UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

UNIDAD ACADEMICA DE LOS CICLOS PROFESIONAL
Y DE POSGRADO

MODELOS ADITIVOS GENERALIZADOS APLICADOS A PROBLEMAS DE CONTAMINACION Y SALUD

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRA EN ESTADISTICA E INVESTIGACION DE OPERACIONES
P R E S E N T A
ADRIANA LOPEZ GARCIA

MEXICO, D. F. 1999

Agradecimientos.

A **Dios** por todo lo que tengo.

A mis padres:

Pilar y Virgilio quienes siempre han mostrado ser los seres más maravillosos de la vida. Por su amor y apoyo incondicional que siempre me han dado.

A mis hermanos:

Ale, Rocío, Lety y Luis por su entusiasmo y esperanza, juntos han significado ser la mano que se extiende en el momento necesario para dar consuelo y amor.

Al compañero de mi vida:

Carlos, por su amor, comprensión y por todo lo que ha llegado a significar en ella.

A mis amigos de siempre:

Gaby, Adriana, Cris, Norma, Salvador, Gabriel, Horte, Yola, Catyn.

Quisiera externar mis agradecimientos:

A la UNAM, en particular al Departamento de Probabilidad y Estadística del IIMAS y a la UACP y P, por el apoyo recibido durante el tiempo en que realice mis estudios.

A mi asesora de tesis la Dra. Silvia Ruíz y a los profesores de la maestría por su enseñanza.

A Juanita, Elida y a todo el cuerpo Administrativo del IIMAS que nos han apoyado siempre a los alumnos de la maestría de una forma desinteresada.

A mis sinodales, por sus valiosos comentarios al trabajo realizado:

Dr. Ignacio Méndez Ramírez
Dra. Silvia Ruíz-Velasco Acosta
M. en C. Leticia Gracia-Medrano
M. en C. Salvador Zamora Muñoz
M. en C. Gabriel Nuñez Antonio

De manera especial quiero agradecer a Salvador y Gabriel por su invaluable e incansable ayuda en la revisión de este trabajo.

Índice General

Introducción

La contaminación del aire es uno de los problemas ambientales más importantes en nuestros días, y es el resultado de las actividades del hombre. Las causas que originan esta contaminación son diversas, pero el mayor índice es provocado por las actividades industriales, vehiculares, comerciales, domésticas y agropecuarias.

Anteriormente, la Ciudad de México enfrentaba de manera primordial el problema del ozono, que dentro de la familia de contaminantes atmosféricos era el que predominaba en esta cuenca, en nuestros días, la contaminación causada por partículas suspendidas se ha unido a la anterior, debido a su preocupante incremento, y por ser potencialmente más dañina a la salud.

En este trabajo se presenta un análisis estadístico en el que se relaciona la presencia de partículas de contaminación en la atmósfera con la inasistencia de los niños a clases por enfermedades respiratorias en un centro preescolar del suroeste de la Ciudad de México.

El análisis se lleva acabo, en primer instancia, a través del ajuste de un modelo aditivo generalizado, con el que se identifica el tipo de relación existente entre inasistencia y contaminación y posteriormente se ajusta un modelo lineal generalizado con la reparametrización sugerida por el modelo aditivo generalizado, derivando así más conclusiones sobre la relación de partículas suspendidas y la inasistencia.

En el capítulo cero, Contaminación y Estadística, se describe brevemente el concepto de contaminación del aire, enfatizando en sus componentes de partículas suspendidas así como en los efectos que tiene sobre la salud. Se introducen además algunos conceptos de Epidemiología y de la relación existente entre esta área y la Estadística.

En el primer capítulo se presenta el modelo lineal generalizado que es una generalización inmediata del modelo de regresión múltiple, así como, sus elementos principales, inferencia y diagnóstico. En este capítulo se da una breve introducción de este modelo y de su utilidad en la modelación estadística.

En el segundo capítulo se hace una revisión del modelo aditivo generalizado, que es una generalización del modelo lineal generalizado y se explica por que, en nuestros días, representa una opción como herramienta para decifrar el tipo de relación que existe entre las variables explicativas y la variable respuesta, con el fin de tener un mejor ajuste, y por tanto, una mejor aproximación a nuestra realidad.

Para ilustrar todos los conceptos revisados en los capítulos anteriores, en el tercer capítulo se presenta una aplicación con datos reales de contaminación del aire, aquí se muestra el uso de los modelos lineales generalizados y del modelo aditivo generalizado conjuntamente.

Finalmente en el cuarto capítulo se dan las conclusiones obtenidas del presente trabajo.

Capítulo 0

Contaminación y Estadística

0.1 Introducción

En este capítulo se da una breve introducción al concepto de contaminación del aire. Haciendo énfasis en uno de sus componentes principales: las partículas suspendidas que han sido estudiadas en los últimos años y han resultado ser un factor influyente en la salud humana. Dado este impacto en la salud, las partículas suspendidas constituyen una de las preocupaciones actuales en las grandes urbes de nuestro planeta y por ende en nuestro país.

Asimismo, se presenta el concepto de epidemiología y su relación con la estadística y la contaminación.

0.2 Contaminación.

La ciudad de México debido a su situación geográfica y a la explosión demográfica que ha venido enfrentando desde hace varios años, ha sido un foco de interés para las asociaciones internacionales del ambiente y, por supuesto, punto de atención para los

epidemiólogos y personas interesadas en conocer las consecuencias que la contaminación provoca en la salud de sus habitantes. Sin embargo, es importante señalar que en nuestro país no sólo la ciudad de México es un punto de preocupación sino toda ciudad densamente poblada y que represente un estado de industrialización elevado. También es interesante mencionar que los efectos de la contaminación en la salud humana no son descubrimientos recientes, al contrario, recordemos el desastre por contaminación en Londres en 1952 (Her Majesty's Public Health Service, 1954) en donde por causa de la contaminación fallecieron cientos de personas.

A nivel nacional, la contaminación del aire se limita a zonas de alta densidad demográfica o industrial. Las emisiones anuales de contaminantes en el país son superiores a 16 millones de toneladas, el 65% es de origen vehicular. En la ciudad de México se genera el 23.5 % de dichas emisiones, en Guadalajara el 3.5% y en Monterrey el 3%. Los otros centros industriales del país generan el 70% restante (página de web www.monografías.com, el aire).

Los principales contaminantes del aire se clasifican en primarios y secundarios.

Los contaminantes primarios: son los que permanecen en la atmósfera tal y como fueron emitidos por la fuente. Para fines de evaluación de la calidad del aire se consideran como contaminantes primarios: el óxido de azufre, el monóxido de carbono, el óxido de nitrógeno, los hidrocarburos y las *partículas.*

La contaminación del aire por partículas consiste de partículas sólidas y líquidas muy pequeñas flotando en el aire como polvo cenizas ollín, partículas metálicas, cemento o polen. La preocupación más grande de las instituciones de salud pública es que las partículas son lo suficientemente pequeñas como para ser inhaladas hasta llegar a las partes más profundas de los pulmones. Estas partículas son menores a 10 micrones (um) en diámetro, lo que corresponde a sólo 1/7 del grosor de un cabello humano y son conocidas como PM10 y estas a su vez incluyen partículas más finas conocidas como PM2.5 consideradas por los científicos de la salud más peligrosas que las partículas PM10, debido a que son lo suficientemente pequeñas como para evitar que el mecanismo de defensa respiratoria del cuerpo las detecte y de esta forma se instalan, sin obstáculo alguno, en las partes más profundas de los pulmones. PM10 y PM2.5 son de los componentes de la contaminación del aire que amenaza tanto a

nuestra salud como a nuestro medio ambiente.

De acuerdo a un análisis realizado por el NRDC (Consejo para la Defensa de los Recursos Naturales de E.E.U.U.), cada año más de 64, 000 personas mueren prematuramente por causas cardiopulmonares relacionadas a la contaminación del aire por partículas.

Para entender lo complejo que es el problema de la contaminación el NRDC en 1995 empleó los resultados de un estudio de 1995 realizado por la Sociedad Americana de Cáncer (SAC) y la escuela de Medicina de Harvard, realizado con datos provenientes de 239 ciudades de E.E.U.U. Este estudio es el más grande y el más amplio que ha mostrado los efectos adversos en la salud causados por la contaminación del aire por partículas. Este estudio utiliza técnicas estadísticas para señalar que factores tales como el hábito de fumar, peso y riesgo de trabajo tienen un efecto nulo sobre los efectos adversos en la salud.

Contaminantes secundarios:

Los contaminantes secundarios son los que han estado sujetos a cambios químicos, o bien, son el producto de la reacción de dos o más contaminantes primarios en la atmósfera. Entre ellos destacan oxidantes fotoquímicos y algunos radicales de corta existencia como el ozono(O_3).

Características del contaminante PM10.

PM10 es la fracción respirable de partículas suspendidas totales (PST) que está constituída por aquellas partículas de diámetro inferior a 10 micras que tienen la particularidad de penetrar en el aparato respiratorio hasta los alvéolos pulmonares.

PM10 es una mezcla de materiales que incluyen humo, ollín, sal, ácidos y metales. También proviene de las reacciones químicas que sufren en la atmósfera los gases emitidos por los vehículos de motor e industriales.

PM10 se encuentra entre los contaminantes del aire más dañinos. Cuando se inhalan, estas partículas invaden las defensas naturales del sistema respiratorio y se alojan en los pulmones.

Los problemas de la salud empiezan cuando el cuerpo reacciona en contra de estas

partículas extrañas. PM10 causa irritación en las vías respiratorias. Su acumulación en los pulmones origina enfermedades como silicosis (enfermedad que produce una fibrosis pulmonar muy grave) y la asbestosis (las partículas se depositan en los bronquiolos y allí producen una reacción fibrosa que se extiende por el tejido intersticial hasta impedir el funcionamiento normal del pulmón). Puede incrementar el número y la severidad de ataques de asma; agrava las enfermedades cardiovasculares, causa o agrava la bronquitis y otras enfermedades pulmonares y reduce la capacidad del cuerpo para atacar a las infecciones.

A pesar de que las partículas pueden causar problemas a toda la población, existen algunos sectores de la misma que son especialmente vulnerables a los efectos de PM10. Esta "población vulnerable" se encuentra conformada por niños, personas de la tercera edad, personas que padecen enfermedades cardiacas, enfermedades pulmonares, atletas y aquéllos que sufren de asma o bronquitis, en los cuales se esperan diversos grados de tolerancia de origen genético-estructural, lo que produce variablidad en las respuestas del organismo a la exposición.

Fuentes principales de producción de las partículas. Las fuentes principales que producen partículas suspendidas son: combustión industrial y doméstica del carbón, combustóleo y diesel; procesos industriales; incendios; erosión eólica y erupciones volcánicas; entre otras.

La American Lung Association recomienda un límite máximo permisible para la presencia de partículas en la atmósfera PM2.5 (partículas de 2.5 um) que corresponde a 18 microgramos por metro cúbico (18 μ/m^3) medido en 24 hrs y 12 μ/m^3 promedio en un año entero.

En la Ciudad de México la red automática de monitoreo ambiental mide la concentración de PST, partículas menores a diez micras (PM10) y metales pesados. Funciona realizando muestreos durante un periodo de 24 hrs cada seis días, excepto en invierno, tiempo en el que se incrementa la frecuencia de muestreo a una vez cada tres días. Actualmente 5 de las estaciones miden la fracción respirable e identifican y cuantifican los metales presentes en PST.

0.3 Epidemiología

La epidemiología es el estudio de patrones en la ocurrencia de enfermedades y los factores que influyen en tales patrones. En la epidemiología un objetivo de estudio es investigar la causa o etiología de un enfermedad o condición fisiológica. Como se sabe, muchas enfermedades no son causadas únicamente por un agente particular sino por una combinación de diferentes circunstancias llamados factores de exposición o factores de riesgo. Estos factores de exposición pueden ser características físicas del individuo tales como la edad y la masa del cuerpo; variables fisiológicas tales como el nivel de colesterol en suero, medidas de función respiratoria, hábitos personales, tales como historia de fumar y dieta; factores socioeconómicos, ambientales etc. Un estudio etiológico trata de encontrar la asociación entre una o varios factores de exposición y un estado de enfermedad particular.

La epidemiología realmente consiste de dos campos, uno que se relaciona con estudios experimentales y otro con estudios observacionales. La confusión entre los dos campos, impide en ocasiones nuestro entendimiento de datos observacionales. Esto es particularmente cierto para epidemiología ambiental, donde las hipótesis a ser examinadas son, con frecuencia, débilmente especificadas. La epidemiología del medio ambiente tiene ciertas diferencias importantes que hacen a la intuición adelantarse en comparación a otros campos de la epidemiología, y sugiere la necesidad del uso de herramientas estadísticas más completas en las investigaciones.

Una gran cantidad de estudios en la epidemiología ambiental han sido enfocados a la contaminación del aire o a la toxicidad del plomo, en ambos casos, las posibles consecuencias son, por ejemplo, enfermedades respiratorias.

Una asociación entre salud humana y la contaminación del aire se ha propuesto desde hace más de 50 años. En muchos paises se han desarrollado programas de salud nacionales y con ello han dado nuevas oportunidades a los investigadores para explorar las relaciones contaminación-enfermedad y contaminación-muerte.

Estudios epidemiológicos.

Los estudios de epidemiología que han sido desarrollados hasta nuestros días, se pueden agrupar en dos grandes clasificaciones según el diseño de estudio.

1) Estudios de exposición aguda, que son comúnmente estudios de series de tiempo y usan cambios cortos de la contaminación del aire sobre el tiempo 1-5 días como la fuente de variabilidad de exposición y

2) Estudios de exposición crónica, que son principalmente de validación cruzada en el diseño y usa un periodo grande de observación (regularmente 1 año o más).

En numerosos estudios de series de tiempo se han observado asociaciones entre las partículas del aire contaminado y diversas enfermedades incluyendo mortalidad, hospitalización por enfermedades respiratorias y del corazón, agravación del asma, incidencia y duración de síntomas respiratorios, funciones pulmonares y actividad física restringida.

La técnica más usada en la modelación de datos es regresión multiple en ocasiones con respuesta categórica. Hay al menos dos razones primarias para que estos modelos sean tan usados en el análisis de datos epidemiológicos sobre los efectos en la salud de la contaminación del aire.

1) Permite la estimación de la asociación de partículas-salud mientras se controla al menos algún otro factor de riesgo.

2) Puede sumar un mayor rigor al análisis al proveer una forma para hacer pruebas de hipótesis y hacer inferencias estadísticas más formales.

0.3.1 Problemática de los estudios más frecuentes en epidemiología.

La dosis-respuesta y el análisis de tendencia en epidemiología se hacen, generalmente, en forma muy simple y frecuentemente de manera ingenua. A lo más, algunos autores llevan acabo pruebas de tendencia usando la prueba de Mantel o ajustando un modelo de regresión con un sólo término de exposición, cuya aproximación puede ser engañosa, por que en esencia se supone que la relación dosis-respuesta o curva de

tendencia sigue un modelo específico (usualmente logístico).

Otros autores dividen el rango de estudio en categorías y observan la tendencia dentro de cada categoría específica, a través de sus coeficientes o riesgos relativos. Un estudio de esta naturaleza puede ser adecuado si los números permiten el uso de categorías que reflejen biológicamente grupos de respuesta homogénea. Con frecuencia, sin embargo, las categorías se eligen en una forma mecánica a través de un algorítmo como, por ejemplo, el método del percentil. Las principales dificultades de los percentiles son más dramáticas cuando muchos sujetos están expuestos en un rango reducido o cuando los efectos de exposición son limitados a un extremo de la escala de exposición, tal como niveles de nutrientes bajos o bien niveles altos de exposición. En tales situaciones individuos con un riesgo elevado de exposición estarán colocados entre miembros que tengan un riesgo menor según la categoría del percentil correspondiente, este riesgo puede algunas veces ser mitigado basándose en percentiles sobre la distribución, mas que la distribución de todos los sujetos, pero sería deseable evitarlo en su totalidad.

Una forma de observar la tendencia dentro de cada categoría es el siguiente: considere un análisis dosis-respuesta de datos categóricos ordinales. Donde se divide el rango observado de exposición x en K categorías, indexado por $k = 1, 2, ..., K$ con $K - 1$ límites internos $c_1, ..., c_{K-1}$. Así, dentro de cada categoría se ajusta una línea completamente horizontal que represente la curva de "dosis-respuesta" para relacionar la exposición a la respuesta dentro de cada categoría. Por ejemplo, en regresión logística categórica simultáneamente se ajustan K categorías-específicas para el modelo logit de riesgo R :

$$\log it(R|x \text{ en la categoría } k) = \alpha_k^* \qquad k = 1, ..., K, \tag{0.1}$$

el cual dice que el rango de exposición de la x no tiene ninguna clase de efecto dentro de las categorías, sin importar qué tan grande sea el efecto entre las categorías.

