I

00324

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

6

FACULTAD DE CIENCIAS

"ANALISIS ESTADISTICO DE UNA POBLACION NEONATAL DE UN HOSPITAL PRIVADO DE LA CIUDAD DE MEXICO"

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

RUBICEL AUSTRIA CRUZ

DIRECTOR DE TESIS: M. en C. JOSE ANTONIO FLORES DIAZ

DIVISION DE ESTUDIOS PROFESIONALES

MEXICO, D. F. 2003

FACULTAD DE CIENCIAS
SECCION ESCOLAR

PAGINACION DISCONTINUA

I-A

UNIVERSIDAD NACIONAL
AUTÓNOMA DE
MÉXICO

DRA. MARÍA DE LOURDES ESTEVA PERALTA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito: "Análisis estadístico de una población neonatal de un hospital privado de la Ciudad de México"

realizado por Rubicel Austria Cruz

con número de cuenta 09324693-8 , quién cubrió los créditos de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	M. en C. José Antonio Flores Díaz
Propietario	M. en A.P. María del Pilar Alonso Reyes
Propietario	Act. Jaime Vázquez Alamilla
Suplente	Act. Marypaola Janett Maya López
Suplente	Act. Gerardo Lucio Chávez Heredia

Consejo Departamental de Matemáticas

M. en C. JOSE ANTONIO FLORES DIAZ

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL
MATEMÁTICAS

I-B

Si uno logra medir lo que está diciendo y lo puede expresar en números, es que sabe lo que dice; pero si no lo puede expresar con números es que el conocimiento que tiene de ello es escaso e insatisfactorio.

- Lord Kelvin

AGRADECIMIENTOS

Expreso mi entero agradecimiento a la Universidad Nacional Autónoma de México, porque me brindó sus aulas y bibliotecas para mi formación académica.

Agradezco a la maternidad del Hospital Español, Sociedad de Beneficencia Española IAP y al Instituto Nacional de Ecología por el interés y asistencia que demostraron en la preparación del presente trabajo.

Mi distinguido reconocimiento al Maestro en Ciencias José Antonio Flores Díaz, por brindarme su conocimiento y el apoyo incondicional.

DEDICATORIA

A mis padres:
Irene Amparo Cruz Herrera
y Rubisel Austria Melo
porque sin ustedes no lo
hubiese logrado

A mis hermanos:
Gloriella, Areté,
Lasubellali y Rodolfo
por su aliento y paciencia
que siempre me brindan

A mis sobrinos:
José Alfredo, Irving Adán
Steve y Rubicel,
por ser la inspiración de
este trabajo.

TESTIMONIO DE GRATITUD

Gracias a todas aquellas personas que me ofrecieron su orientación, y tiempo para continuar el proyecto.

CONTENIDO

Página

INTRODUCCION

El peso al nacer refleja el estado genético y ambiental del bebé, ya que es el punto final del crecimiento intrauterino; tiene un fuerte impacto en la supervivencia neonatal, infantil y posterior, así como en la salud, el crecimiento y el desarrollo. La masa corporal del recién nacido depende de la placenta, de los factores maternos y fetales, así como de una sucesión de influencias constitucionales y ambientales.

Un pronóstico importante para la mortalidad y la morbilidad neonatal es cuando el bebé no alcanza los 2500 gramos en su masa corporal, conocido como recién nacido de bajo peso. Entre las causas principales de este riesgo se encuentran la edad y la talla materna, el consumo de cigarrillos durante el embarazo y el peso de la progenitora antes de la gestación, sin embargo la polución aérea no es considerada generalmente como un posible determinante del resultado del embarazo y pocos han sido los estudios en foros internacionales que han demostrado una asociación entre el bajo peso al nacer y la polución aérea tal es el caso de la República Checa, China y los Estados Unidos de Norteamérica. Los resultados arrojados por los estudios de esta índole son de suma importancia, puesto que los efectos adversos de la contaminación del aire en el peso al nacer resultan irreversibles e incluso concluyen en la muerte neonatal o postnatal. Aunque los datos no puedan usarse para apoyar o desaprobar una inferencia causal entre la contaminación del aire y los resultados adversos del embarazo, arrojó algunas asociaciones que fomentan la investigación.

La contaminación del aire en la Ciudad de México es quizás el problema ambiental de mayor relevancia para la población en general; su notoriedad, aunada a los padecimientos comunes que causa, ha traído como consecuencia una creciente conciencia de su peligrosidad y de la necesidad de resolverla, y debido a que esta es severa y con tendencia a agravarse, resulta impostergable iniciar investigaciones sobre el peso al nacer presuntamente relacionadas con el ambiente, pero debido a la carencia de estudios en el país que respalden la hipótesis de que está relacionado con el impacto de la contaminación ambiental, se sugirió realizar en el presente trabajo una investigación en la Zona Metropolitana del Valle de México, con una población neonatal de la que se tienen datos de suma importancia y se cuenta con los índices de polución aérea de dicha área de estudio. Se encontraron elementos suficientes para sugerir una relación significativa entre el bajo peso al nacer y las partículas suspendidas totales, por lo que se presentó un análisis a fin de determinar una relación estadística descriptiva entre estos elementos, así como la de examinar la oportunidad e intensidad de exposición a este contaminante durante el embarazo en un grupo bien caracterizado, con el uso de interacción de las variables tales como el año de gestación del neonato, la talla del recién nacido y el promedio de la exposición materna durante los primeros treinta y cinco días de gestación a las partículas suspendidas totales, por lo que desde 1993 hasta 1997, todas las mujeres que concibieron a un hijo en la maternidad Hospital Español, Sociedad de Beneficencia Española IAP, fueron registradas y la información individual sobre ambos, madres e hijos, se recolectó en los expedientes clínicos y los datos de contaminación del aire se obtuvieron a través del Instituto Nacional de Ecología dependiente de la Secretaría del Medio Ambiente y Recursos Naturales.

Para el desarrollo del presente trabajo se implementó en primera instancia el paquete STATISTICA versión 5.0 para la construcción de las tablas de contingencia y los modelos de regresión logística y lineal, pero debido a las limitaciones que presentó, se requirió de un software que satisficiera las

necesidades del análisis, por lo que se utilizó STATA versión 7.0, con el cual se logró cumplir las expectativas del estudio.

El presente trabajo se estructuró con tres capítulos, las conclusiones, dos apéndices y un glosario, así como las referencias bibliográficas, hemerográficas e información obtenida del internet. En el primer título se plasmaron los conceptos generales de la pediatría tales como el bajo peso al nacer, sus causas y consecuencias, así como los de la polución aérea; los cuales fueron consultados de la literatura médica y ambiental respectivamente. En la segunda sección se desarrolló la teoría de las tablas de contingencia y de los modelos de regresión lineal y logística, herramientas implementadas en el presente estudio. En el tercer capítulo se asentó la metodología seguida y las metas obtenidas, mostrando el modelo estadístico que describiera la relación existente entre el peso al nacer y las partículas suspendidas totales, así como los comentarios adicionales referentes a la investigación. En los apéndices se incluyó la demostración de los resultados, los programas de computación, los planos, las gráficas y las tablas que fueron complemento de este proyecto, así como otros resultados del marco teórico. En el glosario de términos se integraron conceptos y definiciones médicas y ambientales que auxilian el pleno entendimiento de la tesis. Se consultaron las referencias médicas, ambientales y de estadística aplicada, así como algunos artículos de la red de información.

CAPITULO I

CONCEPTOS GENERALES

1.1. Pediatría

1.1.1. Periodos de crecimiento

El desarrollo prenatal del cuerpo humano se divide en tres períodos principales:

1. Durante los primeros veintiún días de la vida prenatal, se forman las membranas fetales y aparecen las capas germinales; se caracteriza por una intensa multiplicación celular con un escaso aumento del tamaño del embrión.
2. En la etapa embrionaria, que corresponde desde el comienzo de la cuarta hasta el final de la octava semana, se produce una diferenciación y rápido crecimiento, se establecen los sistemas y los órganos principales del cuerpo y la mayor parte de las características de la forma externa. En esta fase el ser humano es muy sensible a cierto tipo de factores adversos tales como las radiaciones, las drogas, el alcohol, las enfermedades infecciosas, que afectan el desarrollo de los órganos y pueden producir malformaciones congénitas, sin embargo el feto no es muy susceptible a la desnutrición materna.
3. En el período fetal, que se extiende desde el final del segundo mes hasta el nacimiento, existe un incremento absoluto y rápido más que de notable diferenciación; las modificaciones de la forma externa del cuerpo se

producen con gran lentitud, a través de los pequeños cambios de la velocidad de crecimiento relativo de los distintos segmentos y partes del cuerpo. También se caracteriza por el acelerado incremento de longitud que es especialmente apreciable durante el tercer, cuarto y quinto mes (5 centímetros por mes aproximadamente), en tanto que el aumento de peso es más llamativo durante los últimos sesenta días de la gestación (aproximadamente 700 gramos por mes) y además de la maduración de los sistemas orgánicos.[1]

1.1.2. Evaluación al nacer

La clasificación del neonato permite inferir sus riesgos de enfermedad, de muerte y de posibles secuelas y obliga a establecer el tratamiento adecuado de inmediato.

1.1.2.1. Calificación de Apgar

Se debe tener precaución al interpretar las calificaciones de Apgar bajas en los niños de muy bajo peso al nacer, menores a 1000 gramos, ya que con frecuencia las presentan, debido en parte a su gran inmadurez, lo cual impide que tengan un tono muscular normal, también porque son pequeños y los predispone a que el choque sea más grave, pero a menudo no están asociados con el aumento en las frecuencias de morbilidad y letalidad.[2]

1.1.2.2. Cociente peso-semanas de gestación

La relación entre el peso y las semanas de gestación, fue calculada a través del cociente:

$$\frac{100 * (\text{peso en gramos})}{(\text{semanas de gestación})^{3}} \qquad [1.1]$$

---

[1] **MOORE**, Keith L., Persaud T.V.N., Shiota Kohei. Atlas de embriología clínica. Traducción de Color atlas of clinical embriology. Traducido por Pérez de Miguelsanz J. Primera Edición. Editorial Médica Panamericana. España. 1996. págs. 115-117.

[2] **MOORE**, op. cit. pág. 120.

Su rango de valores aceptable es de 4.0 a 7.0 y aquellos neonatos cuyo coeficiente peso-semanas de gestación oscila en valores menores a 4, corresponden a aquellos bebés cuyo peso en gramos es relativamente menor a su edad gestacional, y mientras más se acerque este valor al número 1 será mayor el grado de afección reflejado en el peso del recién nacido.

En caso contrario, los neonatos cuyo coeficiente peso-semanas de gestación oscila en valores mayores que 7, corresponden a aquellos niños cuyo peso en gramos fue mayor al relacionado con su edad gestacional.[3]

1.1.2.3. Cociente peso-talla

La relación entre el peso y la longitud del neonato fue calculada a través del cociente:

$$\frac{100 * (\text{peso en gramos})}{(\text{talla en centímetros})^3} \qquad [1.2]$$

El cual ayuda a la identificación de las anomalías del crecimiento fetal; su rango de valores aceptable es de 2.0 a 3.0, así pues si el coeficiente es menor que 19.0 significa que el peso es pequeño en relación con la longitud del bebé, y si el valor es mayor a 28.0 implica que su masa corporal es mayor respecto a la talla del neonato.[4]

1.1.2.4. Por el peso y las semanas de gestación

Haciendo uso de la gráfica del percentil 10 y 90, ver apartado A.4.2., se puede agrupar al recién nacido por su peso de la siguiente manera:

1. Grandes para su edad gestacional: si está por encima del límite superior normal.
2. Recién nacidos con retardo del crecimiento intrauterino: si el recién nacido está por debajo del límite inferior normal.

---

[3] **HAMILTON**, William James, Mossman H. W., Embriología humana: Desarrollo prenatal de la forma y la función. Cuarta Edición. Editorial Intermédica. Argentina. 1973. pág. 205.
[4] **HAMILTON**, op. cit. pág. 206.

3. Si el bebé está dentro de ambos límites, es clasificado como adecuado para su edad gestacional.

También se pueden clasificar a los neonatos según la edad gestacional en la que nacen, por lo que puede ser recién nacidos de:

1. Menos de 36 semanas de gestación, denominados de pretérmino.
2. 36 a 42 semanas de gestación, llamados de término.
3. Más de 42 semanas de gestación, conocidos como postérmino.

Se ha establecido que a menor edad gestacional el riesgo de muerte es mayor, y que con ello se relacionan las dificultades del ajuste a las condiciones de vida extrauterina (la respiración, la temperatura, la situación cardiovascular, el aporte energético, la condición inmunológica y la regulación neurológica entre otras).

Cuando se evalúa a un recién nacido, es conveniente el uso de ambas formas su peso y su edad gestacional.

El neonato de 36 a 42 semanas de gestación tiene un peso promedio de 3400 gramos con una fluctuación que va de 2500 a 4000 gramos y la talla media es de 50 centímetros.[5]

1.1.3. Bajo peso al nacer

1.1.3.1. Definición y concepto

La Organización Mundial de la Salud adoptó en 1950, la figura de menos de 2500 gramos como una definición universal del bajo peso al nacer sin importar cual sea su edad gestacional.

El recién nacido de bajo peso, en América Latina, tiene una frecuencia de alrededor del 10% de todos los nacimientos y está presente en aproximadamente el 70% de los niños que mueren en el periodo neonatal.[6]

---

[5] **HAMILTON**, op. cit. págs. 210-212.
[6] **MOORE**, op. cit. pág. 135.

1.1.3.2. Causas

Entre los probables factores que intervienen en el bajo peso están:

1. La edad de la madre: existe el riesgo cuando la gestante está en las edades extremas (menor de 15 ó mayor de 45 años); ya que el tiempo óptimo para la reproducción oscila entre los 20 y los 35 años.
2. El peso y la talla de la madre: se ha observado en que la mayoría de las progenitoras con estatura menor de 1.50 metros y peso inferior a 45 kilogramos, sus hijos al nacer son de peso bajo y la tasa de neonatos prematuros aumenta a medida que la longitud disminuye.
3. El embarazo múltiple: es la causa de productos de bajo peso al nacer en un número importante, ya que abarcan de un 12 a un 15% del total de estos casos, el riesgo de tener un recién nacido de bajo peso al nacer en embarazos dobles es alrededor de diez veces mayor que en producto único.
4. El número de partos anteriores al embarazo: las primíparas tienen un mayor porcentaje de nacidos de bajo peso aumentando su frecuencia después del segundo embarazo, e igualando o superando las grandes multíparas con gestaciones próximas.
5. El intervalo entre embarazos: se sugiere que el tiempo adecuado sea de 24 meses.
6. El control prenatal: el significado que puede tener el cuidado prenatal en relación con la producción de recién nacidos con peso bajo al nacimiento, es difícil de establecer ya que es frecuente que coincidan una serie de causas potenciales productoras del mismo fenómeno.

La atención médica será menos efectiva en clases sociales bajas, en las áreas en vías de desarrollo, en las primíparas jóvenes y en las grandes multíparas. Indudablemente, ciertos factores patológicos del embarazo, que afectan el producto al nacimiento, pueden ser adecuadamente prevenidos con una buena atención médica.

7. La patología del embarazo: las complicaciones obstétricas del embarazo es la explicación del peso al nacimiento por debajo de 2500 gramos de un 10 a un 15%.

   Los factores como la historia de cesárea anterior, la toxemia, la anemia, las hemorragias, la placenta previa, la diabetes mellitus y la preeclampsia son las complicaciones que más a menudo se correlacionan con el peso bajo al nacer.

   Las infecciones durante el embarazo, ocasionadas principalmente por virus como la rubéola, ocasionan retardo en el crecimiento intrauterino, ya que la infección cruza la placenta e infecta el producto.

8. Los antecedentes genéticos de la madre: los defectos fetales que resultan de las enfermedades o de los factores ambientales hereditarios pueden limitar el desarrollo normal.

9. La nutrición de la madre: la relación de la alimentación materna durante el embarazo y el peso del producto al nacimiento no se ha precisado totalmente, solo se ha encontrado que el consejo dietético puede reducir el índice de frecuencia de los nacidos de menos de 2500 gramos.

10. El hábito de fumar: existe una asociación significativa entre el consumo de tabaco durante el embarazo y el bajo peso al nacer siendo directamente proporcional a la cantidad de cigarrillos consumidos.

11. El consumo de alcohol: la madre que ingiere grandes cantidades de alcohol durante el embarazo puede tener un recién nacido con un síndrome alcohólico fetal caracterizado por el retardo en el crecimiento intrauterino y los defectos congénitos; se ha demostrado que también la progenitora moderadamente bebedora tiene un riesgo mayor de tener hijos de bajo peso.

12. El empleo de drogas: existe suficiente evidencia de los efectos negativos del uso de drogas durante el embarazo; en este caso se puede aconsejar la abstención total y el tratamiento correspondiente en los casos que se detecten.

13. El trabajo de la madre: los conflictos laborales como la falta de apoyo social tangible, así como las trabajadoras con jornadas mayores a 50 horas semanales fueron identificados como factores de riesgo del bajo peso al nacer en los neonatos y entre los agentes de peligro destacan el empleo en las industrias, la postura, el esfuerzo físico maternal, el estrés, el ruido, la exposición a los abrasivos y si la progenitora conduce el automóvil.
14. La violencia en la madre: las progenitoras atacadas tienen cuatro veces más riesgo de tener productos de bajo peso en comparación con las no maltratadas teniendo una diferencia de 560 gramos menos las primeras de las últimas.
15. Los factores psicológicos: la ilegitimidad (hijos de madres solteras), o que viven en unión libre, angustiadas por su situación ya mencionada o por los aspectos económicos, las relaciones familiares inadecuadas, y otros problemas que constantemente tienen a la madre en tensión nerviosa han sido factores que influyen gradualmente en el nacimiento de los niños prematuros y de bajo peso, tal vez debido a consecuencias circulatorias con hipertensión de la progenitora.
16. Los factores socioeconómicos: la cultura médica que tenga la madre, la distribución geográfica, el número de personas en la familia y el grado de escolaridad están ligados con el nivel de desarrollo de cada país, así en los estados latinoamericanos, africanos y asiáticos, los índices de bajo peso, son 4 a 6 veces más elevadas que en países como los europeos, especialmente Suecia, Noruega, Suiza, Holanda y en americanos como los Estados Unidos.
17. Otros factores relacionados que se han encontrado son: los viajes largos sin confort, el número de escaleras para llegar al hogar, la ganancia excesiva o insuficiente y la reducción de peso durante la gestación y los abortos tardíos previos.[7]

[7] **MOORE**, op. cit. págs. 135-145.

1.1.3.3. Consecuencias

Algunos lactantes de bajo peso, no muestran ninguno de los signos clínicos de desnutrición al nacer excepto que el niño es liviano, corto de estatura y corren más riesgos de infección y de muerte que los niños de peso normal.

Entre las patologías que pueden presentar se encuentran:

1. Los problemas respiratorios: debido a que tienen problemas para adaptarse a la vida fuera del útero; si el almacenamiento de glucosa en el feto es bajo, existirá una deficiencia de energía que puede ser utilizada en caso de una emergencia, y si esta deficiencia es severa puede causar convulsiones.
2. La mortalidad por daño cerebral: debido a que tienen desajustes de sal, agua o un nivel bajo de azúcar en la sangre y el sangrado en el cerebro sucede en cerca de una tercera parte de lactantes de peso muy bajo al nacer; menos de 1000 gramos, por lo que presentan un riesgo significativamente mayor de morir en el periodo neonatal y los que sobreviven, de presentar mayores alteraciones del desarrollo neurológico y psíquico, es decir problemas de aprendizaje y de conducta con relación a los nacidos a término y con peso adecuado. Entre los nacidos de bajo la mortalidad es de 20 a 40 veces mayor y la morbilidad es de 10 a 15 veces más grande.
3. La ictericia: ya que sus hígados pueden ser lentos para comenzar a funcionar por sí solos, se vuelven amarillos y un problema severo puede conducir al daño cerebral.
4. La enterocolitis necrotizante: una inflamación severa del intestino que puede dar lugar a la muerte.
5. La temperatura baja: los neonatos de bajo peso al nacer quizás no tengan suficiente grasa para mantener una temperatura corporal normal, la cual puede causar cambios en la química sanguínea y un crecimiento lento.
6. La insuficiencia cardiaca: si la arteria grande llamada el arteriosus ductus deja que la sangre se desvíe de los pulmones incluso antes del nacimiento, produciría la muerte en el neonato.

7. La retinopatía: es un crecimiento anormal de los vasos sanguíneos en el ojo, que puede dar lugar a la visión pobre o a la ceguera.[8]

1.2. Contaminación en la Zona Metropolitana del Valle de México

1.2.1. Condiciones atmosféricas

La Zona Metropolitana del Valle de México está constituida por el Distrito Federal, 53 municipios del Estado de México y un municipio de Hidalgo, se encuentra a 2,240 metros a nivel del mar y ocupa una extensión de 7,860 km^2. El área urbana tiene una superficie de 3,000 km^2 y está habitada por más de 20 millones de personas; circulan aproximadamente 3 millones de vehículos a un promedio de 20 kilómetros por hora, y operan alrededor de 35,000 industrias cuyo total calculado de emisiones a la atmósfera es de 4,916,673 toneladas por año.

México experimenta procesos graves de contaminación ambiental, especialmente en la atmósfera de la Zona Metropolitana del Valle de México y de áreas urbanas circunvecinas. El origen de las principales emisiones contaminantes está en el quehacer del hombre, pero la magnitud del problema reside en la cantidad y variedad de las fuentes contaminantes, la complejidad de las reacciones físico-químicas que se generan en la atmósfera y la cantidad de individuos y entidades que deben participar en la solución. El inventario de emisiones es un instrumento estratégico de gestión ambiental ya que permite identificar quiénes son los agentes productores de contaminación y evaluar el peso específico de cada uno de los sectores; en términos generales, existe una relación entre el volumen de emisión de contaminantes y la calidad del aire en una cuenca atmosférica. Sin embargo, se debe recordar que en las grandes ciudades pueden presentarse variaciones bruscas en los niveles de contaminación de un día a otro, debido principalmente a cambios en las condiciones meteorológicas más que a cambios significativos en la emisión diaria de contaminantes (como fuente natural). Las emisiones a la atmósfera en la gran área urbana mexicana, de productos de la actividad industrial la representan en su mayoría la emisión de

[8] **MOORE**, op. cit. págs. 146-150.

dióxido de azufre y las partículas suspendidas (como fuente fija). Las fuentes móviles son representadas en forma mayoritaria por los vehículos automotores y sus concentraciones en el aire de la Ciudad de México son lo suficientemente elevadas que rebasa las recomendaciones de la norma de calidad del aire nacional y extranjera.[9]

1.2.2. Normas e índices de calidad del aire

El 23 de diciembre de 1994 la Secretaría de Salud publicó las Normas Oficiales Mexicanas para evaluar la calidad del aire ambiente, ver la tabla 1 del apartado A.5.1. Estos índices de calidad del aire establecen los niveles máximos permisibles de concentración de contaminantes que garantizan la protección de la salud de la población en general y también de los grupos más susceptibles, para lo cual se incorpora un margen de seguridad. Las normas son de observancia para las autoridades estatales y municipales que tengan a su cargo el desarrollo y la aplicación de los planes y programas de política ambiental y en particular de calidad del aire. Cabe mencionar que las normas de calidad del aire mexicanas son similares a las de otros países, en particular a las de los Estados Unidos de Norteamérica y Canadá.[10]

1.2.3. Red de monitoreo de la calidad del aire

El Sistema de Monitoreo Atmosférico de la Zona Metropolitana del Valle de México consta de dos redes, una red manual con 19 estaciones para el muestreo y la determinación de partículas suspendidas totales, además, en cinco se hace el muestreo tanto de partículas suspendidas de fracción respirable como partículas menores de 10 micras y formaldehído. La otra parte del sistema es la Red Automática, con 32 estaciones de monitoreo atmosférico, 10 estaciones

---

[9] **INSTITUTO NACIONAL DE ECOLOGÍA.** Centro Nacional de Investigación y Capacitación Ambiental. SEMARNAP. Primer informe sobre la calidad del aire en ciudades mexicanas 1998. Primera Edición. Editorial Dirección General de Gestión e Información Ambiental del Instituto Nacional de Ecología de la SEMARNAP. México. 1997. págs. 5-6.

[10] **INSTITUTO NACIONAL DE ECOLOGÍA.** op. cit. pág. 12.

micrometeorológicas, una torre meteorológica, un radar acústico y una ecosonda, ver la tabla 2 y el plano 1 de los apartados A.5.2. y A.4.1. respectivamente.[11]

1.2.4. Las partículas suspendidas totales

El término partículas suspendidas abarca un amplio rango de sólidos o líquidos sutilmente divididos que pueden estar dispersos en el aire, y que son de tamaño mayor que las moléculas simples (0.0002 micras de diámetro), pero menor que 100 micras, tienen una vida media en estado suspendido que va de unos cuantos segundos a varios meses.

Son una mezcla compleja de partículas líquidas y sólidas en el aire; pueden absorberse y llevar los agentes activos a que entren en contacto con los tejidos respiratorios y pueden ser por sí mismos irritantes en forma moderada y cuando se presentan en cantidades excesivas pueden sobrepasar los mecanismos de limpieza normales del sujeto.

Las partículas en suspensión pueden ser producidas por actividad volcánica, las tormentas de polvo o los vientos fuertes que soplan sobre terreno seco (erosión), por origen biológico (polen, esporas, quistes, incendios forestales, materias fecales), y también por las industrias (procesos de combustión), transformación de otros contaminantes y los vehículos.

Éste contaminante observa una tendencia decreciente en la mayoría de las estaciones, aunque el factor de tolerancia aún sigue siendo rebasado el mayor número de días al año.

También se apreciaron los diferenciales de la contaminación por zonas: comparativamente son más altas las concentraciones en el Nordeste y Sureste que en el Noroeste, Centro y Sudoeste de la Zona Metropolitana del Valle de México durante los años 1992 hasta 1997.[12]

---

[11] **INSTITUTO NACIONAL DE ECOLOGÍA.** op. cit. pág. 25

[12] http://www.lamolina.edu.pe/calidad_ambiental/monitoreoatm.html **Martínez, Ana Patricia. Romieu Isabelle. Centro Panamericano de Ecología Humana y Salud (ECO/OPS). Agencia de Cooperación Técnica de Alemania (GTZ). Departamento del Distrito Federal** Introducción al monitoreo atmosférico. **Capítulo 1: Introducción. Actualizada el 03 de febrero de 1999.**

1.2.4.1. Los efectos en la salud

Los primeros estudios se enfocaron en episodios severos de contaminación aérea, incluyendo un estudio en el Valle de Meuse, Bélgica, en diciembre de 1930, un caso en Donora, Pennsylvania en 1948 y varios sucesos rígidos en Londres, siendo el más notable el que ocurrió en diciembre de 1952, durante el cual los niveles de dióxido de azufre y las concentraciones de humo se elevaron por arriba de 500 $\mu g/m^3$. Las personas que se vieron afectadas fueron, principalmente, aquellas con enfermedades cardiacas y pulmonares preexistentes, los ancianos y los niños menores de cinco años.

Posteriormente, la atención se fijó en los estudios dirigidos a las variaciones moderadas en la mortalidad cotidiana relacionada con los contaminantes dentro de las grandes ciudades y estos estudios epidemiológicos han proporcionado datos cuantitativos de los efectos sutiles sobre la salud asociados a las partículas suspendidas, en niveles comunes en ciudades de los países occidentales, y sugieren que un incremento de 10 $\mu g/m^3$ en las partículas menores de 10 micras se asocia con un incremento diario en la mortalidad de 0.5 al 1.5%; se observaron estos efectos en un amplio límite de niveles de partículas. Varios de los estudios de mortalidad cotidiana también proporcionaron información sobre mortalidad por categorías amplias de causa de muerte; se estimó un incremento en las muertes por problemas cardiovasculares de 0.8 a 1.8% en cada nivel de 10 $\mu g/m^3$ en las partículas menores de 10 micras, y un aumento en las muertes por enfermedades respiratorias con un efecto estimado entre 1.5 y 3.7%.

Se han observado asociaciones generalmente positivas entre la mortalidad y varias mediciones de contaminación por partículas y en especial con las finas; estos estudios estiman que del 2 al 9% de la mortalidad total se asoció con la contaminación por este contaminante.

