Universidad Nacional Autónoma de México

Facultad de Ciencias

Influencia de las comorbilidades y factores socioeconómicos en el riesgo de fallecer por COVID-19 en México

T E S I S

que para obtener el título de:

Matemática

presenta:

Vanessa Itzel Soulé Flores

Asesor:

Dr. Carlos Erwin Rodríguez Hernández-Vela

Ciudad Universitaria, CD. MX., 2022

A mis compañeros de vida.
A.A.S y Pisha
∞

Agradecimientos

Agradezco de todo corazón a cada una de las personas que me apoyaron y motivaron a lo largo de esta etapa, gracias a su apoyo lo logré; fueron mi fuerza y motivación.

Agradezco principalmente a mi compañero de vida, Ángel Arenas Soní por ser mi pilar y ejemplo a seguir. Gracias por el amor y apoyo incondicional que me brindaste día con día; tú me inspiraste y diste fortaleza en los momentos más difíciles... llegaste a mi vida y la cambiaste por completo, te agradezco infinitamente por todo. Este es el primero de muchos logros que tengo la dicha de compartir y vivir a tu lado.

Hago también un agradecimiento especial a aquellas personas fundamentales en esta etapa:

Gracias, Carlos Erwin por guiarme y asesorarme en este proceso.

Gracias, Ramsés por el apoyo y la paciencia brindada.

Gracias, Pedro por todas tus enseñanzas, marcaron mi vida y mi camino.

Gracias a mis sinodales por sus consejos y comentarios.

Gracias a mi familia y amigos por ayudarme y creer en mí.

Por último agradezco sinceramente a CONACyT y al Proyecto UNAM-DGAPA-PAPIIT IG100221 por brindarme apoyo económico con el que pude concluir mis estudios de manera exitosa.

Resumen

La pandemia por COVID–19 ha provocado una necesidad imperante de realizar estudios multidisciplinarios que ayuden a entender la dinámica de la transmisión y lleven a minimizar los efectos negativos en la población.

En esta tesis se analizaron las condiciones sociales y económicas que existen dentro de la población mexicana para determinar el riesgo de fallecimiento por COVID-19. En primer lugar, el estudio se hizo a nivel municipio, por lo que se consideró el riesgo de fallecimiento por COVID-19 en cada uno de los 2,471 municipios en los que se divide la República Mexicana. En un segundo lugar, y debido a que la cantidad de población entre municipios es muy heterogénea, se definieron grupos de municipios con número de habitantes similares. Finalmente, los datos se analizaron por cada ola de la pandemia, pues México ha sufrido cuatro olas y tanto el conocimiento de la enfermedad como las condiciones han cambiado de ola en ola.

Para este trabajo se consultó y combinó información de distintas fuentes. Siendo la más importante la base de datos abierta que actualiza diariamente la Secretaría de Salud y que da cuenta de la evolución de la pandemia de COVID-19 en México (ver [6]). Otro recurso de información fue la base de datos del CENSO Nacional de Población y Vivienda de 2020 publicada por el INEGI (ver [12]), en donde se ofrece un panorama actualizado de la dimensión, estructura y distribución espacial de la población que reside en México, así como sus principales características socioeconómicas y culturales. Finalmente, también se usaron otras dos fuentes de información de INEGI que dan cuenta del nivel de marginación así como de la prevalencia de algunas comorbilidades en la población, en ambos casos a nivel municipal.

Para el análisis de la información se utilizó la regresión lineal múltiple. Con esta herramienta, se identifican los factores socio-económicos que tienen mayor impacto en el riesgo de fallecer por COVID-19. Asimismo, con el modelo obtenido se realizó una predicción para la cuarta ola de la pandemia y se encontró que las defunciones disminuyen a comparación de la tercera ola.

Los resultados muestran que existen factores socioeconómicos que influyen en el riesgo de mortalidad por COVID-19 a nivel municipal, por ejemplo: la educación, los bajos ingresos, la falta de servicios básicos en la vivienda y la falta de acceso a servicios de salud. Además, los resultados cambian dependiendo del tamaño de los municipios en cuanto a densidad poblacional, como es el caso del porcentaje de la población que reside en viviendas sin agua entubada: en poblaciones pequeñas el riesgo de mortalidad disminuye a comparación de las poblaciones grandes donde aumenta; o en el caso del analfabetismo el riesgo aumenta en municipios pequeños. Sin embargo, también se observa que los factores socioeconómicos analizados no tienen un impacto determinante en las defunciones, como sí lo tienen factores relacionados con la salud del paciente, por ejemplo comorbilidades, síntomas graves causados por la enfermedad, etc.

Conocer el impacto de los factores socioeconómicos de una sociedad es esencial para su progreso y bienestar; conocer las características demográficas, sociales y económicas de una población ayuda a determinar la capacidad de resiliencia que tiene una sociedad ante cualquier adversidad, como la que trajo consigo la pandemia por el virus SARS-CoV-2.

Índice general

Introducción

« Nous prouvons par logique, mais nous découvrons par intuition. »
« Probamos por medio de la lógica, pero descubrimos por medio de la intuición. »
- Henri Poincaré

Desde nuestros orígenes, la humanidad ha tratado de entender la complejidad del universo mediante herramientas matemáticas como son los modelos; los cuales nos ayudan a obtener un enfoque para desarrollar aproximaciones sobre el comportamiento de un sistema en particular, además, facilitan el entendimiento del problema analizado y nos permiten encontrar una posible solución.

La vida es impredecible y está en constante cambio, existe un riesgo latente de sufrir algún evento inesperado como ha sucedido a lo largo de la historia, como actualmente con el brote epidemiológico causado por el virus de la familia de coronavirus del Síndrome Respiratorio Agudo Grave (SARS-CoV-2) que desencadena la enfermedad por coronavirus de 2019 (COVID-19). En diciembre de 2019 se detectó el primer caso de COVID-19 en la provincia de Wuhan, China, pasó poco menos de un mes para que se confinara y aislara a toda la población de esa provincia y a principios de enero de 2020 ya había pasado de ser un brote epidémico a una epidemia por su alto grado de propagación. Para marzo del 2020 la Organización Mundial de la Salud (OMS) declaró que la epidemia ya se había convertido en pandemia por encontrarse en más de 100 territorios a nivel mundial.

La enfermedad COVID-19 llegó a México el 28 de febrero de 2020 y para el 23 de marzo de ese año la Secretaría de Salud del Gobierno de México determinó que se implementarían medidas de emergencia sanitarias como el distanciamiento social, la suspensión de actividades no esenciales y medidas básicas de higiene para controlar o mitigar la transmisión y

el contagio del virus SARS–CoV–2. México se encuentra entre los 20 países con más casos confirmados y muertes por COVID-19 en todo el mundo (datos al corte del mes de marzo del 2022). De acuerdo a la OMS las personas con problemas médicos como la obesidad, diabetes, hipertensión y enfermedades pulmonares, entre otros son las más vulnerables a padecer severas complicaciones con el nuevo virus SARS–CoV–2, provocando en el peor de los casos, la muerte. En México más del 70 % de la población padece sobrepeso lo que desencadena diversos problemas de salud como la diabetes e hipertensión ocasionando que nuestro país resulte severamente afectado por esta pandemia. [19]

Se ha comprobado que la salud y el crecimiento económico de un país están intrínsecamente relacionados. En países más desarrollados la esperanza de vida es mayor y tienen mejores condiciones de vida en general, cosa que no sucede en países menos desarrollados; así que dependiendo del nivel socioeconómico que se tenga se puede tener acceso a servicios de calidad en salud, educación, empleo, vivienda, etc. Mediante estudios realizados en México, se ha demostrado la mayor concentración de patologías complejas y de exceso de mortalidad en los lugares o grupos donde prevalecen elevados niveles de marginación y exclusión social [9].

A lo largo de toda la República Mexicana existen grandes contrastes entre zonas geográficas, grupos étnicos y niveles socioeconómicos que junto con el alto índice de contagios y muertes por las condiciones de salud que imperan y desencadenan la transmisión del virus en personas con comorbilidades, resulta necesario estudiar la pandemia desde un punto de vista social y económico además del lado médico.

El objetivo de esta tesis es estudiar la influencia que tienen diversos factores socioeconómicos en el riesgo de mortalidad por COVID–19 en México. Con este fin, primero se reúne y homologa información de diversas fuentes. Después, se da una primera visión mediante el análisis exploratorio de los datos recabados. Finalmente, mediante la técnica estadística del análisis de regresión lineal múltiple se obtiene un modelo con el que se busca describir el riesgo de mortalidad a nivel municipal en las primeras 3 olas de la pandemia en el país.

Capítulo 1

Pandemia por COVID-19 en México

Los datos obtenidos de acuerdo al sitio web oficial[1] del Gobierno de México para la pandemia de COVID-19 desde que se confirmó el primer caso el 28 de febrero del 2020 muestran que los contagios aumentaron drásticamente conforme avanzaba la pandemia. Según el grado de transmisión de la enfermedad se fueron estableciendo las fases epidemiológicas identificadas por las autoridades sanitarias siendo un total de 3 fases, el 24 de marzo se decretó la fase 2 que comprende primordialmente la suspensión de ciertas actividades económicas, la restricción de congregaciones masivas y la recomendación de resguardo domiciliario a la población en general [4].

Para ayudar a mitigar y controlar la pandemia en el país el Gobierno de México implementó el Plan DN-III-E que es un operativo militar de la Secretaría de la Defensa Nacional de México para que el Ejército y Fuerza Aérea Mexicana del país realicen actividades de auxilio a la sociedad afectada por cualquier tipo de desastre. Se llevó acabo la *Vigilancia Centinela* que acepta la realidad de que no se están documentando todos los casos y hace una adaptación para expandir el número con estimaciones que están basadas en la dinámica de la ocurrencia de la enfermedad en términos territoriales, la demanda de atención médica y las características de las personas. [2]

[1] https://www.gob.mx/salud/documentos/datos-abiertos-152127

[2] Conferencia de prensa. Informe diario sobre coronavirus COVID-19 en México (08 de abril de 2020).

Luego de la primera ola de contagios, donde hubo hasta 8,458 casos diarios en julio del 2020, durante 4 meses seguidos se registró una disminución de casos reportados hasta noviembre cuando comenzaron a incrementarse de nuevo los contagios. Después de iniciar el programa de vacunación el 24 de diciembre del 2020, se registró la segunda ola de contagios llegando a su máximo el 21 de enero del 2021 con 22,339 casos registrados ese día. El semáforo epidemiológico cambió de rojo a naranja para poder reactivar la economía tomando las medidas pertinentes, que no fueron suficientes para evitar que se tuviera una tercera ola de contagios que inició el 18 de junio, llegando a su pico el 18 de agosto de 2021 con 28,953 casos registrados en un solo día.

Para agosto del 2021 ya se contaba con una base de datos total de 9,230,161 casos registrados en México, de los cuales 3,200,000 fueron casos positivos con el virus SARS–CoV–2 y el resto de los casos registrados fueron negativos, sospechosos o no confirmados. Hay que tener en cuenta que esta pandemia es un fenómeno inconmensurable, no puede medirse de manera directa y exacta; además cabe señalar que las autoridades de salud pública han admitido que el número de contagios y muertes por COVID–19 es mayor a lo reportado, en los inicios de la pandemia se estimó que era entre 8 y 10 veces mayor; más adelante se estableció un método de estimación considerando el índice de positividad para los sospechosos [6] ya que los registrados eran solo los que se hicieron prueba pero sabemos que hay muchos otros casos donde empezaron con síntomas pero nunca se hicieron una prueba para registrarlos como positivos. Al mes de marzo del 2022 los casos registrados se elevan a 15 millones, de los cuales poco más de 5 millones y medio son casos positivos donde el promedio de edad es de 39 años, afectando por igual a ambos sexos, sin embargo las defunciones confirmadas muestran un predomino del 62 % en hombres con una letalidad alta de $\sim 4.7\,\%$ en México en comparación al 1.5 % de letalidad global. [3]

Por lo tanto, teniendo una alta tasa de contagios y letalidad además de la situación alarmante del país en cuestión de obesidad, existe una imperante necesidad de proyectar escenarios factibles del riesgo de mortalidad por COVID–19 al que está expuesta la población de México. Esta pandemia ha impuesto a cada espacio geográfico necesidades

[3] https://www.gob.mx/salud/documentos/comunicados-tecnicos-diarios-covid19

específicas a la hora de enfrentarla y de prepararse para futuros escenarios.

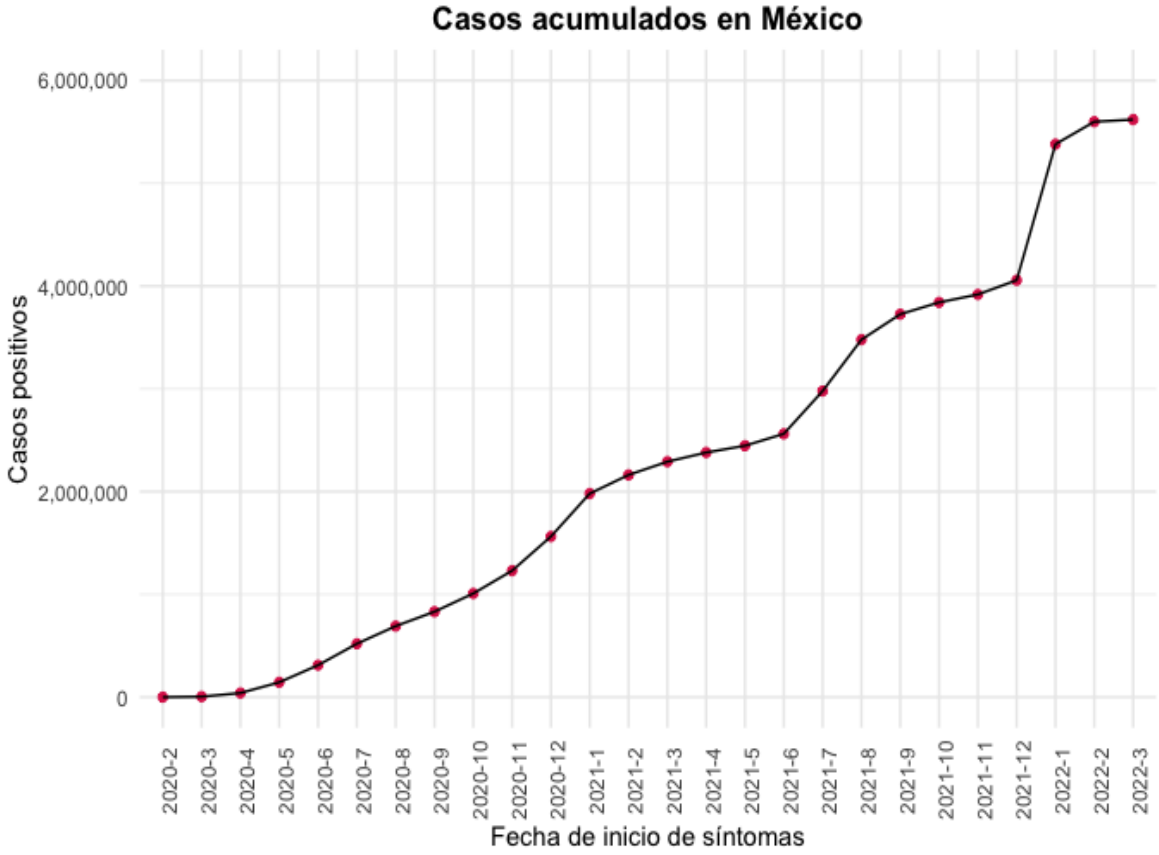


Figura 1.1: Evolución de la pandemia tomando los casos positivos acumulados en México desde el inicio de la pandemia en febrero del 2020 hasta marzo del 2022.

México se encuentra entre los primeros países con una elevada tasa de incidencia, y se refleja observando cómo los contagios se incrementaron drásticamente, por ejemplo el 5 de abril del 2021 había 2,251,705 casos confirmados y en menos de 6 meses la cifra aumentó en más de 1 millón de contagios, el 6 de septiembre del 2021 se llegó a un total de 3,433,511 casos positivos y en cuanto a las cifras sobre los fallecimientos de 204,399 que hubo en abril del 2021 pasaron a 263,470 para septiembre; convirtiéndonos así en el cuarto país con más muertes en todo el mundo (datos al mes de septiembre del 2021) [6].

Posición	País	Casos (Millones)	Muertes (Miles)
1	Estados Unidos	79.64	978
2	India	43.03	521
3	Brasil	30.15	661
4	Francia	26.2	140
5	Alemania	22.84	132
6	Reino Unido	21.64	170
7	Rusia	18.01	372
8	Corea del Sur	15.65	19
9	Italia	15.32	160
10	Turquía	14.96	98
11	España	11.62	103
12	Vietnam	10.25	42
13	Argentina	9.05	128
14	Países Bajos	7.99	22
15	Irán	7.19	140
16	Japón	7.07	28
17	Colombia	6.08	139
18	Indonesia	6.03	155
19	Polonia	5.98	115
20	México	5.72	323

Tabla 1.1: Registros de los casos positivos y defunciones en los 20 países con más casos en todo el mundo. Fuente: WHO Coronavirus (COVID-19) Dashboard. – Registros al mes de abril del 2022. – ver https://covid19.who.int/table

A medida que va avanzando el tiempo, gracias a las evidencias médicas y los avances científicos que han sido implementados para conocer y tratar el virus, hemos aprendido mucho acerca de este, tanto de su tasa de contagio y transmisión, su tasa de mortalidad, síntomas, cual es la gravedad de la enfermedad y como puede ser tratada o controlada. De acuerdo a los datos recabados por el gobierno de México, en el país las principales comorbilidades que estuvieron relacionadas con los casos confirmados fueron la hipertensión con 17.21 %, obesidad con 14.26 %, diabetes con 13.25 % y tabaquismo con 7.33 %, y de igual forma se ha demostrado que el riesgo de enfermar gravemente tras contagiarse del virus SARS–CoV–2 depende de la edad y de las afecciones de salud que tengan las personas. Entonces, conocer a las personas con mayor riesgo de adquirir el virus o presentar un cuadro grave de COVID–19 es crucial para determinar las estrategias políticas y sociales en

términos de salubridad además de prevenir al personal médico para que pueda tener un mejor manejo y control de los pacientes cuando ingresen por alguna complicación.

1.1. Comorbilidades y síntomas

Desde el comienzo de la pandemia se identificó a las comorbilidades preexistentes como uno de los factores que aumentaba el riesgo de fallecimiento por COVID-19. De acuerdo con los hallazgos en torno a las comorbilidades, la enfermedad pulmonar obstructiva crónica (EPOC) tuvo el mayor índice de riesgo con 14.38 %; seguida de la Enfermedad Renal Crónica (ERC) con 10.26 %; la diabetes con 10.12 %; y la hipertensión con 8.954 % [3] las cuales desgraciadamente coinciden con las principales comorbilidades que padece la población mexicana en gran escala.

- **Principales afecciones en los mexicanos: diabetes, obesidad e hipertensión**

  Desde hace 30 años, México se convirtió en uno de los países más afectados en todo el mundo por la epidemia de obesidad, ya que del 70 % de la población que padece sobrepeso, la tercera parte sufre obesidad mórbida [17], provocando que la obesidad sea considerada actualmente como un grave problema de salud en nuestro país. México es ahora el segundo país en todo el mundo con prevalencia de obesidad. En 2006, 2012 y 2018, tanto el sobrepeso como la prevalencia de obesidad se incrementaron de 69.5 % a 71.3 % y luego a 75.2 % en la población de 20 años en adelante, mientras que la tasa de obesidad se elevó de 30 % en 2006 a 32.4 % en 2012 y luego a 36.1 % en 2018. También, México es ahora uno de los países con mayor tasa de obesidad infantil en el mundo con uno de cada tres niños con sobrepeso u obesidad.

  La diabetes, la enfermedad crónica relacionada directamente con la obesidad, se ha propagado rápidamente y en México en 2018 afectó a 10.3 % de la población adulta (arriba de los 20 años de edad), mientras que en 2012 había afectado al 9.2 %.

  Otra enfermedad importante es la presión alta en la sangre o hipertensión que tiene síntomas menos notorios. Sin embargo, si no se trata, se incrementa el riesgo de

tener problemas serios como un ataque al corazón o paros cardiacos. En México, la prevalencia de la hipertensión en 2012 fue de 30.2 % y en 2018 de 32.7 %.

Conocer las afecciones subyacentes de la sociedad permite a los países estimar el porcentaje de la población que pueda desarrollar un cuadro grave debido a las comorbilidades que sufren, facilitando la clasificación de una población en grupos de personas con o sin comorbilidades, por grupos etarios y sexo.

1.2. Factores socioeconómicos

A lo largo de la historia se ha visto que los factores socioeconómicos son determinantes para la supervivencia, mejorar las condiciones de vida y salud además de disminuir el riesgo de mortalidad o contagio de alguna enfermedad como es el caso de la pandemia por COVID–19. Algunos efectos socioeconómicos que tuvo la pandemia en sus inicios en México incluyen la generación de compras de pánico y saqueos de establecimientos, que generan desabasto de productos de limpieza e higiene personal, por otro lado, atravesar la pandemia se convirtió en un reto para la población mexicana con la suspensión de eventos socioculturales o el cierre temporal de comercios, anteponiendo su capacidad para enfrentar la crisis económica que ocasionó esta pandemia.

Los sectores de la población con mayor pobreza y donde se evidencian más las desigualdades son los más afectados, la mayoría de ellos viven del comercio informal; de acuerdo a datos de la Encuesta Nacional de Ocupación y Empleo (ENOE) del 2019 en México el 56.2 % de la población ocupada vive del comercio informal [11], siendo este sector el más afectado con el cierre de los establecimientos no esenciales y el resguardo domiciliario. Además, gracias a la Encuesta Telefónica de Ocupación y Empleo (ETOE) del 2020 se detectó que se perdieron al rededor de 1.1 millones de empleos en los primeros meses de la pandemia, de los cuales el 83.7 % (933 mil) corresponden a trabajadores que percibían entre uno y dos salarios mínimos en el periodo de marzo a junio de 2020 [12].

Para mencionar algunos de los efectos económicos que ha tenido México debido a la pandemia por COVID-19 veamos algunas estadísticas. De acuerdo a Gerardo Esquivel,

miembro de la junta de Gobierno del Banco de México "...el Indicador Global de Actividad Económica (IGAE) de abril de 2020 disminuyó en 17.3 % con respecto a marzo del mismo año, la contracción más grande de toda su historia para un solo mes. Esta caída fue el resultado de una disminución tanto en la actividad industrial (-25 %) como en el sector de servicios (-14 %). Dentro de éstos, la caída más profunda fue en los Servicios de alojamiento temporal y de alimentos y bebidas, que cayeron en un 60 % adicional, seguido del Comercio al menudeo (-31 %), Transporte, correos y almacenamiento (-26 %), Servicios de esparcimiento, culturales y deportivos (-24 %) y Comercio al mayoreo (-15 %). En la industria la caída provino tanto de la Construcción (-33 %) como de las Manufacturas (-31 %)" [5].

Sin embargo, todos estos son ejemplos de impactos económicos y sociales dentro de la sociedad en general pero ***¿cómo influyen los factores socioeconómicos existentes en la población mexicana al riesgo de mortalidad por COVID-19?***

Comencemos por conocer mejor el nivel económico y social que existe en nuestro país; para lo cual analizaremos el censo que realiza el Instituto Nacional de Estadística y Geografía (INEGI) cada 10 años para mantener un control de los factores socioeconómicos, mediante el cual se determina el tamaño de la población que reside en el país, sus condiciones de vivienda, escolaridad, salud, religión y etnicidad. Lo que nos ayuda a conocer el predominio de la población en cada uno de estos sectores, con la finalidad de tomar decisiones, llevar a cabo políticas y mejoras públicas para el crecimiento económico de la población y por lo tanto, mejorar su calidad de vida.

Hay que tener en cuenta que México es un país sumamente poblado, ocupando el lugar número 11 en población mundial con 126,014,024 habitantes, de los cuales 51.2 % corresponden a mujeres y 48.8 % a hombres, teniendo una tasa de población económicamente activa del 62 % y la tasa de participación económica es de 75.8 % hombres y 49 % mujeres. Más aún, de acuerdo al CONEVAL en su reporte del 2020, existe un alto índice de pobreza y pobreza extrema, el 43.9 % de la población vive en condiciones de pobreza y el 8.5 % en pobreza extrema. Además, los niveles de educación en el país son bajos ya que en promedio

se cursan 9.7 años de escolaridad, y un total de 4,456,431 personas son analfabetas, las cuales representan un 4.7 % de la población total, siendo las personas más marginadas del país sin un empleo formal ni seguro social de acuerdo al CENSO 2020. [15]

El promedio de hijas e hijos nacidos vivos de las mujeres de 12 años y más, en 2020 fue de 2.1 por mujer. Además en el país residen 7,364,645 personas que hablan alguna lengua indígena, en términos porcentuales representan el 6.1 % de la población, y si hablamos de religión el 77.7 % de la población se declara católica, 11.2 % se declara protestante o cristiano evangélico, 0.2 % declara otra religión, 2.5 % se declara creyente sin tener una adscripción religiosa y 8.1 % se declara sin religión.