Para ilustrar lo anterior, suponga que x es la ingestión diaria de ácido ascórbico, R es el riesgo de morir, y los límites para x son 20, 50 y 100mg. por día, incluyendo el límite inferior en el intervalo, por ejemplo [0,20). Usando la relación ec.(0.1) se dice

que en el modelo categórico no hay diferencia en el riesgo que se encuentra entre 0 y 20 mg por día, sin embargo, existe un cambio notable en el riesgo cuando se pasa de 20 a 21 mg por día, debido a que la ordenada al origen es diferente para cada intervalo del rango del ácido ascórbico. Esto es biológicamente absurdo dado que 0 mg por día representa un estado relativamente fatal de deficiencia, 20 mg por día no lo representa y la diferencia entre 20 y 21 mg por día es biológicamente trivial. A pesar de que un modelo categórico proporciona los estimadores del riesgo promedio dentro de las categorías, es posible cuestionarse el por qué da un valor tan disparado del riesgo promedio como, por ejemplo, el que se encuentra entre 0 y 20 mg por día de ácido ascórbico. Aún más, en modelos nolineales, el estimador del riesgo promedio que resulta de una regresión con indicadores de categorías puede producir una impresión sesgada de la exposición específica en la curva de dosis-respuesta, una respuesta a este problema es la forma errónea en la que se está procediendo a reparametrizar ya que no se esta tomando en cuenta el verdadero comportamiento de los datos para sugerir los puntos de corte, un ejemplo de esto es el uso de cuartiles para categorizar una variable contínua. Los modelos que se han mencionado, por ejemplo, el modelo logístico es parte de una gama de modelos llamados modelos lineales generalizados los cuales explicaremos con mayor detalle en el siguiente capítulo.

0.3.2 Alternativas estadísticas a la problemática de modelación en epidemiología

Muchos autores han recomendado regresión no-paramétrica como un medio para evitar el problema de la categorización en su totalidad. La regresión no-paramétrica se recomienda usar en aquellos estudios especiales donde no se puede decir con seguridad nada sobre la forma de la tendencia o de la relación enfermedad-exposición (dosis-respuesta). La aplicación de la regresión no-paramétrica había sido obstaculizada por la falta de paquetes computacionales disponibles, sin embargo, este obstáculo está gradualmente desapareciendo por los avances computacionales de hoy en día. Otro problema que presenta la regresión no-paramétrica es que los límites computacionales, debido al número máximo de covariables o individuos se inclina ha ser mucho menor

en comparación con la regresión convencional.

En concreto, el análisis epidemiológico de dosis-respuesta y la tendencia, así como los métodos para control de factores de confusión continuos podrían ser expandidos más alla de un simple análisis categórico y una aproximación lineal (un sólo coeficiente) a estudios que incluyan curvas flexibles que hagan uso de información dentro de las categorías. Tal expansión puede ser acompañada con una pequeña generalización via polinomios fraccionales y regresión con splines. Estos métodos pueden ser especialmente valiosos cuando se anticipe una relación nolineal, como es el caso de estudios de salud, donde los efectos del alcohol, nutrientes y otros factores de estilo de vida se sabe que tienen relaciones no lineales con la respuesta.

La cultura epidemiológica es más observacional que experimental. El mal entendimiento entre observacional y experimental se hace mayor al ajustar factores de riesgo continuos, respuestas multifactoriales y respuestas con una alta variación no explicada por las variables explicativas disponibles. El análisis de estos datos con frecuencia se ve como una prueba de hipótesis en la cual el control estadístico remplaza a la aleatorización; es decir, tales estudios con frecuencia prueban formas restringidas de la hipótesis que está siendo investigada, por ejemplo, la hipótesis de una relación lineal, cuando en realidad no existe una justificación empírica o teórica para sustentar que si una relación existe, entonces deba ser lineal. En este tipo de estudios se sugieren alternativas más flexibles para explorar la asociación, tales como: *suavizamiento no paramétrico y en particular modelos aditivos generalizados*, los cuales representan una opción para atacar a dichos problemas.

En estudios de epidemiología del medio ambiente, se han enfocado en la contaminación del aire o en la toxicidad del plomo. Estos estudios dependen de un gran número de factores de riesgo entre los que se encuentran, factores del medio ambiente que en general no se espera que se encuentren entre los más importantes. La exposición a los contaminantes del medio ambiente en general no son dicotómicos sino continuos. Los otros factores de riesgo también son usualmente medidas continuas.

En epidemiología ambiental, recientemente ha habido un interés creciente sobre estudios que analizan series de tiempo de eventos. Estos eventos pueden ser mortalidad, admisiones a un hospital o síntomas respiratorios. Probablemente dichos eventos

no se encontrarán confundidos con factores de riesgo personales tales como el hábito de fumar, presión arterial, y factores socioeconómicos, ya que estos factores no varían día a día como ocurre con la contaminación del aire. Sin embargo, dichos factores probablemente serán confusores potenciales en estudios comparativos de poblaciones de diferente nivel de contaminación del aire. El interés en estas respuestas ha permitido el uso más general de técnicas de regresión Poisson. Los modelos logísticos han sido usados en epidemiología del medio ambiente. En todos estos modelos es claro que existe una dependecia de la respuesta sobre el clima o la estación, sin embargo, la forma funcional de esta dependencia no es clara. En tales casos se recomienda el uso de técnicas no paramétricas se recomiendan.

Capítulo 1

Modelos Lineales Generalizados

1.1 Introducción

En este capítulo se presenta una breve introducción a los modelos lineales generalizados. Se describen sus elementos, la estimación de sus parámetros y el diagnóstico del modelo.

1.2 Antecedentes

En una gran cantidad de estudios es de interés tratar de modelar los valores esperados de una variable aleatoria (variable respuesta), a través de una relación en la que se puedan incluir aquellas características (variables explicativas) que hacen que este comportamiento se observe. El principal objetivo de este tipo de análisis estadístico es investigar la relación entre la variable respuesta Y y la variable explicativa X. Para investigar esta relación es conveniente construir un modelo pensando en que éste sea capaz de describir tal relación. Por varias décadas el modelo que siempre ha sido propuesto, en el caso particular de datos contínuos, es el modelo lineal de regresión, el cual está representado de la siguiente manera:

$$Y = X\underline{\beta} + \underline{e} \tag{1.1}$$

donde e_i, elementos de $\underline{e}$, son independientes e identicamente distribuídos N(0,σ^2), y por lo tanto$Y = (y_1, ..., y_n)^t$, tiene una distribución normal con

$$\begin{aligned} E[Y] &= X\underline{\beta} \qquad \text{y} \\ Cov[Y] &= \sigma^2 I \end{aligned}$$

$X_{nxp} = (\underline{x}_1, ..., \underline{x}_n)^t$ con $\underline{x}_i^t = (x_{i1}, ..., x_{ip})$, $\underline{\beta}_{px1}$ es el vector de parámetros de regresión y finalmente $\underline{e}_{nx1}$ es un vector aleatorio cuyos elementos son independientes e idénticamente distribuidos, con distribución Normal de media 0 y varianza σ^2.

Una revisión del desarrollo que ha tenido la modelación estadística nos permitirá conocer cómo los modelos lineales se han venido extendiendo más y más a través del paso del tiempo. Una breve descripción de esta historia se enuncia en (Lyndsey,1997).

En primera instancia se tienen los modelos de regresión lineal múltiple, Legendre, Gauss, a principios del siglo XIX (Stigler, 1981,1986). Posteriormente se extienden la idea al conocido análisis de varianza (ANOVA) diseño de experimentos, con la misma distribución normal y con la misma relación entre la media de la variable respuesta y las variables explicativas (Fisher: 1920 - 1935).

El uso de una función de verosimilitud fue el siguiente paso que permitió realizar un estudio general sobre inferencia a partir de cualquier modelo estadístico (Fisher, 1922). Por ejemplo los ensayos de dilución fueron una de las inmediatas consecuencias teniendo como base la distribución Binomial, y la relación entre la media de la variable respuesta y las variables explicativas fué $\log(\mu/1-\mu)$ (Fisher, 1922).

Un estudio profundo de la familia exponencial hizo posible que dentro de ella se agruparan diferentes distribuciones que por su uso en la modelación eran de interés.(Fisher, 1934).

El uso de distribuciones diferentes a la distribución normal era cada vez una necesidad mayor, debido a que el tipo de datos que se tenían ya no sólo eran continuos sino categóricos, de aquí que la relación entre la media de la variable respuesta y las variables explicativas ya no era directa sino era necesario el uso de transformaciones de la media como, por ejemplo, el análisis conocido como probit donde se tiene una relación de la forma $\pi_x = \Phi(\alpha + \beta x)$. con π_x la proporción de supervivencia y Φ es la función de distribución acumulada de la Norma, (Bliss, 1935). Apartir de esto,

el interés fue creciendo y junto con ello aparecieron más y más nuevas formas de relaciones entre la media y las variables explicativas, dependiendo del tipo de datos que se tenía interés por modelar.

Nelder y Wedderburn en 1972 dieron el paso decisivo en el que unificaron la teoría sobre ciertos modelos estadísticos, en particular, para los modelos de regresión, publicando su artículo; Modelos Lineales Generalizados. Ellos mostraron lo siguiente:

- Cuáles de los modelos de regresión más comunes de estadística clásica eran miembros de una misma familia y podían ser tratados de la misma manera.

- Que el estimador máximo verosímil para todos estos modelos se podía obtener usando el mismo algoritmo: mínimos cuadrados ponderados iterados.

De esta manera el análisis probit y el modelo de regresión con error Normal asi como todos aquellos cuya distribución pertenecen a la familia exponencial se podían manejar en forma similar, es decir una sola teoría bastaba para todos ellos.

Más adelante, se demostró que todos los modelos enunciados tenían una distribución perteneciente a la familia de dispersión exponencial que es una generalización de la familia exponencial, con alguna transformación de la media (Jϕrgensen, 1987).

Por la importancia que tiene la familia exponencial en los modelos lineales generalizados, a continuación se presenta un breve resumen de esta familia.

1.2.1 Familia exponencial.

Suponga que se tiene un conjunto de n variables aleatorias independientes, v.a., Z_i $(i = 1, ..n)$ cuya función de probabilidad de Z_i se puede escribir de la siguiente manera

$$f(z_i;\xi_i) = r(z_i)s(\xi_i)\exp[t(z_i)u(\xi_i)] \tag{1.2}$$

si en la anterior ecuación $v(z_i) = \log(r(z_i))$ y $s(\xi_i) = \log(w(\xi_i))$, entonces

$$f(z_i;\xi_i) = \exp[t(z_i)u(\xi_i) + v(z_i) + w(\xi_i)] \tag{1.3}$$

donde ξ_i es el parámetro de localización. Ahora si se aplica la reparametrización $y = t(z)$ y $\theta = u(\xi)$, para obtener la forma canónica de la variable aleatoria, del parámetro y de la familia, entonces la familia exponencial tiene una función de distribución dada por

$$f(y_i;\theta_i) = \exp[y_i\theta_i - b(\theta_i) + c(y_i)] \tag{1.4}$$

donde $b(\theta_i)$ es la constante de normalización de la distribución. Dentro de esta familia se encuentran la distribución Poisson y la distribución Binomial, entre otras.

La familia exponencial se puede generalizar al incluir un parámetro de escala, ϕ, en la distribución, por lo que se observa de la siguiente manera:

$$f(y_i,\theta_i,\phi) = \exp[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)] \tag{1.5}$$

en donde θ_i es la forma canónica del parámetro de localización, $\Lambda(\mu_i) = \theta_i$. A esta nueva familia se le conoce con el nombre de familia de dispersión exponencial.

Es posible ver que dentro de esta familia se encuentra la distribución Normal y la distribución Gamma, entre otras. En caso de que $a_i(\phi) = 1$, es decir, ϕ conocido, se tiene a la familia exponencial de un parámetro cuya forma analítica está dada por la ecuación (1.4).

Para esta familia existe una relación importante entre la media y la varianza (Lyndsey, 1997), que se muestra a continuación y se usa la función de puntajes para llegar a ella.

Sea $L(\theta_i,\phi;y_i) = f(y_i;\theta_i,\phi)$ la verosimilitud para una observación y sea

$$U_i = \frac{\partial \log[L(\theta_i,\phi;y_i)]}{\partial\theta_i}$$

la conocida función de puntajes. De teoría inferencial se puede mostrar que si la distribución en cuestión cumple con ciertas condiciones de regularidad, entonces

$$E[U_i] = 0 \tag{1.6}$$

y

$$Var[U_i] = E[U_i^2] = E[-\frac{\partial U_i}{\partial \theta_i}]. \tag{1.7}$$

Como la familia de dispersión exponencial cumple las condiciones de regularidad la ecuación 1.6 y la ecuación 1.7 son ciertas.

Debido a que

$$\log[L(\theta_i, \phi; y_i)] = \frac{y_i\theta_i - b(\theta_i)}{a_{i(\phi)}} + c(y_i, \phi)$$

entonces

$$U_i = \frac{y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}}{a_{i(\phi)}}$$

usando la ecuación (1.6) se tiene

$$E[Y_i] = \frac{\partial b(\theta_i)}{\partial \theta_i} = \mu_i,$$

así

$$Var[U_i] = \frac{Var[Y_i]}{a_{i(\phi)}^2}$$

utilizando la ecuación (1.7) y el resultado anterior

$$Var[Y_i] = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} a_i(\phi).$$

sea $\tau^2 = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$ que se conoce con el nombre de función varianza, entonces

$$Var[Y_i] = \tau^2 a_i(\phi)$$

donde θ_i es el parámetro de interes, ϕ es el parámetro de dispersión que usualmente se considera como un parámetro de ruido, $a_i(\phi)$ es parte de la función de distribución.

ejemplo: Distribución Binomial.

$$\begin{aligned} f(y_i;\pi_i) &= \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}, \qquad y_i = 0,1,2,...,n_i \\ &= \exp\left\{y_i \log\left[\tfrac{\pi_i}{1-\pi_i}\right] + n_i\log(1-\pi_i) + \log\binom{n_i}{y_i}\right\}, \qquad y_i = 0,1,2,...,n_i \end{aligned} \tag{1.8}$$

en este caso $\theta_i = \log\left[\frac{\pi_i}{1-\pi_i}\right]$, $b(\theta_i) = -n_i\log(1-\pi_i)$, $a_i(\phi) = 1$ y $c(y_i,\phi) = \log\binom{n_i}{y_i}$.

Cuando $\pi_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ entonces $b(\theta_i) = -n_i\log(1+\exp\theta_i)$, $a_i(\phi)$ y $c(y_i)$ permanecen iguales siendo en este caso la función varianza

$$\frac{\partial^2 b(\theta_i)}{\partial\theta_i^2} = n_i\left(\frac{\partial}{\partial\theta_i}\left(\frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right)\right) = n_i\frac{\exp(\theta_i)}{(1+\exp(\theta_i))^2}$$

En la siguiente tabla No.1.1 se presentan las distribuciones más importantes de la familia exponencial junto con su respectiva función varianza.

Distribución	Función varianza
Poisson	$\mu = e^{\theta}$
Binomial	$n\pi(1-\pi) = ne^{\theta}/(1+e^{\theta})^2$
Normal	1
Gamma	$\mu^2 = (-1/\theta)^2$
Gaussiana inversa	$\mu^3 = (-2/\theta)^{\frac{3}{2}}$

Tabla No. 1.1

1.3 Modelos lineales generalizados.

Los modelos lineales generalizados (MLG) surgen de forma natural como una extensión de los modelos lineales. Debido a que se mantiene el supuesto de una relación lineal en las variables explicativas, la diferencia con respecto al modelo de regresión multivariado es la relación con respecto a la media de la variable respuesta en la que ya no es tan directa sino que es una función de ésta y a la vez esta función depende de la distribución que corresponda al comportamiento de la variable respuesta, por ejemplo, Binomial, Poisson, entre otras.

Este modelo está especificado por tres componentes:

1) *Componente aleatorio*; identifica la distribución de probabilidad de la variable respuesta (Y).

2) *Componente sistemático*, el cual especifica una función lineal de variables explicativas (X) que se usa como predictor lineal, y

3) *Función liga*, que describe la relación funcional entre el componente sistemático y el valor esperado del componente aleatorio.

Componente aleatorio. El componente aleatorio de los MLG consiste en observaciones independientes $Y = (Y_1, ..., Y_n)^t$ donde la distribución de Y_i es un miembro de la familia de dispersión exponencial (ecuación 1.5).

Componente sistemático. Si se tiene un vector de p parámetros desconocidos, $\underline{\beta}$, y un conjunto de variables explicativas $X_{nxp} = (\underline{x}_1, ..., \underline{x}_p)^t$ es frecuente suponer un modelo cuya respuesta media varía en una forma lineal. En el caso más simple, el parámetro de localización es una combinación lineal de las variables explicativas, es decir

$$E(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad (i = 1, ..., n).$$

Este modelo lineal para el caso que nos corresponde puede ser generalizado al permitir otras funciones suaves de la media, cuya representación está dada por una función $\eta(\cdot)$ que se le conoce con el nombre de predictor lineal

$$\eta(\cdot) = X\underline{\beta}.$$

Función liga. La relación entre la media de la i-ésima observación y su predictor lineal está dada por una función liga $g(\cdot)$ es decir

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij} \quad i = 1, ..., n \quad j = 1, ..., p. \tag{1.9}$$

esta función liga debe ser una función monótona y diferenciable. A $g(\cdot)$ se le conoce con el nombre de liga canónica cuando se cumple que $\theta = \eta$, donde θ es el parámetro canónico, es decir, la liga canónica es una función que transforma la media a un parámetro de localización canónico de la familia de dispersión exponencial. Con la

función liga canónica, todos los parámetros desconocidos de la estructura lineal tienen estadísticos suficientes si la distribución de respuesta es un miembro de la familia de dispersión exponencial y el parámetro de escala es conocido. Sin embargo la función liga es sólo un artefacto para simplificar los métodos numericos de estimación cuando un modelo involucra una parte lineal. Para modelos de regresión nolineal estrictamente esto pierde sentido (Lindsey, 1997).

