calibración

Muestreo II

Licenciatura en Estadística

2023

sobre la presentación

- Si bien existe una clara diferencia entre estimador (variable aleatoria) y estimación (realización del estimador) vamos a utilizar ambos términos de forma intercambiable.

sobre la presentación

- Presentación teórica en conjunto con ejemplos prácticos.
  - Orientados a encuestas de hogares y personas.
- Los ejemplos prácticos se realizarán en el software estadístico R

introducción

Información que se encuentra disponible para la población objetivo de la encuesta, ya sea:

1. A nivel de cada una de las unidades (x_i), i.e. se encuentra contenida en el marco muestral F
2. A nivel agregado ($\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$), i.e. información que proviene de registros administrativos (RA), censos recientes, encuestas de calidad, etc.

el uso de a información auxiliar

La información auxiliar puede utilizarse:

1. **Diseño muestral**
   - Construir estratos
   - Asignar distintas tasas de muestreo en los estratos
   - Definir probabilidad de selección distintas (e.g. $x_i = \text{MOS}_i$)
   - Muy restrictivo, la información auxiliar debe estar contenida en el marco muestral.
2. **Etapa de estimación**
   - Definir nuevos ponderadores (w_i^*) que sean congruentes con información (variables de control) conocida acerca de la estructura de la población, i.e. la muestra "expandida" coincida con información conocida de la población

información auxiliar en la etapa de estimación

Los requisitos de la información auxiliar en la etapa de estimación son más flexibles

- Conocer simplemente los totales de las variables de control.
  - e.g el total de personas por tramo de edad y sexo.
- Las variables de control deben ser conocida únicamente para los individuos de la muestra.
  - i.e. estar incluida en el formulario de la encuesta o provenir del marco muestral

la idea

- la idea de mantener cerca los ponderadores w_i^* de w_i es para **“pedir prestada”** cualquiera propiedad buena de estimación que los ponderadores w_i tengan.
- si los ponderadores $w_i = 1/\pi_i$ producen estimadores insesgados, y los ponderadores w_i^* se encuentran cerca, producirán estimadores **aproximadamente insesgados**.

calibración

el problema es encontrar un nuevo juego/sistema de ponderadores $w_i^* = g_i \times w_i$ que

1. minimicen una medida de distancia $L(w^*, w)$

$$\text{cambio ponderadores} = \sum_{i \in s} L(w_i^*, w_i)$$

2. y que cumplan la ecuación de calibración

$$\sum_{i \in s} w_i^* \mathbf{x}_i^T = \sum_{i \in U} \mathbf{x}_i^T$$

donde $\mathbf{x}_i^T = (x_{1i}, x_{2i}, ..., x_{Ji})^T$ es el set de variables auxiliares

calibración

una elección de L es la distancia de mínimos cuadrados

$$L(w^*, w) = \sum_{i \in s} (w_i^* - w_i)^2 / w_i$$

minimizando lo anterior sujeto a las ecuaciones de calibración se obtiene el estimador de regresión (GREG)

calibración

Otra elección de L es

$$L(w^*, w) = \sum_{i \in s} (w_i \log(w_i^*/w_i)^* - w_i^* - w_i)$$

de esta forma se obtiene el **estimador raking**.

Potenciales beneficios de la calibración

- **Reducción de los SE de las estimaciones**
  - Si las variables auxiliares de alguna forma explican la variabilidad de interés, es decir, se encuentran correlacionadas.
- **Posible reducción del sesgo por problemas de cobertura.**
- **Reducción del Sesgo ocasionado por la no respuesta (NR)**
  - Si las variables explican de alguna forma la probabilidad/propensión de responder de una unidad (e.g. hogar, persona)
- **Comparabilidad y "estética"**: mismas estimaciones cruzando con otras fuentes. Mejora de la "credibilidad" para los usuarios que no están familiarizados en técnicas de muestreo

regression thinking

El mismo problema de siempre:

- El objetivo estimar el total

$$Y = \sum_{i \in U} y_i$$

- Seleccionamos una muestra aleatoria (s) bajo un diseño muestral cualquiera.
- Los valores de y solo son conocidos para los elementos que encuestamos
- Una vez finalizado la recolección de los datos, estimamos

$$\hat{Y} = \sum_{i \in s} w_i \times y_i$$

regression thinking

Imaginemos que conocemos o podemos construir una variable proxy de y la cual denotamos $\hat{y}$.

