

Notas de Econometría

Pavel Solís
2025

8 Consideraciones para un Proyecto Empírico

Ingredientes de un buen análisis empírico

8.1 Definir la pregunta de investigación

- Definir una pregunta muy específica si se puede contestar con datos
- Formular hipótesis y el modelo a estimar antes de recolectar datos para evitar
 - Olvidar recolectar información de variables importantes
 - Obtener una muestra de una población incorrecta
 - Recolectar datos de un período de tiempo incorrecto
- Verificar que las fuentes de datos existan y que permitan responder la pregunta a tiempo
- La pregunta tiene que ser de interés general con implicaciones de política (país) o local (universidad, comunidad)
 - Discutir con otras personas para evaluar si es de interés
- Temas: economía laboral, finanzas públicas, economía educativa, crecer, consumo (alcohol), variables macro, variables financieras, finanzas corporativas
- Es válido utilizar
 - Datos más recientes, más desagregados
 - Modelos aplicados a nuevas unidades (industria, país, estado)
 - Nueva fuente de datos (datos panel, experimento natural)
 - Nuevos modelos

8.2 Revisión de la literatura

- Todos los artículos deben incluir una revisión de la literatura relevante
- Hay diferentes fuentes para encontrar artículos específicos sobre un tema
 - Journal of Economic Literature (JEL): Diseñó sistema de clasificación por códigos/categorías
 - Journal of Economic Perspectives (JEP): Resúmenes de temas
 - EconLit (suscripción): Búsqueda amplia de muchos journals por autor, tema, palabras
 - Econ Papers: Colección de artículos en economía (publicados y no publicados)

- Social Science Research Network: Artículos de trabajo
- Google Scholar: Artículos publicados o de trabajo
- Lean abstracts de los artículos para ver qué tan relevantes son para su propio trabajo
- Busca también temas relacionados (uso de drogas, uso de alcohol)
- Por preferencia o necesidad
 - Extensión: corta o larga
 - Ubicación: como parte de la introducción o en una sección aparte

8.3 Recolección y análisis de datos

8.3.1 Recolección del tipo de datos apropiado

- Tipos de datos más comunes:
 - Corte transversal, series de tiempo, CT independientes, datos panel
- El tipo de datos a recolectar depende de la naturaleza del análisis
 - ¿Hay suficientes controles para hacer un análisis ceteris paribus razonable?
 - Considerar la endogeneidad de la variable independiente a pesar de los controles
 - Se pueden necesitar datos panel
 - * Individuos o familias (Inegi)
 - * Empresas (Compostat, CNBV)
 - * Ciudades, municipios, estados, países
- Formatos varían
 - Solo impresos (manual, encuesta entre amigos): csv, xlsx, faltantes, strings
 - Digitalizables (bibliotecas, ML)
 - Electrónicos (FMI, BM): csv, xlsx
 - Proporcionados por autores
- Seguir convenciones al guardar datos
 - Corte transversal: Observaciones en filas, variables en columnas
 - Series de tiempo: Más reciente al último
 - Datos panel: Cada unidad con tiempo ordenado

8.3.2 Análisis preliminar de los datos

- Familiarízate con datos a utilizar en el análisis, dedicar tiempo a conocer la base
- Cómo se codifican los datos faltantes
 - Carácter no numérico (-, N/A, NE),
 - Carácter no numérico ("999", "-1"): Cuidado con estadística descriptiva
 - Convertir códigos numéricos en no numéricos (.) para no confundirlos
- Conocer naturaleza de las variables:
 - Binarios, ordinales (calificación crediticia), unidades de medida (monetarias)
 - Tasas (porcentaje, proporción), nominal/real (año base), logaritmos
- Importante tratar de detectar errores:
 - Estadística descriptiva (min, max, media, desvest) para todas las variables
 - Revisar si códigos numéricos para faltantes
 - Unidades de medida de tasas (a veces mezcla entre porcentaje y proporción)
 - Series de tiempo (ajuste estacional, orden cronológico para cambios, variable que indique tiempo)
 - Datos panel (cambios dentro de unidades implican faltantes para primer observación)