Por ejemplo: Sea $x \sim Poisson(\lambda)$, $f(x) = e^{-\lambda x + \log(\lambda) + \log(\frac{1}{x!})} I_{(0,\infty)}(x)$, $a(\phi) = 1$, $b(\theta_i) = \log(\lambda)$, $c(x,\phi) = \log(1/x!)$ y $\theta_i = \lambda$ es el parámetro canónico, como $\mu = E(x) = \lambda$ entonces $\log(\cdot)$ es la función liga canónica por que $log(\mu) = log(\lambda)$.

A continuación se muestran algunas distribuciones junto con su respectiva función liga canónica.

Distribución	Función liga canónica
Poisson	$log(\mu)$
Binomial	$logit = log(\frac{\mu}{1-\mu})$
Normal	μ
Gamma	$\frac{1}{\mu}$
Gaussiana Inversa	$\frac{1}{\mu^2}$

Tabla No. 1.2

En el modelo lineal clásico, la media y el predictor lineal son idénticos, es decir la función liga es la identidad. Sin embargo, en general $g(\mu) = Q(\theta)$ donde Q es una función diferente a la identidad. Esto depende del tipo de datos que se estén estudiando. Por ejemplo, en el caso de una muestra de datos cuya distribución sea Binomial se tiene que $0 < \mu < 1$ y la función liga debe ser aquella que mapee el intervalo (0,1).

Dentro de la bibliografía para datos binomiales, se consideran tres principales funciones liga.

1. Logit $\eta = \log(\frac{\mu}{1-\mu})$
2. Probit $\eta = \Phi^{-1}(\mu)$ con $\Phi(\cdot)$ la función de distribución Normal acumulada.
3. Complementaria $\log(-\log)$ $\eta = \log\{-\log(1-\mu)\}$.

La familia potencia de ligas es importante, al menos para observaciones con media positiva (McCullagh,1989). Esta familia se puede especificar por

$$\eta = \begin{cases} \mu^{\lambda}; & \lambda \neq 0 \\ \log(\mu); & \lambda = 0 \end{cases} . \tag{1.10}$$

1.3.1 Estimación de los parámetros de los modelos lineales generalizados.

En inferencia estadística clásica, los métodos de estimación puntuales más usados son el de máxima verosimilitud y el de mínimos cuadrados. Aquí sólo se presenta la forma de estimar los parámetros por el método de máxima verosimilitud, debido a que es el algorítmo común para los modelos lineales generalizados.

Si se tienen n respuestas independientes de un modelo lineal generalizado, la función de log-verosimilitud, está dada por

$$l(\theta, \phi; y) = \frac{\sum_{i=1}^{n} y_i\theta_i - \sum_{i=1}^{n} b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i) \tag{1.11}$$

donde

$$\frac{\partial b(\theta)}{\partial \theta} = E(Y_i) = \mu_i \tag{1.12}$$

y

$$g(\mu_i) = \underline{x}_i'\underline{\beta} = \eta_i. \tag{1.13}$$

Debido a la suposición de que el error del MLG tiene una distribución que pertenece a la familia exponencial, es posible asegurar que existe el máximo de la función log-verosimilitud $l(\theta, \phi; y)$ y además que se puede encontrar al resolver la ecuación que resulta de derivar la log-verosimilitud con respecto al parámetro de interés e igualarla a cero, esto es debido a que los modelos de la familia exponencial poseen ciertas propiedades de convexidad que en muchos casos garantizan la existencia y unicidad de los estimadores de máxima verosimilitud (Wedderburn, 1976). Usando lo anterior y tomando en cuenta que el parámetro de interés es β_j procedamos a encontrar el

máximo de interés:

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) \qquad j = 1, ..., p \tag{1.14}$$

donde x_{ij} es el j-ésimo elemento de $\underline{x}_i^t$. En general, las ecuaciones $U_j = 0$ no son lineales y tienen que resolverse por métodos numéricos iterativos. Si se usa el método de Newton-Raphson entonces la m-ésima aproximación está dada por

$$\beta^{(m)} = \beta^{(m-1)} - [\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}]^{-1}_{\underline{\beta}=\beta^{(m-1)}} \mathrm{U}^{(m-1)} \tag{1.15}$$

donde

$$[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}]^{-1}_{\underline{\beta}=\beta^{(m-1)}} \tag{1.16}$$

es la matriz de segundas derivadas de l evaluada en $\underline{\beta}= \beta^{(m-1)}$ y $\mathrm{U}^{(m-1)}$ es el vector de la primera derivada $U_j = \frac{\partial l}{\partial \beta_j}$ evaluada en $\underline{\beta}= \beta^{(m-1)}$.

Un método alternativo que en algunas ocasiones es más eficiente que el método de Newton-Raphson es el método de puntajes, que consiste en remplazar la matriz de segundas derivadas en la ecuación (1.15) por la matriz de valores esperados

$$E[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}]. \tag{1.17}$$

Realizando cálculos algebraicos es posible mostrar que la expresión anterior se puede ver como el negativo de la matriz de varianza-covarianza de los U_j 's. Como se sabe a

$$\mathrm{L}_{jk} = E[U_j U_k] = -E[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}] \tag{1.18}$$

son elementos de la matriz de información de Fisher. Para los modelos lineales generalizados esta matriz de información se reduce a

$$\mathrm{L}_{jk} = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \tag{1.19}$$

de aquí que L se puede escribir como

$$\mathrm{L} = X^t W X$$

donde W es una matriz diagonal de nxn con elementos $w_{ij} = \frac{1}{var(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i})^2$.

De esta forma si en la ecuación (1.15) se sustituye a L por la matriz de segundas derivadas como lo indica el método de puntajes, se tiene

$$\beta^{(m)} = \beta^{(m-1)} + [\mathrm{L}^{(m-1)}]^{-1}\mathrm{U}^{(m-1)},$$

donde $\mathrm{L}^{(m-1)}$ denota la matriz de información evaluada en $\beta^{(m-1)}$, multiplicando ambos lados de la igualdad por $\mathrm{L}^{(m-1)}$

$$\mathrm{L}^{(m-1)}\beta^{(m)} = \mathrm{L}^{(m-1)}\beta^{(m-1)} + U^{(m-1)}, \tag{1.20}$$

si se toma la segunda parte de la igualdad anterior y se remplaza la ecuación (1.14) y la ecuación (1.19) resulta que $\beta^{(m)}$ es un vector con elementos

$$\sum_i \sum_k \frac{x_{ij}x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \beta_k^{(m-1)} + \sum_i \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

evaluados en $\beta = \beta^{(m-1)}$, es decir el lado derecho de la ecuación 1.20 se puede escribir como XWz, donde los elementos de z son

$$z_i = \sum_k x_{ik}\beta_k^{(m-1)} + (y_i - \mu_i)(\frac{\partial \mu_i}{\partial \eta_i}).$$

con μ_i y $\frac{\partial \eta_i}{\partial \mu_i}$ evaluado en $\beta^{(m-1)}$. De esta forma la ecuación iterativa para el método de puntajes se puede escribir como:

$$X^t W X \beta^{(m-1)} = X^t W z.$$

Esta ecuación resultante tiene la misma forma que una ecuación normal para un modelo lineal de mínimos cuadrados ponderados. Por ello se tiene que, si se usa el método de puntajes, la forma de estimar los parámetros del MLG es a través de una rutina de modelos lineales de pesos ponderados y la estimación se realizará en forma iterativa pues hay ciertos parámetros que dependen de otros.

1.3.2 Inferencia sobre los parámetros.

Una forma para construir intervalos de confianza y hacer pruebas de hipótesis es através de un estadístico con distribución conocida. En los siguientes renglones se presenta una función del estimador del parámetro de interés ($\hat{\underline{\beta}}$) con la que es posible hacer inferencias.

Distribución muestral para puntajes. El estadístico de puntajes correspondiente a un parámetro β_j se define como la derivada de la función log-verosimilitud con respecto al parámetro de interés, de esta forma para un vector $\underline{\beta}$ de p parámetros, los puntajes están dados por $U_j = \frac{\partial l}{\partial \beta_j}$ $j = 1, ..., p$, donde l es la función de log-verosimilitud.

Por el teorema central de límite se tiene asintóticamente una distribución Normal multivariada para U donde

$$\mathrm{U}=\begin{bmatrix} U_1 \\ \cdot \\ \cdot \\ U_p \end{bmatrix}$$

con media cero y matriz de varianza-covarianza dada por L, (ecuación 1.18). De esta forma

$$U^t \mathrm{L}^{-1} U \sim \chi^2_p$$

con L no singular.

Distribución muestral para estimadores máximo verosímiles. Suponga que la función de log-verosimilitud tiene un máximo que es único en $\hat{\underline{\beta}}$ y este estimador es cercano al verdadero valor del parámetro $\underline{\beta}$. La aproximación de Taylor de primer orden para el vector de puntajes $\mathrm{U}(\underline{\beta})$ sobre el punto $\hat{\underline{\beta}} = \underline{\beta}$ está dado por

$$U(\underline{\beta}) = U(\hat{\underline{\beta}}) + H(\hat{\underline{\beta}})(\underline{\beta} - \hat{\underline{\beta}})$$

en donde $H(\hat{\underline{\beta}})$ denota la matriz de segundas derivadas de la función log-verosimilitud evaluada en $\hat{\underline{\beta}}$. Asintóticamente H está relacionado con la matriz de información de la siguiente manera

$$\mathrm{L} = E(-H) = E(UU^t)$$

por lo que para muestras grandes

$$U(\underline{\beta}) \cong U(\hat{\underline{\beta}}) + \mathrm{L}(\underline{\beta} - \hat{\underline{\beta}}).$$

pero sabemos que $U(\hat{\beta}) = 0$ debido a que $\hat{\underline{\beta}}$ es el punto en donde la log-verosimilitud es cero y por lo tanto su derivada es cero en este punto, de esta forma:

$$(\underline{\beta} - \hat{\underline{\beta}}) = \mathrm{L}^{-1}U(\underline{\beta})$$

con L no singular, en el caso de que L se considere una constante

$$E(\underline{\beta} - \hat{\underline{\beta}}) \cong \mathrm{L}^{-1}E(U(\underline{\beta})) = 0$$

debido a que $E(U(\underline{\beta})) = 0$ (ecuación 1.6) y con ésto se tiene que $\hat{\beta}$ es un estimador insesgado de $\underline{\beta}$.

Por otro lado, la matriz de varianza-covarianza de $\hat{\underline{\beta}}$ es

$$E\left[(\underline{\beta} - \hat{\underline{\beta}})(\underline{\beta} - \hat{\underline{\beta}})^t\right] \cong \mathrm{L}^{-1}E(UU^t)(\mathrm{L}^{-1})^t = \mathrm{L}^{-1}$$

porque $\mathrm{L} = E(UU^t)$ y $(\mathrm{L}^{-1})^t = \mathrm{L}^{-1}$ debido a que L es simétrica. Cuando se tienen muestras grandes es posible usar la siguiente estadística para realizar las inferencias

$$(\underline{\beta} - \hat{\underline{\beta}})^t\mathrm{L}(\underline{\beta} - \hat{\underline{\beta}}) \sim \chi^2_p \tag{1.21}$$

que es conocida como la estadística de Wald, o bien

$$(\hat{\underline{\beta}} - \underline{\beta}) \sim N(0, \mathrm{L}^{-1}). \tag{1.22}$$

En el caso de que Y tenga una distribución Normal los dos resultados anteriores son exactos y en caso de que no, los resultados son asintóticos. Una vez que se tiene la estadística ecuación (1.21) y ecuación (1.22) se procede de manera usual para realizar las inferencias respectivas, intervalos de confianza y pruebas de hipótesis.

1.3.3 Ajuste del modelo

Dentro de la modelación estadística, una de las etapas más importantes es la verificación de los supuestos que se hicieron al construir el modelo; el caso del MLG no es una excepción y por ello, se presenta una forma de verificar los supuestos através de un resumen breve de residuos y gráficas que ayudan a verificar supuestos tales como la función liga, la forma paramétrica de las variables, entre otras cosas.

Uno de los criterios de bondad de ajuste más usados al ajustar un MLG es la devianza, que proviene del cociente de verosimilitudes entre el modelo propuesto y el modelo saturado, es decir, aquel modelo cuyo estimador para la variable respuesta son las mismas observaciones de la variable respuesta sin sufrir ninguna transformación. Sea $l(\mu, \phi; y)$ la log-verosimilitud maximizada sobre $\underline{\beta}$ para un valor fijo del parámetro de dispersión ϕ.

Sean $\hat{\theta} = \theta(\hat{\mu})$ y $\tilde{\theta} = \theta(y)$ estimadores del parámetro canónico del modelo saturado y el modelo propuesto, respectivamente. Suponiendo que $a_i(\phi) = \phi/w_i$, con w_i el peso inicial que varia de observación en observación. La devianza de la familia exponencial de dispersión se puede escribir como

$$\Sigma 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i)\}\phi = D(y;\hat{\mu})\phi, \tag{1.23}$$

que corresponde a la devianza del modelo que se desea ajustar. Las formas de las devianzas para algunas distribuciones importantes se muestran en la siguiente tabla

Normal	$\Sigma(y-\hat{\mu})^2$
Poisson	$2\Sigma\{ylog(\frac{y}{\hat{\mu}}) - (y-\hat{\mu})\}$
Binomial	$2\Sigma\{ylog(\frac{y}{\hat{\mu}}) + (m-y)log(\frac{m-y}{m(1-\hat{\pi})})\}$
(si $m = 1$)	$-\sqrt{2\lvert\log(1-\hat{\pi})\rvert}$
(si $m = 1$ y $y = 1$)	$\sqrt{2\lvert\log(\hat{\pi})\rvert}$
Gamma	$2\Sigma\{ylog(\frac{y}{\hat{\mu}}) + \frac{y-\hat{\mu}}{\hat{\mu}}\}$
Gausiana Inversa	$\Sigma\frac{(y-\hat{\mu})^2}{\hat{\mu}^2 y}$

Tabla No.1.3

Otra medida de discrepancia, para conocer que tan alejado se encuentra el modelo que se porpone con respecto a lo observado se usa la χ^2 de Pearson generalizada

$$X^2 = \Sigma\frac{(y-\hat{\mu})^2}{V(\hat{\mu})} \tag{1.24}$$

donde $V(\hat{\mu})$ es la varianza estimada para la distribución de la que se está tratando. Cuando la distribución del error es Normal la X^2 y la devianza tiene una distribución exacta χ^2, para las otras distribuciones se tiene un resultado asintótico.

Residuos: Para un modelo lineal generalizado se requiere una extensión de la definición de residuo que se pueda aplicar a todas las distribuciones que forman parte de la familia exponencial de dispersión. De esta manera se podrán utilizar para explorar lo adecuado del modelo, función liga, términos del predictor lineal, elección de la función varianza, etc. Para ello, a continuación se definirán diferentes residuos, así como el uso gráfico de algunos de ellos.

Un elemento en el modelo de regresión lineal que es importante en el cálculo de residuos es la matriz de proyección conocida como matriz sombrero, H. En el caso de los modelos lineales generalizados también existe esta matriz que cumple con todas las propiedades de la matriz sombrero de los modelos lineales y está dada por

$$H = V^{\frac{1}{2}} X (X^t V X)^{-1} X^t V^{\frac{1}{2}}$$

en donde

$$V = diag\left[\tau_i^2 \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 a_i(\phi)\right]. \tag{1.25}$$

Residuo de la devianza estandarizada: Estos residuos pueden indicar cuál de las observaciones contribuyen en mayor proporción a la falta de ajuste

$$r_i^D = \frac{\text{sign}(\tilde{\eta}_i - \hat{\eta}_i)\sqrt{d_i}}{\sqrt{1 - h_{ii}}}$$

donde d_i es la contribución de la i-ésima observación a la devianza, $\tilde{\eta}_i$ es el valor de la estructura lineal, η, que maximiza la verosimilitud del modelo saturado y $\hat{\eta}_i$ pertenece al modelo propuesto.

Ejemplos:

1. En el caso donde los errores tienen una distribución Normal los residuos de la devianza estandarizada estarán dados por;

$$r_i^D = \frac{sign(y_i - \hat{\mu}_i)\sqrt{d_i}}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}.$$

2. Para el caso de la regresión logística, donde los errores tienen una distribución Binomial, el residuo de la devianza estandarizada se ve como

$$r_i^D = \frac{\pm\sqrt{2y_i \log(\frac{y_i}{n\hat{\pi}_i}) + 2(n_i - y_i)\log(\frac{n_i - y_i}{n_i - n\hat{\pi}_i})}}{\sqrt{1-h_{ii}}}$$

donde el signo se elige de $y_i - n\hat{\pi}_i$.

