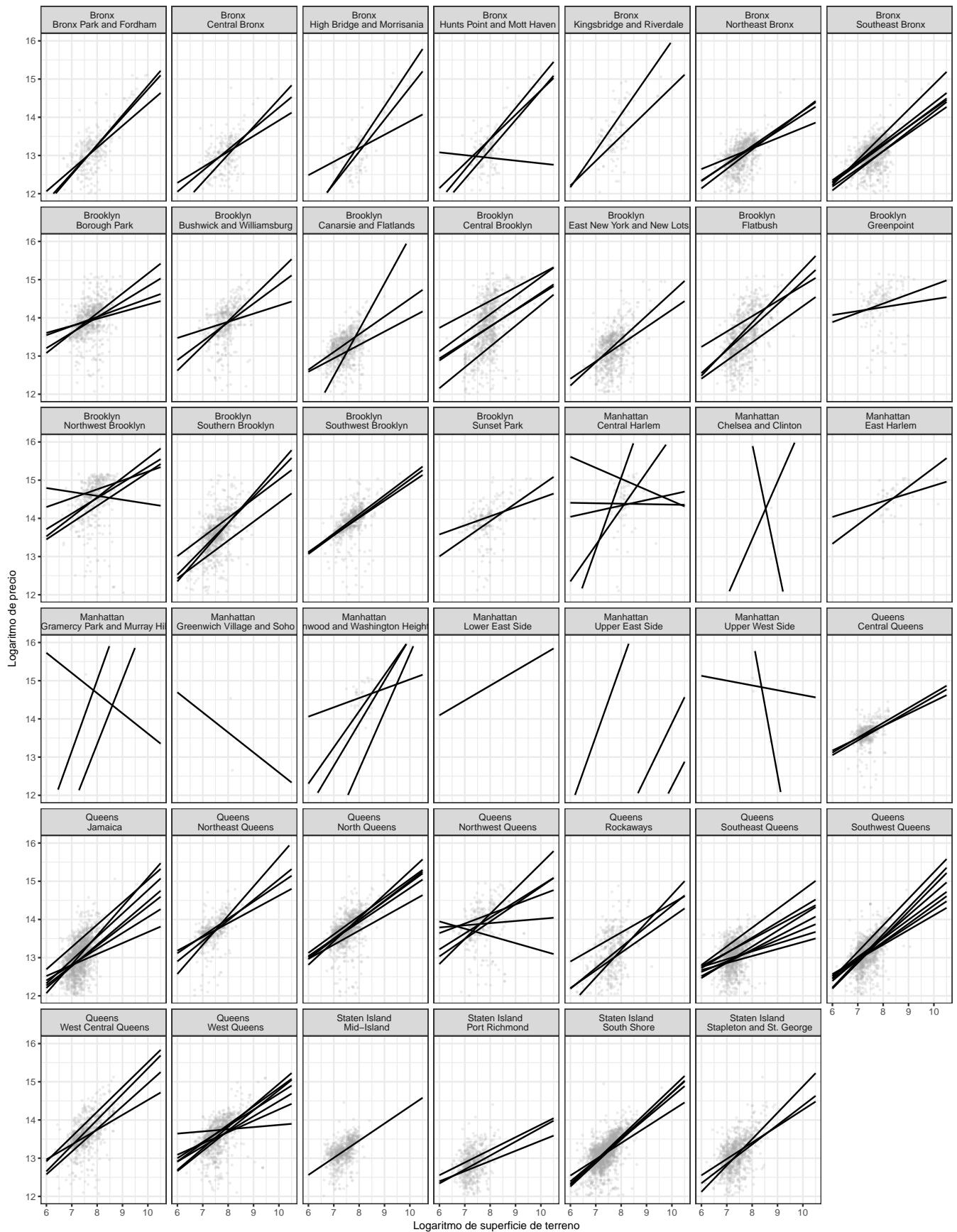


Figura 19: Modelo de unidades independientes: Diagramas de dispersión de logaritmo de superficie de terreno contra logaritmo del precio, separados por vecindario. Dentro de cada gráfica de vecindario, se muestran las líneas de regresión de cada código postal.

