ENCUESTA NACIONAL SOBRE LOS ASPECTOS DE LA VIRTUALIDAD VINCULADOS CON LA PANDEMIA DEL COVID-19 (ENAVIRPA 2021)

ISADORE NABI

PRÓLOGO

La presente investigación es una parte constitutiva de un esfuerzo colectivo realizado por un determinado conjunto de investigadores con la finalidad de realizar mediciones sobre aspectos relacionados a la crisis sanitaria COVID-19 en determinado país, a través de cuestionarios realizados por encuestas por muestreo aleatorio. La encuesta sobre aspectos de la virtualidad relacionados a la pandemia COVID-19 se ha decidido colectivamente llamar de forma abreviada como ENAVIRPA. Esta investigación corresponde en lo fundamental al módulo de afectación económica de dicha encuesta, aunque es complementada con ítems (preguntas) de otros módulos, elaborados por otros(as) investigadores(as), quienes también fueron los que recolectaron la información relativa a ellos.

Este ejercicio académico no necesariamente posee carácter vinculante con la realidad nacional analizada. Lo anterior se debe a que, por un lado, para que lo tuviese, cada etapa de la investigación debería haberse preparado acorde a un conjunto de objetivos comunes y bien definidos (en función de ello se logra un diseño óptimo de los instrumentos), lo que a su vez implica la existencia de un marco teórico que unifique cada uno de los aspectos involucrados en la encuesta, expresados a manera de módulos. Por otro lado, en lo relativo al diseño muestral no se verificó que necesariamente el Muestreo Aleatorio Simple (MAS) fuese la técnica de muestreo óptima para los fines deseados y, de hecho, no lo es; en su lugar debería haberse utilizado un muestreo de tipo estratificado o un derivado del mismo. Tampoco se realizó un análisis psicométrico de los ítems para la validación del instrumento, tanto parcialmente (a nivel de módulo), como a nivel global (considerando conjuntamente todos los módulos). Finalmente, tampoco el tamaño de la muestra fue el ideal para aplicar sobre ésta metodologías de aprendizaje automático, así como también debe señalarse el hecho de que el tamaño de la muestra no puede considerarse representativo del tamaño de la población para

garantizar que la calidad de los datos muestrales sea la ideal en relación a expresar la generalidad de características de los datos poblacionales. Ello no es relevante en cuanto la presente investigación es un ejercicio empírico con la finalidad de comprender teóricamente los tipos de abordaje inferencial que es posible realizar con la información obtenida a través del instrumento psicométrico utilizado: la encuesta.

La investigación aborda en la primera sección la metodología descriptiva, en donde se presentan los resultados fundamentales de carácter descriptivo de tal forma que su información pueda ser capturada de forma instantánea por el(la) lector(a), sin la necesidad de la intervención de texto. En la segunda sección, se diseña y se ejecuta de forma complementaria una metodología para abordar inferencialmente los resultados específicos de la encuesta realizada. En la tercera sección se presentan las conclusiones. Finalmente, en la cuarta sección se presentan las conclusiones y los anexos pertinentes.

El lector observará en la segunda sección que el marco teórico relativo al modelo Probit es desproporcionadamente más pequeño que el relativo al Logit. Lo anterior obedece a la disponibilidad de literatura a favor del modelo Logit, lo que parecería tener que ver, entre otras cuestiones, con la similitud de la distribución logística con la distribución normal, que es de amplio interés aplicado en diversas áreas de las ciencias y las ingenierías.

Adicionalmente, debe señalarse que el criterio de discriminación para incluir módulos complementarios ha seguido un espíritu de análisis de la distribución de las pérdidas sociales (que implican pérdidas multidimensionales) entre los distintos sectores económicos del país, en línea con el análisis clásico de la economía.

Finalmente, por motivos de optimización de espacio se omitió presentar la totalidad de los resultados estadísticos obtenidos tras probar modelos lineales generalizados con las variables dicotómicas generadas por la información capturada por el cuestionario. Sin embargo, estos resultados se ponen a disposición del lector tal y como fueron exportados de SPSS a Word[1]; lo mismo ocurre con lo relativo al procesamiento de los resultados del levantamiento de encuestas y a los procedimientos mecánicos para la construcción de las variables dicotómicas[2].

[1] Véase https://mega.nz/folder/q0cQ2DhY#IOsjkZtrBa_C70j8A9aDRw.
[2] Véase https://mega.nz/folder/iscU0bAb#P_yDdBvhOWv-v7a7-ReCoQ.

I. OBJETIVOS

I. I. OBJETIVO GENERAL

Planear de forma general[3] una encuesta por muestreo vinculada a la crisis sanitaria por COVID-19, con énfasis en la afectación económica vinculada a ella.

I. I. OBJETIVOS ESPECÍFICOS

Analizar los resultados obtenidos (tras la aplicación de la encuesta por muestreo) a través de los modelos lineales generalizados (MLG), con énfasis en los resultados relativos a la afectación económica vinculada a la crisis sanitaria estudiada.

Mostrar la complementariedad existente, en el contexto de los MLG, entre la teoría estadística clásica y la teoría del aprendizaje automático.

[3] Esto comprende desde el planteamiento de sus objetivos, los temas de estudio, confección del cuestionario, etc., hasta la recolección de los datos, procesamiento e informe final de resultados.

II. METODOLOGÍA DESCRIPTIVA

II.I. RESUMEN GENERAL

Metodología		
Población y marco muestral	**Población de estudio**	Personas de 18 años y más, usuarias de la telefonía celular, residentes dentro del
	Marco muestral de la encuesta	Primeros cuatro dígitos activos de los teléfonos celulares activos de las operadoras telefónicas existentes en el país según la Superintendencia de Telecomunicaciones (SUTEL).
	Tamaño del marco muestral	El tamaño del listado completo de las unidades de muestreo no está determinado
	Características	Incluye únicamente teléfonos celulares.
	Fecha de actualización	2021
	Sectores y/o categorías cubiertas actualmente	Se cubren todos los bancos telefónicos activos de los celulares del país.
Muestreo y precisión	**Método de muestreo**	Muestro de bancos telefónicos celulares activos del país, utilizando el procedimiento de Waksberg, para entrevistar personas de 18 años o más.
	Tamaño de la muestra	235 entrevistas
	Cobertura de la muestra	Cubre el 97% de la población de 18 años o más.
	Error de muestreo	6,39%
	Tasa de respuesta	30,0546%
	Tratamiento de la no respuesta	Se realizan cuatro llamadas telefónicas en diferentes días y a diferentes horas para localizar a la persona a entrevistar.
	Sistema de ponderación	No se utilizó ponderación al no variar las proporciones muestrales significativamente de las poblacionales.
Recolección de datos	**Método de encuesta**	Entrevistas telefónicas asistidas por computador (CATI)
	Período de trabajo de campo	Del 21 de junio de 2021 al 26 de junio de 2021
	Presentación de resultados	Entre el 26 de julio de 2021 y el 30 de julio de 2021
Última actualización del metadato		2021

Fuente: Elaboración propia.

II.II. DESCRIPCIÓN GENERAL DEL PROCESO DE INVESTIGACIÓN

II.II. I. GENERALIDADES

El proceso de levantamiento de encuestas fue realizado por diez estudiantes del programa de posgrado de la Universidad de Costa Rica. Inicialmente se recibieron capacitaciones sobre el uso del programa computacional CSPro (programa diseñado para el levantamiento de encuestas federales por parte del gobierno de los Estados Unidos) por parte Gerald Mora (profesor encargado de la asignatura) y Yorlene Quirós (especialista en el uso de CSPro y docente de la UCR), el cual está diseñado para fines de realización de encuestas. Se usó específicamente la versión 7.4 del CSPro.

Luego de las capacitaciones, Yorlene Quirós se encargó de ensamblar los distintos módulos del cuestionario y, luego de ello, se procedió a realizar una prueba piloto buscaba observar el desempeño del cuestionario con una cantidad de encuestados por estudiante igual a tres, equivalente al 10% de lo que se planificó inicialmente encuestar, puesto que en definitiva se buscaba capturar la información de 300 personas, *i.e.*, realizar 300 entrevistas, específicamente 30 por estudiante.

Posteriormente, se realizó el trabajo de campo y se pudieron completar únicamente 235 encuestas, sin embargo, las 30 encuestas correspondientes al módulo de afectación económica fueron completadas exitosamente.

Finalmente, para el caso del trabajo de campo realizado por autor de esta investigación, se realizaron 332 llamadas, el tiempo al teléfono fue de aproximadamente 18 horas, en la mayor parte de casos (por distintos motivos) se omitieron las revisitas (re-llamadas, para este caso), la cantidad de teléfonos ocupados (código 1 en la clasificación del manual para el trabajo de campo) fue de 34, la cantidad de teléfonos que no respondieron fue de 94 (código 2) y el total de llamadas descontando los números inactivos fue de 183. Así, la tasa de respuesta fue de $(1 - (128/183)) \times 100\% = 30.0546\%$.

II.II. II. CRONOGRAMA DE ACTIVIDADES, CRONOLOGÍA GENERAL DEL TRABAJO DE CAMPO, BANCOS TELEFÓNICOS UTILIZADOS Y REGISTRO DE LLAMADAS

II.II. II. I. Cronograma de Actividades

	Junio				Julio				
	31 al 4	7 al 11	14 al 18	21 al 25	28 al 2	5 al 9	12 al 16	19 al 23	26 al 30
Cuestionario	X								
Enviar a Yorlene la carpeta de cada módulo (ZIP) a mas tardar mañana miércoles 9 de junio		X							
Montaje unificado en CSPro por el USES (Yorlene)		X (lunes 14)							
Dropbox			X						
Uso de la versión 7.4 del CSPro (para trabajar CAPI)			X						
Prueba cuestionario (3 c/u)			X						
Envio de los bancos telefónicos			X						
Trabajo de campo (300 entrevistas / 30 c/u)			X	X	X				
Revisión y análisis de datos				X	X	X			
Informe						X	X		
Presentación de informes							X	X	X

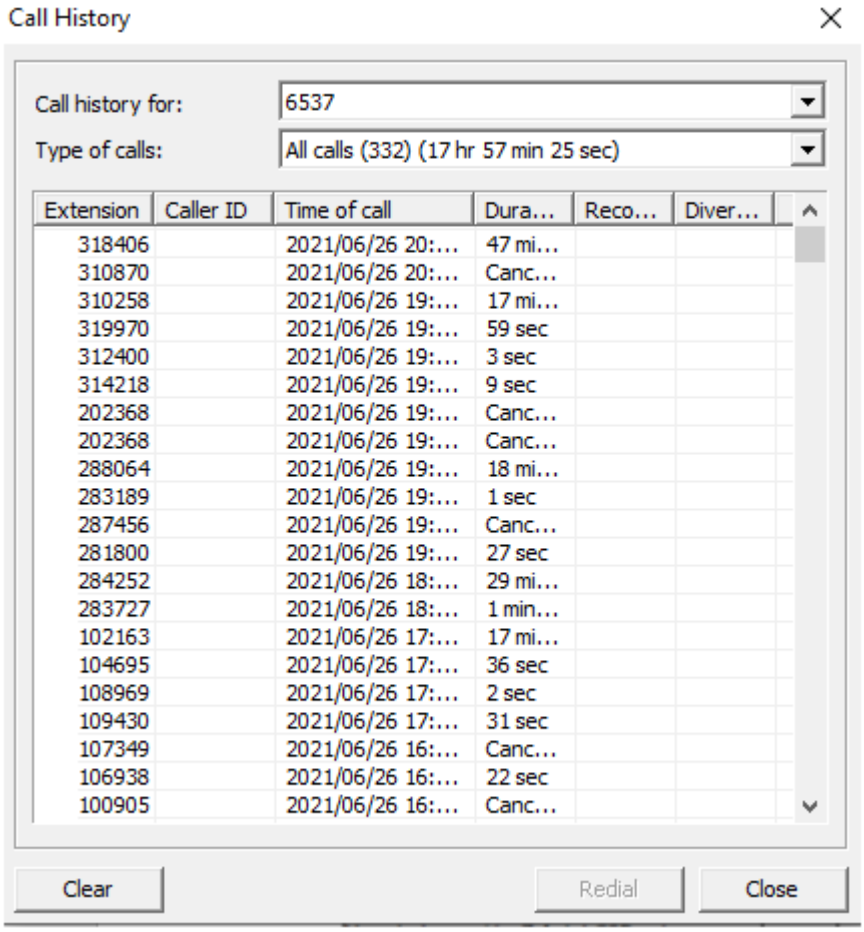
II.II. II. II. Bancos Telefónicos

ID	Nombre	BANCOS TELEFÓNICOS					
		ASIGNADOS		USADOS		NO USADOS	
		INICIAL	FINAL	INICIAL	FINAL	INICIAL	FINAL
1		1	60	1	60		
2		61	120				
3		121	180				
4		181	240				
5		241	300				
6		301	360				
7		361	420	361	394	395	420
8		421	480	421	480		
9		481	540				
10	AE	541	600	541	571	572	600

II.II. II. III. Cronología General del Trabajo de Campo

ID	Nombre	ENTREVISTAS REALIZADAS COMPLETAS (CODIGO 4)														
		TOTAL	DIAS													
			19-jun	20-jun	21-jun	22-jun	23-jun	24-jun	25-jun	26-jun	27-jun	28-jun	29-jun	30-jun	1-jul	2-jul	3-jul
TOTAL GENERAL		**235**	0	1	9	22	32	5	13	22	7	13	34	23	15	13	26
1		**19**	0	0	0	0	1	0	0	6	0	0	2	0	0	3	7
2		**28**	0	0	1	4	3	0	0	2	0	3	3	0	1	1	10
3		**30**	0	0	0	1	7	3	1	2	0	3	4	4	2	3	
4		**30**	0	0	1	1	0	0	3	4	0	1	3	9	1	3	4
5		**21**	0	0	0	3	7	0	0	0	0	0	11	0	0	0	0
6		**16**	0	0	4	0	7	0	0	0	0	0	0	5	0	0	0
7		**16**	0	0	0	3	2	0	0	0	0	6	0	1	4	0	0
8		**24**	0	1	0	1	0	0	0	1	7	0	6	4	0	0	4
9		**21**	0	0	0	2	0	2	1	0	0	0	5	0	7	3	1
10	AE	**30**	0	0	3	7	5	0	8	7							

II.II. II. IV. Registro de Llamadas



II.III. PRESENTACIÓN DEL MÓDULO AFECTACIÓN ECONÓMICA (AE)

AFECTACIÓN ECONÓMICA	
AE1	¿Recibe o recibió usted algún beneficio o de apoyo financiero del gobierno, nacional o local, desde que inició la emergencia sanitaria/cuarentena causada por el COVID19? 1.SÍ 2. NO **(PASE A AE3)** 3. NO SABE/NO RESPONDE **(PASE A AE3)**
AE2	Por favor, indique cuáles de los siguientes beneficios/apoyos usted recibe o recibió **(Respuesta múltiple)** 1. Sí, comida/alimentos 2. Sí, recursos económicos 3. Sí, suministros médicos para prevención (guantes, mascarillas, desinfectante, etc.) 4. Sí, suministros de higiene personal (toallas sanitarias, pañales para bebés, etc.) 5. NO 9. NO SABE/NO RESPONDE

Resultado de la emergencia sanitaria/cuarentena causada por el COVID, por favor indique, cómo se han visto	**Aumentó**	**Sin cambios**	**Disminuyó**	**No es una fuente**

afectados sus recursos personales: Han aumentado, disminuido o no han tenido cambios.					**de ingresos / apoyo**
AE3	Ingresos o ganancias de un trabajo remunerado	4	3	2	1
AE4	Dinero o bienes recibidos de familiares/amigos que viven en otras partes del país.	4	3	2	1
AE5	Dinero o bienes recibidos de familiares/amigos que viven en otro país.	4	3	2	1
AE6	Ingresos de propiedades de alquiler, inversiones o ahorros	4	3	2	1
AE7	Pensiones y/o jubilaciones u otros pagos sociales.	4	3	2	1
AE8	Apoyo del gobierno	4	3	2	1
AE9	Apoyo/donación de organizaciones no gubernamentales, organizaciones de la sociedad civil, fundaciones u otras organizaciones sin fines de lucro	4	3	2	1

AE10 Además de usted ¿algún otro miembro de su hogar ha sufrido algún cambio en el ingreso económico desde que comenzó la emergencia sanitaria/cuarentena causada por el COVID-19?

1. No hay cambio en el ingreso 2. Incremento de ingresos 3. Ingreso disminuido

AE11 Desde que comenzó la emergencia sanitaria/cuarentena causada por el COVID-19 ¿Usted diría que el salario o ingreso total que su familia recibe mensualmente les alcanza o no les alcanza para vivir? **(SONDEE LA MEJOR RESPUESTA)**

1. No les alcanza, tienen grandes dificultades
2. No les alcanza, tiene dificultades
3. Les alcanza justo, sin grandes dificultades
4. Les alcanza bien, pueden ahorrar
9. NS/NR

II.II. PRESENTACIÓN DE LOS ÍTEMS EMPLEADOS DE OTROS MÓDULOS

II.II. I. MÓDULO SOCIODEMOGRÁFICO (SD)

SOCIODEMOGRÁFICAS			
SD1	**ENTREVISTADOR(A): ANOTE SEXO DEL ENTREVISTADO**	1. HOMBRE 2. MUJER	
SD2	Por favor, dígame ¿cuál es su edad?	AÑOS /____/____/	
SD3	¿Es Usted costarricense?	1. SI	2. NO
SD5	¿Cuál es el ingreso total mensual de este hogar? **(SONDEAR)**	1. MENOS DE 200 MIL COLONES 2. 200 MIL A MENOS DE 400 MIL	5. 800 MIL A MENOS DE 1 MILLÓN 6. 1 MILLÓN A MENOS DE 1,5 MILLONES 7. 1,5 MILLONES O MÁS 0. NO SABE / NO RESPONDE

	3. 400 MIL A MENOS DE 600 MIL 4. 600 MIL A MENOS DE 800 MIL

II.II. II. MÓDULO EDUCACIÓN VIRTUAL (E)

RETOS DE EDUCACIÓN VIRTUAL	
E6	¿Cuál es la razón por la que no ha participado o matriculado algún programa de educación/formación virtual/en línea? 1. No cuenta con el tiempo suficiente 2. No cuenta con el dinero suficiente 3. En este momento no está interesado(a) en capacitarse 4. Le faltan destrezas informáticas 5. Tiene baja motivación 6. El ambiente virtual le resulta aburrido 7. Falta de organización del tiempo personal 8. Pospone repetidamente la decisión de iniciar 9. El ambiente virtual le resulta incómodo 10. No desea iniciar solo(a) 11. La educación en línea es muy cara 88. Otro: _____ 99.NS/NR

II.II. III. MÓDULO TECNOLOGÍA (TC)

TECNOLOGÍA – TOD@S	
TC1	¿Cuenta usted con servicio de internet fijo/modem en su casa de habitación? 1. SÍ 2. NO TIENE **(PASE A TC3)** 9. NS/NR **(PASE A TC3)**

II.II. IV. MÓDULO TELETRABAJO (TE)

TELETRABAJO – JORDAN/ CARLOS
Actualmente, durante una semana típica, ¿Cuál es su principal actividad laboral? **SE SELECCIONA LA OPCIÓN QUE MEJOR SE ADECÚE** 1. Trabaja para un patrón (ya sea persona, empresa u hogar. 2. Ayuda en un negocio familiar (sin remuneración) 3. Tengo mi propio negocio y empleo a otras personas 4. Tengo mi propio negocio sin emplear a otras personas

SI RESPONDE DE 5 A 99 PASE A V1	5. No trabaja, es pensionado(a), jubilado(a) 6. No trabaja, se dedica al hogar 7. No trabaja, es estudiante de tiempo completo 8. No trabaja, por una limitación física/mental que me impide trabajar 9. No trabaja, no busca trabajo y no está disponible para trabajar 10. No trabaja, pero está buscando trabajo 88. Otra 99. NS/NR

II.II. V. MÓDULO DE SALUD FÍSICA Y MENTAL (FN)

HÁBITOS DE SALUD FÍSICA Y NUTRICIÓN – LUIS

FN2	Con respecto a su consumo de agua, ¿durante la pandemia ha incrementado la cantidad de agua que bebe?: 1. Sí, ha aumentado 2. No, se ha mantenido igual 3. No, ha disminuido 9. NS//NR

Con respecto a los hábitos de los tiempos de comida, comparando la frecuencia actual con respecto a la frecuencia el año previo a la pandemia, usted considera que han aumentado, disminuido o se mantiene igual:		Se mantiene igual	Aumentó	Disminuyó	NS/NR
FN6	Desayuno	1	2	3	9
FN7	Almuerzo	1	2	3	9
FN8	Cena	1	2	3	9

Con respecto a los alimentos que le voy a mencionar. Comparando la frecuencia actual con respecto a sus hábitos el año previo a la pandemia, ¿usted considera que han aumentado, disminuido o se mantiene igual el consumo de …. **(LEER ALIMENTO)**?		NO LO HA COMIDO	SI			NS/NR
			Mantenido Igual	Aumentado	Disminuido	
FN9	Frutas y vegetales	0	1	2	3	9
FN10	Leguminosas (frijoles, garbanzos, lentejas)	0	1	2	3	9
FN11	Lácteos (leche, yogurt, queso)	0	1	2	3	9
FN12	Harinas (arroz, pasta)	0	1	2	3	9
FN13	Carnes o huevo (res, cerdo, pollo, pescado, embutidos)	0	1	2	3	9
FN14	Alimentos/dulces/bebidas fuentes de azúcar. (galletas, Gaseosas, Hi-C, Powerade, etc.), golosinas, chocolates, confites, helados)	0	1	2	3	9

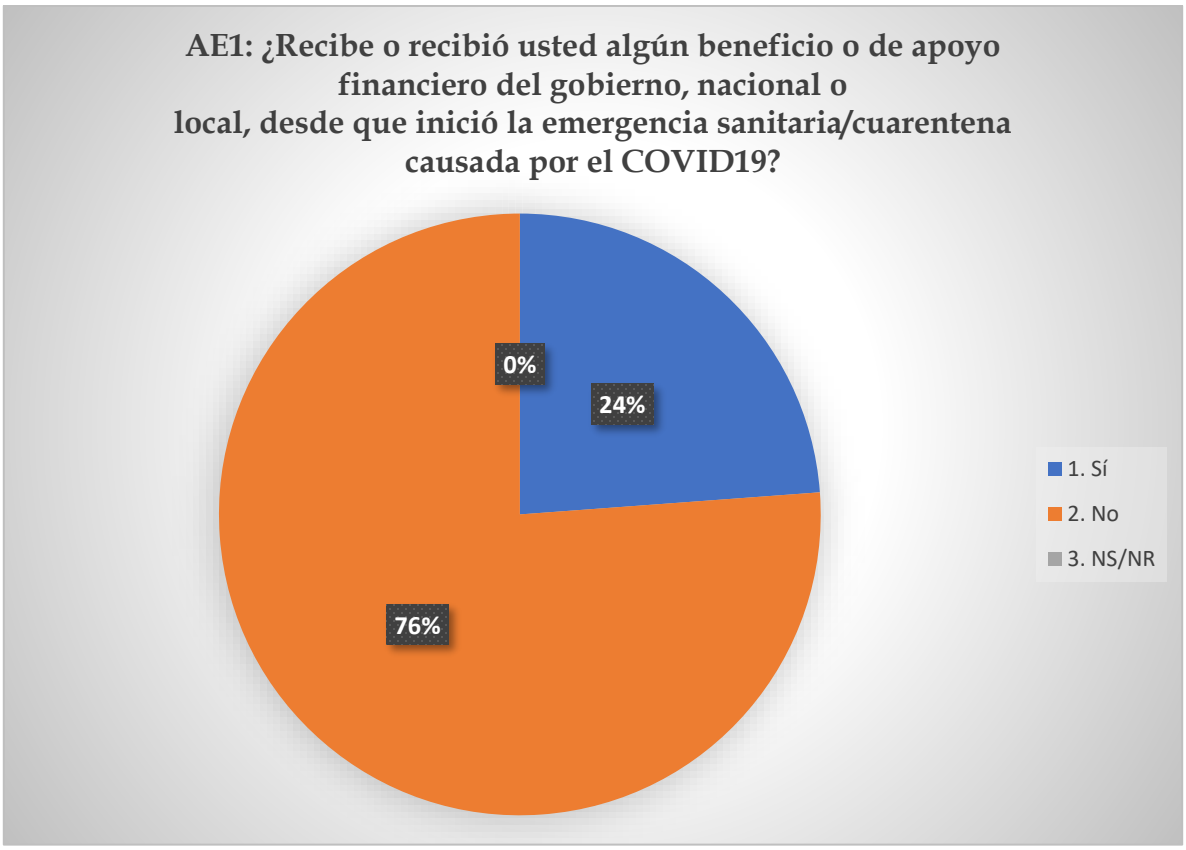
FN15	Grasas y Comidas rápidas (mantequilla, natilla, aderezos, mayonesa) (Pollo frito, papas fritas, pizza, hamburguesas, tacos, repostería, empaquetados (papas tostadas, bolitas de queso, tronaditas, etc.).	0	1	2	3	9

III. PRESENTACIÓN DE RESULTADOS DESCRIPTIVOS

III.I. MÓDULO AE (AFECTACIÓN ECONÓMICA)

III.I. I. ÍTEM AE1

AE1: ¿Recibe o recibió usted algún beneficio o de apoyo financiero del gobierno, nacional o local, desde que inició la emergencia?	
1. Sí	56
2. No	179
3. NS/NR	0
TOTAL	235

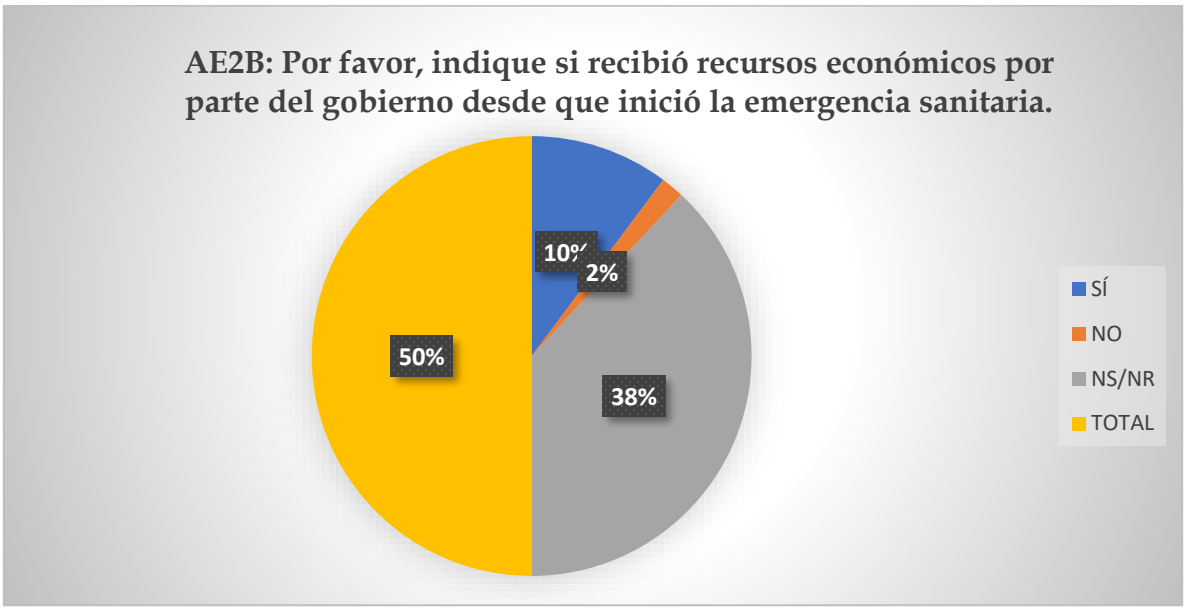


III.I. II. ÍTEM AE2

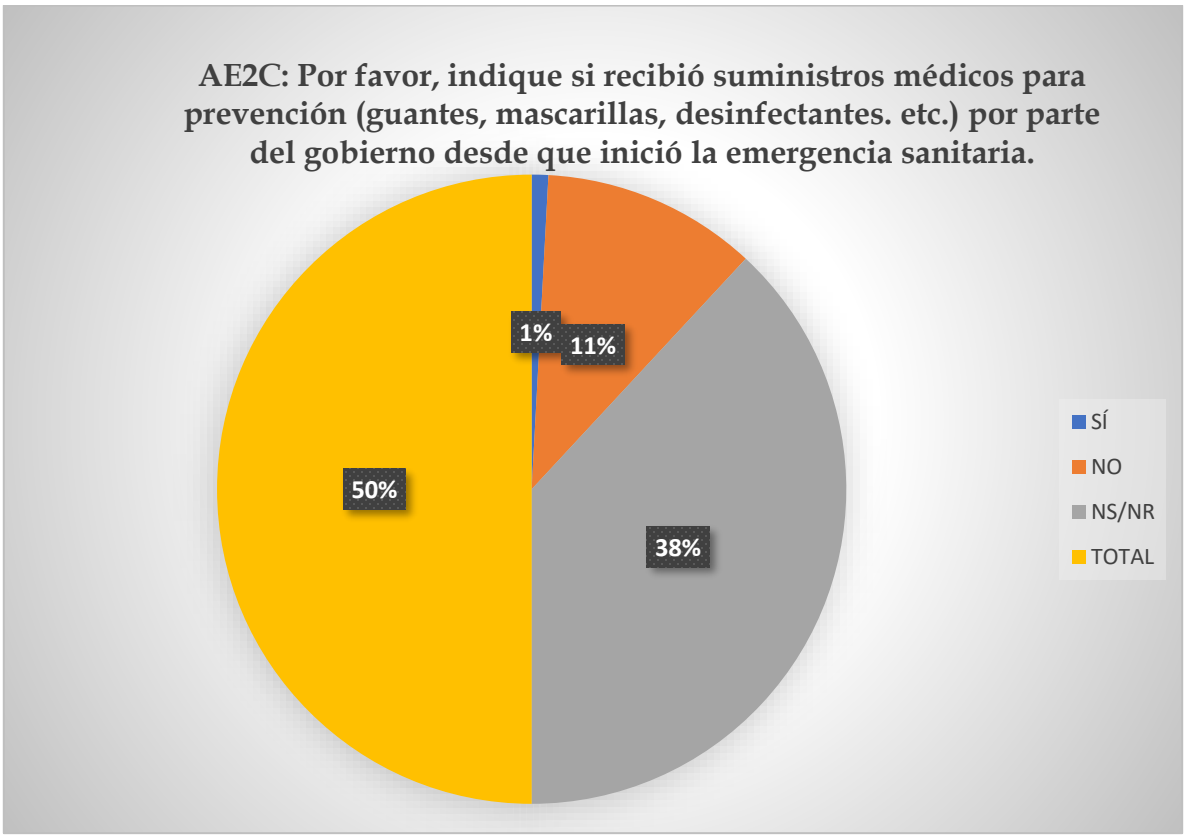
AE2: Por favor, indique cuáles de los siguientes beneficios/apoyos usted recibe o recibió	SÍ	NO	NS/NR	TOTAL
AE2A: Sí, comida/alimentos	23	33	179	235
AE2B: Sí, recursos económicos	48	8	179	235
AE2C: Sí, suministros médicos para prevención (guantes, mascarillas, desinfectante, etc.)	4	52	179	235
AE2D: Sí, suministros de higiene personal (toallas sanitarias, pañales para bebés, etc.)	2	54	179	235