8.4 Análisis econométrico

- Para justificar MCO, discutir si los supuestos MCO se satisfacen en tu modelo
 - ¿El error no está correlacionado con variables independientes?
 - Hay suficientes controles y los que quedan en el error no están correlacionados con los regresores
 - Discutir si fuentes de endogeneidad son un problema serio
 - * Autoselección
 - * Error de medición
 - * Simultaneidad
- Decidir sobre forma funcional
 - Variables en logs, incluir cuadrados para efectos decrecientes, variables binarias solas o interactuadas
- Error incluir variables independientes que tienen valores numéricos pero no significado cuantitativo (variable con ID de ocupación) o variables cualitativas ordenadas
 - Incluir variables dummies
- Si hay heterocedasticidad, utilizar estadísticos ajustados por heterocedasticidad

- Cuando variable dependiente es ordinal, utilizar modelo apropiado (probit, logit binarios u ordenados)
- Con series de tiempo: niveles vs diferencias, modelar estacionalidad, número de rezagos
- Si modelo mal especificado por variable omitida, hacer análisis de mala especificación
 - Dirección del sesgo
- Si se usan variables instrumentales, evaluar validez
 - Resuelve endogeneidad por variable omitida, error de medición y simultaneidad
- Incluir análisis de sensibilidad
 - Estimar modelo original y modificado de forma razonable
 - Idealmente, las conclusiones se mantendrán
- Evaluar si hay valores extremos (algunas observaciones son muy diferentes al resto)
 - ¿Resultados cambian si se excluyen esas observaciones?
 - En caso afirmativo
 - * Ajustar forma funcional
 - * Argumentar que siguen otro modelo
- Con datos panel:
 - Controlar por no observables fijos en el tiempo
 - Errores correlacionados (efectos aleatorios)
 - Incluir rezago de variable dependiente
 - Efectos fijos
- Cuando aplica, se recomienda usar varios métodos econométricos y comparar resultados
 - Generalmente permite determinar cuál de los supuestos puede no cumplirse
- Se pueden llegar a obtener resultados contraintuitivos
 - Hacer *data mining* (probar diferentes modelos, técnicas de estimación o subconjunto de datos hasta obtener algo más en línea con lo esperado) viola los supuestos
 - Si usamos los mismos datos en búsqueda de una especificación, los estimados y las pruebas no son independientes entre ellos
 - Una forma común de cometer ese error: regresión gradual (*stepwise*)
- Minimizar: no buscar varios modelos o métodos de estimación hasta encontrar resultado significativo y solo reportar el resultado original
- Si una variable es estadísticamente significativa en una pequeña fracción de los modelos estimados, probablemente no tiene efecto en la población

8.5 Escribir un artículo empírico

- Un buen artículo combina un análisis de datos cuidadoso y convincente con buenas explicaciones y argumentación
- Muchos artículos siguen la misma estructura

Introducción

- Objetivo, por qué es importante, revisión de literatura, estadísticos o gráficas de resumen, resumen de resultados principales
- Modelos (teórico, econométrico): Modelo teórico, discusión de variables a incluir en el modelo, contexto, ecuación para el modelo poblacional, discusión del método de estimación

Datos

- Descripción de datos usados para que lector pueda obtener y rehacer el análisis, cuestionario utilizado
- Discusión de unidades de medición de las variables
- Tabla de estadísticos descriptivos (min, max, media, desvest) para
 - Facilitar interpretación de los coeficientes
 - Enfatizar unidades de medición
- Mencionar el número de observaciones usadas (años, unidades de corte transversal por año)

Resultados

- Reportar estimados de cualquier modelo mencionado
 - Generalmente en tablas: R^2 , número de observaciones
- Discutir interpretación y fuerza de resultados
 - Signos esperados
 - Significancia estadística (coeficientes vs errores estándar)
 - Magnitudes
 - Diferencias entre métodos de estimación

Conclusiones

- Resumir resultados principales
- Mencionar salvedades de conclusiones hechas
- Sugerencias para futuras investigaciones

8.5.1 Consideraciones finales

- Título debe ser informativo pero no muy largo
- Incluir un abstract de entre 100 y 200 palabras
- Incluir al final todas las referencias citadas
- Escoger unidades de medición (reescalar) para variables (dependiente e independientes) de forma que coeficientes no sean muy grandes o pequeños
 - No usar notación científica ni notación exponencial (2.5×10^5 , $2.5e + 5$) para coeficientes ni errores estándar
 - Limitar número de dígitos después del punto decimal (redondear a 2 o 3 decimales)
- En apéndice resultados adicionales no esenciales