Tarea 1

Fecha de entrega: Miércoles 3 de abril.

Nota: Deberán subir a *Canvas* **dos archivos**. Un **primer archivo** de texto con sus respuestas. Este archivo debe ser auto-suficiente. Esto quiere decir que con revisar este archivo debe ser posible calificar su tarea en su totalidad. Para este archivo pueden utilizar el formato de su preferencia (e.g. pdf, LaTeX, Word u hojas escritas a mano y escaneadas). El **segundo archivo** es un archivo de soporte que genere todos sus resultados. Típicamente este será un archivo de programación, mismo que al ejecutarse replicaría los resultados que estás reportando en tu primer archivo. Dado que ustedes son libres de elegir el software que utilizarán, este puede ser un R-script, Do-File o similar.

ENIGH Salud Alimentaria

El gobierno de México, en su compromiso por abordar los desafíos nutricionales y alimenticios que enfrenta el país, ha decidido emplear la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2020 como herramienta clave para profundizar en el análisis de estas problemáticas, con un enfoque particular en las disparidades entre estratos sociales y entre zonas urbanas y rurales. En este contexto, se ha convocado a la colaboración de tu equipo en Salud Integral Consulting para realizar este análisis.

El objetivo primordial es examinar cómo diversos factores socioeconómicos, como el ingreso corriente, la ubicación geográfica y el gasto en alimentos, están relacionados con la salud alimentaria de los hogares mexicanos. La ENIGH 2020 recaba información sobre los gastos en alimentación para una muestra representativa a nivel nacional de 89,006 hogares.[1] Los gastos que incluye la base de datos son trimestrales y se recopilaron del 21 de agosto al 28 de noviembre de 2020. En Canvas encontrarás la base de datos $ENIGH\_alimentos2020$ y al final de este documento se encuentra el diccionario resumido de las variables necesarias para este análisis.

1. **[Introducción]** Como primer paso en tu análisis, debes proporcionar un panorama general de las variables recopiladas por la ENIGH 2020. Esta revisión inicial te ayudará a formar una comprensión básica de los distintos factores que podrían influir en la seguridad alimentaria en los hogares mexicanos, estableciendo una sólida base para tu reporte.

   (a) La base de datos incluye a las variables *gasto_mon* que representa el gasto corriente monetario total, y *alimentos* que representa el gasto total en alimentos

[1] La base de datos que tienen disponible para está tarea es la verdadera base de datos ENIGH disponible en la página de INEGI.

dentro del hogar. Al gobierno le interesa conocer la proporción de gasto que los hogares mexicanos dirigen hacia la alimentación. Crea la variable *prop_alimen* que represente la proporción mencionada. Haz una gráfica que muestre la distribución de la variable que creaste para cada estrato socio-económico (capturado en la variable *est_socio*) y proporciona una breve descripción de los hallazgos.

(b) Al representante del gobierno que te contrato también le interesa conocer el gasto dirigido a la alimentación per cápita. Utiliza las variables *alimentos* y *tot_integ* Crea una variable adicional *alimentos_pc* que capture esta información e igual haz una gráfica que muestre la distribución de esta variable para cada estrato socio-económico. Proporciona una breve descripción de los hallazgos. [NOTA: notarás que la variable presenta una distribución sesgada con colas largas hacia valores más altos. La transformación logarítmica ayuda a normalizar estos datos y reduce la influencia de los valores atípicos]

(c) Se te solicita crear una tabla con estadísticas descriptivas básicas con todas las variables disponibles en la base de datos [Nota: solo no debes incluir las variables $folioviv$ y *ubica_geo*]. Sí debes de incluir todas las otras variables de la base de datos y, adicionalmente, las variables de interés que creaste en inciso 1(a) y 1(b). La tabla debe incluir el número de observaciones, la media, la desviación estándar, los valores mínimos y máximos para cada variable. Si alguna de las variables es de tipo dummy, señalalo de alguna forma en la tabla. Tal vez deberás transformar algunas variables para realizar el análisis descriptivo de forma que tenga sentido. [R tip: Para hacer la tabla puedes usar el comando *stargazer*].