También se estudió la diferencia en la mortalidad y la contaminación del aire, y este riesgo ajustado era aproximadamente del 15 al 25% mayor en las ciudades en los niveles más altos de contaminación por partículas finas, comparadas con las ciudades con los niveles más bajos, los resultados sugieren

que un incremento promedio de 10 $\mu g/m^3$ en la exposición a las partículas menores de 10 micras se asoció con un incremento de 3% o más en la mortalidad diaria y se observó la relación más fuerte con la enfermedad cardiopulmonar y las muertes por cáncer pulmonar.

El complejo partículas suspendidas totales-dióxido de azufre ha sido encontrado como relacionado con episodios de enfermedad respiratoria aguda especialmente en niños y se sabe que agrava las enfermedades pulmonares preexistentes. El dióxido de azufre pasa del aparato respiratorio a la corriente sanguínea y puede entonces difundirse por todo el cuerpo, donde parece que se elimina a través de las vías urinarias. Las partículas menores de 10 micras incorporan agua, se expanden en el aparato respiratorio y se depositan de un modo diferente del que podía esperarse por sus diámetros cuando se encontraban en el aire, permanecen en la atmósfera durante más tiempo que las partículas más grandes, y son responsables de la reducción de la visibilidad y toman parte en reacciones con otros contaminantes atmosféricos.

Otros estudios han evaluado los efectos de la morbilidad aguda por la contaminación por partículas mediante el examen de las asociaciones a corto plazo de las admisiones hospitalarias o de urgencias, así como cambios en la función pulmonar y los síntomas respiratorios. La mayor parte de los estudios de admisiones hospitalarias por enfermedades respiratorias han observado un incremento del 1 al 4% relacionado a un aumento en las partículas menores de 10 micras de 10 $\mu g/m^3$ en el día de la visita al hospital o 1 a 2 días antes.

En los estudios sobre consultas de urgencia por causas respiratorias los resultados sugieren que un incremento de 10 $\mu g/m^3$ en las partículas menores de 10 micras se relaciona con un incremento del 1 a 4% en las visitas a los servicios de urgencias. El efecto observado sobre los cambios agudos en la función respiratoria por la exposición a las partículas, con un incremento de 10 $\mu g/m^3$ en las partículas menores de 10 micras fue generalmente bajo (menor al 1%). Sin embargo, en los lugares donde el promedio de 24 horas de partículas menores de 10 micras excedía los 150 $\mu g/m^3$ se observaron decrementos de hasta el 7% en la función respiratoria.

Algunos estudios se han enfocado a poblaciones susceptibles, como los asmáticos, y en general, observaron un incremento en los síntomas asociados al asma en relación con las exposiciones a las partículas menores de 10 micras que iban del 1.1 al 11%.

Las concentraciones elevadas de dióxido de azufre pueden producir efectos severos en forma de broncoespasmos, bronquitis química y traqueitis, como se ve en la exposición ocupacional.

Existe una gran variabilidad en la sensibilidad a la exposición del dióxido de azufre entre los individuos, en especial los asmáticos. Los hallazgos de una gran variedad de estudios entre sofocados son consistentes en una relación lineal entre la magnitud del efecto (en términos de un incremento proporcional en la resistencia de las vías aéreas) y la dosis del dióxido de azufre proporcionada a las vías aéreas. Se ha relacionado la exposición a largo plazo con la bronquitis crónica, especialmente en los fumadores de cigarros.[13]

1.2.4.2. Monitoreo de la calidad del aire

Para el muestreo del material sólido que flota en el aire ambiente, se utiliza el método de alto volumen, que consiste en hacer pasar un flujo de aire a gran velocidad, a través de un medio filtrante de fibra de vidrio en el que se retienen las partículas con diámetros dinámicos de entre 0.1 y 100 micrones. En este método es absolutamente indispensable mantener el control y tener conocimiento de la tasa de flujo y del volumen total de aire que se ocupó en el muestreo durante las 24 horas que es, por lo regular, el periodo recomendado para la toma de las muestras. También se requiere conocer el peso del filtro antes y después del muestreo, por lo que este se acondiciona durante 24 horas en una cámara, donde se controla la temperatura y la humedad relativa. Posterior a la determinación de la masa material, la muestra es susceptible de someterse a

[13] http://www.lamolina.edu.pe/calidad_ambiental/monitoreoatm.html op. cit. Capítulo 7: Efectos de la contaminación del aire en la salud.

análisis físico-químicos para el análisis de plomo y otros metales pesados, así como de sulfatos y nitratos.[14]

[14] **INSTITUTO NACIONAL DE ECOLOGÍA.** op. cit. pág. 81

CAPITULO II

MARCO TEÓRICO

2.1. Variables categóricas

2.1.1. Definición y concepto

Para precisar el concepto de variable categórica es necesario tener presente tres ideas fundamentales:

La variable es un símbolo que puede representar cada uno de los elementos (números, vectores, funciones, en otros) de un conjunto y que se utiliza para definir el cambio de un fenómeno o simplemente una correspondencia funcional. La escala de medida estima la variación del fenómeno, es decir es un sistema de clasificación de variables. La categoría proporciona la clasificación de los elementos del conjunto en estudio.

Por lo que una variable categórica es aquella cuya escala de medida es un conjunto de categorías o clasificaciones.[15]

2.1.2. Clasificación

2.1.2.1. Por su escala de medida

Las variables categóricas se clasifican en:

[15] **AGRESTI, Alan.** Categorical data analysis. **Primera Edición. Editorial New York John Wiley & Sons. Estados Unidos de América. 1990. pág. 2.**

1. Nominales las cuales carecen de ordenamiento o este es irrelevante para el análisis estadístico, no existe el orden, la distancia, la proporcionalidad y el cero.
2. Ordinales que tienen niveles indexados, pero la distancia absoluta entre las categorías es desconocida; se plasma como operación empírica básica la idea de concluir si una variable ordinal es más grande que otra, no existe la distancia, ni la proporcionalidad ni el cero pero sí el orden.
3. De intervalo que se determinan de la igualdad de diferencias entre niveles consecutivos, es decir la distancia numérica entre dos valores sucesivos es siempre la misma, por lo que tiene sentido sumar o restar los datos; no existe la proporcionalidad pero si el orden, la distancia y el cero, aunque este último no es absoluto.
4. De razón que tienen como operación empírica básica la determinación de igualdad de los cocientes de niveles; además que introduce el concepto de punto natural que representa el origen de la medida, es decir un punto natural cero, que significa la ausencia de la característica en estudio, lo cual permite comparar los datos por medio de su cociente; existe el orden, la distancia, la proporcionalidad y el cero es absoluto.

2.1.2.2. Cualitativas y cuantitativas

Una variable categórica es cualitativa si los niveles difieren en calidad y a esta clase pertenecen las nominales y ordinales mientras en las cuantitativas se diferencian en cantidad, como son las de razón y de intervalo.

2.1.2.3. Exhaustivas y categorías mutuamente excluyentes

Una clasificación es exhaustiva cuando ofrece suficientes categorías para acomodar a todos y cada uno de los miembros de una población y un conjunto de categorías es mutuamente excluyente cuando cada miembro de una población puede ser ubicado en una y sólo una categoría.

2.1.2.4. Dicotómicas

Son aquellas que únicamente toman dos valores distintos.

2.1.2.5. Continuas y discretas

Una variable categórica es continua si toma cualquier valor en un rango específico, y las discretas tienen valores específicos y a esta clase pertenecen las nominales y ordinales; las de intervalo y de razón pueden pertenecer a cualquier clase.[16]

2.2. Tablas de contingencia

2.2.1. Definición y concepto

Al clasificar un objeto en ambas variables categóricas, se tienen J x K posibles combinaciones de ordenación, por lo que el elemento se sujeta a una función de distribución de probabilidad; y si se colocan los valores que toma en un arreglo de J renglones y K columnas para las variables X y Y respectivamente, entonces las entradas de esta matriz representan los J x K resultados.

Así pues, una tabla de contingencia es un arreglo rectangular que contienen frecuencias de resultados en las celdas elementales y los datos observados denotados como:

$$n_{jk} \qquad [2.1]$$

pertenecen a la j-ésima y k-ésima categoría de la primera y segunda clasificación respectivamente.

Estas frecuencias pueden ser transformadas en proporciones o porcentajes y así estos valores son denotados como:

$$\pi_{jk} \qquad [2.2]$$

que es la probabilidad de que el objeto escogido en forma aleatoria de la población pertenezca a la celda del renglón j y la columna k, es decir la distribución conjunta

[16] **AGRESTI**, Alan. op. cit. págs. 2-4.

de X y Y; por tal razón las categorías de las variables deben ser exhaustivas y mutuamente excluyentes.

Las distribuciones marginales en una tabla de contingencia de J x K son las sumas sobre todos los renglones y columnas de las probabilidades conjuntas, donde:

$$\pi_{j+} = \sum_{k=1}^{K} \pi_{jk} \qquad [2.3]$$

denota la probabilidad marginal para la variable renglón:

$$\pi_{+k} = \sum_{j=1}^{J} \pi_{jk} \qquad [2.4]$$

para la variable columna.[17]

2.2.2. Asociación entre respuestas

2.2.2.1. Definición y concepto

Se analiza una población en dos variables categóricas X y Y, cada una de las cuales tienen J y K niveles de respuesta respectivamente y en sus tablas de contingencia se satisface que para todo valor de j:

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{jk} = 1 \qquad [2.5]$$

2.2.2.2. Distribución condicional

Sean X y Y dos variables dependiente e independiente respectivamente, para un X = x, la distribución condicional del nivel k de Y en el de j para X es denotado por:

[17] **AGRESTI**, Alan. op. cit. págs. 8-9.

$$\pi_{k|j} = \frac{\pi_{jk}}{\pi_{j+}} \qquad [2.6]$$

2.2.2.3. Independencia

Las variables son estocásticamente independientes sí:

$$\pi_{jk} = \pi_{j+}\pi_{+k} \qquad [2.7][18]$$

2.2.2.4. Probabilidades muestrales

Sean p_{jk} la forma de denotar la probabilidad de la distribución conjunta muestral de una tabla de contingencia, y:

$$n_{++} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} \qquad [2.8]$$

la de referirse al tamaño total muestral, se define:

$$p_{jk} = \frac{n_{jk}}{n_{++}} \qquad [2.9]$$

y de aquí que:

$$p_{k|j} = \frac{n_{jk}}{n_{j+}} \qquad [2.10][19]$$

2.3. Momios

2.3.1. Definición y concepto

Sean A y B dos eventos, se define el momio de que B ocurrirá cuando el evento A está presente como:

[18] **AGRESTI**, Alan. op. cit. págs. 9-10.
[19] **AGRESTI**, Alan. op. cit. pág. 41.

$$\Omega_A = \frac{P(B|A)}{P(B^c|A)} \qquad [3.1]$$

en términos más sencillos, el momio es el cociente de la probabilidad, p, de que ocurra cierto evento y de la probabilidad de que este no ocurra, (1 - p), es decir:

$$\Omega = \frac{p}{1-p} \qquad [3.2]^{20}$$

2.3.2. Propiedades

1. La función $\Omega(p) = \frac{p}{1-p} > 0$ si $p \in [0,1)$.

2. $\Omega(p) = \begin{cases} < 1 & p < ½ \\ 1 & p = ½ \\ > 1 & p > ½ \end{cases}$

3. De la función $\Omega(p) = \frac{p}{1-p}$ se obtiene que:

   $$\lim_{0 < \varepsilon \to 0} \Omega(1-\varepsilon) = \lim_{0 < \varepsilon \to 0} \frac{1}{\varepsilon} - 1 = +\infty$$

   y que:

   $$\lim_{0 < \varepsilon \to 0} \Omega(1+\varepsilon) = \lim_{0 < \varepsilon \to 0} -\frac{1}{\varepsilon} - 1 = -\infty$$

   Con base a que los límites no son finitos, existe una discontinuidad de segundo orden en el punto p = 1, es decir, es una asíntota vertical.

4. Desde que $\Omega'(p) = \frac{1}{(1-p)^2} > 0$, se deduce que $\Omega(p)$ es una función monótona creciente en el intervalo [0,1).

5. Dado el valor del momio Ω es posible calcular el valor de la probabilidad por medio de la función:

   $$p = \frac{\Omega}{\Omega + 1} \qquad [3.3]$$

---

[20] **AGRESTI**, Alan. op. cit. pág. 14.

y considerando que $\lim_{\Omega \to 0} \frac{\Omega}{\Omega+1} = 0$, significa que a valor de Ω muy pequeños corresponden valores de p pequeños.

6. La función primitiva del momio está dada por:

$$\int \Omega(p)\, dp = \int \frac{p}{1-p}\, dp = \ln\left[\frac{\exp(1-x)}{1-x}\right] \qquad [3.4]$$

7. La fórmula general para la derivada de orden k se escribe como:

$$\Omega^{(k)}(p) = (-1)^{k+1} k!\,(1-p)^{-(k+1)} \text{ para } k>1 \qquad [3.5]$$

que satisface:

$$-\binom{p}{k} \frac{\Omega^{(k)}(p)}{\Omega^{(k-1)}(p)} = \Omega(p) \qquad [3.6]$$

2.3.3. El momio en las tablas de contingencia

2.3.3.1. Definición

Sean X y Y dos variables categóricas con J y K niveles de respuesta respectivamente, el momio de que Y = k en vez de k' dado que la X = j es:

$$\Omega_{jkk'} = \frac{\pi_{kj}}{\pi_{k'j}} = \frac{\pi_{jk}}{\pi_{jk'}} \qquad [3.7]$$

para todo valor de j, $k \neq k'$.

Cuando son variables dicotómicas, el momio es el cociente entre la probabilidad de estar y no en una categoría.

2.3.3.2. Interpretación

Es el riesgo de que un individuo seleccionado al azar que es observado pertenezca en la categoría de interés en vez de caer en cualquier otra categoría.

Como $\pi_{jk'}, \pi_{jk'} \geq 0$ implica que $\Omega_{jkk'} \geq 0$ pero si $\Omega_{jkk'} > 1$ se tiene que $\pi_{k|j} > \pi_{k'|j}$ o bien, $\pi_{jk} > \pi_{jk'}$ y esto es que en el j-ésimo renglón la respuesta k es $\Omega_{jkk'}$ veces más probable que la respuesta k'.

Si $\Omega_{jkk'} < 1$ implica que $\pi_{jk} < \pi_{jk'}$ y esto es que en el j-ésimo renglón la respuesta k es $\Omega_{jkk'}$ veces menos probable que la respuesta k'.

En una tabla de contingencia, las variables están no asociadas si los momios condicionales son iguales o cercanos unos a otros y de aquí que también sean similares a los momios marginales.[21]

2.3.3.3. Momios muestrales

Para el caso de probabilidades muestrales se tiene que:

$$p_{jk} = \frac{n_{jk}}{n_{++}} \qquad [3.8]^{22}$$

2.4. Cociente de momios

2.4.1. Definición y concepto

Es una medida de riesgo o del grado de asociación entre un factor de antecedente y un resultado, es decir, es la forma básica de la variación para ser explicada.

2.4.2. El cociente de momios en las tablas de contingencia

Considerando una tabla de contingencia de J x K y tomando los momios

$$\Omega_{jkk'} = \frac{\pi_{jk}}{\pi_{jk'}} \qquad [4.1]$$

[21] **AGRESTI**, Alan. op. cit. pág. 15.
[22] **AGRESTI**, Alan. op. cit. pág. 41.

$$\Omega_{j'kk'} = \frac{\pi_{j'k}}{\pi_{j'k'}} \qquad [4.2]$$

para todo valor de $j \neq j'$ y $k \neq k'$, se define el cociente de momios como:

$$\theta_{jj'kk'} = \frac{\Omega_{jkk'}}{\Omega_{j'kk'}} = \frac{\pi_{jk}\pi_{j'k'}}{\pi_{jk'}\pi_{j'k}} \qquad [4.3]$$

Es una medida de asociación que permite comparar el comportamiento de dos poblaciones respecto a una variable categórica.

En particular si j' = j+1 y k' = k+1, $\theta_{j,j+1,k,k+1}$ se le conoce como el cociente de productos cruzados.

Generalmente no es posible obtener una medida de asociación para las tablas de contingencia de J x K a partir del cociente de productos cruzados que siempre se pueda interpretar; una forma de generalizarlo es dividir el arreglo de J x K en submatrices de 2 x 2 y calcularlo para cada una de ellas.

Las probabilidades marginales de estas subtablas también son variables aleatorias y de aquí que los cocientes de productos cruzados solo se interpreten cuando las categorías de las submatrices son independientes o si están asociadas en forma estricta perfecta o débil perfecta.

2.4.2.1. Propiedades

Para el caso de un cociente de momios $\theta_{j,j+1,k,k+1}$ en una tabla de contingencia de asociación entre respuestas se tienen las siguientes propiedades:

1. Es invariante ante el intercambio de renglones y columnas ya que está envuelto en un cuadrado, y éste no es alterado bajo dicha transformación. De aquí que no sea necesario tener configurada en su totalidad la tabla para conocer dicha medida de proporciones.
2. Si sólo se intercambian renglones o columnas entre sí, su valor es igual a $\frac{1}{\theta_{j,j+1,k,k+1}}$.

3. Se dice que es una asociación invariante bajo la transformación $\pi'_{jk} \rightarrow r_j c_k \pi_{jk}$ para cualquier conjunto de números positivos $\{r_j\}, \{c_k\}$ que preservan $\sum_{j=1}^{J}\sum_{k=1}^{K}\pi'_{jk} = 1$ donde π'_{jk} es una probabilidad normalizada.
4. Los cocientes de momios más grandes que 1 indican la covariación directa entre las variables, mientras que los cocientes de momios más pequeños que 1 indican una relación inversa.
5. Puede tomar valores desde 0 hasta ∞, por lo que el logaritmo natural es simétrico con respecto de 0, y de aquí que éste último corra desde -∞ hasta ∞.
6. En una tabla de contingencia de J x K de asociación entre respuestas existe independencia de las probabilidades si y solo si el cociente de dos momios no iguales entre sí es uno.
7. En una tabla de contingencia de J x K de variación de respuesta existe homogeneidad entre las probabilidades conjuntas de un renglón para todas las columnas si y solo si el cociente de dos momios no iguales entre sí es uno.

2.4.2.2. Interpretación

Si $\Omega_{jkk'} \geq \Omega_{j'kk'}$ implica que $\theta_{jj'kk'} > 1$, esto indica que el momio de respuesta k es $\theta_{jj'kk'}$ veces más probable en el renglón j que en el j'.

Si $\Omega_{jkk'} \leq \Omega_{j'kk'}$ implica que $\theta_{jj'kk'} < 1$, esto indica que el momio de respuesta k es $\theta_{jj'kk'}$ veces menos probable en el renglón j que en el j'.

Si el cociente de momios es cero o indeterminado, se debe a que las variables están asociadas en forma perfecta estricta si $\pi_{jk} = \pi_{jk'} = 0$; o bien que estén relacionadas en forma perfecta débil si $\pi_{jk} = 0$ o $\pi_{jk'} = 0$.

En las tablas de contingencia, cualquier otro valor distinto de los anteriores puede interpretarse cuando las frecuencias marginales están fijas.[23]

2.4.2.3. Cociente de momios muestrales

Si se tienen propiedades muestrales, es decir que:

$$p_{jk} = \frac{n_{jk}}{n_{++}} \qquad [4.4]$$

entonces:

$$\Omega_{jkk'} = \frac{n_{jk}}{n_{jk'}} \qquad [4.5]$$

por lo que el cociente de momios esté dado por:

$$\theta_{jj'kk'} = \frac{n_{jk} n_{j'k'}}{n_{jk'} n_{j'k}} \qquad [4.6][24]$$

2.5. El modelo lineal general

2.5.1. Objetivos

La finalidad de la regresión lineal es la de:

1. Especificar un modelo de relación entre las variables a tratar.
2. Utilizar la información muestral acerca de los valores tomados por dichas variables, con el objeto de cuantificar la magnitud de la dependencia entre ellas.
3. Evaluar críticamente la validez de hipótesis propuestas acerca de las relaciones estimadas.
4. Efectuar un ejercicio de seguimiento coyuntural y de previsión de las variables analizadas.

---

[23] **AGRESTI**, Alan. op. cit. págs. 15-16.
[24] **AGRESTI**, Alan. op. cit. pág .42.

Los pasos que el analista sigue para especificar algún modelo de regresión lineal son:

1. Especificar claramente cual es el centro de atención del trabajo empírico.
2. Identificar cuales son los determinantes que explican la evolución de esta variable; por lo que se escoge con cuidado la información estadística relevante para cuantificar tal relación.
3. Se procede a su cuantificación.
4. Se utiliza el modelo de relación estimado, ya sea a efectos de contrastación de algún supuesto teórico, o como elemento de análisis y seguimiento de la variable cuyo comportamiento escogió explicar.

Así la intención del estudio es un modelo de relación entre las variables de interés, denotado como:

$$w = f(x_1, x_2, \ldots, x_k; u/\beta) \qquad [5.1]$$

el cual trata de explicar el comportamiento de una variable de interés utilizando la información proporcionada por un conjunto de k variables explicativas con un significado (biológico en este caso), así como por una parte aleatoria, no observable y, por consiguiente, sin significado conceptual, que se denota como u.

Las variables observables constituyen el vector **x**, de dimensión k x 1, o representado como una fila $\mathbf{x}' = (x_1, x_2, \ldots, x_k)$ (las variables explicativas toman diferentes valores que condicionan o afectan el resultado de la media para las observaciones w), y la relación de dependencia entre la variable explicada y el vector **x** envolverá, generalmente, un vector de parámetros que se denota por β (desconocido).

A fin de analizar empíricamente, es decir, utilizando datos reales, las características de la relación [5.1], se obtiene información muestral, que consiste en una lista ordenada de valores numéricos de las variables w, x_1, x_2,...., x_k.

En una muestra de sección cruzada, diversos agentes (biológicos en este caso) de una naturaleza similar proporcionan la información solicitada en un mismo instante de tiempo.

Considérese una muestra w de N observaciones bajo el supuesto que hayan sido extraídas de una distribución que tiene una media β y varianza σ^2,

también asúmase que los N resultados fueron tomados en forma independiente uno del otro, es decir, la variable aleatoria $w_t \sim (\beta, \sigma^2)$, y la independencia implica que la covarianza $E[(w_t - \beta)(w_s - \beta)] = 0$ para $s \neq t$.

De este modo, se dispone de una lista de relaciones:

$$w_t = f(x_{1t}, x_{2t}, \ldots, x_{kt}; u_t / \beta) \qquad t = 1,2,3,\ldots,T \qquad [5.2]$$

que relacionan los valores correspondientes w_t, x_{1t}, x_{2t}, ... x_{kt} que componen cada una de las T observaciones muestrales.

Entonces los modelos estadísticos son generalizados y escritos en una forma parametrizada y como el análisis de regresión trata con relaciones de dependencia lineal, es decir:

$$w_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \qquad t = 1,2,3,\ldots,T \qquad [5.3]$$

que se denomina modelo de regresión lineal múltiple o modelo lineal general.

En él los componentes del vector β son los coeficientes de las variables explicativas en el modelo lineal.

Dado que β es un parámetro, y bajo el supuesto distribucional de w_t, se tiene que $u_t = w_t - \beta x_t$, es la variable aleatoria u_t llamada término de error del modelo, entra aditivamente en el modelo y no precisa ir acompañada de ningún coeficiente. La variable w se denomina variable endógena, mientras que las x_1, x_2, ..., x_k son llamadas variables explicativas del modelo.

Cabe mencionar que los coeficientes β_1, β_2, ..., β_k recogen la magnitud del impacto de cada una de las variables explicativas sobre la variable endógena.

En algunas ocasiones, el modelo de relación incorpora un término constante:

$$w_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \qquad t = 1,2,3,\ldots,T \qquad [5.4]$$

que se interpreta como acompañado a una primera variable explicativa x_{1t} cuyo valor es siempre igual a 1.[25]

---

[25] **NOVALES**, Cinca Alfonso. "Econometría". Primera Edición. Editorial Mc. Graw Hill. México. 1988. págs. 52-54.

2.5.2. Características del modelo

1. El modelo es estocástico.

La presencia del término de error u hace que la relación entre la variable endógena y las explicativas sea estocástica, como contraposición a la posibilidad de que w hubiese dependido del vector **x** de modo determinista, es decir, a través de una relación funcional exacta.

Se considera que la relación es estocástica ya que:

El modelo [5.4] es sólo una aproximación al verdadero modelo de relación entre w y el vector **x**, que es mucho más complejo y de difícil especificación.

Gran parte de las variables biológicas de interés están sujetas a errores de medida, en muchos casos por inferirse su valor a partir de muestras finitas y, en otros casos, por no ajustarse exactamente al comportamiento biológico que el investigador deseó incorporar en su modelo de regresión lineal.

Se reconoce la posible existencia de otros factores determinantes del comportamiento de w que no se han incluido en el modelo, bien por desconocimiento de dichos factores, o porque no se dispone de observaciones numéricas.

2. El modelo es lineal.

El modelo que relaciona variables endógena y explicativas es lineal en los coeficientes $\boldsymbol{\beta}$, como ocurre en [5.4].

3. Los coeficientes del modelo $\beta_1, \beta_2, \ldots, \beta_k$ son constantes en el tiempo.

El problema de estimación consiste en utilizar la información muestral para asociar valores numéricos a estos k coeficientes.

Si, por el contrario, se permitiera que los coeficientes variasen en el tiempo, el problema de estimación sería más complejo.

4. Existe una relación causal desde las variables explicativas hacia la variable endógena.

Se entiende que la teoría del análisis de regresión subyacente aporta suficientes elementos para sugerir que las variables explicativas **x** influyen sobre la variable w, y no al revés, por lo que a las variables **x** se denominan variables exógenas, es decir, sus valores se toman como datos, ya que no reciben influencia alguna de la variable que se pretende explicar y que, por contraposición, recibe el calificativo endógena.

5. Las variables x no son linealmente dependientes.

Ésta es una de las características menos estricta, pues excluye tan sólo la posibilidad de que alguna de las variables explicativas del modelo lineal pueda escribirse como combinación lineal exacta de las demás.

Algunas variables biológicas muestran algún grado de correlación entre sí, y ello no produce excesivas dificultades, excepto cuando se llega a una situación de dependencia total, que es lo que se excluye al afirmar que las variables explicativas no son linealmente dependientes entre sí.

Si las variables explicativas no son todas independientes entre sí, se producen dificultades cuando se trata de estimar los coeficientes β_1, β_2, ..., β_k del modelo, es decir, de asociar valores numéricos a los mismos, sobre la base de la información muestral disponible, utilizando el procedimiento de mínimos cuadrados, ello se debe a que, al existir correlaciones elevadas entre algunas de las variables explicativas, es difícil desagregar su capacidad explicativa global en las componentes atribuibles a cada una de ellas, esta situación se conoce como multicolinealidad.

6. Las variables x son deterministas.

Si se tuviera la oportunidad de obtener otra muestra, además de la ya disponible, los valores de las variables explicativas serían los mismos que los ya observados en la muestra que ya se tienen, y este supuesto no incluye a la

variable endógena w, ya que es aleatoria y sus valores observados serían diferentes si se pudiera disponer de una muestra distinta.