Residuo estudentizado: Se define como

$$r_i^e = \frac{r_i^D}{\sqrt{\widehat{\text{var}}[Y_i]}} \tag{1.26}$$

que en ocasiones es conocido también como el residuo de Pearson estandarizado debido a que $(y_i - \hat{y}_i)^2 / \widehat{\text{var}}[Y_i]$ es la contribución de la i-ésima observación a la estadística de Pearson.

Métodos gráficos:

Residuos no estandarizados contra residuos para **u** (una variable omitida en el modelo). Esta gráfica conocida como variable-agregada se usa para ver si una cierta variable **u** que fue omitida en el modelo debe ser incluida en el análisis. Lo primero que debe hacerse para construir esta gráfica, es calcular los residuos no estandarizados para **u** tomándola como respuesta y usando los mismos predictores lineales y pesos cuadráticos como si se tratara de y. De esta forma los residuos no estandarizados de y graficados en contra de los residuos **u** no deberán mostrar ningún patron identificable para decir que la omisión de **u** fue correcta.

Variable dependiente ajustada $\hat{y}$ contra el predictor lineal ajustado, $\hat{\eta}$. Si la gráfica resulta ser una linea recta significa que no hay patrón de comportamiento alguno, es

decir, que la función liga fué correcta . Para las funciones liga de la familia potencia ecuación (1.10), una curvatura ascendente significa que es necesaria una liga con potencia mayor que la usada (un valor de λ más grande) y una curvatura descendente significa que es necesario una función liga de potencia menor (un valor de λ más pequeño). Esta gráfica para datos binarios no es informativa por lo que es necesario usar métodos formales.

Uno de estos métodos formales consiste en agregar $\hat{\eta}^2$ como una covariable extra y evaluar la disminución en la devianza. Las verificaciones de la adecuación de la función liga son afectadas inevitablemente por los errores que se pudieran haber cometido en la forma paramétrica de las variables explicativas que son parte del predictor lineal. De esta forma, si esta prueba presenta alguna desviación, puede significar que la función liga es incorrecta ó que la forma paramétrica de las variables explicativas es incorrecta, o bien, ambas situaciones. Para decifrar cuál es el problema que presenta el modelo es necesario apoyarse en las diferentes gráficas de residuos.

La gráfica de residuo parcial esta gráfica es una herramienta importante para determinar cuándo un término βx que se encuentra en el predictor lineal resulta ser mejor si es expresado como $\beta h(x;\theta)$ donde $h(\cdot;\theta)$ es alguna función monótona y el residuo parcial está dado por

$$\mathbf{u} = \hat{y} - \hat{\eta} + \hat{\gamma} x$$

con $\hat{y}$ la variable dependiente ajustada, $\hat{\eta}$ el componente sistemático ajustado y $\hat{\gamma}$ el parámetro estimado para la variable explicativa x. Si la forma paramétrica de x es satisfactoria, la gráfica debería de ser aproximadamente lineal. Si no, su forma sugeriría una alternativa recomendable.

Si se llegara a presentar en la gráfica de residuos una incertidumbre en el patrón de comportamiento entonces un suavizamiento puede ayudar a decifrar tal comportamiento. En particular, la gráfica de residuos parciales, suavizada, puede ser marcadamente útil para datos binarios. Sin embargo distorciones en la gráfica pueden ocurrir si la forma paramétrica de las otras variables explicativas fuera incorrecta.

Una revisión formal comprende en introducir la forma paramétrica de x en una familia $h(x;\theta)$ indexada por θ, conocida como familia potencia, donde ahora el problema consiste en encontrar el valor de θ para el cual x estaría totalmente específicado, para ello se calcula la devianza para una malla de valores de θ con el fin de encontrar la posición del mínimo que da como resultado $\widehat{\theta}$ y de esta forma queda especificado totalmente x. La familia de transformaciones más común es la familia potencia dada por

$$h(x;\theta) = \begin{cases} \frac{x^\theta - 1}{\theta} & \text{para } \theta \neq 0 \\ \log(\theta) & \text{para } \theta = 0 \end{cases}.$$

Una revisión informal consiste en calcular $v = \partial h/\partial\theta_0$, se ajusta el modelo tomando a v como variable dependiente, con el predictor lineal y pesos cuadráticos como si se tratara de y, y se obtienen los residuos. Entonces se grafican los residuos que resultaron del primer ajuste que se realizó sobre y y los que resultaron de ajustar a v como variable dependiente, si se observa una tendencia lineal indica un valor de $\theta \neq \theta_0$ mientras que una gráfica sin patrón indicará que no hay evidencia para suponer que sean diferentes, por lo que en tal caso no es necesario realizar una transformación de la variable independiente en cuestión.

Puntos influyentes y puntos palanca: En este caso la distancia de Cook es útil para examinar la forma en que cada observación afecta al conjunto completo de parámetros estimados. La siguiente ecuación permite la comparación entre los valores ajustados con la observación y sin ésta

$$C_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(i)})^t X^t V X (\hat{\beta} - \hat{\beta}_{(i)})$$

donde $\hat{\beta}_{(i)}$ es el parámetro estimado sin la i-ésima observación. Para evitar hacer el ajuste para cada observación esta distancia se puede aproximar por:

$$C_i \doteq \frac{h_{ii}(r_i^e)^2}{p(1-h_{ii})},$$

sin embargo, esta última expresión se usa de manera más común en la gráfica contra índices (el orden en que fue tomada cada observación).

Lo anterior es un breve resumen del amplio campo de los residuos y sus gráficas usadas para la verificación de supuestos que se hicieron al ajustar el modelo; sin embargo, en este trabajo no es el objetivo mostrar ampliamente todas estas herramientas por lo que si el lector está interesado puede dirigirse a McCullagh, 1989.

Hasta ahora se ha presentado en forma general el concepto de un modelo lineal generalizado. Sin embargo, para el objetivo de este trabajo es necesario ejemplificar este tipo de modelos para dos distribuciones particulares de la variable respuesta, estas son la distribución Binomial y la distribución Poisson.

Distribución Binomial.

En muchos estudios es frecuente encontrar variables cuya respuesta es binaria. Por ejemplo, se podría pensar en estudios en los que la respuesta es estar enfermo o sano. En este caso, para el i-ésimo paciente se tendría la variable aleatoria

$$Z_i = \begin{cases} 1 \text{ si es éxito} & \text{es decir, no padece la enfermedad} \\ 0 \text{ si es falla} & \text{es decir, padece la enfermedad} \end{cases}$$

en donde $P(Z_i = 0) = 1 - \pi_i$ y $P(Z_i = 1) = \pi_i$, en este caso la distribución de la variable aleatoria Z_i tiene una distribución Bernoulli con parámetro π_i. Por lo que la verosimilitud, considerando una m.a. $\{z_i, i = 1, ..., n\}$ es,

$$f(z_i; \pi_i) = \pi_i^{z_i}(1 - \pi_i)^{1-z_i}, \tag{1.27}$$

que se puede reescribir de la siguiente manera

$$f(z_i; \pi_i) = (1 - \pi_i)\exp(z_i \ \log(\frac{\pi_i}{1 - \pi_i})) \tag{1.28}$$

de esta última expresión es posible determinar que la forma canónica del parámetro de localización para esta distribución es $\theta_i = \log\ (\frac{\pi_i}{1-\pi_i})$.

Si se tuviera una muestra de n observaciones con este comportamiento se tendría una distribución Binomial, donde la variable aleatoria sería el número de éxitos en la muestra, es decir $Y = \sum_{i=1}^{n} Z_i$, y cuya forma analítica estaría dada por

$$\begin{aligned} f(y;p) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \exp\left(y \log(\tfrac{\pi_i}{1-\pi_i}) + n\log(1-\pi) + \log\binom{n}{y} \right). \end{aligned} \tag{1.29}$$

Ahora, si se tuvieran N variables independientes $Y_1, Y_2, ..., Y_N$ correspondientes al número de sucesos en N diferentes subgrupos o estratos y cada Y_i tuviera una distribución Binomial con parámetros n_i, π_i la función de log-verosimilitud estaría dada por

$$l(\pi_1, ..., \pi_N; y_1, ..., y_N) = \sum_{i=1}^{N} \left[y_i \log(\tfrac{\pi_i}{1-\pi_i}) + n_i \log\binom{n_i}{y_i} \right]$$

y por tanto la función liga canónica es:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x$$

con x una variable explicativa, de aquí surge la llamada función logit:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \Sigma \beta_j x_{ij} \tag{1.30}$$

y este es un modelo propuesto para describir la respuesta media $\pi = \pi(x)$ a través de un conjunto de variables explicativas $(x_1, ..., x_m) = \underline{x}^t$ usando un modelo lineal generalizado. Como lo que interesa realmente es mostrar una relación que explique la probabilidad por cada subgrupo se tiene

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})} \tag{1.31}$$

si $\eta_i = \Sigma_j \beta_j x_{ij}$

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \tag{1.32}$$

el cual recibe el nombre de modelo logístico lineal, que pertenece a la familia de los MLG.

Distribución Poisson. La distribución Poisson en muchas ocasiones se usa para modelar la ocurrencia de eventos raros, por ejemplo, el número de nuevos casos de cáncer de mama desarrollado en una cierta población y en un cierto periodo de tiempo. La variable Y en este caso es una variable de conteo de eventos ocurridos para cada uno de los subgrupos, que a su vez, están descritos por un conjunto de variables predictoras $x_1, .., x_k$. Si se tienen n subgrupos y se denota con l_i el tiempo de ocurrencia de los eventos de interés para todas las personas en cada subgrupo i, además $\underline{\beta} = (\beta_0, ..., \beta_k)^t$ es un conjunto de parámetros desconocidos y $\lambda(x_i, \underline{\beta})$ alguna función específica de x_i y $\underline{\beta}$, que denota la tasa de falla por unidad de tiempo básico para el subgrupo i, entonces el número esperado de fallas en el i-ésimo grupo es:

$$E(y_i) = \mu_i = l_i \lambda(x_i, \underline{\beta}) \quad i = 1, .., n$$

con $\lambda(x; \underline{\beta}) > 0$. Bajo el supuesto de que Y_i es una v.a. Poisson con media μ_i se tiene

$$P_{Y_i}(y_i, \underline{\beta}) = P(Y_i = y_i; \underline{\beta}) = \frac{[l_i \lambda(x_i, \underline{\beta})]^{y_i} \exp^{-l_i \lambda(x_i, \underline{\beta})}}{y_i!}$$

donde $y_i = 0, 1, ...$ con $i = 1, .., n$. Si se supone que Y_i es independiente de Y_j con i diferente de j, entonces la función de verosimilitud de estas variables está dada por

$$\begin{aligned} L(Y; \beta) &= \Pi_{i=1}^n P_{Y_i}(y_i; \underline{\beta}) \\ &= \frac{\Pi_{i=1}^n [l_i \lambda(x_i, \underline{\beta}]^{y_i} \exp[-\Sigma_{i=1}^n l_i \lambda(x_i, \underline{\beta})]}{\Pi_{i=1}^n y_i!} \end{aligned}$$

donde $E[Y_i] = \mu_i = l_i \lambda(x_i, \underline{\beta}) \quad i = 1, .., n$. Para encontrar el estimador de $\underline{\beta}$ es necesario que se especifique el valor de la función λ, en este caso, $\lambda(x_i, \underline{\beta}) = \sum \underline{x_i} \underline{\beta}$. Por lo general esta especificación se realiza basándose en el conocimiento y experiencia previos entre las relaciones de las variables bajo estudio. Después de darle el valor a esta función se encuentra el valor de $\underline{\beta}$ como se mencionó para el caso general.

Riesgo. En epidemiología la probabilidad de que ocurra una enfermedad durante un periodo de tiempo dado, también conocido como el riesgo de ocurrencia de

la enfermedad, es el número de nuevos casos de la enfermedad que ocurren en este periodo de tiempo, expresado como una proporción de la población en riesgo. En un estudio de cohorte donde una muestra de la población en riesgo es monitoreado para determinar la incidencia de la enfermedad, el riesgo puede ser determinado directamente de la proporcion de individuos en la muestra que desarrolló la enfermedad durante el siguiente periodo. Los riesgos se expresan de manera mas conveniente como tasas de incidencia por unidad de tiempo.

Cuando se comparan riesgos, una diferencia entre riesgos se le conoce con el nombre de diferencia de riesgos. La diferencia de riesgos para individuos de la muestra expuestos a un cierto factor contra no expuestos al mismo factor se define como

$$RD = P_1 - P_2.$$

donde P_i es el riesgo. Otra forma de comparación de dos riesgos es el cociente de riesgos o riesgo relativo.

Una medida de asociación empleada frecuentemente en muchas investigaciones epidemiológicas es el cociente de riesgos o riesgo relativo. El riesgo relativo para expuestos contra no expuestos se define como

$$RR = \frac{P_1}{P_2}.$$

En otras palabras, el riesgo relativo es la posibilidad de enfermedad en aquellas personas expuestas al factor de riesgo, relativo a aquellas que no están expuestas al riesgo. Este representa cuántas veces es más probable que la enfermedad ocurra en el grupo expuesto comparado al grupo no expuesto.

Si un riesgo se expresa en términos de un momio mas que en término de probabilidades, el cociente de momios para los expuestos contra los no expuestos se define como

$$\psi = \frac{P_1/(1-P_1)}{P_2/(1-P_2)}.$$

El ψ es claramente el cociente de momios de la enfermedad en aquellos expuestos al factor riesgo con respecto a los momios de la enfermedad entre aquellos quienes no estan expuestos al factor riesgo.

El propósito principal de estimar riesgos es ser capaz de hacer comparaciones entre el riesgo de la enfermedad en diferentes niveles de un factor de exposición y de esta forma explicar la asociación entre la enfermedad y el factor de exposición.

Es importante contemplar todos las posibles variables que se creen influyen en el riesgo de interés que se pretende estimar, debido a que si no se contemplara una variable de confusión que fuera transcendente en la estimación del cociente de momios, lo que se obtenga como estimación será incorrecto.

Por la forma en que se plantea el riesgo, es natural pensar que un modelo que ayude a dar una estimación del mismo sea el modelo logístico, es decir, una relación de la forma:

$$logit(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1i} +, ..., +\beta_k X_{ki}. \tag{1.33}$$

Una de las razones para usar el modelo logístico lineal en el análisis de datos de estudios etiológicos, es que los coeficientes de las variables explicativas en el modelo son interpretados como el logaritmo del cociente de momios, en el caso de un factor de exposición dicotómico.

Cuando sólo se tiene una variable explicativa con dos niveles, el cociente de momios de la enfermedad para un individuo expuesto relativo a uno no expuesto es

$$\psi = \frac{\frac{Pe}{1-Pe}}{\frac{Pu}{1-Pu}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \tag{1.34}$$

y por lo tanto $\beta_1 = \log(\psi)$ es el logaritmo del cociente de momios.

En el caso de tener un factor de exposición politómico el cociente de momios de la ocurrencia de la enfermedad para individuos en el i-ésimo grupo de exposición relativo al primer grupo no expuesto está dado por

$$\psi = \frac{\frac{P_i}{1-P_i}}{\frac{Pu}{1-Pu}} = \exp(\beta_i). \tag{1.35}$$

Si se trata de un variable explicativa continua, en este caso la interpretación cambia. Por ejemplo, considere un cociente de momios para la enfermedad de un individuo que tiene el valor x de la variable explicativa, de esta forma el cociente de

momios con respecto a un individuo que tiene el valor de la misma variable dado por $x + r$, es el siguiente

$$\psi = \frac{\exp(\beta_0 + \beta_1(x_1 + r))}{\exp(\beta_0 + \beta_1(x_1))} = \exp(r\beta_1) \tag{1.36}$$

aquí $r\hat{\beta}_1$ se interpreta como el cambio estimado en el logaritmo del cociente de momios cuando x se incrementa en r unidades.

Lo anterior muestra que cuando una variable continua C es incluida en el modelo logístico lineal, el riesgo relativo aproximado de la enfermedad cuando X cambia r unidades no depende del valor actual de X. Por ejemplo, si X fuera la edad, el riesgo relativo para un individuo a edad 55 con respecto a un individuo de edad 50 sería el mismo que para una persona de edad 15 con respecto a una de 10. Esto es por que se supone que hay una relación lineal entre la transformación logística de la probabilidad de que ocurra la enfermedad y la edad. Para evitar que esto ocurra se propone: primero verificar si realmente existe una relación lineal, en caso de que no sea propones una categorización de la variable continua, de tal forma que cada categoría corresponda a diferentes valores de la variable continua o bien usar transformaciones de la variable por ejemplo, la variable al cuadrado o la variable al cubo. Cuando se tiene más variables explicativas la interpretación de cada coeficiente se hará suponiendo el mismo nivel para el resto de las variables explicativas.