2. **[Estrato socioeconómico]** De acuerdo con observaciones preliminares y estudios socioeconómicos previos, existe la intuición de que los estratos socioeconómicos más bajos podrían destinar una mayor proporción de su gasto total a la alimentación, reflejando potencialmente diferencias en calidad, variedad o preferencias alimenticias. Este análisis nos permitirá comprender mejor cómo la posición socioeconómica influye en los patrones de consumo alimentario y si realmente se verifica que los estratos más bajos invierten proporcionalmente más en su alimentación en comparación con los estratos más altos. Modifica tu data frame para quedarte únicamente con las viviendas con ingreso trimestral promedio menor a 1,000,000 de pesos (utiliza esta base de datos para lo que resta de la tarea). NOTA: Para la modificación de tu data frame y para la interpretación de la variable *ing_cor* toma en cuenta que esta variable ha sido escalada dividiéndola por 10,000.

(a) Haz dos gráficas: primero, una que ilustre la relación entre el ingreso corriente y la proporción de gasto en alimentos distinguiendo entre estratos socioeconómicos; segundo, una que ilustre la relación entre el ingreso corriente y el gasto en alimentos por integrante de la vivienda distinguiendo entre estratos socioeconómicos.

[R tip: Puedes usar *color* = *est_socio* en ggplot para incluir la distinción entre estratos]. Describe en dos oraciones una conclusión sobre las gráficas.

(b) Te propones analizar con mayor detalle la relación entre el ingreso corriente y las variables *prop_alimen* y *alimento_pc* por lo que te interesa estimar la siguientes especificaciones:

$$prop\_alimen_i = \beta_0 + \beta_1 log(ing\_cor_i) + U_i$$

$$log(alimento\_pc_i) = \beta_0 + \beta_1 log(ing\_cor_i) + U_i$$

Lleva a cabo estas estimaciones y muéstralas junto con un diagrama de dispersión. Al final de la tarea viene una ilustración para ejemplificar cómo debes reportar tu resultado.

(c) Te propones analizar con mayor detalle la relación entre el estrato socioeconómico y las variables *prop_alimen* y *alimento_pc*. Realiza una regresión donde *prop_alimen_i* sea la variable dependiente y *est_socio* la variable explicativa. Reflexiona sobre el uso adecuado de la variable explicativa y la posible necesidad de agregar variables adicionales para capturar su efecto correctamente. Posteriormente, repite el proceso de análisis utilizando el logaritmo de *alimento_pc* como variable dependiente. Reporta los resultados de tus estimaciones con un formato de ecuación [Nota: al final de la tarea viene un ejemplo de como reportar los resultados con formato de ecuación].

(d) Utiliza los resultados de las preguntas 1(a) y 1(b) para argumentar el uso de errores homocedásticos o heterocedásticos en la regresión del inciso (c). No vuelvas a hacer la regresión. Solo ajusta el resultado del inciso anterior de forma adecuada y como respuesta a esta pregunta incluye la reflexión que llevaste a cabo con las preguntas 1(a) y 1(b) para decidir el tipo de errores a utilizar.

3. **[Análisis]** En esta sección de análisis, profundizaremos en el estudio de las variables que pueden influir en la proporción del gasto destinado a alimentos en los hogares mexicanos y el gasto per cápita en alimentos.

(a) Para analizar la relación entre las características sociodemográficas y el gasto en alimentos , se deberán realizar las estimaciones presentadas en la Tabla 1. En la tabla notarás dummies para cada estrato socioeconómico donde $est\_socio2 = 1$ si el hogar pertenece al estrato socioeconómico 2 y $est\_socio2 = 0$ si el hogar pertenece a cualquier otro estráto socioeconómico. Lo mismo para est_socio3 y est_socio4. Los coeficientes estimados en la tabla deben venir acompañados de asteriscos para indicar su nivel de significancia estadística, de la siguiente manera: * para el 10 %, ** para el 5 % y *** para el 1 %. Deberás emplear errores

heterocedásticos para tus estimaciones. [Tip de R: para correr una regresión, utiliza el comando *lm*. Para producir las tablas, usa el paquete *stargazer*] Ojo: Las líneas horizontales en algunas variables (—) significan que NO debes incluir esta variable en la estimación de dicha columna.