7. Los errores son homocedásticos y no existe autocorrelación entre dos errores observados.

Se supone que la esperanza matemática del término de error u_t del modelo es cero, si, por el contrario se tuviera que $E(u_t)=a\neq 0$, este sería un efecto constante y, por ello, determinista sobre w_t, y debería incluirse como parte de la constate β_1 en [5.4]. Este supuesto propone una esperanza matemática nula para cada una de dichas variables aleatorias, ya correspondan a las distintas observaciones de sección cruzada de modo que se pueden recoger todos ellos formulado $E(\mathbf{u})=\mathbf{0}_T$, donde **u** denota el vector formado por todos los términos de error del modelo, así también se tiene que la varianza $E[u_t-E[u_t]]^2=E(u_t^2)=\sigma^2$ para t = 1,2,3,...,T, y además son variables aleatorias independientes e idénticamente distribuidas, por lo que la matriz de covarianzas del vector **u** definida por $E\{[\mathbf{u}-E(\mathbf{u})][\mathbf{u}-E(\mathbf{u})]'\}=E[\mathbf{uu}']$ se tiene que:

$$E[\mathbf{uu}']=E\left[\begin{bmatrix}u_1\\u_2\\\vdots\\u_T\end{bmatrix}\begin{bmatrix}u_1 & u_2 & \cdots & u_T\end{bmatrix}\right]$$

o bien

$$E[\mathbf{uu}']=E\begin{bmatrix}E[u_1^2] & E[u_1u_2] & \cdots & E[u_1u_T]\\E[u_2u_1] & E[u_2^2] & \cdots & E[u_2u_T]\\\vdots & \vdots & \ddots & \vdots\\E[u_Tu_1] & E[u_Tu_2] & \cdots & E[u_T^2]\end{bmatrix}$$

usando las hipótesis (que los términos del error no están correlacionados entre sí, y que la varianza del error del modelo es constante a través de la sección cruzada), se desprende que:

$$Var(\mathbf{u}) = E[\mathbf{uu}'] = E\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_T$$

Como **u** es una matriz de T x 1, entonces $E[\mathbf{uu}']$ es de dimensión T x T, esta matriz es igual a su transpuesta, se dice que es simétrica, y tienen $\frac{T(T+1)}{2}$ elementos diferentes, éste es un número que crece muy rápidamente con el número de observaciones disponibles, haciendo imposible la estimación de todos y cada uno de dichos parámetros, por lo que se hace un supuesto más sencillo de todos, haciendo que los elementos de la matriz sean función de un único parámetro, σ_u^2, por lo que ahora la matriz es escalar, ya que la diagonal principal es constante y los elementos de fuera de la diagonal de la matriz son iguales a cero.

De forma similar para el vector aleatorio w, se tiene que $E[\mathbf{w}] = E[\mathbf{x\beta} + \mathbf{u}] = \mathbf{x\beta}$ y la matriz de covarianzas está dada por $E[(\mathbf{w} - \mathbf{x\beta})(\mathbf{w} - \mathbf{x\beta})'] = E[\mathbf{uu}'] = \sigma^2 I_T$

Nótese que las matrices de covarianza $Var(\mathbf{u}) = \sigma_u^2 I_T$ y $E[(\mathbf{w} - \mathbf{x\beta})(\mathbf{w} - \mathbf{x\beta})']$ están relacionadas.

El supuesto $Var(\mathbf{u}) = \sigma_u^2 I_T$ puede abandonarse en cada una de las dos dimensiones:

Cuando la varianza del término de error es diferente para cada observación muestral, a esta situación se le llama heterocedasticidad.

Cuando, los términos de error correspondientes a periodos diferentes están correlacionados, a esta situación se le conoce como autocorrelación.[26]

---

[26] **NOVALES**, Cinca Alfonso. op. cit. págs. 55-59.

2.5.3. El estimador de mínimos cuadrados ordinarios del vector de parámetros β

2.5.3.1. Estimación

El primer objetivo del análisis de regresión es el de obtener estimaciones de los parámetros desconocidos del modelo, estos son, por un lado, los coeficientes **β** de las variables explicativas y, por otro, los que entran en la matriz de covarianzas del término del error. Estimar consiste en utilizar la información muestral para asignar valores numéricos a dichos parámetros.

Debe notarse que un estimador es, por tanto, una función del espacio muestral (el conjunto de todas las observaciones posibles que pudieron haberse tenido de las variables exógena y endógena) sobre el espacio paramétrico (el conjunto de todos los valores admisibles de los parámetros). Los distintos estimadores posibles difieren unos de otros en el modo de resumir la evidencia muestral para asociar valores numéricos a los parámetros del modelo.

Una vez estimados los coeficientes **β** se puede calcular para cada observación t:

$$\hat{w}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{2t} + \cdots + \hat{\beta}_k x_{kt} \qquad [5.5]$$

en el que las estimaciones $\hat{\beta}_j$ $1 \leq j \leq k$ han sustituido a los verdaderos valores, desconocidos.

La expresión [5.5] representa la estimación, de acuerdo con el modelo de regresión lineal, del valor que debía haber tomado la variable endógena w_t. Habrá siempre una discrepancia entre el valor realmente observado de w_t y la estimación anterior, a la que se denomina residuo correspondiente a dicha observación:

$$\hat{u}_t = \hat{w}_t - \hat{w}_t$$

De este modo se generara una serie de T residuos que, representados en forma matricial, como un vector T x 1, son:

$$\hat{\mathbf{u}} = \mathbf{w} - \hat{\mathbf{w}} = \mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}} \qquad [5.6]$$

Se procede a descomponer el problema de estimación de un modelo de regresión lineal en dos partes: en primer lugar se estiman los coeficientes $\boldsymbol{\beta}$; con ellos se obtiene el vector de residuos $\hat{\mathbf{u}}$ y, a partir de estos, se estiman los parámetros de la matriz de covarianzas.

Bajo el criterio que defina a un estimador sea la minimización de la magnitud de los residuos que dicho estimador genera, y tomando en cuenta que los residuos constituyen un vector T x 1, por lo que no se trata de minimizar un residuo determinado, sino una medida conjunta del tamaño global de todos ellos.

Dado un vector de estimaciones $\hat{\boldsymbol{\beta}}$, se procede a sumar los T residuos por él generados y escoger como estimación aquel vector $\hat{\boldsymbol{\beta}}$ cuya suma de residuos fuese la menor posible.

El estimador de mínimos cuadrados utiliza como criterio la minimización de la norma euclidiana del vector **u**, es decir, de la suma:

$$SR = \sum_{t=1}^{T} \hat{u}_t^2$$

que se denomina en lo sucesivo la suma residual, ésta también puede expresar en notación matricial como:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}}$$

siendo $\hat{\mathbf{u}}$ el vector T x 1 de residuos.

La suma residual es una función de las observaciones muestrales y de las estimaciones $\hat{\boldsymbol{\beta}}$, desde que:

$$\begin{aligned} SR(\hat{\boldsymbol{\beta}}) &= \hat{\mathbf{u}}'\hat{\mathbf{u}} \\ &= (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{w}'\mathbf{w} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{w} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned} \qquad [5.7]$$

Y derivando la forma cuadrática, se tiene que

$$\frac{\partial SR(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}'\mathbf{w} \qquad [5.8]$$

y la solución al problema de minimización de $SR(\hat{\boldsymbol{\beta}})$ requiere, en primer lugar, que este vector gradiente sea igual a cero, es decir, que:

$$\mathbf{X'X\hat{\beta} = X'w} \qquad [5.9]$$

Además debe cumplirse que la matriz de segundas derivadas o matriz hessiana de $SR(\hat{\boldsymbol{\beta}})$ sea definida positiva, la cual es:

$$\frac{\partial^2 SR(\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}'} = \mathbf{X'X}$$

y ésta es siempre semidefinida positiva.

Puesto que $\mathbf{X'X}$ es una matriz de k x k y $\mathbf{X'w}$ un vector k x 1, la ecuación matricial [5.9] es, un sistema de k ecuaciones lineales en los k coeficientes desconocidos $\beta_1, \beta_2, \ldots, \beta_k$. Este sistema se denomina sistema de ecuaciones normales y tiene, generalmente, solución única. Dicha solución al sistema de ecuaciones normales es el estimador de mínimos cuadrados ordinarios del vector $\boldsymbol{\beta}$, que se escribe como:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'w} \qquad [5.10]$$

Dicho estimador está definido de modo único siempre y cuando la matriz producto $\mathbf{X'X}$ sea invertible como puede verse en [5.10], lo cual ocurre siempre que:

Las k columnas de la matriz **X** son linealmente independientes, es decir, siempre que las k variables explicativas del modelo no sean linealmente dependientes entre sí, es decir, se disponga de al menos tantas observaciones como variables explicativas, es decir: $T \geq k$.

Para lograr precisión en la estimación mínimos cuadrática ordinaria es necesario disponer de un número de observaciones notablemente superior al de variables explicativas, es decir, $T > k$.

A la diferencia T-k se le conoce como número de grados de libertad de la estimación; cuando T=k, no se dispone de grados de libertad; las k observaciones determinan exactamente el valor numérico del estimador mínimo cuadrado ordinario, lo que conduce a una suma residual igual a cero.

De este modo, por un lado es mejor disponer de un número grande de observaciones muestrales para obtener mayor precisión y robustez en las

estimaciones; por otro lado, cada observación adicional generará un residuo más y, con ello, la suma residual tenderá generalmente a aumentar.

La intención del investigador consiste en disponer de un elevado número de observaciones y un modelo de regresión lineal que las represente suficientemente bien, de modo que ningún residuo, por sí solo, tenga una contribución importante a la suma residual.[27]

2.5.3.2. Propiedades

1. El vector de parámetro β es un vector aleatorio.

Depende del vector de observaciones de la variable endógena **w**, dependiendo también del vector de término de error **u**:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}) \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned} \qquad [5.11]$$

Nótese que en la expresión anterior se puede utilizar para definir el error de estimación como:

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

Proposición 5.1. Si $E(\mathbf{u})=\mathbf{0}_T$, entonces el estimador mínimo cuadrado ordinario es insesgado, es decir, $E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$.

Demostración:

De [5.11] se tiene que:

$$E(\hat{\boldsymbol{\beta}})=E\left(\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right)=\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}]=\boldsymbol{\beta}$$

Proposición 5.2. Si $Var(\mathbf{u})=\sigma_u^2\mathbf{I}_T$, la matriz de covarianzas del estimador de mínimos cuadrados ordinarios es igual a $Var(\hat{\boldsymbol{\beta}})=\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$.

Demostración:

---

[27] **NOVALES**, Cinca Alfonso. op. cit. págs. 62-67.

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= E\left(\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right) = E\left(\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right) \\ &= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}\mathbf{u}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_u^2\mathbf{I}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Estas dos propiedades implican que el estimador $\hat{\beta}_i$ de uno cualquiera de los coeficientes β_i tiene como esperanza matemática β_i, el verdadero valor del parámetro que se pretende estimar, y como varianza $\sigma_u^2 a_{ii}$, donde a_{ii} es el elemento i-ésimo en la diagonal principal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. Como debe ocurrir con toda matriz de covarianzas, se observa que $(\mathbf{X}'\mathbf{X})^{-1}$ es una matriz simétrica y definida positiva, puesto que su inversa $\mathbf{X}'\mathbf{X}$, también lo es.

La primera propiedad garantiza que, bajo los supuestos, el estimador de mínimos cuadrados ordinarios es insesgado. Ésta es una propiedad interesante si se asume que la ausencia de sesgo es una condición indispensable. Sin embargo, se debe matizar que el insesgo del estimador de mínimos cuadrados ordinarios no garantiza nada acerca de lo próximo o lejano que el valor numérico de este estimador se halla con respecto a los verdaderos valores de los coeficientes.

Que el estimador de mínimos cuadrados ordinarios sea insesgado quiere decir que: si se dispusiera de varias muestras diferentes de datos acerca de las variables w, x_1, x_2, ..., x_k y que se utilizan cada una de dichas muestras para calcular la estimación de mínimos cuadrados ordinarios correspondiente. Si el número de muestras es elevado, entonces el promedio de las estimaciones obtenidas con cada una de ellas será, aproximadamente, el verdadero valor de los coeficientes $\boldsymbol{\beta}$. Sin embargo, en la práctica se dispone de una sola muestra, lo que implica que se tiene una única realización del vector aleatorio $\hat{\boldsymbol{\beta}}$. Con una sola muestra, difícilmente se puede asegurar algo acerca de la distancia entre esa realización de $\hat{\boldsymbol{\beta}}$ y su esperanza matemática poblacional.

Por esta razón, el conocimiento de la matriz de covarianzas es fundamental. Cuando se trabaja con una variable aleatoria, su desviación estándar

proporciona, por su definición, un promedio de la distancia entre una observación cualquiera de dicha variable y su esperanza matemática. Cuando se trabaja con un vector aleatorio como $\hat{\boldsymbol{\beta}}$, estas ideas son válidas de un modo similar.

Por lo que sí se cuenta con una sola muestra, en tales casos que $\hat{\boldsymbol{\beta}}$ sea insesgado no es una propiedad excesivamente interesante, y se debe tratar de minimizar la distancia entre la única realización muestral y su esperanza matemática que es el vector de verdaderos valores de los parámetros $\boldsymbol{\beta}$.

Pero esa distancia viene medida por la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$, que se requiere sea tan pequeña como sea posible. En este caso, ello conduce a que tanto el valor del parámetro $\sigma_{\mathbf{u}}^2$ como el tamaño de la matriz $(\mathbf{X'X})^{-1}$ sean pequeños.

Por el contrario, cuando más próxima esté la matriz $\mathbf{X'X}$ a ser singular (es decir, más cercano a cero sea su determinante), mayor será el determinante de $(\mathbf{X'X})^{-1}$ y, con ello, mayor será la varianza del estimador de mínimos cuadrados ordinarios o, lo que es lo mismo, menor será su precisión.

Obsérvese que el error cuadrático medio del estimador de mínimos cuadrados ordinarios coincide con su matriz de covarianzas:

$$\mathrm{ECM}\left(\hat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left[\hat{\boldsymbol{\beta}}\right]$$

Proposición 5.3. Cada una de las variables explicativas es ortogonal al vector de residuos mínimo cuadráticos, es decir: $\mathbf{X'u} = \mathbf{0_k}$.

Demostración:

$$\begin{aligned}\mathbf{X'u} &= \mathbf{X'}\left(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\\ &= \mathbf{X'w} - \mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'w}\\ &= \mathbf{X'w} - \mathbf{X'w}\\ &= \mathbf{0_k}\end{aligned}$$

Proposición 5.4. Si hay un término independiente en la regresión, entonces la suma de los residuos mínimo cuadráticos es cero, es decir $x_{1t} = 1 \ \forall t \Rightarrow \sum_{t=1}^{T} \hat{u}_t = 0$.

Demostración:

En efecto, si una de las variables es constante, entonces basta aplicar la proposición anterior a esa variable en particular para obtener este resultado.

Proposición 5.5. Teorema de Gauss–Markov

El estimador de mínimos cuadrados ordinarios es una función lineal del vector de observaciones de la variable aleatoria **w**. Un estimador lineal genérico se escribe como:

$$\hat{\boldsymbol{\beta}}^* = \mathbf{f}(\mathbf{X}) + \mathbf{g}(\mathbf{X})\mathbf{w}$$

Utiliza como funciones **f** y **g** en la expresión anterior $\mathbf{f} = \mathbf{0}$ y $\mathbf{g} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Entonces es un estimador lineal insesgado óptimo, en el sentido de que cualquier otro estimador lineal e insesgado tiene una matriz de covarianzas mayor que la del estimador de mínimos cuadrados ordinarios.

Demostración:

Sea $\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{A}}\mathbf{w}$ un estimador lineal de $\boldsymbol{\beta}$, donde $\tilde{\mathbf{A}}$ es una matriz k x T y se define: $\mathbf{A} = \tilde{\mathbf{A}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, de modo que:

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \left(\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{w} = \left(\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta} + \left(\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{u}\end{aligned}$$

y, por tanto, $E\left(\tilde{\boldsymbol{\beta}}\right) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}$. El estimador $\tilde{\boldsymbol{\beta}}$ será insesgado sólo sí la matriz **A** es tal que $\mathbf{AX} = \mathbf{0}_{kxk}$. Con esta condición, el estimador $\tilde{\boldsymbol{\beta}}$ resulta:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{u}$$

y su matriz de covarianzas será:

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\boldsymbol{\beta}}\right) &= E\left(\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right) = E\left\{\left(\left(\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{u}\right)\left(\left(\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{u}\right)'\right\} \\ &= \mathbf{A}\mathbf{A}'\sigma_u^2 + (\mathbf{X}'\mathbf{X})^{-1}\sigma_u^2\end{aligned}$$

Donde se ha utilizado la condición de ausencia de sesgo que $\mathbf{AX} = \mathbf{0}_{kxk}$. Como la matriz **AA'** es semidefinida positiva, se concluye que la diferencia entre las matrices de covarianzas de $\hat{\boldsymbol{\beta}}$ y $\tilde{\boldsymbol{\beta}}$ es una matriz semidefinida positiva, por lo que la primera es igual, si no mayor, que la segunda.

Proposición 5.6. La suma residual se puede expresar como:

$$SR = \mathbf{u}'\mathbf{u} = \mathbf{w}'\mathbf{w} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{w} \qquad [5.12]$$

Demostración:

$$\begin{aligned} SR = \mathbf{u}'\mathbf{u} &= (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{w}'\mathbf{w} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{w} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{w}'\mathbf{w} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{w} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w} \\ &= \mathbf{w}'\mathbf{w} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{w} \end{aligned}$$

Proposición 5.7 La suma residual también se expresa como:

$$SR = \mathbf{u}'\mathbf{u} = \mathbf{w}'\mathbf{w} - \hat{\mathbf{w}}'\hat{\mathbf{y}} = \sum_{t=1}^{T} w_t^2 - \sum_{t=1}^{T} \hat{w}_t^2 \qquad [5.13]$$

es decir, la suma residual es la diferencia entre la suma de cuadrados de las observaciones y la suma de cuadrados de los valores de w_t implicados por el modelo, $\hat{w}_t$.

Demostración:

$$\hat{\mathbf{w}}'\hat{\mathbf{w}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{w}$$

y sustituyendo en la expresión [5.12] se obtiene lo que se pretende demostrar.

Proposición 5.8. El vector de residuos mínimo cuadrados es una transformación lineal del vector término de error:

Demostración:

Si $\mathbf{u} = \mathbf{w} - \mathbf{X}\boldsymbol{\beta}$ y $\hat{\mathbf{w}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ y sea:

$$\begin{aligned} \hat{\mathbf{u}} = \mathbf{w} - \hat{\mathbf{w}} &= \mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{w} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w} \\ &= \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{w} = \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right](\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{u} \\ &= \left[\mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right]\boldsymbol{\beta} + \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{u} \\ &= \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{u} \end{aligned}$$

Sea $\mathbf{M} = \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$ una matriz singular, simétrica e idempotente, entonces $\hat{\mathbf{u}} = \mathbf{Mu}$ es expresada como una función lineal de los errores aleatorios no observables.

Proposición 5.9. El vector de residuos mínimos cuadráticos tiene esperanza cero y matriz de covarianzas $\sigma_{\mathbf{u}}^2\mathbf{M}$.

Demostración:

$$E[\hat{\mathbf{u}}] = E[\mathbf{Mu}] = \mathbf{M}E[\mathbf{u}] = \mathbf{M0} = \mathbf{0}$$

$$Var[\hat{\mathbf{u}}] = Var[\mathbf{Mu}] = \mathbf{M}^2 Var[\mathbf{u}] = \sigma_{\mathbf{u}}^2\mathbf{M}$$ [28]

2.5.4. Suma total, explicada y residual

La suma total es la varianza muestral de la variable endógena multiplicada por T, por lo que es una medida del tamaño de las fluctuaciones experimentadas por dicha variable alrededor de su valor medio. El objeto fundamental de todo modelo de regresión lineal es tratar explicar dichas fluctuaciones y, por consiguiente, la suma total es la cantidad que se pretende explicar, y se denota como:

$$\text{Suma Total (ST)} = \sum_{t=1}^{T}(w_t - w)^2 \qquad [5.14]$$

La suma explicada es el grado de fluctuaciones de la variable $\hat{w}_t$ alrededor del promedio de w, donde es la variable que el modelo genera como w_t. Por lo tanto, la suma explicada es el nivel de fluctuación de la variable w_t que el modelo es capaz de explicar.

$$\text{Suma Explicada (SE)} = \sum_{t=1}^{T}(\hat{w}_t - w)^2 \qquad [5.15]$$

La suma residual, es un indicador del nivel de error del modelo en su intento de explicar la evolución temporal de la variable w_t.

---

[28] **NOVALES**, Cinca Alfonso. op. cit. págs. 62-73.

$$\text{Suma Residual (SR)} = \sum_{t=1}^{T}(w_t - \hat{w}_t)^2 \qquad [5.16]$$

Proposición 5.10. Si entre las variables explicativas hay un término constante, entonces se tiene que: ST = SE + SR.

Demostración:

Tomando la expresión [5.13], $\mathbf{u'u} = \mathbf{w'w} - \hat{\mathbf{w}}'\hat{\mathbf{w}}$, reagrupando los términos, se observa que: $\mathbf{w'w} = \hat{\mathbf{w}}'\hat{\mathbf{w}} + \mathbf{u'u}$ y que restando ambos miembros de la ecuación la cantidad Tw^2, se tiene:

$$\mathbf{w'w} - Tw^2 = \hat{\mathbf{w}}'\hat{\mathbf{w}} + \mathbf{u'u} - Tw^2 \qquad [5.17]$$

Por otro lado:

$$\begin{aligned} ST &= \sum_{t=1}^{T}(w_t - w)^2 \\ &= \sum_{t=1}^{T} w_t^2 - 2w\sum_{t=1}^{T} w_t + Tw^2 \\ &= \mathbf{w'w} - 2wTw + Tw^2 \\ &= \mathbf{w'w} - Tw^2 \end{aligned}$$

de modo que el miembro de la izquierda en [5.17] es precisamente la suma total.

Ahora bien, la suma explicada puede escribirse como:

$$\begin{aligned} SE &= \sum_{t=1}^{T}(\hat{w}_t - w)^2 = \sum_{t=1}^{T}\hat{w}_t^2 - 2w\sum_{t=1}^{T}\hat{w}_t + Tw^2 \\ &= \mathbf{w'w} - 2w\sum_{t=1}^{T}\hat{w}_t + Tw^2 \end{aligned} \qquad [5.18]$$

Por otra parte, puesto que $w_t = \hat{w}_t + \hat{u}_t$, se desprende que:

$$w\sum_{t=1}^{T}\hat{w}_t = w\sum_{t=1}^{T}w_t - w\sum_{t=1}^{T}\hat{u}_t = w\sum_{t=1}^{T}w_t = Tw^2$$

donde ya se utilizó el hecho de que cuando uno de los regresores es constante, entonces $\sum_{t=1}^{T}\hat{u}_t = 0$.

Sustituyendo en [5.18] se tiene que $SE = \mathbf{w}'\mathbf{w} - T\bar{w}^2$ y sustituyendo esta expresión en [5.17] se tiene finalmente el resultado: ST = SE + SR.[29]

2.5.5. Coeficiente de determinación

En un modelo lineal se define el coeficiente de determinación como la cantidad:

$$R^2 = 1 - \frac{SR}{ST}$$

mientras que su raíz cuadrada positiva, cuando existe, se denomina coeficiente de correlación entre $\mathbf{w}$ e $\hat{\mathbf{w}}$.

El coeficiente de determinación es denotado con R^2 que expresa la proporción muestral de la variabilidad en w que es explicada a través del modelo lineal

Proposición 5.11

El coeficiente de determinación es siempre menor o igual que 1.

Si una de las variables explicativas es constante, entonces $R^2 \geq 0$

Demostración:

Puesto que $ST \geq 0$ y $SR \geq 0$ entonces $\frac{SR}{ST} \geq 0$ y de aquí que $-\frac{SR}{ST} \leq 0$

por lo que $R^2 = 1 - \frac{SR}{ST} \leq 1$.

Como ST = SR + SE, se tiene que $ST \geq SR$, por lo que $\frac{SR}{ST} \leq 1$ y de aquí que $R^2 = 1 - \frac{SR}{ST} \geq 0$.

Si el modelo lineal tiene término independiente, el coeficiente de determinación puede calcularse mediante la expresión: $R^2 = \frac{SE}{ST}$.

Estas expresiones no son, sin embargo, válidas en general, por lo que no es correcto utilizar el cociente anterior, como definición del coeficiente de

[29] **NOVALES**, Cinca Alfonso. op. cit. págs. 73-74.

determinación. Del mismo modo, tanto la no negatividad de R^2 como la descomposición de la suma total sólo son válidas en tanto haya un término independiente en el modelo, y no en otro caso. En consecuencia, cuando no hay un término independiente en el modelo, entonces el coeficiente de determinación puede tomar cualquier otro valor, siempre menor o igual a 1, pero incluyendo en su rango todos los reales negativos (a pesar del exponente 2 con el que se denota). Dicho exponente es puramente una convención notacional, y no indica en absoluto que el valor del estadístico R^2 haya de ser necesariamente positivo. Por supuesto que cuando el coeficiente de determinación es negativo, entonces no es posible calcular el coeficiente de correlación lineal, definido como $\sqrt{R^2}$.

Por otra parte, incluso en casos en que R^2 resulte negativo, aún será posible utilizarlo para comparar el grado de poder explicativo de dos modelos, desde que la única diferencia entre dos modelos es la SR, se prefiere aquel modelo que tenga la menor suma residual, es decir, aquel con R^2 mayor, no importa que este estadístico sea positivo o negativo.

Si los dos modelos a comparar tienen un número distinto de variables explicativas, puede probarse que cuando se añade una variable a un modelo, entonces la suma residual siempre disminuye. Por tanto, si uno de los dos modelos contiene las mismas variables que el otro y alguna más (en cuyo caso los dos modelos se dicen anidados), entonces este modelo amplio sería siempre el preferido de acuerdo con el criterio del mayor R^2.

En el análisis de regresión se sugiere la utilización del estadístico conocido como R^2 corregido que se define como:

$$R^2 = 1 - \left(\frac{T-1}{T-k}\right)\left(1 - R^2\right) \qquad [5.19]$$

El interés de este estadístico reside en que, cuando el número de variables explicativas k aumenta, la fracción $\frac{T-1}{T-k}$ también aumenta, mientas que $1-R^2$ disminuye, ya que el coeficiente de determinación R^2 aumenta. En la

definición del R^2 corregido aparece el producto de estos dos factores, la idea es que ambos efectos, el creciente y el decreciente, se compensen aproximadamente, por lo que este estadístico sea una medida de la bondad de ajuste de un modelo de regresión lineal con la propiedad de ser neutral frente a la introducción de variables adicionales. Este estadístico está concebido, por tanto, para la comparación de modelos anidados.

Cabe mencionar que $\frac{T-1}{T-k} \leq 1$, por lo que el rango de valores del R^2 corregido es $[-\infty,1]$, igual que el rango de R^2.[30]

2.5.6. Estimación de σ_u^2

La matriz de covarianzas del estimador de mínimos cuadrados ordinarios es importante por dos razones:

1. Cuando sólo se dispone de una muestra para llevar a cabo la estimación del vector $\boldsymbol{\beta}$ entonces es crucial conocer la matriz de covarianzas del estimador para poder juzgar la exactitud con que la estimación obtenida se aproxima a su esperanza matemática $\tilde{\boldsymbol{\beta}}$.
2. La matriz de covarianzas es precisa para conocer las varianzas y covarianzas de cada elemento del vector $\hat{\boldsymbol{\beta}}$, de modo que se puedan hacer contrastes de hipótesis acerca de valores individuales, o de varios coeficientes, o de combinaciones lineales de ellos.

Pero al ser el parámetro σ_u^2 desconocido, la matriz de covarianzas $\boldsymbol{\beta}$ también es desconocida. Sólo se puede estimar previamente el parámetro σ_u^2 y multiplicar tal estimación por la matriz $(\mathbf{X'X})^{-1}$.

Haciendo uso que $\hat{\mathbf{u}} = \mathbf{Mu}$, por tanto se puede calcular el valor esperado de la suma residual, es decir $E[\hat{\mathbf{u}}'\hat{\mathbf{u}}] = E[\mathbf{u'Mu}]$, ya que $\mathbf{u'Mu}$ es un escalar, o bien,

[30] **NOVALES**, Cinca Alfonso. op. cit. págs. 75-78.

matriz 1 x 1, entonces: $E[\mathbf{u'Mu}] = E[tr(\mathbf{u'Mu})]$ y por las propiedades de la traza desde que es un operador lineal, se desprende que:

$$E[\hat{\mathbf{u}}'\hat{\mathbf{u}}] = tr[E(\mathbf{Muu'})] = tr[\mathbf{M}E(\mathbf{uu'})]$$
$$= tr\left[\mathbf{M}\sigma_u^2 \mathbf{I}_T\right] = \sigma_u^2 \mathbf{I}_T tr[\mathbf{M}]$$

y como:

$$tr[\mathbf{M}] = \left\{tr[\mathbf{I}_T] - tr\left[\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}\right]\right\}$$
$$= \left\{T - tr\left[\mathbf{X'X}(\mathbf{X'X})^{-1}\right]\right\}$$
$$= \{T - tr[\mathbf{I}_k]\} = (T-k)$$

entonces $E[\hat{\mathbf{u}}'\hat{\mathbf{u}}] = \sigma_u^2(T-k)$ y si $\hat{\sigma}_u^2 = E[\mathbf{u'u}]$, de este resultado se deduce que el conciente $\hat{\sigma}_u^2 = \frac{SR}{T-K} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}$ es un estimador insesgado del parámetro σ_u^2, puesto que se tiene:

$$E\left[\hat{\sigma}_u^2\right] = E\left[\frac{SR}{T-K}\right] = \frac{E[\hat{\mathbf{u}}'\hat{\mathbf{u}}]}{T-k} = \frac{1}{(T-k)}(T-k)\hat{\sigma}_u^2 = \hat{\sigma}_u^2.$$

Este cociente se conoce como estimador de mínimos cuadrados de la varianza del término de error σ_u^2.