Si se tiene una variable x dicotómica $\exp(\beta_1)$ representa el cambio de estar con el grupo de no expuestos ($x = 0$) al cambiar al grupo de exposición ($x = 1$).

Ejemplo: si se tiene un modelo Poisson con liga canónica y X es una variable dicotómica, se tiene como resultado

$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_i x_i \\ \exp(\beta_i) &= \frac{y_1}{y_0} = 1.02 \end{aligned}$$

la interpretación será que existe un incremento porcentual del 2% en el valor de y cuando x pasa de un valor cero a un valor uno.

En el caso de tener una variable x politómica, $\exp(\beta_i)$ representa el cambio porcentual de estar con el grupo de no expuestos ($x = 0$) al cambiar al grupo de exposición i ($x = i$).

Cuando x es continua, $\exp(\beta_i)$ representa el cambio porcentual en y cuando x pasa de x a $x+1$, y de la misma manera se puede observar que no depende del valor de x.

"Offset"

Un recurso comúnmente usado en modelos lineales generalizados es el uso de un "offset", el cual es un componente del predictor lineal conocido y con coeficiente uno. Un "offset" es redundante para los modelos lineales con distribución Normal, ya que uno puede simplemente trabajar con los residuos. Un "offset" permite una cierta forma de análisis residual para modelos lineales generalizados; nosotros podemos evaluar la contribución de términos adicionales mientras permanezcan fijos aquellos que ya han sido ajustados. En una situación de muestreo estratificado, los "offsets" pueden ser usados para corregir el muestreo desbalanceado. En el caso del modelo poisson un offset puede ser usado cuando el recorrido de la variable respuesta es finito. El "offset" será el valor máximo que puede alcanzar la variable respuesta.

Capítulo 2

Modelos Aditivos Generalizados

2.1 Introducción.

En el segundo capítulo se realizó una revisión de los modelos lineales generalizados que son una generalización de los modelos lineales. En estos modelos los efectos predictores se suponen lineales en los parámetros. Sin embargo, la distribución de las respuestas no es sólo una distribución Normal sino cualquier elemento de la familia exponencial.

Se vio cómo muchos de los modelos usuales caen dentro de esta clase, por ejemplo, el modelo logístico lineal para datos binarios y modelos log-lineales para datos categóricos.

En este capítulo se presenta una extensión de la clase de modelos lineales generalizados. llamados Modelos Aditivos Generalizados. Se verá que su estimación está basada en un método iterativo de regresión, de la misma manera como sucedió en la estimación del modelo lineal generalizado; en particular, sólo se remplazará el paso de regresión lineal por un paso de regresión aditiva no paramétrica. El algoritmo se conoce con el nombre de *puntajes locales*.

2.2 Suavizadores de dispersión (Scatterplot smoothers).

Se mencionó en el capítulo dos cómo el modelo de regresión lineal ha sido una de las herramientas más utilizada dentro de la estadística debido a las caraterísticas que lo conforman, sin embargo, en muchas ocasiones no es fácil asegurar que la relación entre las variables explicativas y la variable respuesta sea de una forma lineal. A este respecto existen los métodos de diagnóstico ver por ejemplo: Cook y Weisberg, 1982, que nos ayudan a decidir si la relación es o no lineal y además nos pueden sugerir el tipo de relación. De aquí surge la necesidad de una herramienta estadística que permita "hablar a los datos" y de esta forma descubrir qué clase de relación se está presentando. La idea detrás de un *suavizador* (smoothing) es tratar de exponer la dependencia a través de una función entre las variables explicativas y la variable respuesta sin imponer un cierto comportamiento paramétrico lineal del cual no se tiene la certeza.

Se dice que un *suavizador* es una herramienta que resume la tendencia de una respuesta Y como una función de uno o más predictores $x_1, ..., x_n$, este produce un estimador que es menos variable que la misma variable respuesta, de aquí el nombre de suavizador. Una propiedad de un suavizador es su naturaleza no paramétrica, es decir, no supone una forma rígida de la dependencia de Y sobre $(x_1, ..., x_n)$, por esta razón, un suavizador es, con frecuencia, referido como una herramienta para la regresión no paramétrica. En el caso de un sólo predictor se le conoce con el nombre de suavizador de dispersión.

Definición: *Un suavizador de dispersión se define como una función de x, y cuyo resultado es una función $s = S(y|x)$ cuyo dominio son los valores de x.*
donde $y = (y_1, ..., y_n)^t$ es la variables respuesta y $x = (x_1, ..., x_n)^t$ es la variable explicativa o predictor, cada y y x son medidas de las variables Y y X respectivamente. En particular no es necesario suponer que los pares (x_i, y_i) provienen de una muestra aleatoria de alguna distribución conjunta.

Un uso que generalmente se les da a los suavizadores de dispersión es la de estimar

$E(y|x = x_i)$.

El suavizador de dispersión tiene dos usos principales:

1. El primero es el uso descriptivo. Un suavizador de dispersión se pude usar para realzar la apariencia visual de la gráfica Y vs X; es decir, ayuda a identificar visualmente la tendencia de la gráfica.

2. El segundo uso es estimar la dependencia de la media de Y sobre los predictores. De esta forma sirve como un cimiento para la estimación de los modelos aditivos, que se explicarán más adelante.

Al usar un suavizador es necesario tomar en cuenta dos requisitos:

1) Conocer cómo promediar los valores de la respuesta en cada vecindad.

2) Conocer qué tan grande debe de ser la vecindad.

Una primer pregunta que surge al usar un suavizador es ¿qué clase de suavizador se debe usar?. Existe un cierto número de técnicas de suavizamiento y en general no hay una recomendación técnica para elegirlo, debido a que existen pocas comparaciones sistemáticas dentro de la literatura.

La segunda interrogante se relaciona con el problema de sesgo-varianza ya que vecindades grandes hacen que el estimador tenga una varianza pequeña pero un sesgo grande y viceversa cuando se trata de una vecindad pequeña. En los suavizadores el tamaño de la vecindad se expresa directamente a través de un parámetro de suavizamiento.

El suavizador de dispersión es una herramienta básica en el modelo aditivo generalizado, esto se verá más adelante, mientras tanto, se describirán brevemente algunos de los suavizadores más usados.

Tipos de suavizadores de dispersión:

1. **Suavizador binario (bin smoothers).** Este suavizador, conocido con el nombre de regresograma, particiona el rango del predictor en un cierto número de regiones disjuntas y exhaustivas para después promediar las respuestas en cada región. Es decir, se parte en k regiones el dominio de x, $c_0 < ... < c_k$ donde $c_0 = -\infty$ y $c_k = +\infty$. Entonces definimos los índices de los puntos en la región k-ésima como

$R_k = \{i; c_k \leq x_i < c_{k+1}\}; k = 0, .., K-1.$

Así, este suavizador queda definido como:

$$s = S(y|x), s(x_0) = ave_{i \in R_k}(y_i)$$

si $x_0 \in R_k$. y *ave* significa el promedio en la vecindad R_k. Típicamente si el número de regiones es $k = 5$, por ejemplo, se eligen los puntos de corte tales que existan en cada intervalo un número igual de puntos en cada región.

Por la forma en que está definido este suavizador claramente se ve que no es muy suave por que el estimador brinca en cada punto en el que se corta la región.

2. Suavizador de media móvil y suavizador de línea móvil (Running-mean y running line-smoothers).

El suavizador de media móvil necesita que se defina una vecindad alrededor del punto de interés x_0, donde x_0 es un punto en el dominio de x_i. Para esto se eligen k puntos a la derecha y k puntos a la izquierda de x_i.

En caso de que no se pudieran tomar exactamente k, se toman los que están disponibles. A esto se le conoce como vecindad simétrica más cercana. La media-móvil se define como

$$s(x_i) = ave_{j \in N^s(x_i)}(y_j)$$

con ave nuevamente como el promedio pero en este caso en la vecindad $N^s(x_i)$.

Una definición más formal de esta vecindad N^s es la siguiente

$$N^s(x_i) = \{\text{máx}(i-k, 1), ..., i-1, i, i+1, ..., \text{mín}(i+k, n)\}$$

A este suavizador también se le conoce como promedios móviles y es popular en series de tiempo.

En ciertos casos, es común encontrar que este suavizador presenta un problema por el sesgo. Así que una simple generalización de este suavizador es calcular la línea de mínimos cuadrados en lugar de una media en cada vecindad, en este caso se tiene un suavizador *linea móvil* y se define por

$$s(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0$$

donde $\hat{\alpha}$. y $\hat{\beta}$ son los estimadores de mínimos cuadrados con los valores de y_i y x_i que se encuentran en la vecindad $N^s(x_0)$.

Se puede observar cómo el parámetro k controla la apariencia del suavizador de línea móvil y también del de media móvil, valores grandes de k producen curvas suaves mientras que valores pequeños producen curvas más accidentadas. Es conveniente pensar en términos de $w = (2k + 1)/n$ que es la proporción de puntos en cada vecindad y es conocido con el nombre de *span*.

Selección del span y el intercambio de sesgo varianza.

Como se vio anteriormente, el suavizador requiere de la elección de un tamaño de span w. Veamos que ocurre en los casos extremos; primero cuando $w = 0$, $\hat{s}(x_i)$ es solo y_i. en este caso, no se tiene un buen estimador por que tiene una varianza grande y no es suave. Si $w = 2.0$ se tiene una vecindad que contiene a todos los puntos, y por tanto, se tendría en este caso que $\hat{s}(\cdot)$ es la recta de regresión de mínimos cuadrados. En este caso se tiene una función muy suave y por tanto no describirá la curvatura correspondiente a la función de interés, es decir, será sesgada. De esta forma el span deberá ser elegido entre 2.0 y 0 siempre pensando en un intercambio de sesgo y varianza.

Se puede derivar para este propósito un criterio basado en los datos. Si se considera el estimador de $E(Y|X)$ como un minimizador empírico del error cuadrático de predicción

$$PSE = E(Y - s(X))^2$$

o equivalentemente el error cuadrático medio

$$MSE = E(E(Y|X) - s(X))^2.$$

Sea $\hat{s}_w(x_i)$ el suavizador con span de tamaño w en x_i. Entonces la validación cruzada de la suma de cuadrados se define por $CVSS(w) = (1/n) \sum_1^n (y_i - \hat{s}(x_i))^2$. Es posible mostrar que $E(CVSS(w)) \approx PSE$, usando el hecho de que $\hat{s}_w^i(x_i)$ es independiente de y_i. Entonces es razonable elegir el span que produzca el valor mas pequeño de $CVSS(w)$.

3. Suavizador de kernel (Kernel smoothers).

Un suavizador de kernel usa un conjunto de pesos locales explícitamente definidos por el kernel elegido, para producir el estimador en cada punto objetivo. Usualmente un suavizador de kernel usa pesos que decrecen en una forma suave conforme uno se aleja del punto objetivo.

El peso dado en el j-ésimo punto para producir el estimador en x_0 se define por

$$S_{0j} = \frac{C_0}{\lambda} d(|\frac{x_0 - x_j}{\lambda}|)$$

donde $d(t)$ es una función par decreciente en $|t|$, el parámetro λ es el ancho de la ventana y c_0 es una constante que se elije de tal forma que la suma sea uno. Un candidato natural para d es la densidad Normal estándar. Otro kernel popular es el kernel de Epanechnikov dado por:

$$d(t) = \begin{cases} \frac{3}{4}(1-t^2), & \text{para } |t| \leq 1; \\ 0 & \text{en otro caso.} \end{cases}$$

Las investigaciones realizadas a la fecha sugieren que la elección del kernel es relativamente poco importante comparado con la elección del ancho de la banda. En general se dice que el suavizador de Kernel presenta sesgo en el desarrollo de los puntos finales.

4. Suavizador de spline cúbico.

Este suavizador surge como un problema de optimización. El problema que da origen a este tipo de suavizador es el siguiente:

Encontrar la función $f(x)$ entre todas aquellas funciones que tengan segunda derivada continua y que minimizen la suma de cuadrados penalizados del residuo, es decir, aquella función $f(x)$ que minimize la expresión:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{i=1}^{b} \left\{ f''(t) \right\}^2 dt$$

donde λ es una constante fija y $a \leq x_1 \leq \ldots \leq x_n \leq b$. El primer término mide la cercanía a los datos mientras que el segundo término penaliza la curvatura en la función. Es posible mostrar que la solución a esta ecuación es un spline cúbico natural con puntos claves(nudos) en los valores únicos de x_i.

El parámetro λ juega el mismo papel que el span en el suavizador de línea móvil, es el parámetro de suavizamiento y controla el intercambio entre fidelidad de los datos y el suavizamiento. Valores grandes de λ producen curvas más suaves mientras que valores pequeños producen curvas accidentadas. En el caso de que $\lambda \to \infty$ el término de penalización domina, forzando a que $f''(x) = 0$ en cualquier lugar dando como resultado que la solución sea la línea de mínimos cuadrados. Por el otro lado cuando $\lambda = 0$ el término de penalización es poco importante y da como consecuencia que la solución tienda a interpolar los datos con funciones dos veces diferenciables.

Algunos métodos de suavizamiento similares al que se acaba de describir necesitan de la especificación de un parámetro de suavizamiento. En la práctica es común usar técnicas automáticas de selección, tales como validación cruzada las cuales son muy caras de instrumentar computacionalmente hablando.

5. Suavizadores de regresión.

Este tipo de suavizador usa regresiones de promedios móviles para predecir el valor en el centro de la ventana más que promedios móviles. Estos suavizadores se comportan usualmente mejor al final de los datos.

6. Suavizadores para predictores múltiples.

En el caso de tener una regresión múltiple el problema que surge es; ajustar una superficie $p-$dimensional en Y. Para los primeros dos suavizadores descritos anteriormente lo que se necesita únicamente es una definición de la vecindad más cercana de un punto en un espacio de dimensión p. Esto se puede determinar usando la distancia Euclidiana, una vez que se tiene determinada la vecindad la generalización de la media móvil es inmediata sólo resta estimar la superficie en el punto objetivo promediando los valores de la respuesta.

Se han propuesto muchas alternativas para generalizar el suavizador de dispersión para el caso de regresión múltiple. Una generalización inmediata es usar una superficie suavizada. Friedman y Suetzle (1982), entre otros, puntualizan los problemas de dimensionalidad que toman lugar cuando se usan superficies-suavizadoras. Esencialmente por que todos los suavizadores basan sus estimaciones en el promedio dentro de una vecindad. En grandes dimensiones, el encontrar vecindades suficientes no es un problema fácil, por lo que estos mismos autores han propuestos la técnica de

projection pursuit como una alternativa.

2.3 El modelo aditivo generalizado

Los modelos aditivos generalizados son una nueva clase de modelos de regresión flexible, los cuales retienen la interpretabilidad provista por la estructura aditiva, sin forzar a una cierta forma paramétrica rígida de dependencia sobre variables explicativas que no se pueda justificar: Trevor y Tibshirani han sido los pioneros en la modelación aditiva y su libro Modelos Aditivos Generalizados 1990 provee una fuente de información sobre el tema.

El modelo aditivo generalizado en relación a un modelo lineal generalizado, difiere en que en lugar de suponer una dependencia sobre una suma de predictores lineales, se supone que la respuesta tiene una dependencia sobre una suma de funciones de los predictores cuya característica primordial es que son funciones suaves. Es decir, que un predictor aditivo, $\alpha + \sum_j f_j(x_j)$, reemplaza al predictor lineal, $\alpha + \sum x_j\beta_j$. Específicamente, supongamos que la respuesta Y tiene una distribución en la familia de dispersión exponencial con media $\mu = E[Y|X_1, ..., X_p]$ ligado al predictor a través de:

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(x_j). \tag{2.1}$$

La estimación de α y $f_1, .., f_p$ se pueden llevar acabo en forma iterativa. Aquí se presentan dos de las técnicas más usadas para su estimación; una es el procedimiento de puntajes locales, que usa suavizadores de dispersión para generalizar el procedimiento usual de puntajes de Fisher para calcular el estimador máximo verosímil, y el método de estimación de verosimilitud local. Ambas técnicas han sido comparadas en varios documentos a través de ciertos ejemplos y se ha encontrado que las funciones estimadas son muy similares, la única diferencia es la velocidad en el cálculo, dando como resultado que el del método local de puntajes sea más rápido computacionalmente.

Usos del modelo: es posible usar las funciones estimadas para sugerir transfor-

maciones paramétricas de las covariables. Una técnica más tradicional para este propósito es el uso de residuos y gráficas de residuos parciales, en este caso las no-linealidades se detectadan en una de las covariables mientras que las otras se conservan lineales. En el caso del modelo aditivo generalizado se pueden detectar todas las no-linealidades simultáneamente.