Tabla 1: Estimaciones MCO

	Dependent variable:					
	prop_alimen		alimento_pc		ln(alimento_pc)	
	(1)	(2)	(3)	(4)	(5)	(6)
est_socio2		-		-		-
est_socio3		-		-		-
est_socio4		-		-		-
ln(ing_cor)	-		-		-	
urbano						
tot_integ						
prop_mujeres		-		-		-
prop_menores						
prop_p65mas						
sexo_jefe	-		-		-	
Constant						
Observations						
R^2						
Note:	Significativo al * 10 %; ** 5 %; *** 1 %					

(b) Interpreta de manera específica los coeficientes de las siguientes variables de la tabla:

(I) β_0 en la columna (3)
(II) *urbano* en la columna (5).
(III) *pop_mujeres* en la columna (2).
(IV) *sexo_jefe* en la columna (4).
(V) *est_socio*4 en la columna (1)
(VI) *prop_menores* en la columna (5).
(VII) *prop_p*65*mas* en la columna (6).
(VIII) *est_socio*4 en la columna (5).
(IX) ln(*ing_cor*) en la columna (1).
(X) ln(*ing_cor*) en la columna (6).
(XI) *urbano* en la columna (4).
(XII) *tot_integ* en la columna (5).
(XIII) ln(*ing_cor*) en la columna (4).

(c) En la especificación de la columna (5) notarás que se incluye como variable de control *urbano*, que, de acuerdo a la teoría económica, podría influir en la proporción de gasto en alimentos debido a diferencias en acceso a mercados y opciones de alimentos. Describe qué tipo de sesgo se hubiera producido de no incluir la variable *urbano* sobre el coeficiente de las variables relacionadas a *est_socio*. Estima la regresión auxiliar que te permitiría deducir el sesgo del coeficiente *est_socio* que hubiera existido en la estimación de la columna (5) de no haber incluido urbano. Reporta esta regresión con formato de ecuación. ¿Puedes calcular el coeficiente de *est_socio* que hubieras obtenido en la especificación (5) de no haber incluido urbano utilizando la columna (5) y la regresión auxiliar? De ser así, hazlo. Si no puedes, indica qué información te hace falta.

4. **[Alerta]** La región con mayor costo de una dieta saludable en 2020 fue América Latina y el Caribe (3.89 dólares por persona y día) de acuerdo con la Organización de las Naciones Unidas para la Alimentación y la Agricultura en su publicación: EL ESTADO DE LA SEGURIDAD ALIMENTARIA Y LA NUTRICIÓN EN EL MUNDO 2022 [1]. Asumiendo que el periodo trimestral es de 90 días y el tipo de cambio promedio en el 2020 fue de 21.48, el costo de una dieta saludable por persona sería de 7520.148 pesos. Crea una variable dummy que sea 1 si alimento por persona al dia es menor al numero establecido por la FAO y 0 si es igual o mayor.

(a) Haz una gráfica que muestre la media de la variable *alerta* para casa estrato socioeconómico y comenta tus hallazgos.

(b) Para analizar la relación entre la variable de *alerta* y el estrato socioeconómico debes realizar una regresión donde *alerta* sea la variable dependiente y *est_socio* la variable explicativa. Recuerda agregar variables adicionales si lo consideras necesario para capturar su efecto de *est_socio* correctamente. Reporta tu resultado con formato de ecuación.

(c) Realiza las estimaciones de la tabla 2 como en el inciso 3.a.

Tabla 2: Estimaciones MCO

	Dependent variable:	
	alerta	
	(1)	(2)
est_socio2		-
est_socio3		-
est_socio4		-
ln(ing_cor1)	-	
urbano		
prop_mujeres		-
prop_menores		
prop_p65mas		
sexo_jefe	-	
Constant		
Observations		
R^2		
Note:	Significativo al * 10 %; ** 5 %; *** 1 %	

(d) Interpreta de manera específica los coeficientes de las siguientes variables de la tabla 2:
‘

(i) β_0 en la columna (1)
(ii) est_socio4 en la columna (1).
(iii) $prop\_mujeres$ en la columna (1).
(iv) $sexo\_jefe$ en la columna (2).
(v) $prop\_menores$ en la columna (2).
(vi) $prop\_p65mas$ en la columna (2).
(vii) $urbano$ en la columna (2).