Considerando las condiciones de vida dentro de una vivienda existen factores que registran el tipo de suelo o servicios básicos con los que cuenta cada vivienda, como por ejemplo: si cuenta o no con agua entubada, drenaje o excusado y electricidad. Las cifras que registra el CENSO 2020 sobre esto muestran que el 77.6 % de las viviendas registradas cuentan con agua entubada y el 78.1 % tienen drenaje o excusado y 99 % de ellas sí cuentan con electricidad.

Más adelante tomaremos todos estos factores socioeconómicos para analizar su relación e influencia con el riesgo de fallecer si una persona se contagia de COVID-19 ya que se tiene conocimiento de la importancia de los factores de salud del paciente pero no de cómo afecta la situación social o económica en la que viven a este riesgo.

Capítulo 2

Modelos matemáticos

«Essentially, all models are wrong, but some are useful»

- George Box

Un modelo es una representación compacta y sintetizada de algún fenómeno real o no, basado en proposiciones teóricas con sustento científico o empírico. Un modelo matemático es una herramienta científica para expresar relaciones entre las variables y componentes de un sistema complejo, difíciles de observar o predecir, con el fin de estudiar su dinámica y así encontrar soluciones exactas o aproximadas al problema usando el rigor y formalismo matemático, y citando al Dr.Jorge X. Velasco Hernández – *"La construcción del modelo matemático no empieza con la postulación de ecuaciones, sino cuando se plantea un tipo de pregunta que requiere de un enfoque cuantitativo y formal para su repuesta"*.

Ahora bien, la Epidemiología es una disciplina científica en el área de la Medicina que estudia la distribución, frecuencia y factores determinantes de las enfermedades existentes en una población, en la cual existen diversos tipos de modelos epidemiológicos. Un modelo epidemiológico se define como la representación matemática o lógica de la Epidemiología, es decir nos ayudan a determinar la dinámica de la transmisión y propagación de la enfermedad.

Por lo anterior, los modelos matemáticos son la herramienta perfecta para entender las epidemias y tomar las medidas necesarias para reducir los contagios hasta cesar la transmisión.

Los modelos epidemiológicos más utilizados en esta área son los modelos compartimentales, sin embargo, estos tienen un problema para la correcta modelización y predicción de los hechos futuros, debido a la no identificación de los parámetros. Un problema de identificación ocurre cuando distintos valores de los parámetros que caracterizan a un modelo compartimental (en el caso del modelo SIR, la tasa de transmisión y la tasa de recuperación) producen la misma solución que el modelo obtiene como resultado (la predicción del número de infectados)[8].

Por tanto, se optó por un modelo más apropiado para hacer predicciones, tomando como base la Probabilidad y Estadística para predecir el riesgo de mortalidad utilizando la información existente hasta el momento, para obtener tendencias y patrones de la dinámica del virus en los mexicanos y por consiguiente, haciendo el análisis adecuado se pueda obtener una predicción.

El modelo de predicción mediante Probabilidad y Estadística se obtiene al analizar la relación que existe entre las variables en juego. Es decir, se analizan sus componentes estadísticos así como la distribución que tiene y tendrá el fenómeno observado al variar en el tiempo tomando en cuenta información previa y la dinámica que presenta en ese momento.

2.1. Modelos de regresión

Un modelo de regresión es un modelo matemático que busca determinar la relación de una variable con respecto a otras; mientras que el análisis de regresión es la técnica estadística para investigar y modelar la relación entre variables, este tipo de análisis es uno de los más utilizados en numerosos campos de investigación como lo son la Medicina, Física, Química, Biología, Ingeniería, etc. Un aspecto esencial de este tipo de análisis es la recolección, recopilación u obtención de datos; y se dice que todo análisis de regresión es tan bueno como lo son sus datos.

De manera muy general, este tipo de modelos sirven para:

1. Entender o identificar la relación que existe entre las variables.
2. Predicción.

Ahora bien, hay que tener en cuenta que la complejidad matemática del modelo y la precisión de los resultados obtenidos depende mucho de cuánto se sepa acerca del proceso que está siendo estudiado, la cantidad de datos que se tengan, las suposiciones que se hagan y cómo se realice el análisis de los datos.

En los casos donde el objetivo principal es la predicción, los modelos casi siempre son *lineales en los parámetros*, por lo que se les conoce como **modelos lineales**, y aunque los modelos más realistas son frecuentemente *no lineales* algunos pueden ser transformados en modelos lineales para facilitar su estudio.

Hoy en día, se toman medidas y decisiones públicas a nivel nacional e internacional con base en la información científica para salvaguardar la salud de las personas, y es por este hecho que los modelos matemáticos predictivos se han vuelto fundamentales para entender los problemas sociales, económicos, demográficos, ecológicos, etc.

Los modelos de regresión son un tipo de modelos estadísticos que sirven para investigar, modelar y analizar la relación entre variables, entendiéndose por variable cualquier factor que pueda ser cuantificable de una población. [16]

Así que para ir entrando en materia, definamos los elementos que posee un modelo de regresión. Este tipo de modelos están determinados por los siguientes elementos:

- **Variable respuesta o dependiente:** Es nuestra variable de interés, la que buscamos predecir.
- **Variable regresora, explicativa o independiente:** Es aquella variable o variables que explican el comportamiento de la variable dependiente.
- **Coeficientes de regresión:** Son parámetros desconocidos del modelo.
- **Residual:** Es la diferencia que hay entre las observaciones y los valores ajustados.

El objetivo principal del análisis de regresión es estimar los parámetros desconocidos en el modelo de regresión para obtener *los coeficientes de regresión*. Este proceso se basa en el ajuste del modelo a los datos, un método para estimar estos parámetros es el de mínimos cuadrados que como su nombre lo indica, minimiza la suma de cuadrados de los residuales dando una muy buena aproximación de los valores de los parámetros del modelo, permitiendo así hacer predicciones de futuras observaciones.

2.2. Regresión lineal simple

El modelo de regresión lineal simple involucra solamente a una variable regresora (o independiente), la cual tiene relación con la variable dependiente, es decir, el resultado de la variable respuesta cambia en razón constante cuando el valor de la variable regresora crece o decrece. Este modelo de regresión simple queda definido por la siguiente ecuación:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

donde y es la variable respuesta, x es la variable independiente, β_0 y β_1 son constantes desconocidas y por último, ε es una componente de error aleatorio. Además, al ser ésta una ecuación lineal es importante notar que β_0 es la ordenada al origen, β_1 es el coeficiente que determina la pendiente que tiene la recta y ε es una variable aleatoria que representa el error del modelo.

Se supone que los errores tienen esperanza cero, varianza constante desconocida y que los errores no están correlacionados, i.e. $Cor(\varepsilon_i, \varepsilon_j) = 0$ para todo i distinto de j. Adicionalmente, se puede incluir un supuesto distribucional, el más común es que $\varepsilon_i \sim N(0, \sigma^2)$. La normalidad y la correlación cero, implica que los errores son independientes y por lo tanto las variables y_i son también variables aleatorias independientes.

Bajo estos supuestos, la respuesta media de cualquier valor de la variable regresora es:

$$E(y \mid x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x \tag{2.2}$$

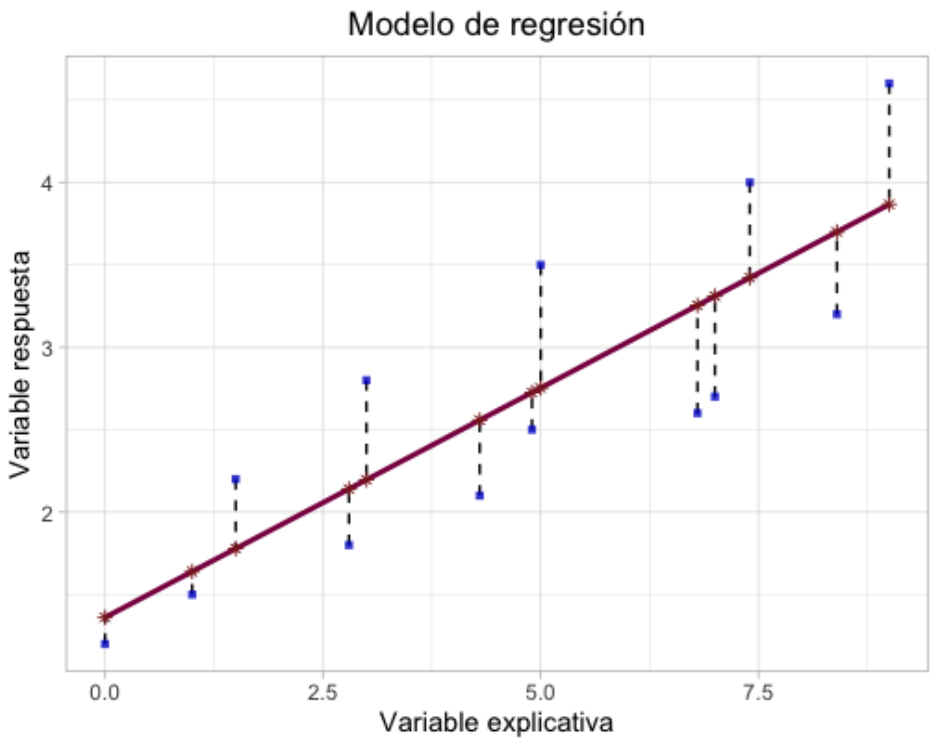


Figura 2.1: Representación gráfica de un modelo de regresión lineal simple con los valores observados y estimados marcados con un punto y un asterisco, respectivamente.

Mientras que la varianza de y dando cualquier valor de x es:

$$Var(y \mid x) = Var(\beta_0 + \beta_1 x + \varepsilon) = Var(\varepsilon) = \sigma^2 \tag{2.3}$$

Así, la media de y es una función lineal de x, aunque la varianza de y no depende del valor de x.

2.2.1. Estimación de los parámetros por mínimos cuadrados

En este caso se define la suma del cuadrado de los residuales, esto es

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Claramente, los parámetros β_0 y β_1 son desconocidos y se busca los valores que hagan que $S(\beta_0, \beta_1)$ sea mínimo. Esto es un problema básico de cálculo: primero se deriva $S(\beta_0, \beta_1)$

con respecto a β_0, luego con respecto a β_1. Segundo, se iguala a cero en ambos casos obteniendo las conocidas **ecuaciones normales**. Finalmente, resolviendo el sistema de ecuaciones, se obtienen los estimadores de los coeficientes de regresión:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

donde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

Y estos estimadores son los que dan como resultado el modelo ajustado de regresión lineal simple:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.4}$$

Es importante notar que desde el enfoque de los mínimos cuadrados no es necesario que los errores sigan una distribución normal, sean independientes, etc. Lo único que se necesita es suponer que existe una relación lineal entre las variables dependiente e independiente, esto es (2.1).

2.2.2. Estimación de los parámetros por máxima verosimilitud

Considerando los supuestos distribucionales descritos al inicio de esta sección, es fácil verificar que $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, para $i = 1, \ldots, n$. Además, como resultados de la independencia de los errores, las y_i's también son independientes entre sí. Entonces, la verosimilitud del modelo estaría dada por

$$L_n(\beta_0, \beta_1, \sigma^2 | y_1, \ldots, y_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}.$$

Al igual que en el caso de mínimos cuadrados, todo se reduce a un problema de cálculo en donde ahora existe un tercer parámetro que se debe estimar, esto es σ^2. Los estimadores para β_0 y β_1 son exactamente los mismo que los que se obtienen por mínimos cuadrados. Pero, el estimador insesgado para la varianza se estima vía:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}.$$

en donde los residuales $\hat{\varepsilon}_i = y_i - \hat{y}_i$, tienen por objetivo estimar a los errores del modelo ε_i.

2.2.3. Verificación de supuestos

Bajo los supuestos distribucionales, es posible obtener la distribución de los coeficientes de regresión (son normales), así como para el estimador de la varianza (ji-cuadrada). Este marco, brinda la posibilidad de obtener intervalos de confianza y realizar pruebas de hipótesis. Pero, en este caso es necesario verificar si los supuestos distribucionales se cumplen o no.

Al realizar el ajuste del modelo siempre es conveniente realizar las siguientes preguntas:

1. ¿El modelo se ajusta bien a los datos?
2. ¿Serán útiles nuestras predicciones con este modelo?
3. ¿Se cumplen todos los supuestos?, si no es así ¿qué tanto afecta eso al ajuste de los datos?

Con la finalidad de corroborar la adecuación del modelo antes de darlo por definitivo. Los residuales juegan un papel importante para evaluar la adecuación del modelo.

El diagnóstico del modelo de regresión aborda la adecuación de un modelo estadístico una vez que se han ajustado los datos. De hecho, el ajuste de un modelo debe verse como un proceso iterativo en el que se ajusta el modelo, se verifican los supuestos y se mejora si se requiere, así hasta llegar a un modelo óptimo.

A continuación se enumeran los supuestos que se deben verificar para determinar si el modelo es adecuado:

1. **Correcta especificación del modelo:** La relación debe ser lineal, esto es:

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i.$$

   Este es nuestro supuesto básico y se debe verificar siempre. La manera común de hacer lo anterior es mediante una gráfica de $(\hat{y}_i, \hat{\varepsilon}_i)$, para $i = 1, \ldots, n$. Lo que se busca es una gráfica nula. En la gráfica nula los residuales tienen que verse como si hubieran sido generados de manera independiente por una variable uniforme con media cero. La varianza de la uniforme puede cambiar para cada i, pero la media debe ser siempre 0.

2. **Independencia:** Los errores del modelo son independientes entre sí. Este supuesto se puede analizar en una gráfica de los residuales contra el tiempo para detectar si hay autocorrelación, ya sea positiva o negativa.

3. **Homocedasticidad:** Este supuesto se debe verificar cuando se supone $Var(\varepsilon_i) = \sigma^2$. Significa que los errores deben tener varianza constante. De nuevo se verifica con una gráfica de $(\hat{y}_i, \hat{\varepsilon}_i)$, para $i = 1, \ldots, n$, lo que se busca es una gráfica nula. En este caso los errores deben tener todos la misma varianza.

4. **Normalidad de los errores:** $\varepsilon_i \sim N(0, \sigma^2)$. Es necesario constatar que los residuales siguen una distribución normal, al menos de forma aproximada. Este supuesto sólo debe verificarse si se suponen los supuestos distribucionales. En este caso se realiza una gráfica conocida como QQ plot. En donde se grafican los cuantiles teóricos de una normal contra los cuantiles de los residuales. Si el supuesto se cumple, debería observarse un buen ajuste entre ambas gráficas.

5. **No hay correlación entre los errores:** Esto es $Cov(\varepsilon_i, \varepsilon_j) = 0, \ \forall \ i \neq j$.

6. **Multicolinealidad (regresión lineal múltiple):** En el caso en el que existan más de dos variables independientes (próxima sección), es importante determinar si

existe una relación lineal entre ellas. En algunos casos, esta relación lineal es tan fuerte que tiene consecuencias importantes. Por un lado, resulta irrelevante tener dos variables que nos den la misma información. Por otro lado, las columnas de la matriz que necesitaremos para calcular los coeficientes del modelo serían linealmente dependientes o tal vez sólo aproximadamente linealmente dependientes. En el primer caso, el sistema de ecuaciones que se tiene que resolver no tiene solución. Mientras que en el segundo se tendrán problemas numéricos que alterarían la solución que se obtenga.

Algunas de las maneras de determinar si se tiene multicolinealidad es mediante el **VIF**: el Factor de Incremento de Varianza. Otras más sencillas son calculando las correlaciones entre variables o simplemente de manera visual mediante la matriz de diagramas de dispersión.

7. **El número de observaciones es mayor que el número de variables independientes (regresión lineal múltiple):** En este caso el sistema de ecuaciones que se tiene que resolver para estimar los coeficientes de regresión tiene muchas soluciones.

2.3. Regresión lineal múltiple

Un modelo de regresión que involucra más de una variable regresora se llama modelo de regresión lineal múltiple (RLM). Y el modelo queda descrito de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i. \tag{2.5}$$

La notación incluye ahora un subíndice i que denota la unidad observacional a partir de la cual las observaciones en y y las k variables independientes fueron tomadas. Hay k variables independientes y $(k+1)$ parámetros que necesitan ser estimados (incluyendo a β_0) y además el tamaño de la muestra se denota por n, $i = 1, 2, \ldots, n$.
Por conveniencia se usará $k' = (k+1)$ suponiendo que $n > k'$, es decir, que se tienen más observaciones que parámetros.

Ahora bien, para expresar este modelo en notación matricial se necesitan cuatro matrices:

Y: El vector columna de tamaño $(n \times 1)$ que consta de las observaciones en la variable dependiente y_i;

X: La matriz de orden $(n \times k')$, formada por una columna cuyas entradas son todas uno, seguida de los k vectores columna de las variables independientes;

$\boldsymbol{\beta}$: Un vector columna de tamaño $(k' \times 1)$ cuyas entradas son los parámetros que serán estimados; y por último

$\boldsymbol{\varepsilon}$: El vector de tamaño $(n \times 1)$ de los errores aleatorios.

Con estas definiciones, el modelo de RLM puede ser escrito como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.6}$$

o bien

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{pmatrix}}_{n\times k'} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{k'\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{n\times 1}$$

Los vectores $\mathbf{Y}$ y $\boldsymbol{\varepsilon}$ son vectores aleatorios, es decir, los elementos de estos vectores son variables aleatorias. Cada elemento β_i de $\boldsymbol{\beta}$ es un coeficiente parcial de regresión que refleja el cambio en la variable dependiente Y_i por unidad de cambio en la i-ésima variable independiente, x_i.

Por otro lado, se supone que $\boldsymbol{\varepsilon}$ tiene esperanza cero (vector) y matriz de varianzas y covarianzas es generalizada a la matriz de varianzas y covarianzas del vector $\boldsymbol{\varepsilon}$. También se puede incluir el supuesto distribucional sobre los errores, en este caso sería el de una normal multivariada.

Nota: La matriz de varianzas y covarianzas de cualquier vector aleatorio de n elementos, se define como una matriz simétrica, cuyos elementos en la diagonal principal son iguales a la varianza de las variables aleatorias; y los (i, j)-ésimos elementos que se hallan fuera de la diagonal principal, son las covarianzas entre ε_i y ε_j.

$$Var(\boldsymbol{\varepsilon}) = \begin{pmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) & \cdots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_1, \varepsilon_2) & Var(\varepsilon_2) & \cdots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & & \vdots \\ Cov(\varepsilon_1, \varepsilon_n) & Cov(\varepsilon_2, \varepsilon_n) & \cdots & Var(\varepsilon_n) \end{pmatrix}$$

Ahora, como ε_i son variables no correlacionadas con varianza común y constante, σ^2, se obtiene la matriz diagonal:

$$Var(\boldsymbol{\varepsilon}) = \begin{pmatrix} Var(\varepsilon_1) & 0 & \cdots & 0 \\ 0 & Var(\varepsilon_2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Var(\varepsilon_n) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

De esta forma, se tiene:

$$\boldsymbol{\varepsilon} \sim N(\bar{0}, \sigma^2 I_n)$$

Así, $\mathbf{Y}$ es un vector aleatorio cuya media es:

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$

Y varianza:

$$Var(\mathbf{Y}) = Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

Bajo los supuestos distribucionales, $\mathbf{Y}$ es una normal multivariada, ya que la distribución de $\boldsymbol{\varepsilon}$ también lo es. Entonces, $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$. En donde el vector de residuales

estimados queda definido como $\hat{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, así la suma de cuadrados de los residuales está dada por:

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

2.3.1. Estimación de los coeficientes de regresión por mínimos cuadrados

Si aplicamos el método de mínimos cuadrados para estimar los coeficientes de regresión, la solución de las ecuaciones normales serán los estimadores por mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Y para determinar el vector $\hat{\boldsymbol{\beta}}$ de estimadores que minimice la distancia entre el valor observado y el estimado se tiene:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.7}$$

Igualando a cero, derivando con respecto a $\boldsymbol{\beta}$ y simplificando obtenemos las **ecuaciones normales**

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}, \tag{2.8}$$

Para resolver las ecuaciones normales se multiplican ambos lados de (2.8) por la inversa de $\mathbf{X}'\mathbf{X}$, así el estimador para $\hat{\boldsymbol{\beta}}$ por mínimos cuadrados es:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}), \tag{2.9}$$

La matriz $(\mathbf{X}'\mathbf{X})^{-1}$ siempre existe si las columnas de X (las variables regresoras) son linealmente independientes. Por lo tanto, el modelo ajustado de regresión queda definido como $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

2.3.2. Estimación de σ^2

Se puede seguir de nuevo el camino de máxima verosimilitud, en este caso se obtendría un estimador para σ^2. Los estimadores para los coeficientes de regresión, serán los mismos

que los obtenidos por mínimos cuadrados. A continuación, sólo se describe el estimador para σ^2 en el caso multivariado.

Al igual que en el caso de una sola variable independiente, en el caso multivariado la suma de residuales se usa para estimar σ^2, i.e.

$$SS_{Res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\hat{e}_i^2 = \hat{e}'\hat{e}$$

Se sustituye $\hat{e} = Y - X\hat{\boldsymbol{\beta}}$ y se obtiene:

$$\begin{aligned} SS_{Res} &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{2.10}$$

dado que $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ la ecuación (2.10) puede escribirse de la siguiente manera:

$$SS_{Res} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \tag{2.11}$$

La suma de cuadrados de residuales tiene $(n - k')$ grados de libertad asociados, puesto que existen k' parámetros en el modelo que se quiere estimar. Así, el **cuadrado medio del error** o **error cuadrático medio** es

$$MS_{Res} = \frac{SS_{Res}}{n - k'} \tag{2.12}$$

Y el valor esperado de MS_{Res} es σ^2. Por lo que un estimador insesgado de σ^2 que depende del modelo sería

$$\hat{\sigma}^2 = MS_{Res} \tag{2.13}$$

Por lo tanto, el **error estándar residual** queda definido como:

$$\hat{\sigma} = \sqrt{MS_{Res}} \tag{2.14}$$

2.4. Validación del modelo

En esta sección se explica la siguiente fase del análisis de regresión que es la comprobación de la adecuación del modelo. En este caso los supuestos distribucionales son esenciales. Y nos sirven para evaluar la calidad del ajuste y la utilidad del modelo de regresión mediante la prueba de significancia de la regresión.

La prueba de significancia de la regresión indica si al menos una de las variables independientes nos ayuda a describir la variable dependiente y se realiza vía la prueba *F–Fisher*. Para verificar la significancia para cada variable independiente, se utiliza la estadística *t–Student*.

2.4.1. Significancia

La prueba de significancia para la regresión se basa en una estadística F–Fisher. Las hipótesis son las siguientes:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \quad \textit{al menos para una } j$$

donde H_0 indica que ninguna variable independiente es útil para describir a la variable dependiente y H_1 indica que al menos hay una variable independiente que nos ayuda a describir a la variable dependiente.

Para obtener la estadística F, la **suma total de cuadrados**, SS_T se divide en la **suma de cuadrados debidos a la regresión**, SS_R, más la **suma de cuadrados de residuales**, SS_{Res}. Así,

$$SS_T = SS_R + SS_{Res}.$$

Esta ecuación es la igualdad fundamental en el análisis de varianza para un modelo de regresión, en donde

$$SS_R = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2. \tag{2.15}$$

y

$$SS_{Res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{2.16}$$

La varianza total de la variable dependiente, está dada por la conocida expresión

$$SS_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \tag{2.17}$$

La varianza total, SS_T, se divide en la cantidad de variabilidad en las observaciones explicada por la línea de regresión (SS_R), más la variación residual que queda sin explicar por la línea de regresión (SS_{Res}).

Es posible demostrar que, bajo el supuesto de normalidad de los residuales y suponiendo H_0 cierta, que

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k')} = \frac{MS_R}{MS_{Res}}, \tag{2.18}$$

sigue una distribución $F_{k,n-k'}$ y si el valor observado de F_0 es grande, da evidencia de que al menos una $\beta_j \neq 0$. Por consiguiente se rechaza la hipótesis nula si

$$F_0 > F_{\alpha,k,n-k'}$$

Si H_0 es cierta, entonces $\dfrac{SS_R}{\sigma^2}$ tiene una distribución χ^2_k con la misma cantidad de grados de libertad que la cantidad de variables regresoras en el modelo.

