Profesor: Manuel Fernández

- David Arboleda
- Valentina Castilla
- Valentina Daza
- Santiago Torres.

Taller 4

Econometría Avanzada, 2022-2
Fecha de Entrega: 2 de diciembre, 2022

Universidad de los Andes

Primer ejercicio

En esta ocasión, un colegio de Bogotá los contrata para estudiar las razones por las cuáles los estudiantes faltan a clases. Para ello, les dan información acerca de las ausencias registradas de 316 estudiantes, junto con algunas de sus características. La información la encuentran en la base de datos *"Ausencia_data"*. Las variables en la base de datos son las siguientes:

- *id*: Identificador único de estudiante.
- *daysabs*: Número de días que el/la estudiante estuvo ausente.
- *female*: Dummy que toma el valor de uno si la estudiante es mujer.
- *mathpr*: Percentil de la nota obtenida por el estudiante en la clase de matemáticas ($1 \equiv$ top 1 % mejor de la clase).
- *langpr*: Percentil de la nota obtenida por el estudiante en la clase de lenguaje ($1 \equiv$ top 1 % mejor de la clase).

a) Presenten un histograma que refleje la distribución del número de días que los estudiantes estuvieron ausentes. Modifiquen el diagrama de manera que sea interpretable como una función de masa de probabilidad. Describan brevemente el comportamiento de la gráfica en el valor 0.

Dada la estructura particular de los datos, su jefa les comenta que encontró unos modelos nuevos que serían potencialmente útiles para lograr su análisis. Para ello, les propone la siguiente breve introducción:

Los modelos lineales generales (GLM) son una alternativa popular a la regresión lineal en circunstancias donde la variable dependiente no es continua o la relación entre X y W no es lineal. Suponga que contamos con una muestra aleatoria $\{(W_i, X_{i,1}, \cdots, X_{i,k})\}_{i=1}^{n}$. Un modelo lineal general está definido por una transformación invertible $g : \mathbb{R} \to \mathbb{R}$, con la cual especificamos una relación funcional de la forma:

$$\mathbb{E}[W_i|X_i] = g^{-1}(X_i^T\beta)$$

donde $X_i^T = (X_{i1}, X_{i2}, \cdots, X_{ik})$ y $\beta^T = (\beta_1, \beta_2, \cdots, \beta_k)$.

Estos modelos son particularmente importantes cuando la variable dependiente es dicótoma o categórica. Así, han encontrado buena acogida entre los científicos de datos por su flexibilidad, fácil estimación e inferencia.

b) Como primera aproximación, veamos que los modelos generales lineales no son desconocidos. Para ello consideren que W_i es una variable dicótoma en $\{0, 1\}$. Especifiquen funciones $g(\cdot)$ tales que

   i) El modelo resultante es un modelo de probabilidad lineal.
   ii) El modelo resultante es un Logit.
   iii) El modelo resultante es un Probit.

   Ayuda: Para esto piensen en la forma que toma $\mathbb{E}[W_i|X_i]$ en cada escenario.

Una instancia muy frecuente de uso de los GLM es para modelar datos de conteo, esto es, cuando la variable dependiente W_i es tal que $W_i \in \mathbb{N}_0 = \{0, 1, 2, \cdots\}$. En este taller exploraremos un tipo de modelo muy importante en actuaría, economía de la salud, epidemiología, psicología, entre otras áreas del conocimiento, conocido como modelos con ceros inflados ("Zero-inflated models").

La idea de los modelos con ceros inflados es modelar datos que provienen de una mezcla de procesos generadores de datos, uno de los cuales siempre produce ceros, mientras que el otro produce datos con valores en los números naturales. Estos modelos son muy útiles para entender fenómenos en donde existe una gran masa de los datos concentrados en el 0, pues explican cuál es la lógica detrás de estas acumulaciones.

Hablando con los profesores del colegio, ellos les comentan que existen dos tipos de estudiantes. El primer tipo son aquellos estudiantes que faltan si y solo si se enferman o tienen algún compromiso importante al cual no pueden faltar. Noten que existe la posibilidad de que estudiantes de este tipo no falten en todo el año, condicional en que no se les presente ninguna de estas circunstancias. Por otra parte, el segundo tipo de estudiantes son aquellos que son obligados por sus padres a asistir al colegio, y que no pueden faltar bajo ninguna circunstancia. Así las cosas, la existencia del segundo tipo de estudiantes puede ser la causa de que exista una gran masa de probabilidad acumulada en el cero.