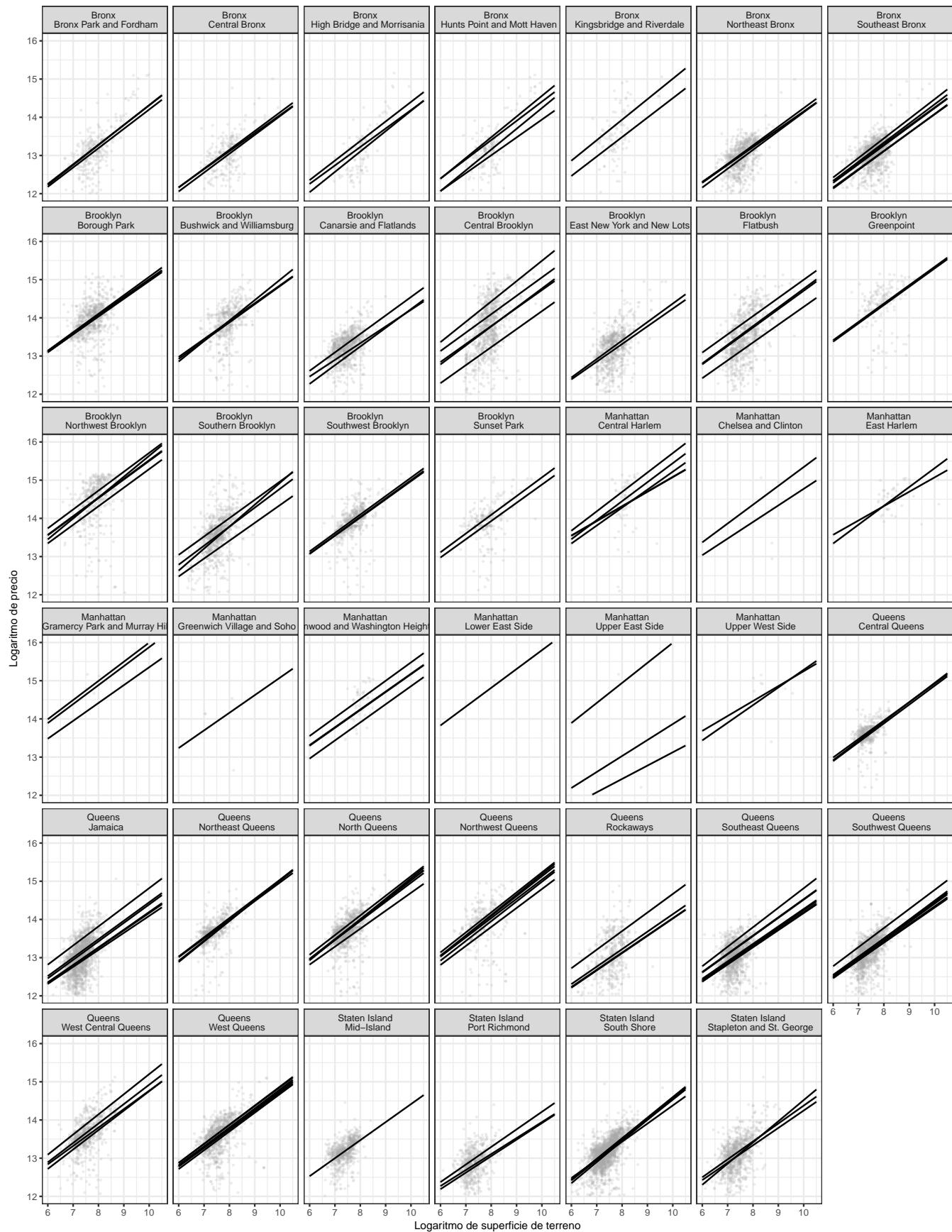


Figura 20: Modelo multinivel: Diagramas de dispersión de logaritmo de superficie de terreno contra logaritmo del precio, separados por vecindario. Dentro de cada gráfica de vecindario, se muestran las líneas de regresión de cada código postal.

Tabla 1: Valores de DIC, RMSE y MAE de cada modelo.

		Entrenamiento		Prueba	
	DIC	MAE	RMSE	MAE	RMSE
Unidades iguales	630,996	287,374	461,166	283,847	446,443
Unidades independientes	7,239,403	188,123	315,006	190,248	324,864
Multinivel	16,716	191,383	324,493	192,564	335,204

Para medir el desempeño de los distintos modelos, se pueden usar distintas medidas. En este trabajo, se utilizan tres: el DIC (*Deviance Information Criterion*), la raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE). Dada su definición, para las tres medidas, es preferible tener un menor valor. Sean y_i los valores observados y \hat{y}_i las estimaciones puntuales para cada $i \in \{1, \dots, n\}$, donde n es el número total de observaciones. Las cantidades antes mencionadas están definidas como

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$\text{DIC} = -2 \log p(y|\theta) + 2 \log h(y),$$

donde $p(y|\theta)$ es la función de verosimilitud y $h(y)$ es una función de estandarización de los datos.

En la tabla 1 se muestran los valores de estas medidas para cada uno de los modelos. El modelo de unidades iguales es claramente inferior, pues los valores son mucho mayores. Entre el modelo de unidades independientes y el multinivel hay competencia, pues el primero tiene mayor DIC pero menores MAE y RMSE. Sin embargo, las diferencias en MAE y RMSE no son muy grandes; y además, viendo los resultados anteriores, el modelo multinivel es más robusto que el de unidades independientes.

Finalmente, en la figura 21 se muestra un mapa de los códigos postales de la ciudad de Nueva York. El color de cada polígono corresponde a la estimación puntual del precio de un departamento de 1 pie cuadrado de acuerdo al modelo multinivel. Esta estimación no es nada más que e^{α_j} , pues el modelo está planteado de la forma $\log(\text{precio}_i) = \alpha_{j[i]} + \beta_{j[i]} \log(\text{tam}_i)$, por lo que si el tamaño es de 1 pie cuadrado, como $\log(1) = 0$, entonces $\log(\text{precio}_i) = \alpha_{j[i]}$, por lo que $\text{precio}_i = \exp(\alpha_{j[i]})$.

En el mapa de arriba de la figura 21 hay algunos espacios blancos, los cuales corresponden a códigos postales sin ventas. Se puede hacer una predicción con base en las distribuciones de las jerarquías que le corresponden, y al hacer esto, se tiene como resultado el mapa de abajo.