III.I. II.I. Ítem AE2A

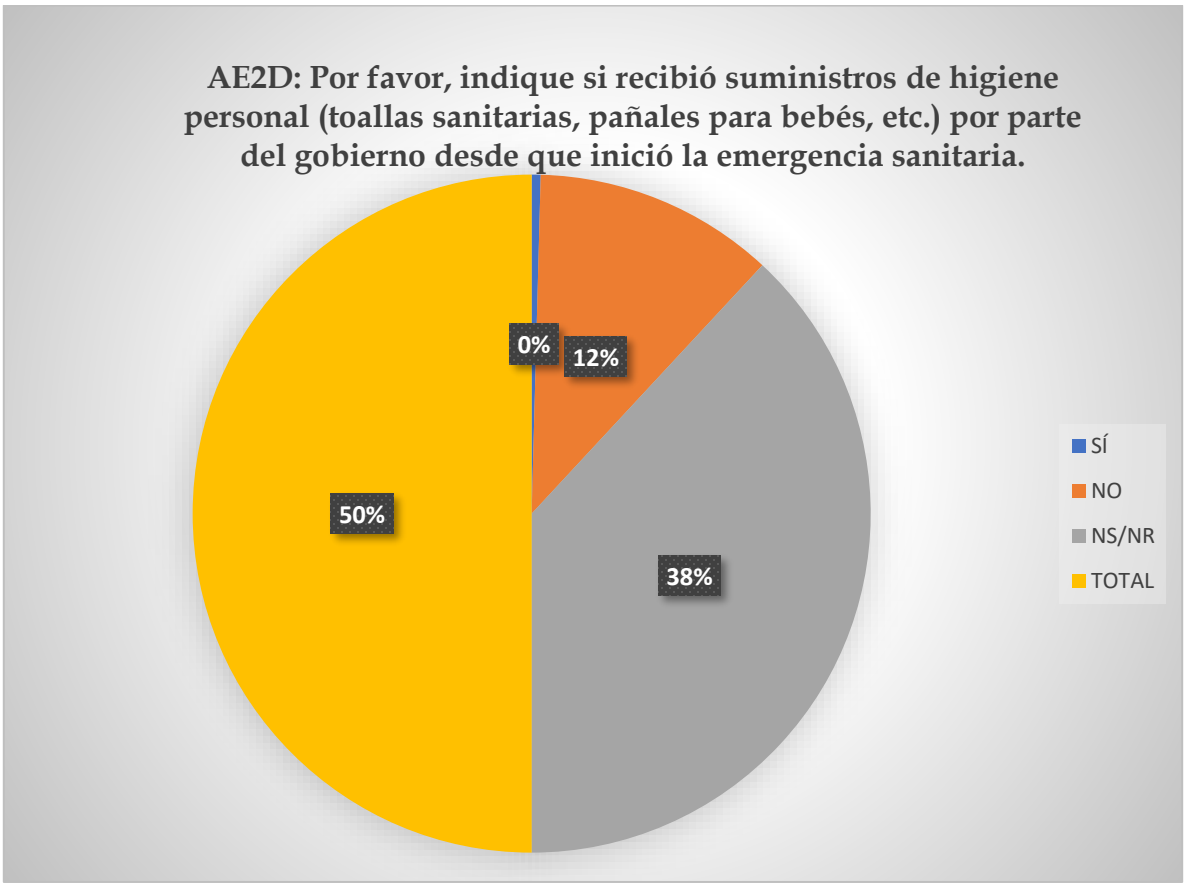
III.I. II. II. Ítem AE2B



III.I. II. III. Ítem AE2C

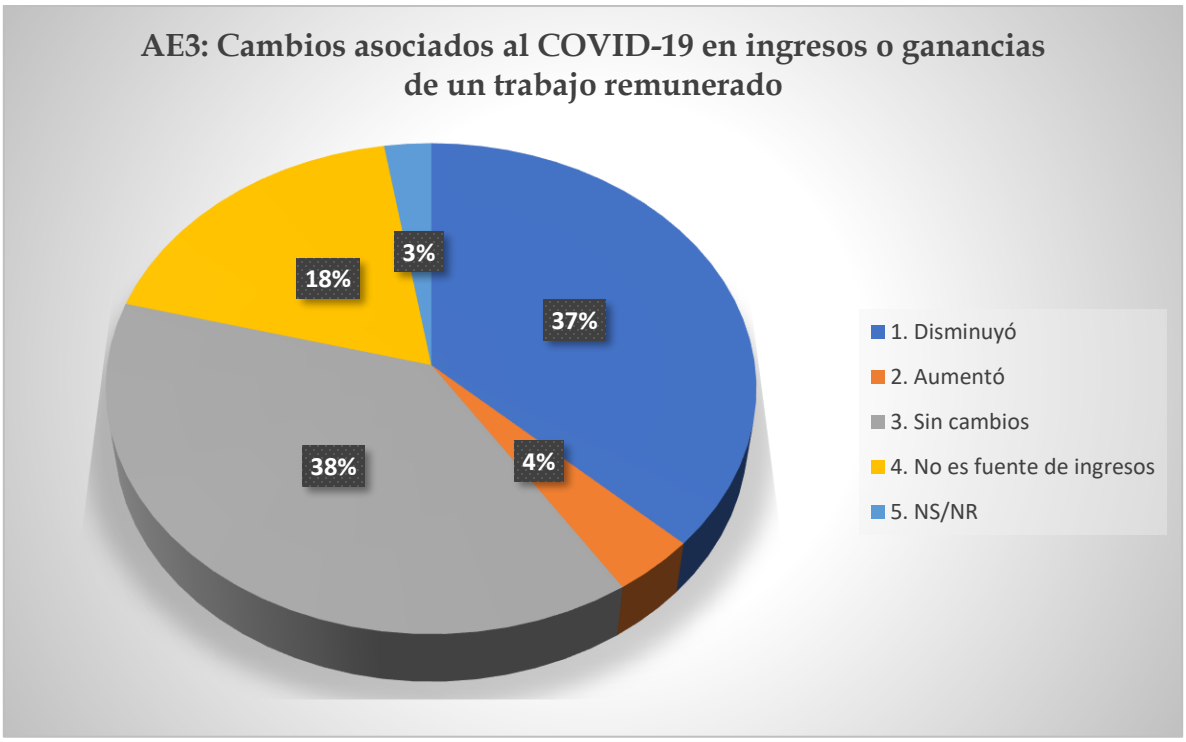


III.I. II. III. Ítem AE2D



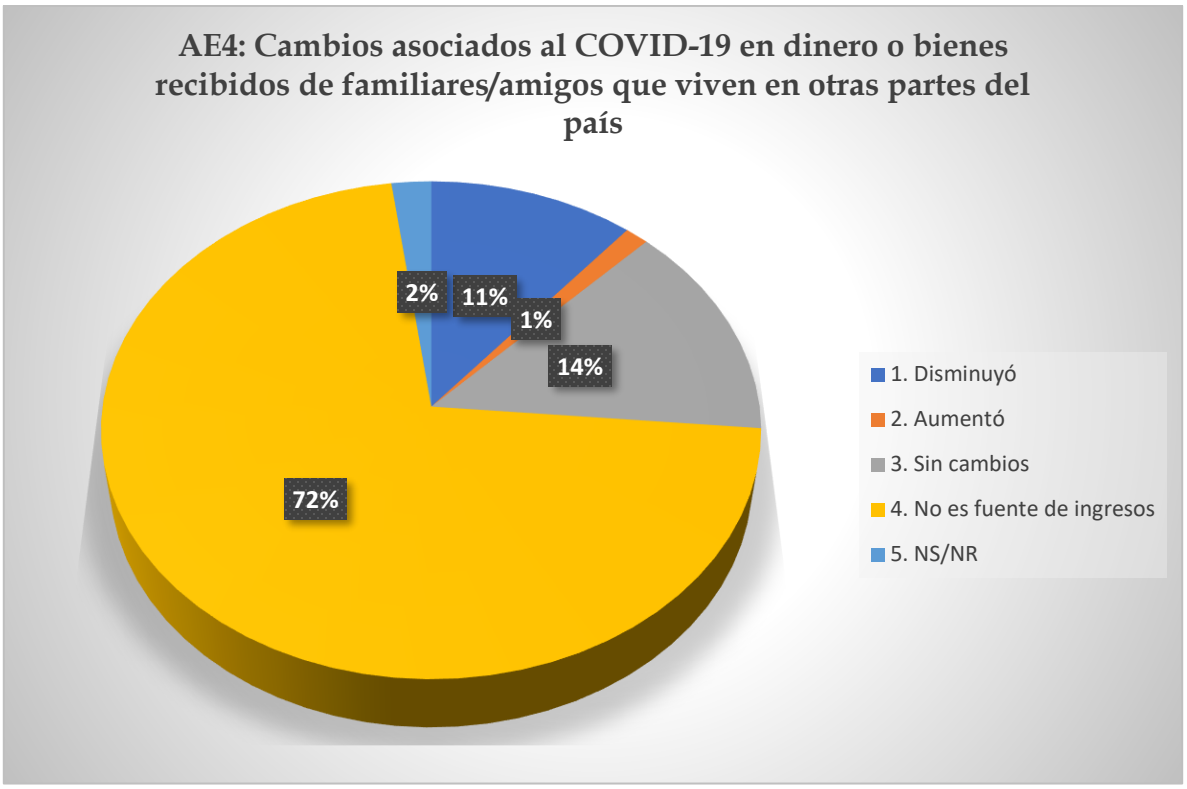
III.I. III. ÍTEM AE3

AE3: Cambios asociados al COVID-19 en ingresos o ganancias de un trabajo remunerado	
1. Disminuyó	87
2. Aumentó	9
3. Sin cambios	90
4. No es fuente de ingresos	43
5. NS/NR	6
TOTAL	235



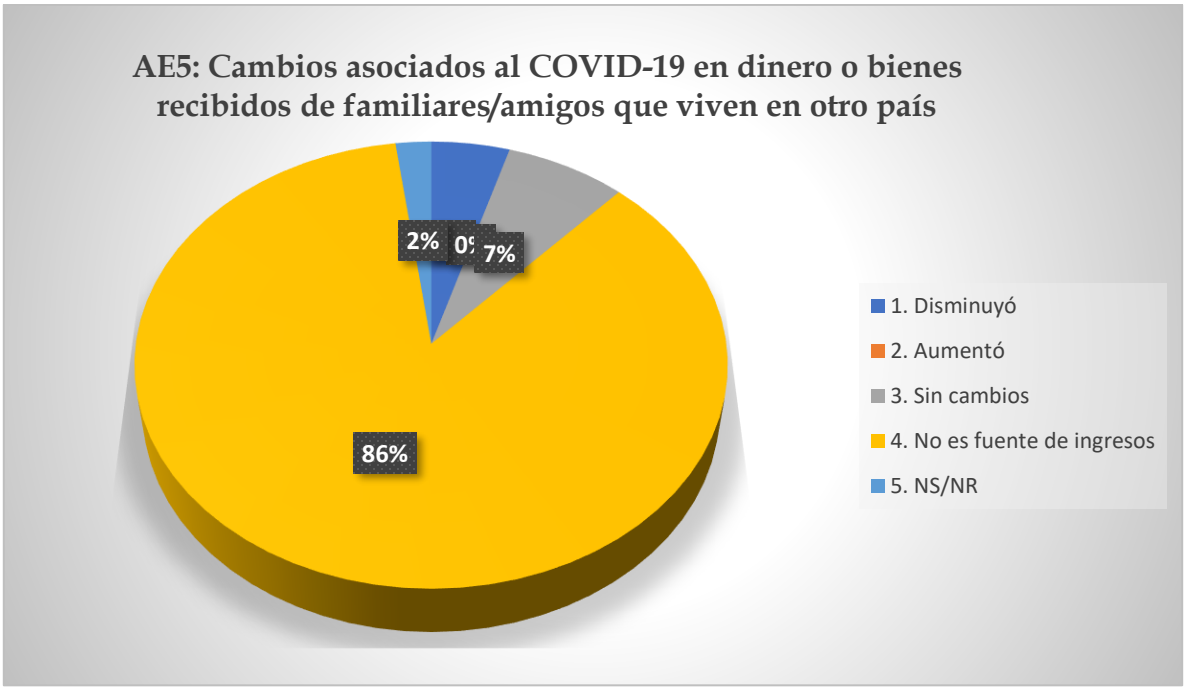
III.I. IV. ÍTEM AE4

AE4: Cambios asociados al COVID-19 en dinero o bienes recibidos de familiares/amigos que viven en otras partes del país	
1. Disminuyó	26
2. Aumentó	3
3. Sin cambios	33
4. No es fuente de ingresos	168
5. NS/NR	5
TOTAL	235



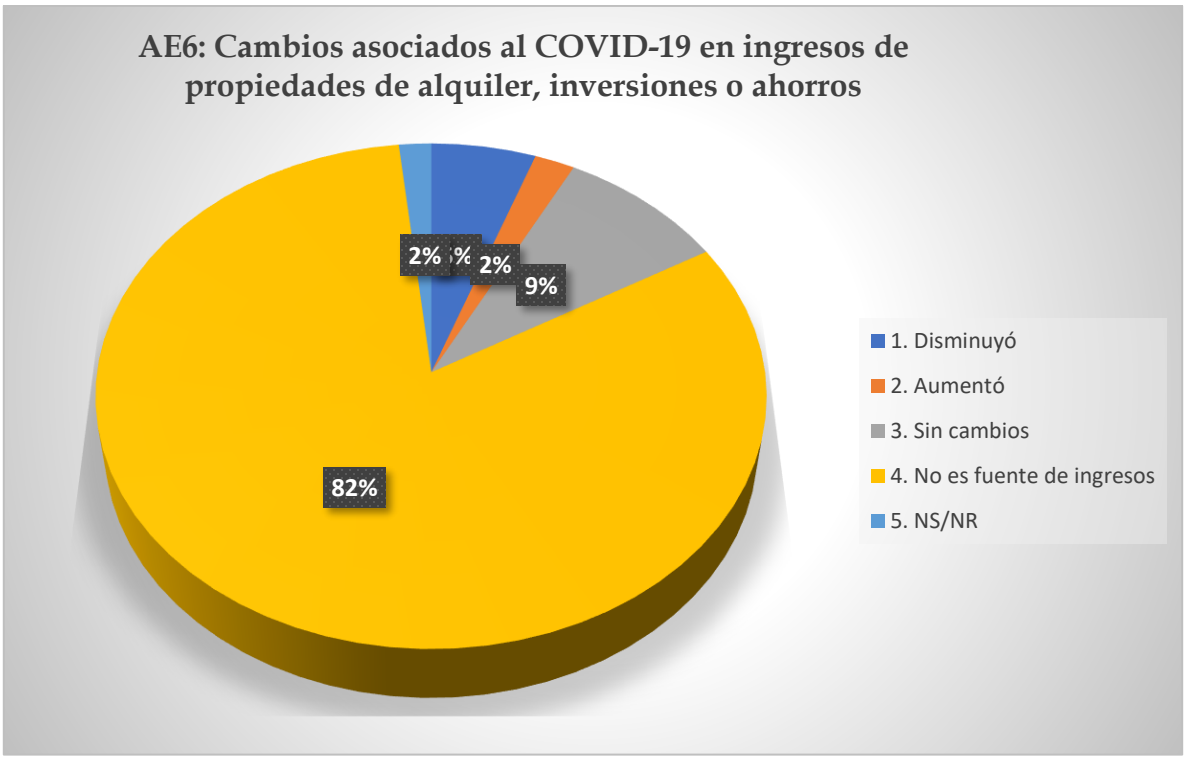
III.I. V. ÍTEM AE5

AE5: Cambios asociados al COVID-19 en dinero o bienes recibidos de familiares/amigos que viven en otro país	
1. Disminuyó	11
2. Aumentó	0
3. Sin cambios	17
4. No es fuente de ingresos	202
5. NS/NR	5
TOTAL	235



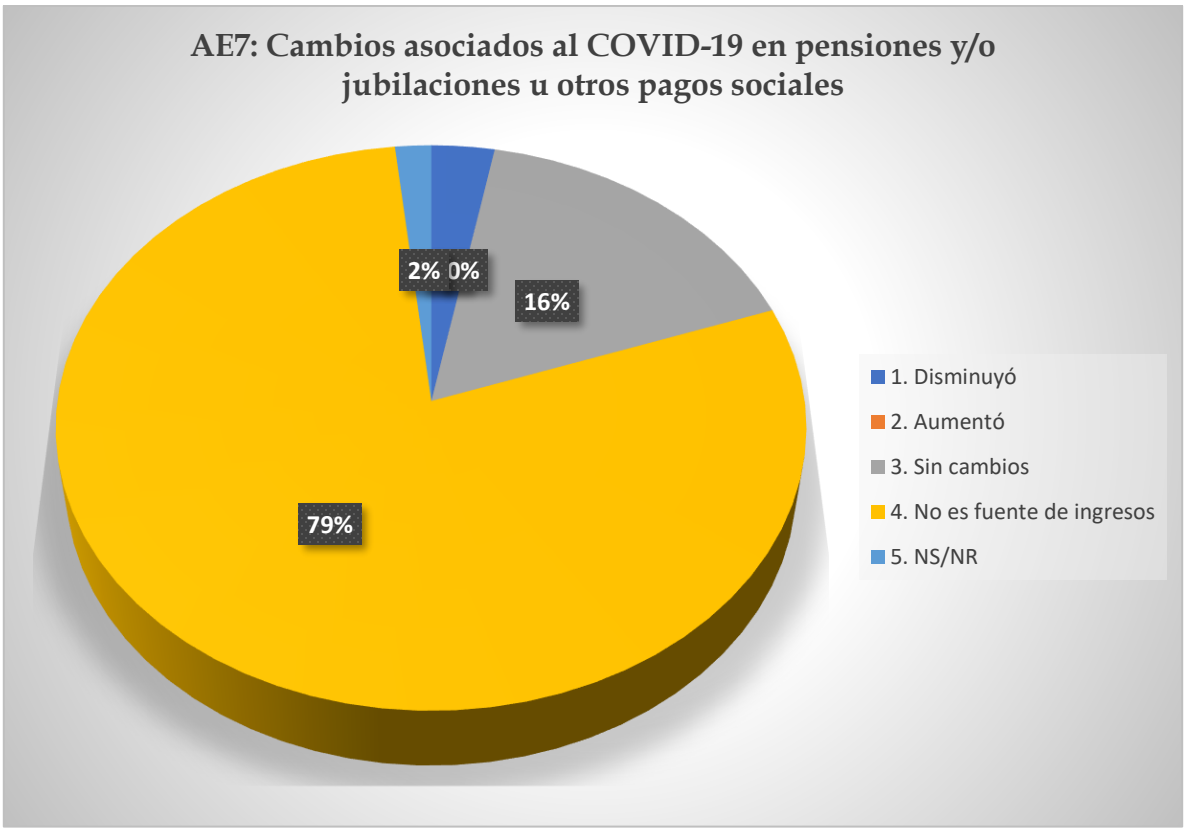
III.I. VI. ÍTEM AE6

AE6: Cambios asociados al COVID-19 en ingresos de propiedades de alquiler, inversiones o ahorros	
1. Disminuyó	13
2. Aumentó	5
3. Sin cambios	21
4. No es fuente de ingresos	192
5. NS/NR	4
TOTAL	235



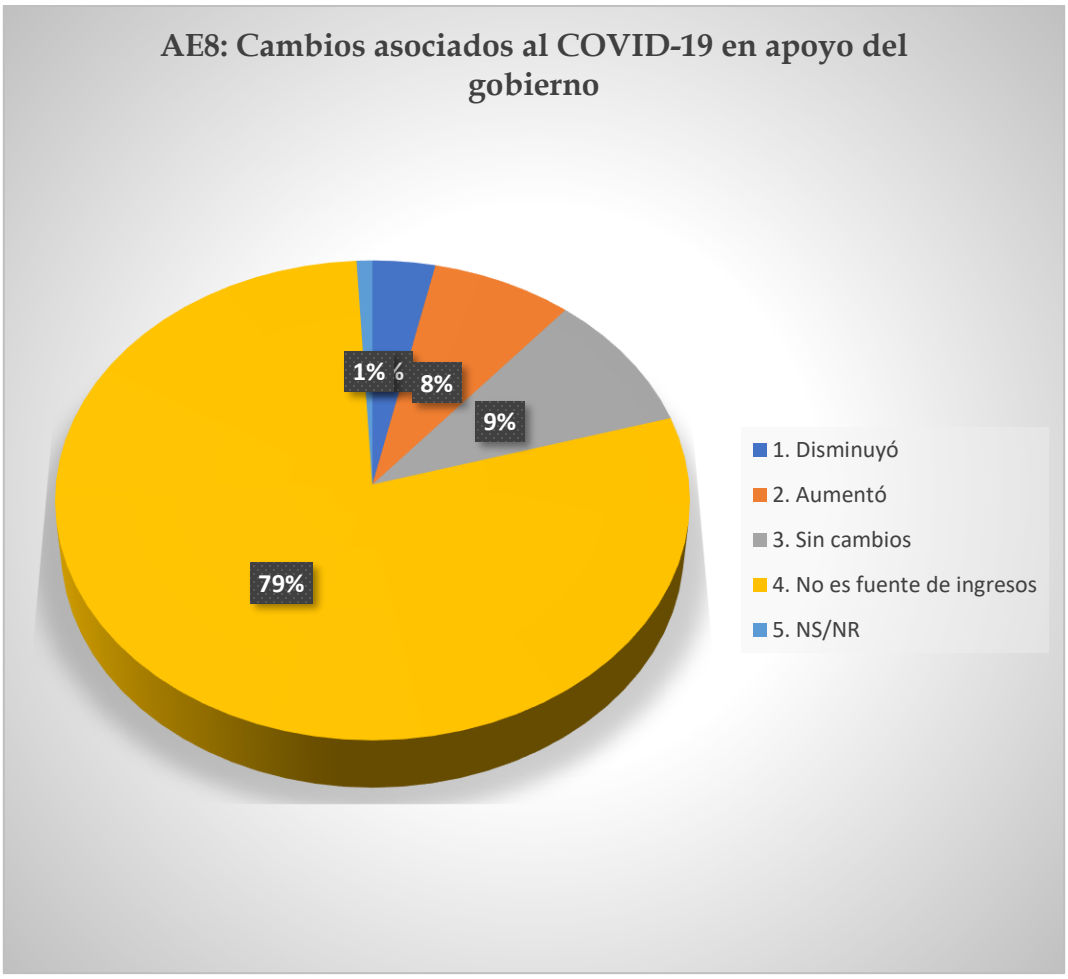
III.I. VII. ÍTEM AE7

AE7: Cambios asociados al COVID-19 en pensiones y/o jubilaciones u otros pagos sociales	
1. Disminuyó	7
2. Aumentó	0
3. Sin cambios	39
4. No es fuente de ingresos	185
5. NS/NR	4
TOTAL	235



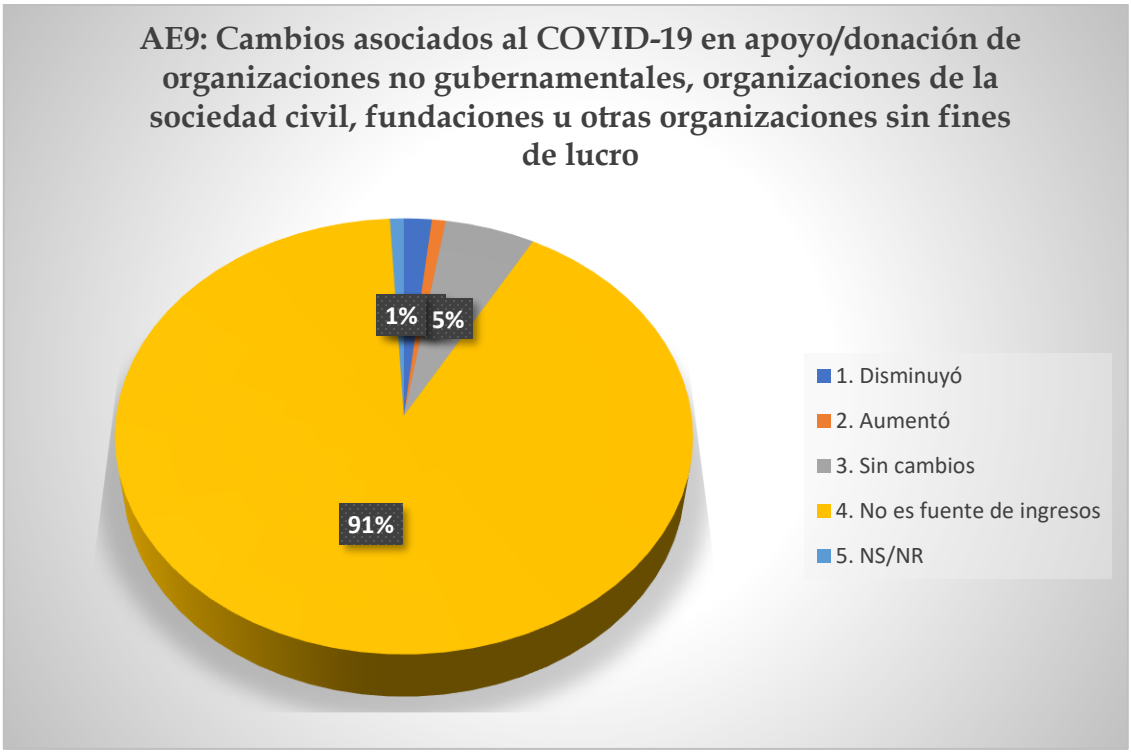
III.I. VII. ÍTEM AE8

AE8: Cambios asociados al COVID-19 en apoyo del gobierno	
1. Disminuyó	8
2. Aumentó	18
3. Sin cambios	22
4. No es fuente de ingresos	185
5. NS/NR	2
TOTAL	235



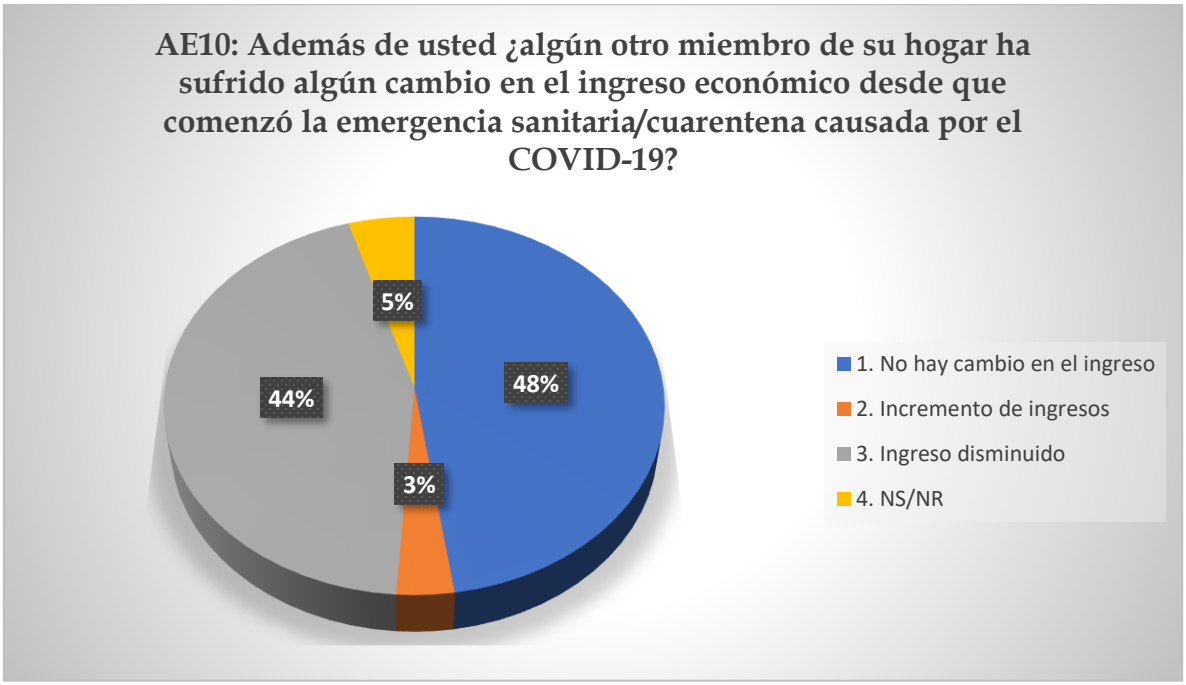
III.I. IX. ÍTEM AE9

AE9: Cambios asociados al COVID-19 en apoyo/donación de organizaciones sin fines de lucro	
1. Disminuyó	4
2. Aumentó	2
3. Sin cambios	13
4. No es fuente de ingresos	214
5. NS/NR	2
TOTAL	235



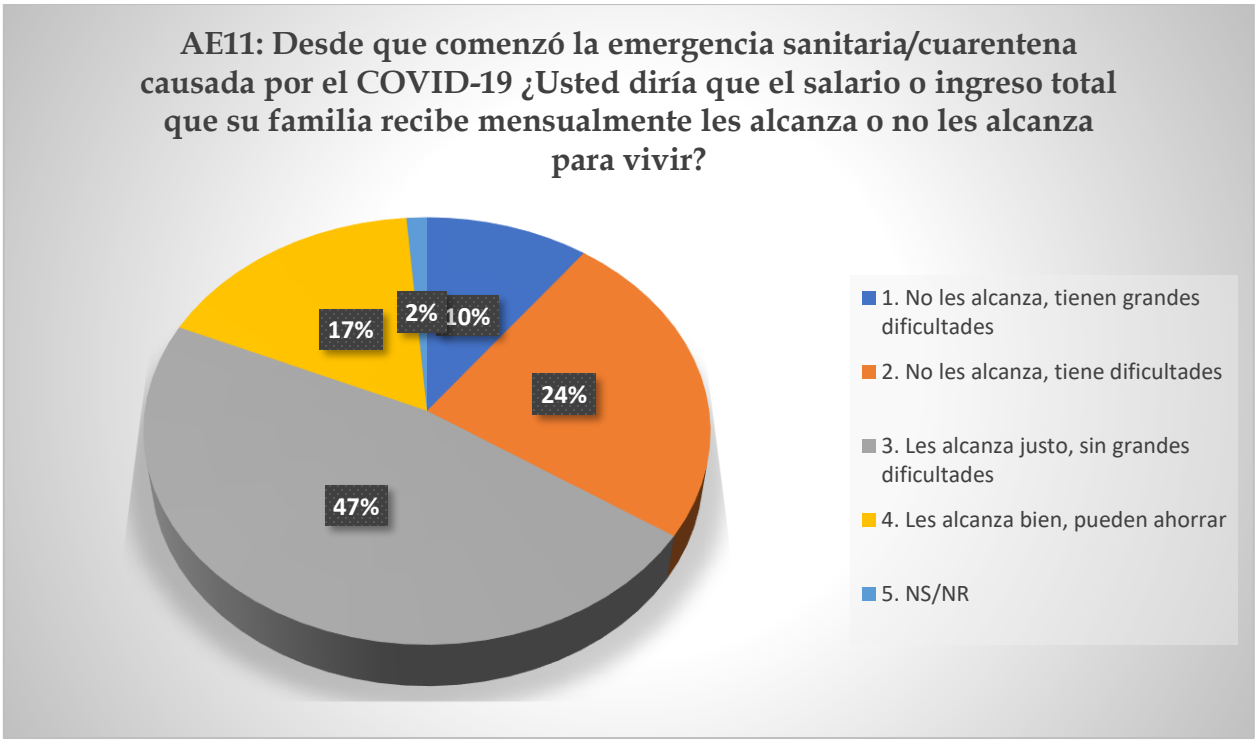
III.I. X. ÍTEM AE10

AE10: Además de usted ¿algún otro miembro de su hogar ha sufrido algún cambio en el ingreso económico?	
1. No hay cambio en el ingreso	112
2. Incremento de ingresos	8
3. Ingreso disminuido	104
4. NS/NR	11
TOTAL	235



III.I. XI. ÍTEM AE11

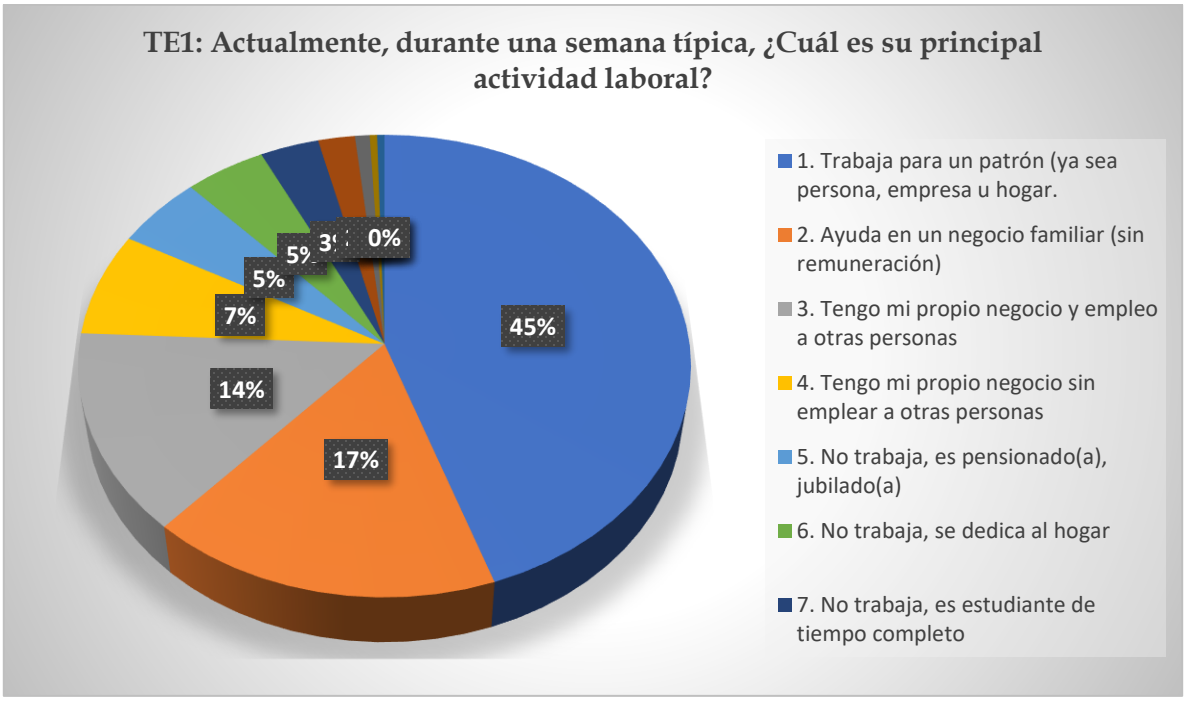
AE11: ¿Usted diría que el salario o ingreso total que su familia recibe mensualmente les alcanza o no les alcanza para vivir?	
1. No les alcanza, tienen grandes dificultades	24
2. No les alcanza, tiene dificultades	57
3. Les alcanza justo, sin grandes dificultades	111
4. Les alcanza bien, pueden ahorrar	40
5. NS/NR	3
TOTAL	235



III.II. MÓDULO TE (TELETRABAJO)

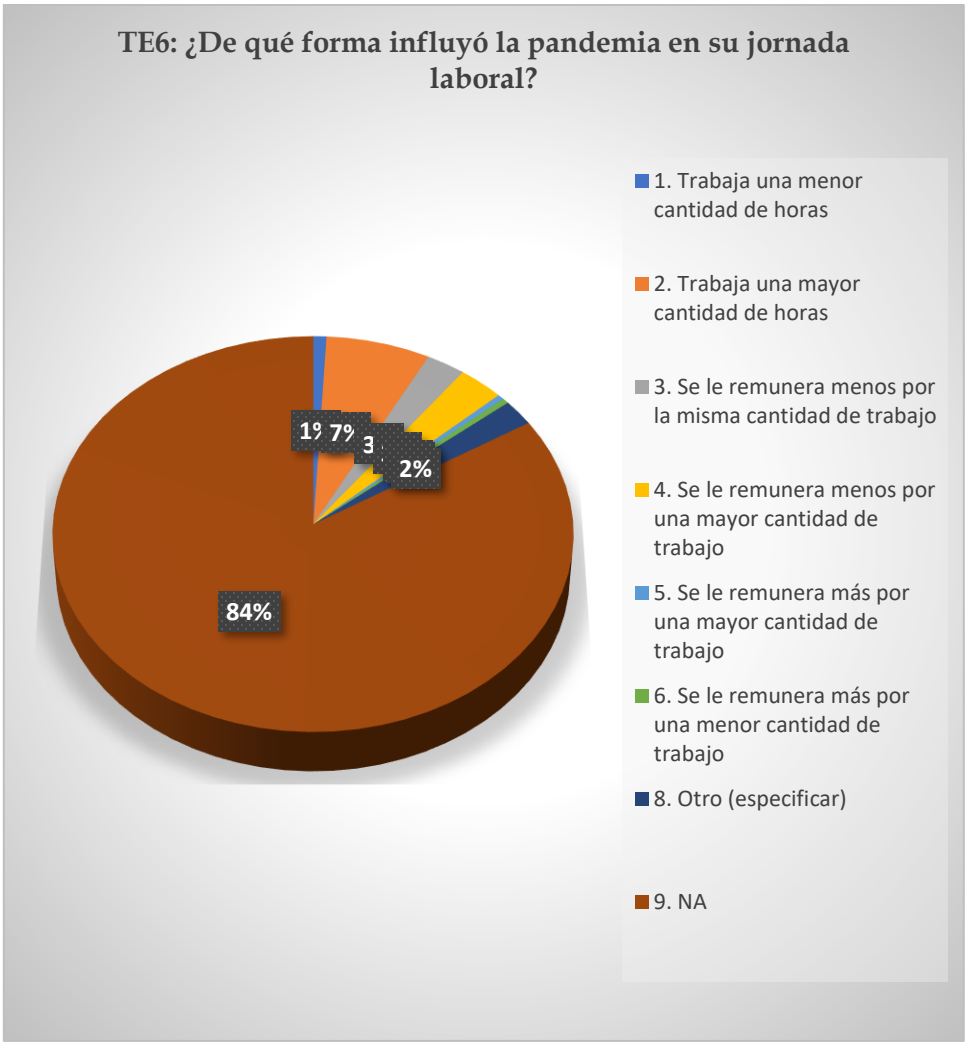
III.I. I. ÍTEM TE1

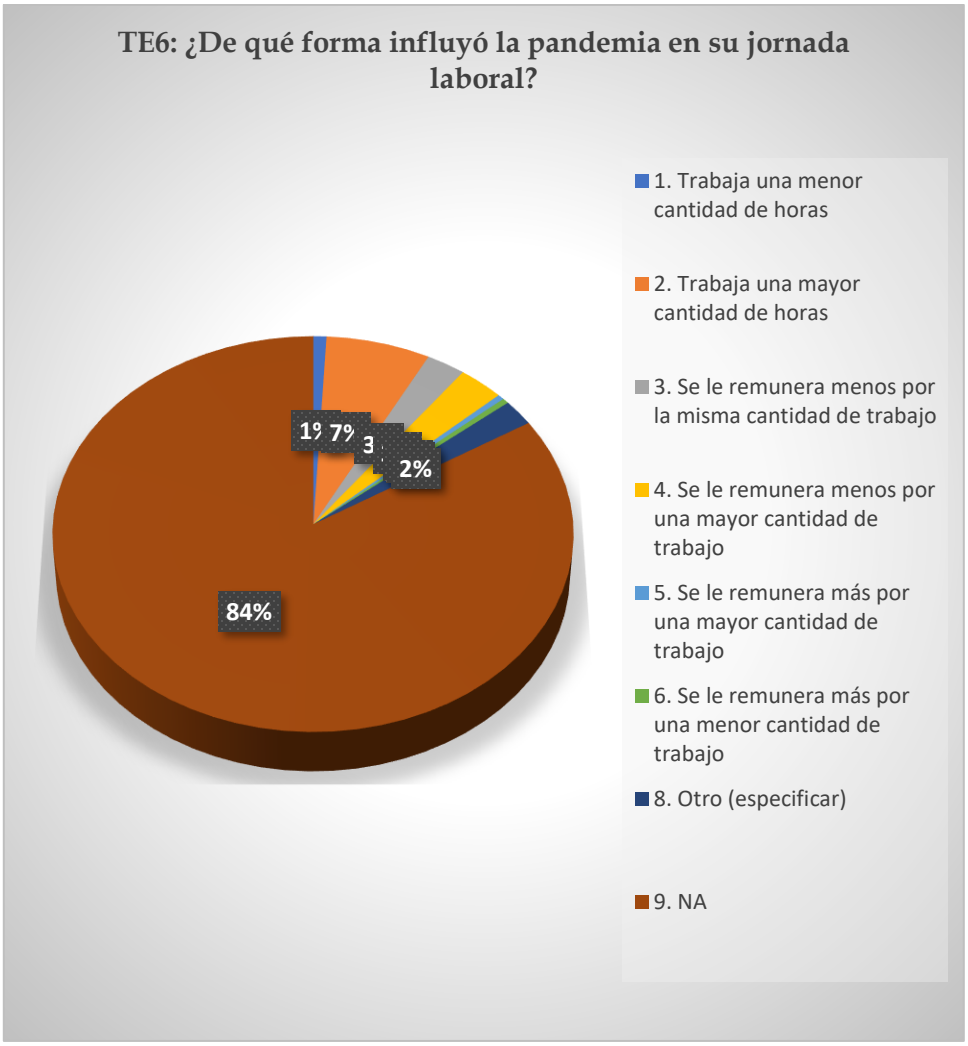
TE1: Actualmente, durante una semana típica, ¿Cuál es su principal actividad laboral?	
1. Trabaja para un patrón (ya sea persona, empresa u hogar.	105
2. Ayuda en un negocio familiar (sin remuneración)	40
3. Tengo mi propio negocio y empleo a otras personas	33
4. Tengo mi propio negocio sin emplear a otras personas	17
5. No trabaja, es pensionado(a), jubilado(a)	12
6. No trabaja, se dedica al hogar	11
7. No trabaja, es estudiante de tiempo completo	8
8. No trabaja, por una limitación física/mental que me impide trabajar	5
9. No trabaja, no busca trabajo y no está disponible para trabajar	2
10. No trabaja, pero está buscando trabajo	1
88. Otra	1
99. NS/NR	0
TOTAL	235



