

Proyecto de Investigación Inferencia Causal

1. Alcance del proyecto

1.1 Objetivo

Desarrollar un mini-paper aplicado y totalmente reproducible en R que responda una pregunta concreta de política pública utilizando al menos un diseño de identificación causal visto en clase (RCT/IV/RD/DiD/Synthetic Control, o combinaciones como DiD+IV).

El énfasis está en que el/la estudiante:

1. **Formule** una pregunta de investigación clara, con población, outcome y unidad de análisis definidos.
2. **Seleccione y justifique** un diseño de identificación consistente con el contexto empírico provisto.
3. **Implemente** el estimador correspondiente con especificaciones correctas (SE/clustering, FE, controles) y pruebas de supuestos pertinentes.
4. **Comunique** hallazgos con magnitudes e interpretación económica y derive implicaciones de política.

1.2 Flujo de trabajo

Este proyecto se organiza en cuatro etapas secuenciales para que el estudiante aplique, de forma guiada y reproducible, las técnicas de identificación causal vistas en clase sobre un conjunto de datos sintéticos diseñado a partir de su propuesta. La ruta comienza con la formulación rigurosa de la pregunta y el marco metodológico, continúa con la entrega de los datos y su contexto, avanza con la elaboración del mini-paper y culmina con la defensa de resultados.

1. *Entrega 1 (Propuesta)*. Documento en PDF que precise la pregunta de investigación, el outcome y las variables de interés, la población y unidad de análisis, el marco metodológico y el diseño tentativo de identificación. Debe indicar, en el escenario ideal, cómo se estimaría el efecto (estimador, especificación y supuestos clave), además de amenazas potenciales y su mitigación.

2. *Entrega de datos sintéticos*. Con base en la propuesta, el docente entrega un paquete de datos y un contexto de política coherentes con la pregunta planteada. El paquete incluirá diccionario de variables y lineamientos para que el método elegido sea aplicable sin restricciones de acceso o confidencialidad.
3. *Entrega 2 (mini-paper)*. Aplicación del método seleccionado al paquete de datos, con una especificación base, diagnósticos y pruebas de supuestos, análisis de robustez y, cuando sea pertinente, heterogeneidad. El documento debe comunicar los resultados con interpretación económica y derivar implicaciones de política.
4. *Defensa de resultados*. Presentación y discusión de hallazgos en la última semana de clase. Se espera que cada estudiante exponga la lógica del diseño, los supuestos y diagnósticos, la magnitud e interpretación del efecto y las principales limitaciones.

2. Acerca de los Datos

Tras la primera entrega (*propuesta del estudiante*), el docente proveerá a cada estudiante una base de datos sintética coherente con la pregunta planteada y un contexto de política (narrativa breve con instituciones, calendario, reglas de elegibilidad o asignación, etc.). El objetivo de los datos sintéticos es permitir el uso riguroso de las técnicas del curso sin restricciones de acceso o confidencialidad.

2.1 Formato y estructura

Cada paquete de datos incluirá un archivo `data.csv` con las variables mínimas ppropuestas por el estudiante más las variables necesarias para desarrollar la estimación que permita resolver la pregunta de investigación:

- Identificadores: `id` (unidad), `t` (tiempo) si panel o series.
- Tratamiento/elegibilidad: `D`, `eligible`, `post`, `treated_at` (si adopción escalonada).
- Potenciales instrumentos (si aplica): `Z` (asignación/lottery, distancia, regla).
- Variable de corte (si RD): `running` y umbral z_0 .
- Outcome principal y secundarios: `Y`, `Y_alt`.
- Covariables y clústeres: `X_k`, `cluster_id` (escuela, municipio, etc.).

Además, el repositorio de los datos contendrá un archivo `README.md` con el contexto, las definiciones de variables, unidad de análisis, periodo y sugerencias de especificación.

2.2 Tipos de diseño contemplados

Para asegurar la coherencia propuesta \rightarrow *datos* \rightarrow *método*, el paquete de datos sintéticos generado por el docente contemplará una estructura que anticipe el diseño de identificación más apropiado. A continuación se describen los tres esqueletos de datos más frecuentes y sus lineamientos de uso.

(A) Corte transversal *Cuándo procede:* una sola ola de información con asignación aleatoria/as-if random (RCT/ITT), presencia de *instrumento* (IV) o elegibilidad por *umbral* (RD nítida/difusa).

(B) Panel o cortes repetidos *Cuándo procede:* cambios de política en el tiempo con grupos tratados y de control; adopción *simultánea* (DiD estático) o *escalonada* (tratamiento en distinto t).

(C) Caso tratado único con donantes *Cuándo procede:* una unidad tratada (o pocas) y un conjunto de donantes con larga ventana pretratamiento.

