N4143 – Análisis de Datos e Inferencia Causal

Otoño 2024
Profesores: José Arenas, Omar Perez.
Equipo Docente: Antonia Aceituno, Guillermo Escobar, Camila Galarce, Diego Guevara, Rienzi Roldan, Karla Toledo y Martín Torrico.

TAREA 1

En esta tarea se abordarán los siguientes temas: construcción de base de datos, tipos de validez, tipo de datos y su análisis descriptivo, visualización, modelos de regresión lineal e inferencia estadística.

Entrega: jueves 18 de abril de 2024, hasta las 23:59 (entregas posteriores verán reducida su nota en 10 décimas por día de atraso).

En esta tarea deben examinar si determinados factores asociados a indicadores de desarrollo y gestión comunal (por ejemplo, superficie de parques urbanos, índice de pobreza CASEN, fondo de incentivo al mejoramiento de la gestión municipal) afectan la presencia y cantidad de viviendas irregulares **por comuna**. Para examinar esta temática se le pide utilizar los datos del SINIM[1] y el Catastro de campamentos comunal[2] del año 2022 (realizado entre el 2021 y 2022).

Esta es una tarea "semi-estructurada", por lo tanto, no existe una única forma de llevarla a cabo. Se evaluará la coherencia del proceso y del análisis. Para completarla deberán hacer lo siguiente:

1) Elaborar una hipótesis sobre el impacto de uno de los indicadores de desarrollo y gestión comunal sobre la cantidad de viviendas irregulares catastradas, por comuna. Deben justificar brevemente su hipótesis utilizando al menos dos referencias de estudios empíricos internacionales que sean relevantes y guarden relación a su hipótesis. Utilicen Google Scholar para buscar artículos científicos, y recuerden referenciar su fuente (10 pts.).
2) Construir una base de datos que permita examinar su hipótesis y que les permita controlar por diversas variables (por ejemplo, población comunal, índice de pobreza comunal, eficiencia de gestión aproximada por FIGEM[3], etc.). Dado que el tiempo y espacio es limitado, se deben enfocar sólo en algunas variables de interés. Expliquen brevemente el proceso de construcción de su base de datos, identificando claramente su unidad de análisis (p.j. persona, comuna, etc.) y periodo (año de los datos) (15 pts.).
3) Realizar un análisis exploratorio de datos (EDA) de sus variables[4]. Además, analicen con algunos gráficos su hipótesis diferenciando por comunas con un alto porcentaje de ruralidad (mayor valor que la mediana de % ruralidad de todas las comunas) y las que no (recuerde el ejemplo de la mortalidad por COVID en Italia y China visto en clases). Utilice los conceptos de visualización vistos en clases para una mejor comunicación de la información. Este análisis descriptivo debe apoyar coherentemente los modelos de análisis que realizarán posteriormente para testear sus hipótesis (25 pts.).
4) Utilizar regresiones lineales que permitan realizar inferencias en relación con su hipótesis. En particular, realicen varios modelos agregando variables explicativas en forma ordenada y justificada (miren el ejemplo de "brecha de género" en las *slides* de la clase "Lecture 4 – MLR"), explicando qué pasa con su relación de interés al agregar cada variable. En al menos un modelo

[1] Link base SINIM: https://datos.sinim.gov.cl/datos_municipales.php
[2] Link base MINVU: https://www.minvu.gob.cl/catastro-campamentos-2022/
[3] Si una comuna recibe FIGEM significa que cumplió con ciertos estándares de eficiencia en la gestión municipal establecidos a nivel central.
[4] Recuerde examinar las variables, por ejemplo, qué significan los valores faltantes o valores extremos, y cómo considerarlos.

agregue e interprete una interacción entre variables que guarde relación con los gráficos realizados en la parte 3. Es importante ir entendiendo los resultados que va agregando cada modelo e interpretar los resultados para cada variable en forma concisa en los modelos que estimen importante (no deben dar la interpretación de cada modelo en forma mecánica, sino que debe tener un sentido con respecto a lo que están estudiando). Además, deben hacerse cargo de los supuestos estadísticos de homocedasticidad y multicolinealidad para el modelo que consideren más relevante (de los modelos sin interacción) - es decir, testear y corregir si fuese necesario (30 pts.).

5) Discutir y concluir en base a la evidencia encontrada, incluyendo cualquier supuesto que hayan realizado para llevar a cabo su análisis. La conclusión debe estar vinculada a su hipótesis inicial. En la discusión, brevemente discutan cada tipo de validez (interna, externa y del constructo) vistas en el curso (20 pts.).

Formato: Máximo 6 páginas, excluyendo referencias. Letra Times New Roman número 12, interlineado 1,15 o 1,5; márgenes normales de página y tamaño carta.

La tarea tiene 100 puntos (60 pts = 4.0).

Algunos aspectos para considerar:

- Las hipótesis no son objetivos – se puede encontrar evidencia a favor o en contra de ésta. Procure que su hipótesis no predetermine su conclusión.
- Es mejor usar supuestos razonables y ser coherente con un análisis "enfocado" que explorar muchas cosas sin robustez ("quien mucho abarca poco aprieta").
- Sea cuidadosa/o con temas de formato, incluyendo ortografía y gramática, lo cual también será considerado en la evaluación.
- Uno de los objetivos de esta tarea es aprender a presentar resultados en forma efectiva, sabiendo seleccionar información que sea relevante y conteste directamente las preguntas del enunciado (es decir, evitar la costumbre de poner todo lo que se hizo para acumular páginas).
- Es posible usar apéndices (y hacer referencias a estos), pero se revisará sólo el contenido de las páginas establecidas en el documento principal.
- Se aconseja no dejar las tareas para el último momento. Por favor no solicitar extensión en los tiempos de entrega.

Recomendaciones técnicas:

- Para citar y referenciar una fuente bibliográfica podría servirles: https://biblioguias.uma.es/citasybibliografia/ejemplosAPA
- Para agrupar (agregar) datos en un solo nivel (por ejemplo, agregar datos de muchas viviendas para dejar un único dato a nivel comunal) Stata tiene el comando "collapse". En él se especifica la operación de agregación (por ejemplo, promediar o sumar) y si existen factores de expansión o ponderadores. Rstudio tiene el siguiente comando para colapsar BD= BD %>% group_by(variable) %>% summarise(cantidad_observaciones = n())
- En Rstudio es recomendable usar la librería dplyr con su operador "%>%" y aplique las funciones de la librería para un mejor manejo de las bases de datos. En particular utilice funciones como: group_by() summarise() select() unique(), entre otras.
- Para examinar multicolinealidad utilice el factor de inflación de la varianza (FIV, o VIF para inglés).
- Para mostrar resultados de múltiples regresiones, utilice un formato de tabla, como la mostrada en el ejemplo del estudio, en las *slides* de clases, de brecha de género en Uber.