Conjunto de problemas 1: Predicción de ingresos

MECA 4107

Fecha de entrega: 26 de junio a las 23:59 en

el Bloque Ñ 10 de junio de 2022

1 Introducción

En el sector público, la declaración exacta de los ingresos individuales es fundamental para calcular los impuestos. Sin embargo, el fraude fiscal de todo tipo siempre ha sido un problema importante. Según el Servicio de Impuestos Internos (IRS), alrededor del 83,6% de los impuestos se pagan voluntariamente y a tiempo en los Estados Unidos.[1]. Una de las causas de este desfase es la infradeclaración de ingresos por parte de los particulares. Un modelo de predicción de ingresos podría ayudar a detectar casos de fraude que podrían conducir a la reducción de la brecha. Además, un modelo de predicción de ingresos puede ayudar a identificar a personas y familias vulnerables que puedan necesitar más ayuda.

El objetivo del conjunto de problemas es aplicar los conceptos que hemos aprendido utilizando datos del mundo "real". Para ello, vamos a raspar de la siguiente página web: https://ignaciomsarmiento. github.io/GEIH2018 sample/. Este sitio web contiene datos de Bogotá de la GEIH de 2018.

1.1 Instrucciones generales

El objetivo principal es construir un modelo predictivo de la renta individual

$$Renta = f(X) + u \tag{1}$$

Donde $Renta$ es el ingreso que recibe un individuo, y X es una matriz que incluye potenciales predictores. En este conjunto de problemas, nos centraremos en $f(X) = X\beta$

1. *Adquisición de datos*
   (a) Raspe los datos que están disponibles en el siguiente sitio web https://ignaciomsarmiento. github.io/GEIH2018 sample/.
   (b) ¿Existe alguna restricción para acceder a estos datos?
   (c) Utilizando un pseudocódigo describa su proceso de adquisición de datos

[1]Ver https://www.irs.gov/newsroom/the-tax-gap

2. *Limpieza de datos*. En este conjunto de problemas, nos centraremos únicamente en las personas empleadas mayores de dieciocho (18) años que trabajan en EE.UU. En esta sección, se va a centrar en la limpieza y descripción de los datos.

    - El conjunto de datos incluye múltiples variables que pueden ayudar a explicar la renta individual. Guiado por su intuición y conocimientos económicos, elija las más relevantes y realice un análisis descriptivo de estas variables. Por ejemplo, puede incluir variables que midan la educación y la experiencia, dadas las implicaciones del modelo de acumulación de capital humano (Becker, 1962, 1964; y Mincer (1962, 1975).
    - Tenga en cuenta que hay muchas observaciones con datos que faltan. Le dejo a usted la tarea de encontrar una manera de manejar estos datos faltantes. En su discusión, describa los pasos que realizó para limpiar los datos y justifique sus decisiones.
    - Como mínimo, debes incluir una tabla de estadísticas descriptivas, pero espero que haya tablas y figuras. Aproveche esta sección para presentar una narración convincente que justifique y defienda sus elecciones de datos. Utilice sus conocimientos profesionales para añadir valor a esta sección. No la presente como una lista "seca" de ingredientes.

3. *Perfil edad-ganancia*. Muchos datos de la economía laboral sugieren que el perfil edad-ganancia del trabajador típico tiene una trayectoria predecible: Los salarios tienden a ser bajos cuando el trabajador es joven; aumentan a medida que el trabajador envejece, alcanzando un máximo alrededor de los 50 años; y la tasa salarial tiende a permanecer estable o a disminuir ligeramente después de los 50 años.

    - En el conjunto de datos, múltiples variables describen los ingresos. Elija la que considere más representativa de los ingresos totales de los trabajadores, justificando su selección.
    - Sobre la base de esta estimación utilizando OLS la ecuación del perfil de edad-ingresos:

$$\textit{Renta} = \beta_1 + \beta_2 \textit{Edad} + \beta_3 \textit{Edad}^2 + u \qquad (2)$$

    - ¿Cómo de bueno es este modelo en el ajuste de la muestra?
    - Trace el perfil de edad-ganancia previsto que implica la ecuación anterior.
    - ¿Cuál es la "edad máxima" que sugiere la ecuación anterior? Utiliza el bootstrap para calcular los errores estándar y construir los intervalos de confianza.

4. *El GAP de ganancias*. La mayoría de los estudios económicos empíricos se interesan por un único parámetro de baja dimensión, pero la determinación de ese parámetro puede requerir la estimación de parámetros adicionales "molestos" para estimar este coeficiente de forma coherente y evitar el sesgo de las variables omitidas. Los responsables políticos llevan mucho tiempo preocupados por la brecha salarial entre hombres y mujeres.