Proposición 5.12. Sí $\mathbf{u} \sim N_T(\mathbf{0}_T, \sigma_u^2\mathbf{I}_T)$, entonces $\hat{\boldsymbol{\beta}} \sim N_k(\boldsymbol{\beta}, \sigma_u^2(\mathbf{X'X})^{-1})$

Demostración:

El estimador $\hat{\boldsymbol{\beta}}$ es una transformación lineal del vector aleatorio $\mathbf{u}$, ya que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{Au}$, donde $\mathbf{A} = (\mathbf{X'X})^{-1}\mathbf{X'}$, por lo que sí $\mathbf{u}$ sigue una distribución Normal, entonces $\hat{\boldsymbol{\beta}}$ también es una distribución Normal, cuya esperanza y varianza se demostraron en la proposición 5.1 y 5.2 respectivamente.

Nótese, sin embargo, la diferente dimensionalidad de las dos distribuciones normales.

Por otra parte, se observa que de la proposición 5.8 que el vector de residuos mínimo cuadrados ordinarios es una transformación lineal del vector de error $\mathbf{u}$, también se tiene que sí $\mathbf{u} \sim N_T(\mathbf{0}_T, \sigma_u^2\mathbf{I}_T)$ entonces:

$$\hat{\mathbf{u}} \sim N_T\left(\mathbf{0}_T, \sigma_u^2\mathbf{M}\right).^{31}$$

2.5.7. El estimador de máxima verosimilitud

Se ha adoptado el criterio de estimación, consistente en escoger los valores de los parámetros $\boldsymbol{\beta}$ y σ_u^2, de modo que se obtenga la menor suma de cuadrados de residuos posibles.

El método de máxima verosimilitud descansa sobre un determinado supuesto acerca del tipo de distribución seguido por el término de error del modelo de regresión lineal. Si el supuesto que se haga acerca de dicha distribución es aproximadamente correcto, se ganará eficiencia utilizando tal información adicional.

Dado el modelo de regresión lineal $\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, y bajo el supuesto que el vector $\mathbf{u}$ sigue una distribución Normal y si la esperanza es $\mathbf{0}_T$ y su varianza la matriz $\sigma_u^2\mathbf{I}_T$. Bajo estas hipótesis, la función de densidad del vector $\mathbf{u}$ es:

$$\begin{aligned} f(\mathbf{u}) &= \frac{1}{(2\pi)^{T/2}} \frac{1}{\left|\sigma_u^2\mathbf{I}_T\right|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{u}'\left(\sigma_u^2\mathbf{I}_T\right)^{-1}\mathbf{u}\right) \\ &= \frac{1}{(2\pi)^{T/2}} \frac{1}{\left(\sigma_u^2\right)^{T/2}} \exp\left(-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{u}\right) \end{aligned} \qquad [5.20]$$

La función de densidad anterior puede transformarse en la función de verosimilitud muestral si se expresa el vector $\mathbf{u}$ como función de las matrices $\mathbf{w}$ y $\mathbf{X}$. Calculando el jacobiano de la transformación que convierte el vector aleatorio $\mathbf{u}$ en el vector aleatorio $\mathbf{w}$ es la matriz:

$$\left(\frac{\partial \mathbf{u}}{\partial \mathbf{w}}\right) = \begin{pmatrix} \frac{\partial u_1}{\partial w_1} & \frac{\partial u_1}{\partial w_2} & \cdots & \frac{\partial u_1}{\partial w_T} \\ \frac{\partial u_2}{\partial w_1} & \frac{\partial u_2}{\partial w_2} & \cdots & \frac{\partial u_2}{\partial w_T} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial u_T}{\partial w_1} & \frac{\partial u_T}{\partial w_2} & \cdots & \frac{\partial u_T}{\partial w_T} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

[31] **NOVALES**, Cinca Alfonso. op. cit. págs. 78-80.

que tiene por determinante la unidad. En consecuencia, la función de verosimilitud se obtiene sustituyendo en [5.20] el vector **u** como función de **w**, para obtener:

$$L\left(\mathbf{w},\boldsymbol{\beta},\sigma_u^2\right)=\frac{1}{(2\pi)^{T/2}}\frac{1}{\left(\sigma_u^2\right)^{T/2}}\exp\left(-\frac{1}{2\sigma_u^2}(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})\right) \qquad [5.21]$$

El estimador de máxima verosimilitud de $\boldsymbol{\beta}$ y σ_u^2 está formado por aquellos valores de estos parámetros que maximizan la función de verosimilitud [5.21]. Por tanto, se trata de maximizar [5.21] con respecto a sus argumentos $\boldsymbol{\beta}$ y σ_u^2, para una muestra dada, y como esto es equivalente a maximizar el logaritmo neperiano de la función de verosimilitud:

$$\text{Ln } L\left(\mathbf{w},\boldsymbol{\beta},\sigma_u^2\right)=-\frac{T}{2}\text{Ln } 2\pi-\frac{T}{2}\text{Ln } \sigma_u^2-\frac{1}{2\sigma_u^2}(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})$$

e igualando a cero las derivadas con respecto a sus argumentos se tiene:

$$\frac{\partial \text{Ln } L}{\partial \boldsymbol{\beta}}=-\frac{1}{2\sigma_u^2}[-2\mathbf{X}'(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})]=\mathbf{0}_k$$

$$\frac{\partial \text{Ln } L}{\partial \sigma_u^2}=-\frac{T}{2\sigma_u^2}+\frac{1}{2\sigma_u^4}(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})=0$$

o que es lo mismo:

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{w} \qquad [5.22]$$

$$\sigma_u^2=\frac{\left(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{T}=\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T} \qquad [5.23]$$

La ecuación [5.22] revela que, bajo el supuesto de Normalidad del término del error, el estimador de máxima verosimilitud del vector $\boldsymbol{\beta}$ coincide con el estimador de mínimos cuadrados ordinarios. En particular, esta equivalencia permite concluir, sin necesidad de demostrar que $E\left(\hat{\boldsymbol{\beta}}_{MV}\right)=\boldsymbol{\beta}$ y $\text{Var}\left(\hat{\boldsymbol{\beta}}_{MV}\right)=\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$.

Por lo contrario, el estimador de máxima verosimilitud del parámetro σ_u^2 difiere del estimador de mínimos cuadrados ordinarios, además de ser sesgado:

$$E\left(\hat{\sigma}_{MV}^2\right)=E\left(\frac{T-k}{T}\hat{\sigma}_{MCO}^2\right)=\frac{T-k}{T}\hat{\sigma}_u^2 \qquad [5.24]$$

si bien es cierto que, al aumentar el tamaño muestral, el estimador de máxima verosimilitud de σ_u^2 tiende al estimador de mínimos cuadrados ordinarios de dicho parámetro, en consecuencia, como muestra [5.24], el sesgo de $\hat{\sigma}_{MV}^2$ tiende a cero al aumentar el tamaño muestral.[32]

2.5.8. Inferencia

2.5.8.1. Introducción

Desde que el estimador de mínimos cuadrados ordinarios es una transformación lineal del vector $\mathbf{u}$, entonces si $\mathbf{u} \sim N_T(\mathbf{0}_T, \sigma_u^2 \mathbf{I}_T)$, se tiene que $\hat{\boldsymbol{\beta}} \sim N_k(\boldsymbol{\beta}, \sigma_u^2 (\mathbf{X}'\mathbf{X})^{-1})$.

Por otro lado $\hat{\mathbf{u}} = \mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{M}\mathbf{u}$, donde $\mathbf{M} = \left[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$ y, por consiguiente $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{w} - \mathbf{M}\mathbf{u} - \mathbf{X}\boldsymbol{\beta} = \mathbf{u} - \mathbf{M}\mathbf{u} = \mathbf{N}\mathbf{u}$, donde $\mathbf{N} = \mathbf{I}_T - \mathbf{M}$, al igual que $\mathbf{M}$, es una matriz simétrica e idempotente, de dimensión T x T. Por ser idempotente, el rango de $\mathbf{N}$ es igual a su traza y se tiene que:

$$\text{Rango}(\mathbf{N}) = \text{traza}(\mathbf{N}) = \text{tr}\left(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) = \text{tr}(\mathbf{I}_k) = k$$

donde se utilizó la propiedad circular de la traza de matrices. Entonces el $\text{Rango}(\mathbf{M}) = T - k$, y se observa que la forma cuadrática:

$$\frac{1}{\sigma_u^2}\mathbf{u}'\mathbf{N}\mathbf{u} = \frac{1}{\sigma_u^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad [5.25]$$

sigue una distribución chi-cuadrado con k grados de libertad, mientras que

$$\mathbf{u}'\mathbf{M}\mathbf{u} = (T-k)\frac{\hat{\sigma}_u^2}{\sigma_u^2} = \frac{1}{\sigma_u^2}\hat{\mathbf{u}}'\hat{\mathbf{u}} \qquad [5.26]$$

sigue una distribución chi-cuadrado con T - k grados de libertad.

Como:

---

[32] **NOVALES**, Cinca Alfonso. op. cit. págs. 82-84.

$$\mathbf{MN} = \mathbf{M}(\mathbf{1}-\mathbf{M}) = \mathbf{M}-\mathbf{M}^2 = \mathbf{M}-\mathbf{M} = \mathbf{0}_{T\times T}$$

entonces las matrices son ortogonales, por lo que las dos formas cuadráticas anteriores son independientes entre sí.

Proposición 5.13. Los estimadores mínimos cuadrados ordinarios del vector $\boldsymbol{\beta}$ y del parámetro σ_u^2 son independientes entre sí.

Demostración:

En [5.25] el único elemento aleatorio que aparece es el vector de estimadores $\hat{\boldsymbol{\beta}}$, mientras que en [5.26] es proporcional a $\frac{\hat{\sigma}_u^2}{\sigma_u^2}$, donde σ_u^2 es una constante desconocida.

Ambas formas son independientes, pero ello sólo puede ocurrir si el vector aleatorio $\hat{\boldsymbol{\beta}}$ es independiente de la variable aleatoria $\hat{\sigma}_u^2$.

Nótese que esto ocurre a pesar de que los vectores $\hat{\boldsymbol{\beta}}$ y $\hat{\mathbf{u}}$ no son independientes $\hat{\mathbf{u}} = \mathbf{w} - \mathbf{X}\boldsymbol{\beta}$. Aún siendo una función del vector de residuos $\hat{\mathbf{u}}$, el estimador $\hat{\sigma}_u^2$ resume la información estadística contenida en dicho vector de tal manera que resulta independiente del vector $\hat{\boldsymbol{\beta}}$.[33]

2.5.8.2. Contraste de hipótesis

Si:

$$\frac{\frac{\mathbf{u}'\mathbf{N}\mathbf{u}}{k}}{\frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{T-k}} = \frac{\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{k}}{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}} = \frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\left[\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}\right](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{k} \qquad [5.27]$$

sigue una distribución $F_{k,T-k}$, por ser el cociente de dos variables chi-cuadrado independientes.[34]

---

[33] **NOVALES**, Cinca Alfonso. op. cit. págs. 113-115.
[34] **NOVALES**, Cinca Alfonso. op. cit. pág. 115.

2.5.8.3. Interpretación del estadístico F

Sea la hipótesis nula $Ho: \boldsymbol{\beta} = \boldsymbol{\beta}^{\circ}$ y si esta fuese correcta, entonces el vector $\hat{\boldsymbol{\beta}}$ no debería ser muy diferente del vector $\boldsymbol{\beta}^{\circ}$ especificado, es decir, se espera que el vector diferencia $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\circ}\right)$ fuese pequeño. La cuestión por dilucidar es lo pequeña que puede ser esta diferencia o, en otras palabras, que distancia puede aceptarse entre el vector $\boldsymbol{\beta}^{\circ}$ y el vector estimado $\hat{\boldsymbol{\beta}}$.

En primer lugar se observa que el estadístico [5.27] está formado por la norma del vector diferencia $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\circ}\right)$, donde las coordenadas están ponderadas de acuerdo con la inversa de la matriz de covarianzas estimada del vector $\hat{\boldsymbol{\beta}}$. De este modo, $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\circ}\right)$ es tanto más importante en el estadístico F cuanto mayor sea la precisión con que se ha estimado el parámetro $\boldsymbol{\beta}$, es decir, cuanto menor sea su varianza. Si un parámetro se estima con poca precisión, entonces valores relativamente apreciables de la diferencia $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\circ}\right)$ recibirán una menor ponderación en el cálculo del estadístico F.

Bajo la hipótesis nula, $\boldsymbol{\beta}^{\circ}$ es el verdadero valor del vector $\boldsymbol{\beta}$ y, como tal, es constante, por lo que la matriz de covarianzas de la diferencia $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\circ}\right)$ coincide con la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$, cuya inversa aparece en [5.27]. Así el estadístico consiste en dividir la norma del vector diferencia por un indicador de su tamaño para decidir si dicha diferencia es grande o no.

Consecuentemente si el valor del estadístico es menor que el valor de las tablas de la distribución $F_{k,T-k}$, entonces la diferencia entre $\hat{\boldsymbol{\beta}}$ y $\boldsymbol{\beta}^{\circ}$ es pequeña, por lo que no se rechaza la hipótesis establecida $Ho: \boldsymbol{\beta} = \boldsymbol{\beta}^{\circ}$ y se rechaza en caso contrario.

Una consecuencia de que $\hat{\boldsymbol{\beta}} \sim N_k\left(\boldsymbol{\beta}, \sigma_u^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$ es que, en particular, $\hat{\beta}_i \sim N\left(\beta_i, \sigma_u^2 a_{ii}\right)$, donde a_{ii} es el i-ésimo elemento en la diagonal principal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. Por tanto:

$$\frac{\hat{\beta}_i - \beta_i}{\sigma_u \cdot a_{ii}} \qquad [5.28]$$

se distribuye como una normal con media cero y varianza uno.

Pero como σ_u^2 es desconocido y haciendo uso de que el cociente $\frac{\hat{u}'u}{\sigma_u^2}$ se distribuye como una chi-cuadrado con T-k grados de libertad e independientemente de cada uno de los parámetros estimados $\hat{\beta}_i$, por tanto el cociente de una normal estándar y la raíz cuadrada de la variable chi-cuadrado dividida por sus grados libertad:

$$\frac{\dfrac{\hat{\beta}_i - \beta_i}{\sigma_u \cdot a_{ii}}}{\sqrt{\dfrac{(T-k)\hat{\sigma}_u^2}{\sigma_u^2}\,(T-k)}} \qquad [5.29]$$

se distribuye como una t de Student con T - k grados de libertad.[35]

2.5.9. Heterocedasticidad

2.5.9.1. Definición

Cuando la matriz de covarianzas del término de error del modelo de regresión lineal deja de tener una estructura escalar, y conserva su forma diagonal, donde los elementos de la diagonal no son iguales entre sí, se dice entonces que el término del error tiene heterocedasticidad, siendo su varianza diferente para las distintas observaciones que integran la muestra.

Si se estiman los parámetros $\hat{\beta}_i$ con la presencia de heterocedasticidad, dicho vector será insesgado y lineal pero no eficiente.[36]

[35] **NOVALES**, Cinca Alfonso. op. cit. págs. 115-117.
[36] **NOVALES**, Cinca Alfonso. op. cit. pág. 193.

2.5.9.2. El contraste de Breusch y Pagan

Bajo el supuesto que la varianza del término de error en cada periodo depende de un vector de variables $\mathbf{z}_t$ de dimensión p, es decir:

$$\sigma_t^2 = h\left(\mathbf{z}_t'\boldsymbol{\alpha}\right) = h\left(\alpha_0 + \alpha_1 z_{1t} + \alpha_2 z_{2t} + \cdots + \alpha_p z_{pt}\right) \qquad [5.30]$$

Nótese que si todos los coeficientes de la combinación lineal $\mathbf{z}'_t\boldsymbol{\alpha}$, excepto el término independiente, fuesen cero, entonces se tendría una situación de ausencia de heterocedasticidad.

Por consiguiente, si se estiman los coeficientes $\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_p$ un contraste de la hipótesis nula de homocedasticidad está dado por la contrastación conjunta de las p restricciones lineales:

$$H_o : \alpha_0 = \alpha_1 = \alpha_2 = \cdots = \alpha_p \qquad [5.31]$$

Este contraste se efectúa como sigue:

1. Se estima por el método de mínimos cuadrados ordinarios el modelo de regresión lineal original y se obtienen los residuos correspondientes.
2. Se obtiene la sección cruzada de residuos normalizados al cuadrado:

$$\hat{e}_t^2 = \frac{\hat{u}_t^2}{\hat{\sigma}_u^2} \text{ con t:1,2,3,...,T} \qquad [5.32]$$

donde $\hat{\sigma}_u^2$ es la estimación de máxima verosimilitud de la varianza del término de error bajo la hipótesis nula de homocedasticidad, es decir $\hat{\sigma}_u^2 = \frac{SR}{T}$, donde SR es la suma residual de la regresión en el inciso 1)

3. Se estima la regresión de $\hat{e}_t^2$ sobre una constante y las variables $z_{1t}, z_{2t}, \cdots, z_{pt}$ y se obtiene la suma explicada en dicha regresión.
4. Bajo la hipótesis nula de homocedasticidad, y supuesta una distribución normal para el término de error, el cociente $\frac{SE}{2}$ calculado para la

regresión en el inciso 3) se distribuye, según crece el tamaño muestral, como una variable chi-cuadrado con p grados de libertad.

La interpretación de este contraste reside en que si los residuos fuesen homocedásticos, entonces las variables z_t no deberían tener ningún poder explicativo acerca de los residuos transformados $\hat{e}_t^2$ y, por consiguiente, SE debería ser pequeña. Si $\frac{SE}{2}$ es mayor que el valor de la chi-cuadrado en las tablas al nivel de significación escogido, entonces se considera suficientemente grande y se rechaza la hipótesis nula de homocedasticidad.

Como medida remedial, se podría dividir cada observación por $\sqrt{z'_t \hat{\alpha}}$ como una aproximación a la desviación típica de cada periodo.

La estimación de mínimos cuadrados ordinarios de este modelo transformado será la estimación de mínimos cuadrados generalizados del modelo original.[37]

2.5.10. Autocorrelación

2.5.10.1. Definición

Cuando la matriz de covarianzas del término de error del modelo de regresión lineal deja de tener una estructura escalar, es decir algunos o todos los elementos fuera de la diagonal principal son distintos de cero.

Estadísticamente, esta situación proviene del hecho de que el término de error del modelo tiene correlación consigo mismo a través del tiempo.

Si se estiman los parámetros $\hat{\beta}_i$ con la presencia de autocorrelación, dicho vector será insesgado y lineal pero no eficiente.[38]

---

[37] **NOVALES**, Cinca Alfonso. op. cit. págs. 201-202.
[38] **NOVALES**, Cinca Alfonso. op. cit. pág. 224

2.5.10.2. El contraste de Durbin-Watson

Si se sospecha que el término de error del modelo de regresión lineal tiene autocorrelación de primer orden:

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad [5.33]$$

donde ε_t no tiene autocorrelación, entonces el estadístico de Durbin-Watson permite contrastar la hipótesis nula de ausencia de autocorrelación, frente a la alternativa ya mencionada.

Dicho estadístico está definido por la expresión:

$$d = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T}\hat{u}_t^2} \qquad [5.34]$$

La interpretación del estadístico es la siguiente: si la autocorrelación de primer orden con coeficiente ρ positivo hace que valores positivos del término de error u_t tiendan a venir seguidos de valores positivos, y valores negativos tiendan asimismo a venir seguidos de valores negativos. La razón es que, excepto por el componente puramente aleatorio ε_t, el término de error u_t es igual a ρu_{t-1}. Aunque el término de error del modelo es una variable aleatoria no observable, se puede disponer de los residuos generados por una estimación de mínimos cuadrados ordinarios y si $\hat{\boldsymbol{\beta}}$ es insesgado, $\hat{u}_t$ también lo será (aunque ineficiente) de u_t.

Por todo ello, con autocorrelación de primer orden con coeficiente positivo, se observarán rachas de valores negativos y positivos de los residuos. En tales condiciones, la diferencia $\hat{u}_t - \hat{u}_{t-1}$ será generalmente menor, en valor absoluto, que el valor de u_t. Como consecuencia $(\hat{u}_t - \hat{u}_{t-1})^2 < \hat{u}_t^2$ y, en consecuencia, el numerador de la función d será "pequeño" en relación con el denominador.

Si, por el contrario, el coeficiente de autocorrelación fuese negativo, entonces el efecto sería el de tener valores positivos de $\hat{u}_t$ seguidos de valores negativos, y recíprocamente. La implicación numérica es que el valor absoluto de las diferencias $\hat{u}_t - \hat{u}_{t-1}$ tenderá a ser mayor que $\hat{u}_t$. Lo mismo tenderá a ocurrir con sus cuadrados y, como consecuencia, el estadístico d tenderá a tomar valores "grandes".

De hecho el estadístico d está más relacionado con el valor del parámetro ρ, desarrollando el numerador del estadístico de Durbin-Watson:

$$d = \frac{\sum_{t=2}^{T}\hat{u}_t^2 + \sum_{t=2}^{T}\hat{u}_{t-1}^2 - 2\sum_{t=2}^{T}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=2}^{T}\hat{u}_t^2} \qquad [5.35]$$

Si el número de observaciones es suficientemente grande, entonces

$$\sum_{t=2}^{T}\hat{u}_t^2 \approx \sum_{t=2}^{T}\hat{u}_{t-1}^2 \qquad [5.36]$$

y el estadístico d puede aproximarse por $d \doteq 2(1-\hat{\rho})$, donde:

$$\hat{\rho} = \frac{\sum_{t=2}^{T}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=2}^{T}\hat{u}_t^2} \qquad [5.37]$$

pero como $-1 \le \rho \le 1$ implica que: $0 \le d \le 4$, con valores próximos a cero cuando exista autocorrelación positiva de primer orden, y valores cercanos a cuatro cuando existe autocorrelación de primer orden con coeficiente negativo. Cuando el término de error es independiente a lo largo del tiempo, $\hat{\rho}$ será casi nulo y el valor del estadístico d será próximo a dos.[39]

[39] **NOVALES**, Cinca Alfonso. op. cit. págs. 228-229.

2.5.10.3. Pruebas asintóticas

Para establecer si existe o no autocorrelación, se define la hipótesis nula $Ho: \rho = 0$, que puede ser probada contra la alternativa $Ha: \rho \neq 0$ o alguna otra.

Bajo supuestos adecuados puede ser probado que $\hat{\rho}$ será aproximadamente distribuida como una normal con media ρ y varianza $\frac{1-\rho^2}{T}$. Entonces la cantidad:

$$z = \frac{\hat{\rho} - \rho}{(1-\rho^2)\,T} \qquad [5.38]$$

tendrá aproximadamente una distribución normal estándar.

Si la hipótesis nula es cierta, este estadístico será:

$$z = T\hat{\rho} \qquad [5.39]$$

y, consecuentemente a un nivel de significancia de 5%, en una prueba de dos colas, se rechaza Ho si:

$$T\hat{\rho} \geq 1.96 \qquad [5.40]^{40}$$

2.5.11. Multicolinealidad

2.5.11.1. Definición

La multicolinealidad aparece cuando las variables explicativas de un modelo de regresión lineal están correlacionadas entre sí, y tienen implicaciones negativas cuando se pretende estimar un modelo lineal por mínimos cuadrados.

Si se estiman los parámetros $\hat{\beta}_i$ con la presencia de multicolinealidad, dicho vector será insesgado y lineal pero no eficiente.[41]

---

[40] **JUDGE**, George G. R. Carter Hill. William E. Griffiths. Helmut Lütkepohl. Tsoung-Chao Lee. "Introduction to the theory and practice of econometrics". Segunda Edición. Editorial John Wiley & Sons. Estados Unidos de América. 1988. pág. 394.
[41] **NOVALES**, Cinca Alfonso. op. cit. pág. 344.

2.5.11.2. Detección de la multicolinealidad

La selección de variables explicativas en un modelo de regresión pudiera ser adecuada, aunque debido a correlaciones importantes entre ellas, las varianzas de los coeficientes estimados pudieran ser excesivamente altas, los intervalos de confianza muy grandes, y la hipótesis de no significación podría aceptarse.

Los coeficientes de determinación obtenidos en las regresiones de cada variable explicativa sobre las restantes son un buen indicador de una posible situación de multicolinealidad.[42]

2.5.12. Regresión con variables dicotómicas

En el análisis de regresión comúnmente ocurre que la variable dependiente está influenciada no solamente por variables que se pueden cuantificar fácilmente a través de alguna escala bien definida, sino también por aquellas que tienen una naturaleza esencialmente cualitativa.[43]

Puesto que tales variables cualitativas generalmente indican la presencia o ausencia de una cualidad o atributo y un método de cuantificarlos consiste en construir variables artificiales que toman los valores de 1 o 0, donde 0 indica la ausencia de un atributo y 1 la presencia o posesión de esa característica. Las variables que asumen tales valores de 0 y 1 se denominan variables dicotómicas o dummy.

La regla general para incluir variables dummies en el modelo de regresión lineal es que si una variable cualitativa tiene m categorías, se introducen únicamente m-1 variables dicotómicas. Si no se sigue la regla, se presentaría Multicolinealidad perfecta.

La asignación de los valores de 1 y 0 a dos categorías es arbitrario, por tanto, al interpretar los modelos que utilizan variables dicotómicas es fundamental saber como se asignaron los valores 1 y 0.

---

[42] **NOVALES**, Cinca Alfonso. op. cit. pág. 355.

[43] **GUAJARATI**, Damonar N. Econometría. Traducción de: Mayorga Torrado Victor Manuel. Segunda Edición. Editorial Mc. Graw Hill. México. 1992. pág. 367

El grupo, categoría o clasificación a la que se asigna el valor de 0, con frecuencia se conoce como la categoría base, fija, de control o de comparación. Es la base, en el sentido de que las comparaciones se hacen con relación a esa categoría.

El coeficiente asociado a la variable dicotómica se puede denominar coeficiente de intersección diferencial porque indica en cuanto difiere el valor del término de la intersección de la categoría que recibe el valor 1 del coeficiente de intersección de la categoría base.[44]

2.6. El modelo de regresión logística

2.6.1. Introducción

Los métodos de regresión han llegado a ser una componente integral de cualquier análisis de datos para describir la relación entre una variable de respuesta y una o más explicativas y es frecuente el caso en que la de resultado es discreta, es decir aquella que toma dos o más valores. En las últimas décadas los modelos de regresión logística han sido, en muchos campos, el método estándar del análisis en esta situación.

Antes de iniciar el estudio de la regresión logística es importante entender que la intención del análisis usando este método es la misma que la de cualquier técnica de diseño de modelos usada en la estadística: encontrar el mejor ajuste y el más parsimonioso.

Lo que distingue el modelo de regresión logística del lineal es que la variable de resultado en el primero es binaria o dicotómica, sin embargo tienen los mismos principios generales.[45]

Otra diferencia consiste en la naturaleza de la relación entre la variable de resultado y la independiente. En cualquier problema de regresión la cantidad clave es el valor promedio de la variable de resultado, dado el de la independiente. Esta cantidad es llamada la media condicional y su expresión está dada por:

---

[44] **GUAJARATI**, Damonar N. op. cit. págs. 373-374

[45] **HOSMER**, David W., Lemeshow Stanley. Applied logistic regression. Primera Edición. Editorial John Wiley & Sons. Estados Unidos de América. 1976. pág. 1

$$E(w\,x) \qquad [6.1]$$

Donde w denota la variable de resultado y x el valor de la independiente, y se lee como "el valor esperado de w, dado el de x". En la regresión lineal se asume que esta media podría ser expresada como una ecuación lineal en x, tal como:

$$E(w\,x)=\beta_0+\beta_1 x+\cdots+\beta_k x_k \qquad [6.2]$$

Esta expresión implica que es posible para $E(w\,x)$ tomar cualquier valor como el rango de x entre $-\infty$ y ∞.