Para saber si el modelo es adecuado, es necesario evaluar la significancia de los coeficientes. La significancia de β_0 casi nunca resulta de gran importancia ya que la intersección no tiene mas interpretación que el valor que toma la variable respuesta cuando $x = 0$, y eso sólo en el caso en el que la variable independiente puede tomar el valor 0. Por otro lado, la significancia de β_i es realmente importante ya que está relacionada con la existencia de una relación lineal entre las variables x_i y y; además nos dice que variables son significativas para el modelo. [23]

La prueba de significancia de la regresión es quizá el paso más importante para evaluar la adecuación del modelo. En este punto ya tenemos la certeza de que la variable dependiente puede describirse de forma lineal usando al menos una de las variables independientes.

Por otro lado, las pruebas con la estadística t nos sirven para probar la significancia de cualquier coeficiente de regresión. Para el modelo de regresión lineal simple, es posible demostrar que la prueba t para β_1 es equivalente a la prueba de significancia de la regresión. Sin embargo, con más variables independientes esto no se cumple.

Para probar la significancia de cualquier coeficiente individual de regresión se realiza la prueba de hipótesis:

$$H_0 : \beta_j = 0,$$
$$H_1 : \beta_j \neq 0,$$

Para realizar este contraste, se utiliza el estadístico de prueba:

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}, \tag{2.19}$$

en donde

$$se(\hat{\beta}_j) = \sqrt{MS_{Res} \cdot C_{jj}},$$

es el **error estándar estimado** [16], donde C_{jj} es el elemento diagonal de $(X'X)^{-1}$ que corresponde a $\hat{\beta}_j$. Y la hipótesis nula se rechazaría si $\mid t \mid > t_{\alpha/2,n-2}$. Mientras que el p-valor está dado por $p-value = P(|t| >= t_{\alpha/2,n-2})$. El $p-valor$ indica la probabilidad de obtener valores de la estadística de prueba más extremos (peores) al observado, suponiendo que H_0 es cierta. Obviamente, p-valores pequeños indican que es poco probable obtener valores peores de la estadística de prueba. Lo que nos da evidencia para rechazar H_0, formalmente se rechazará H_0 al nivel de significancia α si $p - valor < \alpha$.

2.4.2. Bondad de ajuste

Una vez que tenemos certeza de que al menos una variable independiente puede ser usada para describir a la variable dependiente. Nos gustaría tener una medida de la calidad de ajuste de nuestro modelo. Es decir que tan bien describe el modelo a la variable

dependiente. Para ello, una medida muy útil es el coeficiente de determinación:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{2.20}$$

es decir,

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

La estadística R^2 indica la proporción de la varianza de las y_i's explicada por el modelo de regresión. Y, ya que $0 \leq SS_{Res} \leq SS_T \Rightarrow 0 \leq R^2 \leq 1$; los valores cercanos a 1 implican que la mayor parte de la variabilidad de Y está explicada por el modelo de regresión.

En general, R^2 aumenta siempre que se agrega una variable regresora al modelo, independientemente de la contribución de esa variable. En consecuencia, no es buena idea comparar modelos de regresión con diferente número de variables usando la estadística R^2.

Existe una modificación de R^2 conocido como R^2 "ajustada"que se prefiere cuando se trabajan con modelos de regresión múltiple ya que considera una penalización por la cantidad de variables regresoras que se incorporan al modelo.

La R^2 **ajustada** se define como

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-k')}{SS_T/(n-1)}. \tag{2.21}$$

2.4.3. Ajuste del modelo de RLM usando R

Como se ha indicado, el modelo de RLM es una de las técnicas estadísticas más usadas. Por lo tanto, es común ajustar este tipo de modelos usando paquetes estadísticos que cuentan con rutinas para realizar todos los procedimientos que hemos descrito en sólo unas décimas de segundo. En la Tabla 2.1 se despliega una salida típica del lenguaje de programación estadístico `R`. En particular la conocida función `lm`, seguida de la función `summary`. A continuación se describe cada detalle de esta salida. Como marco de referencia se utilizará toda la teoría desarrollada hasta ahora.

El modelo que se intenta ajustar es:

$$y_i = \beta_0 + \beta_1 \cdot GPEST + \beta_2 \cdot PEA + \beta_3 \cdot POV + \beta_4 \cdot NEUMONIA +$$
$$+ \beta_5 \cdot EDAD + \beta_6 \cdot TPACIE + \beta_7 \cdot OBESIDAD + \beta_8 \cdot TABAQUISMO + \varepsilon.$$

	Estimate	Std. Error	t-Value	Pr(>| t |)
(Intercept)	-2.893501	1.354227	-2.137	0.033142
GPEST	-0.049660	0.015914	-3.121	0.001916
PEA	2.895354	1.352912	2.140	0.032861
POV	2.934649	1.353735	2.168	0.030673
NEUMONÍA	2.947863	1.369447	2.153	0.031858
EDAD	0.003313	0.000587	5.644	2.87e-08
TPACIE	0.190211	0.029325	6.486	2.23e-10
OBESIDAD	0.135902	0.035629	3.814	0.000155
TABAQUISMO	0.211912	0.073029	2.902	0.003885
Residual standard error: 0.1404 on 471 degrees of freedom				
Multiple R-squared: 0.3655, Adjusted R-squared: 0.3533				
F-statistic: 30.14 on 8 and 471 DF, p-value: $< 2.2e-16$				

Tabla 2.1: Salida de un ejemplo de modelo de regresión lineal múltiple en R.

En la primer columna se tienen las *variables del modelo de regresión*, la segunda los *estimadores*, es decir, los coeficientes estimados del modelo de RLM, $(\hat{\beta}_j)$, para cada variable. La tercera fila muestra el *error estándar estimado* de cada estimador, denotado como $se(\hat{\beta}_j)$; la cuarta fila muestra el valor de la *estadística t* para realizar el contraste $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$; y la última columna corresponde al p-valor para realizar el mismo contraste.

En la parte baja de la tabla se presentan los coeficientes de determinación, tanto R^2 como R^2_{Adj}. También nos arroja $\hat{\sigma}$ – el error estándar residual – que nos proporciona una medida de la dispersión de los residuales en nuestro modelo; y por último vemos la estadística F junto con su p-valor para realizar la prueba de significancia de la regresión. Finalmente, se indica el número de variables independientes $k = 8$ y el número de observaciones $n = 471$.

Una breve interpretación sería la siguiente: la prueba de significancia de la regresión

indica que se rechaza $H_0 : \beta_1 = \cdots \beta_8 = 0$ para básicamente cualquier $\alpha > 0$. Por lo que al menos una de las variables es significativa. En el caso de las pruebas de significancia para cada β_j, en todos los casos se rechaza la hipótesis nula $H_0 : \beta_j = 0$ a un nivel de significancia de $\alpha = 0.05$. El coeficiente de determinación R^2 indica que el modelo describe aproximadamente 36 % de la variabilidad presente en las y's.

2.5. Predicción

Una de las aplicaciones más importantes de los modelos de regresión es la de predecir el valor de nuevas observaciones que correspondan a valores específicos de las variables independientes. Supongamos que se quiere predecir el valor y_*, asociado a las valores de las variables independientes $x_*^t = [1, x_{*1}, x_{*2}, \ldots, x_{*k}]$. En los datos que se utilizaron para estimar al vector $\hat{\beta}$, no se encuentra el renglón x_0 y obviamente no se conoce y_*. La estimación puntual de la predicción se obtiene fácilmente como $\hat{y}_* = x_*^t \hat{\beta}$. Mientras que el intervalo de predicción del $(1 - \alpha \times 100\,\%)$ estaría dado por:

$$sepred(\hat{y}_* \mid x_*) = \hat{\sigma}\sqrt{1 + x_*^t (X'X)^{-1} x_*} \tag{2.22}$$

definido también como el error estándar de predicción. [23]

Que de manera extensa sería:

$$\hat{y}_* - t_{(n-)}^{1-\alpha/2}\sqrt{MS_{Res}}\sqrt{1 + x_*^t (X'X)^{-1} x_*} \leq y_* \leq \hat{y}_* + t_{(n-)}^{1-\alpha/2}\sqrt{MS_{Res}}\sqrt{1 + x_*^t (X'X)^{-1} x_*}$$

2.6. Selección de variables

Muchos problemas de regresión tienen como objetivo principal evaluar el efecto que tienen las variables independientes en la variable respuesta. En este caso, es poco práctico incluir muchas variables independientes, ya sea por la interpretación de los resultados o para aumentar la precisión de las pruebas y estimaciones.

En otros casos el objetivo de la regresión podría ser la predicción de valores futuros

de una respuesta dados los valores de las variables independientes. Incluir demasiadas variables independientes puede conducir a predicciones inexactas porque el modelo ajustado podría dar cuenta de las peculiaridades de los datos observados que no están presentes en observaciones futuras. Por otro lado, usar un modelo con muy pocas variables independientes también puede conducir a malas predicciones si las variables independientes más importantes son omitidas.

Finalmente, es importante mencionar que siempre es preferible tener modelos con el menor número de variables independientes pues son más fáciles de mantener. En otras palabras, es más sencillo actualizar la información de sólo 4 variables independientes que de 30.

El problema de la selección de variables es uno de los puntos medulares de cualquier análisis de regresión. Existen distintos métodos que se basan en alguna estadística que nos permita comparar modelos con distintas variables independientes y así poder elegir el modelo con el menor número de variables pero que a la vez describa *"mejor"* la dinámica de la variable respuesta.

La mayoría de los métodos de selección de variables consideran alguna de las siguientes estadísticas:

p-valor: Se toma en cuenta el p-valor asociado a cada β_i, se incluyen las variables independientes con los menores p-valores.

AIC: El criterio de información de Akaike se basa en el cálculo de la estadística

$$AIC = n \log(SS_{Res}/n) + 2k,$$

en donde SS_{Res} es una medida de falta de ajuste del modelo, n es el número de observaciones y $2k$ es una penalización por el número de variables independientes que se incluyen en el modelo. Modelos con menor AIC son preferidos.

BIC: El criterio de información de Bayes es muy similar al AIC, pero la penalización

debida al número de variables es distinta, esto es

$$BIC = n\log(SS_{Res}/n) + \log(n)k,$$

Modelos con menor BIC son preferidos.

R^2_{Adj}: Se busca siempre al que tenga mayor R^2_{Adj}.

Existen muchas otras alternativas para comparar modelos, pero en esta tesis nos centraremos en las anteriores. Para conocer más alternativas se puede consultar la siguiente referencia [2].

La manera obvia para elegir el mejor modelo es hacer todas las regresiones posibles (todas las posibles combinaciones de variables independientes) y para cada regresión calcular, por ejemplo, el AIC. Sin embargo, el número de regresiones posibles con k variables independientes es 2^k. Si tenemos 20 variables independientes sería necesario ajustar $2^{20} = 1,048,576$ modelos, lo que quizá sea factible. Sin embargo, si se tienen 30 variables, el número de modelos a considerar es de $2^{30} = 1,073,741,824$, lo que ya no es viable. Por lo anterior, las estrategias más utilizadas para seleccionar variables son los métodos de selección **paso a paso**. Estas estrategias consisten en agregar o eliminar una variable cada vez, comparando el modelo actual con el siguiente basándose en alguno de los criterios mencionados anteriormente. Este tipo de métodos tiene 3 posibilidades:

Forward: Selecciona paso a paso hacia adelante. Empieza con el modelo más simple y va agregando una a una las variables hasta que el modelo deje de mejorar.

Backward: Selecciona paso a paso hacia atrás. Empieza con el modelo que incluye a todas las variables y elimina una a una, hasta que el modelo deje de mejorar.

Both: Selecciona paso a paso en ambas direcciones. Es una combinación de los métodos anteriores.

Dado que los métodos pueden usar criterios y direcciones distintas, es posible que al usar dos métodos de selección distintos se llegue a dos modelos distintos, es por eso que este tipo de análisis debe ser muy cuidadoso para no incurrir en un error de especificación.

2.7. Transformaciones

Si en la validación del modelo se obtuvo un bajo coeficiente de determinación o se detectan violaciones a algunos supuestos como la falta de normalidad o varianza no constante, entonces es recomendable realizar transformaciones para una mejor adecuación del modelo aunque es posible que se pierda el poder de interpretación, es decir, transformar las variables involucradas dentro de un modelo de regresión puede ser una forma de corregir o mejorar el ajuste del modelo yo ayudar a que se cumplan los supuestos de regresión pero con ciertas consecuencias.

Es importante descubrir y corregir una varianza no constante de los residuales. Si no se elimina este problema, los estimadores de mínimos cuadrados seguirán siendo insesgados, pero ya no tendrán la propiedad de la varianza mínima. Esto quiere decir que los coeficientes de regresión tendrán errores estándar mayores que lo necesario. Hacer una transformación suele proporcionar estimados más precisos de los parámetros del modelo y mayor sensibilidad para las pruebas estadísticas.

2.7.1. Transformaciones de potencias

La familia de transformaciones que se usa más a menudo es la *familia potencia* definida por una variable $\mathcal{U}$ estrictamente positiva como

$$\Psi(U, \lambda) = U^{\lambda} \tag{2.23}$$

donde los valores de λ más usados en la práctica se encuentran en el rango $[-1, 1]$ y el valor de $\lambda = 0$ corresponde a la transformación log.
Lo que se busca es lograr una función media que sea lineal en la escala transformada. En un modelo de regresión lineal existen dos reglas empíricas que sirven a menudo para saber qué variable transformar y qué transformación sería conveniente.

Regla del logaritmo Si los valores de un rango de variables oscilan en más de un orden de magnitud y la variable es estrictamente positiva, entonces reemplazamos la variable por su logaritmo y probablemente será una transformación útil.

Regla del rango Si el rango de una variable es considerablemente menor que un orden de magnitud, entonces cualquier transformación de esa variable será inservible.

Para seleccionar una transformación, es conveniente introducir a la *familia de transformaciones de potencia escalada*, definida para una variable X estrictamente positiva como

$$\Psi_s(X,\lambda) = \begin{cases} \dfrac{X^\lambda - 1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ \log(X) & \text{if} \quad \lambda = 0 \end{cases} \tag{2.24}$$

donde $\Psi_s(X,\lambda)$ es continua como una función de λ.

Un estimador $\hat{\lambda}$ es el valor de λ que minimiza a la SS_{Res}. Las transformaciones de potencia escalada preservan la dirección de asociación, en el sentido de que si (X,Y) están positivamente relacionadas, entonces $(\Psi_s(X,\lambda),Y)$ también estarán positivamente relacionadas para todos los valores de λ.

Se pueden transformar tanto la variable respuesta como las variables regresoras ya sea una o todas a la vez; cada variable puede tener su propia transformación y no tiene que ser la misma transformación para todas necesariamente. Sin embargo, hay que tener en cuenta como enfatiza Weisberg "...no todas las transformaciones útiles corresponderán a modelos físicos interpretables." [1]

Por lo tanto se debe considerar qué es lo mejor para el objetivo de la investigación: si hacer la transformación de variables y mejorar el modelo aún perdiendo la interpretación del mismo o quedarnos con el modelo inicial para poder interpretar sus resultados.

2.7.2. Método Box – Cox

Este método es usualmente aplicado para seleccionar la transformación de la variable respuesta pero también se utiliza para seleccionar las transformaciones de las variables regresoras simultáneamente, para esto se usa una versión de la familia potencia ligeramente más complicada llamada *familia potencia modificada*, definida por Box y Cox (1964) para una variable Y estrictamente positiva como:

[1] *Applied Linear Regression*, Weisberg S., pág. 188.

$$\Psi_M(Y, \lambda_y) = \Psi_s(Y, \lambda_y) \times gm(Y)^{1-\lambda_y} \tag{2.25}$$

donde $gm(Y)$ es la media geométrica de la variable no transformada. La multiplicación de la transformación de potencia escalada por $gm(Y)^{1-\lambda}$ garantiza que todas las unidades de $\Psi_M(Y, \lambda_y)$ sean los mismos para todos los valores de λ_y, y así toda la $SS_{Res}(\lambda_y)$ tenga las mismas unidades.

El método de Box–Cox busca transformaciones para cumplir con la normalidad: se elije λ para hacer que los residuales de la regresión $\Psi(Y, \lambda_y)$ en X estén lo más cerca posible a una distribución normal. Más aún, no olvidemos que usar las transformaciones elegidas por los métodos aquí presentados puede no proveer un función media comprensible y no necesariamente van a mejorar las cosas; cuando se cuenta con muchas variables regresoras la estrategia para obtener buenas transformaciones es estar intentando, intentando e intentando muchas transformaciones y quedarnos con la que nos dé los mejores resultados.

Capítulo 3

Bases de datos de trabajo

La principal fuente de información para esta tesis fue la que la Secretaría de Salud genera a partir del Sistema Nacional de Vigilancia Epidemiológica (SINAVE) así como del Sistema de Vigilancia Epidemiológica de Enfermedad Respiratoria Viral (SISVER), y que por diversas disposiciones se pone al alcance de todos los interesados como parte del proyecto de datos abiertos del Gobierno de México. Esta información puede descargase de la página oficial [7]. La base de datos no tiene un nombre en específico, se le conoce como base de datos de la pandemia, base SINAVE/SISVER, base de datos del COVID19 en México, etc.

La base de datos SINAVE/SISVER incluye todas las pruebas procesadas por la red de laboratorios del sistema de salud pública nacional, la cual incluye el Instituto Nacional de Diagnóstico Epidemiológico y de Referencia así como los 32 laboratorios estatales de salud pública para el monitoreo y apoyo epidemiológico. La base de datos no incluye las pruebas realizadas y procesadas en laboratorios privados, las cuales se colectan en otra plataforma.

SINAVE/SISVER contiene información de pruebas, hospitalización y decesos de 5,186 unidades, que incluyen a las 475 unidades del sistema Centinela, de las cuales 3 son privadas, además de otras 4,281 públicas y 430 privadas distribuidas en los tres niveles del sistema de salud. El sistema es de amplio espectro pero no contempla registrar todo caso o deceso en el país[1].

[1] http://covid-19.iimas.unam.mx/

Se trabajó con una base de datos que se actualiza diariamente desde el inicio de la pandemia en nuestro país, al 16 de marzo del 2022 que pesa 2.5 GB y contiene poco más de 15 millones de pacientes registrados, de los cuales 5,619,780 son positivos que fueron los datos utilizados para nuestro análisis, unos pocos sospechosos y el resto negativos. Se trabajó tanto con datos de la CDMX como con datos de toda la República Mexicana, haciendo énfasis en los síntomas, las comorbilidades, fechas de registro, inicio de síntomas y defunción, además de la edad y el sexo, todo con el fin de obtener un buen análisis del comportamiento de la pandemia y de esa forma determinar su dinámica y modelar posibles escenarios que la expliquen.

Por otro lado se trabajó con la base de datos del CENSO de Población y Vivienda 2020, publicada por el INEGI [15] en donde se reporta tanto la dimensión, estructura y distribución espacial de la población que reside en México, como sus principales características socioeconómicas y culturales. En esta base se reporta una población total en los Estados Unidos Mexicanos de 126,014,024 habitantes, llegando a ocupar el 10vo lugar en términos de población total a nivel mundial de acuerdo a las últimas estimaciones de World Population Prospects 2019 [21].

3.1. Primer intento: CDMX

Inicialmente se intentó trabajar con una base de datos a nivel localidad (un nivel de desagregación mayor al de municipio), enfocándonos sólo en la Ciudad de México ya que es el estado con mayor tasa de contagios en todo el país. Para unir tanto la base de datos de la pandemia como la base de datos del CENSO, la clave es generar una ID único de localidad en ambas bases de datos (usando las mismas variables).

Empezamos con la base de datos SINAVE/SISVER. Primero, se filtraron sólo los casos positivos a COVID-19 y en seguida se consideraron sólo los residentes de la CDMX (estado de residencia del paciente con clave INEGI 09). Posteriormente, se generó un ID único de cada localidad concatenando las claves INEGI de municipio y de localidad (ambas se encuentran en la base de datos); se supuso que al ser una base de datos del gobierno federal

habría consistencia conforme a los registros de INEGI, y así cada ID correspondería a una localidad en específico.

Continuando con la base de datos del CENSO, se hizo un filtro eliminando las claves 0000, 9998 y 9999 debido a que INEGI pone los totales de cada localidad y cada municipio, provocando por ende que se multiplique la suma total de la población final. La clave 0000 es donde se engloban las cifras totales a nivel nacional, entidad federativa y municipio; la clave 9998 engloba las cifras totales a nivel nacional, entidad federativa y municipio de las localidades de una vivienda, y la clave 9999 engloba las cifras totales a nivel nacional, entidad federativa y municipio de las localidades de dos viviendas. Sin embargo no hay que olvidar que al hacer todos los filtros correspondientes se pierde cierta información, ya que hay registros con datos no visibles o confidenciales porque existen localidades muy pequeñas donde puede haber una sola casa y se expondrían sus datos personales, así que estos aparecen con un asterisco y al quitar los totales donde están contemplados se pierde esa información.

Una vez filtrada la base del CENSO, se tomaron sólo los registros referentes a la CDMX (clave INEGI de estado 09) y se generó el mismo ID único de localidad (concatenando la clave INEGI de municipio y localidad) que en la base de la pandemia. Así, se procedió a unir ambas bases de datos (función *merge* de R). Sin embargo, se detectó un problema: las claves INEGI de localidad en la base de la pandemia de la CDMX no coinciden con las claves de localidad en la base de datos del CENSO. En la base de datos de la pandemia, en ocasiones, se pusieron claves de localidad, pero en la mayoría de los casos pusieron colonias, calles o zonas enteras como localidades, teniendo un total de 708 “localidades”; mientras que de acuerdo al CENSO 2020 las localidades en la Ciudad de México son únicamente 494. Como resultado al unir las bases se obtuvieron 190 coincidencias solamente.

Debido a esta falta de consistencia y de intentar corregir el problema, sin éxito. Se tomó la decisión de trabajar con la base de datos SINAVE/SISVER a nivel nacional pero ahora a nivel municipio. Suponiendo que a un nivel de desagregación mayor, sí habría consistencia.

3.2. Base de datos de la pandemia a nivel nacional

La base de datos SINAVE/SISVER a la fecha del último análisis realizado (16 de marzo del 2022) tenía información de $15,396,315$ casos totales, incluyendo positivos, sospechosos y negativos.

El total de variables que conforman esta base son 40, de las cuales se utilizaron 18 para hacer nuestro análisis: edad, sexo, tipo de paciente, clasificación del paciente, fecha de defunción, tabaquismo, enfermedades crónicas y comorbilidades. La mayoría de estas variables son dicotómicas indicando presencia o ausencia de la característica. Para poder trabajar con esta base se realizó un filtro como se hizo con la base de la Ciudad de México; se utilizó la variable 'Clasificación Final' que tiene 7 clasificaciones distintas representadas con los números del 1 al 7, de las cuales 3 son de nuestro interés (1, 2 y 3): las clasificaciones 1 y 2 representan los casos de COVID-19 confirmados y la 3 representa los casos de SARS-COV-2 confirmados, obteniendo un total de $5,619,780$ casos positivos hasta el mes de marzo del 2022.

Para obtener nuestra base de datos final se hicieron diferentes filtros además de los ya mencionados y algunas modificaciones pertinentes para poder hacer un análisis con los datos. Se generó un ID único concatenando las claves INEGI de entidad federativa y de municipio, este ID está compuesto por 5 dígitos, los 2 primeros corresponden al estado y los 3 posteriores corresponden al municipio, obteniendo así una lista ordenada de los municipios de cada estado, por ejemplo: el municipio de Aguascalientes del estado de Aguascalientes tiene la clave 01001, el municipio de Ensenada del Estado de Baja California tiene la clave 02001, etc.

Las variables con información de las comorbilidades contienen valores de 'SI' y 'NO' de acuerdo a si padecen alguna de ellas o no, se modificaron por '1s' y '0s' respectivamente, al igual que la variable 'sexo' donde el número 1 representa al sexo femenino y el 0 al sexo masculino; el 'tipo de paciente' se modifica de igual forma con '1s' a los pacientes que fueron hospitalizados y '0s' a los pacientes ambulatorios, y por último NA's para cualquiera de los registros que contenga valores no especificados, se ignora y/o no aplica que vienen

representados por los números 99, 98 y 97, respectivamente. De esta manera, las variables categóricas se convirtieron en variables numéricas a nivel paciente. Finalmente, estos unos y ceros se promediaron y multiplicaron por cien a nivel municipio: obteniendo una base de porcentajes a nivel municipal. En total se obtuvieron $2,437$ municipios registrados en esta base a lo largo de toda la República Mexicana.

Manejar porcentajes a nivel municipal hace que tanto el análisis exploratorio como el de regresión sean más fáciles de realizar e interpretar.

Nuestra base de datos final de la pandemia en México, se muestra a continuación. Se despliegan sólo las primeras 6 variables.