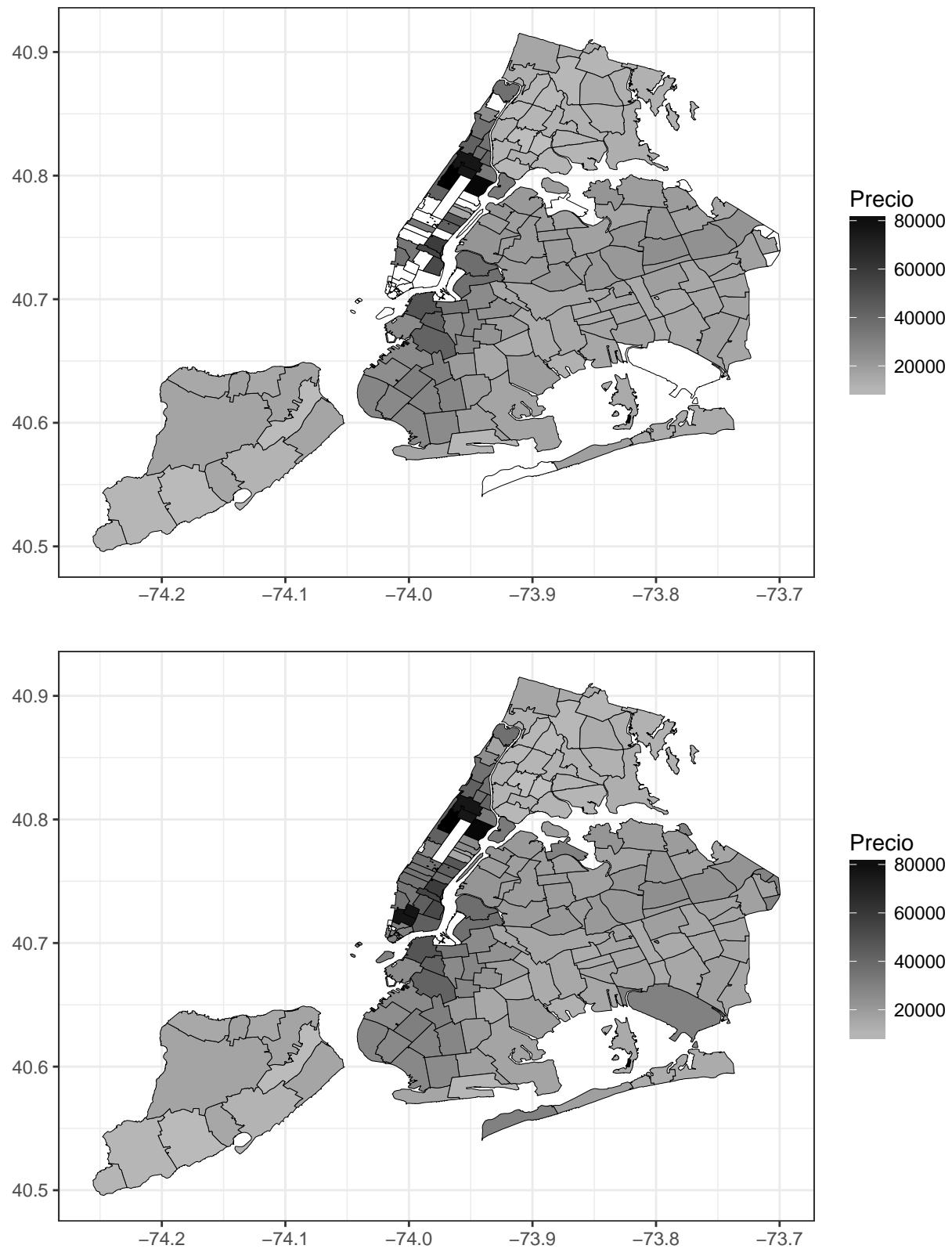


Figura 21: Mapa de códigos postales de la ciudad de Nueva York con su estimación puntual de precio por pie cuadrado con el modelo multinivel

5. Conclusiones

Los tres modelos presentados anteriormente son soluciones a un mismo problema: predecir el precio de una casa en venta. El modelo de unidades iguales no considera la variabilidad de los precios dependiendo en la zona geográfica, lo que lo hace un modelo poco creíble desde su definición. El modelo de unidades independientes incorpora esta variabilidad pero no la capta en su totalidad, y aunque en algunas métricas de evaluación de ajuste salga mejor que los demás modelos, algunos de los resultados arrojados por el modelo no son lógicos. Finalmente, el modelo multinivel también incorpora la variabilidad de los precios pero a distintos niveles geográficos por lo es un modelo más robusto. Otra característica importante es que a diferencia de los otros modelos, el modelo multinivel utiliza toda la información disponible. Por lo que se concluye que el modelo multinivel es el mejor modelo para responder al problema planteado.

Es importante mencionar que aún cuando un modelo presenta mejores métricas de ajuste, no necesariamente es el que tiene mejor interpretabilidad, por lo que en la práctica estadística es recomendable comparar los resultados de un modelo contra la realidad. También creemos que las métricas de ajuste pueden mejorar con más covariables que describan las casas, por ejemplo, número de cuartos, número de baños, estado de la casa, etc. Sin embargo, en este ejercicio estas variables no estaban disponibles.

Asimismo, en este trabajo también se puede observar otra de las aplicaciones de los modelos de regresión: estimación de datos faltantes. Pues mediante el modelo multinivel se realizó una estimación de cual sería el precio de venta de un pie cuadrado en códigos postales que no registrarón ventas.

Referencias

- [1] Stephen P Brooks y Andrew Gelman. «General methods for monitoring convergence of iterative simulations». En: *Journal of computational and graphical statistics* 7.4 (1998), págs. 434-455.
- [2] Andrew Gelman y Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Vol. 1. Cambridge University Press New York, NY, USA, 2007.
- [3] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari y Donald B Rubin. *Bayesian data analysis*. Vol. 2. CRC press Boca Raton, FL, 2014.
- [4] Ross Ihaka y Robert Gentleman. «R: a language for data analysis and graphics». En: *Journal of computational and graphical statistics* 5.3 (1996), págs. 299-314.
- [5] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.
- [6] Martyn Plummer. «JAGS Version 3.3. 0 user manual». En: *International Agency for Research on Cancer, Lyon, France* (2012).
- [7] YS Su y M Yajima. «R2jags: Using R to run ‘JAGS’». R package version 0.5–7». En: Available: CRAN. *R-project.org/package=R2jags*. (September 2015) (2015).

Apéndice: convergencia de cadenas

En esta sección, se ilustra la convergencia del proceso iterativo de MCMC de los modelos. Todos los modelos fueron corridos en JAGS versión 4.2.0⁴ llamándolo desde R [4] a través del paquete **R2jags** [7]. Cada modelo fue corrido con cuatro cadenas con puntos iniciales aleatorios. Para los modelos de unidades iguales y de unidades independientes, el tamaño de muestra era de 4000, mientras que para el modelo multinivel, se utilizó un tamaño de muestra de 8000.

Para cada modelo, se muestra primero una gráfica de cada parámetro y el valor del estadístico de convergencia de Gelman y Rubin \hat{R} [3] [5]. La idea de este estadístico es inicializar varias cadenas en distintos puntos, y después de cierto tiempo, si las cadenas convergieron, deben de tener un valor de \hat{R} cercano a 1. Brooks y Gelman recomiendan como regla de dedo un valor cercano a 1.2 [1]. Después se muestra una gráfica de este mismo estadístico, pero para las distribuciones predictivas del conjunto de entrenamiento y del conjunto de prueba. La mayoría de las cadenas tuvieron un valor menor a 1.2, y los pocos que no, estaban solamente un poco por arriba de ese valor.

También se muestran gráficas del tamaño de muestra efectivo de cada parámetro y de las distribuciones predictivas; con una línea punteada horizontal marcando el tamaño de muestra real. El tamaño de muestra efectivo tiene que ver con la autocorrelación de las cadenas, y es importante cuando se hacen inferencias, sobre todo inferencias de intervalos de probabilidad muy altos. Por ejemplo, si se desea un intervalo de probabilidad de la distribución predictiva al 99 %, entonces se requiere un tamaño de muestra efectivo grande. En general, el tamaño de muestra efectivo no fue chico, aunque hubo varios parámetros que estuvieron lejos del tamaño de muestra observado. Sin embargo, como en este trabajo no se hicieron inferencias tan precisas, no se requería un valor muy grande.