Con los datos dicotómicos la media condicional debe ser mayor o igual a cero y menor o igual a uno, es decir:

$$0 \le E(w\,x) \le 1 \qquad [6.3]$$

El cambio en la $E(w\,x)$ por unidad de cambio en x llega a ser progresivamente pequeño como la media condicional crece de cero a uno. La curva de $E(w\,x)$ es en forma de S y la distribución logística es la indicada para asemejarse.

Existen dos razones para elegir esta distribución: desde el punto de vista matemático, es una función extremadamente flexible y fácil de usar, y permite en sí misma una interpretación biológica. A fin de simplificar la notación, se emplea la cantidad:

$$\pi(x)=E(w\,x) \qquad [6.4]$$

Para representar la media condicional de w dado x cuando la distribución logística es usada. La forma específica del modelo de regresión logística que se usa es la siguiente:

$$\pi(x)=\frac{e^{\beta_0+\beta_1 x+\cdots+\beta_k x}}{1+e^{\beta_0+\beta_1 x+\cdots+\beta_k x}} \qquad [6.5]$$

Una función que es central en el estudio de la regresión logística es la transformación logit. La cual está definida en términos de $\pi(x)$ como sigue:

$$g(x)=\mathrm{Ln}\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1 x+\cdots+\beta_k x_k \qquad [6.6]$$

En el análisis de regresión una observación de la variable dependiente es expresada como:

$$\hat{w}_t = E(w_t x) + u_t \qquad [6.7]$$

donde el segundo término de la ecuación lineal es llamado el error y expresa una desviación de la observación de la media condicional.

En la regresión lineal bajo el supuesto que el error sigue una distribución normal con media cero y varianza constante entre los diferentes niveles de la variable independiente, la distribución condicional de la variable de resultado dado el valor de x también tiene dicha distribución, con media $E(w_t x)$ y varianza constante, pero no pasa lo mismo con la regresión logística, porque se trata de una variable dicotómica, y el error entonces asume dos valores, si $w_t = 1$ entonces $u_t = 1 - E(w_t x)$ y para el caso en que $w_t = 0$ se tiene que $u_t = -E(w_t x)$[46].

Suponiendo que se tiene una variable dependiente dicotómica, y sea π_t la probabilidad de $w_t = 1$ (de que el evento ocurra) y $1 - \pi_t$ es la probabilidad de que $w_t = 0$ (de que no ocurra), y por la definición de esperanza matemática, se tiene que:

$$E[w_t] = 0(1 - \pi_t) + 1(\pi_t) = \pi_t \qquad [6.8]$$

Y al comparar el modelo [5.5] con el [6.8] se tiene que:

$$E(w_t x) = \hat{w}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{2t} + \cdots + \hat{\beta}_k x_{kt} = \pi_t$$

Dado que la probabilidad π_t debe estar entre 0 y 1, se da la restricción [6.3], aunque el método de los mínimos cuadrados ordinarios no requiere que los errores estén normalmente distribuidas, se han supuesto para efectos de inferencia estadística. Sin embargo, el supuesto de normalidad de las perturbaciones no es válido en la regresión logística, debido a que como ocurre con los wy_t, los u_t toman sólo dos valores, es decir:

$$u_t = w_t - (\beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}) \qquad [6.9]$$

Ahora cuando:

---

[46] **HOSMER**, David W., Lemeshow Stanley. op. cit. págs. 5-7

$$w_t = 1,\ u_t = 1-(\beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}) \qquad [6.10]$$

$$w_t = 0,\ u_t = -(\beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}) \qquad [6.11]$$

por lo que no se puede suponer que los errores estén normalmente distribuidos, en realidad, siguen una distribución binomial.

Aún bajo el supuesto que $E(u_t)=0$ y $E(u_s u_t)=0$ para t≠s, es decir no existe correlación serial, no se puede satisfacer la condición de que las perturbaciones sean homocedásticas, esto se desprende de [6.10] y [6.11], ya que la función de probabilidad de los errores está dada por:

$$1-(\beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}) \text{ con } \pi_t \qquad [6.12]$$

$$-(\beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}) \text{con } 1-\pi_t \qquad [6.13]$$

Y por definición de la varianza:

$$Var(u_t) = E[u_t - E(u_t)]^2 = E(u_t^2)$$

Por lo tanto, usando esta distribución probabilística de u_t[47] se obtuvo que:

$$Var(u_t) = E(y_t \mid \mathbf{x})(1-E(y_t \mid \mathbf{x})) = [\pi_t][1-\pi_t] \qquad [6.14]$$

La ecuación [6.14] muestra que la varianza de las perturbaciones es heterocedástica porque depende de la esperanza condicional de w, que está en función del valor que tome **x**.[48]

2.6.2. Estimación de parámetros

Considerando a w_t una variable dicotómica, y denotando los 2 resultados con 0 y 1, se obtiene una variable aleatoria Bernoulli, donde:

$$P(w_t = 1) = \pi(x_t) \qquad [6.15]$$

$$P(w_t = 0) = 1-\pi(x_t) \qquad [6.16]$$

y $\mathbf{x}_t = (x_{1t}, x_{2t}, \cdots, x_{kt})$ es un vector aleatorio, que tiene una escala de medida de intervalo.

---

[47] Distribución bernoulli, ver apéndice B.1.
[48] **GUAJARATI**, Damonar N. op. cit. págs. 407-408

Supóngase además que w_t depende del vector aleatorio $\mathbf{x}_t$, tal que:

$$w_t = g(\mathbf{x}_t) = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt}$$

y que:

$$E(w_t^2) = E[w_t] = \pi(\mathbf{x}_t) \qquad [6.17]$$

entonces la:

$$Var[w_t] = \pi(\mathbf{x}_t)[1 - \pi(\mathbf{x}_t)] \qquad [6.18]$$

y de aquí que se tiene que la variabilidad no es constante, pero esta no depende del vector $\mathbf{x}_t$, sino de la función $\pi(\mathbf{x}_t)$, donde $\pi(\mathbf{x}_t) \in [0,1]$, por lo que: $Var[w_t] \in [0,1]$.

Sea:

$$\pi(\mathbf{x}_t) = \frac{\exp(g(\mathbf{x}_t))}{1 + \exp(g(\mathbf{x}_t))} \qquad [6.19]$$

la función de probabilidad del vector aleatorio $\mathbf{x}_t$ y tomando una muestra de T observaciones independientes de los pares (x_t, w_t) para t = 1,2,3,..,T, y la función de probabilidad del vector w_t está dada por:

$$\varsigma(\mathbf{x}_t) = \pi(\mathbf{x}_t)^{w_t}[1 - \pi(\mathbf{x}_t)]^{1 - w_t} \qquad [6.20][49]$$

para $w_t = 0,1$, t = 1:2,3,..,T.

Nótese que:

$$\frac{\partial \pi(\mathbf{x}_t)}{\partial x_{tk}} = \beta_k \pi(\mathbf{x}_t)[1 - \pi(\mathbf{x}_t)] = \beta_k Var[w_t] \qquad [6.21]$$

por lo que el valor del incremento o decremento de la pendiente de la función de regresión logística depende de las entradas del vector $\boldsymbol{\beta}$.

Entonces la función de máxima verosimilitud se escribe como:

$$L(\boldsymbol{\beta}) = \prod_{t=1}^{T} \varsigma(\mathbf{x}_t) = \prod_{t=1}^{T} \left\{ \pi(\mathbf{x}_t)^{w_t}[1 - \pi(\mathbf{x}_t)]^{1 - w_t} \right\} \qquad [6.22]$$

extrayendo el logaritmo natural de [6.22] se tiene que:

---

[49] **HOSMER**, David W. Jr., Lemeshow Stanley. op. cit.págs. 8-10

$$\log_e L(\beta) = \sum_{t=1}^{T} \{w_t \, Ln\, \pi(x_t) + (1 - w_t) Ln [1 - \pi(x_t)]\}$$

$$= \sum_{t=1}^{T} \left\{ w_t \left(\beta_1 + \sum_{k=2}^{K} \beta_k x_{tk} \right) + Ln[1 - \pi(x_t)] \right\} \qquad [6.23]$$

Y tomando:

$$0 = \frac{\partial Ln\, L(\beta)}{\partial \beta_1}_{\beta_1 = \hat{\beta}_1} = \sum_{t=1}^{T} \{w_t - \pi(x_t)\} \qquad [6.24]$$

$$0 = \frac{\partial Ln\, L(\beta)}{\partial \beta_k}_{\beta_k = \hat{\beta}_k} = \sum_{t=1}^{T} \{x_{tk} [w_t - \pi(x_t)]\} \qquad [6.25]$$

se tiene que:

$$0 = \sum_{t=1}^{T} \{w_t - \pi(x_t)\} = \sum_{t=1}^{T} \{x_{tk} [w_t - \pi(x_t)]\}$$

y resolviendo estas ecuaciones para $\hat{\beta}_1$ y $\hat{\beta}_k$ se obtienen los estimadores de máxima verosimilitud de dichos parámetros.

Debido a la complejidad del sistema de ecuaciones, se implementa el método de Newton - Raphson para dar solución a dicho sistema de ecuaciones. [50]

2.6.3. Propiedades del modelo logit

Dado que $\pi(x_t) \in [0,1]$, a medida que $g(x_t)$ varía entre $-\infty$ y ∞, el logit $L_t = \log_e \left(\frac{\pi(x_t)}{1 - \pi(x_t)} \right)$ está entre $-\infty$ y ∞. Es decir, aunque las probabilidades (necesariamente) se encuentran entre 0 y 1, los logit no tienen esos límites.

Aunque **L** es lineal en x_t, las probabilidades mismas no lo son, es decir, que se acerca a cero a tasas cada vez menores a medida que x_t disminuye y que se aproxima a 1 a velocidades cada vez más lenta a razón que x_t se incrementa.

[50] **HOSMER**, David W. Jr., Lemeshow Stanley. op. cit. págs. 27-28

La interpretación del modelo logit es la siguiente: los coeficientes $\beta_2, \ldots, \beta_k$, son las pendientes que miden el cambio en **L** por un cambio unitario en x_t, es decir muestra como varia la factibilidad del log a favor del riesgo a medida que la variable dependiente cambia una unidad. La intersección β_1 corresponde al valor de la probabilidad en log a favor del riesgo si la variable es 0. Como la mayoría de las interpretaciones de las intersecciones esta puede no tener significado físico.

Dado cierto nivel de la variable dependiente es posible estimar las probabilidades a favor del riesgo así como en rechazo a éste.[51]

[51] **GUAJARATI, Damonar N.** op. cit. pág. 421

CAPÍTULO III

PROCEDIMIENTO Y RESULTADOS

3.1. Motivación y justificación

La necesidad de describir otros factores del peso del neonato lo más acertadamente posible, es de la mayor importancia, ya que desde hace pocos años, el peso del producto al nacer ha sido el elemento esencial para colocarlo en cualquiera de dos categorías:

1. El bajo peso: menos de 2500 gramos.
2. Peso adecuado: mayor o igual a 2500 gramos.

Y varios han sido los parámetros fundamentales para la descripción del peso:

1. Su edad gestacional, es decir el tiempo que el producto ha permanecido en su microambiente natural recibiendo los beneficios de una placenta que pone a su disposición los elementos nutrientes que ha menester, dentro de las limitaciones impuestas por otros factores ecológicos.
2. La talla del neonato: añade la información sobre las condiciones intrauterinas y también influye en el crecimiento subsiguiente, es un valioso indicador de la salud del neonato y tiene una estrecha relación con la longitud del preescolar y del adolescente.

Pero existe otro factor que no se ha considerado, y en la actualidad es necesario indagar acerca de su efecto en el peso al nacer del producto, como es la contaminación ambiental, fue por ello que se planteó este estudio y se recurrió a

la maternidad del Hospital Español, Sociedad de Beneficencia Española IAP, ya que la población ahí atendida es homogénea y corresponde a la clase media de la Ciudad de México.

3.2. Antecedentes

En el año 1996 fue editado el programa nombrado "Teplice Program--The Impact of Air Pollution on Human Health", cuya intención fue la de investigar y analizar el efecto de la polución del aire causada por el dióxido de azufre y las partículas suspendidas totoales en la salud de la población en el distrito de Teplice y Prachatice de la República Checa. El primero es un distrito minero mientras que el otro es una población urbana, y existe un predominio altamente significativo de los síntomas respiratorios adversos y un decremento de la función pulmonar fue más frecuente en el centro minero que en el urbano. Los estudios neuroconductales en los niños indicaron diferencia significativa en ambos distritos, los niños de Teplice fueron referidos para análisis clínicos de aprendizaje o problemas de conducta. Los estudios reproductivos fueron aplicados en hombres y mujeres, los cuales indicaron el predominio de bajo peso al nacer y nacimientos prematuros en el distrito minero más que en el urbano. La salud reproductiva de los hombres fue evaluada en un estudio de semen de los jóvenes, se midió la calidad de este en los dos distritos aunado al factor temporal sugirieron que una exposición a altos niveles de contaminación del aire podría estar asociado con un decremento transitorio en la calidad del líquido eyaculatorio.[52]

En el año 1997 fue publicado el estudio titulado "Association between air pollution and low birth weight: a community- based study", que mostró la relación entre la exposición materna a la contaminación del aire provocada por las partículas suspendidas totales y el dióxido de azufre durante el tercer trimestre de embarazo y el peso del neonato al nacer.[53]

---

[52] **ŠRÁM**, Radim J., Beneš Ivan, Binková Blanka, Dejmek Jan, Horstman Donald, Kotěšovec František, Otto David, Perreault Sally D., Rubeš Jiří, Selevan Sherry G., Skalík Ivan, Stevens Robert K., Lewtas Joellen. Teplice program – the impact of air pollution on human health. Environmental Health Perspectives. Volumen 104. Suplemento 4. Estados Unidos de América. Agosto 1996. págs. 699-714.

[53] **WANG**, Xiaobin, Ding Hui, Ryan Louise, Xu Xiping. Association between air pollution and low birth weight: a community-based study. Environmental Health Perspectives. Volumen 105. Número 5. Estados Unidos de América. Mayo 1997. págs. 514-520.

En el año 1999 se imprimió el programa de nombre "The effect of ambient carbon monoxide on low birth weight among children born in Southern California between 1989 and 1993", que mostró que la exposición materna durante el último trimestre de embarazo al Monóxido de Carbono incrementó el riesgo para el bajo peso al nacer.[54]

En el año 2000 se dio a conocer el análisis llamado "Children's health outdoor air pollution, low birth weight and prematurity", el cual mostró que el bajo peso al nacer y el pretérmino estuvieron asociados con el dióxido de azufre y en menor grado con las partículas suspendidas totales durante el primer trimestre de embarazo.[55]

En el año 2001 se editó el estudio titulado "Relation between ambient air pollution and low birth weight in the Northeastern United States", que sugirió que la exposición durante el tercer trimestre de embarazo al monóxido de carbono y en el segundo trimestre de gestación al dióxido de azufre incrementó el riesgo del bajo peso al nacer. [56]

Considerando que México presenta índices de contaminación semejantes a dicho países donde se realizaron las investigaciones, y que no está exento del bajo peso al nacer, son razones suficientes para plantear una hipótesis acerca de la existencia de una relación de entre los niveles de contaminación del aire y el bajo peso al nacer en la Zona Metropolitana del Valle de México.

3.3. Objetivos

Los estudios epidemiológicos se llevan a cabo con los siguientes objetivos:

1. Determinar si la contaminación del aire o alguna causa de ésta, constituye un riesgo para la salud humana.

---

[54] **RITZ, Beate**, Yu Fei. The effect of ambient carbon monoxide on low birth weight among children born in Southern California between 1989 and 1993. Environmental Health Perspectives. Volumen 107, Número 1. Estados Unidos de América. Enero 1999. págs.17-25.

[55] **BOBAK, Martin**. Children's health outdoor air pollution, low birth weight and prematurity. Environmental Health Perspectives. Volumen 108. Número 2. Estados Unidos de América. Febrero 2000. págs. 173-176.

[56] **MAISONET, Mildred**, Bush Timothy J., Correa Adolfo, Jaakkola Jouni J. K. Relation between ambient air pollution and low birth weight in the Northeastern United States. Environmental Health Perspectives. Volumen 109. Suplemento 3. Estados Unidos de América. Junio 2001. págs. 351-356.

2. Tipificar la relación entre el nivel de exposición y el de respuesta.
3. Examinar las respuestas de la población potencialmente susceptible a la exposición del contaminante.

Por lo que se formularon las siguientes preguntas fundamentales

1. ¿Las partículas suspendidas totales constituyen un riesgo para el peso al nacer?
2. ¿A qué niveles de exposición el riesgo es aceptable?
3. ¿Cuáles son los grupos que necesitan especial consideración debido a su susceptibilidad?

Así pues, el propósito de este estudio fue conïstruir un modelo estadístico descriptivo del bajo peso al nacer a través de las partículas suspendidas totales, así como examinar el tiempo y la intensidad de exposición a este contaminante durante el embarazo y su asociación con el peso del producto haciendo uso de la variable de control: *"talla del neonato"* así como la interacción del *"año de gestación"* y de los "*promedios de exposición de la madre a las partículas suspendidas totales durante distintos periodos del embarazo*".

3.4. Material biológico

Se revisaron los expedientes de los recién nacidos vivos del hospital, durante el periodo comprendido entre el 1° de enero de 1993 al 31 de diciembre de 1997, correspondiente a una muestra de 9323 neonatos.

Se excluyeron del análisis todos aquellos registros en que la residencia de la madre no fue explícitamente el Distrito Federal y su zona conurbana, el desglose de estos casos señaló que 147 no especificaban la zona donde habitaba la progenitora, 2726 hacían referencia a la Zona Metropolitana del Valle de Toluca y 224 a otros estados.

De los neonatos cuya progenitora tuvo como residencia la Zona Metropolitana del Valle de México, se eliminaron los registros de neonatos de pretérmino (294) y los de postérmino (9), así como se descartaron los casos en los que el producto del embarazo fue gemelar (136), por lo que se reduce la muestra a 5787 neonatos, para el análisis estadístico.

3.5. Procedimiento

3.5.1. Origen de la información

Los datos de la progenitora tales como:

1. *La zona de* residencia.
2. *El estado socioeconómico.*
3. *Si se cuenta con alguna póliza de seguro de gastos médicos mayores.*
4. *La descripción de alguna patología o riesgo.*

Fueron consultados directamente de la madre del neonato y asentados en su respectivo expediente clínico.

También la información del neonato como:

5. *El peso:* fue *cuantificado* recurriendo a pesa-bebés "Fairbanks" de 15.5 kilogramos.
6. *La talla*: se midió con el niño en decúbito dorsal y empleando una cinta métrica ahulada inextensible.
7. *Las semanas de gestación:* se calculó a partir del primer día de la última menstruación y se consideró como duración normal del embarazo 36 a 42 semanas.
8. *La descripción de alguna patología o riesgo.*
9. *El apgar al primer y quinto minuto de vida*
10. *La fecha de nacimiento.*

Fue registrada en el expediente clínico para su control.

Cabe mencionar que esta información fue capturada posteriormente en una base de datos del paquete EXCEL.

Respecto de los índices de contaminación empleados en el análisis, fueron cuantificados por el Sistema de Monitoreo Atmosférico de la Zona Metropolitana del Valle de México a través de la red manual desde 1992 hasta 1997 registrados por el Instituto Nacional de Ecología dependiente de la Secretaría del Medio Ambiente, Recursos Naturales y Pesca.

Esta información fue presentada para cada una de las estaciones de monitoreo de la Zona Metropolitana del Valle de México y se resolvió tomar

aquella que fuera representativa del área de estudio, así para la región noroeste, nordeste, centro, sudoeste y sureste fueron Tlalnepantla, Xalostoc, Merced, Pedregal y Cerro de la Estrella respectivamente.

La ubicación de estas estaciones de monitoreo es:

1. Tlalnepantla: *Glorieta* Atlacomulco, Fraccionamiento Tlamez, Código Postal 54070, Tlalnepantla Estado de México.
2. Xalostoc: Distribuidora Volkswagen "Santa Clara", Carretera Pachuca (Emiliano Zapata) kilómetro 13.5 y Calle del Hierro, Xalostoc, Ecatepec Estado de México.
3. La Merced: Centro de Salud "Luis E. Ruiz", Avenida Congreso de la Unión 148, antes Francisco. Morazán y Prolongación de los Carretones, Colonia Merced Balbuena, Código Postal 158600, Delegación Venustiano Carranza.
4. Pedregal: Escuela Primaria "John F. Kennedy", Cañada número 370 y Avenida Crater, Colonia Pedregal de San Ángel, Código Postal 01900, Delegación Álvaro Obregón.
5. Cerro de la Estrella: Planta de Tratamiento de Aguas Negras "Cerro de la Estrella", Departamento del Distrito Federal, Avenida San Lorenzo sin número, Colonia Paraje San Juan Código Postal 090830 Delegación Iztapalapa.

3.5.2. Ajuste de los datos

Los índices de contaminación de las partículas suspendidas totales proporcionados por el Instituto Nacional de Ecología en su mayoría fueron registrados durante cada seis días, debido a que son analizados a través de la Red de Monitoreo Manual con que se cuenta, por lo que recurrió a interpolar los datos usando interpolación lineal, para así poder estimar un valor aproximado por falta de información, cabe aclarar que para cada una de las estaciones empleadas en el estudio, adicionalmente no se registraron los índices de contaminación de las siguientes fechas:

1. Tlalnepantla: 11/01/92, 24/07/96, 30/07/96, 22/09/96 y 28/09/96.

2. Xalostoc: 24/07/96, 30/07/96, 05/08/96, 11/08/96, 17/08/96, 23/08/96, 22/09/96 y 28/09/96.
3. Merced: 24/11/92, 07/01/95, 24/07/96, 30/07/96, 22/09/96 y 28/09/96.
4. Pedregal: 06/04/94, 24/07/96, 30/07/96, 22/09/96, 28/09/96 y 19/02/97.
5. Cerro de la estrella: 09/07/92, 19/12/93, 24/07/96, 30/07/96, 22/09/96 y 28/09/96.

Para calcular los datos faltantes, se implementaron los programas elaborados en MATLAB versión 5.0, que calcularon dichas estimaciones, de los cuales se anexa la sintaxis en los incisos A.3.1.1., A.3.1.2. y A.3.1.3.

3.5.3. Construcción de los promedios

Para examinar la importancia relativa del tiempo y magnitud de riesgo a la contaminación del aire con relación al peso al nacer, se construyeron los promedios de exposición de la madre a las partículas suspendidas totales durante distintos periodos del embarazo:

1. *Desde la fecha de gestación hasta la fecha de* nacimiento *del neonato.*
2. *Durante cada trimestre de embarazo.*
3. *Proximidades a la fecha de nacimiento, es decir, 1, 2, 3, 4, 5, 6, 7 y 8 semanas de gestación antes del parto.*
4. *Proximidades posteriores a la fecha de embarazo, es decir, 1, 2, 3, 4, 5, 6, 7 y 8 semanas de gestación.*

Dichas medias fueron calculadas usando programas elaborados en el lenguaje de programación CLIPPER versión 5.3, de los cuales se anexa la sintaxis en el apartado A.3.2.

La regresión lineal y logística múltiple se implementaron para analizar los efectos de las partículas suspendidas totales en el peso del neonato y el riesgo del bajo peso al nacer respectivamente; y considerando que la talla es un factor determinante del peso al nacer, se incluyó en estos modelos.

3.6. Resultados

En este estudio, la muestra poblacional para el análisis incluyó 5787 nacimientos de producto único, nacidos vivos, de 36 a 42 semanas de gestación, cuya información acerca del nacimiento es completa y definida.

Las características de los nacimientos se resumen en la tabla 3, ver apartado A.5.3., de donde se observó que:

1. El *peso* promedio de los 5787 neonatos fue de 3,095.20 gramos, con una desviación estándar de 407.33 y la tasa de bajo peso del 5.93%. Sin embargo, el promedio del peso y la tasa del bajo peso varió en todas y cada una de las variables.

2. El 16.07, 21.12, 20.03, 18.61, 19.63 y 4.54%, se gestaron en los años 1992, 1993, 1994, 1995, 1996 y 1997 respectivamente.

3. El 21.95, 20.72, 19.20, 18.99 y 19.15% de los bebés, nacieron en los años 1993, 1994, 1995, 1996 y 1997 respectivamente.

4. Las tasas del bajo peso al nacer en los neonatos nacidos en los años 1993, 1994 y 1995 fueron del 1.21, 1.19 y 1.28%, mientras que el 1.12% fue para los periodos 1996 y 1997.

5. Se observaron 343 casos de bajo peso, 248 de los cuales fueron pequeños y los otros 95 adecuados para su edad gestacional, es decir el 72.3% de los neonatos de peso menor a 2500 gramos presentaron retardo en crecimiento intrauterino.

6. 206 neonatos que habían sido agrupados como peso normal, es decir mayor o igual a 2500 gramos, se clasificaron como pequeños para su edad gestacional, esto representó el 45.37% de los casos con síndrome de Cliflord.

7. Mientras que la tasa de bajo peso al nacer fue del 5.93% la del retardo en el crecimiento intrauterino fue del 7.85%.

8. Con respecto al coeficiente [1.1][57] el 95.8% de los casos se agrupó dentro del rango aceptable, es decir, mayor o igual a 3 y menor o igual a 7, mientras el 4.20% cayó fuera de este intervalo.
9. El 44.86% de los neonatos cuyo coeficiente [1.1] cayó fuera del rango aceptable fueron de bajo peso.
10. Del coeficiente [1.2][58] el 97.81% de los casos se clasificó dentro del límite aceptable es decir, mayor o igual a 2 y menor o igual a 3, y el 2.19% restante estuvo fuera de este rango.
11. El 40.16% de los neonatos cuyo coeficiente [1.2] cayó fuera del rango aceptable fueron de bajo peso.
12. También se desprende que para la zona Nordeste que en promedio presenta los más altos índices de contaminación en las partículas suspendidas totales la tasa del bajo peso al nacer fue de 6.60%.
13. Y que para la región sudoeste cuya media de concentración de las partículas suspendidas totales fue la más baja, presentó una tasa del bajo peso al nacer de 4.67% Cabe mencionar que ambos resultados presentan una relación directamente proporcional, ya que a mayor concentración de contaminación se presenta una mayor tasa de bajo peso al nacer.
14. Se encontró una relación directamente proporcional entre las *semanas de gestación* y el *peso* promedio del neonato al nacer.
15. Se halló una relación inversamente proporcional entre el tiempo en que está en el vientre de la madre y el bajo peso al nacer.
16. El 93.59% de los neonatos de bajo peso tuvieron una *talla* menor de 50 centímetros.
17. El 56.27% de los casos registrados de bajo peso al nacer fueron del sexo femenino, y el 43.73% restante fueron varones.