III.I. II. ÍTEM TE6

TE6: ¿De qué forma influyó la pandemia en su jornada laboral?	
1. Trabaja una menor cantidad de horas	2
2. Trabaja una mayor cantidad de horas	16
3. Se le remunera menos por la misma cantidad de trabajo	6
4. Se le remunera menos por una mayor cantidad de trabajo	7
5. Se le remunera más por una mayor cantidad de trabajo	1
6. Se le remunera más por una menor cantidad de trabajo	1
8. Otro (especificar)	5
9. NA	197
TOTAL	235

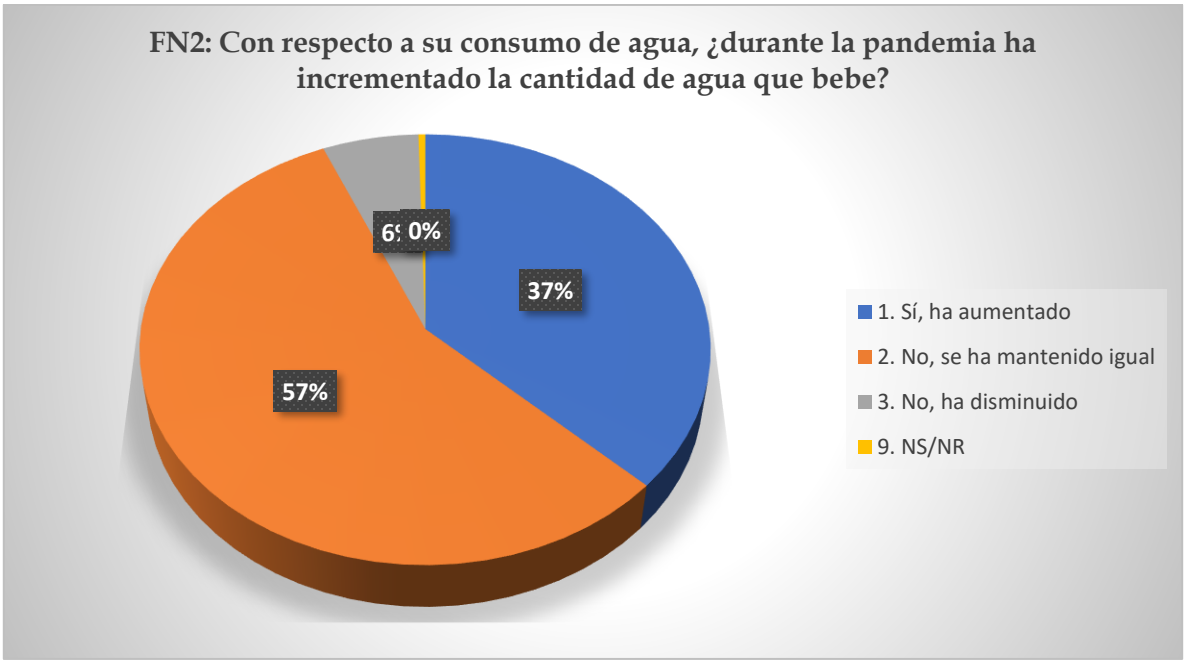

TE6: ¿De qué forma influyó la pandemia en su jornada laboral?
1%
7%
3%
2%
84%
1. Trabaja una menor cantidad de horas
2. Trabaja una mayor cantidad de horas
3. Se le remunera menos por la misma cantidad de trabajo
4. Se le remunera menos por una mayor cantidad de trabajo
5. Se le remunera más por una mayor cantidad de trabajo
6. Se le remunera más por una menor cantidad de trabajo
8. Otro (especificar)
9. NA



III.III. MÓDULO FN (HÁBITOS DE SALUD FÍSICA Y NUTRICIÓN)

III.III. I. ÍTEM FN2

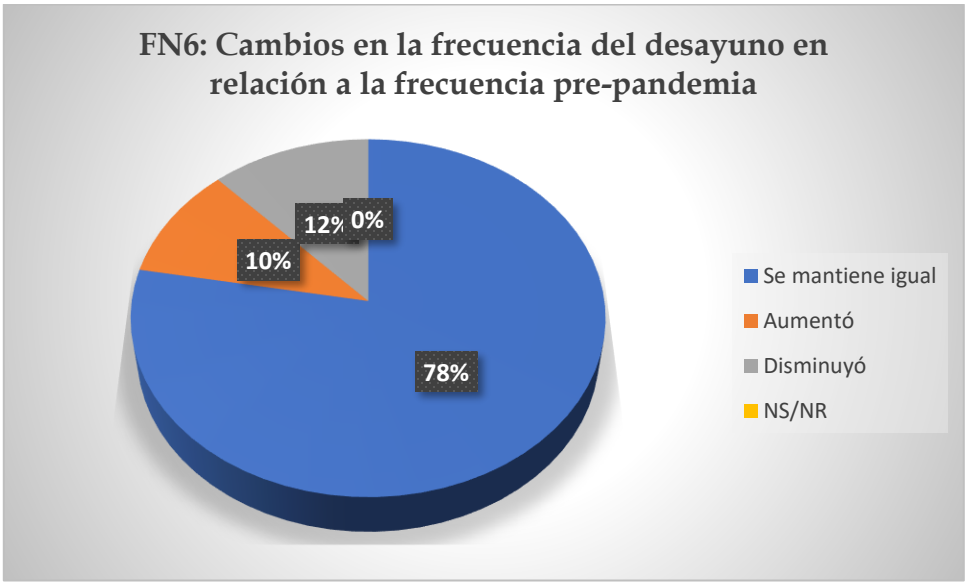
FN2: Con respecto a su consumo de agua, ¿durante la pandemia ha incrementado la cantidad de agua que bebe?	
1. Sí, ha aumentado	87
2. No, se ha mantenido igual	133
3. No, ha disminuido	14
9. NS/NR	1
TOTAL	235



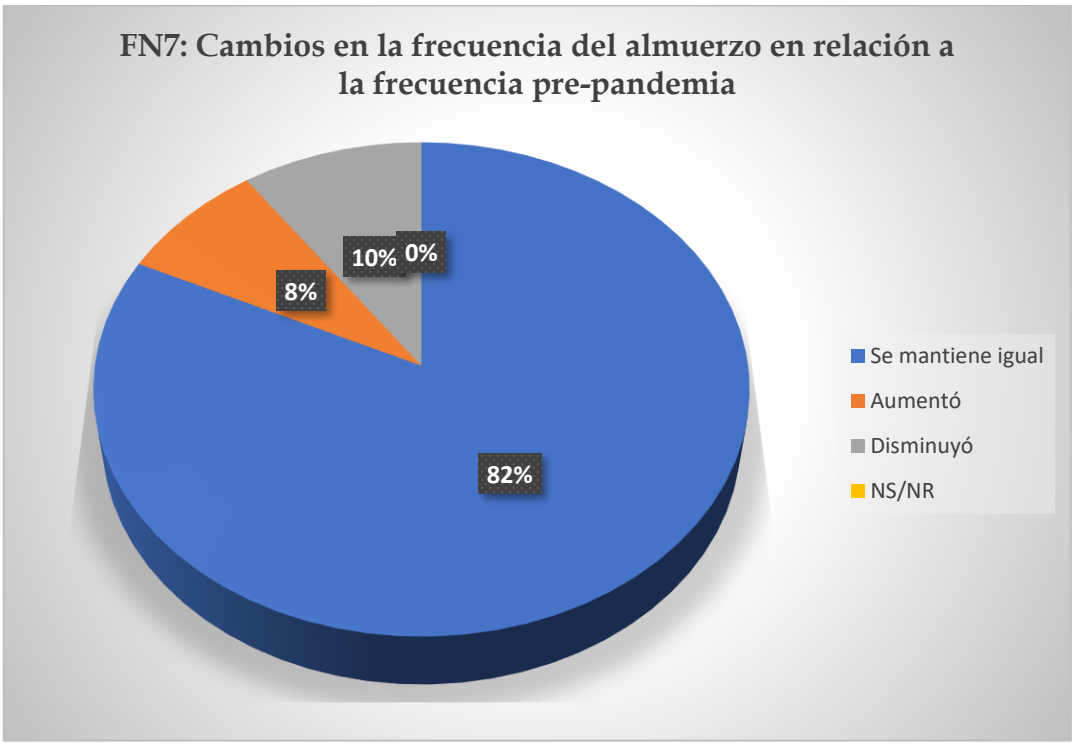
III.III. II. ÍTEMS FN6, FN7, FN8

III.III. II. I. Ítem FN6

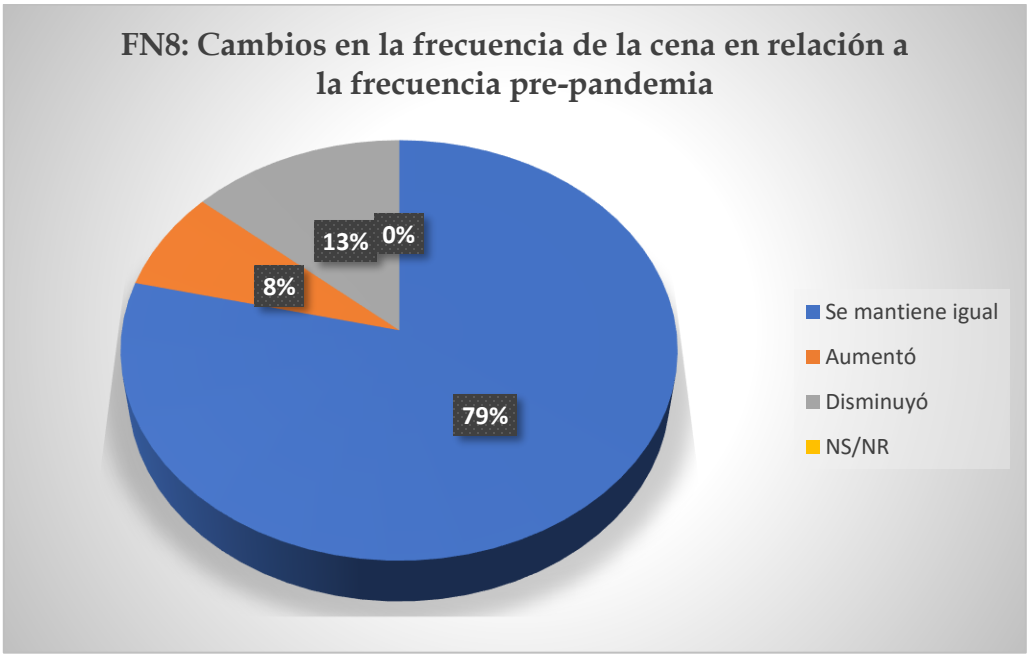
FN6-8: Frecuencia de tiempos de comida pre-pandemia y post-pandemia	Se mantiene igual	Aumentó	Disminuyó	NS/NR	TOTAL
FN6: Desayuno	183	24	28	0	235
FN7: Almuerzo	193	19	23	0	235
FN8: Cena	185	18	32	0	235



III.III. II. II. Ítem FN7



III.III. II. III. Ítem FN8

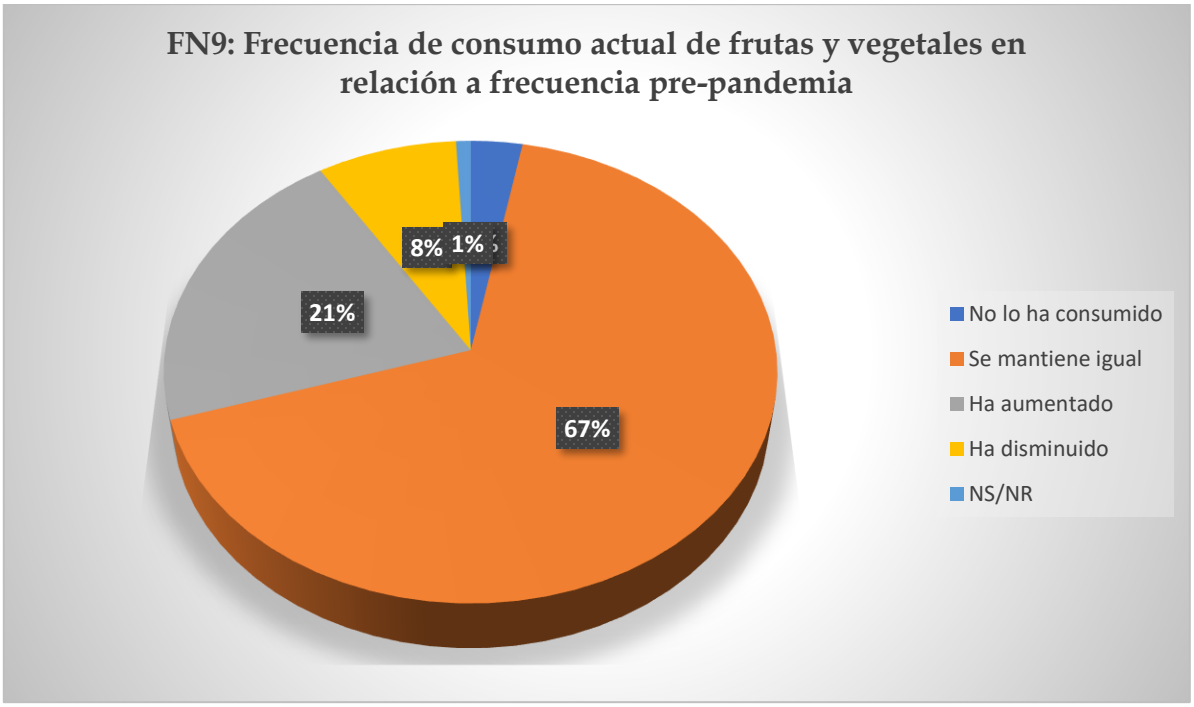


III.III. II. ÍTEMS FN9, FN10, FN11, FN12, FN13, FN14, FN15

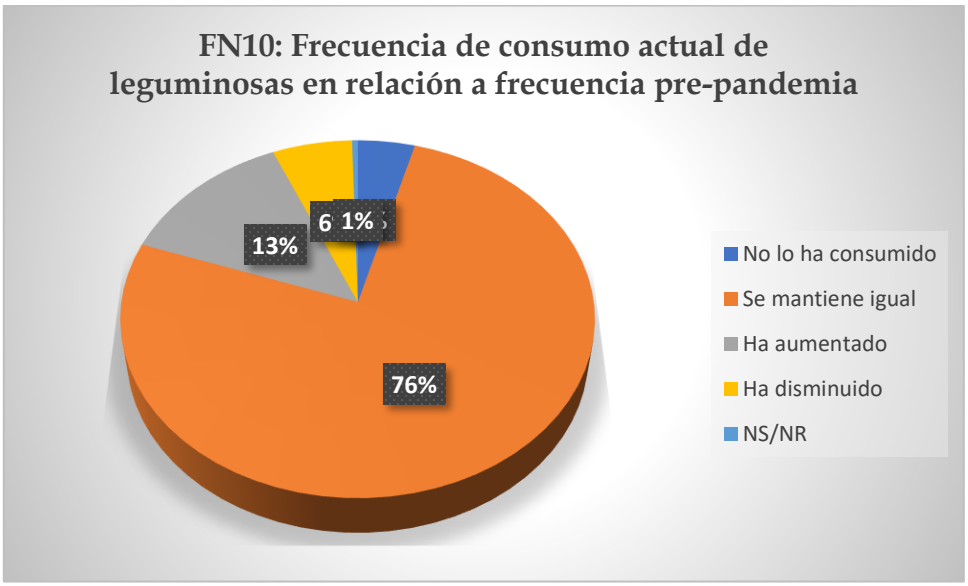
FN9, FN10, FN11, FN12, FN13, FN14, FN15: Comparando la frecuencia actual con respecto a sus hábitos el año previo a la pandemia, ha aumentado, disminuido o se mantiene igual el consumo de:

	No consume	Invariable	Aumentó	Disminuyó	NS/NR	TOTAL
FN9: Frutas y vegetales	7	158	49	19	2	235
FN10: Leguminosas (frijoles, garbanzos, lentejas)	10	179	31	14	1	235
FN11: Lácteos (leche, yogurt, queso)	12	166	26	30	1	235
FN12: Harinas (arroz, pasta)	2	176	28	28	1	235
FN13: Carnes o huevo (res, cerdo, pollo, pescado, embutidos)	1	170	38	25	1	235
FN14: Alimentos y bebidas fuentes de azúcar	42	109	27	56	1	235
FN15: Grasas y comidas rápidas	33	115	23	63	1	235

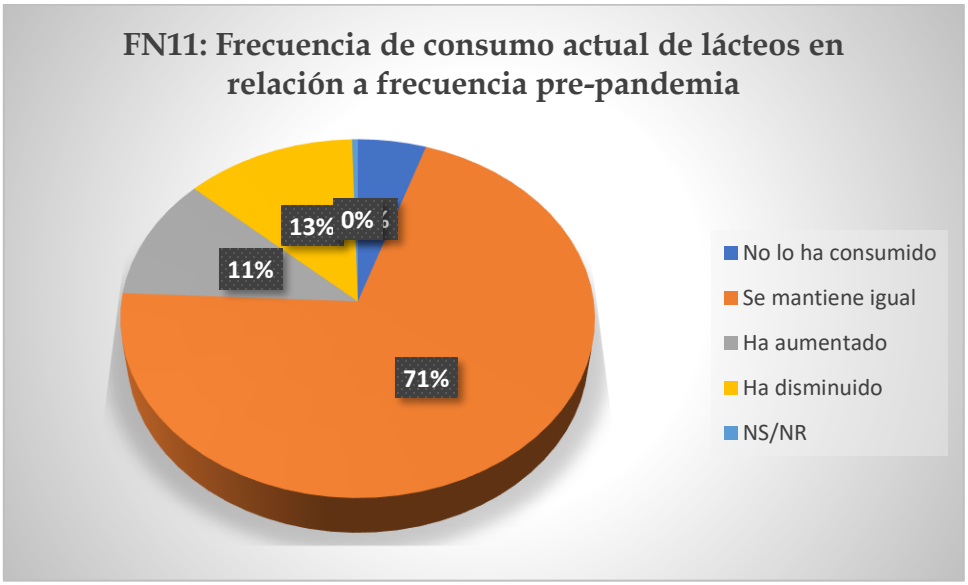
III.III. II. I. Ítem FN9



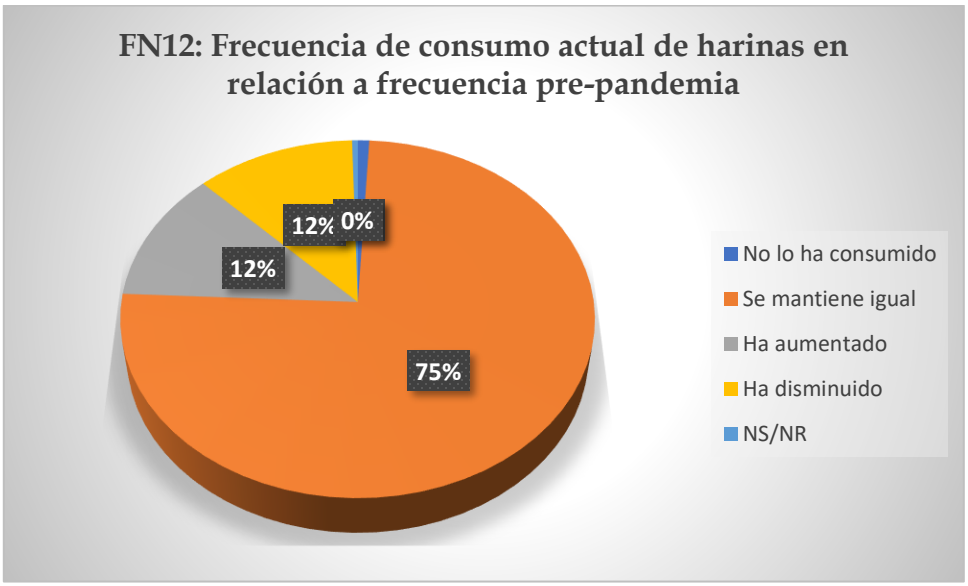
III.III. II. II. Ítem FN10



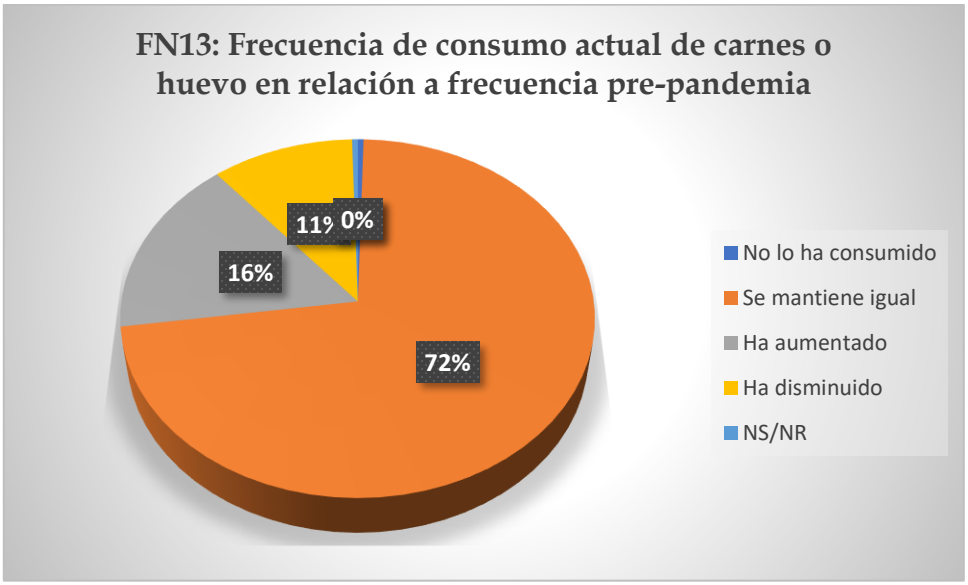
III.III. II. III. Ítem FN11



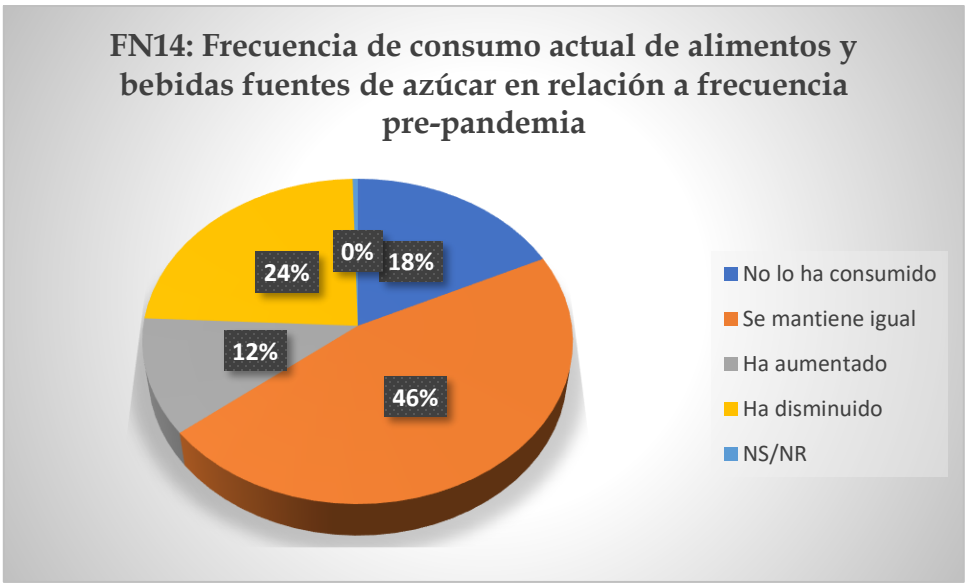
III.III. II. IV. Ítem FN12



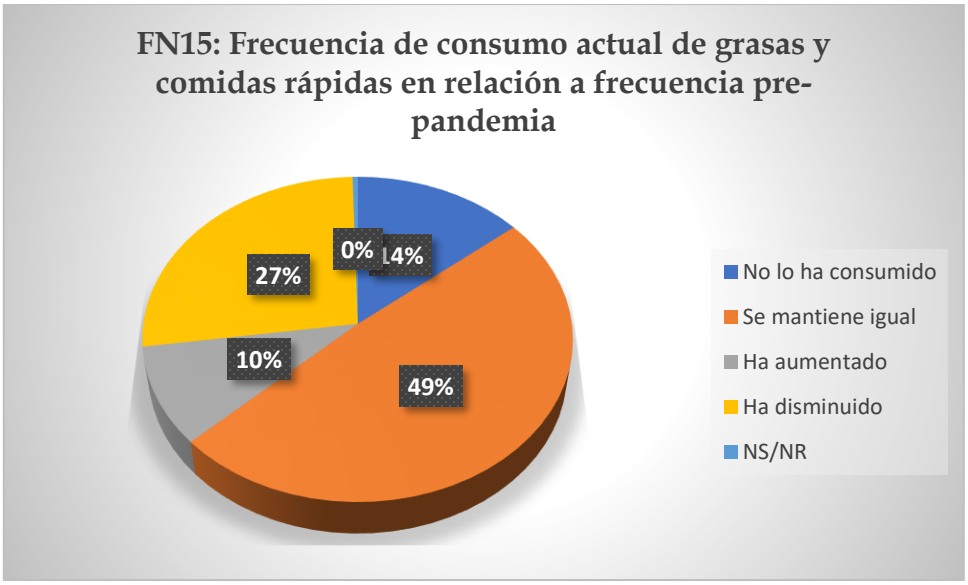
III.III. II. V. Ítem FN13



III.III. II. VI. Ítem FN14



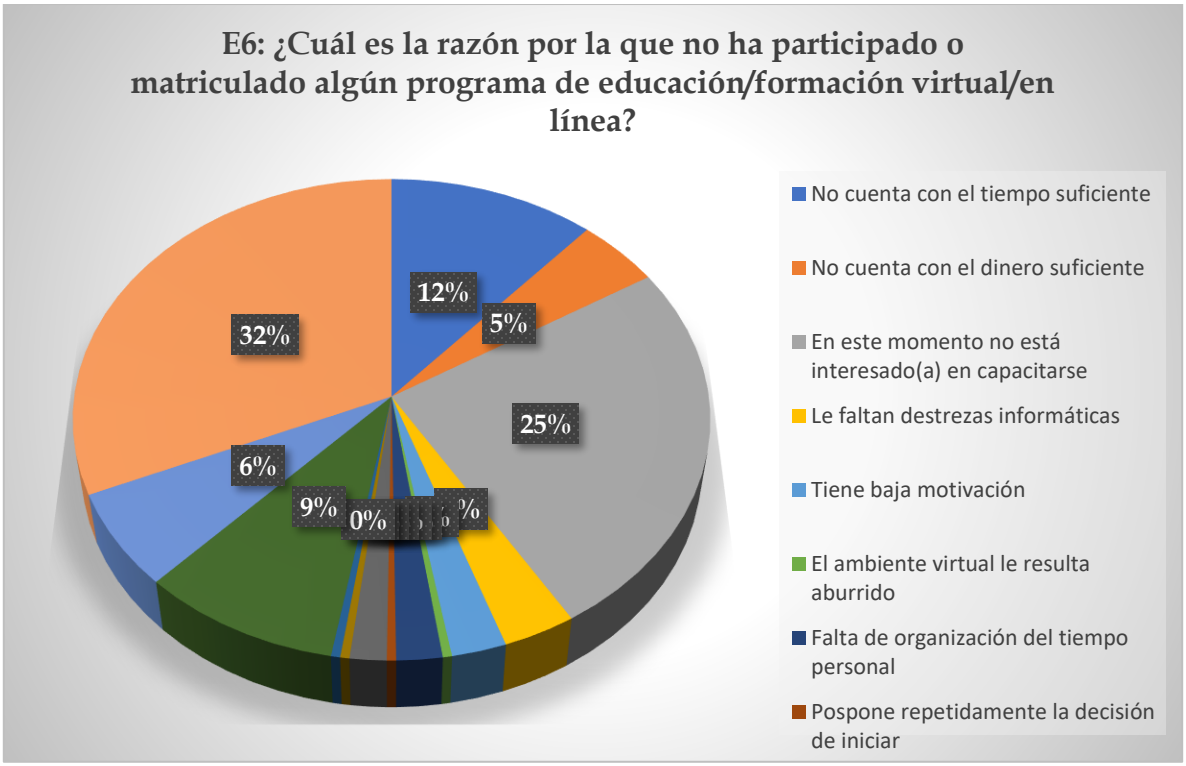
III.III. II. VII. Ítem FN15



III.IV. MÓDULO E (EDUCACIÓN VIRTUAL)

III.IV. I. ÍTEM E6

E6: ¿Cuál es la razón por la que no ha participado o matriculado algún programa de educación/formación virtual/en línea?	
No cuenta con el tiempo suficiente	27
No cuenta con el dinero suficiente	11
En este momento no está interesado(a) en capacitarse	59
Le faltan destrezas informáticas	8
Tiene baja motivación	6
El ambiente virtual le resulta aburrido	1
Falta de organización del tiempo personal	5
Pospone repetidamente la decisión de iniciar	1
El ambiente virtual le resulta incómodo	4
No desea iniciar solo(a)	1
La educación en línea es muy cara	1
Otro:	22
NS/NR	15
NA	74
TOTAL	235



IV. DISEÑO GENERAL DE LA METODOLOGÍA INFERENCIAL Y PRESENTACIÓN DE RESULTADOS

IV.I. INFORMACIÓN A EXTRAER DE LOS RESULTADOS DE LA ENCUESTA

1. Edad.
2. Acceso a internet fijo.
3. Localización geográfica (como urbanidad o ruralidad).
4. Acceso a internet móvil.
5. Nacionalidad.
6. Proporción de personas sin acceso a educación virtual, con énfasis en la frecuencia relativa de la razón económica de esta falta de acceso.
7. Relación con la riqueza social.
8. Variación negativa en la dieta, hidratación y reposo durante la pandemia.
9. Variación negativa en los ingresos del hogar.
10. Sexo.
11. Estrés económico mensual.

La finalidad de que muchas de las variables sean planteadas de forma negativa (como contracción) obedece a que también se desea utilizar modelos estadísticos para variables dicotómicas y para ello es necesario transformar las respuestas a los ítems del cuestionario en variables de respuesta binaria, así como también al hecho de que de antemano se conoce que la distribución sociopolítica del ingreso y la calidad de vida ha empeorado a nivel planetario para la población en general.

IV.II. MECANISMO DE OBTENCIÓN DE LA INFORMACIÓN A TRAVÉS DE LOS ÍTEMS EXTRAÍDOS DEL CUESTIONARIO

1. Edad: usar SD2.
2. Localización geográfica (urbana o rural): usar SD13.
3. Nacionalidad (Nacional o extranjero): SD3.
4. Relación con la riqueza social: si poseen medios de producción o no. Esto se determina si en el ítem TE1 responden con las opciones 2 o 3, porque 4 lo hace un productor directo (no un capitalista) y las demás lo sitúan en alguna forma de proletariado.
5. Proporción de personas sin acceso a educación virtual: calcular la relación (como ratio) entre las respuestas en E6 y todos los casos en que E6 no aplicaba.

6. Variación negativa (disminución) o no, *i.e.*, contracción o no, del flujo mensual de ingresos del hogar: si responden de AE3 a AE9 (en al menos una de ellas) que sí, mientras que las demás se mantienen constantes.
7. Sexo: usar SD1.
8. Acceso a internet fijo: usar TC1.
9. Empeoramiento o no (mejoramiento o se mantuvo igual) de la dieta, la hidratación y el reposo durante la pandemia:

9.1. Empeoramiento o no de la alimentación (enfoque cuantitativo): si responden de FN6 A FN8 de tal forma que denote que la frecuencia de alguno de los tiempos de comida varió negativamente y los otros se mantienen igual; o bien, en caso varían para el mismo sujeto los tres tiempos, si dos de ellos variaron negativamente.

9.2. Empeoramiento o no de la alimentación (enfoque cualitativo): si en FN9, FN10, FN11 y/o FN13 ("y/o" denota que la condición a especificar debe ocurrir en al menos en uno de los ítems especificados) que disminuyó, mientras que FN14 y/o FN15 responden que mejoró o se mantuvo igual, *i.e.*, que en al menos uno de los dos ítems se respondió que aumentó, o se respondió que ambos aumentaron o, en su defecto, se respondió que ambos permanecieron iguales[4].

10. Empeoramiento o no del reposo: usar FN17.
11. Empeoramiento o no de la hidratación: usar FN2.
12. Estrés económico mensual: usar AE11.

[4] Por supuesto, hay muchos grises más allá de este "blanco y negro" metodológico mediante el cual se está planteando abordar los cambios cualitativos en la dieta; sin embargo, de esta forma se minimiza a niveles aceptables (por criterios eminentemente lógicos, aunque no por ello necesariamente menos válidos gnoseológicamente) el error posible. Esto es así porque la sección FN del cuestionario no fue diseñada con la finalidad de medir cambios en la dieta con el nivel de precisión suficiente para que sea posible distinguir con mayor nivel de precisión los cambios cuantitativos y cualitativos en la dieta y ello obedeció a que el objetivo de su diseño no se limitaba exclusivamente a realizar tal medición. Así, queda en evidencia que la génesis gnoseológica de un instrumento impone restricciones metodológicas infranqueables.

IV.III. VARIABLES DICOTÓMICAS (VD) A GENERAR

- VD1: se es nacional o extranjero.
- VD2: se posee capital o no.
- VD3: Empeoró (o no) la dieta (en todas las variedades especificadas en la sección III) en el contexto de la pandemia.
- VD4: Sexo.
- VD5: Acceso (o no) a la educación virtual.
- VD6: Ruralidad o no ruralidad.
- VD7: Adulto o no (menor de veinticinco años o no).
- VD8: Contracción en el flujo mensual de ingresos.