Modelo de unidades iguales

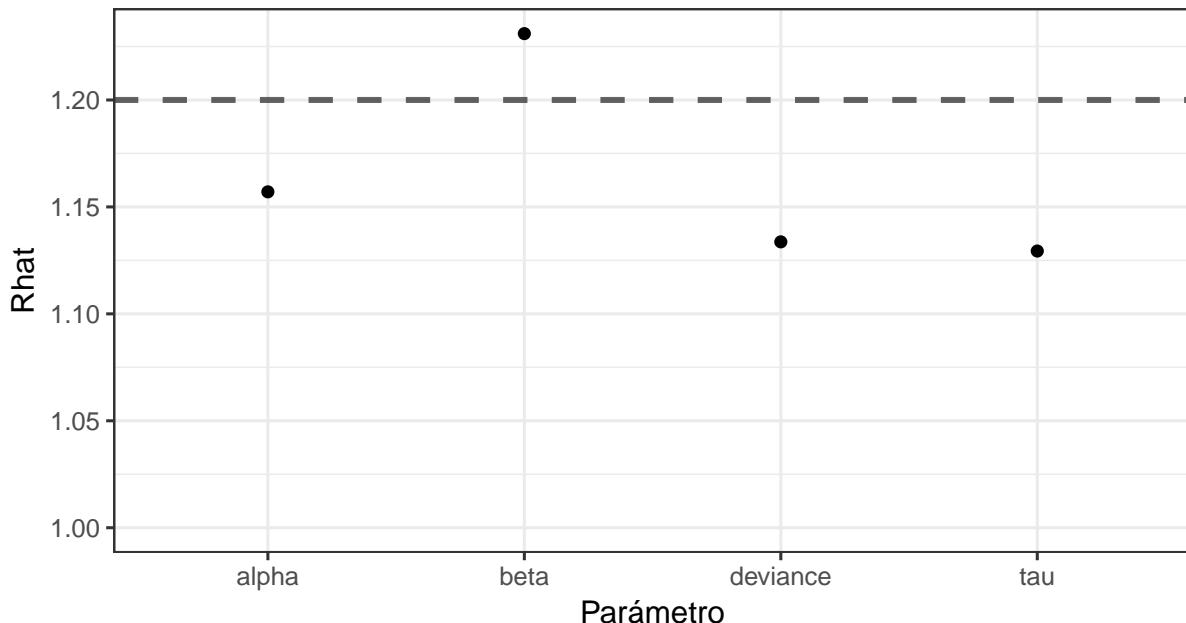


Figura 22: Estadística de convergencia \hat{R} de Gelman y Rubin para cada parámetro del modelo de unidades iguales

⁴<http://mcmc-jags.sourceforge.net/>

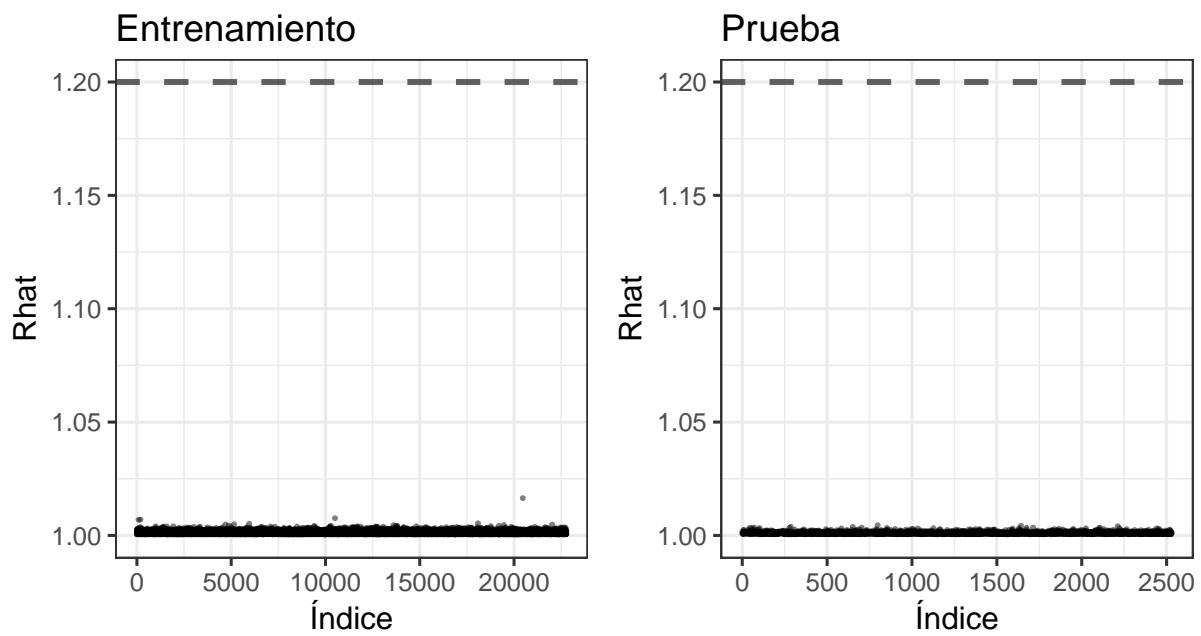


Figura 23: Estadística de convergencia \hat{R} de Gelman y Rubin para estimaciones de la variable respuesta del modelo de unidades iguales