ID	**SEXO**	**EDAD**	**TPACIE**	**DEF**	**NEUMO**	**DIAB**	$\cdots$
01001	50	41.5	19.6	09.01	11.02	12.53	$\cdots$
01002	54.85	41.51	19.83	07.28	10.52	09.91	
01003	56.92	41.21	97.07	02.54	05.08	11.68	
01004	45.11	40.47	23.3	09.02	16.54	14.28	
01005	51.84	40.89	19.16	07.15	11.08	12.81	
01006	53.26	40.80	14.97	06.86	09.34	10.92	
01007	53.06	42.21	22.72	07.48	14.82	15.37	
01008	48.43	43.28	12.50	05.46	08.59	17.18	
01009	53.69	42.80	19.84	06.22	13.22	12.84	
01010	49.27	53.33	42.02	20.28	34.78	26.08	
01011	52.91	41.54	22.26	08.75	10.21	16.78	
02001	51.14	44.23	22.47	13.23	16.76	17.88	
02002	52.35	44.94	26.27	15.25	22.74	18.60	
$\vdots$							$\ddots$

Tabla 3.1: Se muestra solamente una parte de nuestra base de trabajo con los datos de los pacientes contagiados de COVID–19 durante la pandemia en México. En total se trabajó con 19 variables y $2,437$ observaciones.

En la siguiente sección se describen las variables del CENSO, así como el procedimiento para unir las dos bases de datos.

3.3. Censo de Población y Vivienda 2021

El CENSO se conforma de un registro por localidad con los datos de identificación geográfica, así como 222 indicadores con las características de la población, los hogares censales y las viviendas. [15]

Por un lado los indicadores de la población que corresponden a su estructura por Sexo y Edad, Fecundidad, Migración, Etnicidad, Discapacidad, Educación, Características Económicas, Servicios de Salud, Situación Conyugal y Religión. En cuanto a **Hogares** censales la información está relacionada con el número de hogares y su población, de acuerdo con la persona de referencia del hogar. Por otro lado, en lo que respecta a **Vivienda**, destacan factores como: viviendas y ocupantes, material de pisos, número de cuartos, servicios de que disponen (energía eléctrica, agua entubada, sanitario, drenaje) y bienes en la vivienda.

Para construir nuestra base de trabajo primero se extrajeron los sub-totales a nivel municipio de varias variables y se realizaron cocientes, multiplicando por cien. Por ejemplo, para obtener el porcentaje de analfabetas de cada municipio se tomó la población total de 15 años o más analfabeta y se dividió entre la población total de 15 y más años, se multiplicó por cien. Se hizo algo similar para la PEA, población sin seguridad social, población católica, etc. En otros casos simplemente se tomó el sub-total a nivel municipal: grado promedio de escolaridad, promedio de hijos nacidos vivos y promedio de ocupantes en viviendas particulares habitadas. Esto es muy sencillo pues INEGI incluyó estos subtotales en la base de datos a nivel localidad. Finalmente, se obtuvo el ID de cada municipio de la misma manera que con la base de la pandemia, obteniendo así los mismos IDs para poder hacer el *merge* entre la base de datos de la pandemia y la base del CENSO.

En esta base se obtuvieron un total de 2,469 IDs o municipios y además de la variable "población total" se utilizaron 10 variables socioeconómicas que se considera pueden ser de importancia y significativas para la evolución de una persona que se contagie de COVID–19.

Variables socioeconómicas a nivel municipal generadas o seleccionadas	
Porcentaje de población de 60 años y más	Promedio de hijos nacidos vivos
Grado promedio de estudios	Porcentaje de población económicamente activa
Porcentaje de población ocupada	Porcentaje de población analfabeta
Porcentaje de población que habla lengua indígena	Promedio de ocupantes por vivienda
Porcentaje de población sin seguridad social	Porcentaje de población masculina

Es importante recalcar las variables generadas se promediaron (o dividieron) considerando el total de la población que se necesitaba; por ejemplo para la variable de la PEA se tomó el total personas con 12 años o más, en cambio para la variable de la *Población sin afiliación a servicios de salud* se tomó en cuenta la población total.

Además es preciso notar que la variable de *"población total"* es muy heterogénea, evidentemente existen municipios a lo largo de toda la República Mexicana muy distintos entre ellos en cuanto a población se refiere, desde los que tienen una población ínfima de 81 habitantes solamente, hasta los municipios de las grandes ciudades que llegan a tener más de 1 millón de habitantes; por lo que será necesario tomar esto en cuenta para el análisis estadístico que se hará más adelante.

Por consiguiente se obtiene una base de datos como la siguiente:

ID	**POBTOT**	**HNV**	**GPEST**	**PEA**	**ANALF**	**P-HLI**	$\cdots$
01001	948990	1.98	10.84	64.29	01.64	0.20	$\cdots$
01002	51536	2.62	8.54	55.83	03.52	0.04	
01003	58250	2.67	8.05	58.01	04.48	0.13	
01004	17000	2.51	9.08	54.68	03.13	0.04	
01005	129929	2.11	10.22	66.33	02.37	0.12	
01006	47646	2.37	9.77	60.95	02.72	0.11	
01007	57369	2.46	9.60	62.18	03.44	0.39	
01008	9552	2.66	9.24	54.24	02.43	0.23	
01009	22485	2.65	8.56	54.42	04.23	00.08	
01010	20853	2.62	8.49	54.68	03.88	0.06	
01011	61997	2.16	9.21	64.82	02.35	0.20	
02001	443807	1.93	10.34	64.10	0.02	02.76	
02002	1049792	1.95	10.53	63.33	01.61	0.58	
$\vdots$							$\ddots$

Tabla 3.2: Muestra de los datos trabajos del CENSO como ejemplo con las primeras 6 variables de la base de trabajo además del ID.

3.4. Otras bases de datos

Se trabajó también con datos socioeconómicos y de prevalencia de las principales comorbilidades en México de 2 bases más de INEGI: la base del Índice y grado de Marginación por Municipio 2020 (IMM) [13] y la de Prevalencia de Obesidad, Hipertensión y Diabetes para los Municipios de México (2018) [10].

Del IMM se obtuvieron las variables socioeconómicas que determinan el grado de marginación de una población de acuerdo a sus condiciones sociales de salud, vivienda y educativas, mientras que la segunda base nos sirvió para encontrar los municipios con más prevalencia de obesidad, diabetes e hipertensión en el país para ver la relación con los datos de la pandemia y si realmente las comorbilidades impactan de manera significativa a las defunciones o no.

3.4.1. Índice y grado de marginación por municipio 2020

La marginación es un fenómeno multidimensional que considera la exclusión de la población al proceso de desarrollo; producto de la desigualdad en la distribución del progreso que excluye a personas, grupos sociales y/o territorios.

El Índice de Marginación (IM) publicado por la SGCONAPO – Secretaría General del Consejo Nacional de Población – muestra las carencias o déficit en Educación, Vivienda, Ingresos monetarios y en la Distribución de la población en México. Por ejemplo, las personas de 15 años que no cuentan con educación básica enfrentan serias desventajas para desenvolverse en la sociedad; las personas que carecen de servicios sanitarios en su vivienda tienen mayores riesgos de contraer enfermedades, pudiendo convertirse en un problema de salud pública; situaciones como las anteriores se refuerzan en las localidades pequeñas (menores de 5 mil habitantes). Por lo tanto, las variables en esta base son de suma importancia para ver su impacto en el riesgo de mortalidad por COVID–19.

La base contiene los indicadores socioeconómicos que sirvieron para la construcción de los índices de marginación en los diferentes niveles de desagregación geográfica censal posible; en nuestro caso se utilizaron los datos a nivel municipal.

ID	SBASC	SDE	SEE	SAE	VPT	VHAC	⋯
01001	20.37	0.10	0.11	0.38	0.59	10.34	
01002	33.91	2.65	0.49	0.86	1.35	22.94	
01003	42.48	0.37	0.52	0.80	1.04	19.22	
01004	27.70	0.71	0.58	0.66	1.03	22.72	
01005	26.69	0.28	0.35	0.86	1.31	16.40	
01006	25.72	0.54	0.43	1.04	0.71	19.08	
01007	27.09	0.97	0.50	1.03	1.30	20.98	
01008	28.42	2.11	0.92	1.71	0.86	21.57	
01009	35.09	1.39	0.59	1.09	1.43	22.99	
01010	34.76	2.32	1.03	1.78	1.45	21.82	
⋮							⋱

Tabla 3.3: Muestra de la base de datos del IMM 2020 con la que se trabajó.

3.4.2. Prevalencia de obesidad, hipertensión y diabetes para los municipios de México 2018

En México la obesidad es considerada uno de los principales problemas de salud pública, se ha demostrado con el paso de los años que algunas de las comorbilidades asociadas con la obesidad, como la diabetes e hipertensión, contribuyen a un gran porcentaje de mortalidad, discapacidad y muerte prematura en la población. Encontrándose siempre en los primeros 3 lugares de las causas de muerte en nuestro país. [14]

Debido a esto, es vital analizar su influencia en las muertes por COVID–19 y ver si dentro de la población obesa, con diabetes o hipertensión hubo más afectados realmente por padecer esas enfermedades o solo los que llegaban a contagiarse corrieron más riesgo por tener comorbilidades y ser un factor de riesgo.

En un estudio realizado por investigadores de la Universidad Autónoma Metropolitana sobre la probabilidad de supervivencia a la COVID–19 [18] dependiendo el tipo de paciente: el tipo 1 es un paciente promedio, es decir, sin enfermedades graves; el tipo 2 es un paciente con las comorbilidades más comunes en México, y el tipo 3 es un paciente con todas las comorbilidades posibles; encontraron que a partir de su curva de supervivencia se aprecia que en los primeros 14 días no existe una diferencia significativa entre el paciente tipo 2 y un paciente promedio. Sin embargo, el daño aletargado que ocasiona el virus hace

que la probabilidad de supervivencia disminuya considerablemente entre los días 14 y 30, hasta llegar a una probabilidad menor a 10 %. El paciente tipo 3, al igual que el tipo 2 muestra una alta probabilidad de supervivencia en los primeros días después de presentar síntomas; sin embargo, ésta se reduce verticalmente a partir del día 10, llegando al 20 con una probabilidad de apenas 25 %. Desafortunadamente, los pacientes tipo 3 no sobreviven más allá de 25 días, momento en que su probabilidad de supervivencia se censura.

Así que esta base de datos estima la proporción de la población de 20 años y más de edad que padecen enfermedades de Obesidad, Hipertensión y Diabetes para los municipios de México, mediante técnicas de Estimación para Áreas Pequeñas (EAP) utilizando la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2018, y cuyos resultados son confiables a nivel de entidad federativa y del país en su conjunto. De esta forma podemos analizar la influencia de estas variables al riesgo de fallecer por COVID–19.

ID	**POBES**	**PHIPER**	**PDIAB**
01001	31.48	14.94	14.94
01002	32.28	15.32	15.32
01003	40	13.75	13.75
01004	32.59	16.43	16.43
01005	34.73	12.35	12.35
01006	24.05	12.84	12.84
01007	31.74	14.33	14.33
01008	31.30	14.91	14.91
01009	31.95	14.69	14.69
01010	35.05	16.55	16.55
01011	39.3	11.87	11.87.
02001	50.17	22.24	22.24
02002	57.78	25.08	25.08
02003	56.12	20.07	20.07
⋮			

Tabla 3.4: Muestra de la base de datos de las principales comorbilidades en México con la que se trabajó.

Capítulo 4

Análisis de regresión: Pandemia por COVID-19 en México

En este Capítulo se realiza un análisis por grupos poblacionales en las 3 primeras olas de la pandemia, nuestros datos se dividirán utilizando como referencia una variable adicional que será: "FECHASINTOMAS" en la cual se identifica la fecha en que inició la sintomatología del paciente y se usa para determinar cada ola de la pandemia.

4.1. Análisis exploratorio

El objetivo de este tipo de análisis, como su nombre lo dice, es explorar e indagar en nuestra base de datos, conocer bien las variables con las que trabajaremos y encontrar patrones o relaciones importantes entre ellas con la finalidad de poder modelar el fenómeno que se desea estudiar, en este caso la pandemia por el virus SARS–CoV–2.

Para dicho análisis se utilizaron $2,624,272$ observaciones (casos positivos) que se resumieron en $2,437$ observaciones (municipios con casos positivos), 30 variables regresoras y 1 variables respuesta. De esta base se obtuvo una representación gráfica de la dinámica de la pandemia desde su comienzo hasta el fin de la tercera ola, considerando únicamente 3 meses de cada ola de contagios en el país (el mes del pico de contagios, el anterior y el posterior). Con el objetivo de ver cuáles fueron los estados más afectados; considerando

características individuales de las personas que se contagiaron con el virus SARS–CoV–2 que han presentado síntomas, han sido hospitalizados y/o han fallecido.

Primera ola: junio 2020, julio 2020 y agosto 2020.

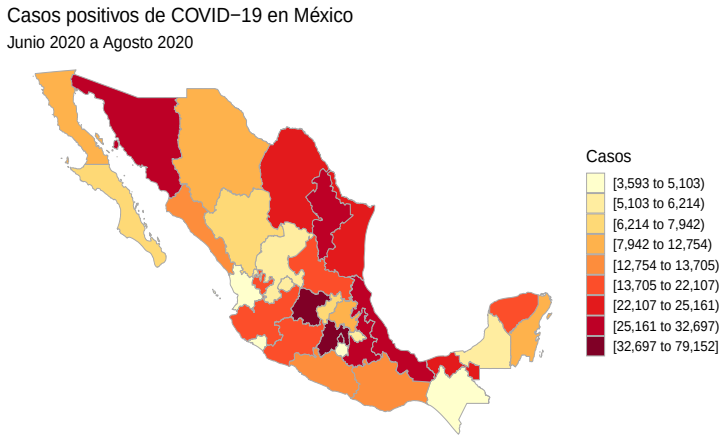
Segunda ola: diciembre de 2020, enero 2021 y febrero 2021.

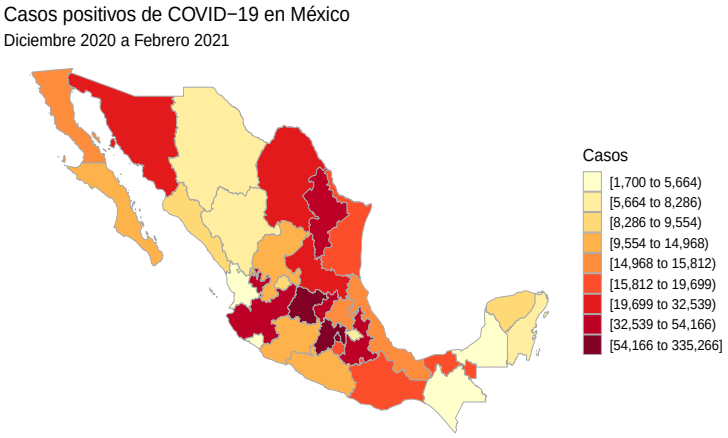
Tercera ola: julio 2021, agosto 2021 y septiembre 2021.

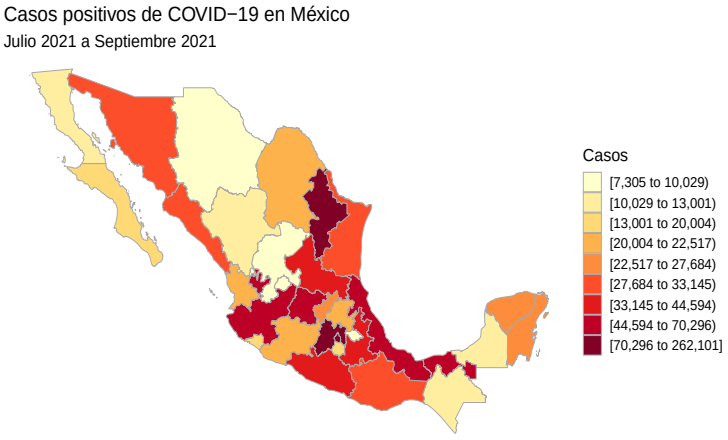
En este primer análisis se trabajó con las 32 entidades federativas del país ya que la representación gráfica es más amigable y se logra un mejor panorama de la dinámica de nuestras variables. A pesar de que se tienen datos a nivel municipal y se hicieron los respectivos análisis y gráficas, los mapas no resultaban comprensibles a primera vista debido a la gran cantidad de municipios que conforman la República Mexicana, $2,471$ para ser exactos. Por lo tanto, para fines prácticos el análisis gráfico principal se puso a nivel estado.

Mediante mapas coropléticos se puede ver como fueron evolucionando los contagios, las hospitalizaciones y las defunciones a lo largo de la República Mexicana en cada ola de la pandemia; encontrando los estados más susceptibles al virus del SARS-CoV-2 al ver cómo fue la evolución de los focos de contagios en el país y después compararlos con los estados con el mayor porcentaje de hospitalizaciones y defunciones.

Se puede observar que hubo un mayor contagio en algunos estados del norte como Sonora, Coahuila, Nuevo León y Tamaulipas; la zona metropolitana del Valle de México y otros como Jalisco, Michoacán, Puebla, Veracruz y Tabasco en los que hubo continuamente más de 30 mil contagios, llegando a tener hasta 300 mil casos positivos en algunos estados.

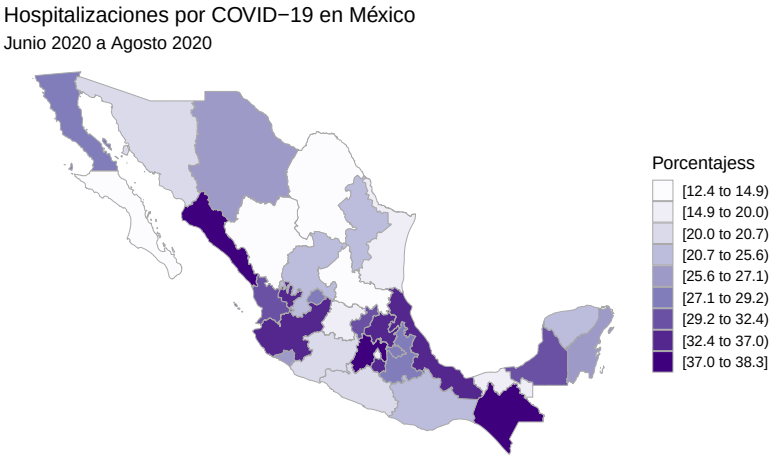


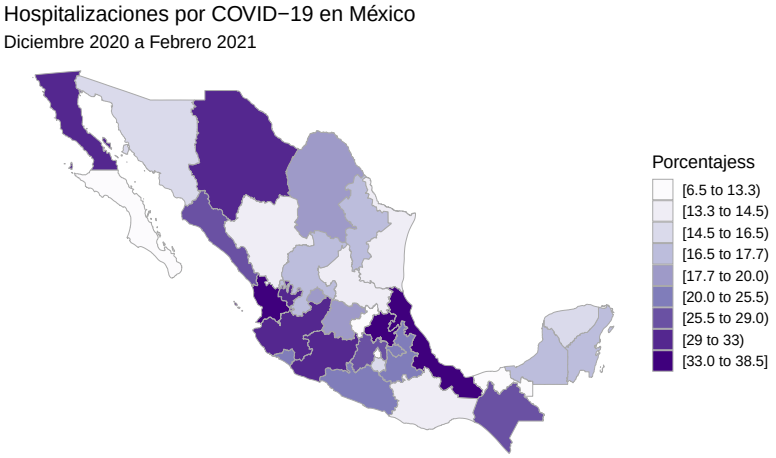


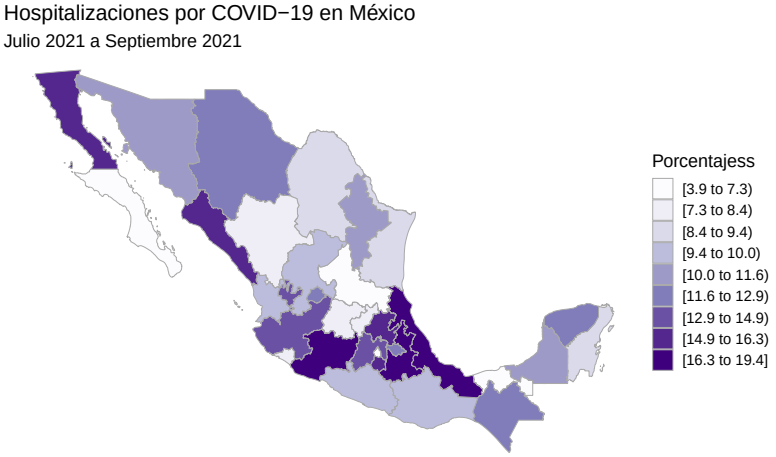


En el caso de las hospitalizaciones podemos ver que la concentración principal se encuentra en los estados de Sinaloa, Jalisco, Estado de México, Ciudad de México, Hidalgo,

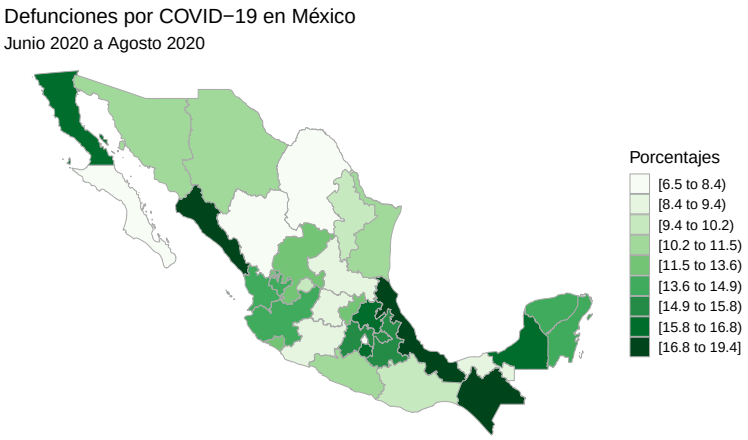
Veracruz, Chiapas, Nayarit, Chihuahua, Nuevo León, Baja California, Campeche y Puebla; y esto se mantiene a diferente escala en las distintas olas de contagios de la pandemia.

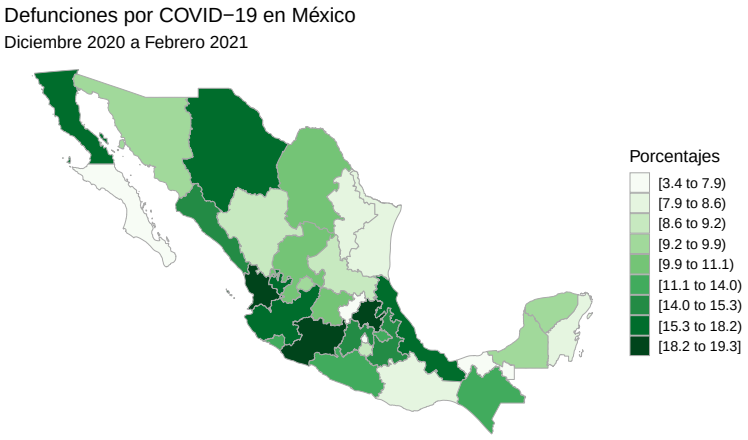


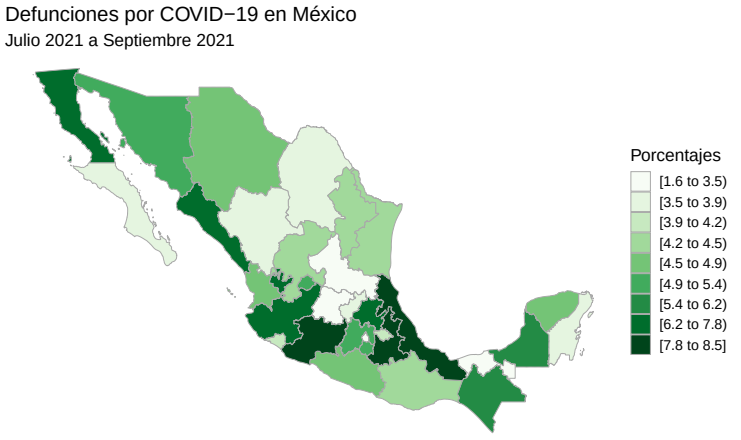




Y finalmente, en el caso de las defunciones los estados que sufrieron más pérdidas humanas fueron Sinaloa, Baja California, Veracruz, Chiapas, Jalisco, Nayarit, Hidalgo, Puebla y Campeche.







De los mapas anteriores se concluye que los estados donde es clara una relación entre los contagios, las hospitalizaciones y las defunciones fueron Veracruz, Jalisco, Hidalgo y Puebla; mientras que los que tuvieron defunciones dado que tuvieron muchas hospitalizaciones fueron Sinaloa, Baja California, Veracruz, Chiapas, Jalisco, Nayarit, Puebla, Hidalgo y Campeche.

Posteriormente se utilizó la información del INEGI de las enfermedades prevalentes en México para conocer cómo está distribuida la obesidad, la diabetes y la hipertensión a lo largo del territorio mexicano e identificar cuáles son los estados con más prevalencia de estas comorbilidades para hacer futuras correlaciones con los datos de la pandemia.