IV.IV. ANÁLISIS COMBINADO DE LA INFORMACIÓN OBTENIDA A TRAVÉS DE LOS ÍTEMS

a. Mediante la combinación de 2, 6, 3, y 7 es posible generar un modelo estadístico para determinar cómo los cambios en el ingreso (en el contexto de pandemia) afectan a las personas según edad, sexo, nacionalidad (nacional o extranjero) y localización geográfica (rural o urbano).
b. Mediante la combinación de 4, 2, 3 y 1 es posible generar un modelo estadístico para determinar cómo la exclusión de la riqueza social varía en función de la edad, sexo, nacionalidad (nacional o extranjero) y localización geográfica (urbano o rural).
c. Con 6 se puede determinar cómo el factor económico excluirá a las personas de las nuevas dinámicas que involucrarán la virtualidad, a pesar que necesitarán incorporarse a ella para mejorar sus opciones de ingreso, lo cual es toda una paradoja.
d. Determinar, fijándose ex ante un umbral de pobreza (expuesto en los anexos), la cantidad de personas que están por debajo de la línea de pobreza.

IV.V. METODOLOGÍAS ESTADÍSTICAS DE ESTUDIO

VI.V. I. ANÁLISIS DE CORRELACIÓN

VI.I.I. Primer Análisis de Correlación

- Tenencia de capital contra condición migratoria.
- Tenencia de capital contra sexo
- Tenencia de capital contra localización sociodemográfica.
- Tenencia de capital contra edad.

VI.I. II. Segundo Análisis de Correlación

- Contracción de ingresos contra acceso a educación virtual.
- Contracción de ingresos contra sexo.

- Contracción de ingreso contra edad.
- Contracción de ingresos contra condición migratoria.
- Contracción de ingresos contra localización sociodemográfica.

VI.V. II. TIPOS DE MODELOS ESTADÍSTICOS A UTILIZAR

a. Modelo Probit.
b. Modelo Tobit.

VI.V. III. COMBINACIONES DE VARIABLES A MODELAR

VI.I.I. Primera Combinación (PC)

$$VD2 = f(VD4, VD7, VD1, VD6)$$

$$VD2 = Tenencia\ de\ Capital$$

$$VD4 = Sexo$$

$$VD1 = Condición\ Migratoria$$

$$VD6 = Ruralidad\ o\ no$$

$$VD7 = Adulto\ o\ no$$

VI.I. II. Segunda Combinación (SC)

$$VD8 = f(VD4, VD7, VD1, VD6, VD5)$$

$$VD8 = Contracción\ de\ ingresos$$

$$VD4 = Sexo$$

$$VD1 = Condición\ Migratoria$$

$$VD6 = Ruralidad\ o\ no$$

$$VD7 = Adulto\ o\ no$$

$$VD5 = Acceso\ a\ educación\ virtual$$

IV.VI. FUNDAMENTO ESTADÍSTICO-MATEMÁTICO DEL DISEÑO METODOLÓGICO GENERAL

IV.VI. I. JUSTIFICACIÓN DEL USO DE LA METODOLOGÍA DE LA ESTADÍSTICA MATEMÁTICA PARA LA OBTENCIÓN DE BUENAS ESTIMACIONES

Como señala (Cochran, 1991, pág. 195), "Uno de los rasgos de la estadística teórica es la creación de una vasta teoría que discute los métodos de obtención de buenas estimaciones a partir de los datos. En el desarrollo de la teoría, específicamente para encuestas de muestreo, se han utilizado poco estos conocimientos, por dos causas principales. Primero, en las encuestas que contienen un gran número de atributos, es una gran ventaja, aunque se disponga de máquinas computadoras, el

poder utilizar procedimientos de estimación que requieran poco más que simples sumas, en tanto que los métodos superiores de estimación de la estadística teórica, como lo son la máxima verosimilitud, podrían necesitar una serie de aproximaciones sucesivas antes de encontrar una estimación (…) La mayoría de los métodos de investigación de la estadística teórica suponen que se conoce la forma funcional de la distribución de frecuencia que sigue a los datos de la muestra, y el método de estimación de estimación está cuidadosamente engranado de acuerdo a este tipo de distribución. En la teoría de encuestas por muestreo se ha preferido hacer, cuando más, algunos supuestos respecto a esta distribución de frecuencia. Esta actitud resulta razonable para tratar con encuestas en las que el tipo de distribución puede variar de un atributo a otro, y cuando no deseamos detenernos a examinarlas todas, antes de decidir cómo hacer cada estimación. En consecuencia, actualmente, las técnicas de estimación para el trabajo de encuestas por muestreo son de alcances restringidos. Ahora consideraremos dos técnicas, el método de razón (…) y el método de regresión línea (…)" Así, "Al igual que la estimación de razón, la regresión lineal se ha diseñado para incrementar la precisión en el uso de una variable auxiliar x_i correlacionada con y_i." (Cochran, 1991, pág. 239).

IV.VI. II. ESTIMACIÓN DEL TAMAÑO DE LA MUESTRA PARA MINIMIZAR EL IMPACTO DE LAS NO-RESPUESTAS EN LA SIGNIFICANCIA ESTADÍSTICA Y EL MARGEN DE ERROR

Como se señala en (Lohr, Sampling: Design and Analysis, 2019, pág. 46), un investigador a menudo mide varias variables y tiene varios objetivos para una encuesta. Cualquiera que diseñe un muestro aleatorio simple (SRS de ahora en adelante, por sus siglas en inglés) debe decidir qué cantidad de error de muestreo en las estimaciones es tolerable y debe equilibrar la precisión de las estimaciones con el costo de la encuesta. Aunque se pueden medir muchas variables, un investigador a menudo puede centrarse en una o dos respuestas que son de interés principal en la encuesta y utilizarlas para estimar el tamaño de la muestra.

Así, el error tolerable en la encuesta está íntimamente relacionado con la especificación del nivel de precisión estadística que se desea tenga la encuesta a través de las probabilidades. Como se verifica en (Lohr, Sampling: Design and Analysis, 2019, pág. 46), el vínculo anterior puede expresarse de la siguiente forma general:

$$\boldsymbol{P}(|\overline{\boldsymbol{y}} - \overline{\boldsymbol{y}}_{\boldsymbol{U}}| \leq \boldsymbol{e}) = \mathbf{1} - \boldsymbol{\alpha}$$

En la expresión anterior, $\overline{\boldsymbol{y}}$ es la estimación realizada de la variable $\boldsymbol{y}$, $\overline{\boldsymbol{y}}_{\boldsymbol{U}}$ son sus valores observados, $\boldsymbol{e}$ es el margen de error de la encuesta, $\mathbf{1}$ es la probabilidad total y $\boldsymbol{\alpha}$ es el nivel de confianza estadística.

(Cochran, 1991, pág. 439) proporciona, considerando la configuración del sesgo menos favorable, una fórmula para determinar un valor de $\boldsymbol{n}$ que garantice que el error obtenido a causa del sesgo generado por las no-respuestas sea menor a un determinado error $\boldsymbol{d}$ que el investigador considere técnicamente aceptable/asumible.

$$\boldsymbol{n = \frac{t_{\alpha}^{2}}{4d(d - W_{2})W_{1}} - 1}$$

En la expresión anterior, como el lector puede verificar en (Cochran, 1991, pág. 439), $\boldsymbol{n}$ es el número es el tamaño de muestra requerido para los fines especificados, $\boldsymbol{t_{\alpha}^{2}}$ (en la investigación fundacional de Birnbaum y Sirken aparece como $\boldsymbol{T_{\alpha}^{2}}$) es el desvío normal correspondiente al riesgo $\boldsymbol{\alpha}$ (nivel de significancia estadística) de que el error exceda $\boldsymbol{d}$ magnitud, $\boldsymbol{d}$ es el margen de error aceptado/asumido por el investigador, $\boldsymbol{W_{1}}$ es la proporción que las respuestas representan del total de llamadas/visitas realizadas y $\boldsymbol{W_{2}}$ es la proporción que las no-respuestas representan del total de llamadas/visitas realizadas. Nótese que no existe valor de $\boldsymbol{n}$ que satisfaga si $\boldsymbol{W_{2} > d}$. Si $\boldsymbol{W_{2} = 0}$, esta ecuación se reduce a $\boldsymbol{n = \frac{T_{\alpha}^{2}}{4d^{2}} - 1}$.

En el marco de la fórmula anterior, la cual es expuesta por Cochran en el lugar referido no especifica en el lugar referido cómo determinar el valor $\boldsymbol{t_{\alpha}^{2}}$; sin embargo, Cochran retoma el trabajo fundacional de Birnbaum y Sirken, quienes señalan en (Birnbaum & Sirken, 1950, pág. 104), quienes el valor en cuestión es aquel que satisface la siguiente identidad:

$$\boldsymbol{\frac{2}{\sqrt{2\pi}} \int_{0}^{t_{\alpha}} e^{-\frac{t^{2}}{2}} dt = 1 - \alpha}$$

El lector inmediatamente habrá notado que la expresión anterior no es más que la fórmula para encontrar los valores de la variable de estandarización $\boldsymbol{Z}$ en la distribución de probabilidad normal estandarizada, cuyo valor $\boldsymbol{Z}$ es igual a 1.96 cuando $\boldsymbol{\alpha = 0.05}$.

En el caso de la encuesta realizada en esta investigación, la tasa de respuestas $\boldsymbol{W_{1} = 0.699454}$ fue porcentualmente de 69.9454%, la tasa de no-respuestas $\boldsymbol{W_{2} = 0.300546}$ fue porcentualmente de 30.0546% y el margen de error preestablecido en términos del número de encuestas realizadas $\boldsymbol{d = 0.0639}$ fue porcentualmente de 6.39%.

$$\boldsymbol{n = \frac{1.96^{2}}{4 * 0.0639(0.0639 - 0.300546)0.699454} - 1 = -91.8015}$$

Como puede observarse, la metodología expuesta por (Birnbaum & Sirken, 1950, pág. 104) y planteada por (Cochran, 1991, pág. 439) arroja, a causa de su misma estructura matemática un valor de $\boldsymbol{n}$ negativo cuando el margen de error es menor que la tasa de no-respuestas. Puesto que, a priori, la situación que genera la negatividad de $\boldsymbol{n}$ a priori no parece ser una singularidad de las prácticas estadísticas cotidianas, se propone aquí que se interprete empíricamente el componente $\frac{t_\alpha^2}{4d(d-W_2)W_1}$ en términos de su valor absoluto, tal y como se hace, por ejemplo, en los análisis neoclásicos de demanda del consumidor con el caso de las elasticidades.

$$\boldsymbol{n = \left| \frac{1.96^2}{4 * 0.0639(0.0639 - 0.300546)0.699454} \right| - 1 = 89.8015}$$

Así, según el índice de Birnbaum & Sirken (de ahora en adelante, IBS), el número de encuestas requerido con un margen de error de 6.39%, a un nivel de confianza del 95% en ese margen de error, sería de aproximadamente de 89 encuestas, un número ampliamente inferior a las 235 encuestas realizadas. Algunos valores de $\boldsymbol{n}$ dados por el método de Birnbaum y Sirken en su investigación pionera son mostrados por Cochran y presentados a continuación.

TABLA 13.3. VALOR MÍNIMO DE n PARA LÍMITE DADO DE ERROR d, CON RIESGO $\alpha = 0.05$

% No-respuesta $100W_2$	d (%) 20	15	10	5
0	24	43	96	384
2	27	50	122	653
4	31	60	166	2000
6	36	75	255	....
8	43	99	521	....
10	53	142	...	....
15	112	...	...	....

Fuente: (Cochran, 1991, pág. 440).

"La tabla anterior cuenta la misma triste historia (…) Si nos contentamos con una estimación imprecisa ($\boldsymbol{d = 20}$), cantidades de no-respuestas hasta de 10% se pueden manejar duplicando el tamaño de la muestra. Sin embargo, cualquier porcentaje considerable de no-respuesta hace imposible o muy costoso alcanzar una precisión altamente garantizada mediante el aumento del tamaño de la muestra entre los entrevistados." (Cochran, 1991, pág. 440).

Sin embargo, como el lector podrá observar, aunque al agregar valor absoluto la interpretación tomada aisladamente tenga lógica, no parece tenerla al contrastarla con la tabla antes expuesta. El lector debe tomar en cuenta que el IBS vio luz en 1950 (hace 71 años), mientras que a la obra de Cochran (específicamente la tercera edición de la versión en inglés) le corresponde el año 1977 (hace 44 años) y mucho ha evolucionado la *praxis estadística*[5] desde entonces. Por ello, conviene emplear metodologías estadísticas más recientes[6] como la presentada por (Lohr, Sampling: Design and Analysis, 2019, pág. 47), cuando el tamaño de la población es desconocido (como es el caso de esta investigación).

$$\boldsymbol{n = \frac{Z_{\frac{\alpha}{2}}^2 * W_1 * W_2}{d^2} = \frac{Z_{\frac{\alpha}{2}}^2 * S^2}{d^2}, siempre\ que\ W_1 * W_2 = S^2}$$

Retomando lo señalado también por (Lohr, Sampling: Design and Analysis, 2019, pág. 47), "En encuestas en las que una de las principales respuestas de interés es una proporción, a menudo es más fácil utilizar esa respuesta para establecer el tamaño de la muestra. Para poblaciones grandes, $\boldsymbol{S^2 \approx p\,(1-p)}$, que alcanza su valor máximo cuando $\boldsymbol{p = \frac{1}{2}}$."

Lo que Lohr implica en la afirmación anterior es una consecuencia teórica de uno de los teoremas más importantes de la Estadística Matemática en particular y del Análisis Real en general, el cual se presenta a continuación en su versión débil.

$$\boldsymbol{\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx}$$

Lo que la expresión anterior indica, conocida como *teorema central del límite* y en la que $\boldsymbol{\Phi(z)}$ es una función de densidad de probabilidad estandarizada, es que la probabilidad de ocurrencia de un evento cualquiera acotado superiormente (*i.e.*, que existe un evento que ocurre al mismo tiempo o posterior al evento de análisis) tenderá hacia la probabilidad de ocurrencia promedio del conjunto de eventos a medida el número de eventos ocurridos tienda a infinito o, para los fines de esta investigación, a medida el tamaño de la muestra sea lo suficientemente grande, lo que a su vez es equivalente a afirmar que lo anterior se cumplirá a medida que el tamaño de la muestra se acerque "lo suficiente" al tamaño de la población; y, como es usual, "lo suficiente" es una cualidad cuya cantidad dependerá del fenómeno concreto que se estudie.

---

[5] La comunión entre teoría y práctica empírica.

[6] Aunque con la finalidad de proporcionar homogeneidad expositiva, se ha mantenido la notación empleada por Cochran en toda la presente sección.

Así, asumiéndose el teorema central del límite, a medida el tamaño de la muestra se acerca al de la población, la distribución de las respuestas tenderá a normalizarse, en ausencia de algún sesgo sistemático; esto puede verificarse realizando un ajuste de distribución sobre el conjunto de datos en conjunto con las pruebas de normalidad que correspondan. Si lo anterior se verifica, el patrón estadístico al que responderá el conjunto de datos se expresa en términos geométricos a través de una distribución espacial del conjunto de datos perfectamente simétrica en relación al valor promedio de este conjunto. Esto significa que se localizarán exactamente la misma cantidad de observaciones a la derecha y a la izquierda de la distribución de probabilidad del conjunto de datos, por lo que la proporción $\boldsymbol{W_1}$ deberá ser igual a la proporción $\boldsymbol{W_2}$ y puesto que $\boldsymbol{W_1 + W_2 = 1}$, entonces $\boldsymbol{W_1 = 0.5}$ y $\boldsymbol{W_2 = 0.5}$.

$$\boldsymbol{n = \frac{1.96^2 * 0.5 * 0.5}{0.0639} = 235.207}$$

Así, se verifica que el número de encuestas alcanzada por el equipo de encuestadores (235 de las 300 planificadas inicialmente) es, en ausencia de conocimiento del tamaño de la población, lo suficientemente grande para tener un nivel de confianza estadística del 95% en el margen de error establecido de 6.39%.

IV.VI. II. I. Tipos de No-Respuesta

IV.VI. II. I. I. Tipos de no-respuesta teóricos

Como se verifica en (Cochran, 1991, pág. 440), los tipos de no-respuesta existentes son los que se presentan a continuación:

1) *No-cubrimientos.* Falla en la localización o en la visita a algunas unidades de la muestra.
2) *Los no-en-casa.* Este grupo contiene personas que residen en el lugar, pero que se encuentran temporalmente fuera de casa.
3) *Incapaz de contestar.* El entrevistado puede no tener información requerida para ciertas preguntas o puede mostrarse poco inclinado a darla. Una redacción hábil del cuestionario y su prueba previa son una salvaguarda.
4) *Los "hueso duro".* Las personas que inexorablemente rechazan ser entrevistados, que están incapacitadas o que están fuera de casa durante todo el tiempo disponible para el trabajo de campo, son las que constituyen este sector. Esto representa una fuente de sesgo que persiste, sin importar cuánto esfuerzo se ponga en la perfección de las listas.

IV.VI. II. I. II. Tipos de no-respuesta en el levantamiento de encuestas realizado según el Manual para el Trabajo de Campo (MTC)

Las no-respuestas son parte de la clasificación especificada en el MTC para los diferentes tipos de resultados de llamar por teléfono a alguno de los números listados en el banco telefónico consta de diez componentes:

1) Teléfono ocupado.
2) Teléfono no responde.
3) Inactivo.
4) Realizada.
5) Pendiente.
6) No realizada por alguna razón.
7) Incompleta.
8) Rechazo.
9) Comercio.
10) Repetido.

Según el MTC[7], página 3, el criterio que sirve para clasificar los componentes anteriores es el siguiente:

1 Ocupado:	el celular marcado avisa que está ocupado.
2 No responde:	el celular timbra pero no responde o responde una contestadora.
3 Inactivo:	el celular da un mensaje de que el teléfono no pertenece a ningún abondado o da un aviso de que el celular está temporalmente suspendido.
4 Realizada:	entrevista completa.
5 Pendiente:	por algún motivo (está trabajando, caminando, almorzando, etc.), la persona a entrevistar no inició la entrevista y llegan al acuerdo de que se hará en otro momento.
6. Norexoa:	entrevista no realizada por otras razones. Puede ser que la persona no habla español, está enferma, tiene problemas por alguna discapacidad que le impide comunicarse apropiadamente, tiene mucha edad, etc.
7 Incompleta:	la entrevista inició y, por algún motivo, no pudo terminarse. Puede ser que la persona no quiso seguir por algún motivo (no quiere opinar, se aburrió, se cansó, etc.). Tendrá que determinarse en cada caso si la entrevista incompleta puede recuperarse o no, lo que dependerá del por qué quedó incompleta.
8 Rechazo:	la persona a entrevistar no quiso responder el cuestionario y lo comunica de manera explícita.
9 Comercio/Menor:	el celular marcado es utilizado con fines comerciales, pues así lo expresó la persona que respondió o así se deduce por la forma en que responde. Menores de edad no serán entrevistados.
10 Repetido:	el celular marcado ya fue llamado, pues la persona que responde así lo hace saber.

IV.VI. III. FACTOR DE EXPANSIÓN

Como se señala en (Instituto Nacional de Estadística y Censos de Costa Rica, 2021), "El factor de expansión es un ponderador que se aplica a cada unidad de estudio en la muestra para obtener una estimación poblacional, y se interpreta como la cantidad de unidades en la población que representa cada unidad en la muestra, ya sea vivienda, hogar o persona."

La definición conceptual anterior tiene su expresión matemática básica, según (Departamento Administrativo Nacional de Estadística, 2003, pág. 12) y en (Instituto Nacional de Estadística y Censos de la República Argentina, 2019, pág.

[7] Véase https://mega.nz/file/y1cxyYpC#l_pz0e3Ba8D6SjHMEKZrpqC8p6-IbV0R175Wncxrz7U.

8), en la multiplicación de las inversas de las probabilidades de inclusión de cada una de las etapas de selección definidas por las necesidades concretas de la investigación. Lo que se planteará a continuación, generalizando la definición anterior, es válido en el contexto del muestreo aleatorio simple (MAS) y en el contexto del muestreo de probabilidad proporcional al tamaño (MPPT), como se señala en (Samuels, 2014).

Sea f_h la probabilidad de que una unidad sociodemográfica de muestro h (provincia, cantón, distrito, etc.) sea seleccionada dentro de la muestra. Sea $f_{(j|h)}$ la probabilidad condicional de que en h sea seleccionado el hogar j. En h pequeñas, todos los hogares $j - ésimos$ pueden ser localizados en un mapa; es una práctica común enlistar los hogares, establecer un punto de inicio aleatorio y tomar una muestra sistemática, por ejemplo, 1 en k.

Sea además $f_{(i|jh)}$ la probabilidad condicional de que dentro del hogar j de h el individuo $i - ésimo$ sea seleccionado. Entonces, la probabilidad total de que el individuo i en j de h sea seleccionado es equivalente al producto de las probabilidades condicionales antes descritas $f_{ijh} = f_h \times f_{(j|h)} \times f_{(i|jh)}$, por lo que el peso final asociado a cada individuo será entonces:

$$W_{ijh} = \frac{1}{f_{ijh}} = \frac{1}{f_h} \times \frac{1}{f_{(j|h)}} \times \frac{1}{f_{(i|jh)}}$$

Si únicamente existe un factor de expansión para cada unidad sociodemográfica de muestreo, entonces la expresión anterior se contrae de la siguiente manera:

$$W_{ijh} = W_{h_*}$$

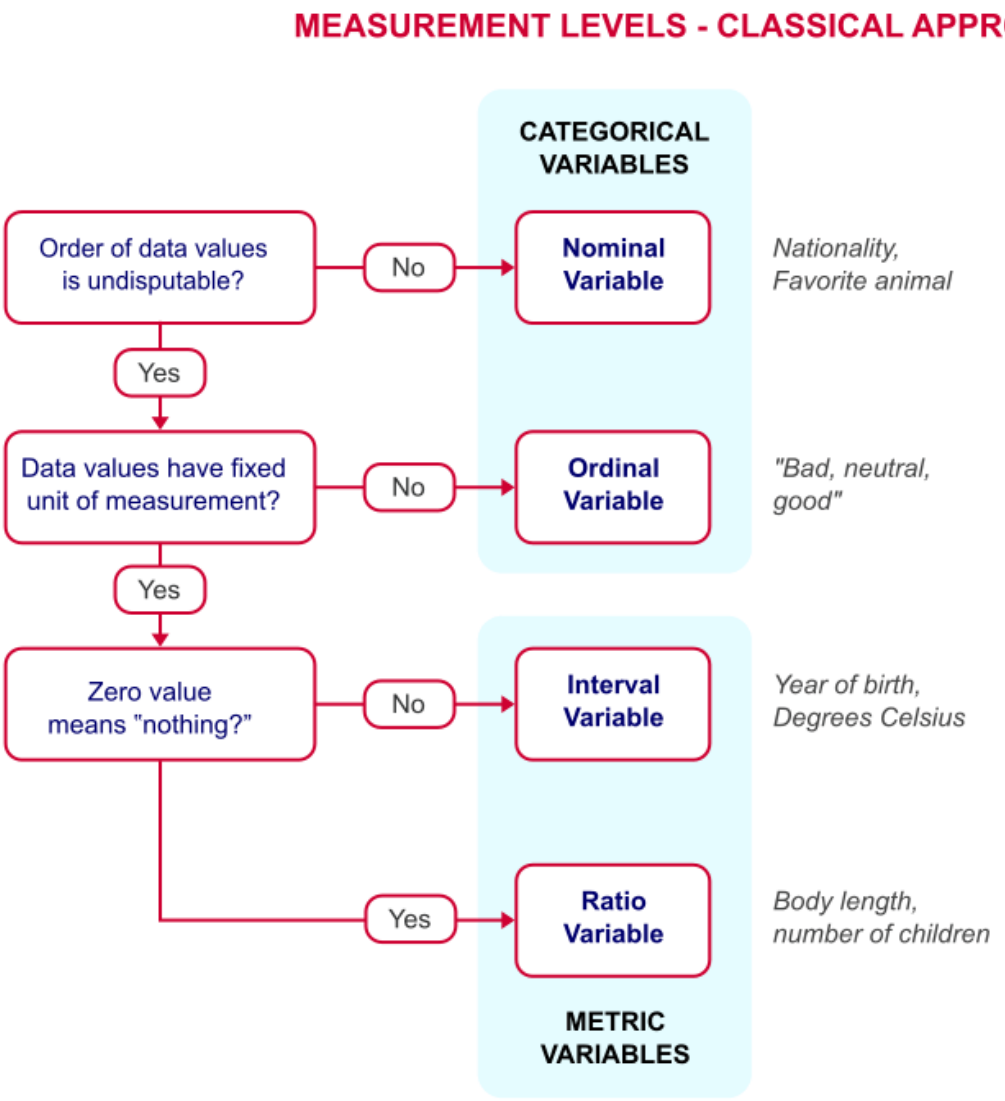
IV.VI. IV. MODELOS LINEALES GENERALIZADOS (MLG)

IV. VI. IV. I. Conceptos Preliminares

1. *Tipos de variable* según escala de medición (Centro Centroamericano de Población, 2021):
    1.1. *Nominal:* Sus valores sólo se pueden clasificar en clases (o categorías), no se pueden ordenar de pequeño a grande o de menos a más. Ejemplos: sexo, estado civil, profesión, ocupación.
    1.2. *Ordinal:* Sus valores se pueden clasificar en categorías y se pueden ordenar en jerarquías con respecto a la característica que se evalúa. Ejemplos: nivel socioeconómico, Apgar, puntaje Apache de Gravedad cardíaca, clase social, lugar en la clase.
    1.3. *De intervalo:* Sus valores tienen un orden natural, es posible cuantificar la diferencia entre dos valores de intervalo. Generalmente tienen unidad de medida. Una variable de intervalo es *discreta* cuando sólo puede tomar un valor entero (por ejemplo: número de

hijos, veces que se consultó al establecimiento de salud); o bien es *continua* si puede tomar cualquier valor en un intervalo (por ejemplo.: peso, talla, índice de masa corporal, etc.).

1.4. *De proporción:* El cero representa la ausencia de la característica que se evalúa. Ejemplos: costo por atención, adecuación peso(edad), etc.



Fuente: (van den Berg, 2021).

2. *Nivel de una variable:* Como se señala en (AMERICAN PSYCHOLOGICAL ASSOCIATION, 2021), en el contexto de los diseños experimentales, es la cantidad, magnitud o categoría de la variable independiente (o de un conjunto de ellas) que está siendo estudiada. Por ejemplo., si un investigador está evaluando el efecto del alcohol en la cognición, cada valor

específico de alcohol incluido en el estudio es un nivel (*i.e.*, 0.0 oz, 0.5 oz, 1.0 oz, 1.5 oz). Complementariamente, (Online Stat Book, 2021) señala que, si un experimento compara un tratamiento experimental con un tratamiento de control, entonces la variable independiente (tipo de tratamiento) tiene dos niveles: experimental y control. Si en un experimento se comparan cinco tipos de dieta, entonces la variable independiente (tipo de dieta) tiene cinco niveles. En general, el número de niveles de una variable independiente es el número de condiciones en las que la variable independiente se evalúa.

3. *Factor:* Como señala la (AMERICAN PYSCHOLOGICAL ASSOCIATION, 2021), un factor puede tener los siguientes significados:
    a) Cualquier cosa que contribuya a un resultado o tenga una relación causal con un fenómeno, evento o acción.
    b) Una influencia subyacente que explica en parte las variaciones en el comportamiento individual.
    c) En el análisis de varianza y otros procedimientos estadísticos, una variable independiente.
    d) En el análisis factorial, una variable latente subyacente no observable que se piensa (junto con otros factores) como responsable de las interrelaciones entre un conjunto de variables.
    e) En matemáticas, un número que se divide sin resto en otro número.

Las definiciones de interés en esta sección de la investigación son las definiciones a), b) y c).

4. *Covariable:* Como señala (Allen, 2017, págs. 282-283), una covariable es una variable continua que se espera que cambie ("varíe") con ("co") la variable de salida/resultado/variable dependiente del estudio. En general, una covariable puede referirse a cualquier variable continua que se espera esté correlacionada con la variable de salida de interés.
5. *Variables Métricas:* Como señala (van den Berg, 2021), este es el nombre que reciben las variables que pueden ser de escala de intervalo (que a su vez pueden ser discretas o continuas) o de razón.
6. *Regresión Logística:* Como señala (AMERICAN PYSCHOLOGY ASSOCIATION, 2021), es una forma de análisis de regresión usada cuando la variable independiente (o variable de salida) sólo puede asumir uno de dos valores categóricos (por ejemplo, aprobar o reprobar) y las predictoras o variables independientes pueden ser tanto categóricas como continuas. Complementariamente, señala (TalkStats, 2011) que la regresión logística multinomial, que es la forma más general de regresión logística, es utilizada para determinar aquellos factores que afectan la presencia o ausencia de una característica cuando la variable dependiente tiene tres o más niveles.

IV. VI. IV. II. MLG y su vínculo genético con los modelos de regresión lineal: la teoría como patrón

IV. VI. IV. II. I. Antecedentes Históricos

Como se señala en (Gujarati & Porter, 2010, pág. 15), fue Francis Galton quien acuñó el término "regresión". En "Family Likeness in Stature", Proceedings of Royal Society, Londres, vol. 40, 1886, pp. 42-72". Ahí planteó que, a pesar de la tendencia de los padres de estatura alta a procrear hijos altos y los padres de estatura baja, hijos bajos, la estatura promedio de los niños de padres de una estatura determinada tendía a desplazarse, o "regresar", a la estatura promedio de la población total. En otras palabras, la estatura de los hijos de padres inusualmente altos o inusualmente bajos tiende a dirigirse a la estatura promedio de la población. Esta ley de regresión universal de Galton fue confirmada por su discípulo y amigo Karl Pearson (junto con A. Lee) en "On the Laws of Inheritance", Biometrika, vol. 2, noviembre de 1903, pp. 357-462. Ahí, se reunen más de mil registros de estaturas de miembros de grupos familiares. Pearson descubrió que la estatura promedio de los hijos de un grupo de padres de estatura alta era menor que la estatura de sus padres, y que la estatura promedio de los hijos de un grupo de padres de estatura baja era mayor que la estatura de sus padres; es decir, se trata de un fenómeno mediante el cual los hijos altos e hijos bajos "regresan" por igual a la estatura promedio de todos los demás. En palabras de Galton, se trata de una "regresión a la mediocridad". La definición moderna de regresión consiste en el "(…) estudio de la dependencia de una variable (variable dependiente) respecto de una o más variables (variables explicativas) con el objetivo de estimar o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las segundas." (Gujarati & Porter, 2010, pág. 15).

Sin embargo, "A pesar de que el análisis de regresión tiene que ver con la dependencia de una variable respecto de otras variables, esto no implica causalidad necesariamente. En palabras de Kendall y Stuart: "Una relación estadística, por más fuerte que y sugerente que sea, nunca podrá establecer una conexión causal nuestras ideas de causalidad deben provenir de estadísticas externas y, en último término, de una u otra teoría." (…) M. G. Kendall y A. Stuart, The Advanced Theory of Statistics, Charles Griffin Publishers, Nueva York, 1961, vol. 2, cap. 26, p. 279." (Gujarati & Porter, 2010, pág. 19). Profundizando un poco más en ello, (Ritchey, 2002, pág. 522) señala que "La existencia de una correlación tan solo denota que las puntuaciones de las dos variables varían de manera conjunta y sistemática en un patrón

predecible. Este descubrimiento por sí mismo no establece causalidad entre las variables. Muchas correlaciones son espurias. Una correlación espuria es aquella que es *conceptualmente falsa, sin sentido o teóricamente sin sentido*, lo cual se ilustra por la correlación entre (…) la tasa de delito en los barrios de la ciudad y la composición racial de una comunidad. Existe una correlación positiva entre el porcentaje de la población minoritaria (por ejemplo, afroamericanos) que viven en barrios y las tasas de crimen. Es decir, para una muestra de comunidades, aquellas con un alto porcentaje de afroamericanos tienden a presentar altas tasas de delito. No obstante, ello sugiere que los afroamericanos son más propensos al comportamiento delictivo, y de hecho, los racistas a menudo citan tal estadística. Esta correlación, sin embargo, resulta espuria. Las tasas de delito son altas en los barrios *pobres* sin tener en cuenta su composición racial, y una parte desproporcionada de los barrios minoritarios son pobres. Es más, la relación entre pobreza y composición racial se debe al racismo, no a la raza biológica. Es decir, ser pobre no tiene nada que ver con la genética. Es la herencia racista de Estados Unidos la que contribuye al hecho de que una parte desproporcionada de los afroamericanos vivan en pobreza, lo cual, a su vez, es un buen predictor de las tasas de delito."[8]

IV.VI. IV. II. II. Mínimos Cuadrados Ordinarios y Regresión

El método de mínimos cuadrados ordinarios y el modelo de regresión lineal no son sinónimos. Como señala (Bhuptani, 2020), hay que comenzar por resaltar primero la diferencia existente entre *regresión lineal* y *ajuste de curvas*. Tener un conjunto de puntos y desear dibujar una curva (línea) a través de ellos que se ajuste lo mejor posible a los mismos es un problema puramente geométrico, es decir, los ejes x e y no tienen interpretación, puesto que lo que en otro contexto serían *datos*, en este son meramente puntos en el espacio cartesiano. Por su parte, la regresión lineal es una inferencia estadística sobre un problema concreto de la realidad. Los valores de y se interpretan según el contexto analítico en que se encuentre el investigador[9]

---

[8] A la explicación anterior hay que añadir que no es el racismo por sí mismo el que genera un nexo entre pobreza y composición racial (al menos no entendido como actitud ideológica frente a las personas afro-descendientes), sino que es la exclusión económica y financiera a la que en general se enfrentan los miembros de la sociedad desprovistos de medios de producción, la cual a su vez se agudiza particularmente con los afro-descendientes dadas las condiciones históricas de esclavitud formal, informal y de marginación social en general a la que los distintos imperios que han existido a lo largo de los diversos modos de producción social han sometido a los pueblos africanos desde los tiempos de la antigua Grecia hasta nuestros días. Merece la pena mencionar, en el contexto del movimiento *Black Lives Matters*, que existen dificultades no triviales para delimitar a qué nos referimos con "afro-descendientes", tomando en cuenta que en 1987 los investigadores Rebecca Cann, Stoneking y Wilson demostraron que el Homo sapiens se originó en África entre 140,000 y 290,000 años atrás y migró de allí al resto del mundo, sustituyendo a los humanos arcaicos; véase (Haskett, 2014) .embargo, para fines de este análisis tómese de punto de partida la época en que las comunidades primitivas ya estaban bien definidas.

[9] Con todas las implicaciones que esto posee.

y con ello se transforman en *datos sobre la variable de interés* para estudio mediante modelos estadísticos, mientras que los valores de x se transforman en *datos adicionales* que se tiene sobre cada elemento de y que podría ser útil para realizar estimaciones sobre su comportamiento, es decir, transformar los *datos* en *información*, información cuyo carácter debe ser estadísticamente significativo para ser empleado en toma de decisiones relevantes en distintas esferas de la realidad. Cuando se hace una regresión lineal, se está tratando de construir un modelo probabilístico que describa la variable y teniendo en cuenta a la variable x, sin embargo, existen múltiples formas de realizar esto. Un modelo lineal supone que y tiene una media diferente para cada valor posible de x. Así, el conjunto de estos valores medios sigue una línea recta con una cierta intersección y una cierta pendiente. Como con cualquier problema de inferencia estadística, se estiman los parámetros desconocidos utilizando la estimación de máxima verosimilitud. Sin embargo, como en este caso los parámetros desconocidos son una intersección y una pendiente, el resultado final de la estimación de máxima verosimilitud es básicamente que se está eligiendo una línea recta que se ajuste mejor a los datos observados, por lo que es así como convergen la regresión lineal con el ajuste de curvas, *i.e.*, la regresión lineal es el resultado de la estimación de máxima verosimilitud del ajuste del modelo, cuando el conjunto de datos tiene un comportamiento lineal.