2.3.1 Estimación de los parámetros del modelo aditivo generalizado.

Estimación del modelo a través del algoritmo de puntajes locales. Consideremos un modelo de regresión basado en una verosimilitud con una covariable. Supongamos además que las parejas de datos $(x_1, y_1), ..., (x_n, y_n)$ son realizaciones independientes de las variables aleatorias X y Y. También supongamos que dado $X = x$, Y tiene una densidad

$$Y \mid X = x \sim h(y, \eta)$$

debido a que η es una función de x, algunas veces se escribe $\eta(x)$, para enfatizar. Denotemos la log-verosimilitud con $l(\eta, Y)$. Para estimar $\eta(\cdot)$ bastará maximizar $\sum_{1}^{n} l(\eta(x_i), y_i)$ sobre $\{\eta(x_1), \eta(x_2), ..., \eta(x_n)\}$. Sin embargo esto no es suficiente porque no obliga al estimador a ser suave, ya que por ejemplo en el modelo logístico $\widehat{\eta}(x_i) = \infty$ si $y_i = 1$ y es -∞ si $y_i = 0$ y las probabilidades estimadas son simplemente las y_i's observadas.

Una solución al problema, es elegir $\widehat{\eta}(\cdot)$ que maximice la esperanza de la log-verosimilitud, es decir

$$E(l(\widehat{\eta}(x), Y)) = \max_{\eta} E(l(\eta(x), Y)) \tag{2.2}$$

donde la esperanza estará dada sobre la distribución conjunta de X y Y (Hastie, 1984). Es decir se está eligiendo el modelo que maximice la log-verosimilitud entre todas las posibles observaciones futuras.

En el caso de la familia exponencial de dispersión, la solución se puede encontrar usando la distancia de Kullback-Leibler como la generalización del error cuadrático.

Esta mide la distancia entre densidades. La distancia entre un modelo con el verdadero parámetro η^* y uno con un parámetro η se define como:

$$k(\eta^*, \eta) = E_{\eta^*}[\log \frac{h(Y, \eta^*)}{h(Y, \eta)}].$$

Esta equivalencia se considera como una medida de distancia entre los dos parámetros η^* y η, o aún la media asociada μ^* y μ. La siguiente descomposición para el error cuadrático y para la distancia de Kullback-Leibler se puede derivar fácilmente.

$$E[(Y - \mu(x))^2] = E[(Y - \mu^*(X))^2] + E[(\mu^*(X) - \mu(X))^2] \quad (2.3)$$

$$EK[Y - \mu(x)] = EK[Y - \mu^*(X)] + EK[\mu^*(X) - \mu(X)] \quad (2.4)$$

donde $\mu(x)$ es la verdadera media condicional. De las expresiones anteriores se puede determinar que el máximo se encuentra cuando $\mu^*(X) = \mu(X)$, siempre y cuando no existan restricciones sobre el valor de la media. En el caso de la distribución Normal, la distancia de Kullback-Leibler es la raíz cuadrada del error cuadrático medio. Debido a que $EK[Y - \mu(X)] = E[l(Y, Y)] - E[l(Y, \mu)]$, por lo que se tiene que ésto es equivalente a maximizar la esperanza de la log-verosimilitud, resultado que coincide con el caso general.

La esperanza de la log-verosimilitud también ha sido utilizada por Brillinger 1977 y Owen 1983.

Derivación de las técnicas de estimación via la esperanza de la log-verosimilitud. Una manera simple de estimar $\eta(\cdot)$ sería suponer una forma simple para $\eta(x)$, igual a $\eta(x) = \beta_0 + \beta_1 x_1$. Entonces se estaría encontrando la función lineal $\eta(x)$ más cercana, en el sentido de la distancia de Kullback-Leibler a $\eta^*(x)$. La esperanza en la ecuación (2.2) podrá ser remplazada por su análoga muestral y la expresión resultante maximizada sobre β_0 y β_1. Esto no es más que la estimación de máxima verosimilitud estándar.

Consideremos ahora que no se quiere suponer una forma paramétrica para $\eta(x)$. Derivando la ecuación (2.2) con respecto a η se tiene

$$E[\frac{\partial l}{\partial \eta}|x]_{\widehat{\eta}(x)} = 0,$$

suponiendo que la esperanza y la derivada se puede intercambiar (condiciones de regularidad) y dado un estimador inicial $\eta(x)$, la primer expansión de la serie de Taylor sobre $\eta(x)$ da el estimador mejorado

$$\eta^1(x) = \eta(x) - \frac{E[\frac{\partial l}{\partial \eta}|x]}{E[\frac{\partial^2 l}{\partial \eta^2}|x]} \tag{2.5}$$

incluyendo a $\eta(x)$ en el operador de esperanza condicional se tiene:

$$\eta^1(x) = E\left[\eta(x) - \frac{\frac{\partial l}{\partial \eta}}{E[\frac{\partial^2 l}{\partial \eta^2}|x)}\middle| x\right]. \tag{2.6}$$

Esto da una manera para estimar $\eta(\cdot)$ en la práctica. Dado un estimador inicial $\eta(x)$ y remplazando la esperanza condicional por un estimador, en particular un suavizador de dispersión, de la ecuación (2.6) se obtiene un nuevo estimador. De esta manera al tener una muestra el algoritmo resulta ser

$$\eta^1(x) = \text{suavizador}\left[\eta(x) - \frac{\frac{\partial l}{\partial \eta}}{\text{suavizador}(\frac{\partial^2 l}{\partial \eta^2}|x)}\right]. \tag{2.7}$$

Debido a que la varianza de cada uno de los términos en los corchetes es aproximadamente $E(\frac{\partial^2 l}{\partial \eta^2})$, el suavizador podría usar pesos tales como suavizador$(\frac{\partial^2 l}{\partial \eta^2}|x)^{-1}$ para una estimación eficiente.

El algoritmo de datos consiste en repetidas iteraciones de la ecuación (2.7) parando cuando el cambio en la devianza sea muy pequeño.

En el caso de la familia exponencial es posible simplificar la ecuación (2.6) de la siguiente manera:

$$\begin{array}{lcl} \frac{\partial l}{\partial \eta} & = & (y-\mu)V^{-1}\left(\frac{\partial \mu}{\partial \eta}\right) \\ \frac{\partial^2 l}{\partial \eta^2} & = & (y-\mu)\frac{\partial}{\partial \eta}\left[V^{-1}\left(\frac{\partial \mu}{\partial \eta}\right)\right] - \left(\frac{\partial \mu}{\partial \eta}\right)^2 V^{-1} \quad \text{y} \\ E\left[\frac{\partial^2 l}{\partial \eta^2}\right] & = & \left(\frac{\partial \mu}{\partial \eta}\right)^2 V^{-1} \end{array}$$

con V la varianza de Y en $\mu = \widehat{\mu}$, entonces

$$\eta^1(x) = E\left[\eta(x) + (Y - \mu)(\frac{\partial \eta}{\partial \mu})\middle| x\right].$$

Haciendo la sustitución por el mismo estimador que en la ecuación (2.7), el cual es un suavizador de dispersión se tiene

$$\eta^1(x) = \text{suavizador}\left[\eta(x) + (Y - \mu)(\frac{\partial \eta}{\partial \mu})\right]$$

con pesos $(\frac{\partial \mu}{\partial \eta})^2 V^{-1}$.

2.3.2 Estimación del modelo en el caso múltiple.

Modelo aditivo Antes de presentar la estimación del modelo para el caso múltiple, se introducirá el modelo aditivo brevemente y el algoritmo que se utiliza para la estimación de sus parámetros, debido a que este algoritmo es una parte fundamental en el algoritmo que se utiliza para la estimación de los parámetros en el modelo aditivo generalizado, conocido como algoritmo de ajuste.

Si se tienen p covariables representadas por el vector $X = (X_1, X_2, ..., X_p)$, un modelo general que especifica $E(Y \mid X = x) = \mu$ y $g(\mu) = \eta(x)$ donde η es una función de p variables es

$$Y = \eta(X) + \varepsilon$$

donde $\eta(x) = E(Y \mid X = x)$, $Var(Y \mid x) = \sigma^2$ y los errrores ε son independientes de X. El objetivo es estimar $\eta(x)$. Si se usa el criterio de mínimos cuadrados, $E(Y - \eta(X))^2$, el mejor estimador para $\eta(x)$ es $E(Y \mid X = x)$. En el caso de una sola covariable la $E(Y \mid X = x)$ es estimado por un suavizador de dispersión. El cual en su forma más cruda es el promedio de aquellos y's en la muestra para los cuales x_i es cercano a x.

Se podría pensar en hacer algo similar para el caso multivariado, es decir, promediar las y_i para las cuales x_i es cercano a x. Sin embargo los suavizadores no funcionan bien en grandes dimensiones.

Como se ha mencionado, la varianza de un estimador depende del número de puntos en la vecindad. Sin embargo, el encontrar vecindades implica un mayor esfuerzo, además que el estimador ya no es local y puede resultar ser severamente sesgado. Esta es la motivación principal para el modelo aditivo $\eta(x) = f_o + \sum f_j(x_j)$.

Cada función se estima por suavizamiento en una sola coordenada; se pueden incluir puntos suficientes en la vecindad para conservar una varianza pequeña de los estimadores y, aún más permanecer local en cada coordenada. Por supuesto, el mismo modelo aditivo puede ser un estimador sesgado de la verdadera superficie de regresión, pero este sesgo es mucho menor que el que se produce por suavizadores en grandes dimensiones. El modelo aditivo es una generalización del modelo lineal estándar y esto permite la interpretación fácil de la contribución de cada variable.

El modelo aditivo sustituye la esperanza condicional de Y expresada de la siguiente manera

$$E(Y|X_1, X_2, ..., X_p) = \beta_0 + \beta_1 X_1 +, ..., +\beta_p X_p \qquad \text{(modelo de regresión usual)}$$

por

$$E(Y|X_1, X_2, ..., X_p) = f_0 + \sum_{j=1}^{p} f_j(x_j) \qquad \text{(modelo aditivo)} \tag{2.8}$$

donde las $f_j(\cdot)'s$ son funciones suaves estandarizadas tales que $E(f_j(X_j)) = 0$ para todo $j = 1, ..., p$.

Para motivar el algoritmo de estimación suponga que se tiene el modelo $Y = f_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$ y que se conoce el valor de $f_1(\cdot), ..., f_{j-1}(\cdot), f_{j+1}(\cdot), ..., f_p(\cdot)$. Si se definen los residuos parciales como

$$R_j = Y - (f_0 + \sum_{k \neq j} f_k(X_k)),$$

entonces $E(R_j|X_j) = f_j(X_j)$ y minimiza $E(Y - f_0 - \sum_{k=1}^{p} f_k(X_k))^2$. Por supuesto, no se conoce el valor de las $f_k(\cdot)'s$, pero esto da una forma de estimación para cada $\widehat{f}_j(\cdot)$ dados los estimadores $\{\widehat{f}_i(\cdot), i \neq j\}$. El procedimiento iterativo que resulta se le conoce como algoritmo de ajuste (Friedman and Suetzle, 1981) y queda expresado como:

Algoritmo de ajuste

Inicialización: $f_0 = E(Y), \qquad f_1^1(\cdot) = f_2^1(\cdot) = ... = f_p^1(\cdot) = 0, \qquad m = 0$

Iteracción: $m = m + 1$

hacer para $\quad j = 1 \quad$ hasta $\quad p$

$$R_j = Y - f_0 - \sum_{k=1}^{j-1} f_k^m(X_k) - \sum_{k=j+1}^{p} f_k^{m-1}(X_k)$$

$$f_j^m(X_j) = E(R_j \mid X_j)$$

parar hasta que:

$$RSS = E(Y - f_0 - \sum_{j=1}^{p} f_j^m(X_j))^2 \qquad \text{comience a decrecer o ya no decrezca.}$$

$f_j^m(\cdot)$ denota el estimador de $f_j(\cdot)$ en la m-ésima interacción. Note que por centrar Y al inicio, se garantiza que $E(f_j^m(X_j)) = 0$ en cada etapa. Es claro que RSS no se incrementa en otro paso del algoritmo y por ello converge. Para muestras finitas, remplazamos la esperanza condicional en el algoritmo de ajuste por sus estimadores, los suavizadores de dispersión. Breiman y Friedman probaron que:

- Para una clase restrictiva de suavizadores, el algoritmo converge
- Para una clase menos restringida, el procedimiento es consistente en media cuadrática.

En regresión múltiple es necesario preocuparse por la colinealidad de las covariables cuando se interpretan los coeficientes de regresión; sin embargo, la "concurvidad" (concurvity), término para denotar la colinealidad en los modelos aditivos, ha tenido peores implicaciones cuando se trata de interpretar las funciones individuales en el modelo aditivo (Hastie, 1990).

Si el propósito del análisis es predecir. Estos problemas son poco importantes.

Ajuste de un modelo aditivo generalizado Por lo dicho en la sección anterior la atención se ve restringida a un modelo aditivo, es decir, a un modelo de la forma

$$\eta(x) = f_0 + \sum_{j=1}^{p} f_j(X_j). \tag{2.9}$$

Si $Z = \eta(x) + (Y - \mu)(\frac{\partial \eta}{\partial \mu})$ en la ecuación (2.6) entonces se tiene un modelo de la forma

$$\eta^1(x) = E[Z|x] \tag{2.10}$$

que es un miembro del modelo aditivo, donde Z juega el papel de Y, así, para estimar los $f_j(\cdot)'s$ se ajusta un modelo de regresión aditiva para Z, tratando a esta como la variable respuesta Y en (2.8). Esta es la motivación principal para el algoritmo general de puntajes locales que para el caso exponencial es el siguiente

Algoritmo general de ajuste (puntajes locales)

1) Inicializar $\alpha = g(\sum_{i=1}^{n} \frac{y_i}{n})$; $f_1^o = ... = f_p^o = 0, m = 0$.
$m \leftarrow m + 1$

2) Construir una variable dependiente

$$z_i = \eta_i^{m-1} + (y_i - \mu_i^{m-1}) \left(\frac{\partial \eta_i}{\partial \mu_i^{m-1}} \right) \quad i = 1, .., n$$

con $\eta_i^{m-1} = \alpha^o + \sum_{j=1}^{p} f_j^{m-1}(x_{ij})$ y $\mu_i^{m-1} = g^{-1}(\eta_i^{m-1})$.

3) Construir los pesos

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i^{m-1}} \right)^2 (V_i)^{-1}$$

donde V es la varianza de Y.

4) Ajustar un modelo aditivo de pesos ponderados en z_i usando el algoritmo de ajuste con pesos w_i, para obtener la función estimada f_j^m, el predictor aditivo η^m, y el valor ajustado μ_1^m.

Calcular el criterio de convergencia

$$\Delta(\eta^1, \eta^o) = \frac{\sum_{j=1}^{p} ||f_j^1 - f_j^o||}{\sum_{j=1}^{p} ||f_j^o||}.$$

Un candidato natural para $||f||$ es $||s||$ que es la longitud del vector de evaluaciones de f en los n puntos muestrales.

5) Repetir el paso 2) remplazando η^o por η^1 hasta que $\Delta(\eta^1, \eta^o)$ sea menor que alguna cota establecida.

El algoritmo anterior está representado para el caso más simple del modelo compuesto para una suma de términos univariados. Sin embargo, se puede extender para el caso general. Cualquier método de regresión multiple puede usarse en el paso de regresión, tal como superficies suavizadas, suavizadores, análisis multifactorial de descomposición de varianza o todos aquellos como componentes del modelo aditivo.

Si la función liga es la identidad entonces $z_i = y_i$ y el procedimiento es simplemente un ajuste aditivo de reasignación de pesos de y_i sobre x_i donde los pesos cambian. El procedimiento de puntajes locales descrito aquí se aplica tanto a modelos de la familia de dispersión exponencial como a otros modelos que no pertenecen a esta familia.

Las funciones suaves producidas por el procedimiento de puntajes locales se pueden usar como una descripción de los datos, para predicción o para sugerir transformaciones de las covariables.

El procedimiento de puntajes locales es similar a otros métodos para estimar los modelos aditivos generalizados, por ejemplo, el de estimación por verosimilitud local (Hastie, 1984 1984). La ventaja del método de puntajes locales sobre los otros métodos, es que es considerablemente más rápido.

Otro algoritmo de estimación de los parámetros es el procedimiento de verosimilitud local que puede verse como un método empírico de maximización de $E[l(\eta(x), Y)]$. En lugar de derivar esta expresión, la escribimos de la siguiente manera $E[l(\eta(x), Y)] = E(E[l(\eta(x), Y)]X = x)$. Por lo que es suficiente maximizar $E[l(\eta(x), Y)|X = x]$ para cada x. La receta correspondiente se encuentra de la siguiente forma; considere la estimación de $\eta(x)$ en algún punto $x = x_i$, entonces un estimador de $E[l(\eta(x), Y)|X = x_i]$ es

$$E[l(\eta(x), Y)|X = x_i] = \frac{1}{k_n} \sum_{j \in N_i} l(\eta(x_j), y_j),$$

donde k_n es el número de datos en N_i .

Los algoritmos descritos aquí pueden usarse en cualquier modelo de regresión basados en una verosimilitud. Como un punto técnico, note que se está ligando el predictor aditivo $\eta = \sum_1^p s_i(X_i)$ a la distribución de Y vía $\eta = g(\mu)$. En algún modelo que no pertenezca a la familia exponencial, μ puede ser una función complicada del parámetro del modelo, o bien, puede no existir. Entonces sería deseable ligar η a algún otro parámetro de la distribución. Esto ocurre en particular en el modelo de Cox, que no se describe en este trabajo, pero es posible encontrar una amplia información en (Hastie,1990).