---

[57] $\frac{100 * (\text{peso en gramos})}{(\text{semanas de gestación})^3}$

[58] $\frac{100 * (\text{peso en gramos})}{(\text{talla en centímetros})^3}$

18. El 10.49% de los neonatos cuyo *apgar al primer minuto de vida* fue menor o igual a siete, fueron recién nacidos de bajo peso.
19. El 23.53% de los casos donde el *apgar al quinto minuto de vida* es menor o igual a siete fueron de bajo peso al nacer.
20. Otra variable que destacó es si *la madre es fumadora*, pues cuando este riesgo estuvo presente la tasa del bajo peso fue del 70.59% y en el otro caso del 5.54%.
21. Si el niño al nacer presentó *algún riesgo o patología*, la tasa del bajo peso fue del 52.87%.
22. Si durante el embarazo, la madre presentó algún antecedente patológico la tasa del bajo peso del neonato fue del 22.17%.
23. La gráfica 2 del apartado A.4.3., mostró el promedio de las partículas suspendidas totales desde abril de 1992 hasta 1997, estos niveles de contaminación del aire fueron el promedio de los niveles diarios de cada mes, sobreimpuestos estos en el peso promedio mensual de todos los neonatos gestados en el mismo periodo, mostraron un patrón temporal para el peso y los niveles de contaminación promedios. De esto se desprendió que existe una relación inversa entre el nivel promedio de contaminación del aire y la media del peso al nacer, y esto es a niveles altos de contaminación corresponde un nivel promedio bajo de peso, y viceversa. Esta relación también la mostró las líneas de tendencia del peso al nacer y las partículas suspendidas totales, además que a una gran variación de los contaminantes existe una variación relativamente pequeña en el nivel promedio del peso al nacer.
24. Para examinar la importancia relativa del tiempo y magnitud de exposición de la madre a las partículas suspendidas totales en relación con el peso al nacer se construyó el promedio de exposición de la madre a este contaminante desde la fecha de embarazo hasta los primeros treinta y cinco días de gestación; en dicho modelo de regresión lineal múltiple se implementó el uso de variables dummies (año de gestación del neonato), la transformación de las dependientes (segunda potencia de la talla del

bebé) y explicativas (raíz cúbica del peso del producto) y el efecto de interacción entre las endógenas (año de gestación y promedio de exposición de la madre a las partículas suspendidas totales desde la fecha de embarazo hasta los primeros treinta y cinco días de gestación).

25. La asociación entre el bajo peso al nacer y las partículas suspendidas totales fue evaluada a través de la regresión logística; se implementó el uso de transformación de las variables dependientes (la cuarta potencia de la talla del neonato y la raíz cúbica del promedio de exposición de la madre al contaminante desde la fecha de gestación hasta los primeros treinta y cinco días posteriores a esta fecha).

3.7. Comentarios adicionales

Durante el análisis preliminar de la información capturada en la base de datos se consideraron las siguientes observaciones:

1. Se resolvió eliminar del análisis los neonatos cuya progenitora haya registrado como lugar de residencia la Zona Metropolitana del Valle de Toluca, esto debido a la carencia de información en los índices de las partículas suspendidas totales en los años 1992, 1993 y 1994, y por consiguiente, también aquellos en que no fue especificada dicha región.
2. Se restaron los bebés de pretérmino, porque la deficiente maduración de sus órganos no produce la ganancia del peso suficiente debido a su edad gestacional.
3. Se eliminaron los recién nacidos posmaduros, porque el peso que presentaron estuvo directamente influenciada por el tiempo en que estuvo en el vientre de su madre.
4. Se excluyeron del análisis los neonatos gemelares y múltiples, debido a que estos presentan menor peso en promedio que los de producto único.
5. Al examinar los riesgos de la madre durante el embarazo asentados en la base de datos, los siguientes registros señalaron ser casos de producto gemelar, sin embargo:

a. En el expediente número 352353, la progenitora registró la Zona Metropolitana del Valle de México y en el 352354 el Valle de Toluca.
b. Respecto del expediente número 362180, no se escribió el par y en el registro próximo anterior o posterior no se encontró coincidencia en la edad gestacional, en la zona de residencia de la madre ni en la fecha de nacimiento.
c. Los expedientes número 365000 y 3365101 no son progresivos tienen una diferencia de 101 registros, sin embargo los demás datos son coincidentes.
d. En el expediente número 368743 se registró la edad gestacional 38.1 y en el número 368744 de 37.5 semanas de gestación.
e. El expediente número 371546 registró como fecha de nacimiento del neonato el día 28/05/96 y el 371547 el 29/05/96.
f. Los expedientes número 372150 y 372351 no son progresivos tienen una diferencia de 201 registros, sin embargo los demás datos son coincidentes.
g. Respecto del expediente número 372832, no se escribió el par y en el registro próximo anterior o posterior no se encontró coincidencia en la edad gestacional, en la zona de residencia de la madre ni en la fecha de nacimiento.
h. Los expedientes número 379800 y 379951 no son progresivos tienen una diferencia de 151 registros, sin embargo los demás datos son coincidentes.

Considerando que los números de expedientes son consecutivos, entonces existe un error en estos registros, por lo que se resolvió eliminar estos casos por ambigüedad en la información.

6. La precisión del primer día de la última menstruación tiene una serie de errores de apreciación por parte de la madre que la inducen a dar datos equívocos con cierta frecuencia, la cual se hace importante a medida que las condiciones culturales y de educación son menos adecuadas,

entonces no es posible determinar exactamente la fecha de la última regla y por consiguiente las semanas de gestación, pero también se tiene presente la gran utilidad que tal estimación está prestando aún a pesar de los errores y se confía que en el futuro la madre pueda dar información más precisa.

7. Con respecto a la variable *peso al nacer* existe una predilección en el último dígito hacia el 0 en un 63.82% y con el número 5 en un 35.86%, siendo el 99.68% de los casos.
8. En la variable *talla del neonato* hay preferencia en la última cifra hacia el número 0 en un 94.12% y el resto conforma el 5.882% de todos los casos.
9. Con respecto a la variable *semanas de gestación* tiene preferencia en el último número hacia el 0 en 53.08%, el otro 46.92% estuvo dividido entre los otros dígitos finales.
10. No se especificó si la información contenida en la variable *zona de residencia de la madre* es la región donde vivió durante el embarazo o es aquella en la cual vivió anterior a la gestación, o bien, si hubo cambio de residencia de la progenitora tampoco fue informado.
11. No se menciona si la información contenida en la variable *tabaquismo en la madre* se refiere a que la progenitora fumó tabaco durante la gestación o antes del embarazo.
12. Al realizar el análisis de regresión del peso al nacer con las variables dependientes: *la edad gestacional, el sexo del neonato, el apgar al primer y al quinto minuto de vida, si se contaba con una póliza de gastos médicos mayores, la zona de residencia de la madre, la edad de la madre y el estatus socio-económico de la madre* se encontró poca evidencia estadística para proceder a analizar el peso del neonato con dichas variables, y sólo con la variable *talla del recién nacido*, hubo una fuerte relación descriptiva, por lo que se implementó el modelo de regresión lineal múltiple con esta variable, cabe mencionar que este dato solo puede ser utilizado cuando el neonato ya nació y no antes, pues de otra

forma se tendría que contar con la estatura de la progenitora y el padre para poder estimar la longitud del bebé.

13. Se construyó el modelo de regresión lineal múltiple con la *raíz cúbica del peso*, se implementó la prueba de Breusch-Pagan y hubo evidencia suficiente para rechazar la hipótesis nula de homocedasticidad, debido a que cuando la *talla del recién nacido* crece la varianza del *peso en gramos* aumenta, por lo que se procedió a ponderar las observaciones y así uniformar la varianza del peso, al comparar dichos resultados se apreció que ésta se incrementó, creció el error estándar de los parámetros estimados, pero al aplicar nuevamente la prueba de Breusch-Pagan no hubo evidencia suficiente para rechazar la hipótesis nula de homocedasticidad.
14. Al realizarse la prueba asintótica para la autocorrelación de primer orden, no se encontró evidencia suficiente para rechazar la hipótesis nula $Ho: \rho = 0$, y esto coincide con la explicación biológica, ya que si el error estimado del peso de un neonato está correlacionado con el del siguiente recién nacido se trataría de un caso de embarazo gemelar o múltiple y fueron eliminados del análisis dichos productos; partiendo de la idea lógica que la progenitora no pudo haber procreado dos o más neonatos consecutivos y que no sean considerados como embarazo gemelar o múltiple, entonces no es posible correlacionar dichos errores de estimación del peso de estos recién nacidos.
15. Se realizaron regresiones lineales entre las variables explicativas del modelo de regresión lineal múltiple y no se encontró una correlación significativa entre las variables: la *talla del neonato* y el *promedio de exposición de la madre a las partículas suspendidas totales desde la fecha de gestación hasta los primeros treinta y cinco días* posteriores a esta fecha y el efecto de interacción del *año de gestación* del recién nacido, y desde que la finalidad del modelo fue la de describir y no la predicción del peso del bebé, quedó justificada la ausencia de multicolinealidad severa.

16. Cabe mencionar que *el nivel socioeconómico de la madre*, no fue una variable estadísticamente significativa, y esto se debe a que el estatus promedio de las madres fue por arriba de la clase media baja, es decir, por tratarse de un hospital privado de México, no presentaron los datos significancia estadística, ya que es una población homogénea con nivel socioeconómico medio.
17. *La edad de la madre*, no desempeñó un papel importante porque no se contó con los datos de las edades en forma total, sólo de aquellos casos donde la edad de la madre cayó en rangos riesgosos, es decir ésta fue menor que 20 años o mayor que 35 años, por esta razón no se pudo evaluar el efecto de esta variable, y que en estudios de foros internacionales se muestra estadísticamente significativa.
18. Para la regresión a través del origen (sin intercepto) el coeficiente de determinación R^2 representa la proporción de la variabilidad de explicación con respecto del origen, este valor no puede ser comparado con el coeficiente de determinación R^2 cuando el intercepto es incluido.

CONCLUSIONES

La muestra poblacional que se implementó en el análisis incluyó 5787 neonatos gestados entre los años 1992 y 1997 y nacidos entre 1993 y 1997, de producto único, de término, es decir de 36 a 42 semanas de gestación, cuya información acerca del nacimiento es completa y definida. Del estudio se desprendieron las siguientes resoluciones:

PRIMERA.- Para los años 1993, 1994, 1995 y 1997 se tomó como referencia el año 1996 y se encontró una asociación entre la masa corporal del recién nacido y la exposición de la madre a la contaminación producida por las partículas suspendidas totales, desde la fecha de gestación hasta los primeros treinta y cinco días embarazo.

SEGUNDA.- El mejor modelo de regresión lineal del *peso al nacer* en gramos del neonato, fue ajustado con la *talla del recién nacido*, el efecto de interacción del *promedio de exposición de la madre a las partículas suspendidas totales desde la fecha de gestación hasta los primeros treinta y cinco días y el año de embarazo.*

TERCERA.- El mejor modelo de regresión logística del *bajo peso al nacer* en gramos del neonato, fue ajustado con la *talla del recién nacido*, y el *promedio de exposición de la madre a las partículas suspendidas totales desde la fecha de gestación hasta los primeros treinta y cinco días de embarazo.*

CUARTA.- La exposición de la *madre* a las partículas suspendidas totales desde la fecha de gestación hasta los primeros treinta y cinco días de embarazo produjo la pérdida de peso en el producto para los años 1993, 1994 1995 y 1997 en relación con el año 1996, por lo que se trata de un riesgo severo ya que durante las tres primeras semanas de la vida prenatal, el producto se encuentra en formación, posteriormente en el período embrionario se establecen todos los sistemas y órganos principales del cuerpo y la mayor parte de las características de la forma externa. Cabe mencionar que en esta etapa el ser humano es muy sensible a cierto tipo de factores adversos los cuales afectan el desarrollo de los órganos y pueden producir malformaciones congénitas.

QUINTA.- Para los neonatos de esta muestra poblacional, la *talla* promedio fue de 50.12 centímetros , ver tabla 4 del apartado A.5.4., y tomando como referencia el año 1996, por cada 100 $\mu g/m^3$ de concentración de las partículas suspendidas totales se pierden en promedio 4.38, 9.40, 11.82 y 20.83 gramos en el peso del bebé cuando éstos nacieron en los años 1993, 1994, 1995 y 1997 respectivamente.

SEXTA.- Para los neonatos de esta muestra poblacional, la *talla* promedio fue de 50.12 centímetros, por cada incremento de 100 $\mu g/m^3$ de concentración de las partículas suspendidas totales, el riesgo de bajo peso en los bebés se incrementó 1.95 veces.

SÉPTIMA.- De manera preventiva para no afectar el peso del neonato, se recomienda no exponerse a niveles altos de las partículas suspendidas totales durante los periodos de embarazo.

OCTAVA.- Es de observarse que en la literatura médica universal existen otras variables que juegan un papel importante en la determinación del riesgo del bajo peso al nacer como son: la edad, el peso y la talla de la madre, los antecedentes genéticos y las patologías presentadas anterior y durante la

gestación, la nutrición y el hábito de fumar de la progenitora así como el intervalo entre los embarazos.

APENDICE

ANEXO A

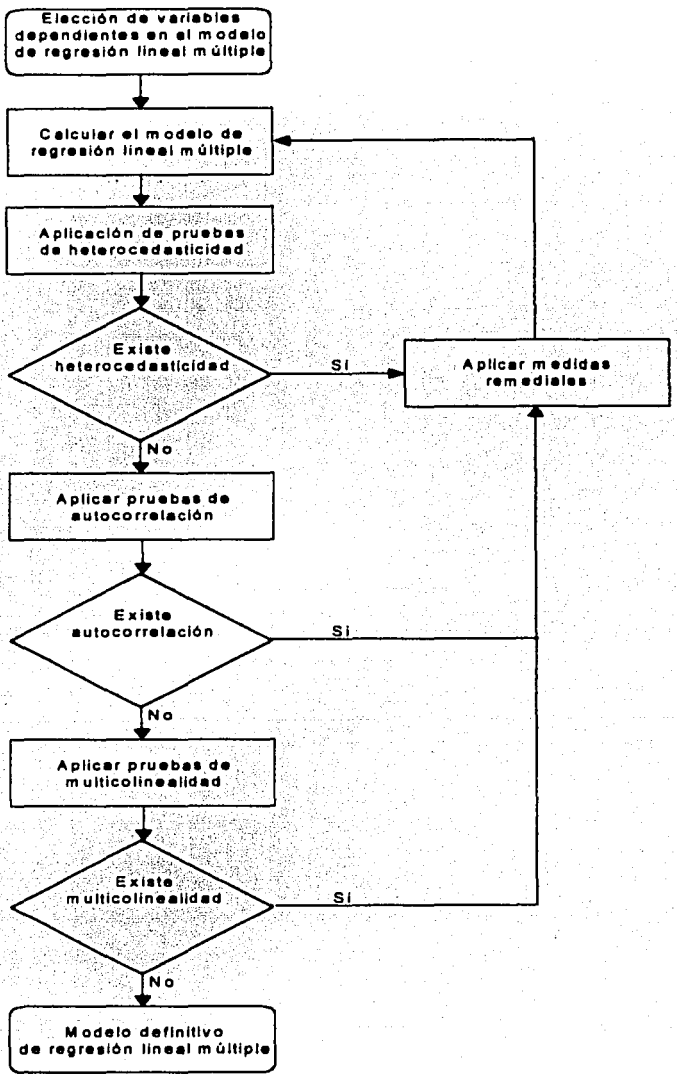
DEMOSTRACIÓN DE LOS RESULTADOS

A.1. Construcción del modelo de regresión lineal múltiple

La metodología seguida para construir el modelo de regresión lineal múltiple fue la siguiente:

1. Se analizaron de forma individual y conjuntas las variables de la base de datos (incluyendo las transformaciones en ellas) en el modelo de regresión lineal univariado y múltiple respectivamente.
2. Se eliminaron aquellas variables que no fueron significativas estadísticamente en los modelos de regresión lineal.
3. Una vez determinadas las variables significativas, se procedió a construir el modelo de regresión lineal preliminar y así iniciar el análisis estadístico.
4. Se aplicó la prueba de Breusch-Pagan para determinar la existencia en el modelo de regresión lineal múltiple de la heterocedasticidad.
5. Se ponderó el modelo para crear un modelo de regresión lineal homocedástico.
6. Se implementó la prueba de Durbin-Watson de autocorrelación de primer orden para determinar la existencia de errores correlacionados.
7. Se utilizaron regresiones lineales para probar la existencia de multicolinealidad entre las variables del modelo de regresión original.

Se vertió el procedimiento en el diagrama de flujo que fue un auxiliar en el paquete estadístico STATA versión 7.0.



Las pantallas del software donde se ejecutaron los comandos implementados para la construcción de los modelos de regresión lineal y logística fueron fotografiadas.

A.1.1. Generalidades de la base de datos

Paso número:

1. *cd d:*

   Se indicó la ubicación del archivo de texto que contiene la información de los neonatos, de la progenitora y de los promedios de exposición de la madre a las partículas suspendidas totales.

2. *set memory 80m*

   Se fijó la memoria requerida para el análisis estadístico a realizar en función del tamaño del archivo de texto.

3. *insheet using datos.txt*

   Se solicitó sea leído el archivo de nombre datos.txt, archivo cuyas características están descritas en la tabla 6 del apartado A.5.6., que corresponde a la base de datos que contiene dicha información para el análisis.

4. *drop if zona==0*

   Se eliminaron los casos donde la zona de residencia de la madre del bebé no estuvo definida (147 observaciones).

5. *drop if zona>=6*

   Se restaron los registros donde la zona de residencia de la progenitora del neonato fue la Zona Metropolitana del Valle de Toluca y provincia (2950 casos).

6. *drop if ege<36*

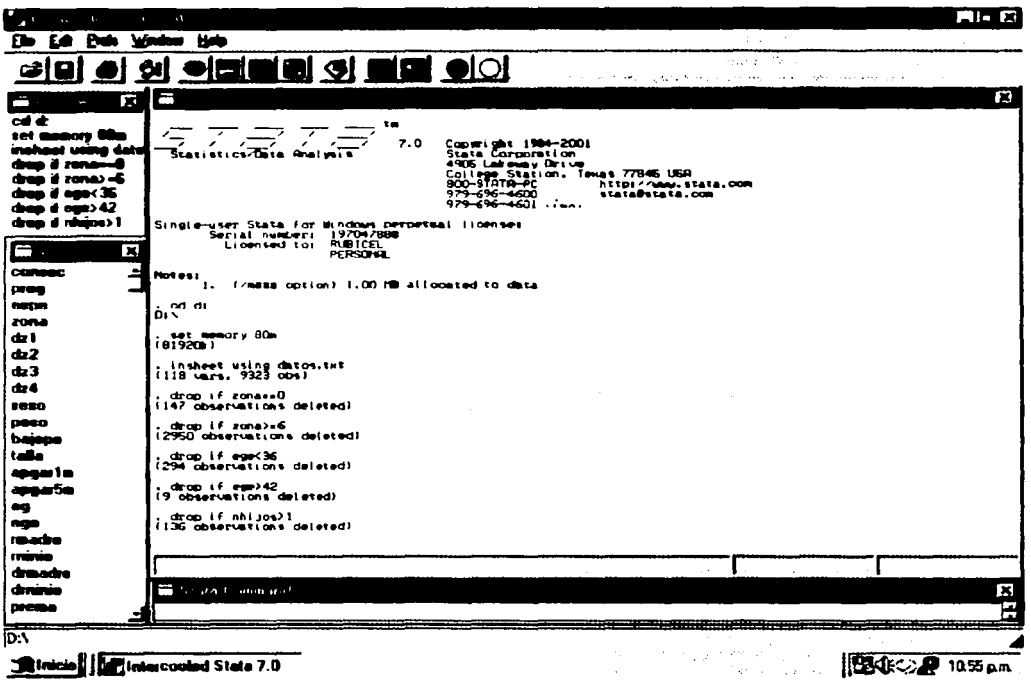
   Se excluyeron los neonatos de pretérmino, menores de 36 semanas de gestación (294 registros).

7. *drop if ege>42*

   Se eliminaron los bebés de postérmino, mayores de 42 semanas de gestación (9 observaciones).

8. *drop if nhijos>1*

   Se descartaron los casos de embarazo gemelar y múltiple (136 neonatos).



A.1.2. Generando las variables implementadas en el modelo

Paso número:

9. *generate r3peso* = peso^*(1/3)*

   Se generó la variable temporal: *raíz cúbica del peso del neonato.*

10. *generate talla2 = talla*^2

    Se construyó la variable temporal: *segunda potencia de la talla del neonato.*

11. *generate da1p5gesp =da1*p5gesp*
12. *generate da2p5gesp =da2*p5gesp*
13. *generate da3p5gesp =da3*p5gesp*

14. *generate da4p5gesp =da4*p5gesp*

15. *generate da5p5gesp =da5*p5gesp*

Se parametrizó el año de gestación del neonato por medio de variables dummies de la siguiente forma:

Para el año 1992, da3=1 y da1=da2=da4=da5=0.

Para el año 1993, da1=1 y da2=da3=da4=da5=0.

Para el año 1994, da2=1 y da1=da3=da4=da5=0.

Para el año 1995, da4=1 y da1=da2=da4=da5=0.

Para el año 1996, da1=da2=da3=da4=da5=0.

Para el año 1997, da5=1 y da1=da2=da3=da4=0.

Así pues, los pasos 11, 12, 13, 14 y 15 generaron el producto del promedio de exposición de la madre a las partículas suspendidas totales durante las primeras cinco semanas de gestación y el año de gestación del neonato.

16. *regress r3peso talla2 da1p5gesp da2p5gesp da3p5gesp da4p5gesp da5p5gesp*

Se implementó la regresión lineal múltiple de la variable dependiente *raíz cúbica del peso del neonato* con *la segunda potencia de la talla del bebé y el producto del año de gestación del bebé con el promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de embarazo.*

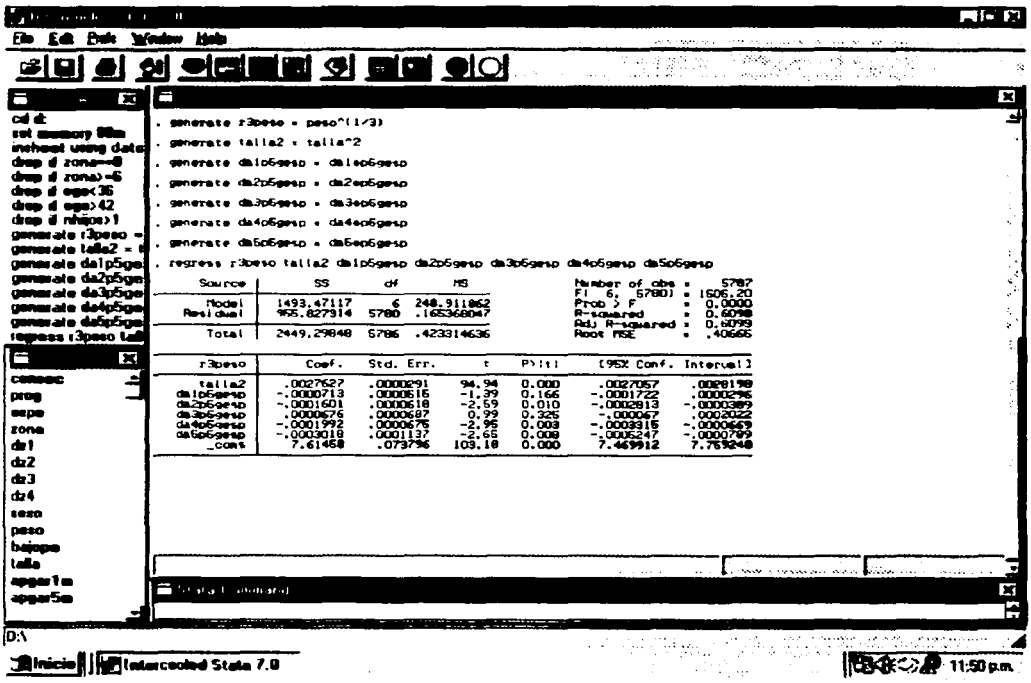
El valor del R-cuadrado del modelo de regresión lineal multivariado fue de 0.6098 y 0.6093 para el R-cuadrado ajustado.

La desviación estándar fue de 0.40665, un valor grande para los valores arrojados de la *raíz cúbica del peso del neonato.*

El coeficiente de la *segunda potencia de la talla del neonato* fue de 0.0027627 con un error estándar de 0.0000291 y el valor de la prueba t fue de 94.94, lo que significa que fue significativa.

Los coeficientes de los productos de las variables dummies con el promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de gestación fueron variantes,

cabe mencionar que dos de estos cinco coeficientes no fueron significativos, pero el resto si lo fue, por lo que se aceptó continuar el análisis estadístico.



A.1.3. La heterocedasticidad en el modelo

Se implementó la prueba de Breusch-Pagan para probar la existencia de la heterocedasticidad en la muestra.

Paso número:

17. *predict r, residual*

Se construyeron los residuos del modelo de regresión lineal múltiple.

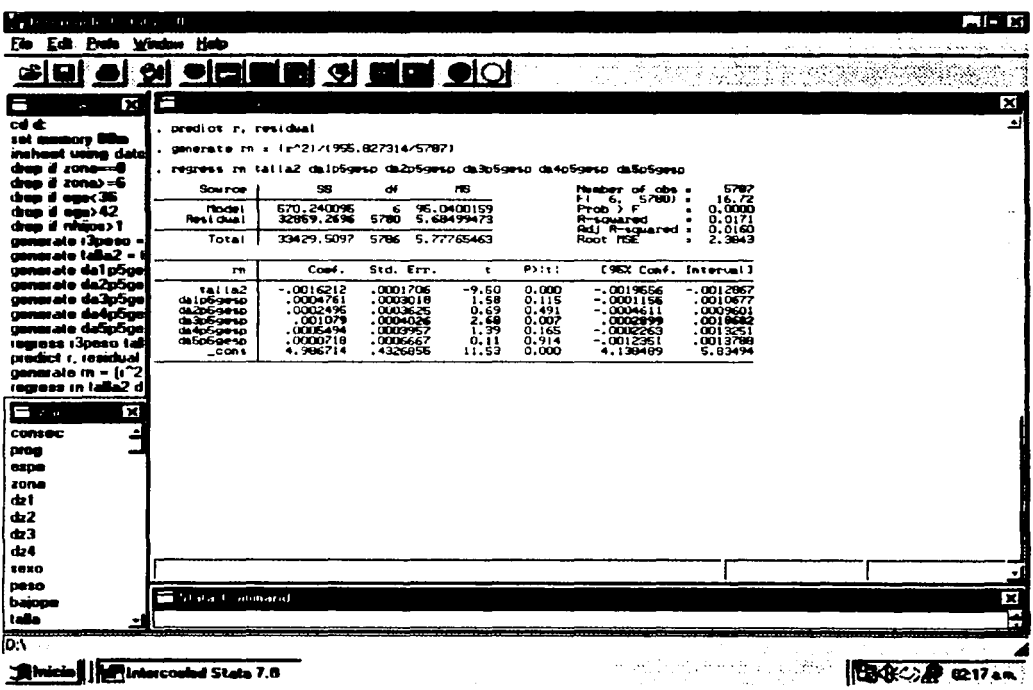
18. *generate rn = (r^2)/(955.827314/5787)*

Se generaron los residuos normalizados, según la prueba que se implementó.

19. *regress rn talla2 da1p5gesp da2p5gesp da3p5gesp da4p5gesp da5p5gesp*

Se implementó la regresión lineal de los residuos normalizados y las variables independientes: *la segunda potencia de la talla del neonato, el producto de exposición de la madre a las partículas suspendidas totales y el año de gestación del bebé.*

Comparando el valor de $q = \frac{570.240095}{2} = 285.120048$ con el de una chi-cuadrado con seis grados de libertad a un 95% de confianza que corresponde a 12.6, se desprende que hubo suficiente evidencia estadística para rechazar la hipótesis nula de homocedasticidad.