Se puede observar en la figura 4.1 y en el mapa 4.2 que los estados con menores proporciones de obesidad se encuentran en la región suroeste y algunos estados del cento norte como Querétaro, Guanajuato y San Luis Potosí. Mientras que los estados con mayor índice de obesidad (por arriba del 50 % del total de su población) son 17 de los 32 estados en el país, los cuáles son Baja California, Baja California Sur, Campeche, Chiapas, Chihuahua, Coahuila, Colima, Michoacán, Morelos, Nayarit, Quintana Roo, Sinaloa, Sonora, Tabasco, Tamaulipas, Veracruz y Yucatán; y de éstos, 8 tienen de igual forma mayor índice de diabetes e hipertensión que como se puede observar en la figura 4.3 y el mapa 4.4 se observa que estas comorbilidades se manifiestan con valores altos en zonas del norte y noreste del país, particularmente en los estados de Baja California, Sonora, Chihuahua, Coahuila, Durango y Nuevo León; y también toda la costa del Golfo de la República Mexicana hasta Campeche. Adicionalmente, en la parte media baja de la República; en estados como Michoacán, Guanajuato, Hidalgo y la Ciudad de México se tienen valores altos e intermedios de estas enfermedades. Y los estados con un índice bajo de diabetes e hipertensión son Sinaloa, Jalisco, Aguascalientes, Puebla, Tlaxcala, Chiapas y Quintana Roo. De los cuales es curioso observar que los estados de Sinaloa y Chiapas tienen un alto índice de obesidad pero no de hipertensión o diabetes. Sin embargo fueron estados que presentaron un porcentaje muy alto de hospitalizaciones y defunciones.

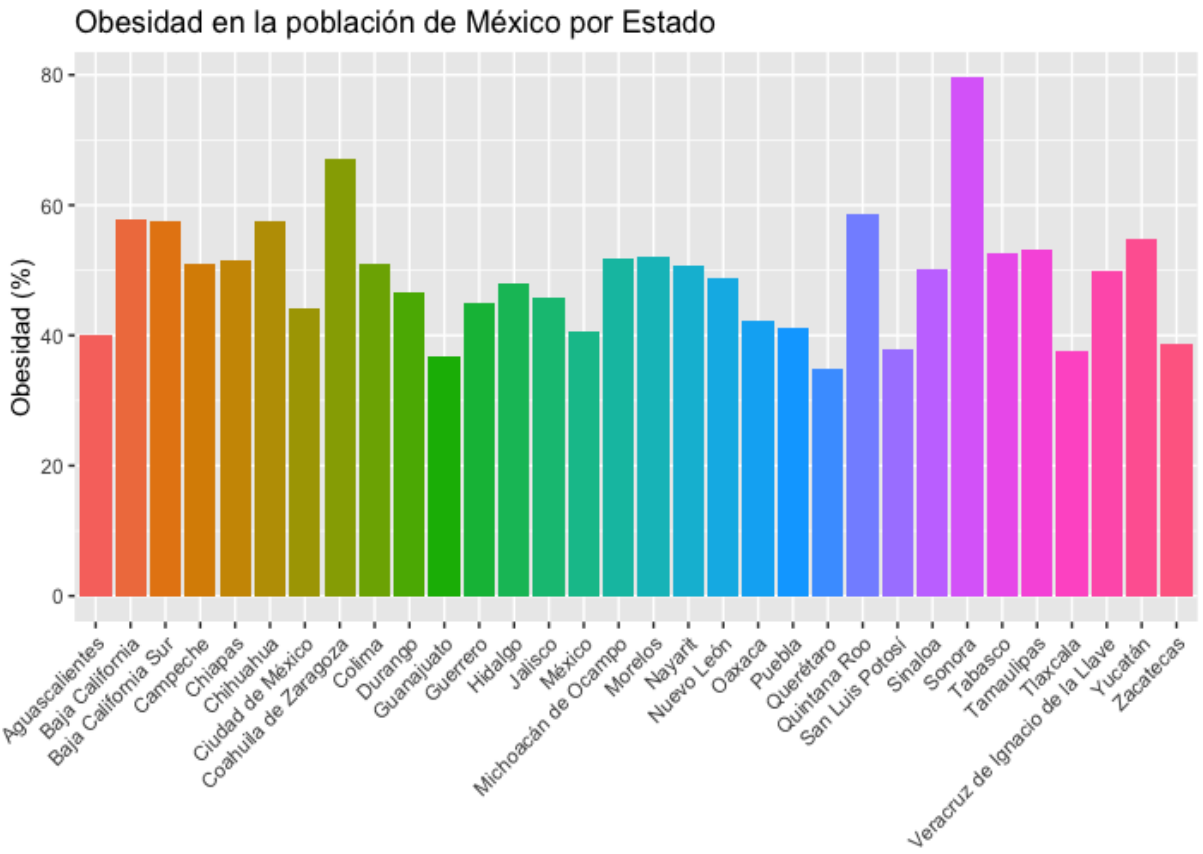


Figura 4.1: Índice de obesidad en la República Mexicana (2018).

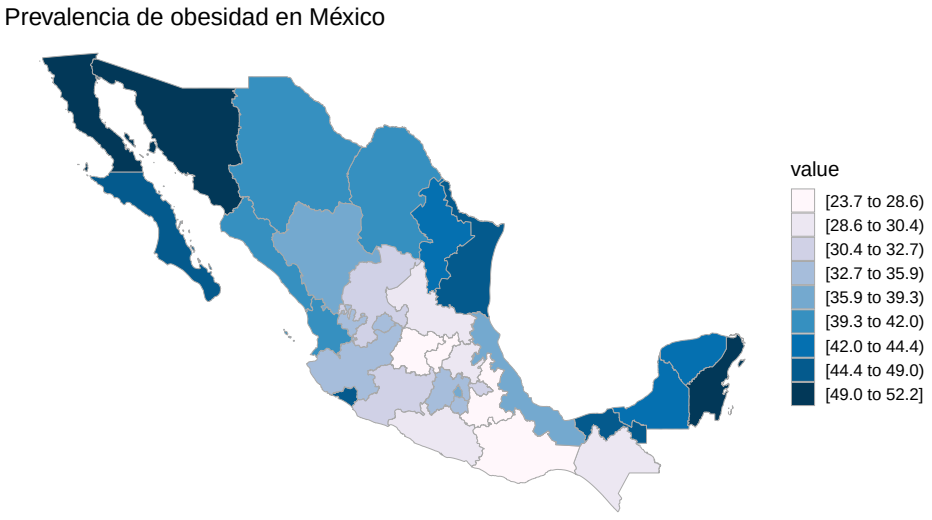


Figura 4.2: Distribución de la prevalencia de obesidad por estado a lo largo de la República Mexicana (2018).

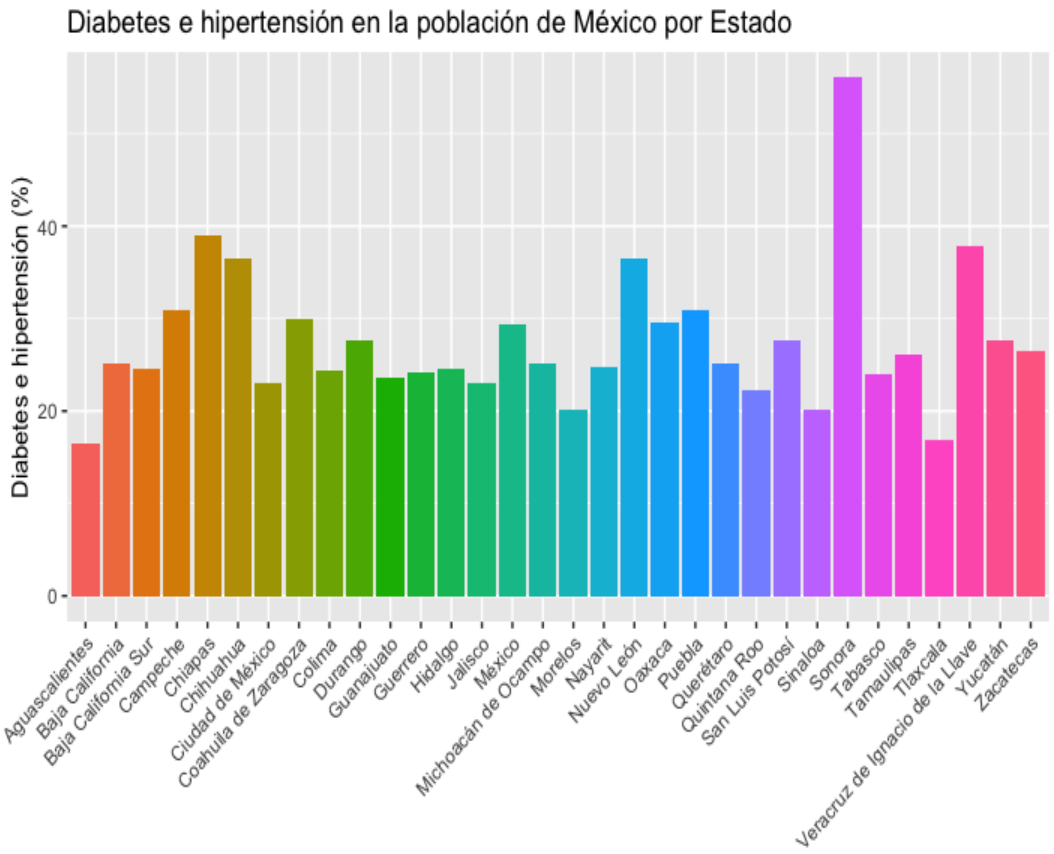


Figura 4.3: Índice de diabetes e hipertensión en la República Mexicana (2018).

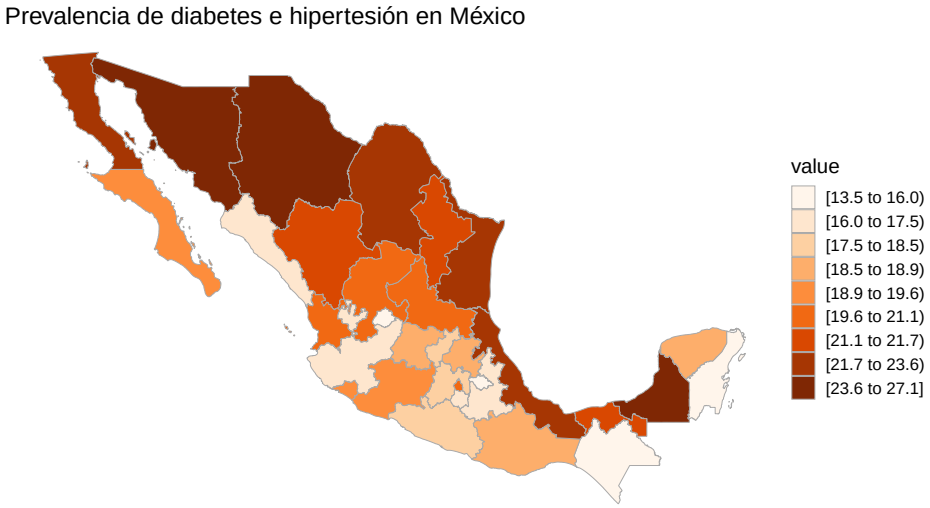


Figura 4.4: Distribución de la prevalencia de diabetes e hipertensión por estado a lo largo de la República Mexicana (2018).

Estos datos son del 2018, sin embargo, sabemos que la prevalencia de sobrepeso y obesidad ha incrementado a nivel mundial en las últimas tres décadas y desde el año 2000 se ha registrado un incremento gradual pero sostenido en el transcurso de los años. [1] Por lo que podemos considerar estos datos factibles para nuestro propósito.

4.2. División de la población para su análisis

Como ya se mencionó antes, para tener un modelo del comportamiento de la pandemia más preciso se decidió trabajar con los municipios debido a que la República Mexicana está integrada de 2,471 municipios de los cuales 2,437 presentaron al menos algún caso positivo del virus SARS–CoV–2.

Sin embargo, el tamaño de la población de los municipios es muy variado. Fijándonos en nuestro mínimo y máximo de la variable *"población total"* podemos observar que existe un municipio solo con 81 habitantes y otro con 1,922,523 habitantes, lo que resulta ser un amplio rango como para analizarlo conjuntamente tomando las mismas consideraciones ya que no es una muestra uniforme y se podrían obtener resultados erróneos. Además, como el objetivo de este trabajo es ver la relevancia de las variables socioeconómicas no sería muy congruente tomar como iguales a los municipios pequeños, que por lo regular son zonas marginadas, a un municipio grande como lo son la mayoría de las capitales en el país. Por lo tanto, se dividió a la población total en grupos poblacionales considerando municipios con características similares, en cuánto a población se refiere y para hacerlo adecuadamente, se utilizó el rango intercuantil que es una forma de dividir los datos en un cierto número de partes en donde en cada una de ellas hay la misma cantidad (o casi la misma) de valores de la variable elegida obteniendo como resultado grupos poblacionales comparables.

4.2.1. Grupos poblacionales

Se decidió hacer el análisis por grupos poblacionales en cada ola de contagios debido a la situación tan compleja que representa la pandemia por el SARS–CoV–2 en la que es

[1] https://ensanut.insp.mx/encuestas/ensanutcontinua2020/informes.php

imprescindible tomar en cuenta el tamaño de la población, para así tener una mejor precisión y asegurar tener poblaciones muestrales aleatorias considerando además las medidas sanitarias que se fueron implementando por parte del gobierno conforme evolucionaba esta pandemia, como el aislamiento, los cierres de distintos sectores, el uso de cubrebocas, la vacunación, etc.

Para fines prácticos se dividió a la población tomando quintiles obteniendo así 5 grupos poblacionales donde cada grupo contempla la misma cantidad de habitantes y dependiendo la ola que se esté analizando contiene cierto número de municipios. A continuación se muestra como es que está distribuida la población en cada grupo:

- **Grupo 1:** Población menor o igual a 3,419 habitantes
- **Grupo 2:** Población mayor a 3,419 habitantes y menor o igual a 9,161 habitantes
- **Grupo 3:** Población mayor a 9,161 habitantes y menor o igual a 19,215 habitantes
- **Grupo 4:** Población mayor a 19,215 habitantes y menor o igual a 45,399 habitantes
- **Grupo 5:** Población mayor a 45,399 habitantes

Primera ola: Junio de 2020 a Agosto de 2020

Grupo	1	2	3	4	5	**Total**
No. de municipios	429	428	428	428	429	2142

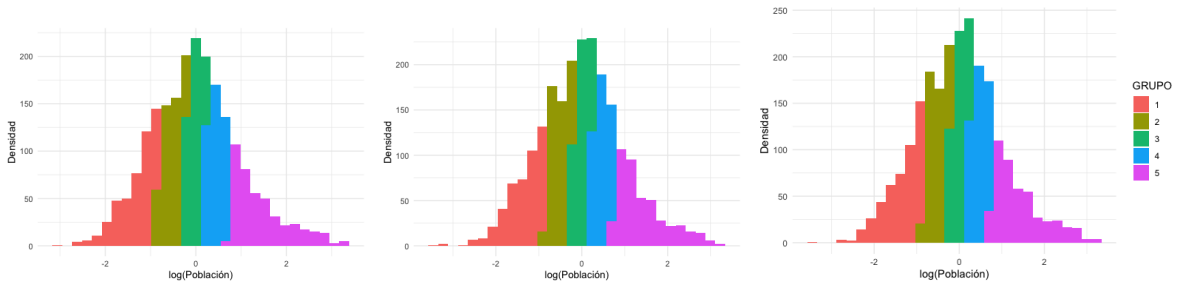
Segunda ola: Diciembre de 2020 a Febrero de 2021

Grupo	1	2	3	4	5	**Total**
No. de municipios	444	444	443	444	444	2219

Tercera ola: Julio de 2021 a Septiembre de 2021

Grupo	1	2	3	4	5	**Total**
No. de municipios	462	461	461	461	461	2306

La distribución de la población por grupos en cada ola se ve de la siguiente manera:



((a)) 1ra ola ((b)) 2da ola ((c)) 3ra ola

Figura 4.5: Distribución de la población contagiada de COVID-19 (con base logarítmica) en México en cada ola de la pandemia.

Posteriormente se realizó un análisis exploratorio de los casos positivos de COVID-19 con factores socioeconómicos obtenidos del CENSO 2020, el IMM 2020 y el ENSANUT 2018 del INEGI. Mediante el diagrama de correlación calculado solo con las variables socioeconómicas dentro de cada ola de contagios dependiendo el grupo poblacional, se puede ver la intensidad de la correlación entre los distintos grupos como se muestra a continuación. Ponemos como ejemplos los diagramas de correlación del *grupo más pequeño* (grupo 1) y del *grupo más grande* (grupo 5). Después de realizar un análisis exploratorio de nuestros datos con la información oficial disponible se lograron identificar algunas variables socioeconómicas significativas para su estudio. Para cada grupo se muestran ciertas relaciones lineales positivas y negativas con algunas variables socioeconómicas como se resume a continuación:

GRUPO 1	
Relación lineal positiva	*edad, tipo paciente, tabaquismo, población económicamente activa, Población que vive en viviendas sin agua entubada y/o drenaje, viviendas particulares con hacinamiento y población ocupada con ingresos de hasta dos salarios mínimos*
Relación lineal negativa	*Sexo, grado promedio de estudios, población que reside en viviendas sin TICs, población que vive en viviendas sin energía eléctrica y/o piso de tierra*

GRUPO 5	
Relación lineal positiva	*edad, tipo de paciente, población sin seguridad social, analfabetismo, población de 15 años o más sin educación básica, población que vive en viviendas sin energía eléctrica, agua entubada y/o piso de tierra, viviendas particulares con hacinamiento, población ocupada con ingresos de hasta dos salarios mínimos*
Relación lineal negativa	*Sexo, tabaquismo, grado promedio de estudios, obesidad, diabetes*

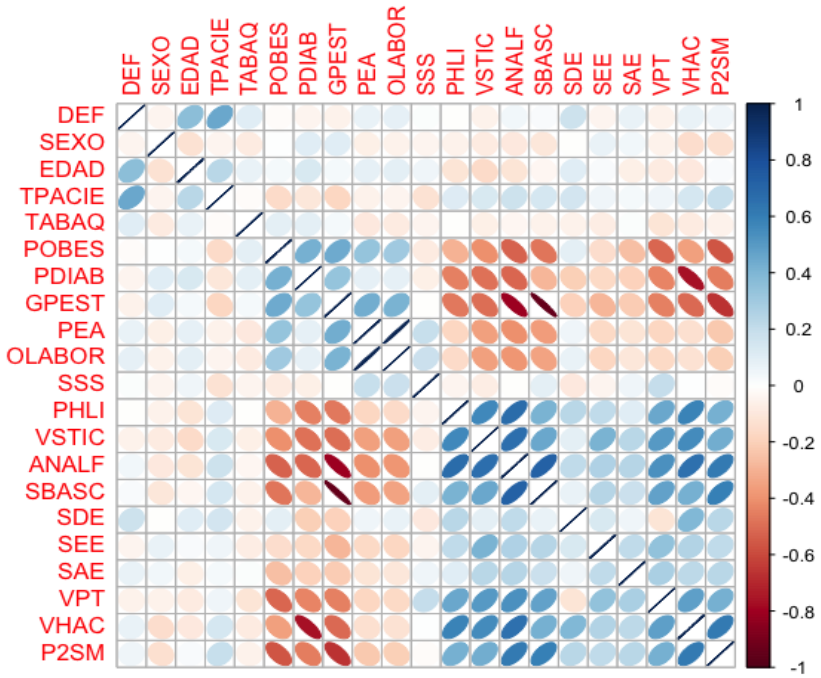


Figura 4.6: Correlación de las defunciones por COVID-19 entre distintos factores socioeconómicos del país calculado para el grupo 1 en la primera ola de contagios.

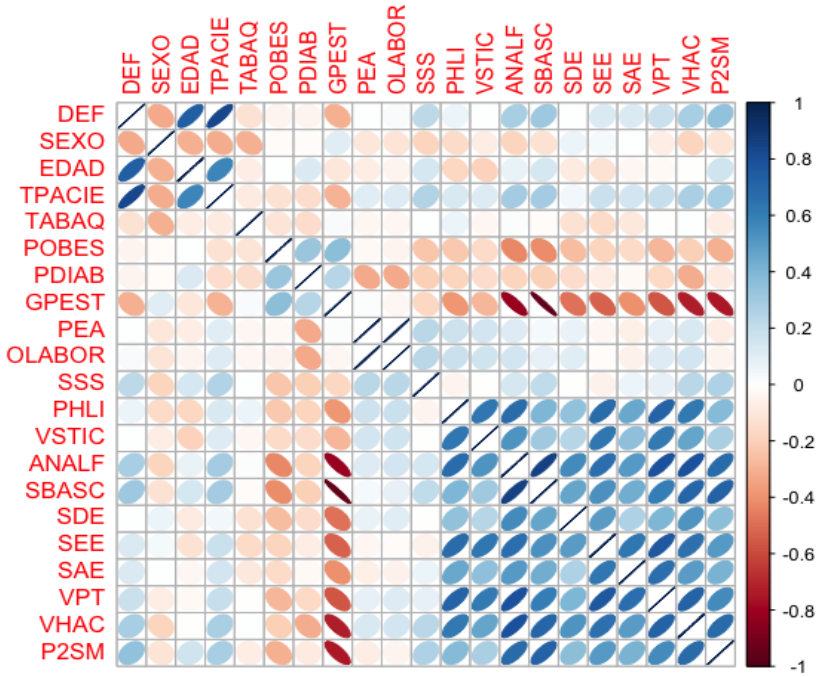
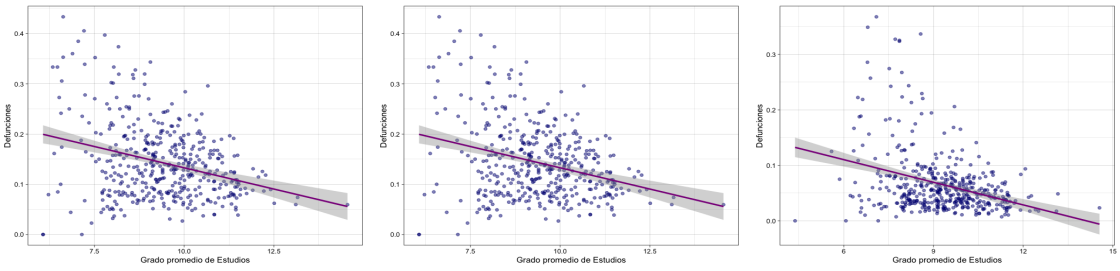


Figura 4.7: Correlación de las defunciones por COVID-19 entre distintos factores socioeconómicos del país calculado para el grupo 5 en la primera ola de contagios.

De las figuras 4.6 y 4.7 se observa que existe una colinealidad casi perfecta entre la variable "*PEA*" que representa el porcentaje de la población de 12 años y más económicamente activa y la variable "*OLABOR*" que representa a su vez el porcentaje de la población de 12 años y más ocupada; y del mismo modo sucede con "*SBASCS*" que es el porcentaje de la población de 15 años o más sin educación básica y "*GPEST*" que representa el grado promedio de escolaridad de las personas de 15 a 130 años de edad, en las que se marca una relación lineal muy fuerte, por lo que se decidió retirar una de las dos variables (en ambos casos) para hacer nuestro modelo adecuadamente y evitar posibles errores de multicolinealidad afectando nuestros resultados.

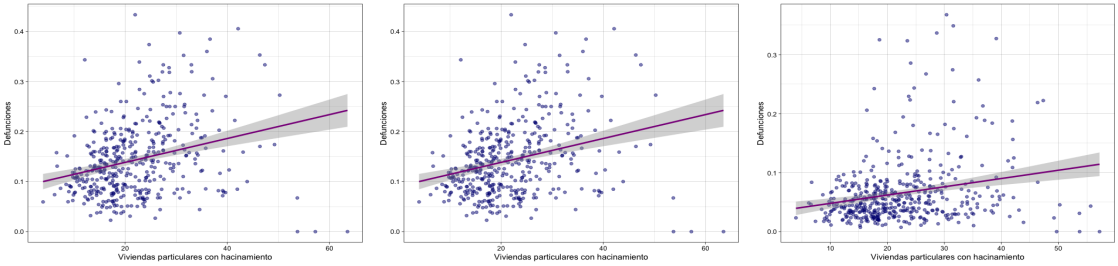
Nuestro análisis se hizo con datos de la República Mexicana separándolos por municipios y como ya se mencionó, se tienen marcadas diferencias tanto sociales, económicas y de salubridad en ellos que se deben tomar en cuenta. Se cuestionó entonces si cada una de esas variables estaban realmente relacionados con las defunciones por la pandemia de COVID-19 y si era así que tan significativos eran dentro del modelo y cuáles variaron conforme fue avanzando la pandemia.