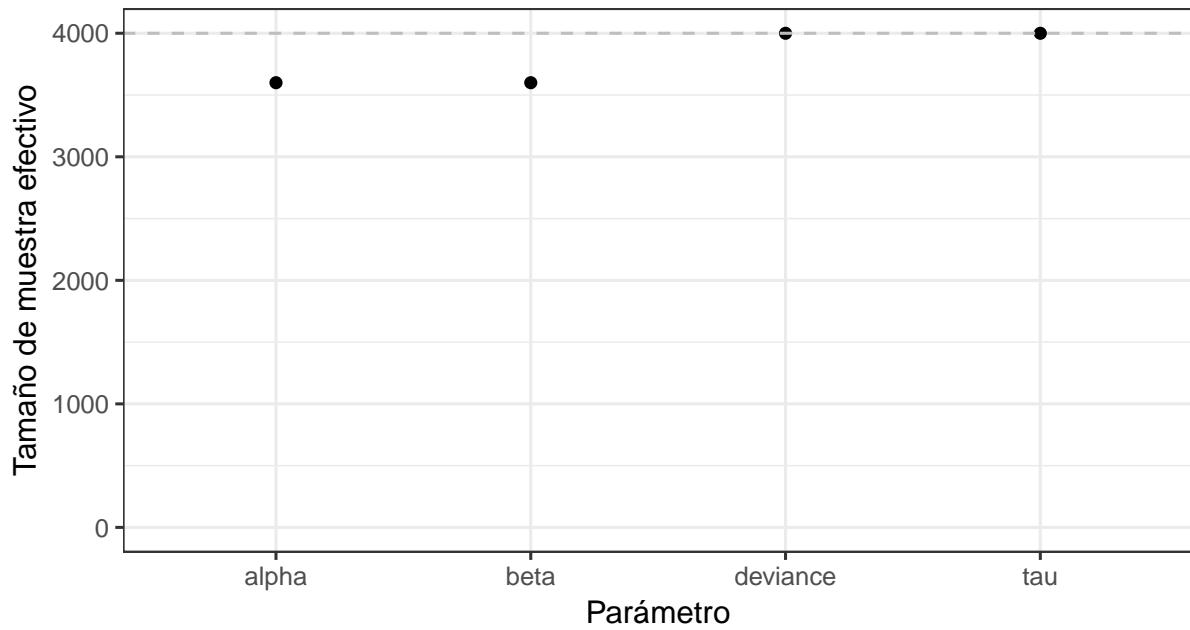


Figura 24: Tamaño de muestra efectivo para cada parámetro del modelo de unidades iguales

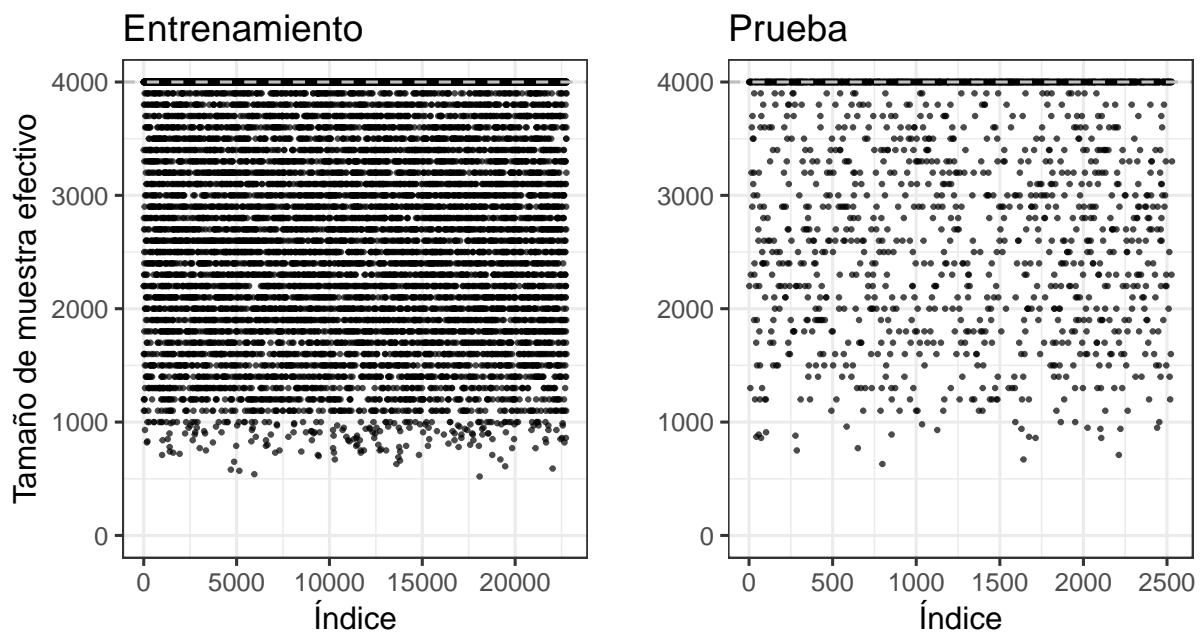


Figura 25: Tamaño de muestra efectivo para cada estimación de la variable respuesta del modelo de unidades iguales

Modelo de unidades independientes

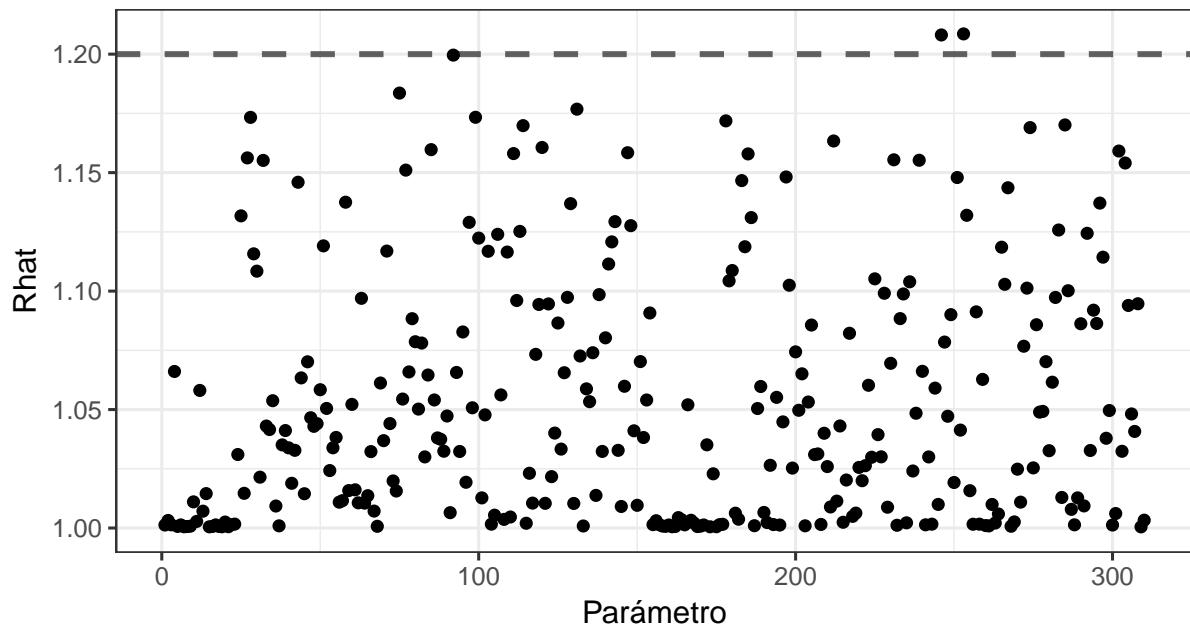


Figura 26: Estadística de convergencia \hat{R} de Gelman y Rubin para cada parámetro del modelo de unidades independientes