2.3 Acompañamiento y soporte

Las tutorías tendrán un enfoque exclusivamente conceptual y metodológico. Su propósito es afinar la aproximación a la pregunta con los datos, verificar la coherencia propuesta \rightarrow *datos* \rightarrow *método*. No se resolverán dudas de programación ni se hará depuración de scripts. Estas sesiones se realizarán por estudiante, después de la entrega de los datos y antes de la defensa, con una duración estimada de veinticinco a treinta minutos.

3. Acerca de los Entregables

3.1 Sesión 1 — Propuesta (1–2 páginas en PDF)

Esta primera entrega es una hoja en prosa, sin resultados ni código. Debe formular una pregunta de investigación que pueda responderse con una regresión simple y explicitar una hipótesis breve sustentada en la literatura o en el contexto (un párrafo a lo sumo). Indique con precisión la población a la que se refiere la inferencia, el período de interés y el nivel de observación (por ejemplo, persona, firma, escuela o municipio). En esta etapa no se usan datos ni se ejecuta código.

El núcleo del documento es la especificación de la regresión que correría asumiendo una base de datos ideal (con todas las variables necesarias): escriba la ecuación y describa sus componentes, por ejemplo $Y_i = \alpha + \beta X_i + \varepsilon_i$. Importante: no limite su planteamiento por la disponibilidad actual de datos; piense en las variables que le permitirían estimar la regresión

ideal (más adelante se discutirá cuál aproximación metodológica permite aproximarse a la variable de interés). En ese marco, defina en texto qué es Y (resultado que desea explicar) y qué es X (variable cuyo efecto desea medir sobre Y); detalle cómo se medirían o construirían, en qué unidad y período, y qué signo o magnitud espera para β según la hipótesis. Si considera indispensable incluir algún control, méncionelo y justifique su función en una frase; la expectativa principal es una relación Y – X clara y bien definida.

Finalmente, explique por qué, en la práctica, no podría estimar exactamente esa regresión ideal (por ejemplo, dilemas éticos, restricciones de medición o presencia de endogeneidad), sin proponer aún estrategias alternativas. Para orientar el formato y el contenido, se compartirá un ejemplo de entregable que ilustra cómo redactar la pregunta, plantear la hipótesis y presentar la especificación de la regresión en una página (descargar ejemplo [aquí](#)).

Fecha y método de entrega: La propuesta deberá enviarse en formato PDF al correo efmartinez@icesi.edu.co antes de las 08:00 horas del lunes 29 de septiembre, con el asunto: Propuesta Curso Inferencia Causal.

3.2 Entrega 2 — Mini-paper

El texto debe ser conciso —máximo 10 páginas sin anexos— y centrado en la evidencia empírica. Debe incluir, como mínimo, Resumen (Abstract), Introducción, Datos y metodología, Resultados y Discusión final. Evite texto de relleno: cada párrafo debe aportar a la narrativa y al argumento empírico.

3.2.1 Sesiones

Introducción (1–2 páginas): Piense esta sección como un “abstract extendido” en 4–6 párrafos, escrito en prosa clara y sin tecnicismos innecesarios. El objetivo es que, al terminar la introducción, el lector sepa con exactitud cuál es la pregunta, a quién aplica, qué estima el documento, qué encuentra y por qué importa para política y para la literatura.

Párrafo 1 — Motivación y pregunta. Abra con 2–3 oraciones que describan el problema de política (contexto, magnitud, por qué importa) y cierre con la pregunta de investigación en una sola oración, formulada de modo que pueda contrastarse con una regresión.

Párrafo 2 — Qué hace el documento. Declare en presente lo que el trabajo hace: estima el efecto de X sobre Y , reporta su magnitud en unidades interpretables y señala por qué el resultado es relevante para una decisión de política. Mantenga el foco en qué se mide y qué se aprende, sin detallar aún el cómo.

Párrafo 3 — Metodología en una frase. Describa en una o dos oraciones la aproximación empírica, la población o área de estudio y la ventana temporal, sin tecnicismos: “Aprovechamos la variación exógena del programa Z para identificar el efecto de X sobre Y ; estimamos la

especificación principal en [población] durante [período] al nivel de [unidad de observación].” Los pormenores se desarrollan en la sección de metodología.

Párrafo 4 — Resultados principales. Anticipe el hallazgo central con magnitud y unidades comprensibles (no más de dos números clave) y una línea sobre su interpretación económica. Evite tablas, figuras o discusiones extensas aquí.

Párrafo 5 — Aporte y organización. Ubique el trabajo respecto a la literatura en una o dos frases (qué vacío llena o cómo extiende evidencia existente) y cierre con el mapa del documento: “La sección 2 describe los datos; la sección 3 presenta la metodología; la sección 4 reporta resultados; la sección 5 discute implicaciones y limitaciones.”