Mediante diagramas de dispersión entre las defunciones y las variables socioeconómicas que parecen ser más relevantes al observar el diagrama de correlaciones del *grupo 5*, el cual se decide tomar como referencia por ser el de mayor tamaño poblacional, y ajustando un modelo lineal general que se grafica solamente para ver si existe realmente una relación lineal o no; se encontró lo siguiente:



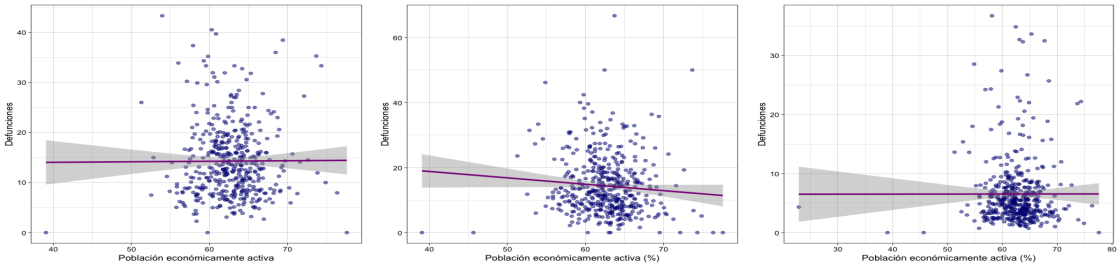
((a)) 1ra ola ((b)) 2da ola ((c)) 3ra ola

Figura 4.8: Relación de las defunciones por COVID–19 con el grado promedio de estudios.



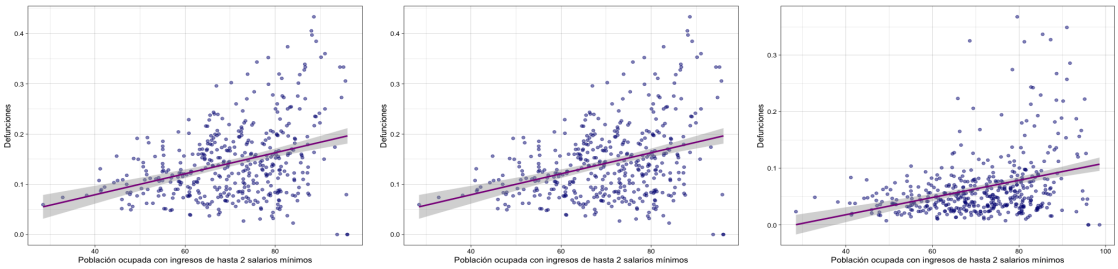
((a)) 1ra ola ((b)) 2da ola ((c)) 3ra ola

Figura 4.9: Relación de las defunciones por COVID–19 y la población que vive en viviendas con hacinamiento.



((a)) 1ra ola ((b)) 2da ola ((c)) 3ra ola

Figura 4.10: Relación de las muertes por COVID–19 y la población económicamente activa.



((a)) 1ra ola ((b)) 2da ola ((c)) 3ra ola

Figura 4.11: Relación de las defunciones por COVID–19 y la población ocupada con ingresos de hasta 2 salarios mínimos.

Podemos concluir a grandes rasgos que el riesgo de mortalidad por COVID–19 aumenta si las personas viven en viviendas particulares con hacinamiento o si tienen un ingreso igual o menor a 2 salarios mínimos; mientras que el riesgo de mortalidad disminuye conforme aumenta el grado promedio de estudios. Y por otro lado, al parecer la población económicamente activa no está linealmente relacionada con las defunciones, sin embargo, es claro que sí existe una influencia en cuanto a las defunciones que hubo por COVID–19

en cierto rango de municipios donde tienen $\sim 65\,\%$ de la población económicamente activa pero este riesgo disminuyó drásticamente en la tercera ola.

4.3. Ajuste del modelo

La identificación o ajuste de un modelo de regresión se refiere básicamente a obtener las variables que mejor describan los datos y modelen el riesgo de mortalidad por COVID-19, obteniendo además los coeficientes que resuelven el modelo ajustado, es decir el vector $\hat{\boldsymbol{\beta}}$.

A continuación resumiré los resultados de los modelos obtenidos con "R"; pero antes que nada hay que considerar lo que ya se sabe de la pandemia, las variables: edad, tipo de paciente, neumonía, asma, inmunosupresión, enfermedad cardiovascular, insuficiencia renal crónica y tabaquismo son sumamente significativas y prácticamente determinan la evolución del paciente como se ve reflejado en los casos que hubo en cada estado y municipio del país y como se confirmó en nuestro modelo tomando todas nuestras variables, las de la pandemia y las socioeconómicas. En particular lo podemos comprobar con el modelo de la ola 1 con los 5 grupos poblacionales que se ve en la tabla 4.1, además se obtienen resultados similares en la segunda y tercer ola de la pandemia.

En este trabajo buscamos la relación con las variables socioeconómicas, por lo que se omiten las variables significativas antes mencionadas con el fin de quedarnos solo con las de nuestro objeto de estudio debido a que está clara la influencia de las otras variables pero no conocemos la influencia de los factores socioeconómicos, por lo tanto para hacer el modelo en cuestión que nos interesa se toman las defunciones como nuestra variable respuesta y a todas nuestras variables socioeconómicas como nuestras variables regresoras obtenidas del CENSO, del IMM y además las 3 enfermedades prevalentes en México que son la hipertensión, diabetes y obesidad. Finalizando con los resultados que se muestran en las tablas 4.2, 4.3 y 4.4.

Se ajustaron modelos para los tres primeros picos de la pandemia, considerando los cinco grupos de municipios definidos conforme a su población. No obstante, únicamente se presentan los coeficientes de regresión para los mejores modelos de acuerdo al criterio

Tabla 4.1: Comparación de modelos contemplando todas las variables: 1ra ola

	Variable dependiente:				
	Defunciones				
	(1)	(2)	(3)	(4)	(5)
SEXO				0.051 (0.034)	
EDAD	0.397*** (0.075)	0.322*** (0.087)	0.943*** (0.090)	0.470*** (0.098)	0.658*** (0.059)
TPACIE	0.106*** (0.036)	0.110*** (0.041)	0.079** (0.039)	0.178*** (0.031)	0.210*** (0.017)
NEUMO	0.300*** (0.039)	0.328*** (0.046)	0.267*** (0.042)	0.248*** (0.033)	0.079*** (0.021)
ASMA	−0.180* (0.099)	0.273 (0.169)			
DIAB			−0.072 (0.046)	−0.168*** (0.045)	0.075* (0.040)
INMUSU			−0.201 (0.144)		
CARDIO			−0.241** (0.099)		0.163* (0.096)
EPOC			0.271*** (0.071)	−0.392*** (0.112)	−0.224** (0.096)
RENALC	0.168** (0.084)		0.183* (0.105)		
TABAQ	0.172** (0.086)		−0.240*** (0.073)		−0.058** (0.028)
SDE	0.526** (0.257)				−0.128* (0.068)
HIPER		0.145*** (0.042)		0.196*** (0.042)	
OBES		−0.129*** (0.041)			
POBES		0.219* (0.123)	0.272*** (0.105)		
PDIAB		0.471 (0.315)			
PHNV		−5.281* (2.800)			−2.446** (1.235)
POV		4.926 (3.017)	4.429*** (1.614)		
SEE	−2.046*** (0.745)	−0.557** (0.233)		−0.385*** (0.115)	
SAE	0.251* (0.131)		0.140** (0.063)		
GPEST					−0.665** (0.283)
PEA					−0.104** (0.044)
SSS					−0.032 (0.021)
VHAC	0.205 (0.131)	0.282** (0.126)			0.107*** (0.030)
VPT			−0.172* (0.088)		
P2SM	−0.238* (0.129)		0.155** (0.069)	0.067* (0.040)	0.036* (0.021)
P5000		0.039 (0.025)			−0.017* (0.010)
Constant	0.631 (9.708)	−42.481** (17.093)	−73.091*** (10.058)	−25.513*** (5.793)	−9.977 (6.409)
Observations	429	428	428	428	429
R^2	0.405	0.501	0.578	0.635	0.836
Adjusted R^2	0.389	0.485	0.564	0.627	0.830
Residual Std. Error	21.198	16.257	11.656	7.870	3.050
F Statistic	25.786***	31.945***	40.486***	80.787***	140.266***

Note: *p<0.1; **p<0.05; ***p<0.01

de información de Akaike (AIC) (con una selección de variables automatizada paso a paso en ambas direcciones), puesto que nos interesa obtener un modelo para hacer predicciones además de interpretar y este es un método de selección de variables más adecuado ya que nos proporciona un mayor número de variables; mientras que el criterio de información de Bayes (BIC) tiene más penalizaciones para seleccionar las variables, haciéndolo óptimo solamente cuando se busca obtener un modelo fácil e interpretable.

Se muestran los coeficientes de Regresión Lineal Múltiple para 15 modelos (5 por cada pico de la pandemia) con el objetivo de describir el riesgo de fallecimiento por COVID-19 en cada uno de los 2,437 municipios de México donde se registró al menos un caso confirmado del virus SARS-CoV-2. Además se puede observar el nivel de significancia de cada uno junto con sus respectivos R^2 para ver cuánto porcentaje de los datos se modelan considerando estas variables.

Tabla 4.2: Comparación de modelos con variables socioeconómicas: 1a ola

	Variable dependiente:				
	Defunciones				
	(1)	(2)	(3)	(4)	(5)
PDIAB	−0.692* (0.375)				
PEA	0.306* (0.180)			−0.179 (0.120)	
SSS				0.110* (0.065)	0.094** (0.042)
P60YM	0.899** (0.379)			0.588*** (0.207)	
POBES		0.303* (0.160)	0.677*** (0.183)		
PHNV		−7.612* (4.096)			
GPEST		−2.757** (1.302)	−3.849*** (1.340)		
PHLI					−0.104*** (0.034)
SDE	1.126*** (0.297)	0.695*** (0.224)			−0.531*** (0.132)
SEE	−1.575* (0.925)	−1.340*** (0.383)		−0.370* (0.191)	
VHAC					0.223*** (0.072)
ANALF			−0.605*** (0.221)		0.491*** (0.150)
POBMAS			−2.197** (1.061)		
SAE	0.293* (0.161)		0.133 (0.084)		
POV					−3.862*** (1.166)
P2SM		0.348** (0.137)	0.343*** (0.111)	0.204*** (0.064)	0.116*** (0.039)
Constant	−0.634 (10.633)	21.505 (26.381)	111.978** (55.307)	1.721 (9.905)	11.755*** (4.120)
Observations	429	428	428	428	429
R^2	0.060	0.089	0.084	0.054	0.214
Adjusted R^2	0.047	0.076	0.071	0.043	0.201
Residual Std. Error	26.479	21.779	17.019	12.608	6.613
F Statistic	4.506***	6.847***	6.435***	4.837***	16.365***

Note: *p<0.1; **p<0.05; ***p<0.01

Tabla 4.3: Comparación de modelos con variables socioeconómicas: 2da ola

	Variable dependiente:				
	Defunciones				
	(1)	(2)	(3)	(4)	(5)
POBES	0.373** (0.158)	0.247* (0.148)	0.345* (0.185)	0.330*** (0.114)	
PHNV	10.915** (4.571)				12.035*** (2.228)
VHAC		−0.245* (0.134)			−0.224*** (0.072)
GPEST	8.816*** (2.547)	−3.906*** (1.097)	−4.129*** (1.302)		
SSS	−0.228* (0.125)				0.340*** (0.049)
PEA			0.234* (0.130)		
ANALF	1.271*** (0.346)		−0.440* (0.232)		
PHLI				−0.093*** (0.031)	
POV	−7.590** (3.607)			7.875*** (2.616)	−2.501* (1.454)
SDE					−0.425*** (0.150)
SAE	−0.388** (0.169)				0.124* (0.069)
P2SM	0.526** (0.217)				
P5000				0.078*** (0.024)	0.053** (0.025)
P60YM			0.420* (0.254)	1.269*** (0.305)	
POBMAS			−1.839* (1.073)		
Constant	−111.784*** (42.096)	50.139*** (10.600)	117.604** (56.096)	−41.898*** (14.969)	−8.235 (5.528)
Observations	444	444	443	444	444
R^2	0.072	0.029	0.047	0.093	0.214
Adjusted R^2	0.055	0.023	0.034	0.083	0.201
Residual Std. Error	29.452	23.999	17.433	12.719	8.196
F Statistic	4.213***	4.408***	3.594***	9.005***	16.939***

Note: *p<0.1; **p<0.05; ***p<0.01

Tabla 4.4: Comparación de modelos con variables socioeconómicas: 3ra ola

	Variable dependiente:				
	Defunciones				
	(1)	(2)	(3)	(4)	(5)
POBES	−0.400*** (0.141)			0.319*** (0.101)	
PDIAB	0.684* (0.398)				
ANALF	0.526*** (0.181)		−0.632*** (0.229)		
PHNV				3.295 (2.324)	7.694*** (0.896)
POBMAS	2.310*** (0.716)			−1.610*** (0.503)	
VPT	−0.276* (0.152)			−0.121* (0.072)	
GPEST		−2.631*** (0.711)	−5.201*** (1.106)	−3.481*** (0.594)	
PEA			0.170 (0.105)	0.206** (0.082)	
SSS		0.218*** (0.067)	0.299*** (0.071)	0.240*** (0.045)	0.184*** (0.027)
P60YM				0.369 (0.234)	0.139* (0.076)
POV		3.762* (1.971)		3.522* (1.899)	
SAE		−0.209*** (0.079)			0.090** (0.036)
SDE				−0.121 (0.078)	−0.371*** (0.073)
VHAC	0.317* (0.183)	−0.253** (0.114)		−0.240*** (0.078)	
PHLI			0.056 (0.038)		
P2SM		0.162* (0.093)	0.154* (0.093)		
P5000				0.079*** (0.020)	
Constant	−115.367*** (34.148)	6.903 (11.829)	29.754* (15.378)	65.262*** (25.185)	−16.417*** (2.180)
Observations	462	461	461	461	461
R^2	0.064	0.082	0.141	0.274	0.253
Adjusted R^2	0.052	0.070	0.129	0.254	0.245
Residual Std. Error	22.685	14.747	14.246	8.737	4.864
F Statistic	5.186***	6.789***	12.378***	14.086***	30.866***

Note: *p<0.1; **p<0.05; ***p<0.01

En la tabla 4.5 se muestran las variables socioeconómicas significativas con sus respectivos coeficientes de regresión estimados del modelo de regresión lineal múltiple con el objetivo de observar con mejor claridad los cambios que hubo en cada pico de la pandemia y qué variables fueron explicativas en cada uno.

Asimismo, recordemos que cada uno de los coeficientes de regresión estimados de las variables predictoras son las pendientes del modelo de regresión lineal múltiple. Y su interpretación sería la siguiente:

Si fijamos una variable predictora (cualquiera que sea de nuestro interés) y el resto de las variables se mantienen constantes, entonces la variable respuesta ($\mathbf{Y}$) varía en promedio tantas unidades como indica la pendiente de la variable predictora.

Veamos un par de ejemplos, en el caso de la primera ola en los grupos 2 y 3 si aumentara en una unidad el *grado promedio de escolaridad* entonces las defunciones disminuirían en 2.76 % y 3.85 % respectivamente. En caso contrario, fijándonos ahora en el grupo 4 en los 3 picos, si el *porcentaje de la población de 60 años y más* aumentará en 1 % las defunciones aumentarían 0.59 % en la primera ola, 1.27 % en la segunda ola y 0.37 % en la tercera ola. Algo similar sucede en algunos grupos en las distintas olas de la pandemia si aumentara en 1 % el *porcentaje de la población ocupada con ingresos menores o iguales a 2 salarios mínimos* las defunciones aumentarían.

Por otra parte en la tercera ola, si nos fijamos en el analfabetismo y este aumenta podemos ver que en el grupo 1 las defunciones aumentarían pero en el grupo 3 disminuirían; esto podemos explicarlo analizando un poco más a fondo considerando un punto de vista más "social" ya que por lo regular en los municipios (pequeños) con poca población suele haber más desinformación y menos acceso a sistemas dignos de salud o educación, provocando un desconocimiento o ignorancia en cuestiones de salud; en municipios más grandes se puede tener acceso más fácil a la información al igual que algún tipo de ayuda para el bienestar o la seguridad social disminuyendo el riesgo de mortalidad.

VARIABLES SOCIOECONÓMICAS	PICOS DE LA PANDEMIA														
	1 Junio 2020 - Agosto 2020					**2** Diciembre 2020 - Febrero 2021					**3** Julio 2021 - Septiembre 2021				
Grado promedio de escolaridad de la población de 15 años o más		-2.76	-3.85			8.82	-3.91	-4.13				-2.63	-5.20	-3.48	
% Población económicamente activa con respecto a la población de 15 años o más	0.31			-0.18				0.23					0.17	0.21	
% Población habla lengua indígena					-0.10				-0.09				0.06		
% Población sin seguridad social				0.11	0.09	-0.23				0.34		0.22	0.30	0.24	0.18
Promedio de ocupantes por vivienda					-3.86	-7.59			7.88	-2.50		3.76		3.52	
Promedio de hijos nacidos vivos		-7.61				10.92				12.04				3.30	7.69
Analfabetismo			-0.61		0.49	1.27		-0.44			0.53		-0.63		
% Ocupantes en viviendas particulares sin drenaje ni excusado	1.13	0.70			-0.53					-0.43				-0.12	-0.37
% Ocupantes en viviendas particulares sin energía eléctrica	-1.58	-1.34		-0.37											
% Ocupantes en viviendas particulares sin agua entubada	0.29		0.13			-0.39				0.12		-0.21			0.09
% Ocupantes en viviendas particulares con piso de tierra											-0.28			-0.12	
% Viviendas particulares con hacinamiento					0.22		-0.25			-0.22	0.32	-0.25		-0.24	
% Población en localidades con menos de 5000 habitantes									0.08	0.05				0.08	
% Población de 60 años y más	0.90			0.59				0.42	1.27					0.37	0.14
% Población masculina			-2.20					-1.84			2.31			-1.61	
% Población ocupada con ingresos de hasta 2 salarios mínimos		0.35	0.34	0.20	0.12	0.53						0.16	0.15		
% Población con diabetes	-0.69										0.68				
% Población con obesidad		0.30	0.68			0.37	0.25	0.35	0.33		-0.40			0.32	
	G1	**G2**	**G3**	**G4**	**G5**	**G1**	**G2**	**G3**	**G4**	**G5**	**G1**	**G2**	**G3**	**G4**	**G5**

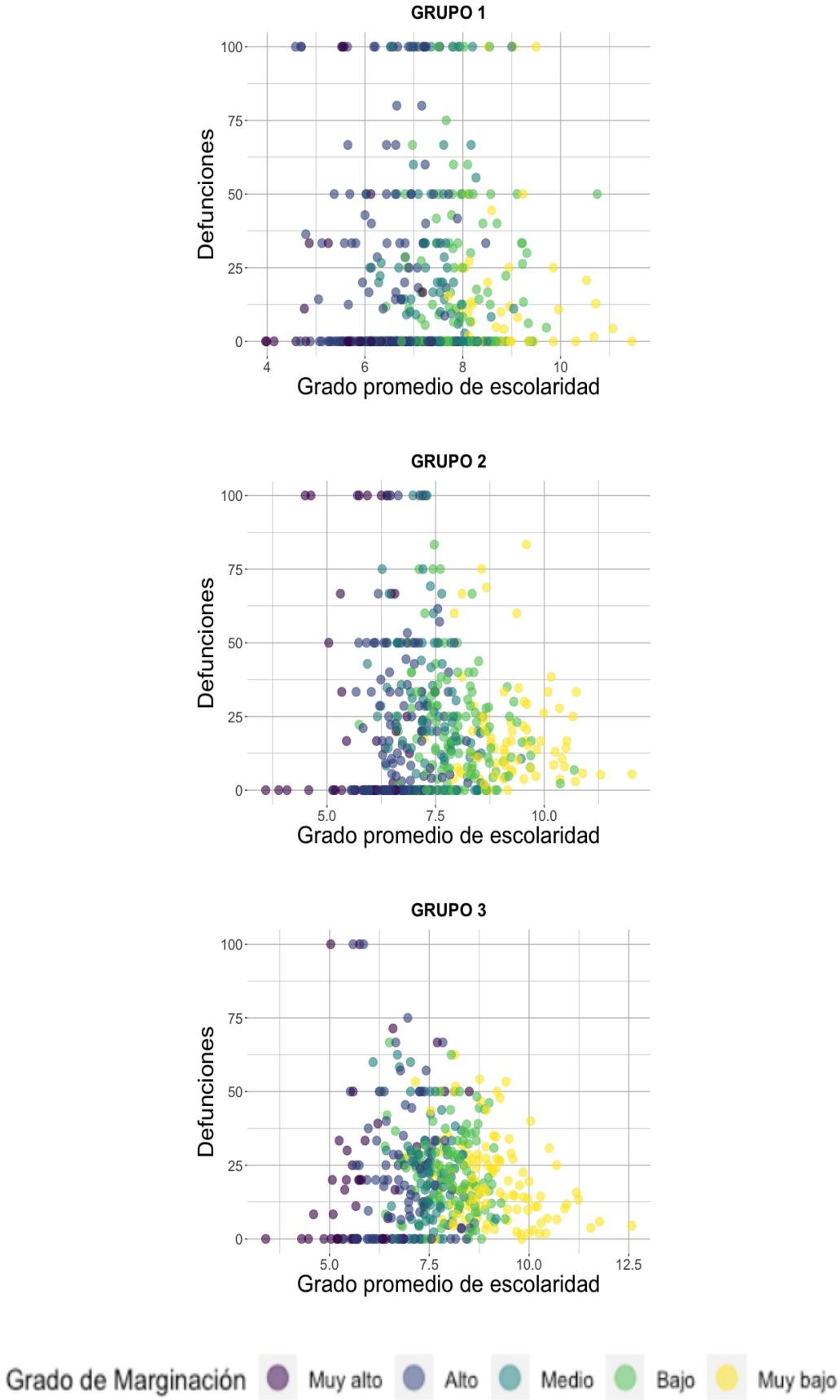
G1 = GRUPO 1	Población menor o igual a 3,419 habitantes
G2 = GRUPO 2	Población mayor a 3,419 habitantes y menor o igual a 9,161 habitantes
G3 = GRUPO 3	Población mayor a 9,161 habitantes y menor o igual a 19,215 habitantes
G4 = GRUPO 4	Población mayor a 19,215 habitantes y menor o igual a 45,399 habitantes
G5 = GRUPO 5	Población mayor a 45,399 habitantes

Tabla 4.5: Tabla de las variables socioeconómicas significativas en cada ola de contagios de la pandemia dividida por los grupos poblacionales del modelo, donde cada valor representa el coeficiente de regresión.

Por último notemos que existen valores de los coeficientes muy discordantes en algunos grupos o picos de la pandemia como es el caso del *grado promedio de escolaridad* que en la segunda ola parece ser que si aumentara una unidad el promedio de escolaridad, entonces las defunciones aumentarían 8.82 % lo cual no parece tener sentido, porque es de esperarse que mientras uno tenga más educación será más consciente y acatará más las medidas sanitarias que se dicten; lo que implica un menor riesgo de contagio y por tanto una disminución en las defunciones y esto se puede notar en los grupos poblacionales 2 y 3 donde las defunciones disminuirían en 3.91 % y 4.13 % respectivamente. Pero entonces, ¿Por qué en el grupo 1 aumentan?

Lejos de hacer suposiciones veamos cómo se ve la relación entre las defunciones por COVID – 19 y el grado promedio de escolaridad en cada uno de estos 3 grupos en la segunda ola de la pandemia. Primero vamos a clasificar a nuestros municipios mediante el grado de marginación que es calculado por el INEGI tomando en cuenta factores socioeconómicos donde clasifica el nivel de marginación de los municipios en 5 categorías distintas: “Muy alto”, “Alto”, “Medio”, “Bajo” y “Muy bajo”; y de esta forma podremos ver más fácilmente la relación que hay entre el *grado promedio de escolaridad* y las *defunciones*.

Podemos observar dicha relación en el gráfico de dispersión del grupo 1 donde se ven muy dispersos los puntos y además hay una cantidad considerable de ellos en el 100 % de las defunciones y en casi todos los grados de escolaridad. En el caso del grupo 2 vemos que las defunciones de 100 % ya disminuyeron y los puntos se acercan más entre sí quedando la mayoría por debajo del 50 % de las defunciones en un rango aproximado entre 6 y 10 grados promedio de escolaridad. Finalmente en el grupo 3 vemos claramente una disminución en las defunciones mientras mayor es el grado promedio de estudios y el índice de marginación es bajo. Entonces, debido a que el modelo selecciona las variables de acuerdo al ajuste de los datos, en el grupo 1 tenemos mucha dispersión y no se tiene una tendencia con la que se pueda determinar que mientras mayor sea el grado de estudios menor es el riesgo de mortalidad por COVID-19, sino al contrario se ve que mientras más aumenta el grado promedio de escolaridad aumenta el porcentaje de defunciones y por eso el coeficiente de regresión que proporciona el modelo es positivo para con el riesgo de fallecer.