2.3.3 Algunos ejemplos

El modelo gaussiano Para este modelo, $\eta = \mu$ entonces la ecuación (2.7) se simplifica a $\eta^1(x)$ =suavizador[y], el algoritmo de puntajes locales se reduce a un suavizador de línea móvil de y sobre x.

El procedimiento de verosimilitud local también arroja este resultado ya que el estimador máximo verosimil local es $\widehat{\eta}(x_i) = \widehat{\beta}_{0i} + \widehat{\beta}_{1i}x_i$, donde $\widehat{\beta}_{0i}$ y $\widehat{\beta}_{1i}$ son los estimadores de mínimos cuadrados para los puntos en la vecindad N_i.

El modelo logístico Un modelo de respuesta binomial supone que la proporción de sucesos Y es tal que $\eta(x)Y|x \sim Bin(n(x), p(x))$ donde $Bin(n(x), p(x))$ se refiere a la distribución Binomial con parámetros $n(x)$ y $p(x)$. La distribución Binomial es un miembro de la familia exponencial con liga canónica $g(p(x)) = \log \frac{p(x)}{1-p(x)} = \eta(x)$. En el modelo logístico lineal suponemos $\eta(x_i) = \beta_0 + \beta_1 x_i$ y los parámetros son estimados por máxima verosimilitud usando la técnica de puntajes de Fisher. La extensión suave de este modelo generaliza la relación liga $\log \frac{p(x)}{1-p(x)} = \eta(x)$. El paso de puntaje local es

$$\eta^1(x) = \text{suavizador}[\eta^0(x) + \frac{y - p^0(x)}{p^0(x)(1 - p^0(x))}]$$

con pesos $n(x)p^0(x)(1 - p^0(x))$ ya que $\frac{\partial \eta}{\partial \mu_0} = \frac{1}{p^0(x)(1-p^0(x))}$, y $\mu_0 = p^0(x)$.

2.3.4 Inferencia de los parámetros

Bandas del error estándar.

Cada paso del método de puntajes locales consiste de un ciclo de ajuste aplicado a la variable dependiente z, con pesos $A = (\frac{\partial \mu_i}{\partial \eta_i})^2 V_i^{-1}$ dados por la matriz de información estimada. Además, si R es el operador de pesos en el ajuste aditivo y $\widehat{u} = \frac{\partial l}{\partial \widehat{\eta}}$, entonces en convergencia

$$\widehat{\eta} = R(\widehat{\eta} + A^{-1}\widehat{u})$$

sea $z = R(\widehat{\eta} + A^{-1}\widehat{u})$ entonces

$$\hat{\eta} = Rz.$$

La idea ahora es aproximar a z por una cantidad asintóticamente equivalente z_0, suponiendo que el modelo es consistente. Para los cálculos, estas cantidades se aproximarán a la covarianza de z por la de z_0 . Calculando la expansión de primer orden de $\hat{u}$ sobre el verdadero valor de η_0 se obtiene que $z \approx z_0 = \eta_0 + A_0^{-1}u_0$ el cual tiene media η_0 y varianza $A_0^{-1}\phi \approx A^{-1}\phi$. Como el predictor aditivo ajustado es $\hat{\eta} = Rz$ donde z tiene covarianza asintótica $A_0^{-1}\phi$ y como R no es un operador lineal debido a su dependencia a $\hat{\mu}$ y a y a través de los pesos, entonces es necesario usar una versión asintótica de R_0. De esta forma se tiene

$$\begin{aligned} cov(\hat{\eta}) &\approx R_0 A_0^{-1} R_0^T \phi \\ &\approx RA^{-1}R^T\phi. \end{aligned}$$

Similarmente,

$$cov(\hat{f}_j) \approx R_j A^{-1} R_j^T \phi,$$

donde R_j es la matriz que produce $\hat{f}_j$ de z. En algunos modelos tales como el logístico, el parámetro de dispersión ϕ es conocido y es igual a uno; en otros modelos, puede ser estimado, por ejemplo, por la media del error de la devianza (Hastie and Tibshirani 1990).

Las condiciones usuales de regularidad son necesarias, incluyendo consistencia. Consistencia implícitamente requiere que la cantidad de suavizamiento decrezca en una tasa apropiada. Estos argumentos ayudan a mostrar que $\hat{\eta}$ tiene una distribución asintóticamente $N(\eta_0, R_0 A_0^{-1} R_0^T \phi)$. Recientemente Gu (1989) desarrolló un argumento Bayesiano para aproximar la distribución posterior de η, para una situación en particular que él consideró, en la cual usa suavizadores splines, y obtuvo $\zeta(\eta|y) \approx N(\hat{\eta}, RA^{-1}\phi)$.

Grados de libertad Para generalizar los resultados para el modelo aditivo cuyos errores tienen una distribución Normal, es conveniente usar la aproximación asintótica de la devianza

$$\begin{aligned} D(y,\mu) &\approx (y-\widehat{\mu})^T A^{-1}(y-\widehat{\mu}) \\ &\approx (z-\widehat{\eta})^T A(z-\widehat{\eta}). \end{aligned}$$

Aplicando la misma definición de grados de libertad del error a esta última expresión se tiene

$$df^{\text{err}} = n - tr(2R - R^T ARA^{-1}).$$

Si el modelo es insesgado entonces $E(D) \approx df^{err}\phi$. De más interés es la diferencia en la devianza entre dos modelos anidados. Suponga que $\widehat{\eta}_1$ y $\widehat{\eta}_2$ difieren por un sólo término, por ejemplo, un término no paramétrico de x_j. Si el modelo más pequeño, $\widehat{\eta}_1$, es correcto, entonces

$$\begin{aligned} ED(\widehat{\eta}_2;\widehat{\eta}_1)/\phi &= E\{D(y;\widehat{\eta}_1) - D(y;\widehat{\eta}_2)\}/\phi \\ &\approx tr(2R_1 - R_1^T A_1 R_1 A_1^{-1}) - tr(2R_2 - R_2^T A_2 R_2 A_2^{-1}) \\ &= df^{err}(\widehat{\eta}_1) - df^{err}(\widehat{\eta}_2) \\ &= df_j^{err}, \end{aligned}$$

debido a que el término del sesgo se cancela. Si el parámetro de dispersión ϕ es conocido, como en el caso de la Binomial o la Poisson entonces es posible aproximar la distribución asintótica de $D(\widehat{\eta}_2;\widehat{\eta}_1)$ por una distribución $\chi^2_{df_j^{err}}$. Cuando el parámetro de ruido es desconocido, una distribución F es más apropiada.

Otra manera es usar la distancia de Kullback-Leibler para medir los errores de predicción

$$PE = E\left\{\frac{1}{n}\sum_{i=1}^{n} D(Y_i^0;\widehat{\mu}_i)\right\} \tag{2.11}$$

donde Y_i^0 tiene la misma distribución que y_i, y $\widehat{\mu}_i = \widehat{\mu}(x^i)$ es el ajuste aditivo basado en y. El estadístico AIC se define como

$$AIC = D(y;\widehat{\mu})/n + 2df\phi/n$$

donde $df = tr(R)$ hace que AIC sea asintóticamente insesgado para el error de predicción en (2.11). Esta cantidad tiene la forma del criterio de información de Akaike, por ello el nombre de AIC.

Selección del parámetro de suavizamiento Existe un número de métodos posibles para la selección automática de los parámetros de suavizamiento en un modelo aditivo generalizado. Se discutirán algunos estudios aquí.

Suponga que el modelo usa suavizadores lineales $S_1, ..., S_p$ con correspondientes parámetros se suavizamiento $\lambda_1, ..., \lambda_p$. Sea $\widehat{\mu}_i^{-1}$ el valor ajustado en el $i-$ésimo punto, obtenido al dejar el $i-$ésimo punto fuera de la muestra. Entonces la devianza de la validación-cruzada se define por:

$$CV = \frac{1}{n} \sum_{i=1}^{n} D(y_i; \widehat{\mu}_i^{-1}).$$

En principio uno podría minimizar esta cantidad sobre $\lambda_1, .., \lambda_p$, pero esto sería computacionalmente caro, requiriendo n aplicaciones del procedimiento de puntajes locales por cada valor de $\lambda_1, .., \lambda_p$.

Como en el caso más simple hay un número de aproximaciones para validación cruzada que son mas confiables computacionalmente. Como antes, denotemos por R los pesos del ajuste-aditivo correspondiente a la última interacción del procedimiento de puntajes locales Se define una devianza generalizada de validación-cruzada

$$GCV = \frac{\frac{1}{n} \sum_{i=1}^{n} D(y_i; \widehat{\mu}_i^{-1})}{\{1 - tr(R)/n\}^2}.$$

Esto es posbile gracias a que en el último paso del paso de puntajes locales se puede considerar como un ajuste aditivo de pesos ponderados.

Estimación de la función liga En los modelos hasta ahora se ha supuesto que la función liga es conocida, es decir, la función entre la media de la respuesta y el predictor aditivo. En algunos casos para incrementar la flexibilidad de los modelos aditivos generalizados es de interés estimar la función liga a partir de los datos. Por ejemplo, la función logit presenta una cierta simetría; $g(\mu) = \log\{\mu/(1-\mu)\}$, $g(1-\mu) = -g(\mu)$. No es claro inicialmente cuánta simetría se puede suponer, y funciones ligas asimétricas tales como el complementario $\log - \log$, puede ser una buena alternatival.

Existen estudios que han sugerido funciones liga paramétricas, su estimación ha sido desarrollada por modelos lineales generalizados, y puede ser fácilmente incorpo-

rada a la estimación de un modelo aditivo generalizado. La estimación no-paramétrica de la función liga es posible a través de un procedimiento de Gauss-Newton llevando a una versión generalizada del procedimiento de puntajes locales. Se ha experimentado con esta idea y se han obtenido éxitos limitados (Hastie, 1984); el algoritmo presenta ciertas inestabilidades cuando la estimación de la función liga llega a ser muy plana o lisa (relajada). La regresión monótona de splines puede proveer una base más estable para la estimación.

2.3.5 Ajuste del modelo

Análisis de devianza

La devianza o el cociente de verosimilitud para un modelo ajustado $\widehat{\mu}$ esta definido por

$$D(y;\widehat{\mu}) = 2\{l(\mu_{\max};y) - l(\widehat{\mu};y)\}$$

donde $\mu_{\max}$ es el valor del parámetro que maximiza $l(\mu;y)$ para todo μ.

La devianza juega el papel de la suma de cuadrados de residuos para el modelo lineal generalizado, y puede usarse para juzgar el ajuste del modelo y para comparar modelos, en este caso la distribución de la devianza se sabe que tiene una distribución asintótica χ^2,

Para el caso de los modelos aditivos generalizados y modelos no paramétricos la devianza sigue teniendo sentido como un medio para juzgar el ajuste del modelo a través de la comparación de modelos. Sin embargo, la distribución de la devianza para los modelos aditivos generalizados aún no está justificada teóricamente y la distribución χ^2 se ha tomado sólo como referencia.

Hay otras formas de hacer inferencia sobre ajustes aditivos. El modelo aditivo no-paramétrico puede verse como un diagnóstico para identificar formas funcionales. Las funciones ajustadas pueden usarse para inspirar reparametrizaciones parsimoniosas de alguna variable, por ejemplo, el uso de logaritmos, inversas, términos polinómicos, etc. Otro medio es el uso del bootstrap para estimar la distribución de la devianza para una variable explicativa bajo la hipótesis nula de que no esté presente.

Capítulo 3

Aplicación

3.0.6 Introducción

En este capítulo se ejemplifica la teoría que se ha desarrollado, tanto para modelos lineales generalizados como para modelos aditivos generalizados (capítulo 1 y 2), con la ayuda de una base de datos observacionales de inasistencia a clases provenientes de una escuela del suroeste de la Ciudad de México y de los índices de contaminación de la estación de monitoreo ubicada en la misma escuela. La idea principal es mostrar al modelo aditivo generalizado como herramienta de diagnóstico para conocer el tipo de relación que existe entre la variable respuesta y las variables explicativas, para posteriormente usar esta información en el ajuste de un modelo lineal generalizado.

3.0.7 Presentación del problema y la base de datos

Los datos del apéndice 1 corresponden a una muestra, de tamaño 189, tomada en la escuela Alexander Bain ubicada al suroeste de la Ciudad de México. La población en estudio fueron niños de nivel preescolar de edad entre 3 y 5 años. El objetivo general es conocer si existe una relación entre las inasistencias que se presentaron en la muestra y los índices de partículas de contaminación $pm25$.

El objetivo particular del estudio es identificar los factores de riesgo que influyen en el ausentismo, y por tanto, la investigación natural es modelar el ausentismo como una función de los predictores. Las medidas fueron realizadas diariamente. Se tienen

un total de 189 observaciones con los registros en todas sus covariables completas. Los datos que se muestran en el apéndice 1 fueron proporcionados por la Dra. Silvia Ruíz quien fue un consultor en un proyecto del cual estos datos se tomaron.

Por qué del estudio. La información que se presenta en el apéndice 1, corresponde a la zona suroeste de la ciudad de México como se ha venido mencionando. Dentro del Colegio Alexander Bain se encuentra localizada una estación de monitoreo con domicilio en el Pedregal de San Angel,. Esta región de la Ciudad se considera como una de las menos críticas, pues con frecuencia presenta los índices más bajos con respecto a contaminación de partículas $pm25$. Y además se anexa la información de que los niños que asisten a esta escuela son de un estrato social alto, por lo que realmente esperaríamos que la influencia de la cantidad de partículas en enfermedades respiratorias sea casi nula, ya que de lo contrario se estaría en un estado crítico de afectación para la población en general.

Descripción de la base de datos. La respuesta, referida como y, es el número de niños (0,1,2,3,..) de nivel preescolar de la edad entre 3 y 5 años que no asistieron a clases por motivo de un padecimiento respiratorio. Los predictores disponibles son: la variable continua $pm254$ que corresponde a la cantidad promedio de partículas (ppm) que se observaron durante cuatro días antes del registro de la observación, temperatura ($tempmin$, de tipo continuo) medida en grados centígrados correspondiente a la temperatura mínima observada en el día de el estudio, época del año referida como *periodo* ($calor = 1$ que va desde septiembre a febrero y $frío = 0$ que va desde marzo hasta agosto, tipo categórico) y día de la semana en el que se tomó la muestra *(lunes=1, martes=2, miércoles=3, jueves=4, viernes=5, sábado=6, domingo=7,* de tipo categórica*)* referida como *diad* y finalmente actuando como offset el número total de niños que asistieron a clase el día anterior referido como $offset$.

3.0.8 Identificación del modelo a ajustar

El ajuste de un modelo lineal generalizado con respuesta Poisson tiene sentido en nuestro análisis debido a que se tiene un conjunto de covariables (*diad*, *pm*254, *tempmin*, *periodo*) que se creen pueden explicar a la variable respuesta (y), aunado a esto se tiene una variable respuesta de conteos en un periodo de tiempo en este caso es el día.

Se ajusta un modelo aditivo generalizado con el fin de tener información sobre el tipo de relación que existe entre las variables explicativas (*pm*254, *tempmin*, *periodo*, *diad*) y la variable respuesta y. Debido a que la variable respuesta es una variable discreta y referida a conteos se modela su comportamiento a través de una variable Poisson, así el modelo aditivo generalizado que se ajusta es con una liga logarítmica, recordando que un factor de corrección será incluído en este caso como offset (capítulo 1).

Al ajustar un modelo aditivo generalizado se utilizó como suavizadores a s = suavizador de spline y lo = suavizador de regresión local, ambos con diferentes parámetros de suavizamiento (ver gráfica 1-4), eligiendo finalmente a lo con un span de un medio debido a que gráficamente se podía observar mejor los puntos de cambio de la variable en cuestión (*pm*254). Los ajustes que se presentan en este capítulo se realizaron en Splus y los comandos usados se muestran en el apéndice 2.

$$gam(formula = y \sim lo(tempmin, span = 1/2) + diad + periodo + lo(pm254, span = 1/2) + offset(log(ofset)), family = poisson) \tag{3.1}$$

Residuos de la devianza:

Min	1Q	Mediana	3Q	Max
-1.680259	-0.7502419	-0.4134448	-.001913014	3.56031

Devianza nula: 226.1341 on 188 grados de libertad

Devianza residual: 134.6841 on 173.0721 grados de libertad

	gl	gl Npar	Chisq	P(chi)
(Intercept)	1			
lo(tempmin,span=1/2)	1	2.7	0.71484	0.8338343
periodo	6			
diad	1			
lo(pm254,span=1/2)	1	3.2	18.09717	.0005272

Tabla 3.1

De esta salida (tabla 3.1) se puede observar, en primer lugar, cómo la inclusión de las variables explicativas *tempmin*, *periodo*, *diad*, *pm*254 mejoran el ajuste ya que se observa una disminución de la devianza con respecto al modelo nulo el cual explica a la respuesta con la media de la propia variable respuesta observada (y).

