MINISTERIO
DE ECONOMÍA, COMERCIO
Y EMPRESA

INE
Instituto Nacional de Estadística

Oposición al Cuerpo Superior de Estadísticos de Estado

Cuarto Ejercicio

Convocatoria de la oferta pública de empleo de 2023

Resolución de 14 de diciembre de 2023, de la Subsecretaría, por la que se convocan procesos selectivos para ingreso, por el sistema de acceso libre y promoción interna, en el Cuerpo Superior de Estadísticos del Estado. (BOE 27 de Diciembre de 2023)

Especialidad I: Estadística-Ciencia de Datos

Cuestión 1. Sea un municipio con un total de N =8.000 alumnos distribuidos en $M = 16$ centros escolares de tamaños desiguales $(N_i, \forall i = 1 \ldots, 16)$. Se desea estimar la proporción P de estudiantes que practican algún deporte fuera del horario escolar para ajustar los planes deportivos municipales.

Para alcanzar estos objetivos, se utiliza un muestreo de conglomerados con submuestreo con probabilidades iguales y sin reposición en ambas etapas. En la primera etapa, se seleccionan $m = 4$ centros, y en la segunda etapa, se seleccionan los alumnos con una fracción de muestreo de $f_2 = 0,10$.

La información disponible sobre los centros muestreados es la siguiente:

Centro i	N_i	$\hat{P}_i$	$N_i\hat{P}_i$	$N_i\hat{P}_i(1-\hat{P}_i)$
1	700	0,4	280	168
2	500	0,3	150	105
3	400	0,5	200	100
4	900	0,6	540	216
Total	2.500	1,8	1.170	589

donde $\hat{P}_i$ es la estimación de la propoción de alumnos practicando algún deporte fuera del horario escolar en cada conglomerado muestreado.

Se pide:

a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

b) Consideramos otro estimador de P, definido por el promedio de los porcentajes observados en los centros: $\hat{P}^* = \frac{1}{4}\sum_{i=1}^{4}\hat{P}_i$

1) Demuestre que: $Sesgo(\hat{P}^*) = -\frac{1}{N}\sum_{i=1}^{M} P_i(N_i - \frac{N}{M}) = -\frac{M}{N}Cov(N_i, P_i)$

2) Deduzca las condiciones para las que $\hat{P}^*$ sería un estimador insesgado de P.

c) Suponga que, para obtener el estimador $\hat{\theta}$ de cierta característica del alumnado, se han aplicado varios procedimientos (reponderación por falta de respuesta, calibración con variables auxiliares, ...). Por esta razón, se ha decidido estimar la varianza utilizando el método indirecto de Jackknife. Indique cómo proceder para calcular la varianza con este método en este contexto, proporcionando las expresiones necesarias para la estimación de la varianza.

Cuestión 2. Para la obtención de ciertas estimaciones en una población de tamaño N, se utiliza un muestreo bifásico. Responda a las cuestiones planteadas para cada caso:

a) Para estimar la media de una característica $\bar{Y}$, primero se seleccionan 2.000 elementos para determinar su estrato, mediante un muestreo aleatorio simple sin reemplazamiento. De éstas, corresponden al primer estrato 1.300 y 700 al segundo. En una segunda fase, se lleva a cabo un muestreo aleatorio simple estratificado, seleccionando 150 elementos del primer estrato y 100 en el segundo, a quienes se les solicita medir la característica de interés. Los resultados se muestran en el siguiente cuadro:

Estrato h	$\bar{y}_h$	s^2_{yh}
1	5,4	3,75
2	9,2	6,9

donde $\bar{y}_h$ representa la media muestral en el estrato h de la segunda fase y s^2_{yh} es la varianza muestral del estrato h en la segunda fase. Además, se supone que el tamaño poblacional N es mucho mayor que el tamaño muestral de primera fase.

1) Calcule una estimación insesgada de la media $\bar{Y}$.
2) Estime su varianza.

b) Se desea estimar el total de una característica Y en una población, donde las probabilidades de selección de las unidades son proporcionales a una medida de su tamaño, denotada por M_i. Sin embargo, dado que no se conocen a priori los tamaños de las unidades de la población, se extrae en una primera fase una muestra de tamaño $n^{'}$, con probabilidades iguales y sin reposición, para obtener información acerca de los tamaños $M_1, \ldots, M_i, \ldots, M_{n'}$, siendo $M^{'} = \sum_i^{n^{'}} M_i$. En la segunda fase se obtiene una submuestra de tamaño $n < n^{'}$, con probabilidades proporcionales al tamaño y con reposición. Se pide:

1) Demuestre que $\hat{Y} = \sum_{i=1}^{n} \frac{N}{n} \frac{M^{'}}{n^{'}} \frac{X_i}{M_i}$ es un estimador insesgado del total poblacional Y.
2) Obtenga de forma razonada la varianza de dicho estimador $V(\hat{Y})$.

c) ¿Sería posible utilizar el muestreo bifásico para estimar la media de una característica, $\bar{Y}$, mediante el uso de un estimador de razón, utilizando la información de una variable auxiliar X disponible para todas las unidades muestreadas en primera fase? Exponga de manera concisa cómo se llevaría a cabo cada fase para obtener el estimador de razón.