Una vez planteado lo anterior, como se señala en la fuente citada, es posible pensar en la regresión lineal como una metodología que utiliza la herramienta del ajuste de curvas (en el caso de la regresión lineal simple, específicamente de una línea recta –o de un hiperplano, si es una regresión lineal múltiple–) mediante un conjunto de puntos que llamamos cualitativamente datos. Sin embargo, existen muchas estrategias posibles para ajustar una línea a través de un conjunto de puntos; por ejemplo, en el contexto de la Ciencia de Datos, existen técnicas para entrenar un modelo lineal que no usa mínimos cuadrados lineales, como se señala en (StackExchange Cross Validated, 2017). A nivel de la Estadística Matemática Clásica existen diferentes metodologías para realizar el ajuste de curvas, entre ellas se encuentran:

a) Tomar el punto más a la izquierda y el más a la derecha y dibujar una línea entre ellos.
b) Calcular las pendientes de las líneas que conectan cada par de puntos y calcular la pendiente promedio, dibujando una línea con esta pendiente que pase por el punto en el promedio de los valores de x y el promedio de los valores de y.
c) Se puede encontrar la línea para la cual hay un número igual de puntos sobre la línea y debajo de la línea.

d) Es posible dibujar una línea y luego, para cada uno de los puntos de datos, medir la distancia vertical entre el punto y la línea y sumarlos; la línea ajustada sería aquella donde esta suma de distancias es lo más pequeña posible.
e) También se puede dibujar una línea y luego, para cada uno de los puntos de datos, medir la distancia vertical entre el punto y la línea, elevarlos al cuadrado y sumarlos; la línea ajustada sería aquella donde esta suma de distancias es lo más pequeña posible.

La última estrategia se llama *mínimos cuadrados ordinarios*, de ahora en adelante *MCO*, y su nombre proviene del hecho que se está buscando minimizar la suma de los errores de predicción al cuadrado. Sin embargo, a pesar de que los *MCO* son la técnica más popular que emplea la metodología del Análisis de Regresión, en lo que respecta al ajuste de una línea a través de un conjunto de puntos, cualquiera de las otras estrategias es igualmente válida. Las tres primeras estrategias las inventó el autor a manera de ejemplo y probablemente no funcionen adecuadamente; sin embargo, la cuarta es una estrategia real llamada *desviaciones menos absolutas* y es preferida por algunas personas por sobre mínimos cuadrados.

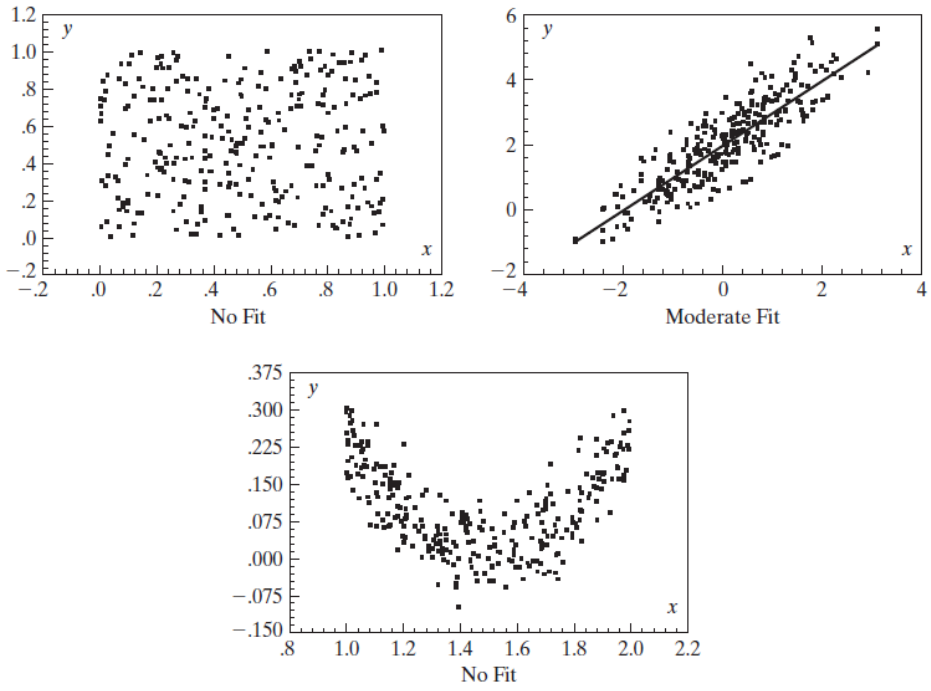
Cabe preguntarse por qué si no es la única técnica entonces es la más utilizada. La razón es que, al resolver el problema de regresión lineal estadística, una suposición de modelado muy común es que por cada valor posible de x, la cantidad y se distribuye normalmente con una media que es lineal en x. Por lo tanto, la función de verosimilitud es esencialmente un producto de funciones de densidad de probabilidad de la distribución normal. Así mismo, el autor señala que se estiman los parámetros desconocidos (y, por lo tanto, se encuentra la recta de mejor ajuste al conjunto de observaciones) maximizando la función de verosimilitud. Si se observa cómo es el producto de funciones de densidad de probabilidad normales, el lector notará que maximizar esta expresión es equivalente a minimizar la suma de los errores al cuadrado. Es decir, la línea que se obtiene realizando el ajuste de la curva a través de mínimos cuadrados es equivalente a la línea que obtiene realizando una regresión lineal utilizando un modelo distribuido normalmente.

De esta forma, puede observarse que el análisis de regresión es una metodología, mientras que los mínimos cuadrados con una técnica empleada por esta metodología. Mucho menos debe identificarse una regresión lineal con la técnica de mínimos cuadrados, puesto que, por ejemplo, existen distintos tipos de análisis de regresión, entre ellos el análisis de regresión lineal. Sin embargo, los mínimos cuadrados son una de las técnicas posibles en la regresión lineal para encontrar la línea recta de mejor ajuste al conjunto de datos del que se dispone. Así, se presentan a continuación dos figuras, la primera permite ver los diferentes ajustes que un conjunto de datos muestrales puede tener respecto a una recta y cómo esta se

convierte en la recta de mejor ajuste, mientras que la segunda permite visualizar la descomposición de cada y_i en la regresión lineal.

The original fitting criterion, the sum of squared residuals, suggests a measure of the fit of the regression line to the data. However, as can easily be verified, the sum of squared residuals can be scaled arbitrarily just by multiplying all the values of y by the desired scale factor. Since the fitted values of the regression are based on the values of **x**, we might ask instead whether *variation* in **x** is a good predictor of *variation* in y. Figure 3.3 shows three possible cases for a simple linear regression model. The measure of fit described here embodies both the fitting criterion and the covariation of y and **x**.

FIGURE 3.3 Sample Data.



Fuente: (Greene, 2012, pág. 79).

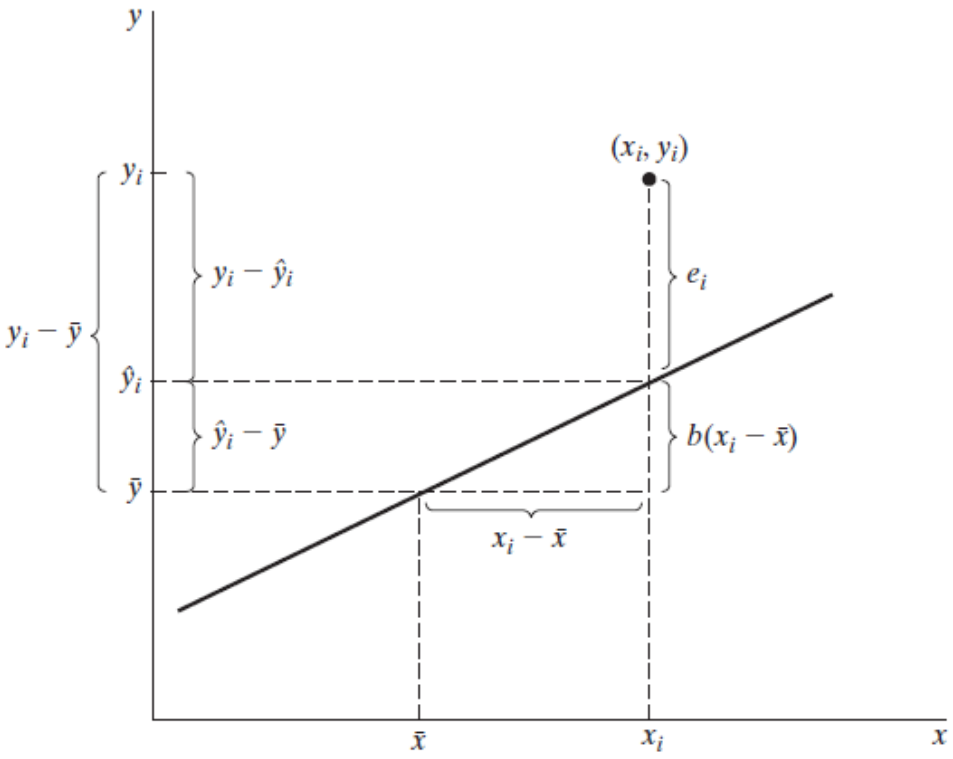


FIGURE 3.4 Decomposition of y_i.

Fuente: (Greene, 2012, pág. 80).

En suma, pueden entenderse los mínimos cuadrados como una técnica en el análisis de regresión para aproximar la solución de sistemas sobre determinados (i.e., conjuntos de ecuaciones en las que hay más ecuaciones que incógnitas) minimizando la suma de los cuadrados de los residuos hechos en los resultados de cada ecuación. Por su parte, los mínimos cuadrados lineales es una técnica de aproximación de funciones lineales a conjuntos de datos por la técnica de mínimos cuadrados. Es un conjunto de formulaciones para resolver problemas estadísticos relacionados con la regresión lineal, incluidas las variantes para residuos ordinarios (no ponderados), ponderados y generalizados (correlacionados). Los métodos numéricos para mínimos cuadrados lineales incluyen la inversión de la matriz de las ecuaciones normales y los métodos de descomposición ortogonal. Por su parte, los mínimos cuadrados ordinarios son una de las 3 técnicas más comunes (más no las únicas) dentro de la familia de técnicas conocida como "Mínimos Cuadrados Lineales", que a su vez pertenece a la familia de técnicas conocida como "Mínimos Cuadrados". Los mínimos cuadrados ordinarios son utilizados en el contexto del modelo clásico de regresión lineal para estimar sus parámetros desconocidos.

En línea con (McCullagh & Nelder, 1989, pág. 4), deben considerarse a las diferentes teorías estadísticas como descripciones de determinados patrones que es

posible identificar que siguen los números en la vida real, patrones los cuales en alguna medida pueden sustituir al conjunto de datos en sí mismos (puesto que estos patrones numéricos describen patrones geométricos, es decir, relativo a las formas que adoptan los fenómenos naturales y/o sociales) a través de determinados valores numéricos concretos en los que se cristalizan dichos patrones. Es por ello que según las características empíricas del conjunto de datos en concreto que se estudie (obtenido de la medición de fenómenos naturales y/o sociales) los parámetros $\boldsymbol{\beta}$ generados por tal conjunto de datos tomarán diferentes valores y precisamente de este hecho empírico es que se formulan las teorizaciones estadísticas-matemáticas que actualmente se conocen como *familias de distribuciones de probabilidad*.

El modelo básico de regresión lineal $y = \alpha + \beta x$ conecta dos variables x e y vía el par de parámetros (α, β) y define una relación entre ambas que describe geométricamente una línea recta.

VI.VI. IV. II. III. Familias Exponenciales

Formalmente, según (Patil & Shorrock, 1965, pág. 94), una familia exponencial se define como aquella familia $\{X_\omega : \omega \in \Omega\}$, en donde Ω es un espacio de variables estocásticas real-evaluadas X_ω en el que cada una de estas tiene asociada una función de densidad (o de masa) de probabilidad F_ω y en el que la derivada de la función respecto a x tiene la forma $dF_\omega(x) = \left\{\frac{e^{\omega x}}{f(\omega)}\right\} dv(x)$ y $\int e^{\omega x} dv(x)$.

En la definición anterior, v es una función que sirve como medida[10] en las integrales de Lebesgue–Stieltjes[11] y es una función no decreciente de variable real, mientras que ω es el conjunto de parámetros que en la notación usual suelen encontrarse como θ.

IV. VI. IV. III. Los Componentes del Modelo Lineal Generalizado

IV. VI. IV. III. I. El modelo lineal clásico como punto de partida

Como señalan (McCullagh & Nelder, 1989, pág. 26), los modelos lineales generalizados son una extensión de los modelos clásicos, por lo que estos últimos representan el punto de partida de la exposición.

Un vector de observaciones y que posee n componentes se asume que es una realización de la variable aleatoria Y cuyos componentes están independientemente distribuidos con medias μ. El componente sistemático del modelo es una especificación del vector μ en términos de un pequeño número de

[10] Como señala (Kolmogórov & Fomin, 1978)

[11] Es una de las generalizaciones posibles de la integral de Riemann y de Stieltjes, fundamentada bajo el marco formal de la Teoría de la Medida fundada por Henri Lebesgue, en sentido en que lo están las integrales de Lebesgue.

parámetros desconocidos $\beta_1, \beta_2, \dots, \beta_p$. En el caso de los modelos lineales ordinarios, esta especificación toma la siguiente forma:

$$\mu = \sum_{1}^{p} x_j \beta_j$$

En la expresión anterior, los β son parámetros cuyos valores son usualmente desconocidos y deben ser estimados a partir del conjunto de datos. Si se indizan las observaciones mediante la letra i, es posible entonces expresar al componente sistemático del modelo de la siguiente manera:

$$E(Y_i) = \mu_i = \sum_{1}^{p} x_{ij} \beta_j \, ; i = 1,2, \dots, n,$$

En la expresión anterior, x_{ij} es el valor de la $j - ésima\ covarianza$ de la observación i, que en este caso es una variable en sí misma (por ello es que se habla de $j - ésima$ covarianza). En notación matricial, en donde μ es una matriz de orden $n \times 1$, X una matriz de orden $n \times p$ y β una matriz de orden $p \times 1$, es posible escribir lo anterior como $\mu = X\beta$, en donde X es la matriz modelo y β el vector de parámetros; la estructura del componente sistemático asume que las covarianzas que perturban a la media son conocidas y pueden ser medidas con efectividad y libre de errores, lo cual también debe verificarse con el conjunto de datos del que se disponga en la medida en que sea posible. Para el caso de la parte estocástica, se asume independencia entre sus elementos y varianza constante de los errores. Estos supuestos son fuertes a nivel teórico y deben verificarse, en tanto sea posible, de los datos mismos.

Algunas especializaciones del modelo lineal clásico asumen supuestos más fuertes (restrictivos) como que los errores de estimación siguen una distribución normal con varianza constante σ^2. Sintetizando lo visto sobre el modelo lineal clásico, este puede ser expresado como $E(Y) = \mu$, donde $\mu = X\beta$.

VI. VI. IV. III. I. La generalización del modelo lineal clásico

Como señalan (Nelder & Wedderburn, 1972, pág. 370), "Los modelos lineales habitualmente incorporan componentes tanto sistemáticos como aleatorios (error), y los errores generalmente se asume que tienen distribuciones normales. La técnica analítica asociada es la teoría de mínimos cuadrados, que en su forma clásica asumía solo un componente de error; las extensiones para errores múltiples se han desarrollado principalmente para el análisis de experimentos diseñados y datos de encuestas. Las técnicas desarrolladas para datos no normales incluyen análisis Probit, donde una variable binomial tiene un parámetro relacionado con una distribución de tolerancia subyacente asumida, y tablas de contingencia, donde la

distribución es multinomial y la parte sistemática del modelo, generalmente multiplicativa."

Así, la generalización del modelo lineal clásico es posible puesto que "En ambos ejemplos hay un aspecto lineal del modelo; por lo tanto, en el análisis Probit, el parámetro p es una función de la tolerancia Y, que en sí misma es lineal sobre la dosis (o alguna función de la misma), y en una tabla de contingencia con un modelo multiplicativo, el logaritmo de la probabilidad esperada se asume lineal al clasificar los factores que definen la mesa. Por tanto, para ambos, la parte sistemática del modelo tiene una base lineal. En otra extensión (Nelder, 1968) se usa cierta transformación para producir errores normales, y se usa una transformación diferente de los valores esperados para producir linealidad. Hasta ahora hemos mencionado modelos asociados con las distribuciones normal, binomial y multinomial (esta última puede considerarse como un conjunto de distribuciones de Poisson con restricciones). Una clase adicional se basa en la distribución χ^2 o gamma y surge en la estimación de los componentes de la varianza a partir de formas cuadráticas independientes derivadas de las observaciones originales. Nuevamente, el componente sistemático del modelo tiene una estructura lineal (…) En esta investigación, nosotros desarrollamos una clase de modelos lineales generalizados, los cuales incluyen todos los ejemplos anteriores, y proporcionamos un proceso unificado para ajustarlos con base en la verosimilitud. Este procedimiento es una generalización del procedimiento bien-conocido descrito por Finney (1952) para máxima verosimilitud en el contexto del análisis Probit" (Nelder & Wedderburn, 1972, págs. 370-372).

Para simplificar la transición del modelo lineal clásico al generalizado, se debe reestructurar sutilmente $E(Y) = X\beta$ para producir la siguiente especificación de tres partes.

VI. VI. IV. III. II. Componentes del MLG

VI. VI. IV. III. II. I. El componente estocástico

Así, con base en (Nelder & Wedderburn, 1972, pág. 371), supóngase que las observaciones que conforman el conjunto de datos pueden ser descrita por una función de densidad (o de masa) de probabilidad π de la siguiente forma:

$$\pi(z;\ \theta, \phi) = \exp\left[\alpha(\phi)\{z\theta - g(\theta) + h(z)\} + \beta(\phi, z)\right]$$

En donde α, g, h y β son conocidas, así como también $\alpha(\phi) > 0$ tal que para un valor fijo de ϕ se tiene una familia exponencial descrita por $\pi(z;\ \theta, \bar{\phi})$. θ representa los parámetros de la distribución de la variable dependiente descrita por el conjunto de observaciones $z \in Z$ y ϕ es un parámetro de dispersión, usualmente asociado a la varianza de las distribuciones de probabilidad, aunque también

puede ser, por ejemplo, el parámetro p para el caso de una distribución gamma. La media de z se denota como μ. Para el caso de vectores de variables aleatorias o vectores estocásticos Y, los componentes de Y tienen distribuciones independientes e idénticas (*iid*) con esperanza matemática $E(Y) = X\beta = \mu$ y varianza constante σ^2 o cualquier otro parámetro de dispersión.

VI. VI. IV. III. II. I. El componente sistemático
Como se señala en (McCullagh & Nelder, 1989, pág. 26), las covarianzas $x_1, x_2, \dots, x_p$ producen un predictor lineal η definido como $\sum_1^p x_j\beta_j$. En este sentido, se señala en (Nelder & Wedderburn, 1972, pág. 372) que las variables independientes pueden ser cuantitativas y producir un único valor para la variable x en el modelo, pueden ser cualitativas y producir un conjunto de valores de x conformado por la opción 0 y la opción 1, o puede ser una mezcla de ambos tipos.

VI. VI. IV. III. II. I. El enlace entre el componente aleatorio y el componente sistemático: el enlace canónico
Con este enlace se garantiza que $\mu = \eta$. Esta generalización introduce, como puede verificarse, un nuevo símbolo η para el predictor lineal y el tercer componente, para luego especificar que μ y η son de hecho idénticos. Si se escribe $\eta_i = g(\mu_i)$, entonces $g(\cdot)$ será llamada *función enlace*.

Como se adelantó, preliminarmente los modelos lineales clásicos estaban restringidos al componente estocástico y a la identidad entre el predictor lineal y la media. Los modelos lineales generalizados permiten dos extensiones:

1) La distribución de probabilidad que sigue el componente estocástico ya no está restringida únicamente a ser normal, sino que puede ser generada por los miembros de una familia de funciones exponenciales diferente de la normal.
2) La función de enlace especificada en el tercer componente puede ser cualquier función monotónica diferenciable.

Complementariamente, acorde con (McCullagh & Nelder, 1989, pág. 31), se debe mencionar que la función enlace relaciona al predictor lineal η con el valor esperado μ de un conjunto de datos y. En los modelos lineales clásicos la media y el predictor lineal eran idénticos, y el enlace identidad es plausible en el sentido de que tanto η como μ pueden tomar valores que pertenecen a los números reales. Sin embargo, cuando se está trabajando con conteos y la distribución de Poisson, se debe tener $\mu > 0$, por lo que el enlace identidad es menos atractivo, en parte porque η puede ser negativa mientras que μ no debe serlo. Modelos de conteo basados en la independencia en datos de clasificación cruzada conduce naturalmente a efectos multiplicativos, y esto está expresado en el *log-enlace* de la forma $\eta = \log(\mu)$, cuya inversa es $\mu = e^n$. Además, los efectos aditivos que

contribuyen a η se convierten en efectos multiplicativos que contribuyen a μ, y μ es necesariamente positiva. El tratamiento matemático dado a cinco distribuciones de probabilidad, en el contexto de los modelos lineales generalizados, se presentan a continuación.

Table 2.1 *Characteristics of some common univariate distributions in the exponential family*[†]

	Normal	*Poisson*	*Binomial*	*Gamma*	*Inverse Gaussian*
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
Range of y	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0, \infty)$	$(0, \infty)$
Dispersion parameter: ϕ	$\phi = \sigma^2$	1	$1/m$	$\phi = \nu^{-1}$	$\phi = \sigma^2$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y; \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log y!$	$\log\binom{m}{my}$	$\nu\log(\nu y) - \log y$ $-\log\Gamma(\nu)$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$
$\mu(\theta) = E(Y; \theta)$	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
Canonical link: $\theta(\mu)$	identity	log	logit	reciprocal	$1/\mu^2$
Variance function: $V(\mu)$	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

[†]The mean-value parameter is denoted by μ, or by π for the binomial distribution.
The parameterization of the gamma distribution is such that its variance is μ^2/ν.
The canonical parameter, denoted by θ, is defined by (2.4). The relationship between μ and θ is given in lines 6 and 7 of the Table.

Fuente: (McCullagh & Nelder, 1989, pág. 30).

VI. VI. IV. III. III. Proceso de Ajuste del Modelo

VI. VI. IV. III. III. I. Fundamento Estadístico-Matemático Preliminar

El ajuste de una simple relación lineal entre los elementos x y los elementos y requiere que sea utilizado un par de valores en particular (a, b), seleccionado del conjunto de todos los posibles pares de valores que pueden adoptar los parámetros (α, β), que genere valores estimados $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ que se ajusten (en el sentido de generar una figura geométrica -en este caso una línea recta-) mejor a los valores observados $y_1, y_2, \dots, y_n$.

Con la finalidad de cuantificar las diferencias en el ajuste entre los valores estimados $\hat{y}_n$ y los valores observados y_n, se deben medir las distancias entre tales conjuntos de valores. En general, estas distancias (referidas en textos como el de McCullagh y Nelder como discrepancias) buscan cuantificarse con la finalidad de optimizar el proceso de selección del patrón geométrico teórico que mejor describa el patrón observado descrito por el conjunto de datos. Para cuantificar tales distancias se pueden utilizar diferentes tipos de métricas (criterios de medición), selección que a su vez responderá a las características específicas del conjunto de datos, sobre lo cual se abordarán brevemente a continuación algunos aspectos.

Cada uno de los tipos de métrica que es posible utilizar para medir distancias se conoce, en el contexto del Análisis Matemático en general y del Análisis Real en particular (campo al que pertenece a su vez la Estadística Matemática), como *función distancia* en espacios topológicos equipados con métrica. Los espacios topológicos son estructuras matemáticas en las cuales existen reglas de agrupación o *topología* de los elementos de un determinado conjunto en subconjuntos, en donde tales reglas de agrupación expresan las relaciones entre los elementos que conforman el conjunto y a partir de las cuales se generan los demás tipos de relaciones entre tales elementos, las cuales geométricamente representan sus distancias relativas, como se verifica en los anexos correspondientes al estudio de los grafos desde la perspectiva de los isomorfismos. Por otro lado, la métrica o función distancia es la función utilizada para medir las distancias absolutas entre los elementos de un conjunto o, para este caso, las distancias absolutas de las observaciones que conforman un conjunto de datos. Los distintos tipos de métrica son siempre generalizaciones de la métrica de espacios euclidianos, expresada en su forma básica en el ampliamente conocido *teorema de Pitágoras*.

Con la finalidad de ofrecer una exposición integral de lo presentado por (McCullagh & Nelder, 1989, pág. 5) en relación a las normas L_1 y L_2, específicamente para comprender cómo es que estos conceptos matemáticos son el fundamento del ajuste del modelo de regresión utilizado en cuanto son fundamento teórico y aplicado de la métrica empleada en las mediciones de

cualquier índole, se abordarán sintéticamente algunos conceptos del Análisis Matemático y el Álgebra Abstracta.

El primer concepto a estudiar es el de *función convexa*. Como señalan (Kolmogórov & Fomin, 1978, pág. 140), una *funcional*[12] *convexa* no negativa ρ, definida sobre un espacio lineal real L, se llama convexa si cumple las siguientes condiciones:

1) $\rho(x+y) =\leq \rho(x)+\rho(y)$, para todo x, y que pertenece a L.
2) $\rho(\alpha x) = \alpha\rho(x)$, para todo $\alpha \geq 0$
3) No se admite que el valor de $\rho(x)$ es finito para todo x que pertenece a L, es decir, se admite el caso en que $\rho(x) = +\infty$ para algunos x que pertenecen a L.

El segundo concepto a estudiar es el de *norma*. Como señala (Kolmogórov & Fomin, 1978, pág. 149), una funcional convexa finita ρ, definida sobre L, se llama *norma* cuando verifica las siguientes dos condiciones, adicionales a las de convexidad:

1) $\rho(x) = 0$, sólo si $x = 0$.
2) $\rho(\alpha x) = |\alpha|\rho(x)$, para todo α.

Sintetizando las condiciones vistas en el primer y segundo concepto, se obtienen:

1) $\rho(x) \geq 0$, con la particularidad de que $\rho(x) = 0$ sólo si $x = 0$.
2) $\rho(x+y) \leq \rho(x)+\rho(y)$, para todo x, y que pertenece a L.
3) $\rho(\alpha x) = |\alpha|\rho(x)$, para todo α.

Generalizando las tres condiciones anteriores, se tiene

$$\left\|\sum_{k=1}^{n} \alpha_k x_k\right\| \leq \sum_{k=1}^{n} |\alpha_k|\|x_k\|$$

Así, en el contexto de los *espacios topológicos lineales* (conocidos en cursos de matemática elemental como espacios vectoriales) es posible, en conjunto con el cumplimiento de otras condiciones de carácter más general (global) que no se especificarán aquí, derivar de la función norma una función métrica. La importancia de que la métrica sea inducida por una norma, a nivel de los espacios euclidianos y sus generalizaciones más usadas (los espacios de Hilbert), proviene del hecho de que así se garantiza que la función métrica posea la característica de ser invariante ante traslaciones, *i.e.*, una *métrica invariante ante traslaciones* a la que suele hacer alusión la literatura bajo el nombre de *principio de traslación invariante*, así como también la tercera propiedad relativa a los escalares α. Lo que en

[12] Un funcional es una generalización del concepto de función, específicamente es una función de funciones.

términos cotidianos significa que una función métrica sea invariante ante traslaciones es que dicha función, al realizar mediciones de cualquier índole sobre algún objeto localizado en la estructura matemática en cuestión, no arrojará mediciones diferentes cuando el objeto sea trasladado de un lugar a otro dentro de la misma estructura matemática, que para este caso son los espacios topológicos lineales antes mencionados.

Como puede verificarse en señala (Kolmogórov & Fomin, 1978, pág. 52)[13], las características de la función métrica d de espacios euclidianos, conocida también como *función distancia* d de espacios euclidianos, adicionales son las siguientes:

1) Tomando para los elementos de un conjunto arbitrario
$$d(x,y) = \begin{cases} 0, si\ x = y \\ 1, si\ x \neq y \end{cases}$$
se obtendrá, evidentemente, un espacio métrico que puede ser denominado espacio de puntos aislados.
2) El conjunto de los números reales con la distancia $d(x,y) = |x - y|$ forma el espacio métrico R^1.
3) El conjunto de grupos ordenados de n números reales $x = (x_1, x_2, \dots, x_n)$ con la distancia $d(x,y) = \sqrt{\sum_{k=1}^{n}(y_k - x_k)^2}$ se denomina *espacio aritmético euclídeo de n dimensiones* R^n. Esta condición es conocida también como cumplimiento del *axioma triangular*.

Como se verifica en (Lipschutz, 1992, pág. 51)[14], las tres condiciones anteriores pueden expresarse de forma general como se presenta a continuación:

a) $d(x,y) = 0$, (para $x = y$)[15]

b) $d(x,y) = \rho(y,x) > 0$, (para $x \neq y$)

c) $d(x,y) \leq \rho(x,z) + \rho(z,y)$, (para $z \neq x, y$)[16]

d) $d(x,y) \geq 0$, (positividad, deducida de las tres propiedades anteriores).

---

[13] La obra citada no hace diferenciación entre la función métrica y la función norma en términos de notación, puesto que para ambas utiliza ρ. En esta investigación hemos sustituido la ρ métrica por d, ello con la finalidad de mostrar la mecánica operativa de cómo la métrica es inducida por una norma.

[14] Aunque su formato de presentación se ha generalizado por cuenta propia para estandarizar notación y nivel de abstracción matemática con las referencias tomadas de (Kolmogórov & Fomin, 1978) (que es el libro de referencia que se usa en esta investigación en lo relativo al Análisis Matemático). Ello con la finalidad de ofrecer mayor claridad y fluidez en la narrativa.

[15] También conocida como *Identidad de los Indiscernibles*.

[16] Esta condición es precisamente la que garantiza que la métrica o función distancia sea invariante ante traslaciones.

El concepto de espacio real R^n tiene su generalización natural en el concepto de espacios L_p. Como se verifica en (Wikipedia, 2021), los espacios L_p son espacios de funciones (*i.e.*, espacios más abstractos a los usuales en donde las coordenadas están dadas en términos de funcionales y no de números reales) definidos como generalización natural de la norma ρ para espacios vectoriales de dimensión finita. Estos espacios también son conocidos como *espacios de Lebesgue*. De ahí que se hable de norma L_1, norma L_2 y hasta norma L_p. ¿Cómo exactamente es que la métrica es inducida por una norma?

Sean $x, y, z \in V$, donde x, y, z son variables, V un espacio topológico lineal, $d(x, z)$ una métrica entre un punto x y un punto z, y sea también $\rho(x - y)$ la función norma de interés. Mostrar cómo una métrica es inducida por una norma es equivalente a demostrar que al igualar matemáticamente la expresión $d(x, z)$ con la expresión $\rho(x - y)$ y desarrollar la expresión resultante, es posible obtener un resultado, conocido también como *métrica inducida* (por una norma), que expresa una función métrica que posee las propiedades de una función métrica usual, lo cual se verifica si partiendo de la norma de interés es posible arribar a métrica invariante ante traslaciones[17]. Esto se muestra a continuación.

$$
\begin{aligned}
d(x, z) = \rho(x - y) &= \rho\big(x + (-z)\big) \\
&= \rho\big((x - y) + (y - z)\big) \\
&\leq \rho(x - y) + \rho(y - z) \\
&= \rho(x - y) + |-1|\rho(y - z) \\
&= \rho(x, y) + \rho(y, z) \\
&= d(x, y) + d(y, z)
\end{aligned}
$$

Así, como señalan (McCullagh & Nelder, 1989, pág. 5), seleccionar el modelo estadístico óptimo para determinado conjunto de datos se hace bajo el criterio de cuál es el que genera un conjunto de datos estimado más cercano al conjunto de datos observado (los que han sido capturados a través de mediciones a fenómenos de la realidad), lo cual equivale a determinar la discrepancia cuantitativa (en términos de sus distancias dentro del espacio de Lebesgue en el que se estudian) que existe entre cada uno de los elementos del conjunto de datos estimado y los del conjunto de datos observado, lo cual los econometristas suelen conocer a nivel empírico como *residuo de la regresión* y a nivel teórico como *término de perturbación estocástica* o *término de error*. Esta discrepancia se cuantifica mediante alguna función norma, cuya selección dependerá de las características específicas del

[17] Véase (Perry, 2014).

conjunto de datos que se estudie (cuyas características vienen conferidas por las del fenómeno que cuantifican y por el diseño mismo del instrumento de medición con el que se capturaron estos datos); sin embargo, estas normas poseen todas limitantes en común que condiciona su uso, las cuales se expondrán más adelante.

VI. VI. IV. III. III. II. Introducción

Como señalan (McCullagh & Nelder, 1989, pág. 5), existen desde la norma L_1 definida como $S_1(y, \hat{y}) = \sum|y - \hat{y}|$ hasta la norma L_p definida como $S_\infty(y, \hat{y}) = \max_i|y - \hat{y}|$. Sin embargo, los espacios que usualmente son de más interés aplicado, así como también aquellos en que se realiza el ajuste clásico por mínimos cuadrados, son los espacios L_2 normados, conocidos también a nivel estadístico como espacios de desviaciones cuadráticas, que son los espacios aritméticos euclidianos $n - dimensionales$ antes definidos a nivel métrico. A continuación, se define la norma de estos espacios.

$$S_2(y, \hat{y}) = \sum (y_i - \hat{y}_i)^2$$

Como se mencionó, la validez de las estimaciones realizadas tiene determinados elementos en común con independencia del orden de la norma que se utilice, orden que indica precisamente la potencia a la que se eleva la diferencia entre las localizaciones de los elementos de los conjuntos de datos (del estimado o generado con la regresión y del observado). Estos elementos en común son tres:

1) En primer lugar, todas las mediciones (expresadas en las observaciones) han sido realizadas bajo la misma escala física.
2) Las observaciones son independientes entre sí o, al menos, "(…) ellas son en algún sentido intercambiables, lo que un trato imparcial de los componentes." (McCullagh & Nelder, 1989, pág. 5). Esta noción puede generalizarse a lo que es posible concebir como *intercambiabilidad estocástica*, que generaliza la noción de *independencia estocástica*.
3) Cada una de las desviaciones debe ser independiente del valor esperado del conjunto de observaciones.

Como se deduce de lo anterior, cada una de las normas $L\ p - ésimas$ se corresponde con un determinado criterio estadístico, específicamente con la potencia a la que se elevan las distancias entre los valores observados y los valores estimados, puesto que así se realiza la medición de los residuos o errores en el contexto de los modelos de mínimos cuadrados. En este sentido, (Wikipedia, 2021) señala que el método IRLS se utiliza para encontrar las estimaciones de máxima verosimilitud de un modelo lineal generalizado como una forma de mitigar la influencia de valores atípicos en un conjunto de datos normalmente distribuido.

Por ejemplo, minimizando los errores mínimos absolutos (expresado en la norma L_1) en lugar de los errores de mínimos cuadrados (expresado en la norma L_2).