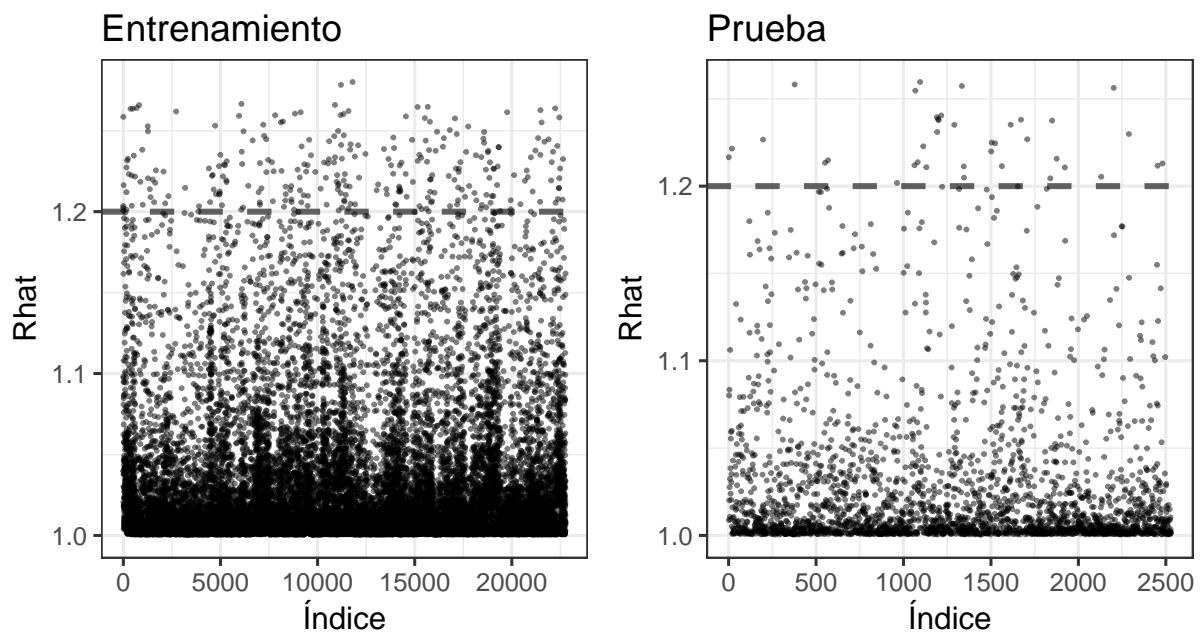


Figura 27: Estadística de convergencia \hat{R} de Gelman y Rubin para estimaciones de la variable respuesta del modelo de unidades independientes

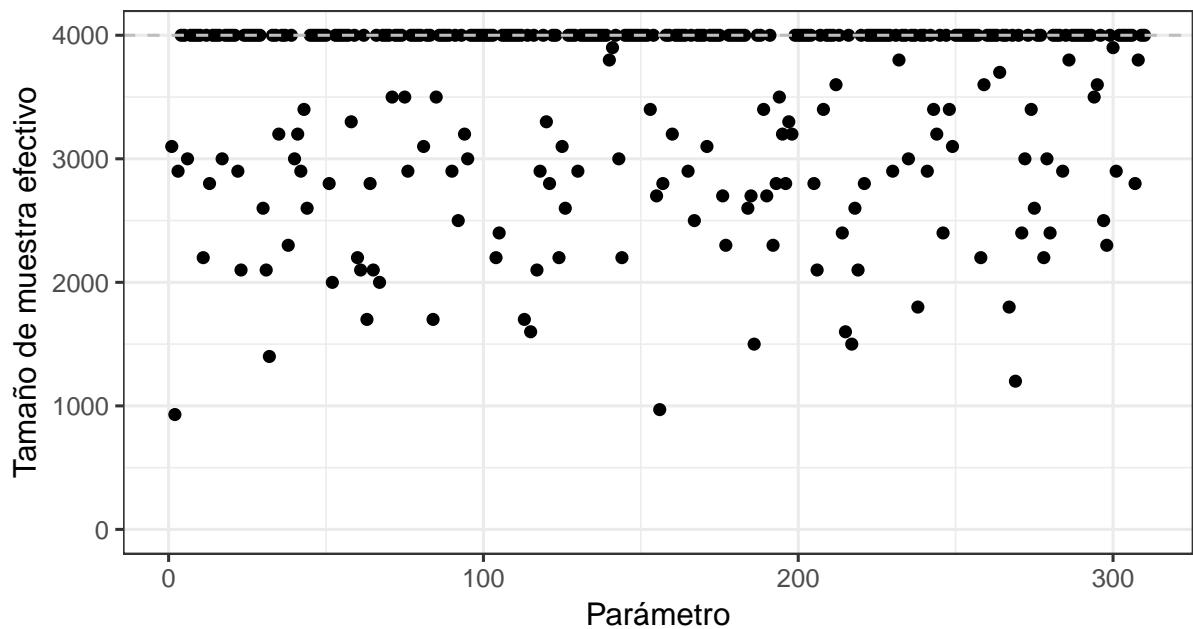


Figura 28: Tamaño de muestra efectivo para cada parámetro del modelo de unidades independientes