Para formalizar esta intuición, podemos pensar en el siguiente modelo matemático: Hay 2 tipos de estudiantes. Sea T_i una variable aleatoria binaria que toma el valor de uno si el estudiante es de tipo 2 (esto es, que nunca falta a clase). Suponemos que $T_i \sim$ Bernoulli(π_i), para $0 < \pi_i < 1$ desconocido. Además, sea E_i el número de eventos, tales como enfermedades o compromisos, que tiene el estudiante i en un año. Supondremos que estos eventos siguen un proceso de Poisson estándar, por lo cual $E_i \sim$ Poisson(λ_i), donde λ_i es la tasa a la cual estos sucesos ocurren. Así las cosas, el número de faltas que tiene un estudiante i, el cual llamaremos Y_i sigue el proceso:

$$Y_i = (1 - T_i)E_i \tag{1}$$

c) Siguiendo (1), calcule la función de masa de probabilidad para Y_i.

Ayuda: Recuerden que la función de masa de probabilidad[1] de una variable aleatoria $W \sim$ Poisson(λ) es

$$p(w) = \mathbb{P}(W = w) = \frac{e^{-\lambda}\lambda^w}{w!} \qquad w = 0, 1, 2, \cdots$$

Ahora incorporaremos la noción de los modelos generales lineales para entender qué factores inciden en que un estudiante sea de tipo 1 y tipo 2, así como de que le ocurran eventos por los cuales dejaría de asistir al colegio. Sean $X_i = (X_{i1}, \cdots, X_{ik})^T$ y $Z_i = (Z_{i1}, \cdots, Z_{ir})^T$, dos vectores (no necesariamente iguales) de características de un individuo i. Suponemos entonces que:

$$\ln(\lambda_i) = \ln(\mathbb{E}[E_i|X_i]) = X_i^T\beta \tag{2}$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \ln\left(\frac{\mathbb{E}[T_i|Z_i]}{1-\mathbb{E}[T_i|Z_i]}\right) = Z_i^T\gamma \tag{3}$$

donde $\beta^T = (\beta_1, \beta_2, \cdots, \beta_k)$ y $\gamma^T = (\beta_1, \beta_2, \cdots, \beta_r)$ son dos vectores desconocidos de parámetros a estimar.

Debido a la naturaleza del problema, el método de estimación más sencillo de ejecutar es el de máxima verosimilitud. Para ello, necesitamos conocer la función de máxima verosimilitud:

d) Hagan lo siguiente:

i) Demuestren que la función de log-verosimilitud $l(\beta, \gamma|X, Z)$ asociada al problema, bajo el supuesto que $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ es una muestra i.i.d, donde $Y_i|X_i, Z_i$ sigue la distribución encontrada en c), pero usando las formas funcionales de los parámetros del modelo dadas en (2) y (3), es

$$l(\beta, \gamma|X, Z, Y) = \sum_{i=1}^{n} \left[D_i \ln(\exp(Z_i^T\gamma) + \exp(-\exp(X_i'\beta)))\right] + (1 - D_i)[y_i(X_i^T\beta) - \exp(X_i^T\beta) - \ln(y_i!)] - \ln[1 + \exp(Z_i^T\gamma)]$$

donde D_i es una variable dicótoma auxiliar que indica si la persona i nunca estuvo ausente.

ii) Deriven un sistema de ecuaciones cuya solución resulta en los estimadores de máxima verosimilitud (no intenten resolver el sistema).

[1] Recuerden Y_i es una variable aleatoria discreta. Por tanto, hablamos de su función de probabilidad o función de masa de probabilidad. Cuando tenemos una variable aleatoria continua, en cambio, hablamos de su función de densidad de probabilidad.

Ayuda: Para el ii), no se olviden que β, γ son vectores.

El modelo que acabamos de construir se conoce como el modelo de Poisson con ceros inflados (ZIP-Zero inflated Poisson), y es implementable tanto en Stata (comando *zip*) como en R (*zeroinfl* del paquete *pscl*).

e) Usando los datos provistos,

i) Estimen el modelo ZIP. usando que

$$X_i^T \beta = \beta_0 + \beta_1 \text{mathpr}_i + \beta_2 \text{langpr}_i + \beta_3 \text{female}_i$$

y

$$Z_i^T \gamma = \gamma_0 + \gamma_1 \text{mathpr}_i + \gamma_2 \text{langpr}_i$$

Presenten los resultados en una tabla.

ii) Interpreten $\hat{\gamma}_1$.

Segundo ejercicio

Aumentar la participación política de las mujeres constituye uno de los grandes retos de las sociedades modernas. Una mayor representatividad de las mujeres en los cuerpos legislativo permite que problemáticas históricamente obviadas y que afligen de manera diferenciada al sexo femenino sean puestas en el foco de la discusión, lo cual favorece la implementación de políticas y programas que atienden dichas problemáticas y que permiten forjar el camino hacia sociedades más equitativas. No obstante, debido a desigualdades estructurales, aumentar la participación política de las mujeres suele ser difícil, por lo que múltiples países han recurrido a leyes de cuotas para intentar dar un impulso inicial a la representación femenina. Por las razones previamente mencionadas, dichas leyes, además de incrementar la participación política femenina, suelen incidir sobre variables asociadas al bienestar de las mujeres. En particular, para el presente caso de estudio a ustedes les interesa conocer el impacto de las leyes de cuotas sobre la mortalidad materna. Para ello, ustedes cuentan con la base *partfem.dta*, la cual conforma un panel anual para 119 países entre 1990 y 2015. Dicha base contiene las siguientes variables:

Variable	**Descripción**
lnmmrt	Logaritmo natural del número de muerte de mujeres hasta 42 días después del parto
womparl	$\%$ de mujeres en el parlamento
country	País
year	Año
quota	$= 1$ si el país tiene una ley de cuotas
lngdp	Logaritmo natural del PIB

1. Para comenzar, reporte en una tabla cuáles países de la muestra tienen leyes de cuotas y en qué año las adoptaron. Creen una variable que contenga esta información.

2. Ahora, suponga que se desea estudiar el impacto de la ley de cuotas para algunos países específicos de la muestra. En especial, ustedes quieren estimar el efecto de la política sobre la mortalidad materna en Djibouti, Jordan y Kenya. Para ello, un compañero les sugiere emplear la metodología de control sintético, aplicada individualmente a cada uno de los países de interés. En este sentido, sigan los siguientes pasos para cada uno de los países mencionados:

a) Construyan, por medio del comando *synth*, el control sintético del país correspondiente. Seleccionen las variables empleadas para la predicción y argumenten su elección.

b) Recuperen los pesos de cada uno de los países del pool de donantes y construyan un histograma de dichos pesos. Interpreten sus resultados.

c) Reporten en una gráfica (i) la serie de tiempo real y el control sintético, y (ii) la evolución de la diferencia entre estos. Basado en estas, argumenten si la estrategia de identificación parece o no ser válida.

3. Como alternativa al control sintético, otra compañera les hace notar que ustedes cuentan con varias unidades que son tratadas en momentos diferentes del tiempo, lo que permitiría emplear una metodología de estudio de eventos para recuperar un estimado del efecto promedio de tratamiento. Ustedes deciden atender esta sugerencia. Estimen un modelo de estudio de eventos dinámicos[2] por MCO[3] que les permita recuperar el efecto de la ley de cuotas sobre la mortalidad materna. Reporten sus resultados en forma de una gráfica en la que muestren los coeficientes de los leads y lags. ¿Qué pueden decir sobre la validez de esta estrategia?

4. Recientemente, una alternativa a los dos enfoques anteriores, conocida como diferencias en diferencias sintéticas (*SDiD*, por sus siglas en inglés) ha venido tomando fuerza. Supongan que se cuenta con un panel de N unidades y T periodos. Sea Y_{it} el outcome de interés, α_i el efecto fijo de la unidad i, β_t el efecto fijo de periodo, y D_{it} la indicadora de si la unidad i es tratada en el periodo t. En el modelo de TWFE básico, se resuelve el problema

$$\min_{\tau,\mu,\alpha,\beta} \sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau)^2$$

Donde τ captura el efecto causal de interes. En cambio, el estimador de control sintético se obtiene resolviendo

$$\min_{\tau,\mu,\beta} \sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \mu - \beta_t - D_{it}\tau)^2\hat{w}_i^{CS}$$

Donde $\hat{w}_i^{CS}$ son pesos óptimos estimados. La metodología de SDiD propone resolver el problema alternativo:

$$\min_{\tau,\mu,\alpha,\beta} \sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau)^2\hat{w}_i^{sdid}\hat{\lambda}_t^{sdid}$$

En este caso, los pesos de unidad $\hat{w}_i^{sdid}$ están diseñados para que el outcome promedio de las unidades tratadas sea aproximadamente paralelo a un promedio ponderado de los outcomes de las unidades de control. Por su parte, los pesos temporales $\hat{\lambda}_t^{sdid}$ están diseñados para que el outcome promedio post-tratamiento de cada unidad de control difiera por solo una constante del promedio ponderado de los outcome pre-tratamiento de la misma unidad de control. En últimas, estos pesos temporales permiten eliminar o reducir el impacto de periodos de tiempo que son muy diferentes de los periodos post-tratamiento.

   a) Demuestren que los estimadores de control sintético y diferencias en diferencias son un caso particular (i.e. iguales bajo alguna restricción) del estimador de diferencias en diferencias sintéticas. Interpreten.
   b) Por medio del comando *sdid* de Stata, obtengan el estimador de diferencias en diferencias sintéticas del efecto promedio de la ley de cuotas sobre la mortalidad materna. Construya, para cada cohorte de tratamiento, un gráfico en el que muestre la serie de la mortalidad materna promedio observada y el control sintético construido para dicha cohorte (*Pista:* exploren la opción *graph*). Interpreten.

[2] Esto es, incluyan los leads y lags de la variable de tratamiento.
[3] Para propósitos del ejercicio, ignore los problemas de esta metodología.