VI. VI. IV. III. III. III. Método de los Mínimos Cuadrados de Reponderación Iterativa (IRLS)

Como se adelantó, el proceso de ajuste del modelo se lleva a cabo mediante el método de máxima verosimilitud. En este sentido, se señala en (Nelder & Wedderburn, 1972, págs. 372-373) que la solución a las ecuaciones de máxima verosimilitud generadas para el caso de $n - ésimas$ variables estocásticas contenidas dentro del vector estocástico Y es equivalente a un procedimiento iterativo por mínimos cuadrados ponderados con una función de ponderación:

$$w = \frac{\left(\frac{d\mu}{dY}\right)^2}{V}$$

En la expresión anterior, w son las ponderaciones, $\frac{d\mu}{dY}$ es el diferencial del valor esperada respecto de Y, mientras que V es la varianza de las observaciones, en donde μ, Y y V expresan estimaciones derivadas del conjunto de datos con el que se cuenta. Además, Y (que es el conjunto de datos pronosticado, no el observado) es expresada en el método IRLS a través de la expresión $y = Y + \frac{z-\mu}{\frac{d\mu}{dY}}$, conocida como *working Probit*.

VI. VI. IV. III. III. IV. Valores Semilla de la Simulación

Como señalan (Nelder & Wedderburn, 1972, pág. 374)En la práctica, podemos obtener un buen procedimiento de partida para la iteración de la siguiente manera: tomar como primera aproximación $\mu = z$ y calcular Y a partir de ese valor; luego calcúlese w como se definió y establézcase $y = Y$. Luego obténgase la primera aproximación a los β's por regresión. El método puede necesitar una ligera modificación para tratar con valores extremos de z. Por ejemplo, con la distribución binomial probablemente será adecuado reemplazar instancias de $z = 0$ o $z = n$ con $z = \frac{1}{2}$ y $z = -\frac{1}{2}$ en aquellos casos en los que, por ejemplo, las transformaciones Probit y Logit, $\mu = 0$ o $\mu = n$ conducirían a obtener valores infinitos para Y.

VI. VI. IV. III. III. V. Funcionamiento del Algoritmo IRLS

VI. VI. IV. III. III. V. I. Fundamento matemático

Como se señala en (Burrus, 2021), La distancia de una observación cualquiera respecto a la medida de tendencia central de su distribución se conoce como desviación. En el contexto de los modelos predictivos, las desviaciones se conocen como residuos, denotados mediante la letra e. Estos residuos en el contexto de la regresión lineal se estiman matricialmente de la forma:

$$e = Ax - b \quad (1)$$

Como ya es sabido, la ecuación de la norma puede tomar la siguiente forma:

$$\|e\|_p = \left(\sum_n |e(n)|^p\right)^{1/p} \quad (2)$$

Una regresión consiste, en general y sintéticamente, en ajustar un conjunto de datos a una función, la función de mejor ajuste, mediante la minimización de los residuos elevados a alguna potencia, usualmente al cuadrado. Si existe una solución óptima o aproximadamente óptima para minimizar el error de la primera ecuación matricial, entonces es posible alcanzar este objetivo mediante la minimización de la segunda ecuación, correspondiente a la ecuación de la norma del residuo.

Se sabe que la fórmula para encontrar el residuo cuadrado mínimo ponderado es:

$$\|W_e\|_{p=2}^2 = \sum_n w_n^2 |e(n)|^2 \quad (3)$$

Así, existen demostraciones matemáticas que prueban que minimizar la norma del residuo equivale a minimizar la siguiente ecuación:

$$\|e\|_p = \sum_n (w(n)^2 |e(n)|^2)^{1/2} \quad (4)$$

Para realizar exitosamente la minimización de las distancias a través de la norma, se necesita encontrar un valor óptimo de x, el cual se calcula de la siguiente manera:

$$x = [A^{\mathbf{T}}\mathrm{W}^{\mathbf{T}}\mathrm{WA}]^{-1}A^{\mathbf{T}}\mathrm{W}^{\mathbf{T}}\mathrm{Wb} \quad (5)$$

Finalmente, factorizando las ecuaciones (3) y (4), se obtiene el algoritmo IRLS:

$$\|e\|_p = \left(\sum_n |e(n)|^{(p-2)} |e(n)|^2\right)^{1/p} \quad (6)$$

Para realizar la transformación anterior a la estructura de ecuaciones ya descrita, las ponderaciones $w(n)$ son estimadas de la siguiente manera:

$$w(n) = e(n)^{\frac{(p-2)}{2}} \quad (7)$$

VI. VI. IV. III. III. V. II. Funcionamiento mecánico del Algoritmo IRLS

Paso I. En el punto inicial t, los pesos asignados en la ecuación (5) se corresponden con el caso clásico en el análisis de regresión.

Paso II. Se estiman los residuos de la ecuación (2).

Paso III. Si el resultado del paso anterior es la no-correspondencia de los residuos obtenidos con los residuos óptimos, el algoritmo procede bajo el criterio de minimización ya planteado (que es el que obedece para asignar y reasignar los pesos a x) a reestimar los residuos en el punto $t + I$, asignando nuevas ponderaciones a la diagonal principal de la matriz de ponderaciones o pesos W, mediante la utilización de (5), definiendo los nuevos pesos mediante la identidad (6).

Paso IV. El procedimiento descrito anteriormente se repite hasta que finalmente las iteraciones converjan a los valores óptimos de los residuos (correspondientes a la minimización óptima de las distancias).

VI. VI. IV. III. III. V. III. Análisis del funcionamiento del Algoritmo IRLS

Se realizan las ponderaciones a la norma y no a la métrica, puesto que la lógica de los espacios euclidianos es que la métrica es inducida por una norma, entonces la métrica es una consecuencia de la norma y no a la inversa (en los espacios definidos). Lo anterior ofrece adicionalmente ciertas ventajas, como por ejemplo ampliar los tipos de métrica para los cuales es válida la iteración algorítmica (mediante la ramificación de métricas que se pueden obtener con una norma en espacios en que la métrica es inducida por una norma). Ponderar los valores de x implica un re-escalamiento (que captura la intuición geométrica de cambiar las medidas de una determinada figura), lo que hace que decrezcan las distancias de las x_i (asignándole nuevas localizaciones a los valores x_i su nueva posición es el valor resultante de multiplicar el valor x_i por la ponderación w_i), es decir, que decrezca la varianza de x. Reducir la varianza es reducir las distancias respecto de la media, lo que implica que los elementos x_i están cada vez más próximos entre sí; lo anterior es una de las implicaciones lógicas del Teorema Central del Límite, que garantiza en ausencia de error sistemático la convergencia a una distribución de probabilidad normal. El incremento asintótico (progresivo y de largo plazo) de la proximidad entre los x_i es una consecuencia de que el algoritmo IRLS es, como su nombre lo indica, un algoritmo de modelado por una función iterada de la forma $\{x, f(x), f(f(x)), \dots\}$ y debido al teorema de la aplicación contractiva de Banach, el cual garantiza que un mapeo funcional contractivo (que es el tipo de mapeo

realizado con las sucesiones de funciones iteradas) es siempre convergente, por lo que la sucesión de funciones iteradas por las que está compuesto el algoritmo IRLS será, también, siempre convergente, cuando el punto fijo atractivo x_0 exista. Por ello, siempre que exista solución al sistema que puede obtenerse mediante mínimos cuadrados ordinarios (esta solución es el punto fijo atractivo), sea una solución exacta o una solución *sparse*[18], es posible aplicar este algoritmo.

VI. VI. IV. III. III. V. IV. Estadísticos Suficientes

Como señalan (Nelder & Wedderburn, 1972, pág. 374), un caso de especial importancia en la estimación estadística de este modelo ocurre cuando el valor del parámetro θ de la distribución del componente aleatorio y el valor pronosticado Y por el modelo lineal coinciden. En el caso descrito antes, tanto $L = zY - g(Y) + h(z)$ como $\frac{\partial L}{\partial \beta_i} = \alpha(\phi)(z-\mu)x_i$ [con base en $\frac{\partial L}{\partial \theta} = \alpha(\phi)(z-\mu)$], *i.e.*, las ecuaciones de máxima verosimilitud. Estas pueden resumirse entonces como una suma indizada a las observaciones $\sum_k (z-\hat{\mu})\, x_{ik} = 0$.

Por tanto, se obtiene la equivalencia $\sum_k z_k x_{ik} = \sum_k \hat{\mu} x_{ik}$. Para variables independientes cualitativas, esto implica que los totales marginales ajustados[19] con respecto a esa variable será igual a los observados.

De la expresión para L es posible observar que las cantidades $\sum_k z_k x_{ik}$ son conjuntos de estadísticos suficientes (que cumplen las condiciones de máxima verosimilitud establecidas). Además, en $\frac{\partial L}{\partial \beta_i} = \alpha(\phi)\left\{\frac{\left(\frac{d\mu}{dY}\right)}{V+(z-\mu)\left(\frac{d^2\theta}{dY^2}\right)}\right\}$, $\frac{d^2\theta}{dY^2} = 0$, por lo que $\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} = E\left(\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}\right) = -\alpha(\phi)\left(\frac{\left(\frac{d\mu}{dY}\right)}{V}\right)x_i x_j$.

---

[18] Una solución *sparse*, significa que la mayoría de los coeficientes de x son ceros, solo unos pocos son distintos de ceros, lo que implica que a medida la densidad o la masa de probabilidad crecen, el modelo necesita cada vez menos variables independientes para explicar a la variable dependiente; a nivel empírico esto se observa en modelos estadísticos para datos de panel en los que se utiliza en los procesos de optimización estadística únicamente algunas columnas. Una solución *sparse* es deseable puesto que con ella se podría evitar un sobreajuste del modelo estadístico en cuanto obliga al algoritmo IRLS a que optimice con menos variables y aun así lograr el resultado óptimo o cercano al óptimo.

[19] Como se señala en (Wikipedia, 2021), el *procedimiento de ajuste iterativo* (conocido también como *ajuste biproporcional* o *biproporción* en Estadística, *algoritmo RAS* en Economía en el contexto del análisis insumo-producto, *raking* en Encuestas por Muestreo y escalamiento matricial en Ciencias de la Computación) es una operación para encontrar la matriz ajustada X (*i.e.*, la matriz del conjunto de datos pronosticados) que es la más cercana a la matriz inicial Z, pero con los totales fila y los totales columna de una matriz objetivo Y (lo que proporciona las restricciones al problema de optimización; el interior de la matriz Y es desconocido. La matriz ajustada de la forma $X = PZQ$, en donde P y Q son matrices diagonales tal que X tiene los márgenes (*i.e.*, los totales por filas y columnas) de Y.

Cuando θ es también la media de la distribución, *i.e.*, $\mu = \theta = Y$, se tiene entonces el modelo lineal usual con errores normales, que para $g'(\theta) = \theta$ se obtiene el resultado $g(\theta) = \frac{1}{2}\theta^2 + constante$ que determina de forma única la distribución como normal de varianza $\frac{1}{\alpha(\phi)}$ como resultado del siguiente teorema:

(Patil & Shorrock, 1965, pág. 94), *teorema 1*: Es conocido que entre todas las familias del tipo exponencial la forma funcional de la función generadora $f(\omega)$ (una forma de codificar una sucesión infinita de elementos a_n tratándolos como si fuesen coeficientes de una serie de potencias formal -*i.e.*, sin consideraciones sobre convergencia-) caracteriza a la familia. Generalizando la afirmación anterior, sea S una sucesión con un punto límite en Ω_v, en donde v es una función no decreciente real evaluada que sirve como medida del espacio Ω. Si $\mu(\omega)$ se encuentra dentro de S, entonces la familia está determinada dentro de todas las familias exponenciales posibles.

Finalmente, la subclase de modelos para los cuales existen estadísticos suficientes fue descubierto por Cox (1968), mientras que Dempster (1971) extendió los resultados de Cox para incluir varias variables dependientes.

IV. VI. V. MODELO PROBIT

Este modelo, tal como señalan (McCullagah & Nelder, 1989, pág. 31), posee la siguiente función enlace:

$$\eta = \Phi^{-1}(\mu)$$

En la expresión anterior, conocida también como *función Probit*, $\Phi(\cdot)$ es la distribución normal acumulada, por lo que $\Phi^{-1}(\cdot)$ es la inversa de esa función, mientras que $\Phi^{-1}(\mu)$ es esta inversa en función de μ, que no es otra cosa que el valor esperado de un determinado conjunto de datos, *i.e.*, su media o esperanza matemática.

La función Probit arroja el cálculo 'inverso' al obtenido utilizando la función de densidad normal estándar. Así, esta función calcula los valores de una variable aleatoria normal estándar asociados con una probabilidad acumulada especificada.

Como se señala en (Wooldridge, 2010, pág. 565), el modelo econométrico conocido bajo el nombre de *modelo Probit* adopta, al igual que todo modelo de respuesta binaria, a la siguiente forma funcional:

$$P(y = 1|\boldsymbol{x}) = G(\boldsymbol{x\beta}) \equiv p(\boldsymbol{x})$$

Como se puede verificar en (McCullagah & Nelder, 1989, pág. 27), en la expresión anterior el conjunto de los y representa al componente aleatorio Y que sigue una distribución normal (con ampliaciones a más tipos de distribuciones), mientras que

$\boldsymbol{x}$ es una matriz de $1 \times K$ dimensiones que representa la matriz que contiene al componente (conjunto de datos) sistemático que produce al predictor lineal η (equivalente a $G(\cdot)$ en la notación usada por Wooldridge).

IV. VI. V. MODELO LOGIT

IV. VI. V. I. Introducción

Como se señala en (Aldrich & Nelson, 1984, págs. 30-31), la inferencia estadística comienza por asumir que el modelo que se va a estimar y utilizar para hacer inferencias está correctamente especificado. La presunción, *i.e.*, el supuesto de partida, es que la teoría estadística-matemática correspondiente a tal o cual modelo estadístico es la que justifica el uso del mismo. Sin embargo, a lo planteado por los autores hay que agregar que es aún más importante que las propiedades reales del fenómeno a estudiar (establecidas por el marco científico mediante el cual se estudia) deben corresponderse en una magnitud mínima necesaria y suficiente con las propiedades matemáticas de tal o cual modelo estadístico. Los autores señalan que es bastante fácil demostrar que la especificación incorrecta del modelo tiene implicaciones realmente sustanciales, ya que todas las propiedades estadísticas de las estimaciones pueden destruirse. Para decirlo sin rodeos, la especificación incorrecta del modelo conduce a respuestas incorrectas.

Los autores también elaboran una maravilla gnoseológica en su argumentación, relativa a la justificación del difundido uso del supuesto de linealidad, estableciendo una versión modificada de la navaja de Occam, una que no implica reduccionismo filosófico, como sí lo suele ser la que utilizan, por ejemplo, los bayesianos subjetivos en los modelos parsimoniosos (y fue en ese sentido en el que la criticó también Albert Einstein):

"¿Por qué es tan popular la especificación lineal? Hay dos razones básicas (y relacionadas). En la práctica, los modelos lineales son matemáticamente simples, por lo que los estadísticos han podido aprender mucho sobre ellos, y se han escrito programas de computadora para hacer la estimación. Sobre bases teóricas, la simplicidad conduce a su adopción, justificada por una versión de la navaja de Occam: en ausencia de una guía teórica en sentido contrario, comience asumiendo el caso más simple. Así, la Navaja de Occam, por implicación, diría: Con alguna orientación teórica en sentido contrario, no asuma el caso más simple." (Aldrich & Nelson, 1984, pág. 31).

IV. VI. V. II. La Función Enlace Canónico Logit

Para estudiar la función logit es de relevancia fundamental comprender previamente el concepto de *cuota estadística*. Así, como se señala en (Wikipedia, 2021), se conoce como cuota estadística, o simplemente *cuota*, a la medida de la verosimilitud de un evento/resultado, calculada como el cociente entre el número de veces que ocurre un resultado y el número de veces que no ocurre[20].

Como señalan los autores en (Aldrich & Nelson, 1984, pág. 32), el problema (desde la perspectiva matemática) con la especificación del modelo de probabilidad lineal es que $\sum b_k X_{ik}$ se usa para aproximar una cuota, P_i [$P_i \equiv P\ (Y_i = 1)$], restringida a ser de 0 a 1, mientras que $\sum b_k X_{ik}$ no está tan restringido. Una forma de abordar este problema es transformar P_i para eliminar una o ambas restricciones. Para el caso dicotómico, es posible eliminar el límite superior, $P_i = 1$, observando la relación $\frac{P_i}{1-P_i}$. Esta relación debe ser positiva (ya que $0 < P_i < 1$, que efectivamente es una restricción), pero no hay límite superior explícito en el cociente antes expuesto. El resultado de lo cual puede ser cualquier número real desde el infinito negativo hasta el positivo.

Luego se asume que la variable transformada es una función lineal de X, como se señala en (Aldrich & Nelson, 1984, pág. 32) y se presenta a continuación:

$$\log\left[\frac{P_i}{1-P_i}\right] = \sum b_k X_{ik} \equiv Z_i$$

A esto, (McCullagah & Nelder, 1989, pág. 30) le llaman *enlace canónico logit*.

Este enlace canónico fue definido como tal originalmente en (McCullagah & Nelder, 1989, pág. 31) y exposición genética, la cual se realizará respetando la notación original (que difiere en notación en cuanto $\mu = P$), permite un estudio natural y completo de este enlace canónico, en cuanto lo integra armónicamente con lo expuesto en la sección relativa a los modelos lineales generalizados. Así, la función enlace canónico η empleada por el modelo logit se expresa genéticamente:

$$\eta = \log\left(\frac{\mu}{1-\mu}\right)$$

Como señalan (McCullagah & Nelder, 1989, pág. 31), es importante conocer la forma funcional de la familia de potencias de las funciones enlace, en el contexto de los modelos lineales generalizados, al menos para observaciones cuya media es positiva. Esta familia puede especificarse mediante $\eta = \frac{(\mu^{\lambda}-1)}{\lambda}$ con la característica de que $\lim_{\lambda \to 0} \eta = \log(\mu)$, aunque también puede especificarse mediante el

[20] Este concepto, desde la Filosofía Estadística, es en esencia de espíritu frecuentista.

truncamiento $\eta = \{\mu^{\lambda}, cuando\ \lambda \neq 0;\ \log(\mu)\ cuando\ \lambda = 0\}$. La primera forma de especificación tiene la ventaja de experimentar una transición suave (del inglés "smooth"[21]), aunque con ambas hay que realizar acciones complementarias especiales en el caso de $\lambda = 0$.

Como señala (Wikipedia, 2021), los logaritmos de las cuotas estadísticas anteriores pueden expresarse en términos de probabilidades, estimando la función inversa del enlace canónico logit arriba presentado. Es precisamente a esta inversa que se le conoce como *función logística*, alrededor de la cual se realiza la *regresión logística*. Así, es posible aplicar antilogaritmos y posteriormente realizar manipulación algebraica para aislar (despejar) el monomio P_i que se presenta a continuación:

$$P_i = \frac{e^{Z_i}}{1 + e^{Z_i}}$$

La expresión anterior es conocida como la *función logística*, la cual es continua y puede tomar valores únicamente entre 0 y 1. Es cercana a 0 cuando Z_i tiende a infinito negativo, incrementa monotónicamente con Z_i y es cercana a 1 cuando Z_i tiende a infinito positivo. Forma así una curva suave en forma de S, la cual es simétrica alrededor del punto Z_i tal como se muestra a continuación.

[21] Como se señala en (Weisstein, Smooth Function, 2021), una función suave es aquella para cuyas derivadas de orden superior (primera, segunda, tercera, …) derivadas son continuas sobre algún dominio (*i.e.*, existen) en el cual se estudia la función. Se puede decir entonces que una función es suave sobre algún intervalo restringido como (a, b) o $[a, b]$. El número de derivadas que deben existir para que la función sea considerada suave depende del problema que se estudie, sin embargo, puede variar de dos a infinito. Una función para la cual existen derivadas en todos los órdenes y en la que además estas son todas continuas, se conoce como *función-C-infinito*.

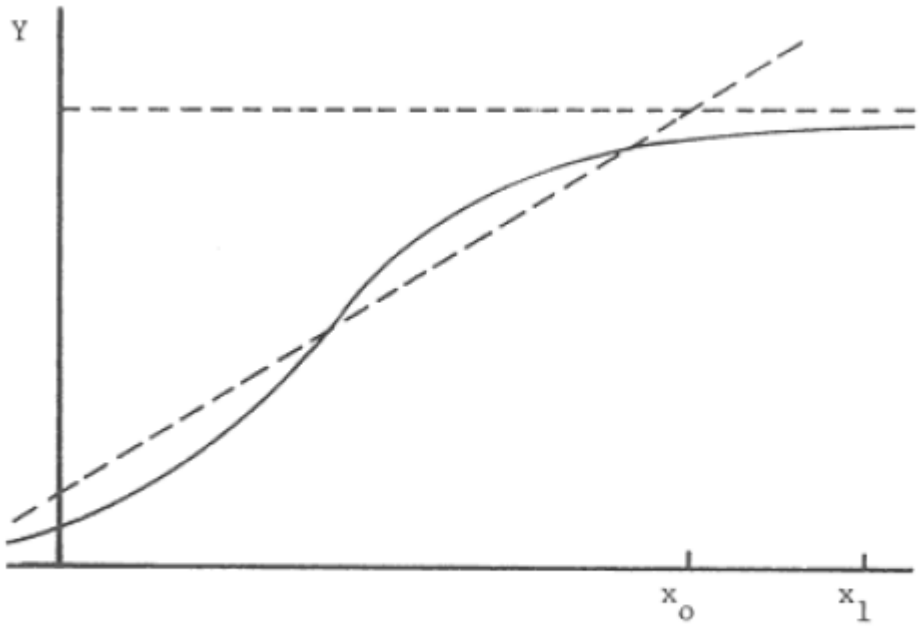


Figure 1
Sigmoid Versus Linear Specifications

Fuente: (Aldrich & Nelson, 1984, pág. 27).

Así, a diferencia del caso de especificación lineal, esto satisface la restricción de encontrarse entre 0 y 1 sin necesidad de también restringir Z_i, es decir, sin restringir $\sum b_k X_{ik}$.

Como señalan los autores, "Estas características de la función descrita en la ecuación 2.4 la convierten en una alternativa atractiva al modelo de probabilidad lineal para variables dependientes dicotómicas. Puede ser bastante razonable, pero ¿por qué este? ¿Por qué no otros? Después de todo, es tan arbitrario como escoger la linealidad. De hecho, hay un número infinito de alternativas a la ecuación 2.4 y, al igual que con la ecuación 2.4, algunas de ellas se han desarrollado como modelos alternativos para la estimación. Siete de estos se muestran en la Figura 3, y los describiremos brevemente para proporcionar una indicación del amplio "menú" de opciones disponibles." (Aldrich & Nelson, 1984, pág. 32).

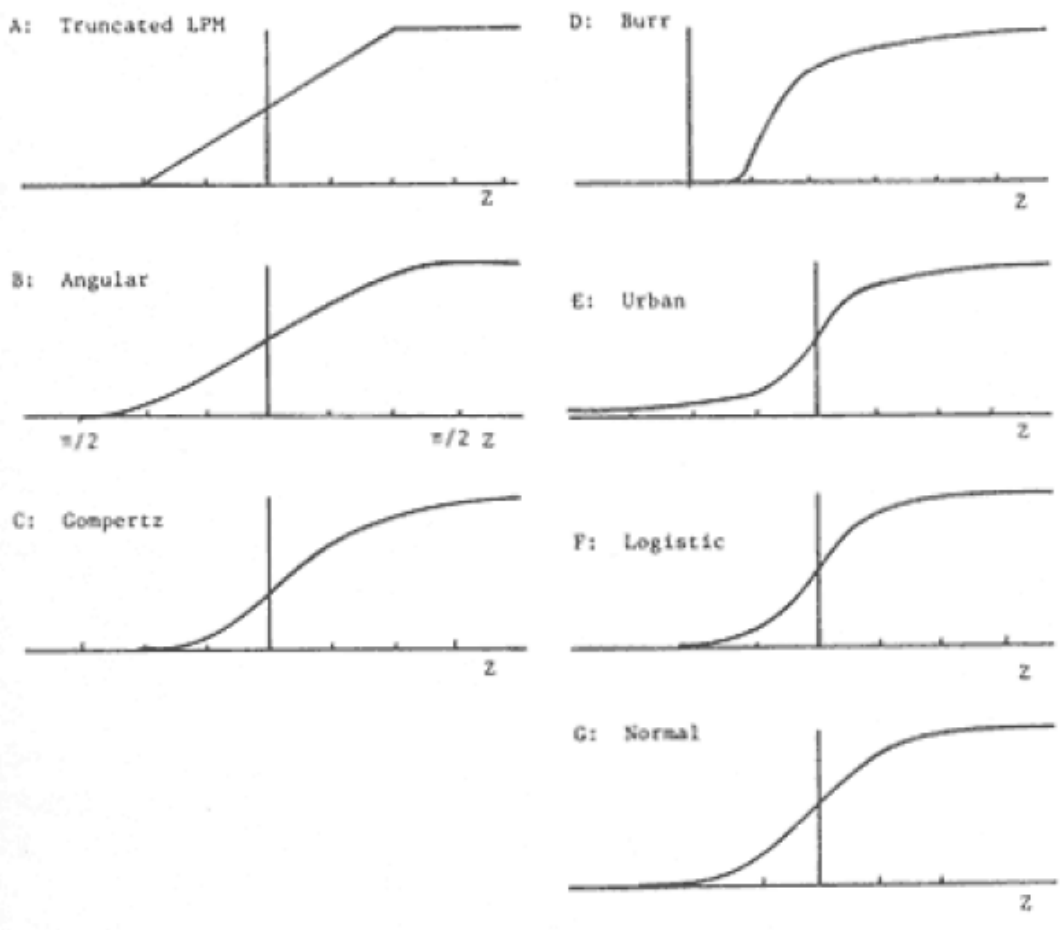


Figure 3
Graphs of Alternative Specifications

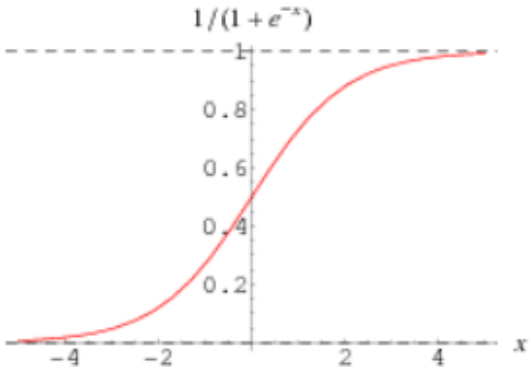
Fuente: (Aldrich & Nelson, 1984, pág. 33).

"Las curvas logísticas y normales son tan similares que producen resultados esencialmente idénticos. En la práctica, arrojan probabilidades de elección estimadas que difieren en menos de 0,02 y que pueden distinguirse, en el sentido de significancia estadística, solo con muestras muy grandes. La elección entre ellos, por lo tanto, gira en torno a preocupaciones prácticas como la disponibilidad y flexibilidad de los programas de computadora y las preferencias y experiencias personales. Estos dos han recibido la mayor atención por parte de investigadores (y programadores de computadoras). Es decir, están mejor desarrollados que los otros ejemplos, se usan mucho más ampliamente y están disponibles en programas de computadora con mucha más frecuencia. Su importancia es tal que centraremos la mayor parte de nuestra atención en ellos. Las curvas logísticas y normales son tan similares que producen resultados esencialmente idénticos. En la práctica, arrojan probabilidades de elección estimadas que difieren en menos de 0,02 y que

pueden distinguirse, en el sentido de significancia estadística, solo con muestras muy grandes. La elección entre ellos, por lo tanto, gira en torno a preocupaciones prácticas como la disponibilidad y flexibilidad de los programas de computadora y las preferencias y experiencias personales. Estos dos han recibido la mayor atención por parte de investigadores (y programadores de computadoras). Es decir, están mejor desarrollados que los otros ejemplos, se usan mucho más ampliamente y están disponibles en programas de computadora con mucha más frecuencia. Su importancia es tal que centraremos la mayor parte de nuestra atención en ellos." (Aldrich & Nelson, 1984, pág. 34).

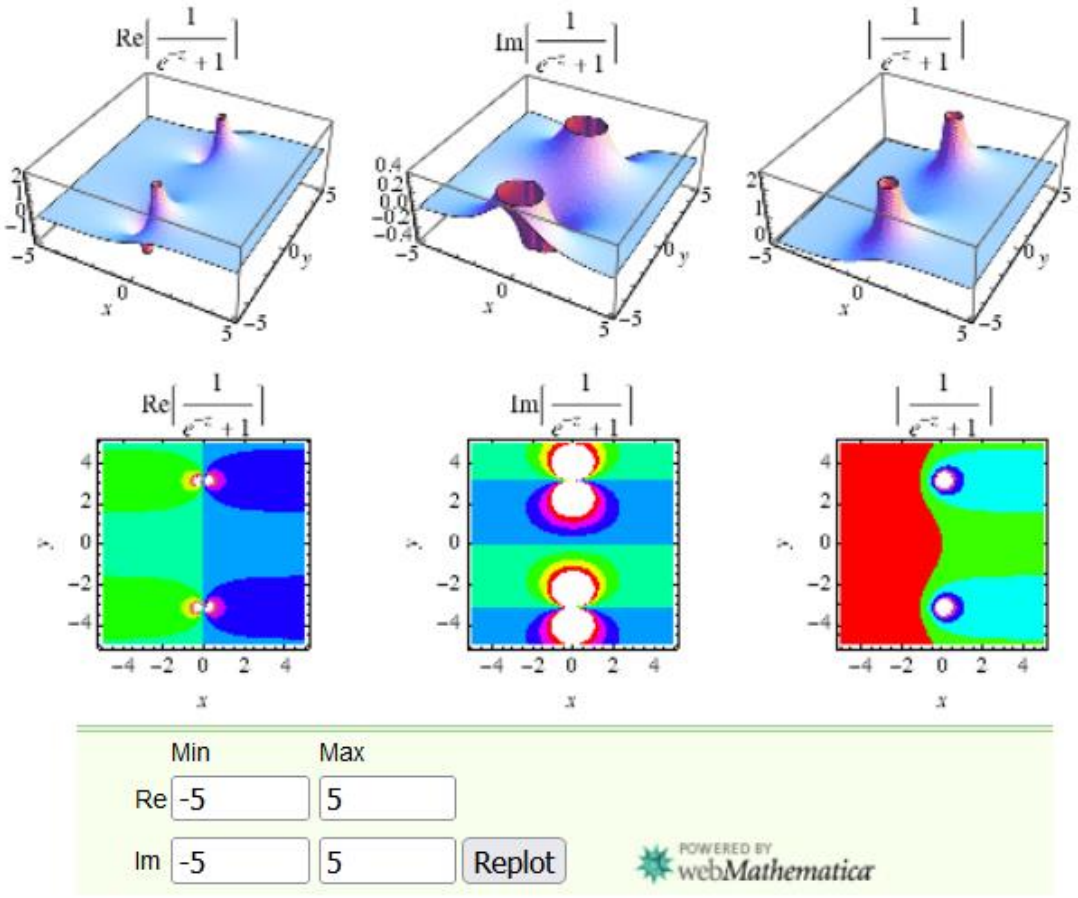
Señala además (Liao, 1994, pág. 11), el *modelo logit* o *regresión logística*[22] asume que el conjunto de datos sigue una distribución binomial. Agrega también que "Para interpretar los resultados de un modelo logit de manera significativa, el modelo en sí debe ser capaz de explicar la variable de respuesta significativamente mejor que el modelo con la intersección solamente. Esto es cierto para todos los modelos lineales generalizados. En un modelo de regresión clásico, se utiliza una prueba F; En un modelo logit (y otros modelos de probabilidad), la prueba más comúnmente utilizada es el estadístico de razón de verosimilitud, que sigue aproximadamente la distribución chi-cuadrado (ver Aldrich y Nelson, 1984; Greene, 1990; McCullagh y Nelder, 1989; entre otros)." (Liao, 1994, págs. 12-13).

Esta distribución adopta empíricamente formas como las que se presentan a continuación:



Fuente: (Weisstein, Sigmoid Function, 2021).

[22] Se le conoce como modelo logit porque la regresión logística emplea como función enlace



Fuente: (Weisstein, Sigmoid Function, 2021).

IV. VI. VI. INTERPRETACIÓN GENERAL DE LOS PRINCIPALES ESTADÍSTICOS EN LOS MODELOS LINEALES GENERALIZADOS

Finalmente, la interpretación de los resultados estadísticos de las regresiones logísticas debe realizarse, en el contexto aplicado de la investigación clínica (y de forma equivalente en otros contextos), tal como se presenta a continuación:

"Es necesario que los médicos pongan su atención en los siguientes estadígrafos y su interpretación cuando se encuentren frente a trabajos que aplican la regresión logística múltiple: a) ajustes del modelo y pruebas de bondad de ajuste; b) prueba de Omnibús; c) valor de la verosimilitud, d) prueba de Hosmer y Lebeshow; y e) tabla de clasificación. Atendidos los anteriores, el de mayor relevancia es el coeficiente de determinación: R^2 de Cox y Snell y R^2 de Nagelkerke. Cualquier coeficiente de determinación pretende estimar en qué grado una variable independiente o un conjunto de ellas pueden explicar la varianza de la variable dependiente. El R^2 de Cox y Snell es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras. Su valor fluctúa entre 0 y 1, pero en la

práctica no llega a 1. El R^2 de Nagelkerke es una transformación del R^2 de Cox y Snell. Este estadígrafo corrige la escala del estadístico para cubrir el rango completo de 0 a 1. Si el R^2 de Cox y Snell estimado fuera de 0,021 implicaría que las variables independientes empleadas en el modelo de regresión solamente explican el 2,1 % de la varianza de la variable dependiente y esto pudiera ocurrir con valores de riesgo altos y variables muy significativas. Como consecuencia, el investigador tiene la obligación de informar el coeficiente de determinación para que el clínico, a su vez, pueda tomar las decisiones pertinentes. El clínico, y solo él, decide si esta información puede serle útil o no. Por otra parte, cuando no se reportan estos indicadores, los resultados de la investigación están incompletos y las conclusiones presentadas en el artículo pueden estar sesgadas. Sobre la base de lo anterior, es necesario que los autores presenten de forma rutinaria estos indicadores que permitan la justificación de su aplicación clínica y los clínicos puedan tomar la mejor decisión al respecto. Los árbitros deben exigir la presentación de estas estimaciones de forma rutinaria." (Díaz-Narváez, 2017, pág. 1505).

IV. VI. VI. VARIABLES DICOTÓMICAS CONSTRUIDAS

Códigos de preguntas usadas	*Regla de Construcción*	*Nombre de la variable binaria*
SD1	Hombre = 1, Mujer = 0	SD1 Binaria
TE1	Si responden a TE1 con 2 o 3 = 1, en otro caso = 0	Tenencia de Capital (o no)
SD3	Sí = 1, No = 0	Nacional (o no)
SD2	Respuesta > 25 = 1, caso contrario = 0)	Adulto (o no)
E1	Sí = 1, No = 0	Acceso a educación virtual (o no)
FN6, FN7 y FN8	Respuestas negativas > respuestas neutras + respuestas positivas	Empeoramiento frecuencia (o no) de dieta
AE3-AE9	Fuentes contraídas > fuentes expandidas	Contracción (o no) del ingreso
FN9-FN15	Disminuciones totales de FN9-FN13 > Disminuciones totales FN14-FN15	Empeoramiento (o no) calidad de dieta

IV. VI. VII. TEORÍA DEL APRENDIZAJE ESTADÍSTICO

IV. VI. VII. I. Definición General de Aprendizaje Estadístico

Según (James, Witten, Hastie, & Tibshirani, 2013, pág. 17), en esencia, el aprendizaje estadístico hace referencia a un conjunto de abordajes para predecir una función f a partir de conjuntos de variables dependientes e independientes. Complementariamente, se señala en (Hastie, Tibshirani, & Friedman, 2017, págs. xi-xii) que esto desempeña en la actualidad un papel fundamental en áreas como la agricultura, la industria, el almacenamiento de datos (que dio origen a la minería de datos), en la bioinformática, en la medicina y en muchos otros campos del conocimiento humano, en donde los modelos estadísticos son utilizados exclusivamente con fines predictivos partiendo de un determinado conjunto de datos, con independencia de otros factores. A esto se le conoce como *aprender de los datos*. De la evolución de los procesos de aprendizaje de datos surge el campo multidisciplinario conocido como *Aprendizaje Automático* (*Machine Learning*, en inglés). Así, los objetivos de este campo consisten en clasificar y predecir conjuntos de datos, para lo cual utilizan el marco teórico de la estadística matemática.