Cuestión 3. Las siguientes estimaciones se calcularon a partir de una muestra de 7.634 mujeres encuestadas en una Encuesta General de Hogares. La variable dependiente toma el valor 1 si la mujer tiene un empleo remunerado, y 0 en caso contrario.

	MCO	Logit	Probit
ALTO	0.093	0.423	0.259
	(0.015)	(0.071)	(0.043)
NOCUAL	–0.210	–0.898	–0.554
	(0.013)	(0.056)	(0.035)
EDAD	0.038	0.173	0.107
	(0.003)	(0.124)	(0.008)
EDAD2	–0.051	–0.230	–0.142
	(0.003)	(0.069)	(0.009)
CAS	0,024	0.103	0.063
	(0.009)	(0.057)	(0.035)
Constante	-0.068	-2.587	-1.593
	(0.049)	(0.225)	(0.137)

Donde ALTO es uno si el encuestado tiene una cualificación educativa superior, cero en caso contrario; NOCUAL es uno si el encuestado no tiene cualificaciones, cero en caso contrario; EDAD es la edad en años; EDAD2 es (edad x edad)/100; CAS es uno si está casado, cero en caso contrario.

Los errores estándar calculados de forma convencional se indican entre paréntesis para los resultados de mínimos cuadrados ordinarios (MCO) y los errores estándar asintóticos entre paréntesis en los demás casos.

a) En el contexto de los modelos GLM, indique qué papel desempeña la función de enlace y cuáles son las funciones de enlace teóricas para cada uno de los modelos estimados, es decir, para MCO, Logit, Probit.

b) Explique brevemente cómo se calculan las estimaciones Probit cuando el modelo no tiene intercepto y sólo tiene una variable explicativa. Indique qué propiedades generales tienen las correspondientes estimaciones.

c) Para los tres conjuntos de estimaciones, pruebe la hipótesis nula de que el coeficiente de CAS es cero. ¿Qué estadístico de contraste consideraría más fiable? Explíquelo.

d) Utilizando las estimaciones MCO y Probit, calcule las probabilidades estimadas de estar desempleada para una mujer casada de 40 años con estudios superiores.

e) Compruebe la hipótesis de significación conjunta del modelo probit, sabiendo que:

$$lnL_R = -416{,}01$$

$$lnL_{NR} = -321{,}25$$

donde lnL_R y lnL_{NR} son los logaritmos de la verosimilitud de los modelos probit restringido y no restringido, respectivamente. Utilice un nivel de significatividad del 5 por ciento.

Cuestión 4. Para promover la vida social y turística de un municipio, determinada Asociación Cultural y Turística ofrece desde 2023 un programa de excursiones anuales para sus miembros de dos tipos, unas de carácter nacional y otras de carácter internacional.

La implementación de este programa se realiza siguiendo las pautas que se establecen a continuación:

- Cada año, los asociados que aún no participan en el programa de excursiones tienen una probabilidad de 0,4 de unirse a las excursiones nacionales y de 0,3 a las internacionales, el año siguiente. Si no se integran al programa, seguirán sin participar en él.
- El 60 % de los asociados que han participado en excursiones nacionales un determinado año continuarán el próximo año participando en las mismas, mientras que el 70 % de los que realizaron las excursiones internacionales repetirá en éstas. El resto de participantes, cambiará al otro tipo de excursión, manteniéndose para siempre en el programa de la asociación.

Considerando que en 2023 ningún miembro participaba en estas excursiones programadas, se pide:

a) Modelar el proceso estocástico asociado utilizando una cadena de Markov, describiendo sus principales componentes y propiedades.

b) Establecer la proporción de miembros de la asociación que durante el año 2025 disfrutarán de cada uno de los tipos de excursiones.

c) Determinar el comportamiento a largo plazo de la proporción de miembros que realizará cada uno de los tipos de excursiones. Asimismo, calcular el tiempo medio que tarda en estar en el programa un asociado que no estaba en él de partida.

