**IN4143 – Análisis de Datos e Inferencia Causal**

**Primavera 2024**
**Profesor: Omar Pérez, Daniel Schwartz.**
**Auxiliares: Antonia Aceituno, Camila Galarce.**
**Ayudantes: Guillermo Escobar, Bastián Medina.**

## 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, inferencia estadística, interpretación y discusión en el contexto de la literatura científica existente.

Entrega: jueves 12 de septiembre de 2024, hasta las 23:59 (entregas posteriores verán reducida su nota en 10 décimas por día de atraso).

Esta tarea tiene como objetivo examinar los factores que influyen en la tasa de natalidad a nivel comunal en Chile, utilizando datos de la CASEN 2022 [link], datos del CENSO 2017 [link] y datos de tasa de natalidad [link]. 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. Ud. deberá hacer lo siguiente:

1) Desarrolle una hipótesis sobre la relación entre algún factor socioeconómico o demográfico y la tasa de natalidad en las comunas de Chile[1]. Justifique su hipótesis utilizando al menos tres referencias de estudios empíricos internacionales relevantes. Utilice bases de datos académicas como Google Scholar y asegúrese de citar correctamente sus fuentes. (10 puntos)

2) Construya una base de datos que permita examinar la hipótesis planteada, incluyendo variables como tasa de natalidad, población, ingreso promedio, horas trabajadas, pobreza multidimensional, urbanización, entre otras. Luego, debe colapsar los datos por comuna (ver recomendaciones técnicas) Explique brevemente el proceso de construcción de la base de datos, justificando la selección de variables y describiendo la unidad de análisis. Incluya variables que le permitan controlar por factores relevantes a su hipótesis (ver parte 5). (10 puntos)

3) Realice un análisis exploratorio de datos (EDA), agregando tablas con las principales variables en su análisis. Discuta cualquier transformación necesaria, si existen errores, problemas de datos faltantes, puntos de influencia o valores atípicos (outliers), y justifique sus decisiones para preparar los datos para el análisis. (5 puntos)

4) Analice gráficamente la relación de la variable de interés escogida en su hipótesis y la tasa de natalidad. Además, analice gráficamente esta relación usando una partición (2 o 3) por alguna dimensión que sea relevante para el análisis (por ejemplo, si depende de si es una comuna tiene alta tasa de mujeres). Utilice los conceptos de visualización vistos en clases para apoyar el análisis. Este análisis descriptivo debe apoyar coherentemente los modelos que Ud. realizará posteriormente en su análisis de regresiones (ver partes 5 y 6). (10 puntos)

5) Estime varios modelos de regresión lineal para examinar su hipótesis, agregando variables explicativas de manera progresiva y justificada (piense en el ejemplo de "brecha de género" visto en clases), explicando qué pasa con su relación de interés al agregar cada variable. Es importante ir entendiendo los resultados que va entregando cada modelo e interpretar los resultados para cada variable en forma concisa en los modelos que Ud. estime importante (no debe dar la interpretación de cada modelo en forma mecánica, sino que debe tener un sentido con respecto a lo que está

[1] Tal como se vio en clases, una hipótesis puede plantearse como el efecto de A hacia B para una población o bajo una condición X.

estudiando)[2]. Además, realice pruebas de homocedasticidad y multicolinealidad en el modelo que considere más relevante y discuta los resultados. Aplique correcciones si es necesario. (25 puntos)

6) En al menos un modelo, agregue e interprete una interacción entre variables que guarde relación con los gráficos realizados en la parte 4. (5 puntos)

7) Discuta los resultados de su análisis en relación con su hipótesis inicial, utilizando la literatura científica como marco de referencia. Evalúe la coherencia de sus hallazgos con estudios previos, destacando tanto similitudes como diferencias. (10 puntos)

8) Evalúe la validez interna, externa y del constructo de su estudio, considerando las limitaciones de los datos, los supuestos del modelo y el contexto en el cual se realiza el análisis. (10 puntos)

9) Realice la parte 5 utilizando GPT (puede poner los resultados en un Anexo). Discuta sobre las diferencias en conclusiones y posibles limitaciones del eso de este tipo de herramientas en este caso en particular. (10 puntos)

Formato: Máximo 6 páginas, excluyendo referencias y Anexos. Letra Times New Roman número 12, interlineado 1,15 o 1,5; márgenes normales de página y tamaño carta. Asegúrese de que su tarea esté bien escrita, libre de errores gramaticales y ortográficos. La claridad en la presentación de los resultados es crucial. (5 puntos).

La tarea tiene 100 puntos (60 pts = 4.0). Además, cada tutor sugerirá hitos de cumplimiento a lo largo de la tarea para terminarla en el plazo establecido.

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”).
- La construcción de la base de datos toma tiempo, y este gran esfuerzo se refleja de manera acotada en el informe. Esto es esperable y sucede en cualquier trabajo en lo cual es importante tener buenos "cimientos" para realizar el análisis y sacar conclusiones.
- 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).
- Utilice responsablemente GPT o herramientas similares, y que no sustituya su aprendizaje. Son herramientas importantes para ayudar a esclarecer conceptos y mejorar redacción. **Señale en su tarea qué partes fueron apoyadas por estas herramientas (Ej: "Usé GPT para elegir entre diversas hipótesis").**
- Es posible usar apéndices/anexos (y hacer referencias a estos), pero se revisará sólo el contenido de las páginas establecidas en el documento principal, por lo que los resultados, como gráficos y modelos de regresión deben estar incluidos en el cuerpo del informe.
- 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:

[2] Por ejemplo, “Para X1, se observa que, al aumentar X1 en 1, Y disminuye en 1,3%. Esta disminución se vuelve no distinguible de cero una vez que...”

https://biblioguias.uma.es/citasybibliografia/ejemplosAPA

- Para agrupar (agregar) datos en un solo nivel (por ejemplo, agregar datos de muchas familias para dejar un único dato a nivel comunal), en Rstudio la librería "dplyr" usa el comando "group_by()". En él se especifica la operación de agregación (por ejemplo, promediar o sumar). Un ejemplo sería:

  BD= BD %>% group_by(parámetro ) %>% summarise(weighted.mean(x, w))
- 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(), 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, visto en clases, de brecha de género luego de un MBA. Para esto es útil la libreria "stargazer" en Rstudio. Un ejemplo de cómo usarla sería:

  ```
  library(stargazer)
  stargazer(modelo_1 , modelo_2 , type = "text", df=FALSE)
  ```