- Estimar la diferencia de ingresos incondicional

$$log(Ingresos) = \beta_1 + \beta_2 Mujer + u \quad (3)$$

- ¿Cómo debemos interpretar el coeficiente β_2 ? ¿Cómo de bueno es este modelo en el ajuste de la muestra?
- Estimar y trazar el perfil de edad-ingresos previsto por género. ¿Tienen los hombres y las mujeres de [illegible] el mismo intercepto y las mismas pendientes?
- ¿Cuál es la "edad máxima" implícita por género? Utiliza el bootstrap para calcular los errores estándar y construir los intervalos de confianza. ¿Se solapan estos intervalos de confianza?
- *¿Igual salario por igual trabajo?* Un lema habitual es "a igual trabajo, igual salario". Una forma de interpretarlo es que, para empleados con características similares de trabajador y de trabajo, no debería existir ninguna brecha salarial de género. Estime una brecha salarial condicional que incorpore variables de control como características similares del trabajador y del puesto de trabajo (X).
  (a) Estimar la brecha condicional de ingresos $log(Income) = \beta_1 + \beta_2 Mujer + \theta X + u$
  (b) Utilice el FWL para repetir la estimación anterior, en la que el interés se encuentra en β_2 . ¿Obtiene las mismas estimaciones?
  (c) ¿Cómo debemos interpretar el coeficiente β_2 ? ¿Cómo de bueno es este modelo en el ajuste de la muestra? ¿Se reduce la brecha? ¿Es esto una prueba de que la brecha es un problema de selección y no un "problema de discriminación"?

5. *Predicción de los beneficios*. Ahora pasamos a la predicción. En la sección anterior has construido un par de modelos utilizando tus conocimientos como economista aplicado, la tarea aquí es evaluar el poder de predicción de estos modelos.

   (a) Divida la muestra en dos muestras: una de entrenamiento (70%) y otra de prueba (30%). No olvides establecer una semilla (en R, set.seed(10101), donde 10101 es la semilla).
      i. Estime un modelo que sólo incluya una constante. Esta será la referencia.
      ii. Vuelve a estimar tus modelos anteriores
      iii. En las secciones anteriores, los modelos estimados tenían diferentes transformaciones de la variable dependiente. En este punto, explore también otras transformaciones de sus variables independientes. Por ejemplo, puede incluir términos polinómicos de ciertos controles o interacciones de éstos. Pruebe al menos cinco (5) modelos que vayan aumentando en complejidad.
      iv. Informe y compare el error medio de predicción de todos los modelos que ha estimado anteriormente. Discute el modelo con el menor error medio de predicción.
      v. Para el modelo con el menor error medio de predicción, calcule la estadística de apalancamiento para cada observación de la muestra de prueba. ¿Existen valores atípicos, es decir, observaciones con un alto apalancamiento que impulsan los resultados? ¿Son estos valores atípicos personas potenciales que la DIAN debería investigar, o son simplemente el

producto de un modelo defectuoso?

(b) Repita el punto anterior pero utilice la validación cruzada K-fold. Comente las similitudes/diferencias de utilizar este enfoque.

(c) *LOOCV*. Con su modelo de predicción preferido (el que tenga el menor error medio de predicción) realice el siguiente ejercicio:

i. Escribe un bucle que haga lo siguiente:
   - Estimar el modelo de regresión utilizando todas las observaciones menos la i*-ésima*.
   - Calcule el error de predicción para la i*-ésima* observación, es decir, ($y_i - \hat{y}$)
   - Calcule la media de los números obtenidos en el paso anterior para obtener el error cuadrático medio medio. Esto se conoce como el estadístico Leave-One-Out Cross-Validation (LOOCV).

ii. Compare los resultados con los obtenidos en el cálculo de la estadística de apalancamiento

2 Directrices adicionales

Espero las siguientes cosas del conjunto de problemas, la omisión de cualquiera de estas pautas será penalizada.

- Debe entregar su documento en bloque neon.
- El documento debe apuntar e incluir un enlace a su repositorio de GitHub.
- Debe seguir la plantilla del repositorio.
- El repositorio debería incluir un archivo README. Un buen LÉEME ayuda a que su proyecto se destaque de otros proyectos y es el primer archivo que una persona ve cuando se encuentra con su repositorio. Por lo tanto, este archivo debe ser lo suficientemente detallado como para centrarse en tu proyecto y en cómo lo hace, pero no tan largo como para perder la atención del lector. Por ejemplo, Project Awesome tiene una lista curada de READMEs interesantes.
- El repositorio debe tener al menos cinco (5) contribuciones sustanciales de cada miembro del equipo.
- Las tablas, las figuras y la redacción deben ser lo más ordenadas posible. Etiquete todas las variables que incluya. Toda variable, estadística, etc., incluida en las figuras o tablas debe describirse en el texto.
- Tu código debe ser legible e incluir comentarios. En la codificación, como en la escritura, un buen estilo de codificación es fundamental para que el código sea legible. Te animo a seguir la guía de estilo de Tidyverse.