Universidad de Los Andes

Big Data and Machine Learning for Applied Economics

MECA 4107

Junio 26,2022

Bogotá D.C, Colombia

**PROBLEM SET 1: PREDICTING INCOME**

Maria Valeria Gaona Guevara

Código: 202214418

Correo: [mv.gaona@uniandes.edu.co](mailto:mv.gaona@uniandes.edu.co)

Andrea Margarita Beleño Hernández

Código: 200620739

Correo: [a.beleno@uniandes.edu.co](mailto:a.beleno@uniandes.edu.co)

**Resumen:** El siguiente documento presenta XXXXXXXX. El repositorio GitHub del siguiente documento se encuentra en el siguiente link:

1. **Información General**
   1. *Adquisición de datos.*

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

*Ingreso= f(*X) *+ u*

* + 1. Extraer los datos que se encuentran disponibles en el siguiente sitio web [https://ignaciomsarmiento.github.io/GEIH2018 muestra/](https://ignaciomsarmiento.github.io/GEIH2018%20muestra/).
    2. ¿Existen restricciones para acceder o extraer estos datos?

Para acceder a los datos es necesario esperar un momento, ya que los datos no se cargan automáticamente. De acuerdo con lo anterior, si existen restricciones para extraer los datos debido al inconveniente mencionado anteriormente. Por lo tanto, cuando se corre el código de extracción de datos en R, toma la página web como vacía y no carga la base de datos.

* + 1. Usando pseudocódigo, describa su proceso de adquisición de datos
  1. *Limpieza de datos*
  2. *Perfil edad-ingresos*

Con base en esta estimación utilizando OLS, la ecuación del perfil de edad-ingresos:

*Ingreso* = β1 + β2Edad + β3Edad^2 + u (2)

* + 1. ¿Qué tan bueno es este modelo en el ajuste de la muestra?
    2. Grafique el perfil predicho de ingresos por edad implícito en la ecuación anterior.
    3. ¿Cuál es la “edad pico” sugerida por la ecuación anterior? Usar bootstrap para calcular los errores estándar y construir los intervalos de confianza.
  1. *La brecha de ganancias*

Estimar la brecha de ingresos incondicional:

*log(ingreso)* = β1 + β2Femenino+ u

* + 1. ¿Cómo debemos interpretar el coeficiente β2? ¿Qué tan bueno es este modelo en samplefit?
    2. Estimar y trazar el perfil de edad-ingresos pronosticado por género. ¿Los hombres y las mujeres en Bogotá tienen la misma intersección y pendientes?
    3. ¿Cuál es la “edad pico” implícita por género?. Utilice bootstrap para calcular los errores estándar y construir los intervalos de confianza. ¿Se superponen estos intervalos de confianza?
    4. Estime una brecha de ingresos condicional que incorpore variables de control tales como características similares del trabajador y del puesto (X).
       1. (a) Estime la brecha de ingresos condicional

log(*Ingreso*) = β1 + β2Femenino +θX + u

* + - 1. Use FWL para repetir la estimación anterior, donde el interés radica en β2. ¿Obtiene las mismas estimaciones?
      2. ¿Cómo debemos interpretar el coeficiente β2? ¿Qué tan bueno es este modelo en ajuste de muestra? ¿Se reduce la brecha? ¿Es esta evidencia de que la brecha es un problema de selección y no un "problema de discriminación"?
  1. Predicción de ganancias.
     1. Divida la muestra en dos muestras: una muestra de entrenamiento (70%) y una muestra de prueba (30%). No olvide establecer una semilla (en R, set.seed(10101), donde 10101 es la semilla).
        1. Estime un modelo que solo incluya una constante. Este será el punto de referencia.
        2. Estime nuevamente sus modelos anteriores
        3. En las secciones anteriores, los modelos estimados tenían diferentes transformaciones de la variable dependiente. En este punto, explora también otras transformaciones de tus variables independientes. Por ejemplo, puede incluir términos polinómicos de ciertos controles o interacciones de estos. Pruebe al menos cinco (5) modelos que aumentan en complejidad.
        4. Informe y compare el error de predicción promedio de todos los modelos que estimó anteriormente. Discuta el modelo con el error de predicción promedio más bajo.
        5. Para el modelo con el error de predicción promedio más bajo, calcule la estadística de apalancamiento para cada observación en la muestra de prueba. ¿Hay valores atípicos, es decir, observaciones con un alto apalancamiento que impulsen los resultados? ¿Son estos valores atípicos personas potenciales que la DIAN debería investigar, o son simplemente el producto de un modelo defectuoso?
     2. Repita el punto anterior pero use la validación cruzada K-fold. Comente las similitudes/diferencias del uso de este enfoque.
     3. Con su modelo predicho preferido (el que tiene el promedio más bajo error de predicción) realice el siguiente ejercicio:

Escribe un bucle que haga lo siguiente:

* + - 1. Estimar el modelo de regresión utilizando todas las observaciones menos la i − ésima.
      2. Calcular el error de predicción para la i − ésima observación, es decir, (yi− yˆi)
      3. Calcular el promedio de los números obtenidos en el paso anterior para obtener el error cuadrático medio. Esto se conoce como la estadística de validación cruzada Leave-One-Out (LOOCV).
      4. Compare los resultados con los obtenidos en el cálculo de la estadística de apalancamiento