GRUPO 1
Defunciones
100
75
50
25
0
4
6
8
10
Grado promedio de escolaridad
GRUPO 2
Defunciones
100
75
50
25
0
5.0
7.5
10.0
Grado promedio de escolaridad
GRUPO 3
Defunciones
100
75
50
25
0
5.0
7.5
10.0
12.5
Grado promedio de escolaridad
Grado de Marginación
Muy alto
Alto
Medio
Bajo
Muy bajo

4.4. Validación del modelo

En esta sección vamos a utilizar los resultados obtenidos del modelo de la tercera ola de la pandemia con el grupo 5 poblacional para validar y hacer una correcta interpretación del modelo. Los supuestos a validar son los que se vieron en la Sección 2 y que a continuación expongo de manera concisa y resumida, aunque es preciso notar que en nuestro caso al ser un modelo obtenido por mínimos cuadrados y que seleccionó variables por el método AIC no necesitamos ningún supuesto más que el de linealidad. Sin embargo por formalidad sí se realizó la verificación de los supuestos y se encuentra de manera explícita en el apéndice 5.

En resumen tenemos que la especificación del modelo es la correcta y la relación es lineal debido a que no se muestra algún patrón específico en las gráficas de residuales contra cada variable regresora, no hay presencia de multicolinealidad porque todos los valores del Factor de Inflación de Varianza (VIF) son menores a 5, de hecho se encuentran entre 1 y 2, por lo que no hay muestra de que exista una relación entre las variables regresoras, lo que provocaría una inestabilidad e imprecisión en los estimadores y que además tendrían una varianza alta, provocando a su vez que los coeficientes de regresión no sean significativos cuando sí lo son o se salgan del intervalo de confianza, pero este no es nuestro caso.

Luego se procedió a comprobar las hipótesis de normalidad y homocedasticidad realizando un estudio de los residuales; era de esperarse que no se cumplieran porque tenemos una muestra aleatoria pero con valores reales de una sociedad donde sus condiciones socioeconómicas no están distribuidas homogéneamente, de hecho en general es muy extraño que se dé la varianza constante en problemas reales.

Ahora bien, en nuestro caso el que no se cumplan estos supuestos no nos afecta porque con ellos se tiene la confianza para hacer la selección de variables por medio del *p-value* de los coeficientes pero no se utilizó este método; nuestra selección de variables se hizo por el método AIC lo que significa que no omitimos variables importantes ni introdujimos variables que no lo son, como hubiera sucedido con el *p-value* provocando errores de

especificación o sobrestimando los parámetros.

Por otra parte, es importante explicar que estos problemas de falta de normalidad y heterocedasticidad se pueden solucionar aplicando alguna transformación de variable o variables pero perderíamos la capacidad de interpretar lo que sucede, e interpretar los resultados es uno de nuestros objetivos en esta tesis. [1]

Sin embargo, como veremos más adelante en la Sección 4.6 probaremos qué sucede si aplicamos transformaciones a la variable respuesta y las variables regresoras.

4.5. Prueba de significancia de la regresión

En la tabla 4.4 se encuentra el modelo que se usó para la prueba de significancia que es el modelo obtenido para el grupo 5 y para que podamos analizarlo mejor se muestra también a continuación con todos sus valores estadísticos importantes.

Call:				
lm(formula = DEF ~PHNV + SSS + P60YM + SDE + SAE, data = GRUPO5)				
Residuals:				
Min	1Q	Median	3Q	Max
-11.8420	-2.7052	-0.6417	1.6977	24.7847
Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-16.41708	2.18022	-7.530	2.75e-13 ***
PHNV	7.69417	0.89565	8.591	<2e-16 ***
SSS	0.18374	0.02687	6.837	2.61e-11 ***
P60YM	0.13898	0.07570	1.836	0.0670 .
SDE	-0.37108	0.07273	-5.102	4.94e-07 ***
SAE	0.08994	0.03595	2.502	0.0127 *
—				
Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
Residual standard error:	4.864 on 455 degrees of freedom			
Multiple R-squared:	0.2533,	Adjusted R-squared:	0.2451	
F-statistic:	30.87 on 5 and 455 DF,			p-value: <2.2e-16

Tabla 4.6: Modelo de regresión lineal múltiple del grupo 5 poblacional en la tercera ola de la pandemia.

Si observamos los resultados podemos corroborar que el modelo es significativo porque a un nivel de significancia del 5 % la prueba $F = 30.87$ con un *p-value* menor a 0.01 por lo que el ajuste a los datos es bueno y se puede aceptar que el modelo no es por azar. El coeficiente de determinación $R^2_{Adj} = 0.2451$, es decir, el 24.5 % de la variación en

las defunciones por COVID-19 en el grupo poblacional más grande de nuestro análisis se puede explicar con 5 de nuestras variables socioeconómicas: el promedio de hijos nacidos vivos, el porcentaje de la población sin seguridad social, la población de 60 años y más, el porcentaje de ocupantes en viviendas particulares habitadas sin drenaje ni excusado y el porcentaje de ocupantes en viviendas particulares habitadas sin agua entubada, esto con una diferencia de $\pm 4.8\,\%$ de defunciones calculadas.

Es preciso notar que nuestra R^2 es pequeña pero mientras se cumplan las condiciones estadísticas esto no significa que sea un mal modelo, al ser un problema de modelización complejo este coeficiente se acepta y se considera que el modelo proporciona información estadísticamente significativa.

4.6. Transformaciones

Veamos si vale la pena hacer transformaciones a las variables de nuestro modelo, utilizando el paquete `Car` de R. Primero veamos que transformar solamente la variable respuesta no nos proporciona ninguna mejoría en el ajuste del modelo, puesto que los valores calculados por la función *InvResPlot* de `R` muestran que el valor de $\hat{\lambda}$ es muy cercano a $\lambda = 1$ lo que significa que no es de utilidad aplicar alguna transformación.

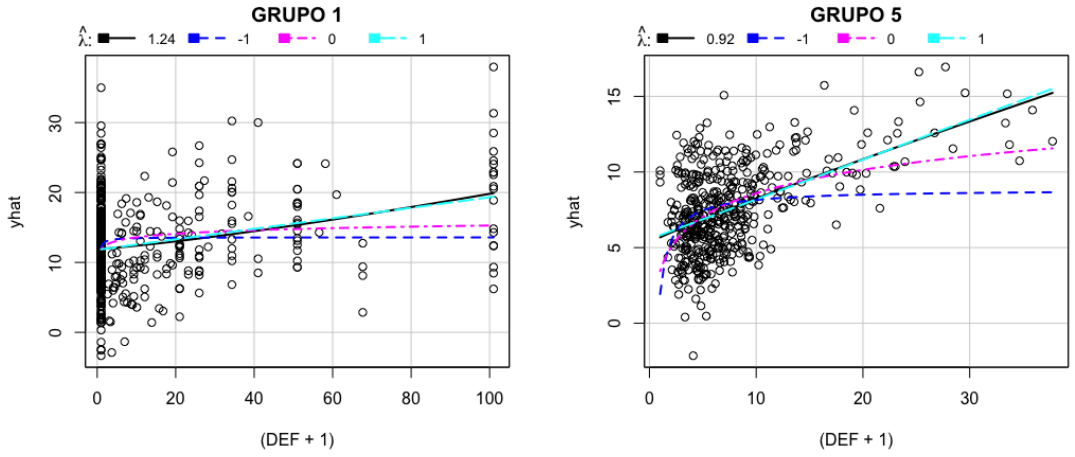


Figura 4.12: Resulatdos de las ransformaciones de la variable respuesta en el grupo 1 y grupo 5.

Al hacer un análisis de las variables regresoras y aplicando la función *powerTransform* de `R` obtenemos que las potencias ideales para transformas nuestras variables son las que se muestran en la Tabla 4.7, por lo tanto, el modelo ajustado con las transformaciones de las variables regresoras elevadas a la potencia sugerida queda descrito como se puede ver en la Tabla 4.8 donde podemos observar que el modelo es significativo aunque el ajuste es menor con un coeficiente de determinación $R^2_{Adj} = 0.2175$ en comparación al $R^2_{Adj} = 0.2451$ del modelo inicial sin transformaciones (tabla 4.6), donde además las variables significativas resultan ser otras, en este caso tenemos que las variables significativas son: el promedio de hijos nacidos vivos, (%) población sin seguridad social, (%) población de 60 años y más, (%) población en viviendas sin drenaje ni excusa y (%) población en localidades con menos de 5000 habitantes. Esto aunado a que con este modelo se pierde la capacidad de interpretación porque no es fácil explicar que significa o que representa en términos reales (%) población sin seguridad social elevada a la potencia 0.3325.

bcPower Transformations to Multinormality				
	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
POBES	0.4254	0.50	0.1390	0.7117
PDIAB	0.6728	0.50	0.4904	0.8553
PHNV	-0.0050	0.00	-0.2929	0.2828
GPEST	1.4659	1.47	1.2037	1.7281
PEA	3.1724	3.17	2.6076	3.7372
SSS	0.3325	0.50	0.1001	0.5650
POBMAS	-1.7253	1.00	-4.4823	1.0317
P60YM	0.7058	0.71	0.5181	0.8936
PHLI	-0.0718	-0.07	-0.1202	-0.0235
ANALF	-0.0008	0.00	-0.0633	0.0618
POV	-0.9252	-1.00	-1.3961	-0.4544
SDE	0.0096	0.00	-0.0205	0.0398
SEE	-0.0430	-0.04	-0.0803	-0.0058
SAE	0.0167	0.00	-0.0285	0.0618
VPT	0.0054	0.00	-0.0491	0.0599
VHAC	0.1722	0.17	0.0434	0.3010
P2SM	1.5174	1.52	1.1458	1.8890
P5000	0.4009	0.40	0.3330	0.4687

Likelihood ratio test that transformation parameters are equal to 0 (all log transformations)		
	LRT df	pval
LR test, lambda = (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)	810.2532 18	<2.22e-16
Likelihood ratio test that no transformations are needed		
	LRT df	pval
LR test, lambda = (1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1)	8024.125 18	<2.22e-16

Tabla 4.7: Transformaciones potencias a las variables regresoras.

Call:
lm(formula = DEF ~ I(POBES^(0.4254)) + I(PDIAB^(0.6728)) + log(PHNV) + I(GPEST^(1.4659)) + I(PEA^(3.1724)) + I(SSS^(0.3325)) + I(POBMAS^(-1.7253)) + I(PHLI^(-0.0718)) + log(ANALF) + I(P60YM^(0.71)) + I(POV^(-0.9252)) + log(SDE) + I(SEE^(-0.043)) + log(SAE) + log(VPT) + I(VHAC^(0.1722)) + I(P5000^(0.4009)) + I(P2SM^(1.52)), data = GRUPO5)

Residuals:

Min	1Q	Median	3Q	Max
-11.8420	-2.7052	-0.6417	1.6977	24.7847

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-1.444e+01	2.131e+01	-0.678	0.49821
POBES^(0.4254)	9.007e-01	1.115e+00	0.808	0.41975
PDIAB^(0.6728)	-4.894e-02	2.972e-01	-0.165	0.86926
log(PHNV)	1.451e+01	4.769e+00	3.043	0.00248 **
GPEST^(1.4659)	-4.816e-02	1.537e-01	-0.313	0.75415
PEA^(3.1724)	1.894e-06	2.714e-06	0.698	0.48562
SSS^(0.3325)	4.791e+00	8.700e-01	5.507	6.20e-08 ***
POBMAS^(-1.7253)	3.162e+03	8.814e+03	0.359	0.71992
PHLI^(-0.0718)	-2.133e+00	2.686e+00	-0.794	0.42742
log(ANALF)	1.559e-01	1.183e+00	0.132	0.89518
P60YM^(0.71)	5.360e-01	3.817e-01	1.404	0.16098
POV^(-0.9252)	4.723e-01	1.700e+01	0.028	0.97785
log(SDE)	-1.418e+00	3.300e-01	-4.297	2.13e-05 ***
SEE^(-0.043)	-1.115e+01	9.926e+00	-1.123	0.26198
log(SAE)	-1.933e-02	3.051e-01	-0.063	0.94952
log(VPT)	-1.979e-01	4.969e-01	-0.398	0.69063
VHAC^(0.1722)	-4.591e+00	5.082e+00	-0.903	0.36680
P5000^(0.4009)	8.549e-01	3.228e-01	2.648	0.00838 **
P2SM^(1.52)	4.116e-03	2.563e-03	1.606	0.10902

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.952 on 442 degrees of freedom
Multiple R-squared: 0.2481, Adjusted R-squared: 0.2175
F-statistic: 8.102 on 18 and 442 DF, p-value: <2.2e-16

Tabla 4.8: Modelo de RLM transformado del grupo 5 en la tercera ola de la pandemia.

Una vez que transformamos por un lado solamente la variable respuesta y vimos que no hay mejoría, y por otro lado solamente las variables regresoras y tampoco hay mejoría, si transformamos la variable respuesta aplicando de nuevo la función *InvResPlot* al modelo que considera las variables regresoras transformadas obtenemos el valor de $\hat{\lambda} = 0.69$ para la variable respuesta que se ajusta un poco mejor a los datos como se ve en la figura 4.13. Entonces, aplicando esa transformación al modelo observamos una pequeña mejoría en el R^2_{Adj} y de igual forma se disminuyó la desviación estándar, sin embargo sigue siendo menor el ajuste en comparación con el primero modelo que consideramos (4.6).

Por lo tanto, podemos concluir que realizar la transformación de variables en nuestro caso no es de utilidad ya que no se mejora el ajuste del modelo ni ayuda para cumplir los supuestos (ver Apéndice 5) por lo que se decidió seguir con el modelo inicial.

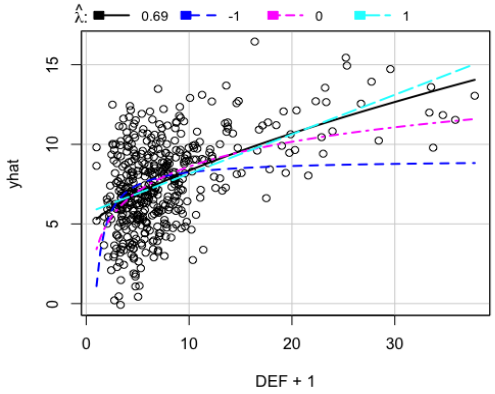


Figura 4.13: Valores estimados para la transformación de la variable respuesta con las variables regresoras ya transformadas.

Call:
lm(formula = I(DEF^(0.69)) ~ I(POBES^(0.4254)) + I(PDIAB^(0.6728)) + log(PHNV) + I(GPEST^(1.4659)) + I(PEA^(3.1724)) + I(SSS^(0.3325)) + I(POBMAS^(-1.7253)) + I(PHLI^(-0.0718)) + log(ANALF) + I(P60YM^(0.71)) + I(POV^(-0.9252)) + log(SDE) + I(SEE^(-0.043)) + log(SAE) + log(VPT) + I(VHAC^(0.1722)) + I(P5000^(0.4009)) + I(P2SM^(1.52)), data = GRUPO5)

Residuals:

Min	1Q	Median	3Q	Max
-11.8420	-2.7052	-0.6417	1.6977	24.7847

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.743e+00	7.131e+00	-0.525	0.59991
I(POBES^(0.4254))	3.997e-01	3.732e-01	1.071	0.28479
I(PDIAB^(0.6728))	-3.858e-02	9.945e-02	-0.388	0.69826
log(PHNV)	5.094e+00	1.596e+00	3.192	0.00151 **
I(GPEST^(1.4659))	-1.334e-02	5.143e-02	-0.259	0.79548
I(PEA^(3.1724))	5.001e-07	9.082e-07	0.551	0.58215
I(SSS^(0.3325))	1.749e+00	2.912e-01	6.008	3.92e-09 ***
I(POBMAS^(-1.7253))	9.434e+02	2.950e+03	0.320	0.74926
I(PHLI^(-0.0718))	-8.083e-01	8.988e-01	-0.899	0.36899
log(ANALF)	4.083e-02	3.957e-01	0.103	0.91787
I(P60YM^(0.71))	1.934e-01	1.277e-01	1.514	0.13080
I(POV^(-0.9252))	9.970e-01	5.690e+00	0.175	0.86099
log(SDE)	-4.482e-01	1.104e-01	-4.059	5.84e-05 ***
I(SEE^(-0.043))	-3.970e+00	3.322e+00	-1.195	0.23272
log(SAE)	-4.113e-02	1.021e-01	-0.403	0.68725
log(VPT)	-5.303e-02	1.663e-01	-0.319	0.74995
I(VHAC^(0.1722))	-1.882e+00	1.701e+00	-1.106	0.26914
I(P5000^(0.4009))	2.892e-01	1.080e-01	2.677	0.00770 **
I(P2SM^(1.52))	1.404e-03	8.578e-04	1.637	0.10240

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.657 on 442 degrees of freedom
Multiple R-squared: 0.2531, Adjusted R-squared: 0.2227
F-statistic: 8.322 on 18 and 442 DF, p-value: <2.2e-16

Tabla 4.9: Modelo de RLM con la variable respuesta y las variables regresoras transformadas del grupo 5 en la tercera ola de la pandemia.

4.7. Predicción

Con base en el modelo obtenido del grupo 5 en la tercera ola que se explicó anteriormente, se hizo la predicción con los datos de la cuarta ola de la pandemia que corresponden a los meses de "enero, febrero y marzo del 2022" para determinar la probabilidad de riesgo de fallecer por COVID–19 modelada con las 5 variables socioeconómicas relevantes de nuestro modelo base.

De igual forma, estas mismas predicciones se pueden hacer para cada grupo poblacional considerando sus modelos; en este caso veremos la predicción para el grupo 5 que es el grupo con mayor densidad poblacional.

Como se muestra en la figura 4.14 obtuvimos que nuestra predicción se ajusta bien y cae dentro de los intervalos de confianza de predicción, que serían las líneas 'azul marino'. El modelo ajustado con la tercer ola es el marcado en color 'rosa', y es interesante notar que nuestra predicción que se ve en color 'turquesa', queda por debajo de los valores que se obtuvieron en la tercer ola. Es decir, nuestra predicción nos dice que en la cuarta ola habrá más defunciones que las que hubo en la tercer ola. Esta predicción nos dice que si y solo si las variables socioeconómicas que influyan en el modelo disminuyen, entonces se tendrán menos defunciones en la 4ta ola. Aunque es importante notar y recalcar que los tiempos no son comparables; los tiempos entre olas pasan muy rápido en comparación al tiempo que tardan en cambiar las condiciones socioeconómicas de una población. Además, no hay que olvidar que existen otros factores que no se consideraron dentro de nuestro análisis como lo es la vacunación, la mayoría de la población ya cuenta con al menos 1 dosis de la vacuna contra el virus, lo cual implicó una reducción considerable de las muertes por COVID–19. Sin embargo, esto nos sirve para identificar y conocer los factores socioeconómicos que son relevantes o pueden influir en una situación similar.

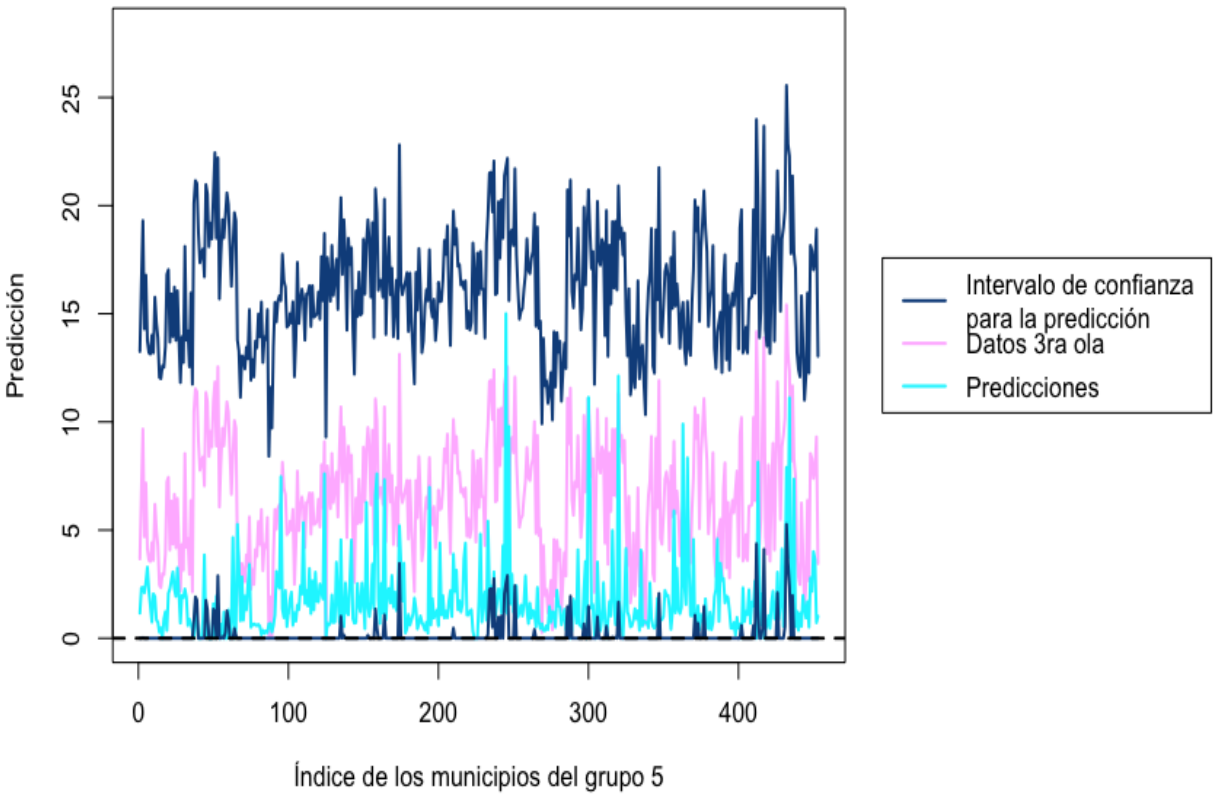


Figura 4.14: Predicción del riesgo de mortalidad por COVID-19 en la cuarta ola de la pandemia con nuestro modelo de regresión descrito en la sección 4.5

Capítulo 5

Estudio de los factores socioeconómicos en la probabilidad de fallecimiento por COVID – 19 en México

Con el análisis exploratorio, los modelos prueba que se hicieron y el modelo final que obtuvimos para hacer predicciones, mostraron que los factores socioeconómicos sí influyen en la mortalidad por el SARS-COV-2 a los mexicanos. Sin embargo la influencia es muy baja y no se ve tan marcada como se esperaba; aún tomando en cuenta diversas consideraciones dentro de la población y la evolución de la pandemia.

Dentro de nuestros modelos obtenidos de cada ola de la pandemia que se observan en las tablas (4.2, 4.3, 4.4) vimos que los factores socioeconómicas más relevantes que influyeron en cada uno de los grupos poblacionales que se analizaron, fueron las variables relacionadas con la educación como el **grado promedio de estudios** donde se mostró que mientras más alto fuera el nivel de educación disminuye el riesgo de fallecer por COVID-19 en especial en los grupos **2**, **3** y **4** donde los municipios tienen una población mayor a 3,419 habitantes y menor a 45,399. Por otra parte, las variables relacionadas con el empleo también influyeron pero de manera inversa a este riesgo, es decir, si el **(%) población económicamente activa** aumentara en los municipios donde la población no es ni muy grande ni muy reducida (al menos en la segunda y tercera ola) aumentaría el

riesgo de fallecer dentro de esta población. En el caso del **(%) población ocupada con ingresos de hasta 2 salarios mínimos** resultó muy afectada durante la primera ola de la pandemia, que fue cuando se afectó más el sector económico dejando sin ingresos a muchas personas que vivían de la economía informal, que por lo mismo no cuentan con seguridad social para atenderse en dado caso de que se contagiaran; y aunado a la ignorancia y desconocimiento del virus, provocó muchos decesos dentro de este porcentaje de la población principalmente en la primera ola.

Los factores sociales fueron sumamente importantes para determinar como sería la evolución del paciente llegando a fallecer en la mayoría de los casos si no se contaba con algún servicio de salud, por eso la variable del **(%) población sin seguridad social** resultó muy significativa sobre todo en los municipios grandes en términos de población, de la misma forma, si el **promedio de hijos nacidos vivos** es alto o si se vive en localidades con menos de 5,000 habitantes (que por lo general son zonas marginadas) se corre un mayor riesgo de fallecer por COVID-19. Y si consideramos las condiciones de vivienda, las personas que residen en viviendas sin agua entubada resultan más afectadas en esta pandemia en municipios donde la población es muy grande ya que el virus se transmite al tener contacto con él (mientras más grande la población mayor es el riesgo de contacto) siendo indispensable tener agua para lavarse las manos constantemente pero en viviendas donde no cuentan con agua para hacerlo, aumenta el riesgo.