Como señalan (StackExchange Data Science, 2016) y (StackOverFlow, 2014), en general, la estadística se preocupa más por inferir parámetros (lo que implica validar que estadísticos muestrales se corresponden con sus versiones poblaciones), mientras que, en el aprendizaje automático, la predicción y la clasificación son el objetivo final.

Con respecto a la predicción, las ciencias de la estadística y el aprendizaje automático comenzaron a resolver casi el mismo problema desde diferentes perspectivas. Básicamente, la estadística asume que los datos fueron producidos por un determinado modelo estocástico. Así, desde una perspectiva estadística, se asume un modelo y, dados varios supuestos, se tratan los errores y se infieren los parámetros del modelo y otras cuestiones.

El aprendizaje automático nace bajo una visión informática de la manipulación de los datos. Por ello, los modelos son algorítmicos y, por lo general, se requieren muy pocas suposiciones con respecto a los datos. Es por ello que, como se adelantó, también usa, al igual que la Estadística, las herramientas del análisis funcional, como por ejemplo al construir los espacios de hipótesis (en el mismo sentido en que fueron planteados por Jerzey Neyman y Egon Pearson), así como también al hablar del sesgo de aprendizaje (conocido también como sesgo de inducción, sesgo de aprendizaje automático o sesgo de inteligencia artificial).

Lo anterior ha contribuido en buena medida a que, a pesar de que las dos ciencias no parecieran terminar de converger en términos gnoseológicos por su diferente aparentemente diferente genética filosófica (el espíritu conceptual bajo el cual nacieron), metodológicamente cada vez existe una mayor convergencia, expresada en que ambas comparten cada vez mayor cantidad de conocimientos y técnicas comunes. Existe por supuesto una base material a este hecho, la cual radica en que en que los problemas que enfrentaban tenían en común que podían ser resueltos mediante la determinación de tal o cual patrón geométrico del conjunto de datos (como se adelantó al introducir el marco teórico de los GLM), sin embargo, en los albores del aprendizaje automático esta compatibilidad de instrumentos no fue tan marcada, como ahora que el aprendizaje automático tiende a abordarse cada vez más desde una perspectiva estadística. Complementariamente, el aprendizaje se clasifica en aprendizaje supervisado, aprendizaje no supervisado, aprendizaje en línea y aprendizaje por refuerzo. Ejemplos de aprendizaje no supervisado incluyen el análisis de agrupaciones y asociaciones.

IV. VI. VII. II. Positivos y Negativos en la Predicción / Clasificación

En el contexto de los modelos de variables dicotómicas, conocidos también como modelos de respuesta binaria, existen dos tipos de positivo y dos tipos de negativo.

- *Verdaderos Positivos:* cuando el valor de las observaciones es "Sí" y el valor predicho por el modelo es "Sí".
- *Falso Positivo:* cuando el valor de las observaciones es "No" y el valor predicho por el modelo es "Sí".
- *Verdadero Negativo:* cuando el valor de las observaciones es "No" y el valor predicho por el modelo es "No".
- *Falso Negativo:* cuando el valor de las observaciones es "Sí" y el valor predicho por el modelo es "No".

Véase el siguiente ejemplo en el contexto de la detección de tumores benignos o malignos:

Verdadero positivo (VP): • Realidad: Maligno • Predicción del modelo de AA: Maligno • **Número de resultados de VP: 1**	**Falso positivo (FP):** • Realidad: Benigno • Predicción del modelo de AA: Maligno • **Número de resultados de FP: 1**
Falso negativo (FN): • Realidad: Maligno • Predicción del modelo de AA: Benigno • **Número de resultados de FN: 8**	**Verdadero negativo (VN):** • Realidad: Benigno • Predicción del modelo de AA: Benigno • **Número de resultados de VN: 90**

Fuente: (Google Developers, 2021).

IV. VI. VII. III. Matriz de Confusión

Como se señala en (James, Witten, Hastie, & Tibshirani, 2013, pág. 145), una matriz de confusión compara las predicciones del modelo de aprendizaje estadístico seleccionado con los verdaderos valores contenidos en las observaciones de entrenamiento del conjunto de datos (*default data set*). Los elementos en la diagonal de la matriz representan observaciones cuyos valores fueron correctamente predichos/clasificados por el modelo, mientras que fuera de la diagonal de la matriz se encuentran aquellos valores que fueron inadecuadamente predichos/clasificados. A continuación, se presenta una matriz de confusión para el caso de un análisis de discriminante lineal.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for* 23 *individuals who did not default and for* 252 *individuals who did default.*

Fuente: (James, Witten, Hastie, & Tibshirani, 2013, pág. 145).

Finalmente, cabe decir, con base en (Barrios, 2019), que una matriz de confusión es esencialmente una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado.

IV. VI. VII. IV. Exactitud, Tasa de Error, Sensibilidad, Especificidad, Precisión y Predicción Negativa del Modelo de Aprendizaje

IV. VI. VII. IV. I. Exactitud

Como se señala en (Google Developers, 2021), la exactitud es una métrica para evaluar modelos de clasificación. Informalmente, la exactitud es la fracción de predicciones que el modelo realizó correctamente. Formalmente, la exactitud tiene la siguiente:

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}$$

IV. VI. VII. IV. II. Tasa de Error del Entrenamiento

Como se señala en (James, Witten, Hastie, & Tibshirani, 2013, pág. 37), el abordaje más común para estimar la precisión del modelo de entrenamiento $\hat{f}$ es el

coeficiente conocido como *tasa de error del entrenamiento*, equivalente al cociente entre los errores cometidos si se aplica $\hat{f}$ a las observaciones de entrenamiento.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

En donde $\hat{y}_i$ es la etiqueta de clase predicha (el valor de la variable dependiente pronosticado) para la $i-$ é$sima$ observación usando $\hat{f}$, mientras que y_i es su valor real (el contenido en el conjunto de datos de entrenamiento). En la expresión anterior, $I(y_i \neq \hat{y}_i)$ es la función indicatriz que toma el valor 1 cuando la predicción es incorrecta y toma el valor 0 cuando la predicción es correcta. En términos de positivos y negativos lo anterior equivale a decir:

$$\frac{Falsos\ positivos\ +\ Falsos\ negativos}{Total\ de\ predicciones}$$

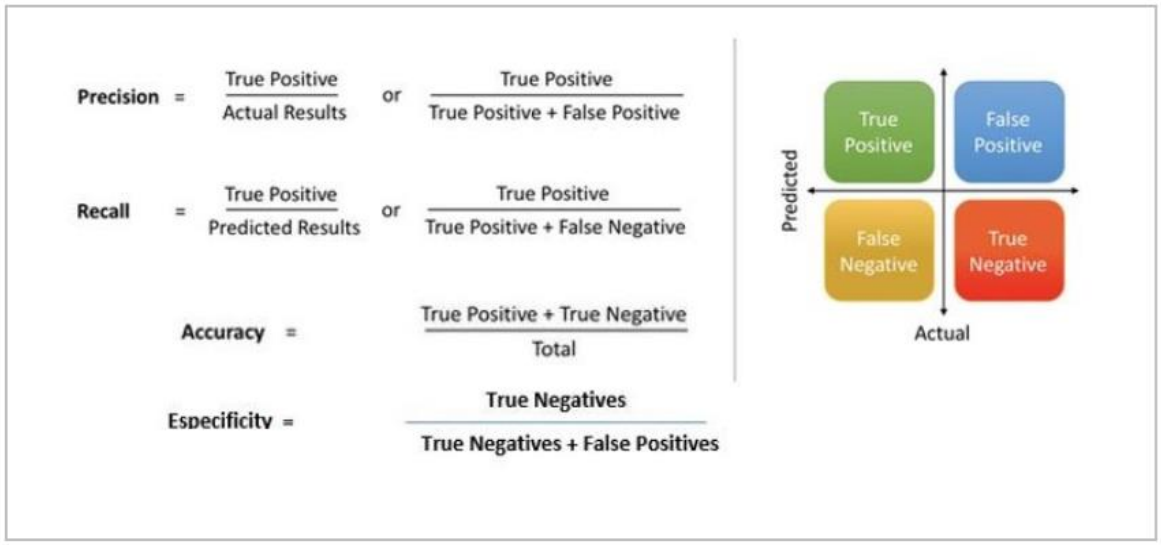
IV. VI. VII. IV. III. Sensibilidad y Especificidad

Como señala (James, Witten, Hastie, & Tibshirani, 2013, pág. 145), el rendimiento específico de la clase también es importante en medicina y biología, donde los términos *sensibilidad* y *especificidad* caracterizan el desempeño de sensibilidad y especificidad un clasificador o prueba de detección. La sensibilidad es el porcentaje de verdaderos positivos que se identifican en relación a los positivos reales totales. Por su parte, la especificidad es el porcentaje de verdaderos negativos que se identifican correctamente en relación a los negativos reales totales.

IV. VI. VII. IV. IV. Precisión (Valor Predictivo Positivo) y Valor Predictivo Negativo

Como señala (Barrios, 2019), el concepto de *precisión* refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión, mayor la precisión. Se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos). Este concepto, como se señala en (Wikipedia, 2021), es conocido a menudo como *valor predictivo positivo*. Su equivalente inverso es el concepto de *valor predictivo negativo*, estimado de forma equivalente, aunque considerando los negativos.

IV. VI. VII. IV. V. Sumario



Fuente: (Barrios, 2019).

Matriz de Confusion		Predicho			
		Negativo	Positivo		
Real	Negativo	a	b	*Verdadero Negativo (True negative rate)*	a/(a+b)
	Positivo	c	d	*Exactitud*	d/(c+d)
		Sensibilidad	*Especifidad*		
		d/(d+c)	a/(a+b)	**Precisión**=(a+d)/(a+b+c+d)	

Figura 3: Matriz de confusión con otras métricas de evaluación.

a: es el número de predicciones correctas de clase negativa (negativos reales)

b: es el número de predicciones incorrectas de clase positiva (falsos positivos)

c: es el número de predicciones incorrectas de clase negativa (falsos negativos)

d: es el número de predicciones correctas de clase positiva (positivos reales)

Fuente: (Barrios, 2019).

IV. VI. VII. V. Modelos Lineales Generalizados desde la Teoría del Aprendizaje Estadístico

Como señalan (StackExchange Data Science, 2016) y (StackOverFlow, 2014), los modelos lineales generalizados son un desarrollo estadístico. Sin embargo, los nuevos tratamientos bayesianos ponen este algoritmo también en el campo de juego del aprendizaje automático. Entonces creo que ambas afirmaciones podrían ser correctas, ya que la interpretación y el tratamiento de cómo funciona podrían ser diferentes.

La distinción sutil entre modelos estadísticos y modelos de aprendizaje automático es que, en los modelos estadísticos, usted decide explícitamente la estructura de la ecuación de salida antes de construir el modelo. El modelo está construido para calcular los parámetros/coeficientes.

Tómense precisamente los GLM, que son modelos estadísticos y, por consiguiente, útiles para verificar que los modelos estadísticos y las técnicas de aprendizaje automático no son mutuamente excluyentes.

$$y = a_1x_1 + a_2x_2 + a_3x_3$$

Las variables independientes son x_1, x_2 y x_3, mientras que los coeficientes a determinar son a_1, a_2 y a_3. Así, se define la estructura de su ecuación de esta manera antes de construir el modelo y calcule a_1, a_2 y a_3. Si se cree que y está correlacionada de alguna manera con x_2 de forma no lineal, puede probarse una transformación como la siguiente:

$$y = a_1x_1 + a_2(x_2)^2 + a_3x_3$$

Evidentemente, la transformación anterior implica imponer una restricción en términos de la estructura de salida. En el caso de los modelos de aprendizaje automático, rara vez se especifica la estructura de salida y los algoritmos, como los árboles de decisión, son intrínsecamente no lineales y funcionan de manera eficiente. Simplemente se parte de un conjunto de datos con una variable dependiente conocida (etiqueta), se "entrena el modelo" su modelo y luego se aplica al conjunto de datos para intenta predecir un número real, como por ejemplo el precio de una casa[23].

---

[23] En este sentido, una aplicación industrial exitosa de los GLM puede encontrarse en http://www.kdd.org/kdd2016/papers/files/adf0562-zhangA.pdf y contribuir a explicar por qué los modelos lineales generalizados son considerados por muchos como una técnica del aprendizaje automático, aun cuando en términos históricos y teóricos no lo son. Se afirma lo anterior puesto que, por ejemplo, muchos afirman que la regresión logística no es en realidad una regresión, lo que justifican planteando que, por lo general, solo se usa para la predicción binaria, lo que es idónea para tareas de clasificación. Por supuesto, estas creencias son refutadas por los mismos orígenes

Específicamente, los GLM, así como cualquier metodología estadística de regresión pertenece al aprendizaje supervisado, por cuanto los datos que tiene incluyen tanto la entrada como la salida, por ponerlo en algunos términos. Entonces, por ejemplo, si se tiene un conjunto de datos para, supóngase, las ventas de automóviles en un concesionario. Se tiene, para cada coche, características como marca, modelo, precio, color, descuento, etc., pero también se tiene el número de ventas de cada coche. Si esta tarea no estuviera supervisada, se tendría un conjunto de datos que incluye, tal vez, solo la marca, el modelo, el precio, el color, etc. (no el número real de ventas) y lo mejor que se puede hacer es agrupar los datos. El ejemplo no es perfecto, pero tiene como objetivo transmitir el panorama general. Una buena pregunta que debe hacerse al decidir si un método está supervisado o no es preguntarse "¿Se cuenta con alguna forma de juzgar la calidad de una entrada?". Si se cuenta con datos de regresión lineal, la respuesta es afirmativa. Simplemente se evalúa el valor de la función (en este caso, la función lineal) de los datos de entrada para estimar la salida. No es así en el otro caso. También se supervisa la regresión logística.

IV. VI. VII. PRESENTACIÓN DE ALGUNOS RESULTADOS ESTADÍSTICOS CON SPSS

VI. VI.VII. I. Matriz de Correlaciones

```
CORRELATIONS
  /VARIABLES=Sexo TenenciaCapital EmpeoramientoDieta EstadoMigratorio
Adulto_o_joven
    AccesoEducacionVirtual Contracción_o_no_ingresos
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Correlaciones

Notas

Salida creada	17-JUL-2021 16:43:58

históricos y teóricos de los GLM, los cuales han sido estudiados en esta investigación. Por supuesto, de forma artificial puede concebirse en el contexto del aprendizaje automático como un método de clasificación, aunque durante el entrenamiento lo que hace es predecir si un valor pertenece a una clasificación o no, lo que prueba en última instancia muestra cómo los orígenes antes referido determinan la naturaleza del método estadístico.

Comentarios		
Entrada	Conjunto de datos activo	ConjuntoDatos1
	Filtro	<ninguno>
	Ponderación	<ninguno>
	Segmentar archivo	<ninguno>
	N de filas en el archivo de datos de trabajo	235
Manejo de valores perdidos	Definición de perdidos	Los valores perdidos definidos por el usuario se tratan como perdidos.
	Casos utilizados	Las estadísticas para cada par de variables se basan en todos los casos con datos válidos para dicho par.
Sintaxis		CORRELATIONS /VARIABLES=Sexo TenenciaCapital EmpeoramientoDieta EstadoMigratorio Adulto_o_joven AccesoEducacionVirtual Contracción_o_no_ingresos /PRINT=TWOTAIL NOSIG /MISSING=PAIRWISE.
Recursos	Tiempo de procesador	00:00:00,08
	Tiempo transcurrido	00:00:00,45

Correlaciones

		Sexo	TenenciaCapital	EmpeoramientoDieta
Sexo	Correlación de Pearson	1	.129*	.110
	Sig. (bilateral)		.048	.092
	N	235	235	235
TenenciaCapital	Correlación de Pearson	.129*	1	.011
	Sig. (bilateral)	.048		.865
	N	235	235	235
EmpeoramientoDieta	Correlación de Pearson	.110	.011	1
	Sig. (bilateral)	.092	.865	
	N	235	235	235
EstadoMigratorio	Correlación de Pearson	.006	.069	-.011
	Sig. (bilateral)	.925	.289	.863
	N	235	235	235
Adulto_o_joven	Correlación de Pearson	-.125	.085	-.091
	Sig. (bilateral)	.056	.195	.166
	N	235	235	235
AccesoEducacionVirtual	Correlación de Pearson	-.063	-.107	-.004
	Sig. (bilateral)	.333	.102	.954
	N	235	235	235
Contracción_o_no_ingresos	Correlación de Pearson	.029	.070	-.063
	Sig. (bilateral)	.653	.286	.340
	N	235	235	235

Correlaciones

		EstadoMigratorio	Adulto_o_joven
Sexo	Correlación de Pearson	.006	-.125
	Sig. (bilateral)	.925	.056
	N	235	235
TenenciaCapital	Correlación de Pearson	.069	.085

	Sig. (bilateral)	.289	.195
	N	235	235
EmpeoramientoDieta	Correlación de Pearson	-.011	-.091
	Sig. (bilateral)	.863	.166
	N	235	235
EstadoMigratorio	Correlación de Pearson	1	.014
	Sig. (bilateral)		.828
	N	235	235
Adulto_o_joven	Correlación de Pearson	.014	1
	Sig. (bilateral)	.828	
	N	235	235
AccesoEducacionVirtual	Correlación de Pearson	.084	-.262**
	Sig. (bilateral)	.200	.000
	N	235	235
Contracción_o_no_ingresos	Correlación de Pearson	-.024	.046
	Sig. (bilateral)	.717	.487
	N	235	235

Correlaciones

		AccesoEducacionVirtual	Contracción_o_no_ingresos
Sexo	Correlación de Pearson	-.063	.029
	Sig. (bilateral)	.333	.653
	N	235	235
TenenciaCapital	Correlación de Pearson	-.107	.070
	Sig. (bilateral)	.102	.286
	N	235	235
EmpeoramientoDieta	Correlación de Pearson	-.004	-.063
	Sig. (bilateral)	.954	.340
	N	235	235
EstadoMigratorio	Correlación de Pearson	.084	-.024
	Sig. (bilateral)	.200	.717
	N	235	235
Adulto_o_joven	Correlación de Pearson	-.262**	.046
	Sig. (bilateral)	.000	.487
	N	235	235
AccesoEducacionVirtual	Correlación de Pearson	1	-.018
	Sig. (bilateral)		.781

	N	235	235
Contracción_o_no_ingresos	Correlación de Pearson	-.018	1
	Sig. (bilateral)	.781	
	N	235	235

*. La correlación es significativa en el nivel 0,05 (bilateral).
**. La correlación es significativa en el nivel 0,01 (bilateral).

VI. VI.VII. II. Resultados Estadísticos en SPSS para el Modelo Probit

```
* Modelos lineales generalizados.
GENLIN TenenciaCapital (REFERENCE=LAST) BY Sexo EstadoMigratorio
Adulto_o_joven (ORDER=ASCENDING)
   WITH AccesoEducacionVirtual
  /MODEL Sexo EstadoMigratorio Adulto_o_joven AccesoEducacionVirtual
INTERCEPT=YES
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL
MAXITERATIONS=100 MAXSTEPHALVING=5
   PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012
ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
   LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Modelos lineales generalizados

Notas

Salida creada		15-JUL-2021 16:08:01
Comentarios		
Entrada	Conjunto de datos activo	ConjuntoDatos1
	Filtro	<ninguno>
	Ponderación	<ninguno>
	Segmentar archivo	<ninguno>
	N de filas en el archivo de datos de trabajo	235

Gestión de valores perdidos	Definición de perdidos	Los valores perdidos definidos por el usuario para las variables de factor, sujeto e intra-sujetos se tratan como perdidos.
	Casos utilizados	Las estadísticas se basan en casos con datos válidos para todas las variables del modelo.
Manejo de ponderaciones		no aplicable

Sintaxis	GENLIN TenenciaCapital (REFERENCE=LAST) BY Sexo EstadoMigratorio Adulto_o_joven (ORDER=ASCENDING) WITH AccesoEducacionVirtual /MODEL Sexo EstadoMigratorio Adulto_o_joven AccesoEducacionVirtual INTERCEPT=YES DISTRIBUTION=BINOMIAL LINK=LOGIT /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL /MISSING CLASSMISSING=EXCLUDE /PRINT CPS DESCRIPTIVES

		MODELINFO FIT SUMMARY SOLUTION.
Recursos	Tiempo de procesador	00:00:00,11
	Tiempo transcurrido	00:00:00,46

Información de modelo

Variable dependiente	TenenciaCapital[a]
Distribución de probabilidad	Binomial
Función de enlace	Logit

a. El procedimiento modela 0 como la respuesta, tratando a 1 como la categoría de referencia.

Resumen de procesamiento de casos

	N	Porcentaje
Incluido	235	100.0%
Excluido	0	0.0%
Total	235	100.0%

Información de variable categórica

			N	Porcentaje
Variable dependiente	TenenciaCapital	0	224	95.3%
		1	11	4.7%
		Total	235	100.0%
Factor	Sexo	0	132	56.2%
		1	103	43.8%
		Total	235	100.0%
	EstadoMigrato	0	21	8.9%

	rio	1	214	91.1%
		Total	235	100.0%
	Adulto_o_joven	0	30	12.8%
		1	205	87.2%
		Total	235	100.0%

Información de variable continua

		N	Mínimo	Máximo	Media	Desv. Desviación
Covariable	AccesoEducacionVirtual	235	0	1	.31	.465

Bondad de ajuste[a]

	Valor	gl	Valor/gl
Desvianza	.877	8	.110
Desvianza escalada	.877	8	
Chi-cuadrado de Pearson	.568	8	.071
Chi-cuadrado de Pearson escalado	.568	8	
Logaritmo de verosimilitud[b]	-4.728		
Criterio de información Akaike (AIC)	19.455		
AIC corregido para muestras finitas (AICC)	19.717		
Criterio de información bayesiana (BIC)	36.753		
AIC coherente (CAIC)	41.753		

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual[a]

a. Los criterios de información están en un formato de cuanto más pequeño mejor.
b. La función de logaritmo de la verosimilitud completa se visualiza y utiliza en el cálculo de los criterios de información.

Prueba ómnibus[a]

Chi-cuadrado de razón de verosimilitud	gl	Sig.
12.035	4	.017

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual[a]
a. Compara el modelo ajustado con el modelo de sólo intersección.

Pruebas de efectos del modelo

Origen	Tipo III Chi-cuadrado de Wald	gl	Sig.
(Intersección)	.000	1	.998
Sexo	3.654	1	.056
EstadoMigratorio	.000	1	.999
Adulto_o_joven	.000	1	.999
AccesoEducacionVirtual	1.624	1	.203

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual

Estimaciones de parámetro

Parámetro	B	Desv. Error	95% de intervalo de confianza de Wald		Contraste de hipótesis Chi-cuadrado de Wald
			Inferior	Superior	
(Intersección)	1.953	.3886	1.191	2.714	25.245
[Sexo=0]	1.334	.6978	-.034	2.701	3.654
[Sexo=1]	0[a]	.	.	.	.
[EstadoMigratorio=0]	19.714	16272.3273	-31873.461	31912.890	.000
[EstadoMigratorio=1]	0[a]	.	.	.	.
[Adulto_o_joven=0]	19.493	13449.1683	-26340.392	26379.378	.000
[Adulto_o_joven=1]	0[a]	.	.	.	.
AccesoEducacionVirtual	1.360	1.0677	-.732	3.453	1.624
(Escala)	1[b]				

Estimaciones de parámetro

Parámetro	Contraste de hipótesis gl	Sig.
(Intersección)	1	.000
[Sexo=0]	1	.056
[Sexo=1]	.	.
[EstadoMigratorio=0]	1	.999
[EstadoMigratorio=1]	.	.

[Adulto_o_joven=0]	1	.999
[Adulto_o_joven=1]	.	.
AccesoEducacionVirtual	1	.203
(Escala)		

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual
a. Definido en cero porque este parámetro es redundante.
b. Fijado en el valor visualizado.

VI. VI.VII. III. Resultados Estadísticos en SPSS para el Modelo Logit

```
* Modelos lineales generalizados.
GENLIN TenenciaCapital (REFERENCE=LAST) BY Sexo EstadoMigratorio
Adulto_o_joven (ORDER=ASCENDING)
   WITH AccesoEducacionVirtual
  /MODEL Sexo EstadoMigratorio Adulto_o_joven AccesoEducacionVirtual
INTERCEPT=YES
 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL
MAXITERATIONS=100 MAXSTEPHALVING=5
   PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012
ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD
   LIKELIHOOD=FULL
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Notas

Salida creada		15-JUL-2021 16:14:09
Comentarios		
Entrada	Conjunto de datos activo	ConjuntoDatos1
	Filtro	<ninguno>
	Ponderación	<ninguno>
	Segmentar archivo	<ninguno>
	N de filas en el archivo de datos de trabajo	235

Gestión de valores perdidos	Definición de perdidos	Los valores perdidos definidos por el usuario para las variables de factor, sujeto e intra-sujetos se tratan como perdidos.
	Casos utilizados	Las estadísticas se basan en casos con datos válidos para todas las variables del modelo.
Manejo de ponderaciones		no aplicable

Sintaxis	GENLIN TenenciaCapital (REFERENCE=LAST) BY Sexo EstadoMigratorio Adulto_o_joven (ORDER=ASCENDING) WITH AccesoEducacionVirtual /MODEL Sexo EstadoMigratorio Adulto_o_joven AccesoEducacionVirtual INTERCEPT=YES DISTRIBUTION=BINOMIAL LINK=LOGIT /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100 MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 CITYPE=WALD LIKELIHOOD=FULL /MISSING CLASSMISSING=EXCLUDE /PRINT CPS DESCRIPTIVES

		MODELINFO FIT SUMMARY SOLUTION.
Recursos	Tiempo de procesador	00:00:00,08
	Tiempo transcurrido	00:00:00,26

Información de modelo

Variable dependiente	TenenciaCapital[a]
Distribución de probabilidad	Binomial
Función de enlace	Logit

a. El procedimiento modela 0 como la respuesta, tratando a 1 como la categoría de referencia.

Resumen de procesamiento de casos

	N	Porcentaje
Incluido	235	100.0%
Excluido	0	0.0%
Total	235	100.0%

Información de variable categórica

			N	Porcentaje
Variable dependiente	TenenciaCapital	0	224	95.3%
		1	11	4.7%
		Total	235	100.0%
Factor	Sexo	0	132	56.2%
		1	103	43.8%
		Total	235	100.0%
	EstadoMigrato	0	21	8.9%

	rio	1	214	91.1%
		Total	235	100.0%
	Adulto_o_joven	0	30	12.8%
		1	205	87.2%
		Total	235	100.0%

Información de variable continua

		N	Mínimo	Máximo	Media	Desv. Desviación
Covariable	AccesoEducacionVirtual	235	0	1	.31	.465

Bondad de ajuste[a]

	Valor	gl	Valor/gl
Desvianza	.877	8	.110
Desvianza escalada	.877	8	
Chi-cuadrado de Pearson	.568	8	.071
Chi-cuadrado de Pearson escalado	.568	8	
Logaritmo de verosimilitud[b]	-4.728		
Criterio de información Akaike (AIC)	19.455		
AIC corregido para muestras finitas (AICC)	19.717		
Criterio de información bayesiana (BIC)	36.753		

AIC coherente (CAIC)	41.753		

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual[a]
a. Los criterios de información están en un formato de cuanto más pequeño mejor.
b. La función de logaritmo de la verosimilitud completa se visualiza y utiliza en el cálculo de los criterios de información.

Prueba ómnibus[a]

Chi-cuadrado de razón de verosimilitud	gl	Sig.
12.035	4	.017

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual[a]
a. Compara el modelo ajustado con el modelo de sólo intersección.

Pruebas de efectos del modelo

Origen	Tipo III Chi-cuadrado de Wald	gl	Sig.
(Intersección)	.000	1	.998
Sexo	3.654	1	.056
EstadoMigratorio	.000	1	.999
Adulto_o_joven	.000	1	.999
AccesoEducacionVirtual	1.624	1	.203

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual

Estimaciones de parámetro

Parámetro	B	Desv. Error	95% de intervalo de confianza de Wald Inferior	Superior	Contraste de hipótesis Chi-cuadrado de Wald
(Intersección)	1.953	.3886	1.191	2.714	25.245
[Sexo=0]	1.334	.6978	-.034	2.701	3.654
[Sexo=1]	0[a]	.	.	.	.
[EstadoMigratorio=0]	19.714	16272.3273	-31873.461	31912.890	.000
[EstadoMigratorio=1]	0[a]	.	.	.	.
[Adulto_o_joven=0]	19.493	13449.1683	-26340.392	26379.378	.000
[Adulto_o_joven=1]	0[a]	.	.	.	.
AccesoEducacionVirtual	1.360	1.0677	-.732	3.453	1.624
(Escala)	1[b]				

Estimaciones de parámetro

Parámetro	Contraste de hipótesis gl	Sig.

(Intersección)	1	.000
[Sexo=0]	1	.056
[Sexo=1]	.	.
[EstadoMigratorio=0]	1	.999
[EstadoMigratorio=1]	.	.
[Adulto_o_joven=0]	1	.999
[Adulto_o_joven=1]	.	.
AccesoEducacionVirtual	1	.203
(Escala)		

Variable dependiente: TenenciaCapital
Modelo: (Intersección), Sexo, EstadoMigratorio, Adulto_o_joven, AccesoEducacionVirtual
a. Definido en cero porque este parámetro es redundante.

b. Fijado en el valor visualizado.

VI. VI.VII. III. Resultados Estadísticos en R: Capacidad Predictiva PC-Logit

VI. VI. VII. III. I. El Código de Programación en R

```
setwd("C:/Users/User/Desktop/Carpeta de Estudio/Maestría Profesional en Estadística/Semestre I-2021/Introducción a las Encuestas por Muestreo/Investigación Final")


m_julian <- read.csv("ENAVIRPA2021.csv",
                     stringsAsFactors = TRUE)
str(m_julian)
names(m_julian)
need <- c("TC1", "TC3", "E6",
"TE1","TE6","FN2","FN6","FN7","FN8","FN9","FN10","FN11","FN12","FN13","FN14","FN15","FN17","FN18","AE1","AE2A","AE2B","AE2C","AE2D","AE3","AE4","AE5","AE6","AE7","AE8","AE9","AE10","AE11","SD1","SD2","SD3","SD5","SD10","SD11","SD12","SD13")
sub_m_julian <- m_julian[, need]
attach(sub_m_julian)


names(sub_m_julian)
str(sub_m_julian)


enavirpa <- read.csv("m_julian.csv")
cod <- c("TC1", "TC3", "E6",
"TE1","TE6","FN2","FN6","FN7","FN8","FN9","FN10","FN11","FN12","FN13","FN14","FN15","FN17","FN18","AE1","AE2A","AE2B","AE2C","AE2D","AE3","AE4","AE5","AE6","AE7","AE8","AE9","AE10","AE11","SD1","SD2","SD3","SD5","SD10","SD11","SD12","SD13")
```

```
####### Creación de las nuevas variables



names(sub_m_julian)
## Varible VD1, se busca determinar si tiene o no internet##

#sub_m_julian$VD1 <- ifelse((TC1 == "si" | TC3 == "Pospago"), 1, 0)
#sub_m_julian$VD1 <- factor(sub_m_julian$VD1,
#                           levels = c(0,1),
#                           labels = c("No internet", "Si internet"))



#table(sub_m_julian$TC1,sub_m_julian$VD1 )

## Creación de la variable VD2, se busca determinar si es nacional o extranjero ##

table(sub_m_julian$SD3)

sub_m_julian$VD1 <-  ifelse( SD3 == "Sí", 1, 0)

sub_m_julian$VD1 <- factor(sub_m_julian$VD1,
                           levels = c(0,1),
                           labels = c("Extranjero", "Nacional"))
table(sub_m_julian$VD1)
str(sub_m_julian$VD1)
```

Creación de la variable VD3, se busca determinar si tiene o no medios de producción, esto se determinara con base en si posee un negocio propio, ya sea sin emplear o empleado a otros.

```
summary(sub_m_julian$TE1)


sub_m_julian$VD2 <-  ifelse(TE1 == "Tengo mi propio negocio sin emplear a otras personas" | TE1=="Tengo mi propio negocio y empleo a otras personas", 1, 0)

head(sub_m_julian$VD2)

sub_m_julian$VD2 <- factor(sub_m_julian$VD2,

                           levels = c(0,1),

                           labels = c("No MP","Si MP"))


table(sub_m_julian$VD2)

str(sub_m_julian$VD2)

table(sub_m_julian$VD2, sub_m_julian$VD1)
```

Creación de la variable VD4, la que busca determinar el empeoramiento de la dieta o no

```
names(sub_m_julian)

head(sub_m_julian[, c(7:16)])


table(sub_m_julian$FN6)

table(sub_m_julian$FN7)

table(sub_m_julian$FN8)

table(sub_m_julian$FN9)

table(sub_m_julian$FN10)
```

```
table(sub_m_julian$FN11)

table(sub_m_julian$FN12)

table(sub_m_julian$FN13)

table(sub_m_julian$FN14)

table(sub_m_julian$FN15)


sub_m_julian$VD3 <- ifelse(FN6 =="Disminuyó"|FN7 == "Disminuyó"  | FN8
=="Disminuyó"| FN9 == "Ha disminuido"  |FN10 == "Ha disminuido"| FN11 ==
"Ha disminuido"| FN12 == "Ha disminuido"| FN13 == "Ha disminuido"|FN14
=="Ha disminuido"|FN15 == "Ha disminuido",0,1)


sub_m_julian$VD3 <- factor(sub_m_julian$VD3,
                           levels = c(0,1),
                           labels = c("Empeoro", "Se mantuvo/Mejoro"))


table(sub_m_julian$VD3, sub_m_julian$VD2)




### creación de la nueva variable VD4, el sexo-


sub_m_julian$VD4 <- factor((ifelse(sub_m_julian$SD1=="Hombre", 0,1)),
                           levels = c(0,1),
                           labels = c("H", "M"))
sub_m_julian$VD4
```

```
## Creación de la variable VD5, busca determinar si esta o no por debajo de un umbral de pobreza

summary(SD5)


sub_m_julian$VD5 <-  ifelse(E6=="No cuenta con el tiempo suficiente" | E6 == "No cuenta con el dinero suficiente",0,1)

sub_m_julian$VD5 <- factor(sub_m_julian$VD5,

                    levels = c(0,1),

                    labels = c("No", "Si"))

#sub_m_julian$VD5 <- ifelse(SD5 =="Menos de 200 mil colones", 1,0)


#sub_m_julian$VD5 <- factor(sub_m_julian$VD5,

#                    levels = c(0,1),

#                    labels = c("No", "Si"))

table(sub_m_julian$VD5)

table(sub_m_julian$VD5, sub_m_julian$VD4)

table(sub_m_julian$VD4, sub_m_julian$VD5)


str(SD1)

class(SD1)

# la variable VD6 sera una reconversión de la variable SD1


View(sub_m_julian)

#sub_m_julian$VD6 <- factor((ifelse(sub_m_julian$SD1=="Hombre", 0,1)),

 #                   levels = c(0,1),

  #                  labels = c("H", "M"))

#sub_m_julian$VD6
```

```
# la variable VD7 menor de edad
#
sub_m_julian$SD2 <- as.numeric(sub_m_julian$SD2)
class(sub_m_julian$SD2)
sub_m_julian$VD7 <- ifelse(sub_m_julian$SD2<25, 0, 1)
head(sub_m_julian[, c("SD2", "VD7")])
sub_m_julian$VD7 <- factor(sub_m_julian$VD7,
                           levels = c(0,1),
                           labels = c("menor de edad", "mayor de edad"))




summary(TE6)


(1-(197/235))*100


#la variable no es representativa puesto que solo el 16.2% de las observaciones tiene respuesta válida para la misma.