Cuestión 5. Suponga que le piden implementar un código que incluya los algoritmos Lasso y random forest para un caso de clasificación. Escriba los pasos a seguir, preferiblemente en código Python, R o pseudocódigo, para llevar a cabo el proyecto completo. Las características del proyecto son las siguientes:

- Objetivo: Utilizar las características del empleado/a para predecir si abandonarán la compañía. Datos disponibles en el archivo "abandono.csv"
- Variable target: Attrition (valores " Yes" , "No")
- Features (covariables): 30 variables (var1 a var30)
- Número de observaciones: 1470

En el código debe tener en cuenta lo siguiente:

a) Teniendo en cuenta que la prevalencia de abandono general está en torno al 16 %, ¿qué métricas de clasificación considera adecuadas para evaluar el desempeño de los algoritmos?.

b) Entre los features (covariables) hay variables de carácter cuantitativo y otras de carácter cualitativo (tanto nominal como ordinal). Algunas variables de carácter cuantitativo pueden contener outliers, tener fuerte asimetría y, habitualmente, estarán en unidades distintas. ¿Qué transformaciones considera adecuadas en las variables de carácter cuantitativo para lograr que los algoritmos puedan tener mejores resultados? Suponga que las variables cuantitativas son las 15 primeras (var1 a var15).

c) En las variables de carácter nominal (var16 a var25), ¿es necesario crear dummies por categorías en ambos algoritmos?. Justifique su respuesta.

d) Suponga que le informan de que algunas de las categorías de las variables cualitativas var17 y var18 tienen muy pocas observaciones. ¿Qué efectos tendría esta situación sobre el entrenamiento y validación de algoritmos y cómo podríamos resolverlo?

e) ¿Cómo se pueden codificar las variables ordinales var26 a var30 en ambos algoritmos?

f) Su supervisora le dice que debe realizarse un proceso de tuneado sobre los hiperparámetros de cada algoritmo. ¿Por qué cree que le pide esto?. Describa detalladamente los pasos para llevar a cabo este proceso teniendo en cuenta que se debe utilizar validación cruzada con la muestra de training. Tenga en cuenta, además, que todos los resultados deben ser reproducibles.

g) ¿Cómo evaluaría la posible existencia de overfitting en los algoritmos planteados?

Cuestión 6. Considere las siguientes relaciones correspondientes a una base de datos de una determinada universidad:

- Student(snum: integer, sname: string, major: string, level: string, age: integer)
- Class(name: string, meets at: string, room: string, fid: integer)
- Enrolled(snum: integer, cname: string)
- Faculty(fid: integer, fname: string, deptid: integer)

La relación Enrolled tiene un registro por cada par estudiante-clase tal que el estudiante está inscrito en la clase. El atributo major en la relación Student se refiere al grado que cursa el estudiante mientras que el atributo fname en la relación Faculty se refiere al nombre del profesor que imparte el curso. No se deben imprimir duplicados en ninguna de las respuestas correspondientes a consultas SQL.

a) Crear las tablas mediante comandos SQL incluyendo las claves y las referencias de integridad necesarias.

b) Escribir código SQL para imprimir los nombres de todos los estudiantes de level = 'JR' que están inscritos en una clase impartida por 'I. Teach'

c) Escribir código SQL para imprimir la edad del estudiante de más edad que esté inscrito en 'Historia' o en un curso impartido por 'I. Teach'

d) Escribir código SQL para imprimir los nombres de los estudiantes inscritos en el mayor número de clases.

Cuestión 7. Ciertos mensajes llegan a una instalación de comunicaciones de acuerdo con un proceso Poisson a razón de 2 por hora. La instalación consta de tres canales, cada mensaje llega a un canal libre, si los tres están libres, o se pierde si todos los canales están ocupados. El tiempo que los mensajes permanecen en un canal es una variable aleatoria que depende de las condiciones meteorológicas al momento de llegada. Si el mensaje llega cuando las condiciones son "buenas", su tiempo de procesamiento es una variable aleatoria con función de distribución:

$$F(x) = x, \quad 0 < x < 1$$

Mientras que si las condiciones son "malas", su tiempo de procesamiento tiene la función de distribución:

$$F(x) = x^3, \quad 0 < x < 1$$

Al principio las condiciones son buenas y, posteriormente, se van alternando en períodos buenos y malos: los buenos tienen una duración fija de 2 horas y los malos de una hora.

a) Realice un esbozo de cuáles serían los pasos para simular la distribución del número de mensajes perdidos hasta el instante $T = 100$.

b) Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python ó R), que permita obtener, mediante simulación, la distribución del número de mensajes perdidos hasta el instante T = 100. Se puede utilizar una función random() que genere valores con distribución $U(0, 1)$.

Cuestión 8. Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python ó R), de un algoritmo para calcular la media recortada (trimmed mean). El algoritmo debe aceptar como parámetros de entrada una secuencia de números y el parámetro alpha de la media recortada. No se puede utilizar ninguna función tipo sort predefinida sino que la ordenación debe estar incluida en el algoritmo y ser eficiente. Justificar la complejidad algorítmica de todos los pasos del procedimiento descrito.