Al total lo podemos escribir como:

$$Y = \sum_{i\in U} \hat{y}_i + \sum_{i\in U} y_i - \sum_{i\in U} \hat{y}_i = \sum_{i\in U} \hat{y}_i + \sum_{i\in U} (y_i - \hat{y}_i)$$

donde $\sum_{i\in U} (y_i - \hat{y}_i)$ es desconocido y lo tenemos que estimar

regression thinking

Decisiones a tomar:

1. cómo elegimos los valores proxy $\hat{y}$?
    - Una opción: por medio de una regresión lineal

$$E_m(y_i) = \beta_0 + \beta_1 x_{1i} + ... + \beta_J x_{Ji} = \mathbf{x}_i^T \boldsymbol{\beta}$$

    donde las variables x es información auxiliar disponible acerca de los individuos

2. cómo estimamos $\sum_{i \in U} (y_i - \hat{y}_i)$?
    - Una opción: utilizando los ponderadores $w_i = 1/\pi_i$

$$\sum_{i \in s} w_i \times (y_i - \hat{y}_i)$$

regression thinking

Si se utiliza un modelo de regresión para la construcción de los valores $\hat{y}_i$y se utilizan los ponderadores originales w_i construye el estimador de regresión (GREG)

$$\hat{Y}^{\text{GREG}} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} w_i \times e_i$$

donde $e_i = (y_i - \hat{y}_i)$ son los errores estimados del modelo en la muestra.

regression thinking

Un estimador de la error estándar utilizando Taylor es:

$$\widehat{\mathrm{SE}}^2(\hat{Y}^{\mathrm{GREG}}) = \widehat{\mathrm{var}}(\hat{Y}^{\mathrm{GREG}}) = \widehat{\mathrm{var}}(\sum_{i \in s} w_i \times e_i)$$

Si la información auxiliar utilizada para la construcción del estimador está correlacionadas con la varia y, es decir, el modelo tiene un buen poder de ajuste, el SE de $\hat{Y}^{\mathrm{GREG}}$ sera pequeña en comparación con el estimador HT

regression thinking

$$\hat{Y}^{\text{GREG}} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} w_i \times e_i$$

- Si se utiliza únicamente el primer término estamos utilizando un estimador basado en diseño. En SAE este estimador es denominado estimador sintético
- El segundo término protege al estimador si el modelo no es correcto.Si el modelo no ajusta bien, el estimador únicamente tendrá mayor SE.

regression thinking

El estimador de regresión lineal lo podemos escribir como una suma ponderada

$$\hat{Y}^{\text{GREG}} = \sum_{i \in s} w_i^* \times y_i$$

donde $w_i^* = g_i \times w_i$ con

$$g_i = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{i \in s} w_i \times \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{x}_i$$

regression thinking

Los ponderadores w_i^* cumplen con las ecuaciones de calibración

$$\sum_{i \in s} w_i^* \mathbf{x}_i^T = \sum_{i \in U} \mathbf{x}_i^T$$

Observación: si bien los estimadores de regresión son calibrados, su enfoque es únicamente reducir la varianza de las estimaciones por medio de un modelo, es decir, realizar predicciones $\hat{y}_i$.

enfoque de regresión (regression thinking)

conclusiones:

1. Si las variables auxiliares utilizadas para la construcción del modelo explican las variables de interés de la encuesta se reducirán los SE de las estimaciones.
2. Genera ponderadores calibrados que cumplen con las ecuaciones de calibración