5. **[Extensión de Análisis]**

(a) Un compañero de tu consultoría intuye que a niveles bajos de ingreso, una mayor proporción del presupuesto familiar se destina a necesidades básicas, como la alimentación, dado que son gastos esenciales. Sin embargo, a medida que el ingreso aumenta, las familias podrían comenzar a asignar una menor proporción de su gasto a alimentos, privilegiando otros tipos de consumo o inversión. Finalmente, en niveles aún más altos de ingreso, podría observarse un aumento en la proporción del gasto en alimentos nuevamente, posiblemente debido a la adquisición de productos alimenticios de mayor calidad o más costosos, reflejando preferencias por bienestar o estatus. Esta dinámica sugiere una relación no lineal entre la proporción del gasto en alimentos y el ingreso corriente ¿Qué especificación estimarías para capturar la intuición de tu compañero? Realiza la estimación que propones e incluye una gráfica adecuada dada tu estimación.

(b) Se te pide considerar la posibilidad de que la proporción del gasto en alimentos varía entre las diferentes entidades federativas.

(i) Identifica la entidad federativa de cada hogar en la base de datos y crea una variable dummy *cdmx* que tome el valor de 1 si el hogar pertenece a la Ciudad de México (la clave de *ubica_loc* inicia con los dígitos 09) y 0 de lo contrario ¿Qué implicaciones tiene incluir la variable *cdmx* en su modelo y cómo podrían interpretar su relación con la proporción del gasto destinado a alimentos?

(ii) Se te pide analizar el contraste de la proporción de gastos que se destina a los alimentos entre la CDMX y el Estado de México (la clave *ubica_loc* del EDOMEX inicia con los dígitos 15). Debes modificar la especificación del inciso anterior para reflejar este interés. Interpreta cómo esta nueva especificación te permite analizar y comparar los efectos específcios entre estas dos entidades vecinas.

(iii) A continuación, plantea y estima un modelo que incluya una interacción entre la variable dummy cdmx y el estrato socioeconómico para examinar si hay diferencias en la proporción del gasto en alimentos entre los hogares de la Ciudad de México y los del resto del país, controlando por el tamaño del hogar y la presencia de menores. Reporta tus hallazgos en una tabla y proporciona una interpretación específica de los coeficientes asociados con la o las interacciones relevantes que hayas agregado. ¿Qué podrían sugerir los coeficientes significativos de interacción?

(c) Reflexionando sobre los desafíos de asegurar una dieta adecuada para los hogares en diferentes entornos, el gobierno desea investigar la interacción entre la urbanización y la estructura familiar, en particular, la presencia de menores. Desean comprender cómo la localización urbana o rural de un hogar y la cantidad de menores que viven en él pueden conjuntamente influir en la probabilidad de que el gasto alimentario por persona caiga por debajo del umbral definido por la FAO para una dieta saludable, lo cual se indica con la variable *alerta*.

(i) Formulen una hipótesis nula (H0) que establezca que la relación entre la proporción de menores y la probabilidad de *alerta* no es heterogenea entre hogares en el sector urbano y rural. En este inciso debes indicar la especificación (sin controles adicionales) que estimarías y plantea la prueba de hipótesis que te interesa para resolver la pregunta.

(ii) Estima la especificación del inciso anterior. Reporta el resultado gráficamente junto con un diagrama de dispersión como en el caso de la pregunta 2(b). Indica si se rechaza la prueba hipótesis de la pregunta anterior, y en caso de que así sea, ¿con que valor-p?

Tabla 3: Descripción de variables.