A.1.4. Estructura de la varianza de la raíz cúbica del peso del neonato

Paso número:

20. *generate r2 = r^2*

    Se generaron los residuos elevados al cuadrado de la regresión lineal múltiple de la *raíz cúbica del peso sobre la segunda potencia de la talla del neonato y el producto de las variables dummies del año de gestación con el promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de embarazo.*

21. *generate talla3 = talla^3*

    Se construyó la variable temporal *tercera potencia de la talla del neonato.*

22. *generate talla4 = talla^4*

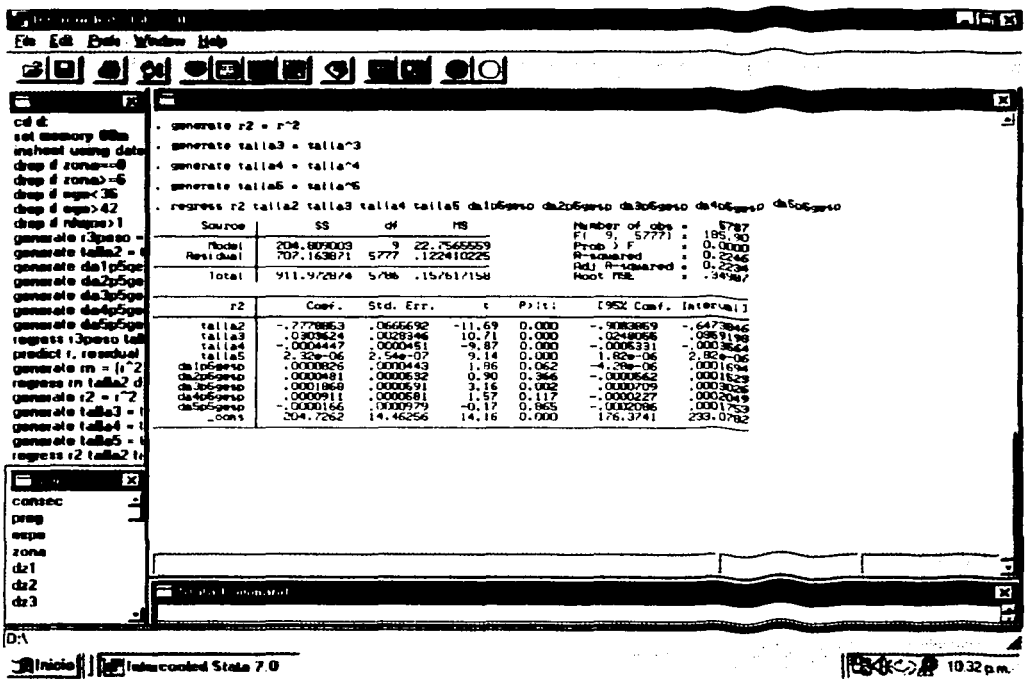
    Se generó la variable temporal *cuarta potencia de la talla del recién nacido.*

23. *generate talla5 = talla^5*

    Se diseñó la variable temporal *quinta potencia de la talla del bebé.*

24. *regress r2 talla2 talla3 talla4 talla5 da1p5gesp da2p5gesp da3p5gesp da4p5gesp da5p5gesp*

    Se implementó la regresión lineal múltiple de los residuos elevados al cuadrado, como variable dependiente, la *segunda, la tercera, la cuarta, la quinta potencia de la talla del neonato, y el producto de las variables dummies del año de gestación con el promedio de exposición de la madre durante los primeros treinta y cinco días del embarazo* como variables independientes a fin de conocer la estructura de la varianza de *la raíz cúbica del peso del bebé.*



A.1.5. Corrección de la heterocedasticidad en el modelo

Paso número:

25. *predict obs*

Se generaron los valores de predicción de la varianza para cada uno de los registros en el análisis.

26. *generate p = 1/(obs^(1/2))*

Se creó la ponderación de las variables extrayendo la inversa de la raíz cuadrada de dichos valores de predicción generados en el punto anterior.

27. *generate r3pesop = r3peso*p*

Se generó la variable ponderada de la *raíz cúbica del peso.*

28. *generate talla2p = talla2*p*

Se produjo la variable ponderada de *la segunda potencia de la talla del neonato.*

29. *generate da1p5gespp = da1p5gesp*p*
30. *generate da2p5gespp = da2p5gesp*p*
31. *generate da3p5gespp = da3p5gesp*p*
32. *generate da4p5gespp = da4p5gesp*p*
33. *generate da5p5gespp = da5p5gesp*p*

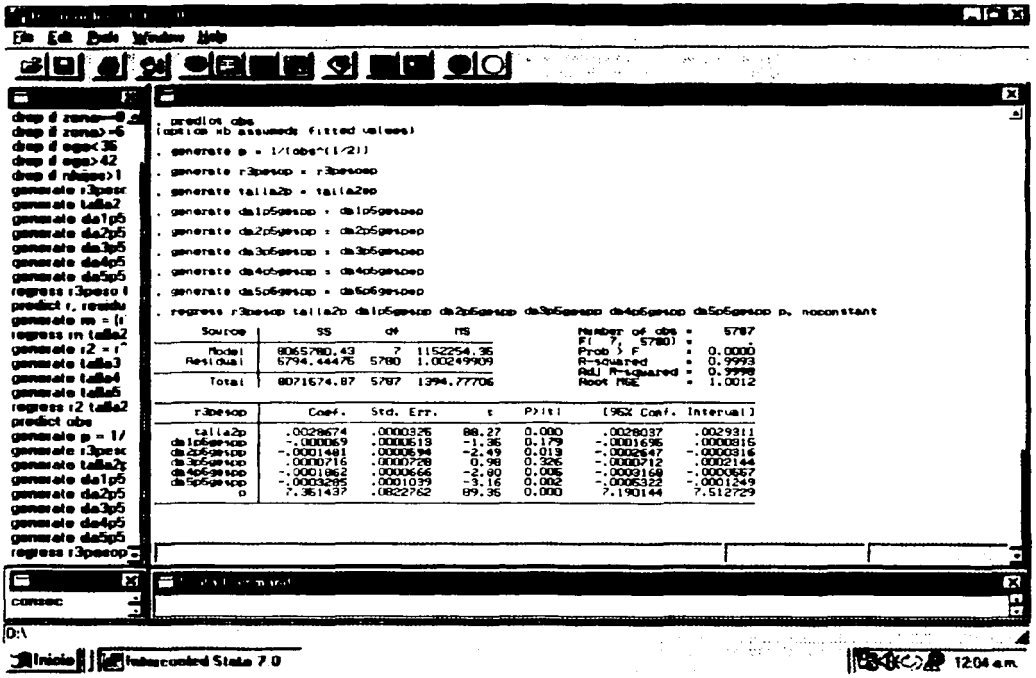
    Se diseñaron las variables ponderadas del producto del promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de gestación y *las variables dummies del año de gestación.*

34. *regress r3pesop talla2p da1p5gespp da2p5gespp da3p5gespp da4p5gespp da5p5gespp p, noconstant*

    Se implementó la regresión lineal de las variables ponderadas, sin el término constante, modelo de regresión lineal múltiple generalizado.

    Para la regresión a través del origen, el R-cuadrado representa la proporción del grado de variabilidad explicada a través del origen, este valor no puede ser comparado con el R-cuadrado cuando el intercepto está incluido.

    Comparando las varianzas de las regresiones lineales con y sin ponderación se descartó que es mayor cuando se pondera, se debió a que la varianza de la raíz cúbica del peso decrece conforme la talla del neonato aumenta, por lo que el error estándar del término independiente y de la longitud del bebé se incrementaron también, y en el caso del error estándar del coeficiente que explica el promedio de exposición de la madre al contaminante disminuyó.



A.1.6. Verificando que el modelo ponderado sea homocedástico

Paso número:

35. *predict res, residual*

Se generaron los residuos del modelo de regresión lineal múltiple ponderado.

36. *generate resn = (res^2)(8065780.43/5787)*

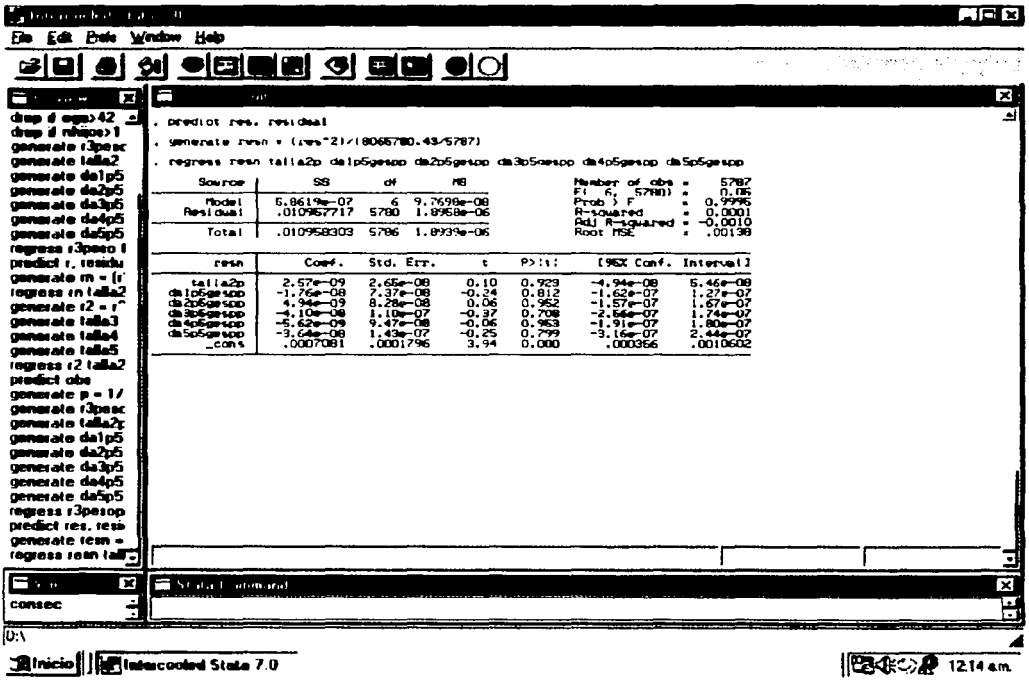
Se construyen los residuos normalizados a fin de implementar la prueba de Breusch-Pagan.

37. *regress resn talla2p da1p5gespp da2p5gespp da3p5gespp da4p5gespp da5p5gespp*

Se implementa la regresión lineal sobre los residuos normalizados y las variables independientes ponderadas: ***la segunda potencia de la talla del neonato y el producto de las variables dummies del año de gestación con***

el promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de gestación.

De esto se desprendió que el valor correspondiente a $q = \frac{5.8619e-07}{2} = 2.93095e-07$ y comparándolo con el de una chi-cuadrado con seis grados de libertad a un 95% de confianza que corresponde a 12.6, se concluyó que no existió suficiente evidencia estadística para rechazar la hipótesis nula de homocedasticidad.



A.1.7. Autocorrelación de primer orden en el modelo

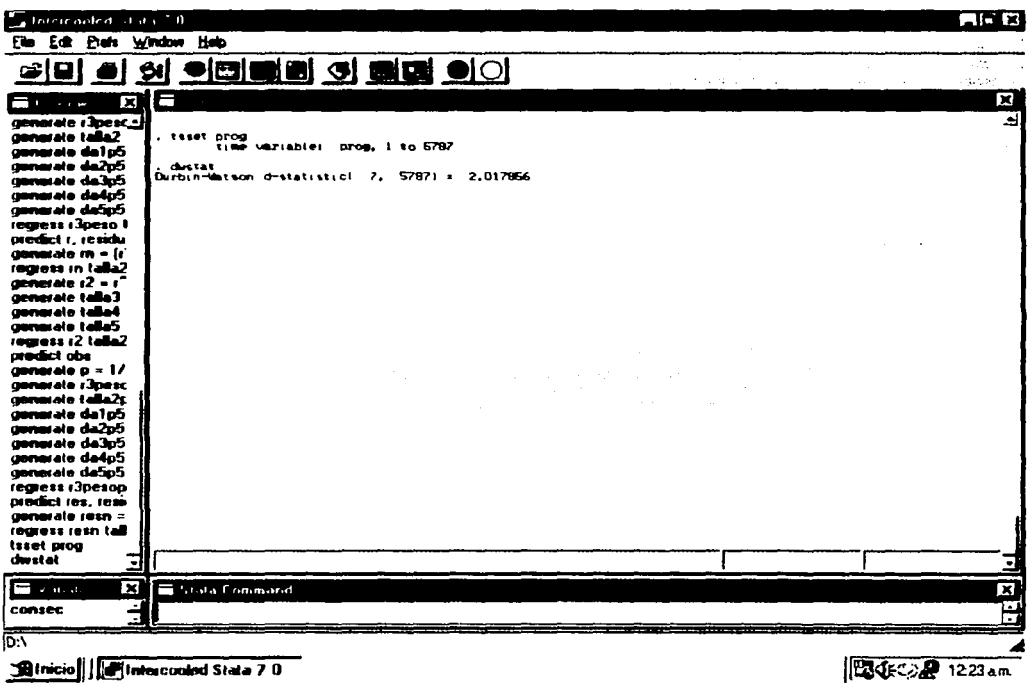
Paso número:

38. tsset *prog*

Se estableció la variable *prog* como auxiliar en el análisis de la autocorrelación de primer orden, que corresponde a un número consecutivo de la muestra reducida, 5787 casos.

39. dwstat

Se calculó el estadístico de Durbin-Watson. Se implementó la prueba asintótica para probar la existencia de autocorrelación de primer orden, con un valor de ρ de –0.008928 por lo que se obtuvo una prueba z de 0.6791738, que para una prueba de significación del 5% (z = 1.96), no existió evidencia estadística para rechazar la hipótesis nula de ausencia de autocorrelación.



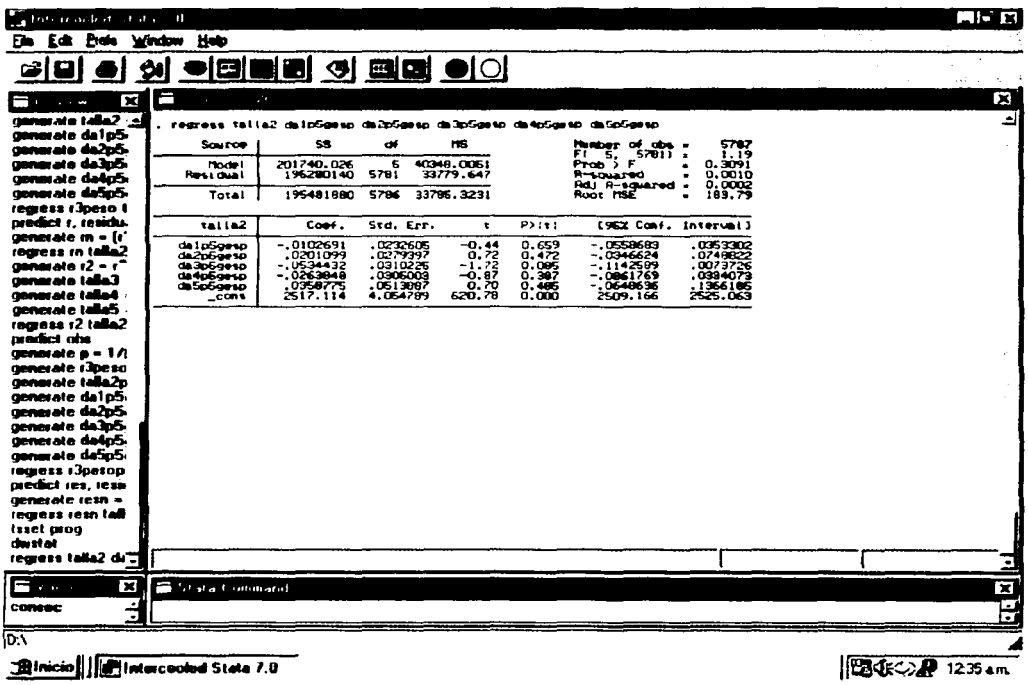
A.1.8. Ausencia de la multicolinealidad severa en el modelo

Paso número:

40. *regress talla2* da1p5gesp *da2p5gesp da3p5gesp da4p5gesp da5p5gesp*

Se implementaron regresiones lineales entre las variables ponderadas: *la segunda potencia de la talla del neonato y el producto de las variables dummies del año de gestación con el promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de embarazo.*

No se encontró una *relación* estadística significativa entre estas variables no ponderadas, por lo que se puede afirmar que no existe multicolinealidad entre ellas.



A.2. Construcción del modelo de regresión logística múltiple

A.2.1. Generando las variables implementadas en el modelo

Paso número:

41. *generate talla4 = talla^4*

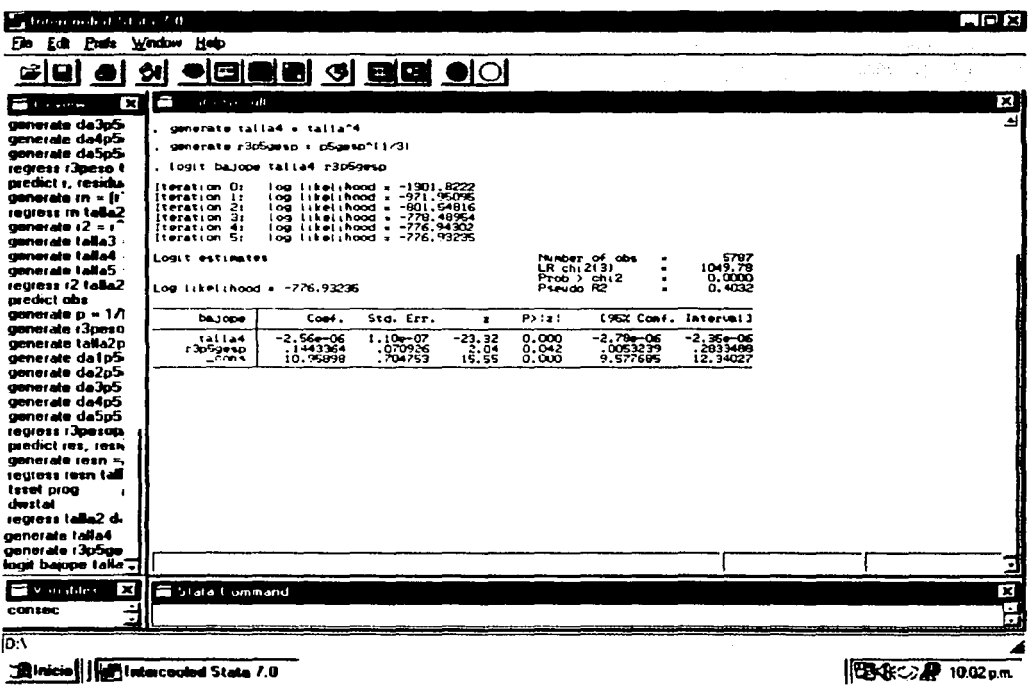
Se generó la variable cuarta potencia de la talla del neonato.

42. generate r3p5gesp = p5gesp^(1/3)

Se generó la variable raíz cúbica del promedio de exposición de la madre a las partículas suspendidas totales durante las cinco primeras semanas de gestación.

43. *logit bajope talla4 r3p5gesp*

Se construyó el modelo de regresión logística múltiple con las variables generadas.



A.3. Programas auxiliares

A.3.1. Interpolación lineal

La sintaxis corresponde a la estructura de los programas elaborados en MATLAB versión 5.0, que fueron implementados para calcular los valores faltantes en el archivo que contiene los registros de las partículas suspendidas totales.

A.3.1.1. Programa 1

Arroja un vector que corresponde a los índices de las partículas suspendidas totales diarios.

```
clc
clear
diary on
x(:,1)=[ ]; %progresivo del día conocido
x(:,2)=[ ]; %fecha conocida
y=[ ]; %índice de contaminación conocido
m1=x(1,2);
d=x(1,1);
[t b]=size(x);
e=x(t);
for h=d:6:e
m(h:h+5)=[m1,m1+1,m1+2,m1+3,m1+4,m1+5];
m1=m1+6;
end
m=m';
m(:,2)=m;
for g=d:e
m(g,1)=g;
end
s=1;
r=1;
```

```
while s<t
w=x(r:r+1,2);
z=y(r:r+1);
c=inter(w,z);
p1=x(r,1);
p2=x(r+1,1);
pval(p1:p2)=horner(c,w,m(p1:p2,2));
r=r+1;
s=s+1;
end
pval=pval';
pval(d:length(pval))
diary off
```

A.3.1.2. Programa 2

Función que calcula el valor interpolado dado dos números.

```
function pval=horner(c,x,z)
n=length(c);
pval=c(n)*ones(size(z));
for k=n-1:-1:1
pval=(z-x(k)).*pval + c(k);
end
```

A.3.1.3. Programa 3

Crea la función que interpolará los datos del archivo de registros de los índices de las particulas suspendidas totales.

```
function c=inter(x,y)
n=length(x);
c=zeros(n,1);
c(1)=y(1);
```

```
if n>1
  c(2:n)=inter(x(2:n),(y(2:n)-y(1))./(x(2:n)-x(1)));
end
```

A.3.2. Promedios de exposición de la madre

La sintaxis corresponde a la estructura de los programas elaborados en CLIPPER versión 5.3, implementados para calcular los promedios de exposición de la madre a las partículas suspendidas totales durante diversos periodos del embarazo, alternándose las variables según lo especifican los bloques que se encuentran descritos en la tabla 5 del apartado A.5.5.

```
CLS
PUBLIC FECHA1:=' / / ', ZONA1:=0, CONTA:=0, CONTAR:=0,
CONTADOR:=0, bloque 1
SELECT 1
USE DATOS
INDEX ON FECHGES TO DATOS
SELECT 2
USE PST
INDEX ON FECHA TO PST
SELECT 1
GO TOP
DO WHILE !EOF()
FECHA1=1->bloque 2
ZONA1=1->ZONA
bloque 3
CONTADOR:=0
CONTAR:=0
SELECT 2
GO TOP
SEEK FECHA1
```

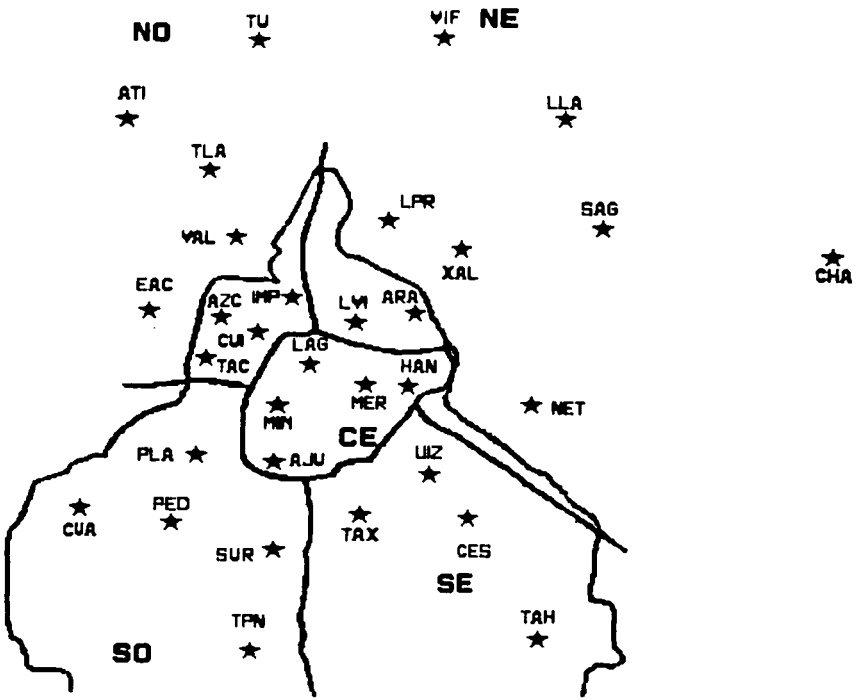
```
IF FOUND()
DO WHILE CONTADOR <= bloque 4
IF ZONA1=1
SUMATLA()
ENDIF
IF ZONA1=2
SUMAXAL()
ENDIF
IF ZONA1=3
SUMAMER()
ENDIF
IF ZONA1=4
SUMAPED()
ENDIF
IF ZONA1=5
SUMACES()
ENDIF
SELECT 2
SKIP
CONTADOR=CONTADOR+1
ENDDO
OPERA()
CONTA:=0
CONTAR:=0
ENDIF
SELECT 1
SKIP
IF !EOF()
FECHA1=1->bloque 2
ZONA1=1->ZONA
bloque 3
```

```
ELSE
EXIT
ENDIF
ENDDO
CLOSE DATA
RETURN
FUNCTION SUMATLA()
CONTA=CONTA+2->TLA
RETURN
FUNCTION SUMAXAL()
CONTA=CONTA+2->XAL
RETURN
FUNCTION SUMAMER()
CONTA=CONTA+2->MER
RETURN
FUNCTION SUMAPED()
CONTA=CONTA+2->PED
RETURN
FUNCTION SUMACES()
CONTA=CONTA+2->CES
RETURN
FUNCTION OPERA()
SELECT 1
bloque 5
CONTAR=CONTA/bloque 6
REPLACE 1->bloque 7 WITH CONTAR
```

A.4. Planos y gráficas

A.4.1. Plano 1

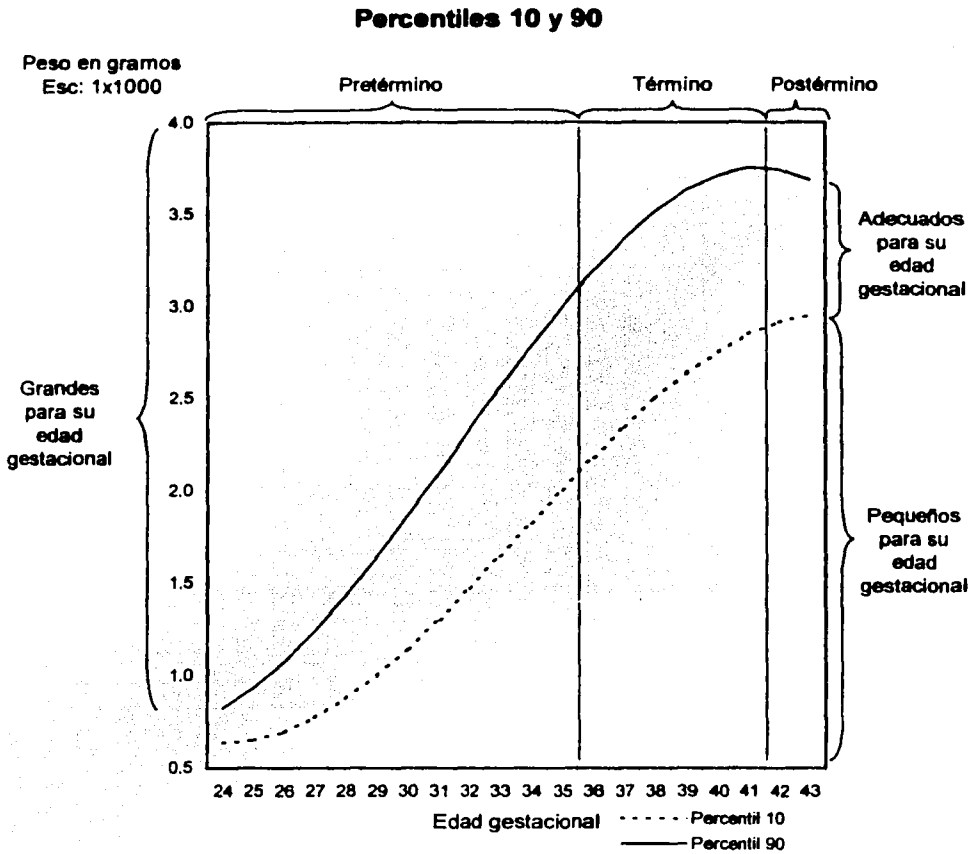
Distribución del sistema de monitoreo atmosférico de la Zona Metropolitana del Valle de México[59]



Muestra las estaciones de la Red Automática de Monitoreo Atmosférico (RAMA) de la Zona Metropolitana del Valle de México, cuya nomenclatura está descrita en la tabla 2 del apartado A.5.2.

[59] **INSTITUTO NACIONAL DE ECOLOGÍA.** op. cit. pág. 25.

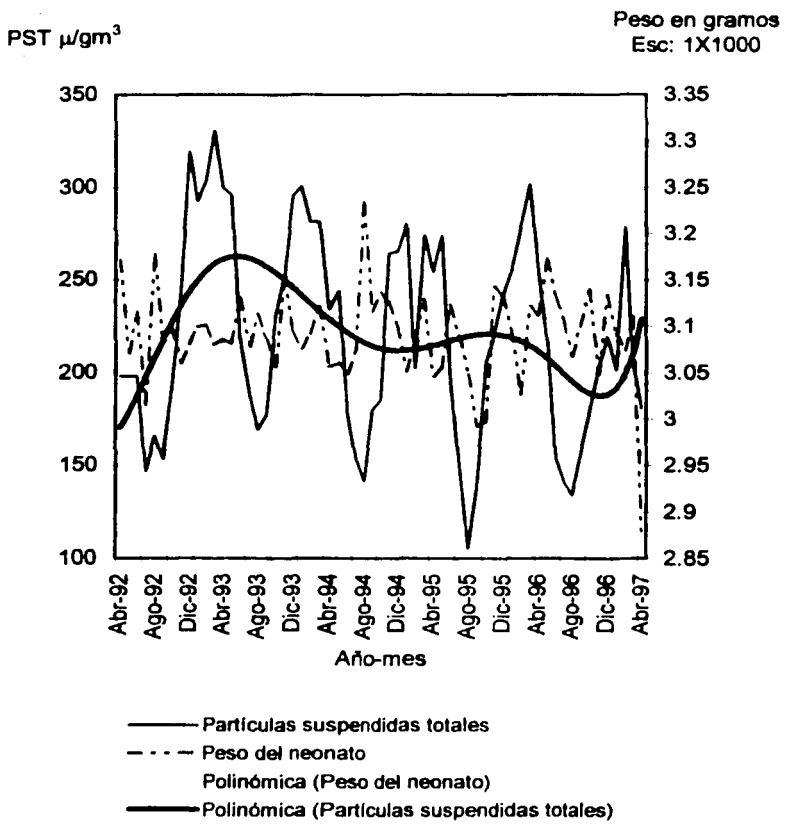
A.4.2. Gráfica 1



Representa en el eje vertical el peso de nacimiento del niño y en el horizontal las semanas de gestación, las curvas dibujadas en ella muestran el límite superior e inferior normal de peso a cada edad gestacional (percentil 90 y 10 respectivamente).