# Variable VD8, Acceso a la educación virtual
#summary(E6)


#sub_m_julian$VD8 <-  ifelse(E6=="No cuenta con el tiempo suficiente" | E6 == "No cuenta con el dinero suficiente",0,1)
```

```
#sub_m_julian$VD8 <- factor(sub_m_julian$VD8,
 #                   levels = c(0,1),
  #                  labels = c("No", "Si"))
27+11
#table(sub_m_julian$VD8, sub_m_julian$VD5)


#summary(sub_m_julian$VD1)


#nueva variable VD8


attach(sub_m_julian)
head(sub_m_julian[, c("AE3", "AE4", "AE5",
"AE6","AE7","AE8","AE9","AE10","AE11")])


sub_m_julian$VD8 <- ifelse(AE3=="Disminuyó" | AE4 == "Disminuyó" | AE5 ==
"Disminuyó" | AE6 =="Disminuyó" | AE7 == "Disminuyó" |AE8 == "Disminuyó" |
AE9 == "Disminuyó"| AE10 == "Ingreso disminuido" | AE11 == "No les alcanza,
tiene dificultades" | AE11 == "No les alcanza, tienen grandes dificultades", 0, 1)


sub_m_julian$VD8 <- factor(sub_m_julian$VD8,
                    levels = c(0,1),
                    labels = c("contracción", "no contraccion"))


summary(sub_m_julian$VD1) # estado migratorio
summary(sub_m_julian$VD2) # tenencia de Medios de producción
summary(sub_m_julian$VD3) # Empeoramiento de la dieta
summary(sub_m_julian$VD4) # Sexo
```

```
summary(sub_m_julian$VD5) # Acceso o no a la educación virtual
summary(sub_m_julian$VD7) # Adultos
summary(sub_m_julian$VD8) # Contracción de ingreso
write.csv(sub_m_julian, "enavirpa_modificado.csv")
str(sub_m_julian$VD2)
summary(sub_m_julian$VD2)



formula <- sub_m_julian$VD2 ~
sub_m_julian$VD4+sub_m_julian$VD7+sub_m_julian$VD1


?glm


ft_1 <- glm(formula = formula, data = sub_m_julian, family = "binomial")


Pred_1 <- predict(ft_1, type = "response")
summary(Pred_1)
plot(Pred_1, sub_m_julian$VD2)
plot(sub_m_julian$VD2)
str(sub_m_julian$VD2)
head(sub_m_julian$VD2)
Pred_1_1 <- ifelse(Pred_1>0.25, 1,0)
class(Pred_1_1)
sub_m_julian$pred_1_1 <- factor(Pred_1_1,
                                levels = c(0,1),
                                labels = c("No MediosP", "Si MediosP"))
```

```
head(sub_m_julian[,c("VD2","pred_1_1")])

table(sub_m_julian$VD2)

table(sub_m_julian$pred_1_1)


observado_P <- sum(sub_m_julian$VD2 =="Si MP")

observado_N <- sum(sub_m_julian$VD2 =="No MP")

predict_P <- sum(sub_m_julian$pred_1_1 == "Si MediosP" )

predict_N <- sum(sub_m_julian$pred_1_1 == "No MediosP" )

total <- nrow(sub_m_julian)


data.frame(observado_P, observado_N, predict_P, predict_N)


VP <- sum(sub_m_julian$VD2 =="Si MP" & sub_m_julian$pred_1_1 == "Si
MediosP")

VN <- sum(sub_m_julian$VD2 =="No MP" & sub_m_julian$pred_1_1 == "No
MediosP")

FP <- sum(sub_m_julian$VD2 =="No MP" & sub_m_julian$pred_1_1 == "Si
MediosP")

FN <- sum(sub_m_julian$VD2 =="Si MP" & sub_m_julian$pred_1_1 == "No
MediosP")


VP

VN

FP

FN

matrix(c(VP, FN,FP, VN ), ncol=2 )


exactitud <- (VP+VN)/total
```

tasa_error <- (FP + FN)/total

sensibilidad <- VP/observado_P

especificidad <- VN/observado_N

precision <- VP/predict_P

pred_neg <- VN/predict_N

data.frame(exactitud, tasa_error, sensibilidad, especificidad, precision, pred_neg)

VI. VI. VII. III. I. Resumen de los resultados Estadísticos en R

```
> data.frame(observado_P, observado_N, predict_P, predict_N)
  observado_P observado_N predict_P predict_N
1          44         191        51       184
>
> VP <- sum(sub_m_julian$VD2 =="Si MP" & sub_m_julian$pred_1_1 == "Si MediosP")
> VN <- sum(sub_m_julian$VD2 =="No MP" & sub_m_julian$pred_1_1 == "No MediosP")
> FP <- sum(sub_m_julian$VD2 =="No MP" & sub_m_julian$pred_1_1 == "Si MediosP")
> FN <- sum(sub_m_julian$VD2 =="Si MP" & sub_m_julian$pred_1_1 == "No MediosP")
>
> VP
[1] 18
> VN
[1] 158
> FP
[1] 33
> FN
[1] 26
> matrix(c(VP, FN,FP, VN ), ncol=2 )
     [,1] [,2]
[1,]   18   33
[2,]   26  158
>
> exactitud <- (VP+VN)/total
> tasa_error <- (FP + FN)/total
> sensibilidad <- VP/observado_P
> especificidad <- VN/observado_N
> precision <- VP/predict_P
> pred_neg <- VN/predict_N
>
> data.frame(exactitud, tasa_error, sensibilidad, especificidad, precision, pred_ne
g)
  exactitud tasa_error sensibilidad especificidad precision  pred_neg
1 0.7489362  0.2510638    0.4090909     0.8272251 0.3529412 0.8586957
```

IV.VI. LIMITACIONES DEL DISEÑO METODOLÓGICO GENERAL EN RELACIÓN AL PROCESO DE ENCUESTAS POR MUESTREO REALIZADO

IV.VI. I. LIMITACIONES COLECTIVAS EN LA CANTIDAD DE ENCUESTAS COMPLETADAS

Como se adelantó antes, por motivos relativos a las condiciones personales de cada uno de los encuestadores, de las 300 encuestas planificadas a completar,

únicamente se lograron completar 235, lo que equivale a aproximadamente al 78% de las encuestas que se estipuló completar inicialmente.

IV.VI. I. LIMITACIONES DE CARÁCTER TEMPORAL RELATIVAS A ESTA INVESTIGACIÓN

Como el lector se habrá dado cuenta, el diseño inferencial inicial contemplaba construir una variable dicotómica con la información sociodemográfica (provincia, cantón y distrito), específicamente una variable dicotómica que estableciese si una persona pertenecía a la zona rural (1) o a la zona urbana (0). Por restricciones temporales fue imposible realizar la significativa cantidad de trabajo adicional que representaba su construcción empírica; sin embargo, se considera aquí que esta es una propuesta que puede servir para estudios posteriores, razón por la cual no se excluyó en la presentación de resultados.

Lo anterior se afirma en cuanto la determinación de la urbanidad o ruralidad de una región del país es posible de realizar únicamente conociendo el criterio de clasificación que establece el INEC a priori para realizar el censo (el último realizado hace alrededor de una década y este año se planea realizar el siguiente), tal y como se verifica en (Instituto Nacional de Estadística y Censos de Costa Rica, 2016, pág. 13): "La asignación del área urbana y rural se hace a partir del Censo Nacional de Población y Vivienda. En el último censo del año 2011 la calificación de urbano y rural se hizo para cada Unidad Geoestadística Mínima (UGM), según sus características."[24]

[24] A su vez, defínase UGM como "(…) "... el espacio geográfico de forma poligonal (Manzana o cuadra) y de superficie variable. Está constituido por un grupo de viviendas, edificios, predios, lotes o terrenos de uso habitacional, comercial, industrial, de servicios, entre otros. [...] está

Estos criterios se presentan en (Instituto Nacional de Estadística y Censos de Costa Rica, 2016, pág. 15), tal como se muestra a continuación:

TABLA 2

Codificación para la calificación de grado de urbanización del distrito

Codificación de los distritos	Código
Urbano	**1**
Predominantemente Urbano	2
Rural	**3**
Predominantemente Rural	4

Fuente: (Instituto Nacional de Estadística y Censos de Costa Rica, 2016, pág. 15).

En ese mismo documento, de la página 18 a la 30, se presentan todos distritos de Costa Rica junto con su código de grado de urbanización pertinente. Sin embargo, al presentarse la información en un documento PDF, forzosamente la digitación de los códigos correspondientes a los 488 municipios debe hacerse manualmente, junto con la digitación del nombre del distrito en cuestión. Una vez construida esa tabla, sólo se debía configurar alguna sintaxis en R o en Excel que permita clasificar binariamente los distritos con "1" si su código de urbanización es 3 o 4, mientras que "0" si su código de urbanización es 1 o 2. Esto, por los motivos expuestos no pudo llevarse a cabo en esta investigación.

El mismo efecto suscitaron las restricciones temporales en cuanto al interés por estudiar cualitativamente el desmejoramiento en la calidad de la alimentación e hidratación de las personas (en términos de los cambios negativos -o no- en relación al estándar que nutricionalmente se considera una dieta saludable), sin embargo, también se presenta el mecanismo para construir tal variable dicotómica, lo cual también puede ser relevante para futuros estudios.

---

delimitada por calles, veredas, cercas, quebradas, áreas de cultivos y otros elementos." (INEC, 2011)" (…)" en (Instituto Nacional de Estadística y Censos de Costa Rica, 2016, pág. 13).

IV.VI. I. LIMITACIONES TÉCNICAS

Como se adelantó en el prólogo de esta investigación, este ejercicio académico no necesariamente posee carácter vinculante con la realidad costarricense. Para que lo tuviese cada etapa de la investigación debería haberse preparado acorde a un conjunto de objetivos comunes y bien definidos (en función de ello se logra un diseño óptimo de los instrumentos), lo que a su vez implica un marco teórico que unifique cada uno de los aspectos involucrados en la encuesta, expresados a manera de módulos. Por otro lado, en lo relativo al diseño muestral no se verificó que necesariamente el Muestreo Aleatorio Simple (MAS) fuese la técnica de muestreo óptima para los fines deseados. Tampoco se realizó un análisis psicométrico de los ítems para la validación del instrumento.

En congruencia con lo anterior, el lector podrá verificar en la sección pertinente que los resultados estadísticos obtenidos tras la utilización de la metodología inferencial, en particular, los modelos lineales generalizados, no son los deseados en un escenario de interés profesional, sin embargo, ello no es relevante en cuanto este componente de la investigación sólo representa un ejercicio empírico con fines académicos para mostrar qué tipos de abordaje inferencial se puede realizar con la información obtenida a través de instrumentos psicométricos, tales como la encuesta. Así, era conocido con antelación que la encuesta no fue diseñada tomando en consideración lo antes descrito, así como también era conocido que el marco teórico que envolvía a la encuesta era diferente gnoseológica y metodológica a lo que específicamente el módulo AE (Afectación Económica) quería capturar[25]. Por ello, los resultados obtenidos, con independencia de una casi seguramente necesaria revisión de la concepción de las variables dicotómicas (elaboradas bajo fuertes restricciones temporales y con fines de prueba), no deben extrañar ni preocupar.

[25] Incluso la misma construcción del módulo era diferente a lo que se deseaba capturar, con la finalidad que se adaptara a los demás módulos del cuestionario.

IV.VI. I. LIMITACIONES RELATIVAS A LA BRECHA TEÓRICA EXISTENTE ENTRE LA ESTADÍSTICA CLÁSICA Y LA TEORÍA DEL APRENDIZAJE ESTADÍSTICO

Como deriva de lo expuesto en la sección VI.VI.VII, existen ciertas brechas entre la teoría del aprendizaje estadístico y la teoría estadística clásica. A medida transcurra el tiempo y con ello la evolución de las ciencias, la diferencia entre tales brechas tenderá a converger a cero, siempre que se logren conciliar las diferencias gnoseológicas entre ambos campos. Realizado a nivel general, constituye en sí mismo una investigación fundacional el abordar las razones últimas (gnoseológicas) por las cuales, como se deriva de los resultados particulares aquí presentados, la idoneidad de los resultados estadísticos no determina la idoneidad de los resultados de aprendizaje.

Sin embargo, el hecho concreto de que los coeficientes estadísticos obtenidos en esta investigación con métodos de la Teoría Estadística Clásica sean radicalmente lejanos óptimo (tómese cualquiera que se tome), mientras que los obtenidos con métodos del Aprendizaje Automático estén al menos al nivel promedio de lo que en general se podría calificar (a juicio de aquellos que entrenan modelos) como un buen modelo, quizás no sea una tarea titánica, al menos si se aborda superficialmente.

Como se adelantó, las metodologías estadísticas, sin importar cuál sea, busca encontrar patrones matemáticos en los datos, los cuales en términos visuales son en última instancia patrones geométricos (y esto está ligado con los orígenes históricos de las matemáticas como geometría). Estos patrones geométricos pueden (o no) estar vinculado en alguna magnitud en términos lógicos con algún marco conceptual para comprender la realidad objetiva, un ejemplo de esto es el contraste existente entre correlación y causalidad. Sin embargo, también este patrón geométrico puede extenderse (o no) al conjunto de datos global (población) de los que proviene el conjunto de datos con los que cuenta el investigador (muestra). Este segundo escenario puede ser el caso presenciado en esta investigación en relación al contraste de la calidad entre los coeficientes obtenidos en cada una de las metodologías, sin embargo, al no existir un marco teórico general sobre ello, no es posible garantizar con certeza que esta sea la causa de tal divergencia.

V. CONCLUSIONES

El estrés económico, medido en términos del grado de dificultades económicas que una persona experimenta para llegar a fin de mes y cuya construcción se especificó con detalle en la sección IV.II, muestra que un 47% de las personas encuestadas está al límite, mientras que un 34% está por debajo del límite (del cual un 10% estaba en estado crítico).

El ingreso disminuyó para un 44% de las personas que respondieron a esa pregunta y sólo un 17% puede ahorrar. En este sentido, el apoyo de alguna índole de organizaciones sin fines de lucro (por ejemplo, las ONG) es prácticamente inexistente, puesto que para el 91% no es fuente de apoyo. Por su parte, un 24% de las personas ha recibido ayuda del gobierno, fundamentalmente con comida (30% de las personas del 24% anterior que sí especificaron la ayuda recibida) y recursos económicos (62% de las personas del 24% anterior que sí especificaron la ayuda recibida), en donde el recurso económico fue en la generalidad de casos el bono *Proteger* dado por el gobierno actual durante la pandemia[26].

En términos de la participación en la riqueza social, el 82% no osee medios de producción, puesto que para un 82% de las personas las propiedades de alquiler, inversiones y/o ahorros no representan una fuente de ingresos; mientras que las variaciones en las remuneraciones laborales fueron negativas para un 37% de los encuestados y únicamente fueron positivas para un 4%. Sólo para un 12% de las personas las remesas del extranjero son fuente de ingresos, en donde de ese 12% el 39% sufrió contracciones en su ingreso, *i.e.*, 11 personas de las 28 para quienes las remesas del extranjero sí son una fuente de ingresos.

Finalmente, en lo relativo al módulo de afectación económica, el 45% de las personas encuestadas trabaja para un patrón, únicamente una personas de las encuestadas se encuentra desempleada (lo cual contrasta con la realidad nacional) de las 38 personas que respondieron que la pandemia había afectado su jornada laboral (el resto aparecen como NA debido a esa pregunta se configuró como parte del módulo "Teletrabajo", por lo que en caso de que la persona no realizara teletrabajo -esta era la primera pregunta del módulo- automáticamente se omitía el resto del módulo), para 16 de ellas les tocaba trabajar más por el mismo nivel de remuneración.

[26] Esto afirmado con base en lo conversado por el investigador en las 30 encuestas que le tocó realizar y por comentarios realizados en sesiones grupales de trabajo por otros investigadores.

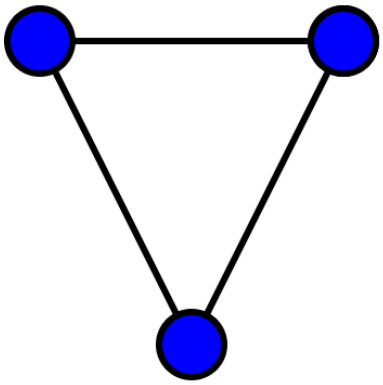
Complementariamente, para un 37% de las personas incrementó la hidratación diaria, mientras que un 57% duerme más, la actividad física incrementó para un 26% y disminuyó para un 24%, un 5% de los encuestados no recibe educación virtual por falta de recursos económicos y un 12% por falta de tiempo (de nuevo, al igual que antes existe una enorme cantidad de NA que se explican por las razones expuestas en el párrafo anterior), las fuentes de consumo de azúcar han disminuido para un 12% de las personas (y se han mantenido igual para un 46%) y las fuentes de grasas para un 10% (y se han mantenido igual para un 49%).

Tomando de referencia la línea de pobreza construida (presentada en la sección VI.II), se estimó una cantidad de personas en la pobreza equivalente al 17% de los encuestados. En general, con independencia de que la muestra no sea representativa de la población por los motivos expuestos en diversos pasajes de esta investigación, fue posible percibir a través de la comunicación misma con los sujetos entrevistados que los costos de la crisis sanitaria COVID-19 han sido asumidos en términos generales por las personas de menores ingresos, lo cual es congruente con todos los análisis y proyecciones realizados por instituciones oficiales costarricenses y también de carácter supranacional como por Comisión Económica para América Latina y el Caribe y el Fondo Monetario Internacional, no sólo para Costa Rica sino para América Latina y el mundo en general. Esto no fue percibido solamente por el autor de esta investigación sino también otras personas que trabajaron en investigaciones relativas a sus propios módulos (lo cual fue comentado en las sesiones de discusión grupal) en conjunto con el encargado del curso de *Introducción a las Encuestas por Muestreo*. En síntesis, existe una convergencia en términos generales y de carácter cualitativo entre las proyecciones económicas (que son las de principal interés dada la naturaleza del módulo aquí trabajado) realizadas a nivel nacional e internacional sobre Costa Rica y el mundo, en relación a los resultados obtenidos en ese mismo apartado en el levantamiento de encuestas realizado.

VI. ANEXOS

VI. I. DISTANCIAS TOPOLÓGICAS COMO DISTANCIAS RELATIVAS DESDE LOS ISOMORFISMOS DE GRAFOS Y LA TEORÍA DEL COMPORTAMIENTO COLECTIVO DE ANIMALES

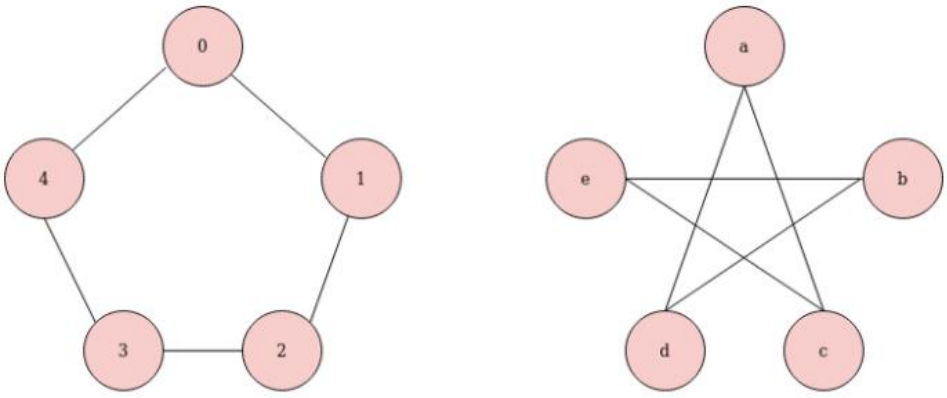
En teoría de grafos, se define como *grafo* al par $G = (V, E)$, en donde V es el conjunto de aquellos elementos que son vértices y E es el conjunto de pares de vértices cuyos elementos se denominan aristas. A continuación, se presenta un ejemplo simple de grafo con tres vértices (círculos azules) y tres aristas (líneas rectas negras), específicamente un triángulo rectángulo visto como grafo.



Fuente: (Wikimedia, 2021).

Un isomorfismo entre dos grafos G_1 y G_2 es una relación funcional biyectiva (*i.e.*, que establece una relación uno-a-uno entre los elementos de dos conjuntos) entre los vértices de G_1 y G_2 de la forma $f: V(G_1) \to V(G_2)$, en la que cualesquiera dos vértices $u, v \in G_1$ son adyacentes (relación entre dos vértices en la que ambos son extremos de la misma arista) si y solo si sus reflejos o imágenes matemáticas $f(u)$ y $f(v)$ son adyacentes en G_2. La característica fundamental de un isomorfismo de grafo es que es una relación funcional biyectiva que preserva las aristas que caracterizan al grafo. Que esta transformación matemática preserve las aristas implica que las distancias entre los vértices, analizados estos "de dos en dos", no cambian.

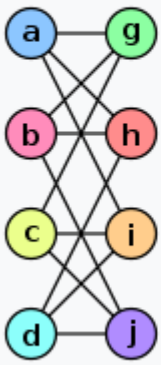
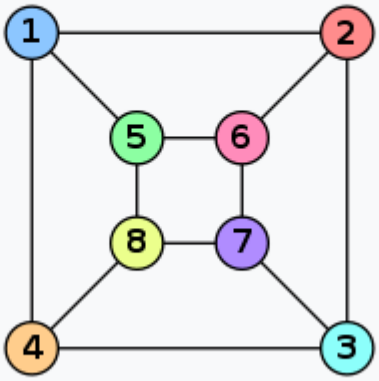
Son precisamente estas distancias a las que se les conoce como *distancias relativas* dentro de la estructura matemática, en contraste con las distancias absolutas que son medidas como distancias de los vértices considerados individualmente. Un ejemplo de ello se muestra a continuación.



G1 G2

Fuente: (Jose, 2020).

Los dos grafos anteriores son isomórficos entre sí, *i.e.*, poseen la misma estructura interna o estructura topológica. A continuación, se presenta un ejemplo numérico de ello, en consonancia con lo anteriormente expuesto.

GRAFO G_1	GRAFO G_2	Isomorfismo entre G_1 y G_2
		$f(a) = 1$ $f(b) = 6$ $f(c) = 8$ $f(d) = 3$ $f(g) = 5$ $f(h) = 2$ $f(i) = 4$ $f(j) = 7$

Fuente: (Wikipedia, 2021).

Como señala el Instituto de Sistemas Complejos de Italia (Instituto dei Sistemi Complessi, 2021), las diferencias concretas entre las distancias topológicas y las distancias métricas pueden observarse con nitidez en lo relativo al desarrollo teórico y aplicado de modelos que explican el comportamiento colectivo de animales, como lo son bandadas de aves, bancos de peces, etc. Esto es un equivalente concreto a nivel biológico del concepto matemático abstracto de la

manera en que se agrupan en subconjuntos los elementos de un determinado conjunto).

VI. II. CRITERIO DE NACIONES UNIDAS PARA MEDICIÓN DE LA POBREZA

En Alemania, por ejemplo, como se señala en (Köhler, 2016, pág. 10), se considera que una persona se encuentra en riesgo de caer en la pobreza si su ingreso no supera el 60% del ingreso medio de la sociedad. Este es un proceso ampliamente aceptado para estimar la línea de pobreza y, de hecho, como se verifica en (UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, 2017, pág. 13).

Según (Instituto Nacional de Estadística y Censos de Costa Rica, 2019), en la hoja C1 del documento Excel correspondiente a los cuadros de ingreso de la ENIGH 2018, el ingreso corriente bruto sin valor locativo medio (que compone el 83.3% del ingreso total) en Costa Rica es de ¢951,826.827, por lo que toda persona cuyo ingreso no supere los ¢571,096.2 sería considerada pobre bajo la metodología de medición de pobreza recomendaba por las Naciones Unidas y generalmente utilizada en Alemania y demás países industrializados. Este criterio se llamará en esta investigación, de ahora en adelante, *criterio UNECE*.

Línea de Pobreza UNECE-CR	**Aproximando:**
¢571,096.2	¢600,000

Tramos de Ingreso Mensual	*# de personas en tramo*
< ¢200,000	42
>= ¢200,000 y < 400,000	53
>= ¢400,000 y < 600,000	24
>= ¢600,000 y < 800,000	17
>= ¢800,000 y < 1,000,000	12
>= ¢1,000,000 y < 1,500,000	18
>= 1,500,000	14
NS/NR	55
TOTAL	235

TOTAL DE RESPUESTAS VÁLIDAS	180
TOTAL DE PERSONAS >= ¢600,000	61
(personas >= ¢600,000)/(total de respuestas válidas)	0,338888889

NOTA: Por supuesto, existe evidentemente una sobreestimación en aproximadamente ¢30,000; sin embargo, asúmase arbitrariamente que esta sobreestimación genera un error del 50% (en relación a la realidad de los encuestados, no la nacional) y se obtendría de igual forma un nivel de pobreza del 17%. Por supuesto, esto bajo estándares de países industrializados.

VII. REFERENCIAS

Aldrich, J. H., & Nelson, F. D. (1984). *Linear Probability, Logit, and Probit Models.* Beverly Hills: Sage University Papers Series. Quantitative Applications in the Social Sciences.

Allen, M. (2017). *The SAGE Encyclopedia of COMMUNICATION RESEARCH METHODS.* London: SAGE Publications, Inc.

AMERICAN PSYCHOLOGICAL ASSOCIATION. (2021, Julio 15). *level*. Retrieved from APA Dictionary of Pyschology: https://dictionary.apa.org/level

AMERICAN PYSCHOLOGICAL ASSOCIATION. (2021, Julio 15). *factor*. Retrieved from APA Dictionary of Pyschology: https://dictionary.apa.org/factor

AMERICAN PYSCHOLOGY ASSOCIATION. (2021, Julio 15). *logistic regression (LR)*. Retrieved from APA Dictionary of Pyschology: https://dictionary.apa.org/logistic-regression

Barrios, J. (2019, Julio 19). *La matriz de confusión y sus métricas* . Retrieved from Health BIG DATA: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Bhuptani, R. (2020, Julio 13). *Quora*. Retrieved from What is the difference between linear regression and least squares?: https://www.quora.com/What-is-the-difference-between-linear-regression-and-least-squares

Birnbaum, Z. W., & Sirken, M. G. (1950, Marzo). Bias Due to Non-Availability in Sampling Surveys. *Journal of the American Statistical Association, 45*(249), 98-111.

Burrus, C. S. (2021, Julio 7). *Iterative Reweighted Least Squares.* Retrieved from https://cnx.org/exports/92b90377-2b34-49e4-b26f-7fe572db78a1@12.pdf/iterative-reweighted-least-squares-12.pdf

Centro Centroamericano de Población. (2021, Abril 28). *Variables y escalas de medición*. Retrieved from Universidad de Costa Rica: https://ccp.ucr.ac.cr/cursos/epidistancia/contenido/2_escmed.html

Cochran, W. G. (1991). *Técnicas de Muestreo.* México, D.F.: Compañía Editorial Continental.

Departamento Administrativo Nacional de Estadística. (2003). *Metodología de Diseño Muestral.* Bogotá: Dirección Sistema Nacional de Información Estadística. Retrieved from https://www.dane.gov.co/files/EDI/anexos_generales/Metodologia_diseno_muestral_anexo1.pdf?phpMyAdmin=a9ticq8rv198vhk5e8cck52r11

Díaz-Narváez, V. P. (2017). Regresión logística y decisiones clínicas. *Nutrición Hospitalaria, 34*(6), 1505-1505. Retrieved from https://scielo.isciii.es/pdf/nh/v34n6/36_diaz.pdf

Google Developers. (2021, Julio 19). *Clasificación: Exactitud*. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/accuracy

Greene, W. (2012). *Econometric Analysis* (Séptima ed.). Harlow, Essex, England: Pearson Education Limited.

Gujarati, D., & Porter, D. (2010, Julio 8). *Econometría* (Quinta ed.). México, D.F.: McGrawHill Educación. Retrieved from Homocedasticidad.

Haskett, D. R. (2014, Octubre 10). *"Mitochondrial DNA and Human Evolution" (1987), by "Mitochondrial DNA and Human Evolution" (1987), by Rebecca Louise Cann, Mark Stoneking, and Allan Charles Wilson*. Retrieved from The Embryo Project Encyclopedia: https://embryo.asu.edu/pages/mitochondrial-dna-and-human-evolution-1987-rebecca-louise-cann-mark-stoneking-and-allan

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (Segunda ed.). New York: Springer.

Instituto dei Sistemi Complessi. (2021, Febrero 27). *Topolical vs Metric Distance*. Retrieved from Biological Systems: https://www.isc.cnr.it/research/topics/physical-biology/biological-systems/topological-vs-metric-distance/

Instituto Nacional de Estadística y Censos de Costa Rica. (2016, Julio). *Manual de Clasificación Geográfica con Fines Estadísticos de Costa Rica.* Retrieved from Biblioteca Virtual: https://www.inec.cr/sites/default/files/documetos-biblioteca-virtual/meinstitucionalmcgfecr.pdf

Instituto Nacional de Estadística y Censos de Costa Rica. (2019). *ENIGH. 2018. Cuadros sobre ingresos de los hogares.* San José: INEC. Retrieved from https://www.inec.cr/sites/default/files/documetos-biblioteca-virtual/reenigh2018-ingreso.xlsx

Instituto Nacional de Estadística y Censos de Costa Rica. (2021, 7 14). *Factor de Expansión*. Retrieved from INEC: https://www.inec.cr/sites/default/files/_book/F.html

Instituto Nacional de Estadística y Censos de la República Argentina. (2019). *Encuesta de Actividades de Niños, Niñas y Adolescentes 2016-2017. Factores de expansión, estimación y cálculo de los errores por muestra para el dominio rural.*

Buenos Aires: Ministerio de Hacienda. Retrieved from https://www.indec.gob.ar/ftp/cuadros/menusuperior/eanna/anexo_bases_eanna_rural.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Jose, K. (2020, Junio 27). *Graph Theory | Isomorphic Trees*. Retrieved from Towards Data Science: https://towardsdatascience.com/graph-theory-isomorphic-trees-7d48aa577e46

Köhler, T. (2016). Income and Wealth Poverty in Germany. *SOEP papers on Multidisciplinary Panel Data Research*, 1-48. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.540534.de/diw_sp0857.pdf

Kolmogórov, A. N., & Fomin, S. V. (1978). *Elementos de la Teoría de Funciones y del Análisis Funcional* (Tercera ed.). (q. e.-m. Traducido del ruso por Carlos Vega, Trans.) Moscú: MIR.

Liao, T. F. (1994). *INTERPRETING PROBABILITY MODELS. Logit, Probit, and Other Generalized Linear Models.* Iowa: Sage University Papers Series. Quantitative Applications in the Social Sciences.

Lipschutz, S. (1992). *Álgebra Lineal.* Madrid: McGraw-Hill.

Lohr, S. L. (2019). *Sampling: Design and Analysis* (Segunda ed.). Boca Raton: CRC Press.

Lohr, S. L. (2019). *Sampling: Design and Analysis* (Segunda ed.). Boca Raton: CRC Press.

McCullagah, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Segunda ed.). London: Chapman and Hall.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Segunda ed.). London: Chapman and Hall.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*(3), 370-384.

Online Stat Book. (2021, Julio 15). *Levels of an Independent Variable*. Retrieved from Independent and dependent variables: https://onlinestatbook.com/2/introduction/variables.html

Patil, G. P., & Shorrock, R. (1965). On Certain Properties of the Exponential-type Families. *Journal of the Royal Statistical, 27*(1), 94-99.

Perry, J. (2014, Abril 2). *NORM TO/FROM METRIC*. Retrieved from The University of Southern Mississippi: https://www.math.usm.edu/perry/old_classes/mat681sp14/norm_and_metric.pdf

Ritchey, F. (2002). *ESTADÍSTICA PARA LAS CIENCIAS SOCIALES. El potencial de la imaginación estadística.* México, D.F.: McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.

Samuels, S. (2014, 11 19). *Can I get to an approximation of the population with knowledge of the expansion factor?* Retrieved from Cross Validated. StackExchange: https://stats.stackexchange.com/questions/124750/can-i-get-to-an-approximation-of-the-population-with-knowledge-of-the-expansion

StackExchange Cross Validated. (2017, Febrero 2). *"Least Squares" and "Linear Regression", are they synonyms?* Retrieved from What is the difference between least squares and linear regression? Is it the same thing?: https://stats.stackexchange.com/questions/259525/least-squares-and-linear-regression-are-they-synonyms

StackExchange Data Science. (2016, Junio 19). *Is GLM a statistical or machine learning model?* Retrieved from https://datascience.stackexchange.com/questions/488/is-glm-a-statistical-or-machine-learning-model

StackOverFlow. (2014, Marzo 15). *Supervised Learning, Unsupervised Learning, Regression*. Retrieved from https://stackoverflow.com/questions/22419136/supervised-learning-unsupervised-learning-regression

TalkStats. (2011, Noviembre 29). *SPSS*. Retrieved from Forums: http://www.talkstats.com/threads/what-is-the-difference-between-a-factor-and-a-covariate-for-multinomial-logistic-reg.21864/

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE. (2017). *Guide on Poverty Measure.* New York and Geneva: UNITED NATIONS. Retrieved from https://ec.europa.eu/eurostat/ramon/statmanuals/files/UNECE_Guide_on_Poverty_Measurement.pdf

van den Berg, R. G. (2021, Julio 15). *Measurement Levels – What and Why?* Retrieved from SPSS Tutorials: https://www.spss-tutorials.com/measurement-levels/

Weisstein, E. W. (2021, Julio 15). *Sigmoid Function*. Retrieved from MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/SigmoidFunction.html

Weisstein, E. W. (2021, Mayo 21). *Sigmoid Function*. Retrieved from MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/SigmoidFunction.html

Weisstein, E. W. (2021, Mayo 18). *Smooth Function*. Retrieved from Wolfram MathWorld - A Wolfram Web Resource: https://mathworld.wolfram.com/SmoothFunction.html

Wikimedia. (2021, Abril 6). *Commons*. Retrieved from Wikipedia: https://upload.wikimedia.org/wikipedia/commons/b/bf/Undirected.svg

Wikipedia. (2021, Julio 6). *Graph isomorphism*. Retrieved from Morphism: https://en.wikipedia.org/wiki/Graph_isomorphism

Wikipedia. (2021, Mayo 21). *Iterative proportional fitting*. Retrieved from Statistical algorithms: https://en.wikipedia.org/wiki/Iterative_proportional_fitting

Wikipedia. (2021, Febrero 25). *Iteratively reweighted least squares*. Retrieved from Least squares: https://en.wikipedia.org/wiki/Iteratively_reweighted_least_squares

Wikipedia. (2021, Julio 13). *Logistic function*. Retrieved from Growth curves: https://en.wikipedia.org/wiki/Logistic_function

Wikipedia. (2021, Mayo 22). *Logistic regression*. Retrieved from Regression models: https://en.wikipedia.org/wiki/Logistic_regression

Wikipedia. (2021, Junio 14). *Logit*. Retrieved from Special functions: https://en.wikipedia.org/wiki/Logistic_function

Wikipedia. (2021, Julio 8). *Lp space*. Retrieved from Measure theory: https://www.wikiwand.com/en/Lp_space

Wikipedia. (2021, Abril 15). *Odds*. Retrieved from Wagering: https://en.wikipedia.org/wiki/Odds

Wikipedia. (2021, Julio 10). *Precision and recall*. Retrieved from Bioinformatics: https://en.wikipedia.org/wiki/Precision_and_recall

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data* (Segunda ed.). Cambridge, Massachusetts: MIT Press.