Variable	Descripción
folioviv	Identificador de la entidad federativa.
ubic_geo	Contiene la ubicación geográfica de la vivienda. Los dos primeros dígitos representan la clave de la entidad y los siguientes tres la clave del municipio.
tam_loc	1 = Localidades con 100 000 y más habitantes, 2 = Localidades con 15 000 a 99 999 habitantes, 3 = Localidades con 2 500 a 14 999 habitantes, 4 = Localidades con menos de 2 500 habitantes
est_socio	Clasificación de las viviendas del país de acuerdo a ciertas características socioeconómicas de las personas que las habitan, así como características físicas y el equipamiento de las mismas. 1 = Bajo, 2 = Medio bajo, 3 = Medio alto, 4 = Alto
ing_cor	Ingreso corriente: suma de los ingresos por trabajo, los provenientes de rentas, de transferencias, de estimación del alquiler y de otros ingresos. La variable de ingreso ha sido escalada dividiéndola por 10,000.
sexo_jefe	Distinción biológica que clasifica al jefe del hogar en hombre o mujer. 1 = Hombre, 0 = Mujer.
tot_integ	Número de personas pertenecientes a este hogar
gasto_mon	Gasto corriente monetario
alimentos	Los gastos en bienes de consumo no duradero que realizan día a día los integrantes del hogar en alimentos, bebidas y tabaco.
urbano	1=localidad urbana (más de 2,500 habitantes), 0=localidad rural
prop_mujeres	proporción de mujeres por nacimiento del total de integrantes de la vivienda.
prop_menores	proporción de niños de 11 o menos años de edad del total de integrantes de la vivienda.
prop_p65mas	proporción de habitantes de la vivienda con 65 años de edad o más
mujeres	Número de mujeres por nacimiento en la vivienda
menores	Número de integrantes de la vivienda de 11 o menos años de edad.
p65mas	Número de integrantes de la vivienda con 65 o más años de edad.

Ejemplo de reportando el resultado de una estimación con un diagrama de dispersión:

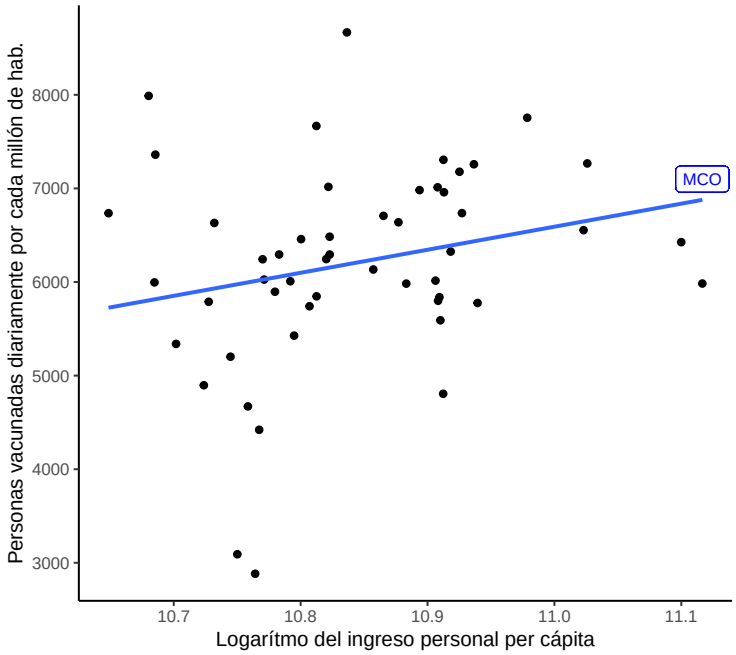

Figura 1: Scatterplot entre $dvaxx\_per\_mill$ y $\ln inc\_pc$

Ejemplo de reportando el resultado de una estimación con formato de ecuación:

$$\ln(ing\_trim_i) = \underset{(0.0044)}{9.601} - \underset{(0.0069)}{0.411}\, rural$$

Referencias

[1] Food and Agriculture Organization of the United Nations (FAO). El estado de la seguridad alimentaria y la nutrición en el mundo 2022. https://www.fao.org/3/cc0639es/online/sofi-2022/cost-affordability-healthy-diet.html, 2022. Accedido: 2022-12-03.