Datos y metodología (3–4 páginas): Esta sección debe convencer al lector de que el efecto estimado es causal. Describa con precisión la unidad de análisis, el período y las variables clave (tratamiento o elegibilidad, *outcome* y covariables) y explique el vínculo institucional que genera variación exógena plausible en el tratamiento: por qué la regla, programa o shock utilizado induce asignación “buena” (p.ej., umbral, sorteo, cambio de política en fecha conocida) y, por tanto, no está correlacionada con shocks no observados del *outcome*. El objetivo es mostrar que, dadas las reglas del contexto, la comparación es creíble para inferencia causal.

Exponga de forma breve los supuestos de identificación y cómo se verificarán empíricamente, y presente la especificación principal en una ecuación compacta (estimando de interés, muestra efectiva, efectos fijos si corresponden y nivel de agrupación de errores estándar). Indique los diagnósticos que el lector verá según el diseño (balance/attrition en experimentos, primera etapa y F en IV, manipulación y continuidad en RD, tendencias paralelas y *event-study* en DiD) y cierre señalando el plan mínimo de robustez (ventanas/bandas alternativas, especificaciones alternativas, exclusión de *outliers*). Evite detalles de implementación de código: la trazabilidad técnica quedará en el repositorio reproducible.

Resultados (2-3 páginas): Presente primero la tabla central con el efecto principal y su interpretación económica. Añada los diagnósticos mínimos que respaldan la estimación y uno o dos ejercicios de robustez razonables.

Discusión final (1-2 páginas): Resuma qué se aprende, límites del ejercicio y la implicación más directa para política.

3.2.2 Reproducibilidad de resultados

El análisis deberá ejecutarse desde un único script documentado (por ejemplo, `00_run.R`) que reproduzca íntegramente todas las tablas y figuras del manuscrito en un entorno limpio. Use rutas relativas y registre versiones del software. La ejecución del script debe generar todos los insumos del documento sin intervención manual. Exporte las salidas a carpetas dedicadas (por ejemplo, `output/tables` y `output/figures`) con nombres cortos y fácilmente

asociables al texto (p. ej., `tab_main.tex`, `fig_eventstudy.pdf`). Evite rutas absolutas y archivos editados manualmente.

3.2.3 Fecha y método de entrega:

Deberá enviar el PDF y el script de estimaciones al correo efmartinez@icesi.edu.co antes de las 08:00 horas del lunes 20 de octubre, con el asunto: **Mini-paper Curso Inferencia Causal**.

3.3 Defensa de resultados

La defensa busca que el estudiantado aprenda a presentar un artículo académico. Cada persona dispondrá de un bloque de 30 minutos: 20–25 minutos de presentación sin interrupciones y 5 minutos de preguntas. La exposición debe contar una historia con una regresión: formular con claridad la pregunta y la hipótesis, situar brevemente los datos y la unidad de análisis, enunciar la especificación principal en términos simples, mostrar el resultado central con su magnitud e interpretación económica y cerrar con una implicación concreta para política.

La evaluación se centrará en la capacidad de construir una narrativa coherente entre pregunta, datos y regresión; en la claridad y corrección de la especificación presentada; en la interpretación económica del efecto; en la mención concisa de los diagnósticos mínimos que sostienen la estimación; en el manejo del tiempo y en la solvencia para responder preguntas. Les dejo estos [tips](#) para hacer una buena presentación.

4. Rúbrica (100 pts)

- **Entrega 1 (10%)**
 - Pregunta e hipótesis claras; población, período y nivel de observación definidos (4 pts).
 - Especificación propuesta $Y = \alpha + \beta X + \varepsilon$: definición precisa de Y , X y controles indispensables; factibilidad (4 pts).
 - Presentación y cumplimiento de formato/fecha (2 pts).
- **Entrega 2 (60%)**
 - Identificación: coherencia datos–método y supuestos explícitos en el contexto (18 pts).
 - Implementación: especificación correcta, FE si aplica, nivel de clustering, muestra efectiva (12 pts).

- Evidencia/diagnósticos/robustez: verificación de supuestos y al menos dos ejercicios de sensibilidad (12 pts).
- Interpretación y discusión: magnitudes comprensibles, alcances y limitaciones, implicación de política (10 pts).
- Reproducibilidad: script único, rutas relativas, `renv/sessionInfo()`, README e insumos generados automáticamente (6 pts).
- Gestión: entrega a tiempo, nomenclatura y organización del repo (2 pts).
- **Defensa (30%)**
 - Narrativa: historia coherente pregunta–datos–regresión–resultado (10 pts).
 - Claridad técnica de la especificación y de los diagnósticos clave (8 pts).
 - Interpretación económica y relevancia para política (6 pts).
 - Visuales y manejo del tiempo; solvencia en preguntas (6 pts).

Nota Se permite el uso de IA con declaración explícita de qué se usó y para qué.