En cuanto a las comorbilidades, de las 3 más importantes que se encuentran presentes en la población mexicana (obesidad, diabetes e hipertensión), solamente el **(%) de la población obesa** fue significativo y un factor clave para aumentar el riesgo de fallecer, además siendo un país donde el 70 % de su población sufre sobrepeso se corre un mayor riesgo de que aumenten las defunciones al contagiarse del virus SARS-CoV-2. Sin embargo, cuando se vio más afectado este porcentaje de la población fue durante la segunda ola de contagios que coincide con las fechas cuando aún no se vacunaba la mayoría de la población y donde la propagación fue mayor ya que se registraron más contagios, como era de esperarse, a comparación de la primera ola.

De manera más general podemos ver nuestros resultados en la Tabla 5.1 que muestra las variables socioeconómicas que estuvieron más presentes en cada ola de la pandemia (en al menos 2 grupos poblacionales) que influyeron de alguna u otra forma a las defunciones por COVID-19, con la finalidad de que sean considerados como factor de riesgo de mortalidad.

Variables \ **Fechas**	**1ra ola** Junio 2020 - Agosto 2020	**2da ola** Diciembre 2020 - Febrero 2021	**3ra ola** Julio 2021 - Septiembre 2021
Grado promedio de escolaridad de la población de 15 años o más	✓	✓	✓
% Población económicamente activa de la población de 15 años o más	✓		✓
% Población sin seguridad social	✓	✓	✓
Promedio de ocupantes por vivienda		✓	✓
Promedio de hijos nacidos vivos		✓	✓
% Población habla lengua indígena			
% Población con analfabetismo	✓	✓	✓
% Ocupantes en viviendas particulares sin drenaje ni excusado	✓		✓
% Ocupantes en viviendas particulares sin energía eléctrica	✓		
% Ocupantes en viviendas particulares sin agua entubada	✓	✓	✓
% Ocupantes en viviendas particulares con piso de tierra			
% Viviendas particulares con hacinamiento		✓	✓
% Población en localidades con menos de 5000 habitantes		✓	
% Población masculina			✓
% Población de 60 años y más	✓	✓	✓
% Población ocupada con ingresos de hasta a 2 salarios mínimos	✓	✓	✓
% Población con diabetes			
% Población con obesidad	✓	✓	✓

Tabla 5.1: Tabla con las variables socioeconómicas más presentes en nuestros modelos en cada ola de la pandemia por COVID-19.

Por lo que podemos concluir que las variables socioeconómicas más significativas e influyentes en las defunciones por estar presente en las 3 olas de contagios fueron el *grado promedio de estudios*, la *población sin seguridad social*, la *población analfabeta*, el *porcentaje de ocupantes en viviendas sin agua entubada*, la *población de 60 años y más*, la *población*

ocupada con ingresos de hasta 2 salarios mínimos, y la *población con obesidad*. Las cuales influyeron para un mayor o menor número de defunciones de acuerdo a cada ola de la pandemia y a cada grupo poblacional.

Y analizándolo en cada ola de la pandemia veamos como fue la evolución y que variables influyeron en las defunciones por la enfermedad COVID-19 causada por el SARS-CoV-2 de acuerdo a las medidas y conocimientos que se tenían en cada periodo:

Primera ola: En este periodo apenas se tenía conocimiento de cómo se propagaba el virus, aún no había vacunas y las medidas sanitarias que implementó el gobierno de México con la fase 2 de la epidemia fueron: distanciamiento social, suspensión de actividades no esenciales del sector público, privado y social, confinamiento, además del uso obligatorio del cubrebocas, medidas de higiene básicas y de desinfección en la entrada de inmuebles. Además el gobierno de México inició con un programa de conferencias de prensa a cargo de la Secretaría de Salud que presentaba el epidemiólogo Hugo López Gatell con la finalidad de informar como evolucionaba la pandemia en el país, qué medidas se irían tomando y ayudar al esclarecimiento de las dudas que tuviera la población sobre la enfermedad adquirida por el virus SARS-CoV-2.

El porcentaje de los pacientes ambulatorios de 18 a 59 años durante este periodo fue de 64.5 % y la población masculina fue la más afectada. [20]

Ahora gracias a nuestro modelo podemos determinar que las variables influyentes en esta ola fueron el *grado promedio de escolaridad*, la *población económicamente activa*, la *población sin seguridad social*, la *población analfabeta*, el *porcentaje de ocupantes en viviendas particulares habitadas sin drenaje ni excusado*, el *porcentaje de ocupantes en viviendas particulares habitadas sin energía eléctrica*, el *porcentaje de ocupantes en viviendas particulares habitadas sin agua entubada*, la *población de 60 años y más*, la *población ocupada con ingresos de hasta 2 salarios mínimos* y la *población con obesidad*.

Segunda ola: En este segundo pico ya se tenía más conocimiento del virus y fue cuando inició la vacunación para adultos mayores en el mes de diciembre, había un desconfinamiento al aperturar gradualmente el sector económico en el país pero con una reducción de la capacidad en restaurantes, plazas y lugares públicos y siguen las medidas de higiene, desinfección y el uso obligatorio del cubrebocas. Sin embargo, el número de contagios y defunciones fue mucho mayor en este periodo con respecto a la primera ola y de acuerdo a los informes del gobierno, el grupo etario de 20 a 29 años fue el más afectado. Además se observó un aumento de los casos en la población femenina durante este periodo y también del porcentaje de los pacientes ambulatorios de 18 a 59 años llegando al 70.3 % del total de casos. [20]

Las variables influyentes en esta ola fueron el *grado promedio de escolaridad*, la *población sin seguridad social*, el *promedio de ocupantes por vivienda*, el *promedio de hijos nacidos vivos*, la *población analfabeta*, el *porcentaje de ocupantes en viviendas particulares habitadas sin agua entubada*, las *viviendas particulares con hacinamiento*, la *población que vive en localidades con menos de 5,000 habitantes*, la *población de 60 años y más*, la *población ocupada con ingresos de hasta 2 salarios mínimos* y la *población con obesidad*.

Tercera ola: En este periodo ya se encontraba vacunada al menos con 1 dosis casi el 50 % de la población mexicana [22] y las medidas sanitarias que persisten solo son el uso obligatorio del cubrebocas, higiene y desinfección. Durante este periodo el porcentaje de los pacientes ambulatorios de 18 a 59 años fue de 79.4 % y la población masculina fue la más afectada pero solo con 1 % más casos que la población femenina, sin embargo, este grupo de edad representa el 70.5 % de todas las hospitalizaciones, cifra muy por encima de la primera (52.5 %) y segunda (45.3 %) ola. [20]

Las variables influyentes en esta ola fueron el *grado promedio de escolaridad*, la *población económicamente activa*, el *porcentaje de la población sin seguridad social*, el *promedio de ocupantes por vivienda*, el *promedio de hijos nacidos vivos*, el *porcentaje de la población analfabeta*, el *porcentaje de ocupantes en viviendas particulares ha-*

bitadas sin drenaje ni excusado, el *porcentaje de ocupantes en viviendas particulares habitadas sin agua entubada*, las *viviendas particulares con hacinamiento*, la *población masculina*, la *población de 60 años y más*, la *población ocupada con ingresos de hasta 2 salarios mínimos* y la *población con obesidad*.

Para finalizar con esta tabla (5.1) podemos observar que las variables socioeconómicas significativas fueron aumentando conforme iba evolucionando la pandemia, a pesar del conocimiento del virus SARS–CoV–2 en general y la vacunación implementada, no obstante esto puede deberse al aumento de casos positivos de COVID–19 que hubo en la segunda y tercer ola con respecto a la primera, ya que en la primera ola se registraron un total de 543,927 casos confirmados, en la segunda ola un total de 928,080 y en la tercera ola se registraron 1,152,077 casos positivos.

Ahora bien, volviendo al diagrama de correlación (4.7) es importante notar que hay una variable socioeconómica muy relevante que influye en otras y esta es el **grado promedio de estudios**, haciendo énfasis en la variable "GPEST" vemos que de 20 variables, en 15 tiene una relación lineal negativa por lo que es evidente que afecta a esas variables y por lo tanto aumenta el riesgo de mortalidad ya que como vimos posteriormente en nuestros modelos, las variables como el analfabetismo, las variables del porcentaje de la población en vivienda sin servicios básicos o con hacinamiento y el porcentaje de la población con hasta 2 salarios mínimos son variables que sí influyen en las defunciones por COVID–19.

Conclusiones

De acuerdo a los resultados obtenidos se encontró que los factores socioeconómicos sí influyen en el riesgo de fallecer al contagiarse del virus pero no son determinantes, tanto pueden afectar como no, y se debe a que hubo muchos factores más que influyen y no se tomaron en cuenta en este análisis como la movilidad de la población afectada, la vacunación, la inmunidad (momentánea) adquirida y también el continuo conocimiento del virus que fue adquiriendo la población. Sin embargo, los datos demuestran que dependiendo el periodo de la pandemia que sea y el municipio donde se encuentre una persona, es decir, si es un municipio chico o grande en cuanto a población se refiere pueden afectarle distintos factores socioeconómicos, como por ejemplo: el (%) población sin seguridad social o con ingresos de hasta 2 salarios mínimos o si cuenta o no con agua entubada en su vivienda.

Se identificó a los grupos poblacionales más afectados dependiendo su condición social y económica, encontrando que el factor socioeconómico más influyente es el: **grado promedio de estudios**; mientras más estudios se tengan menor será el riesgo de fallecer, principalmente en los grupos medianos (2 y 3) donde en promedio ese riesgo disminuiría en 3.75 % si se aumentara en una unidad el grado promedio de estudios. En el caso de no contar con seguridad social y/o encontrarse en condiciones de pobreza (ingreso menor a 2 salarios mínimos o vivir en viviendas sin servicios básicos) aumenta el riesgo de fallecer en cierta cantidad dependiendo el grupo poblacional y la ola. Y si viven en localidades con menos de 5,000 habitantes o se tiene un promedio de hijos muy alto también aumenta el riesgo de fallecer. También se detectó que de las 3 comorbilidades más importantes y que están presentes en nuestro país, solo las personas con obesidad son las que resultaron más afectadas corriendo un riesgo de 0.37 % si aumenta en 1 % el porcentaje de esta población.

La predicción con los factores socioeconómicos resultó correcta para la 4ta ola, a pesar del bajo coeficiente de determinación, con la cual se predijo que habría menos defunciones en comparación con la tercera ola. Ahora que ya pasó la 4ta ola podemos constatar que sí hubo menos defunciones pero fue por otros factores como la vacunación de la población, al menos con una dosis para esas fechas, y como sabemos a pesar de que se tuvo una variante del virus SARS–CoV–2 más contagiosa (la variante ÓMICRON) era considerablemente menos mortal en comparación con otras variantes. Queda por determinar si alguna de las variables que resultaron significativas en los modelos realmente aumentaron o disminuyeron durante la 4ta ola para considerarlas como factores determinantes.

Por lo tanto, se concluye que este trabajo fue exitoso al identificar los grupos poblacionales más afectados dependiendo su condición social y económica encontrando que el factor socioeconómico más influyente es el **grado promedio de estudios**, ya que dependiendo del grado de estudios que se tenga disminuye el porcentaje de otras variables socioeconómicas como: el analfabetismo, el (%) población en viviendas sin servicios básicos o con hacinamiento y el (%) población con hasta 2 salarios mínimos, que como vimos en los modelos obtenidos son variables significativas que resultaron relevantes en el riesgo de mortalidad por el virus SARS–CoV–2.

La pandemia representa un problema muy complejo ya que las condiciones entre cada ola fueron cambiando en el tiempo, sin embargo, las condiciones sociales y económicas no cambian tan rápido. Por lo cual se necesitan enfocar esfuerzos en mejorar la educación a nivel nacional y la situación general en la que vive la población para que tenga una mejor resiliencia ante cualquier adversidad.

Bibliografía

[1] M. Allen. *Understanding Regression Analysis*. Plenum Press, NY, 1 edition, 1997.

[2] K. Burnham and D. Anderson. *Model Selection and Multimodel Inference*. Springer, New York, NY, second edition, 2002.

[3] B. Calixto-Calderón, M. F. Vázquez-González, R. Martínez Peláez, J. R. Bermeo-Escalona, V. García, L. J. Mena, G. Maestre, J. R. Parra-Michel, L. A. Ceja Bravo, and P. L. López-de Alba. Pre-existing comorbidity, the highest risk factor for poor prognosis of COVID-19 among the mexican population. *Nova Scientia*, 13(e), Abril 2021.

[4] S. de Salud. 098. medidas de seguridad sanitaria https://www.gob.mx/salud/prensa/098-medidas-de-seguridad-sanitaria?idiom=es.

[5] G. Esquivel. Los impactos económicos de la pandemia en méxico. pages 1–18, Julio 2020.

[6] Gobierno de México. Página oficial del CONACYT del gobierno de México para datos COVID-19
https://datos.covid-19.conacyt.mx/.

[7] Gobierno de México. Página oficial referente al COVID-19 del gobierno de México https://coronavirus.gob.mx/.

[8] J. González and E. San Martín. Muchas curvas, misma información: Sobre la indeterminación del modelo sir y su uso en el contexto de la pandemia del COVID–19. *SOCHE*, junio 2020.

[9] G. J. V.-L. M. G. González-Pérez, S. Romero-Valle, A. Vega-López, and C. E. Cabrera-Pivaral. Exclusión social e inequidad en salud en méxico: Un análisis socio–espacial. *Revista de salud pública*, 10(1), 2008.

[10] INEGI. Prevalencia de obesidad, hipertensión y diabetes para los municipios de méxico 2018, 2018.

[11] INEGI. Encuesta nacional de ocupación y empleo (ENOE), 2019.

[12] INEGI. Encuesta telef'onica de ocupación y empleo (ETOE), 2020.

[13] INEGI. Índice de marginación por municipio 2020 (IMM), 2020.

[14] INEGI. Estadística de defunciones registradas 2020. (Nota Técnica), 2021.

[15] INEGI. Principales resultados por localidad (ITER). Censo de Población y Vivienda 2020, 2021.

[16] D. Montgomery, E. Peck, and G. Vining. *Introduction to linear regression analysis*. Wiley, 2012.

[17] OCDE. Organización para la cooperación y el desarrollo económico http://www.oecd.org/centrodemexico/estadisticas/.

[18] M. A. Peréz Méndez and R. Gutiérrez Rodríguez. Modelando la supervivencia a la COVID-19 en méxico. *Denarius*, 1(40), Agosto 2021.

[19] J. Rivera Dommarco, M. Colchero, M. Fuentes, T. González de Cosío Martínez, C. Aguilar, G. Hernández, and S. Barquera. *La obesidad en México. Estado de la política pública y recomendaciones para su prevención y control*. Instituto Nacional de Salud Pública, Cuernavaca, México, 2018.

[20] Secretaría de Salud. *27 Informe epidemiológico de la situación de COVID–19*. Dirección de Información epidemiológica, Julio 2021.

[21] United Nations. World population prospects 2019 https://population.un.org/wpp/.

[22] University of Oxford. Coronavirus (COVID-19) vaccinations https://ourworldindata.org/covid-vaccinations?country=MEX.

[23] S. Wesberg. *Applied Linear Regression.* Wiley, 4 edition, 1980.

Lista de Variables

Nota: Entiéndase como **paciente** a un individuo contagiado del virus SARS-CoV-2.

ID - Número de identificación de referencia de cada municipio en la República Mexicana.

SEXO - Sexo con el que se identificó al paciente.

EDAD - Edad del paciente.

TPACIE - Tipo de paciente: ambulatorio u hospitalizado.

DEF - Identifica si el paciente falleció o no.

NEUMO - Identifica si al paciente se le diagnosticó con neumonía.

DIAB - Identifica si el paciente tiene un diagnóstico de diabetes.

ASMA - Identifica si el paciente tiene un diagnóstico de asma.

INMUSU - Identifica si el paciente presenta inmunosupresión.

HIPER - Identifica si el paciente tiene un diagnóstico de hipertensión.

CARDIO - Identifica si el paciente tiene un diagnóstico de enfermedades cardiovasculares.

EPOC - Identifica si el paciente tiene un diagnóstico de EPOC.

OBES - Identifica si el paciente tiene un diagnóstico de obesidad.

RENALC - Identifica si el paciente tiene diagnóstico de insuficiencia renal crónica.

TABAQ - Identifica si el paciente tiene hábito de tabaquismo.

CLASIFICACIONFINAL - Identifica si el paciente es un caso de COVID-19 según el catálogo CLASIFICACIONFINAL.

FECHASINTOMAS - Identifica la fecha en que inició la sintomatología del paciente.

POBTOT - Población total por municipios en la República Mexicana.

PHNV - Promedio de hijas e hijos nacidos vivos de las mujeres de 12 a 130 años de edad.

GPEST - Grado promedio de escolaridad de las personas de 15 a 130 años de edad.

PEA - Porcentaje de la población de 12 años y más económicamente activa.

OLABOR - Porcentaje de la población de 12 años y más ocupada.

SSS - Porcentaje de la población que no está afiliada a servicios médicos en ninguna institución pública o privada.

PHLI - Porcentaje de la población de 3 años y más que habla alguna lengua indígena y habla español.

ANALF - Porcentaje de la población de 15 años y más analfabeta.

POV - Promedio de ocupantes por vivienda.

VSTIC - Viviendas particulares habitadas sin tecnologías de la información y de la comunicación (TIC)

SBASC - Porcentaje de la población de 15 años o más sin educación básica.

SDE - Porcentaje de ocupantes en viviendas particulares sin drenaje ni excusado.

SEE - Porcentaje de ocupantes en viviendas particulares sin energía eléctrica.

SAE - Porcentaje de ocupantes en viviendas particulares sin agua entubada.

VPT - Porcentaje de ocupantes en viviendas particulares con piso de tierra.

VHAC - Porcentaje de viviendas particulares con hacinamiento.

P5000 - Porcentaje de la población en localidades con menos de 5 000 habitantes.

P2SM - Porcentaje de la población ocupada con ingresos menores a 2 salarios mínimos.

POBES - Porcentaje de la prevalencia de obesidad en la población.

PDIAB - Porcentaje de la prevalencia de diabetes en la población.

PHIPER - Porcentaje de la prevalencia de hipertensión en la población.

Verificación de supuestos

Esta sección es solamente para ver lo que pasa con nuestro modelo, que supuestos se cumplen y cuales.

1. **Correcta especificación del modelo:** Es decir si es lineal o no.

   La especificación del modelo es correcta debido a que se construyó bajo la definición del modelo de regresión lineal 2.5.

2. **No hay colinealidad perfecta:** No existe multicolinealidad entre las variables regresoras como lo podemos corroborar con el Factor de Inflación de Varianza.

VIF				
PHNV	**SSS**	**P60YM**	**SDE**	**SAE**
1.52	1.02	1.04	1.47	1.28

3. **El número de observaciones es mayor que el número de variables:** Se trabajó con un total de 30 variables y 2,624,272 observaciones (casos positivos en total) que se resumieron en 2,437 (observaciones) municipios.

4. **Los predictores son valores fijos:** Como se definió al principio en el capítulo 2, las variables dentro del vector X no son aleatorias.

5. **Homocedasticidad:** Por medio del primer gráfico en la figura 1 podemos notar que tenemos una heterocedasticidad – varianza creciente en nuestros errores ya que no se distribuyen de forma aleatoria alrededor del cero.

6. **Normalidad de los errores:** Por medio del gráfico QQ-plot se puede ver que tiene una ligera tendencia normal pero con una desviación en la cola derecha y esto se debe a que hay problemas de normalidad.

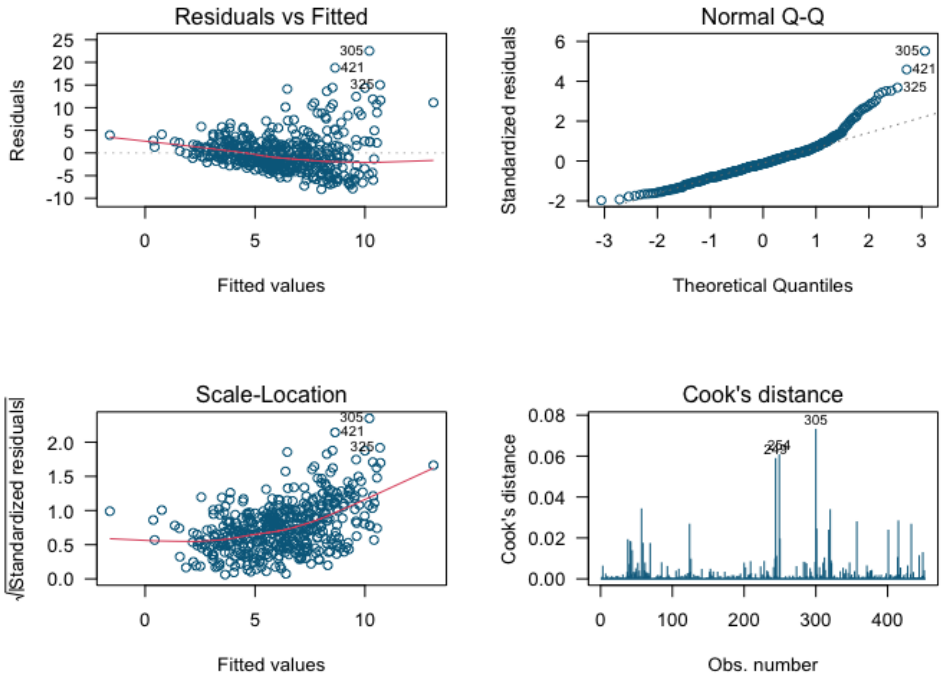


Figura 1: Gráficos para verificar si hay homocedasticidad y normalidad de los errores, además de ver si existen outliers.

7. **No hay correlación entre los errores:** Como podemos corroborar en la tabla 2 los residuales no son independientes, sin embargo, como se dispone de una muestra suficientemente grande de observaciones se puede ignorar el que no se cumpla este supuesto.

Durbin-Watson test	
data:	AICm5s
DW = 1.4969	p-value = 1.964e-08
alternative hypothesis:	true autocorrelation is greater than 0

Tabla 2: Resultados de la prueba Durbin–Watson para nuestro modelo.

La **distancia de Cook** es una estadística para identificar *outliers* o valores atípicos dentro de nuestra muestra. Sin embargo, al ser una muestra de una población real es inevitable no tenerlos.

Verificación de supuestos para modelo transformado

Resumiremos la verificación de supuestos al ver el siguiente gráfico donde se pueden corroborar los supuestos más importantes, donde es fácil notar que la varianza sigue siendo creciente y existe falta de normalidad .

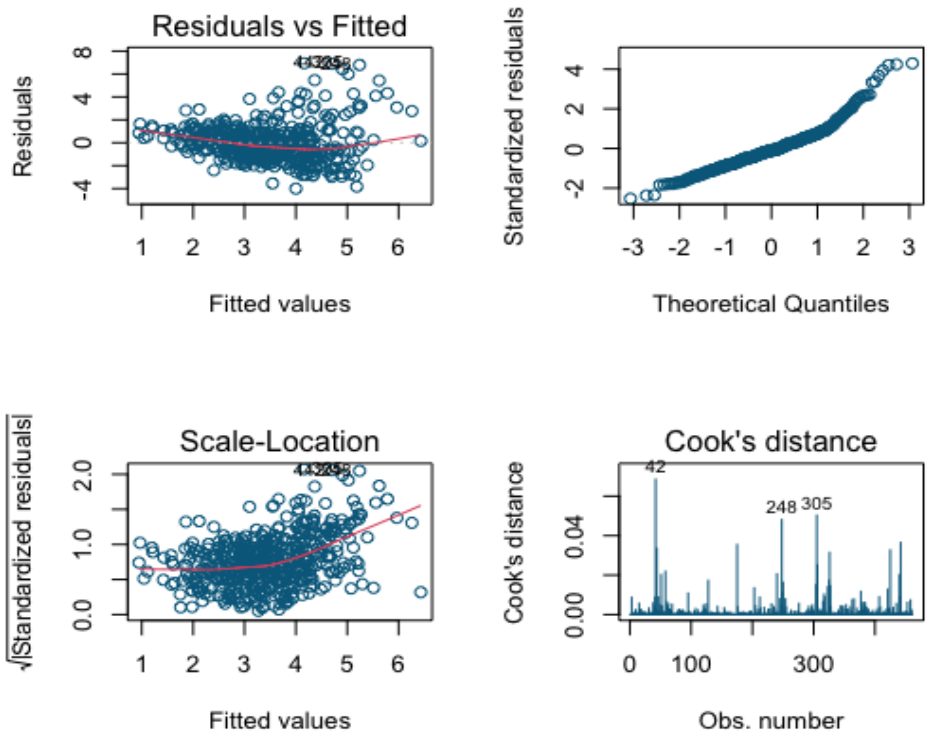


Figura 2: Gráficos para verificar si hay homocedasticidad y normalidad de los errores, además de ver si existen outliers.