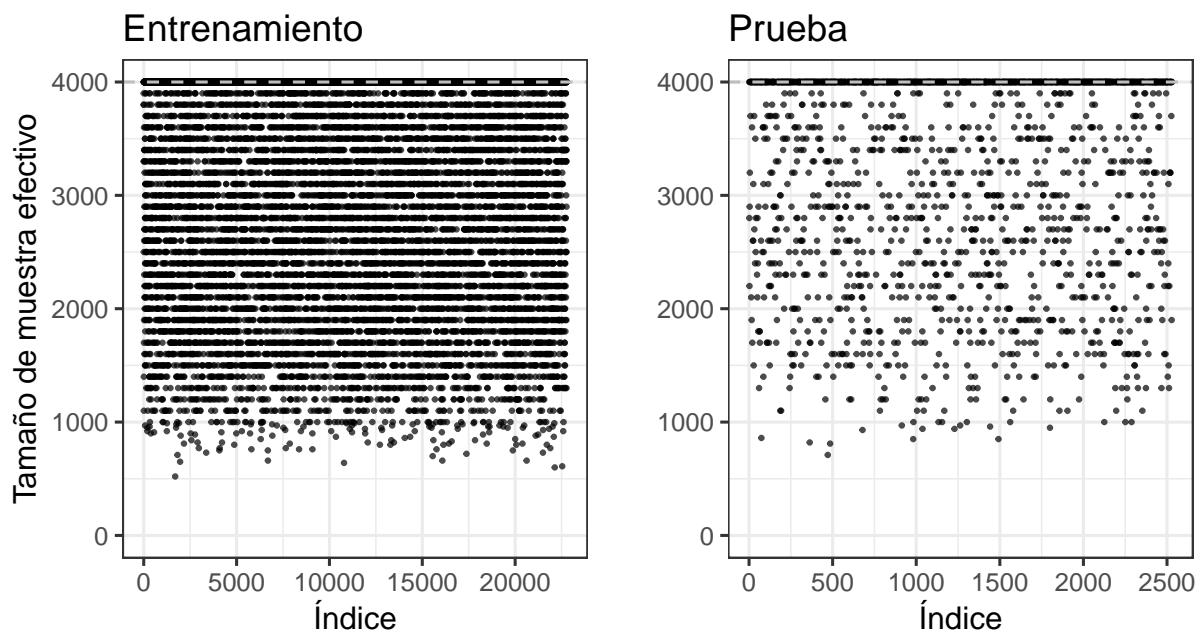


Figura 29: Tamaño de muestra efectivo para cada estimación de la variable respuesta del modelo de unidades independientes

Modelo multinivel

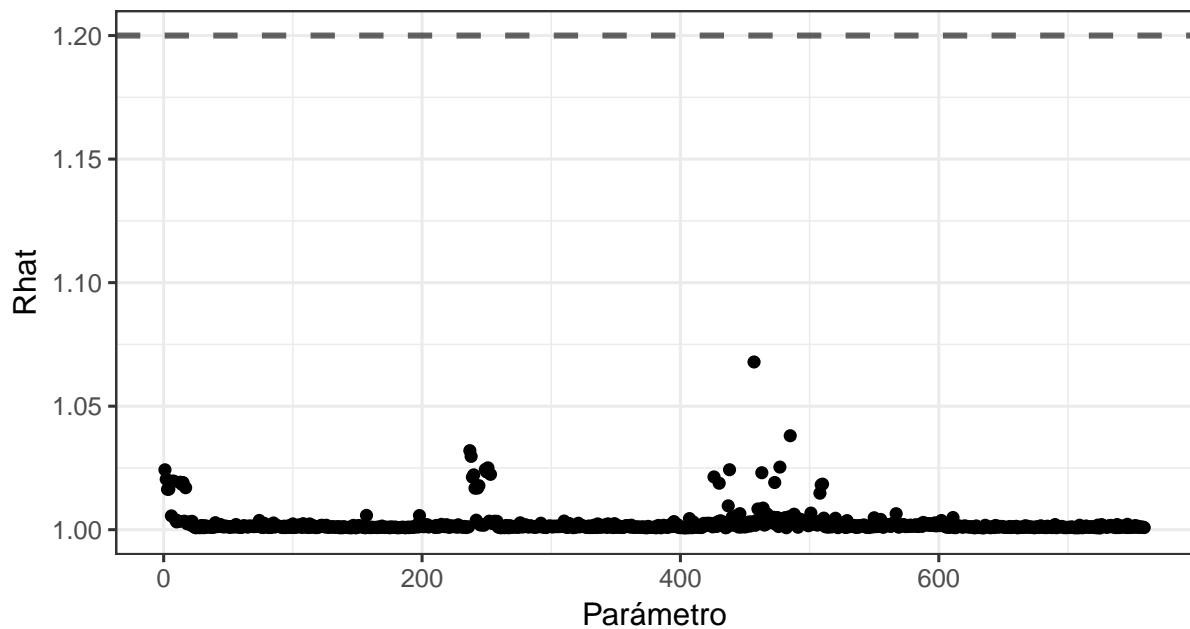


Figura 30: Estadística de convergencia \hat{R} de Gelman y Rubin para cada parámetro del modelo multinivel