A.4.3. Gráfica 2

Media del peso del neonato vs concentración de las PST



Muestra la relación gráfica existente entre los promedios de las partículas suspendidas totales y del peso del neonato al nacer.

A.5. Tablas

A.5.1. Tabla 1

Valores normados para los contaminantes[60]

Contaminante	Valores límite		
	Exposición aguda		Exposición crónica
	Concentración y tiempo promedio	Frecuencia máxima aceptable	(Para protección de la salud de la población susceptible)
Ozono (O_3)	0.11 ppm (1 hora)	1 vez cada 3 años	-
Dióxido de azufre (SO_2)	0.13 ppm (24 horas)	1 vez al año	0.03 ppm (*)
Dióxido de nitrógeno (NO_2)	0.21 ppm (1 hora)	1 vez al año	-
Monóxido de carbono (CO)	11 ppm (8 horas)	1 vez al año	-
Partículas suspendidas totales (PST)	260 $\mu g/m^3$ (24 horas)	1 vez al año	75 $\mu g/m^3$ (*)
Partículas fracción respirable (PM10)	150 $\mu g/m^3$ (24 horas)	1 vez al año	50 $\mu g/m^3$ (*)
Plomo (Pb)	-	-	1.5 $\mu g/m^3$ (**)

(*) Promedio aritmético anual

(**) Promedio aritmético de 3 meses

[60] **INSTITUTO NACIONAL DE ECOLOGÍA.** op. cit. pág. 12

A.5.2. Tabla 2

Estaciones de la Red Automática de Monitoreo Atmosférico de la Zona Metropolitana del Valle de México[61]

Estación	Clave	Estación	Clave
Aragón	ARA	Laureles	LLA
Atizapán	ATI	Merced	MER
Azcapotzalco	AZC	Netzahualcóyotl	NET
Benito Juárez	BJU	Pedregal	PED
Cerro de la Estrella	CES	Plateros	PLA
Chapingo	CHA	San Agustín	SAG
Coacalco	VIF	Santa Ursula	SUR
Cuajimalpa	CUA	Tacuba	TAC
Cuitláhuac	CUI	Taxqueña	TAX
ENEP-Acatlán	EAC	Tláhuac	TAH
Hangares	HAN	Tlalnepantla	TLA
I.M.P.	IMP	Tlalpan	TPN
Insurgentes	MIN	Tultitlán	TU
La Presa	LPR	UAM Iztapalapa	UIZ
La Villa	LVI	Vallejo	VAL
Lagunilla	LAG	Xalostoc	XAL

La tabla contiene la nomenclatura del plano 1 "Distribución del sistema de monitoreo atmosféricos de la Zona Metropolitana de la Ciudad de México" del apartado A.4.1.

[61] **INSTITUTO NACIONAL DE ECOLOGÍA.** op. cit. pág. 26.

A.5.3. Tabla 3

Estadísticas descriptivas de la madre y del neonato

Variable	N°	Peso al nacer		Bajo peso al nacer	
		Media	Desv. Est.	N°	Media
		Datos de la madre			
Edad					
15	1	3,150.00	0.00	0	0.00
16	4	3,077.50	258.54	0	0.00
17	1	2,690.00	0.00	0	0.00
19	3	2,633.33	472.48	1	33.33
≥20 y ≤34	5346	3,093.64	407.13	323	6.04
35	11	2,822.73	388.13	2	18.18
36	133	3,180.30	392.58	1	0.75
37	105	3,090.03	380.85	3	2.86
38	88	3,079.83	435.44	7	7.95
39	43	3,133.02	428.39	2	4.65
40	21	3,060.71	340.30	2	9.52
41	13	3,203.85	469.07	0	0.00
42	13	3,137.31	569.40	2	15.38
43	3	3,035.00	69.46	0	0.00
45	2	3,595.00	21.21	0	0.00
Fuma					
No	5753	3,099.64	403.42	319	5.54
Si	34	2,342.50	368.98	24	70.59
Riesgos o patologías					
No	5327	3,114.02	389.75	241	4.52

Variable	N°	Peso al nacer		Bajo peso al nacer	
		Media	Desv. Est.	N°	Media
Si	460	2,877.17	526.40	102	22.17
Zona de residencia					
Noroeste	1129	3,107.59	417.50	61	5.40
Nordeste	636	3,109.27	416.78	42	6.60
Centro	1546	3,089.61	406.40	98	6.34
Sudoeste	984	3,107.94	399.13	46	4.67
Sureste	1492	3,077.20	401.47	96	6.43
Nivel socioeconómico					
0	3	2,610.00	115.33	1	33.33
1	293	3,078.82	382.20	17	5.80
2	625	3,116.26	428.30	36	5.76
3	1092	3,098.89	400.08	56	5.13
4	1553	3,086.84	399.13	99	6.37
5	1815	3,086.05	414.00	117	6.45
6	383	3,145.10	407.55	16	4.18
7	23	3,073.70	442.94	1	4.35
Cuenta con póliza de gastos médicos mayores					
No	687	3,071.96	398.08	46	6.70
Si	5100	3,098.33	408.50	297	5.82
		Datos del neonato			
Año de gestación					
1993	1270	3,093.37	404.20	70	5.51
1994	1199	3,095.73	402.94	69	5.75
1995	1111	3,100.95	413.51	74	6.66
1996	1099	3,076.72	398.50	65	5.91

Variable	N°	Peso al nacer		Bajo peso al nacer	
		Media	Desv. Est.	N°	Media
1997	1108	3,109.26	417.95	65	5.87
Año de nacimiento					
1992	930	3,097.47	409.49	49	5.27
1993	1222	3,096.27	397.33	65	5.32
1994	1159	3,102.42	413.40	76	6.56
1995	1077	3,077.52	399.34	71	6.59
1996	1136	3,105.04	416.37	65	5.72
1997	263	3,080.17	412.70	17	6.46
Edad gestacional					
≥36 y <37	237	2,667.64	455.86	78	32.91
≥37 y <38	668	2,888.71	399.43	95	14.22
≥38 y <39	1688	3,067.40	380.97	85	5.04
≥39 y <40	1798	3,147.27	379.87	58	3.23
≥40 y <41	1095	3,218.32	370.97	22	2.01
≥41 y <42	258	3,289.40	365.53	4	1.55
≥42 y <43	43	3,272.67	483.34	1	2.33
Talla					
≥34 y <36	1	2,855.00	0.00	0	0.00
≥36 y <38	1	1,120.00	0.00	1	100.00
≥38 y <40	2	2,065.00	1,378.86	1	50.00
≥40 y <41	6	2,533.33	710.31	3	50.00
≥41 y <42	2	1,407.50	88.39	2	100.00
≥42 y <43	6	1,942.50	325.17	5	83.33
≥43 y <44	7	1,927.14	284.78	7	100.00
≥44 y <45	18	2,186.67	406.29	16	88.89

Variable	N°	Peso al nacer		Bajo peso al nacer	
		Media	Desv. Est.	N°	Media
≥45 y <46	41	2,275.12	298.10	33	80.49
≥46 y <47	93	2,437.35	254.20	54	58.06
≥47 y <48	223	2,555.73	246.33	82	36.77
≥48 y <49	515	2,689.86	228.26	83	16.12
≥49 y <50	933	2,852.59	219.31	34	3.64
≥50 y <51	1668	3,085.02	241.05	17	1.02
≥51 y <52	1159	3,288.85	245.87	4	0.35
≥52 y <53	669	3,428.89	269.52	0	0.00
≥53 y <54	297	3,607.46	328.24	1	0.34
≥54 y <55	88	3,792.66	293.96	0	0.00
≥55 y <56	41	3,886.20	275.07	0	0.00
≥56 y <57	14	4,078.93	260.63	0	0.00
≥57	3	3,836.67	471.07	0	0.00
Sexo					
Masculino	2957	3,144.59	414.11	150	5.07
Femenino	2830	3,043.58	393.63	193	6.82
Apgar al primer minuto de vida					
1	2	2,565.00	1,237.44	1	50.00
2	5	2,385.00	1,131.84	3	60.00
3	8	3,101.88	372.83	0	0.00
4	12	2,997.08	461.27	1	8.33
5	40	3,029.30	487.95	5	12.50
6	108	3,042.22	471.28	14	12.96
7	397	3,043.66	469.55	36	9.07
8	2756	3,090.18	410.04	173	6.28
9	2458	3,114.83	383.15	110	4.48

Variable	N°	Peso al nacer		Bajo peso al nacer	
		Media	Desv. Est.	N°	Media
10	1	3,200.00	0.00	0	0.00
Apgar al quinto minuto de vida					
4	4	2,400.00	1,263.28	2	50.00
5	2	2,565.00	1,237.44	1	50.00
6	3	2,466.67	331.26	1	33.33
7	25	2,994.60	576.67	4	16.00
8	343	3,006.05	483.80	43	12.54
9	5253	3,102.07	398.32	281	5.35
10	157	3,112.45	401.87	11	7.01
Riesgos o patologías					
No	5700	3,104.79	395.65	297	5.21
Si	87	2,466.44	619.05	46	52.87
Síndrome de prematurez					
No	5556	3,113.12	395.07	267	4.81
Si	231	2,663.96	457.98	76	32.90
Tamaño para su edad gestacional					
Adecuado	4771	3,079.36	289.46	95	1.99
Grande	562	3,792.32	236.48	0	0.00
Pequeño	454	2,398.63	275.37	248	54.63
Total	**5787**	**3,095.20**	**407.33**	**343**	**5.93**

A.5.4. Tabla 4

Estadísticas descriptivas de las variables empleadas en los modelos de regresión de regresión lineal y logística

Promedio	Des. Est.	Intervalo de confianza		Valor	
		Inferior -0.95	Superior 0.95	Mín	Máx.
Zona de residencia de la madre					
3.19	1.43	3.15	3.22	1	5
Peso en gramos del neonato					
3,095.20	407.33	3,084.70	3,105.69	1,090	4,670
Bajo peso al nacer					
0.06	0.24	0.05	0.07	0	1
Talla en centímetros					
50.12	1.86	50.07	50.16	34	57
Edad gestacional					
38.82	1.18	38.79	38.86	36	42.6
Promedio de exposición de la madre a las partículas suspendidas totales durante los primeros treinta y cinco días de gestación					
210.85	105.99	208.12	213.58	45	602

A.5.5. Tabla 5

Bloques de variables empleadas en los programas auxiliares para el cálculo de promedios de exposición de la madre a las partículas suspendidas totales

Bloque						
1	2	3	4	5	6	7
			Promedio			
Durante el primer trimestre de gestación						
	Fechges		89		90	pptp
Durante el segundo trimestre de gestación						
	Ist		89		90	pstp
Durante el tercer trimestre de gestación						
dias:=0 diasn:=0	Itt	Dias=1->Diasf	Dias	Diasn=(M->Dias)+1	Diasn	pttp
Durante toda la gestación						
dias:=0 diasn:=0	Fechges	Dias=1->Diasg1	Dias	Diasn=Diasg	Diasn	ptgp
Durante la última semana de gestación						
	F1nac		6		7	p1nacp
Durante las dos últimas semanas de gestación						
	F2nac		13		14	p2nacp
Durante las tres últimas semanas de gestación						
	F3nac		20		21	p3nacp
Durante las cuatro últimas semanas de gestación						
	F4nac		27		28	p4nacp
Durante las cinco últimas semanas de gestación						
	F5nac		34		35	p5nacp

Bloque						
1	2	3	4	5	6	7
Durante las seis últimas semanas de gestación						
	F6nac		41		42	p6nacp
Durante las siete últimas semanas de gestación						
	F7nac		48		49	p7nacp
Durante las ocho últimas semanas de gestación						
	F8nac		55		56	p8nacp
Durante la primera semana de gestación						
	Fechges		6		7	p1gesp
Durante las dos primeras semanas de gestación						
	Fechges		13		14	p2gesp
Durante las tres últimas semanas de gestación						
	Fechges		20		21	p3gesp
Durante las cuatro últimas semanas de gestación						
	Fechges		27		28	p4gesp
Durante las cinco últimas semanas de gestación						
	Fechges		34		35	p5gesp
Durante las seis últimas semanas de gestación						
	Fechges		41		42	p6gesp
Durante las siete últimas semanas de gestación						
	Fechges		48		49	p7gesp
Durante las ocho últimas semanas de gestación						
	Fechges		55		56	p8gesp

A.5.6. Tabla 6

Estructura de la base de datos

prog ---------- PROG

type :	numeric (int)				
range :	[-9,5787]		units:	1	
unique values:	5788		codeed missing:	0 / 9323	
mean:	1792.96				
std. dev:	1927.88				
percentiles:	10%	25%	50%	75%	90%
	-9	-9	1126	3457	4855

explicación : variable progresiva auxiliar para el análisis de autocorrelación
-9 indica que el registro no fue incluido en el análisis

zona ---------- ZONA

type :	numeric (byte)				
range :	[0,9]		units:	1	
unique values:	10		codeed missing:	0 / 9323	
mean:	4.24381				
std. dev:	2.14685				
percentiles:	10%	25%	50%	75%	90%

explicación : zona de residencia de la madre
Incluidas en el análisis:
1 noroeste
2 nordeste
3 centro

4 sudoeste
5 sureste
No incluidas en el análisis:
0 desconocida
6-9 Zona Metropolitana del Valle de Toluca

peso ---------- PESO

type :	numeric (int)				
range :	[435,4720]		units:	1	
unique values:	636		codeed missing:	0 / 9323	
mean:	3036.1				
std. dev:	478.828				
percentiles:	10%	25%	50%	75%	90%
	2485	2775	3060	3345	3600
explicación :	peso en gramos del neonato				

bajope ---------- BAJOPE

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
8364	0		
959	1		
explicación :	0 si es mayor o igual de 2500 gramos		
	1 si es de bajo peso		

talla ---------- TALLA

type :	numeric (float)		
range :	[27,58]	units:	0.1
unique values:	55	codeed missing:	0 / 9323
mean:	49.8148		

std. dev:	2.40069				
percentiles:	10%	25%	50%	75%	90%
	47	49	50	51	52
explicación :	longitud del neonato en centímetros				

ege ---------- EGE

type :	numeric (byte)		
range :	[24,44]	units:	1
unique values:	21	codeed missing:	0 / 9323
mean:	38.3606		
std. dev:	1.73692		

percentiles:	10%	25%	50%	75%	90%
	37	38	39	39	40

explicación : edad gestacional del neonato (semanas completas)

nhijos ---------- NHIJOS

type :	numeric (byte)		
range :	[1,3]	units:	1
unique values:	3	codeed missing:	0 / 9323
tabulation:	Freq. Value		
9030	1		
278	2		
15	3		

explicación : número de hijos en el parto

aniog ---------- ANIOG

type :	numeric (int)		
range :	[1992,1997]	units:	1
unique values:	6	codeed missing:	0 / 9323
tabulation:	Freq. Value		
1509	1992		

1988	1993
1866	1994
1702	1995
1803	1996
455	1997

explicación : año de gestación del neonato

da1 ---------- DA1

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
7335	0		
1988	1		

explicación : variable dummy del año de gestación del neonato

da2 ---------- DA2

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
7457	0		
1866	1		

explicación : variable dummy del año de gestación del neonato

da3 ---------- DA3

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
7814	0		

1509	1				
explicación :	variable dummy del año de gestación del neonato				

da4 ---------- DA4

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
7621	0		
1702	1		
explicación :	variable dummy del año de gestación del neonato		

da5 ---------- DA5

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
8868	0		
455	1		
explicación :	variable dummy del año de gestación del neonato		

	dz1	dz2	dz3	dz4	dz5
1992	0	0	1	0	0
1993	1	0	0	0	0
1994	0	1	0	0	0
1995	0	0	0	1	0
1996	0	0	0	0	0
1997	0	0	0	0	1

p5gesp ---------- P5GESP

type :	numeric (int)		
range :	[0,602]	units:	1

unique values:	488		codeed missing:	0 / 9323	
mean:	141.356				
std. dev:	132.138				
percentiles:	10%	25%	50%	75%	90%
	0	0	134	227	320

explicación : promedio de exposición durante las primeras cinco semanas de gestación

p5gesp3 ---------- P5GESP3

type :	numeric (byte)		
range :	[1,3]	units:	1
unique values:	3	codeed missing:	0 / 9323
tabulation:	Freq. Value		
5142	1		
2076	2		
2105	3		

explicación : tertil del promedio de exposición durante las primeras cinco semanas de gestación

dp5gesp1 ---------- DP5GESP1

type :	numeric (byte)		
range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
7247	0		
2076	1		

explicación : variable dummy del tertil del promedio de exposición durante las primeras cinco semanas de gestación

dp5gesp2 ---------- DP5GESP2

type : numeric (byte)

range :	[0,1]	units:	1
unique values:	2	codeed missing:	0 / 9323
tabulation:	Freq. Value		
7218	0		
2105	1		
explicación :	variable dummy del tertil del promedio de exposición durante las primeras cinco semanas de gestación		

	dp5gesp1	dp5gesp2
1	0	0
2	1	0
3	0	1

ANEXO B

OTROS RESULTADOS DEL MARCO TEÓRICO

B.1. Función de probabilidad Bernoulli

B.1.1. Definición

Supóngase que el resultado de un ensayo o experimento puede ser clasificado como: "éxito" o "fracaso".

Sea X = 1 cuando el resultado es un éxito y X = 0 cuando el resultado es un fracaso, entonces la función probabilidad de X está dada por:

$$P(X=x) = \omega_x(p) = \begin{cases} p^x(1-p)^{1-x} & x=0,1 \qquad 0<p<1 \\ 0 & \text{en otro caso} \end{cases}$$

Donde p corresponde a la probabilidad de que en el ensayo el resultado sea un "éxito".[62]

B.1.2. Función generadora

B.1.2.1. De momentos factorial

Sea $z \in \mathbf{C}$ tal que $|z|<1$, entonces la función generadora de momentos factorial está dada por:

$$E\left[z^x\right] = zp + (1-p).$$[63]

B.1.2.2. Característica

Sea $\theta \in \mathbf{C}$, entonces la función característica está dada por:

$$M_x(\theta) = E\left[e^{\theta x}\right] = e^{\theta}p + (1-p).$$

Nótese que $E\left[x^k\right] = p$ para todo valor de $k \in \mathbf{N}$.[64]

B.1.3. Momentos

B.1.3.1. Esperanza

Haciendo uso de la función característica se obtiene que la esperanza está dada por $E[x] = p$.

B.1.3.2. Varianza

Haciendo uso de la función característica se obtiene que la varianza está dada por $Var[x] = p(1-p)$.[65]

---

[62] **MOOD, Alexander M., Graybill Franklin A., Boes Duane C.** Introduction to the Theory of Statistics. **Tercera Edición. Editorial Mc. Graw Hill. Estados Unidos de América. 1974. pág. 87**
[63] **MOOD, Alexander M., Graybill Franklin A., Boes Duane C. op. cit. pág. 77**
[64] **MOOD, Alexander M., Graybill Franklin A., Boes Duane C. op. cit. pág. 78**
[65] **MOOD, Alexander M., Graybill Franklin A., Boes Duane C. op. cit. pág. 88**

GLOSARIO

Calificación de Apgar: Es un sistema de evaluación sencillo, de aplicación rápida y suficiente veracidad, que permite establecer el estado general del neonato e inferir de la calificación las medidas aconsejables; toma en cuenta cinco parámetros: frecuencia cardiaca, esfuerzo respiratorio, coloración de tegumentos, tono muscular y respuesta refleja. Cada parámetro recibe una calificación de 0, cuando está ausente, de 1, si existe pero deficiente, y de 2, si se expresa en su forma normal, de manera que la suma de las cinco calificaciones da un global entre 0 (niño muerto) a 10 (en óptima condición).

Así pues la clasificación se da cómo:

a. Calificación de 8 a 10: el neonato en mejores condiciones está vigoroso, rosado y llora.
b. Calificación de 5 a 7: los bebés moderadamente deprimidos aparecen cianótico, con respiraciones lentas e irregulares, pero poseen un buen tono muscular y reflejos.
c. Calificación de 4 ó menos: el niño intensamente deprimido está pálido, o azul, apneico, lacio, con una frecuencia cardiaca lenta.

Este método debe aplicarse también a los cinco minutos de vida, pues se ha comprobado que la calificación tiene buena relación con el riesgo de muerte inmediata así como de secuelas neurológicas tardías, y esto es porque cuanto más haya durado la asfixia, tanto más probable será que ocurra la muerte o en su defecto las consecuencias en las neuronas, por lo que la prueba de la calificación de Apgar a los cinco minutos constituye un indicador más útil de pronóstico neonatal y a largo plazo.

Contaminante: Son los gases o las partículas suspendidas en la atmósfera diferentes a la composición normal del aire.

Crecimiento fetal retardado: Se considera que un feto presenta un retardo en el crecimiento intrauterino cuando su peso es inferior al que le correspondería tener para su edad gestacional, es decir, son aquellos que están por debajo de la curva del percentil 10 sin importar su edad fetal, ver la gráfica 1 del apartado A.4.2.; también es conocido como el síndrome de Clifford.

Embarazo: Es el periodo de tiempo comprendido desde la fecundidad del óvulo hasta el parto, su duración aproximada es de 280 días (36 a 42 semanas de gestación) 10 meses lunares o casi 9 meses del calendario solar.

Edad gestacional: Es el tiempo que transcurre desde el momento de la concepción hasta que el niño nace y su determinación es fundamental para realizar un diagnóstico correcto de madurez e instaurar eventualmente un tratamiento oportuno y adecuado.

Estimación de la edad embrionaria: Entre los métodos más comunes está la edad menstrual, medida desde el primer día del último período menstrual.

Fecha de parto: La fecha de parto esperada es 266 días o 38 semanas después de la fecundación, también 280 días o 40 semanas después de la fecha de última regla, cabe mencionar que alrededor del 12% de los bebés nacen una o dos semanas después de la fecha probable de parto.

Fecha de la última regla: Se obtiene calculando el tiempo transcurrido desde el primer día del último ciclo menstrual hasta el nacimiento; suele ser el dato más seguro cuando la información es precisa, sin embargo no siempre esto ocurre porque la madre confunde las fechas o el embarazo se inicia durante una amenorrea de lactancia.

Intervalo entre embarazos: es el lapso transcurrido entre el fin de una gestación o aborto y el inicio de un nuevo embarazo.

Multípara: Mujer que ha tenido más de un parto.

Nacimiento: Fue definido por la Organización Mundial de la Salud como la "expulsión o extracción completa del cuerpo de la madre

independientemente de la duración del embarazo de un producto de la concepción que después de tal separación respira o manifiesta cualquier otro signo de vida, latido del corazón, pulsación del cordón umbilical o contracción efectiva de algún músculo sometido a la acción de la voluntad, haya o no haya sido cortado el cordón umbilical y esté o no adherido a la placenta."

Postérmino: Es el recién nacido cuya gestación duró más de 42 semanas (294 días), a partir del primer día de la última menstruación de acuerdo con la Organización Mundial de la Salud y la Federación Internacional de Obstetricia y Ginecología.

Posmaduro: Es el recién nacido de postérmino con retardo del crecimiento intrauterino.

Prematuridad: Es aquel recién nacido que nace antes de las 36 semanas de gestación, calculadas desde el primer día de la última menstruación, también son conocidos como neonatos de pretérmino.

Primípara: Mujer que por primera vez concibe a un hijo.

Talla del neonato: Es la longitud del recién nacido, es otro indicador de la salud fetal, añade información sobre las condiciones intrauterinas y también influye en el crecimiento subsiguiente.

BIBLIOGRAFIA

AGRESTI, Alan. Categorical data analysis. Primera Edición. Editorial New York John Wiley & Sons. Estados Unidos de América. 1990. 558 páginas.

GUAJARATI, Damonar N. Econometría. Traducción de: Mayorga Torrado Víctor Manuel. Segunda Edición. Editorial Mc. Graw Hill. México. 1992. 597 páginas.

HAMILTON, William James, Mossman H. W., Embriología humana: Desarrollo prenatal de la forma y la función. Cuarta Edición. Editorial Intermédica. Argentina. 1973. 667 páginas.

HOSMER, David W. Jr., Lemeshow Stanley. Applied logistic regression. Primera Edición. Editorial John Wiley & Sons. Estados Unidos de América. 1989. 308 páginas.

JUDGE, George G. R., Hill Carterl, Griffiths William E., Lütkepohl Helmut, Lee Tsoung-Chao. Introduction to the theory and practice of econometrics. Segunda Edición. Editorial John Wiley & Sons. Estados Unidos de América. 1988. 1024 páginas.

NOVALES, Cinca Alfonso. Econometría. Primera Edición. Editorial Mc. Graw Hill. México. 1988. 486 páginas.

MOOD, Alexander M., Graybill Franklin A., Boes Duane C. Introduction to the Theory of Statistics. Tercera Edición. Editorial Mc. Graw Hill. Estados Unidos de América. 1974. 564 páginas.

MOORE, Keith L., Persaud T.V.N., Shiota Kohei. Atlas de embriología clínica. Traducción de Color atlas of clinical embriology. Traducido por Pérez de

Miguelsanz J. Primera Edición. Editorial Médica Panamericana. España. 1996. 243 páginas.

HEMEROGRAFIA

BOBAK, Martin. Children's health outdoor air pollution, low birth weight and prematurity. Environmental Health Perspectives. Volumen 108. Número 2. Estados Unidos de América. Febrero 2000. Páginas 173-176.

INSTITUTO NACIONAL DE ECOLOGÍA. Centro Nacional de Investigación y Capacitación Ambiental. SEMARNAP. Primer informe sobre la calidad del aire en ciudades mexicanas 1996. Primera Edición. Editorial Dirección General de Gestión e Información Ambiental del Instituto Nacional de Ecología de la SEMARNAP. México. 1997. 96 páginas.

MAISONET, Mildred, Bush Timothy J., Correa Adolfo, Jaakkola Jouni J. K. Relation between ambient air pollution and low birth weight in the Northeastern United States. Environmental Health Perspectives. Volumen 109. Suplemento 3. Estados Unidos de América. Junio 2001. Páginas 351-356.

ŠRÁM, Radim J., Beneš Ivan, Binková Blanka, Dejmek Jan, Horstman Donald, Kotěšovec František, Otto David, Perreault Sally D., Rubeš Jiří, Selevan Sherry G., Skalík Ivan, Stevens Robert K., Lewtas Joellen. Teplice program – the impact of air pollution on human health. Environmental Health Perspectives. Volumen 104. Suplemento 4. Estados Unidos de América. Agosto 1996. Páginas 699-714.

WANG, Xiaobin, Ding Hui, Ryan Louise, Xu Xiping. Association between air pollution and low birth weight: a community- based study. Environmental Health Perspectives. Volumen 105. Número 5. Estados Unidos de América. Mayo 1997. Páginas 514-520.

RITZ, Beate, Yu Fei. The effect of ambient carbon monoxide on low birth weight among children born in Southern California between 1989 and 1993. Environmental Health Perspectives. Volumen 107, Número 1. Estados Unidos de América. Enero 1999. Páginas 17-25.

PAGINA DE INTERNET

http://www.lamolina.edu.pe/calidad_ambiental/monitoreoatm.html
Martínez, Ana Patricia. Romieu Isabelle. Centro Panamericano de Ecología Humana y Salud (ECO/OPS). Agencia de Cooperación Técnica de Alemania (GTZ). Departamento del Distrito Federal Introducción al monitoreo atmosférico. Capítulo 1: Introducción. Capítulo 7: Efectos de la contaminación del aire en la salud. Actualizada el 03 de febrero de 1999.