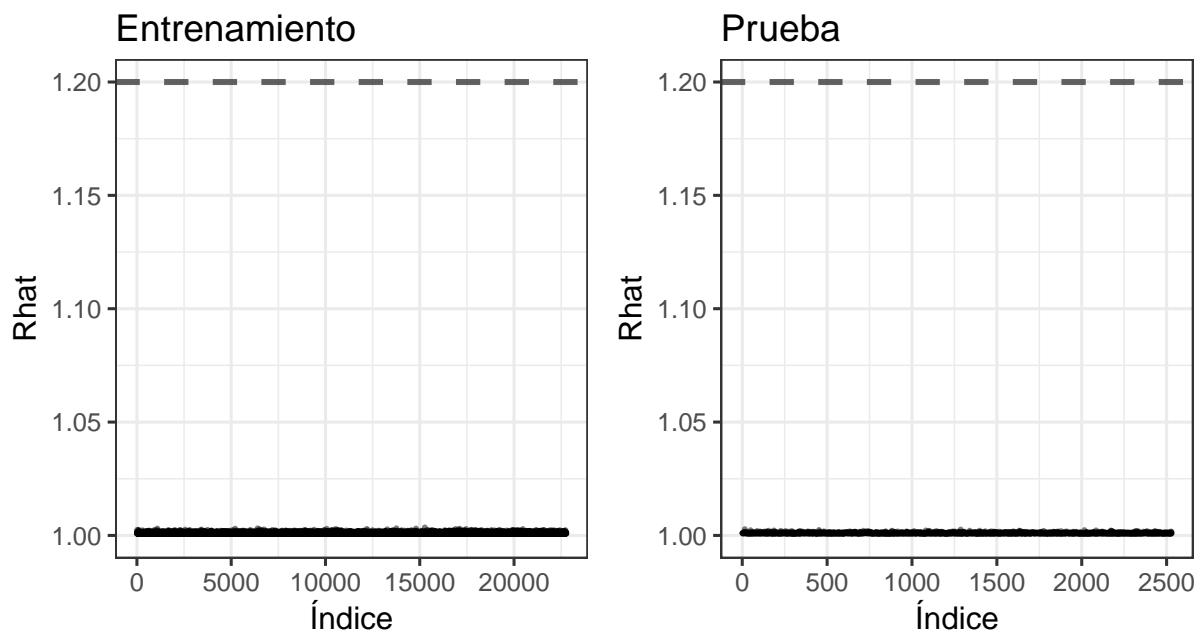


Figura 31: Estadística de convergencia \hat{R} de Gelman y Rubin para estimaciones de la variable respuesta del modelo multinivel

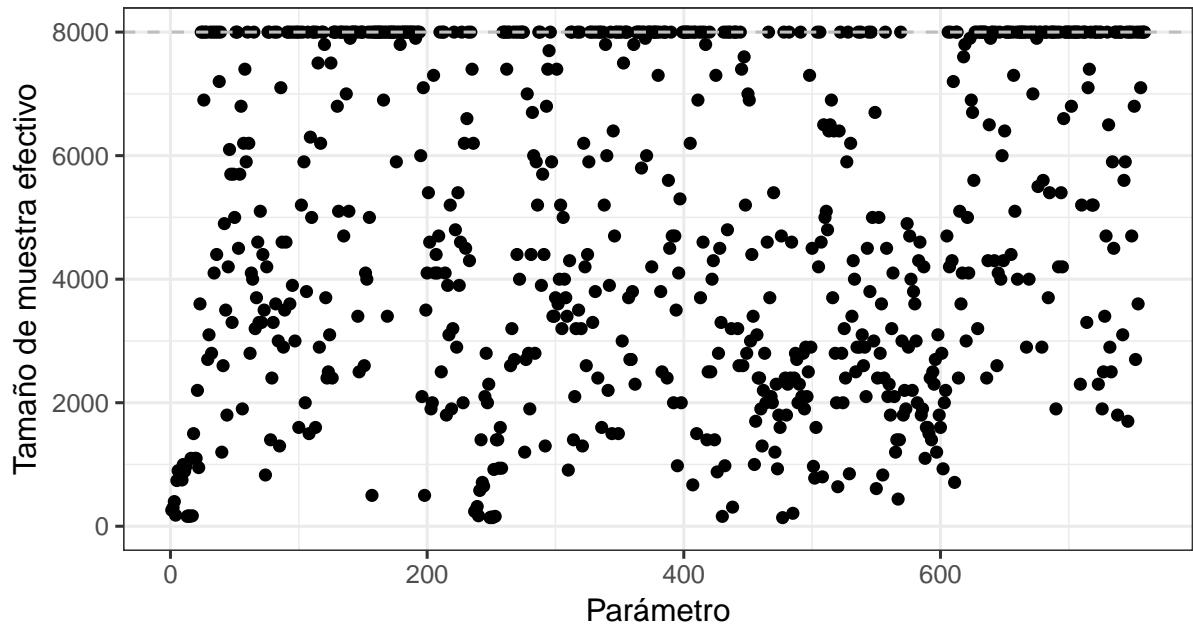


Figura 32: Tamaño de muestra efectivo para cada parámetro del modelo multinivel

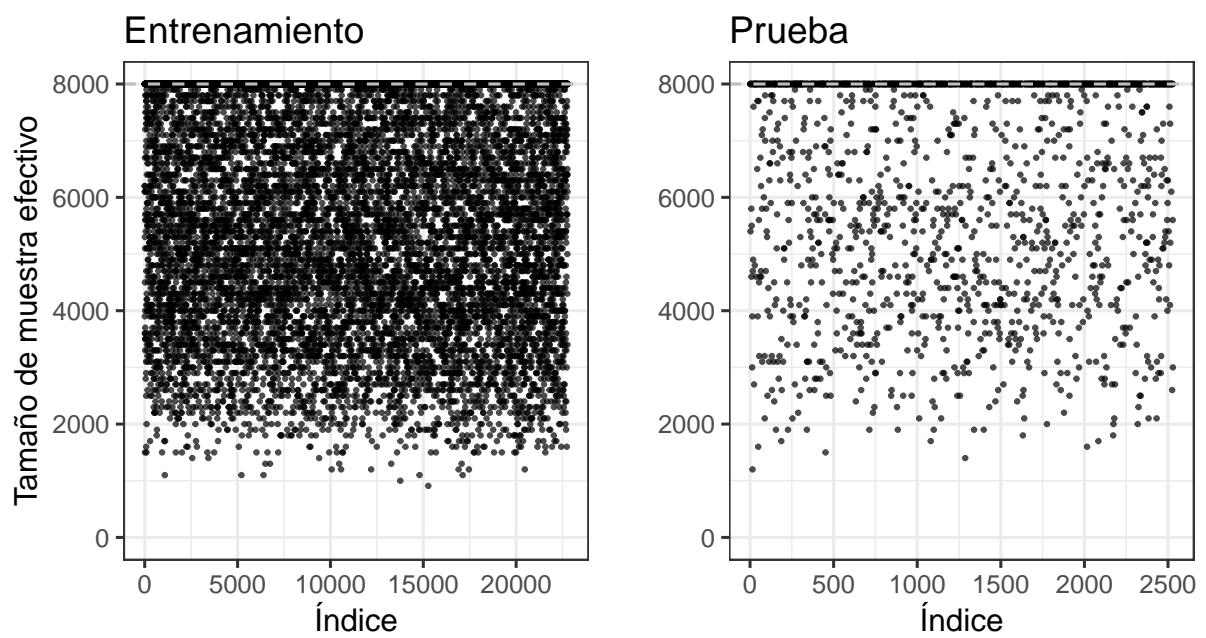


Figura 33: Tamaño de muestra efectivo para cada estimación de la variable respuesta del modelo multinivel