Continuando con el análisis de la tabla 3.1, se observa una columna con el nombre de *P(chi)*, que es un tipo de prueba para determinar el efecto de cada una de las funciones nolineales; esta prueba consiste en que para cada término no-paramétrico en este caso *tempmin* y *pm*254, el componente nolineal se iguala a cero y su parte paramétrica (lineal) se incluye en el modelo, finalmente se vuelve a ajustar el modelo por mínimos cuadrados ponderados, tomando a los otros componentes como fijos. De esta forma el cambio en la estadística χ^2 de Pearson se registra para cada término que se borra. La información que se pudo obtener proviene de Chambers, 1992. Así, usando esta estadística se puede decir que la variable *tempmin* entra al modelo de forma lineal y la variable *pm*254 con una alta significancia debe de ser incluída en el modelo de forma nolineal.

Para corroborar lo anterior, se realizan las gráficas de variables explicativas vs residuos parciales, de las cuales podemos observar que el comportamiento de la variable explicativa *tempmin* dentro del modelo es lineal (gráfica 4) y el comportamiento de la variable explicativa *pm*254 es no-lineal (gráfica 3), además gracias a estas gráficas es posible sugerir el tipo de reparametrización que se puede realizar para posteriormente ajustar un modelo lineal generalizado.

Como se mencionó en el párrafo anterior, a través de estas gráficas es posible dar una reparametrización de las variables, así que, regresándonos a la gráfica 3 se intentará proponer una reparametrización.

Analizando la gráfica (3) podemos observar un comportamiento lineal casi constante al principio hasta el punto 27 (aproximadamente) para posteriormente tener un incremento lineal hasta el punto 31 (aproximadamente) y un descenso lineal también hasta el punto 36.5 (aproximadamente) y finalmente un comportamiento lineal monótono aparentemente constante. Así, después de este análisis gráfico lo que se trata es de particionar el rango de la variable $pm254$ precisamente en los puntos donde se observan cambios en el comportamiento de la gráfica. De esta manera, debido a que el modelo aditivo generalizado en este caso sólo se está usando de una manera descriptiva, nos llevamos la reparametrización que se propone al ajuste de un modelo lineal generalizado.

Así la salida en Splus del modelo lineal generalizado con la reparametrización se ve como

$$glm(formula = y\tilde{}\ I(pm254 * cc1) + tempmin + diad + periodo + offset(log(ofset)), family = poisson) \tag{3.2}$$

Residuos de la Devianza:

Min	1Q	Mediana	3Q	Max
-1.835839	-0.7606424	-0.4050795	-0.007205255	3.574944

	valor	error estándar	t valor	p-valor
intercept	-9.0034	2.76717786	-3.2536485	0.0014
I(pm254*cc1)c1	.03970509	.07720520	0.5142800	0.6077
I(pm254*cc1)c2	0.07372882	0.06393920	1.1531083	0.2504
I(pm254*cc1)c3	0.04147918	0.05478827	0.7570813	0.4500
I(pm254*cc1)c4	0..04368905	0.04663350	0.9368596	0.3501
diad1	-1.55516005	0.51193746	-3.0377930	0.0027
diad2	0.22336933	0.20802805	1.0737462	0.2844
diad3	0.13064033	0.12822042	1.0188731	0.3097
diad4	0.17214446	0.08152646	2.1115165	0.0361
diad5	0.13035093	0.05746493	2.2683560	0.0245
diad6	-1.20885023	2.01002204	-0.6014114	0.5483
periodod	0.51083269	0.15704883	3.2526996	0.0014
tempmin	-0.03161610	0.05549717	-0.5696886	0.5696

Tabla 3.2

(El parámetro de dispersión se tomará igual a 1)

Devianza nula: 226.1341 on 188 grados de libertad
Devianza residual: 139.8934 on 176 grados de libertad
Número de iteracciones por el método de puntajes de Fisher: 8

Donde *cc*1 es una matriz de variables dummy creada para cortar el rango en donde se había observado que la variable *pm*254 va cambiando su comportamiento, es decir, para la primer columna de *cc*1, *c*1, se conforma de unos en aquellos renglones que corresponden a valores de la variable *pm*254 menores o iguales a 27 y cero en el resto, las demas columnas de *cc*1, *c*2, *c*3, *c*4, se formaron de una manera similar a c1, a diferencia de que los unos se colocaron en los intervalos $27 < pm254 \leq 31$, $31 < pm254 \leq 37$, $37 < pm254$, respectivamente.

De la salida correspondiente al ajuste del modelo 3.2 se puede ver que la devianza residual es mayor a la que se obtuvo con el modelo 3.1 esto se debe a que este modelo contempla la forma exacta del comportamiento de la variable *pm*254 a comparación del modelo 3.2 que sólo trata de dar una aproximación del comportamiento, en el caso de que se tuviera la forma funcional exacta de la variable, el resultado sería el mismo en cuanto a valor de las devianzas.

La tabla 3.2 incluye también los niveles de significancia para cada variable explicativa que se encuentra en el modelo, $p-$valor, de ahí podemos ver que la variable temperatura muestra un $p-$valor$=0.5696$ lo cual nos está indicando que no es importante para el modelo, con respecto a la variable *diad* se tiene que al menos el nivel 1 (*diad*1) presenta un p-valor significativo para el modelo, de esta forma (Hosmer, 1989) queda incluída dentro del modelo de interés la variable *diad*. La variable *pm*254 presenta el p-valor más pequeño en el segundo rango correspondiente a p-valor$=0.2504$ y el resto presentan p-valores más grandes, en estudios observaciones con variables de contaminación se considera una relación importante si se presenta un p-valor$=.20$, en nuestro caso se actuó de manera mas laxa y de esta forma se considera una relación importante para el modelo. Así, el modelo final es:

$$glm(formula = y^{\sim} I(pm254 * cc1) + diad + periodo + offset(log(ofset)), family = poisson) \tag{3.3}$$

Residuos de la devianza:

Min	1Q	Mediana	3Q	Max
-1.816145	-0.775422	-0.4065277	-0.007146371	3.508091

Coeficientes	Valor	Error estandar	t valor	p-valor
(Intercept)	-9.24928822	2.75259731	-3.3602039	0.0010
I(pm254*cc1)c1	0.03957183	0.07741124	0.5111898	0.6099
I(pm254*cc1)c2	0.07416093	0.06414785	1.1560938	0.2492
I(pm254*cc1)c3	0.04079573	0.05493857	0.7425698	0.4587
I(pm254*cc1)c4	0.04312598	0.04677600	0.9219681	0.3578
periodod	0.54939481	0.14271214	3.8496712	0.0002
diad1	-1.55452659	0.51185357	-3.0370533	0.0010
diad2	0.22082284	0.20794769	1.0619153	0.0028
diad3	0.13281781	0.12820376	1.0359900	0.2897
diad4	0.17744396	0.08104268	2.1895126	0.0299
diad5	0.13264968	0.05737224	2.3120884	0.0219
diad6	-1.21258770	2.02813817	-0.5978822	0.5507

Tabla 3.3

(Parámetro de Dispersión para la familia Poisson se toma igual 1)

Devianza nula: 226.1341 on 188 grados de libertad

Devianza residual: 140.219 on 177 grados de libertad

Número de iteracciones por el método de puntajes de Fisher: 8

Comparando la devianza de este último modelo con el modelo (3.2) y comparando con la respectiva χ^2 se comprueba que el cambio no es significativo por ello se puede decir que efectivamente la variable explicativa *tempmin* no es un factor importante en la modelación de ausentismo por contaminación.

Antes de realizar las inferencias respectivas realizemos un análisis de residuos para asegurarnos de un buen ajuste.

Para ello se realizarán diferentes gráficas de residuos las cuales se encuentran explicadas en el capítulo 1.

En primer instancia se construye la gráfica de residuos no estandarizados para el modelo final vs los residuos estandarizados del modelo:

$glm(temp\min \tilde{} I(pm254 * cc1) + diad + periodo)$ esta gráfica sirve para verificar que realmente la omision de la variable tempmin no juega un papel importante en el modelo como lo mostró el cambio en la devianza, y precisamente el resultado de tal aseveración se comprueba en la gráfica (3) donde se ve un comportamiento disperso, es decir no hay ningun patrón de comportamiento identificable.

Para la verificación de la liga se realiza también la gráfica (5) en donde se observa una línea recta, comportamiento que nos dice que la función liga fue bien elegida para el modelo en estudio.

De esta manera y debido a que el análisis de residuos no reflejó ningún tipo de problema en el modelo ajustado (3.3) continuamos con la interpretación de los parámetros.

Retomando los resultados de la tabla 3.3 se tiene lo siguiente.

Coeficientes	Valor	$\exp(\hat{\beta})$
I(pm254*cc1)c1	0.03957183	1.040365
I(pm254*cc1)c2	0.07416093	1.07698
I(pm254*cc1)c3	0.04079573	1.041639
I(pm254*cc1)c4	0.04312598	1.044069
periodod	0.54939481	1.732204
diad1	-1.55452659	0.2112894
diad2	0.22082284	1.247102
diad3	0.13281781	1.142042
diad4	0.17744396	1.194161
diad5	0.13264968	1.14185
diad6	-1.21258770	0.2974266

Tabla 3.4

De la salida anterior se tiene que el riesgo de inasistencia por enfermedad respiratoria en el día martes es mayor que el resto, es decir hay un incremento de ausentismo en el día martes del 24% con respecto al domingo, esto es coherente debido a que en muchas ocasiones la actividad en la ciudad de México en los primeros días de la semana se ve incrementada. Con respecto a los niveles de partículas suspendidas, se

observa que el riesgo es mayor en el segundo rango que en el resto de los rangos de la variable $pm254$ este comportamiento ya se observaba en la gráfica que se realizó precisamente para identificar el tipo de reparametrización que era necesaria. De esta manera, se puede decir que después de una cierta cantidad de partículas suspendidas en el aire que afectan a la inasistencia, en este caso el primer nivel, es necesario una mayor cantidad de partículas suspendidas, que fue el segundo y último nivel, para que afecte de nueva cuenta a la inasistencia y en nuestro caso a la salud de los niños. Estas conclusiones se han realizado sin tomar encuenta que se está hablando de intervalos de tiempo, ahora considerando esto y fijando un cambio en 5 unidades en cada intervalos tendríamos que:

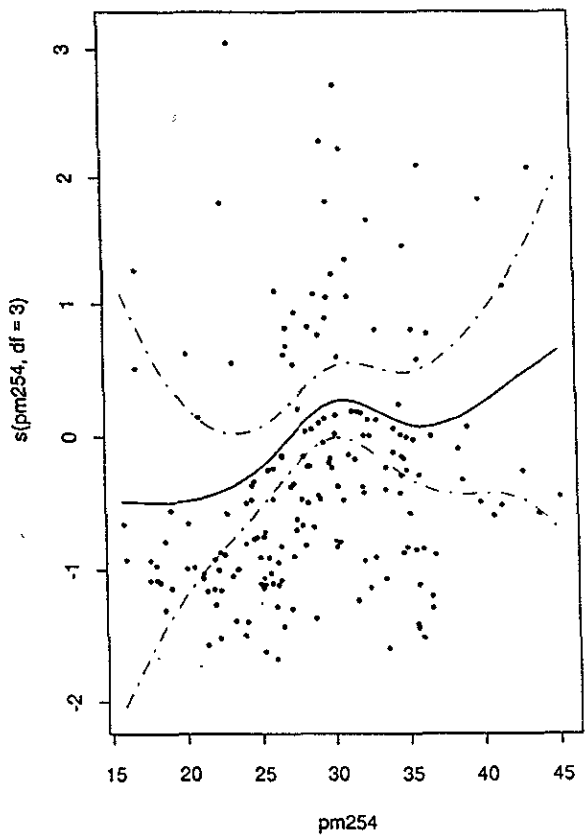
(pm254*cc1)c1	(pm254*cc1)c2	(pm254*cc1)c3	(pm254*cc1)c4
1.218791	1.4489	1.226272	1.240643

Tabla 3.5

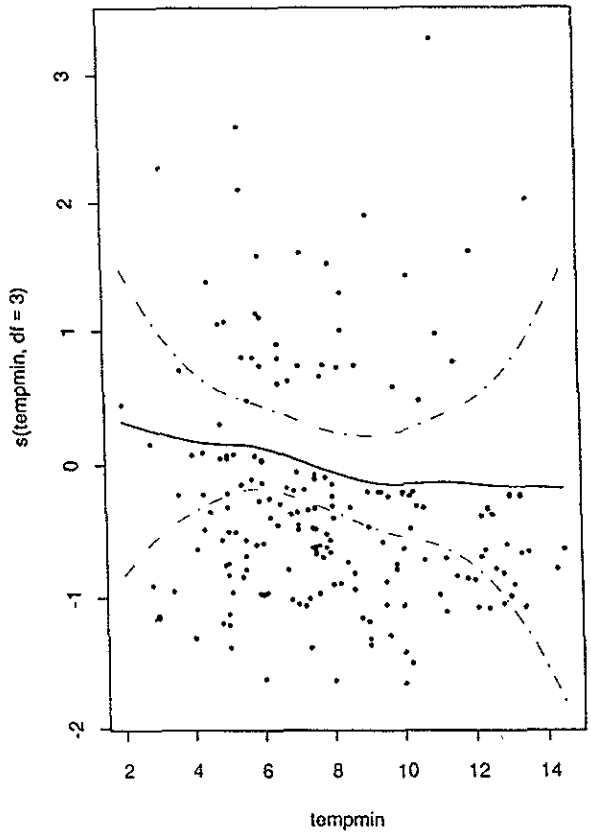
lo que nos estaría diciendo es que: si nos movemos de un punto fijo del segundo intervalo, de la variable $pm254$, 5 unidades el riesgo presenta un incremento de ausentismo de un 44% en ese intervalo.

Con respecto al periodo del año, en el que se encuentra la población, como bien se sabe, afecta definitivamente al ausentismo, ya que en temporada de frío las enfermedades respiratorias se incrementan, por lo que este último resultado tiene coherencia con estudios anteriores.

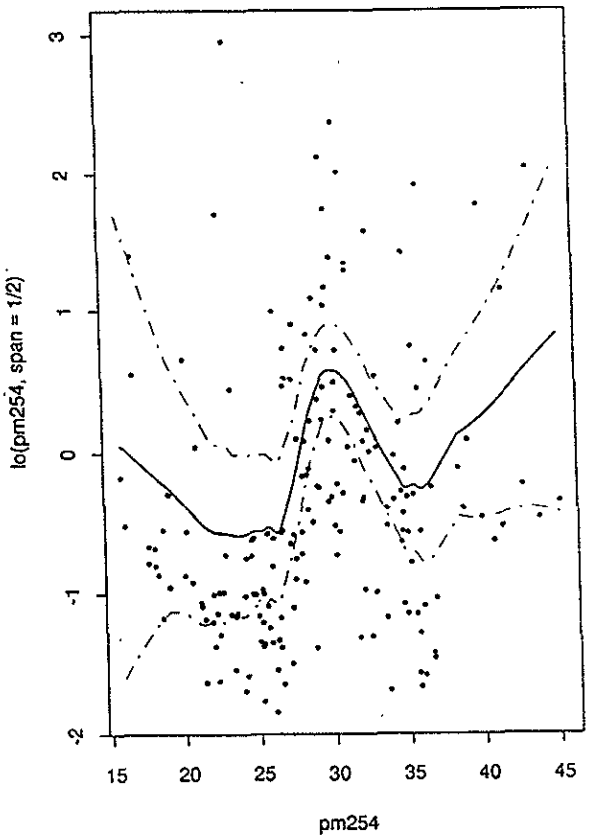
GRAFICA No.1



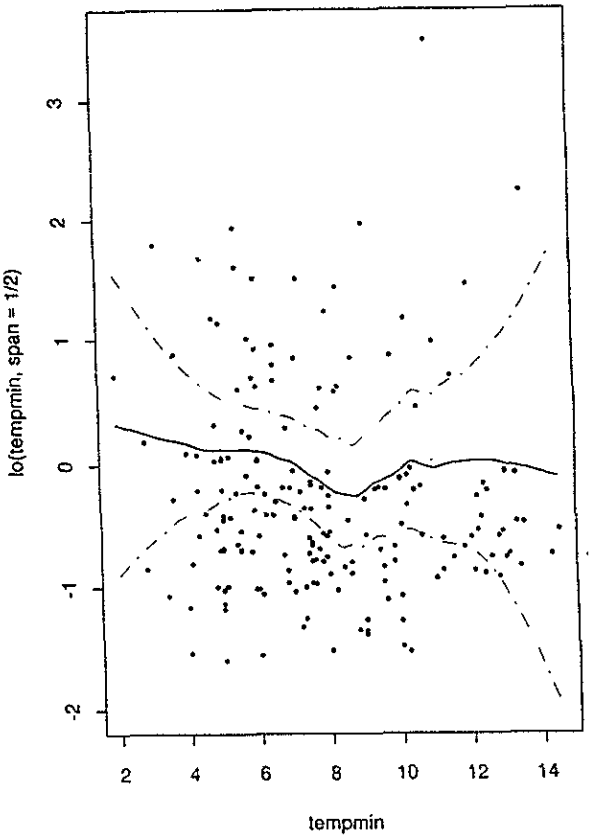
GRAFICA No.2



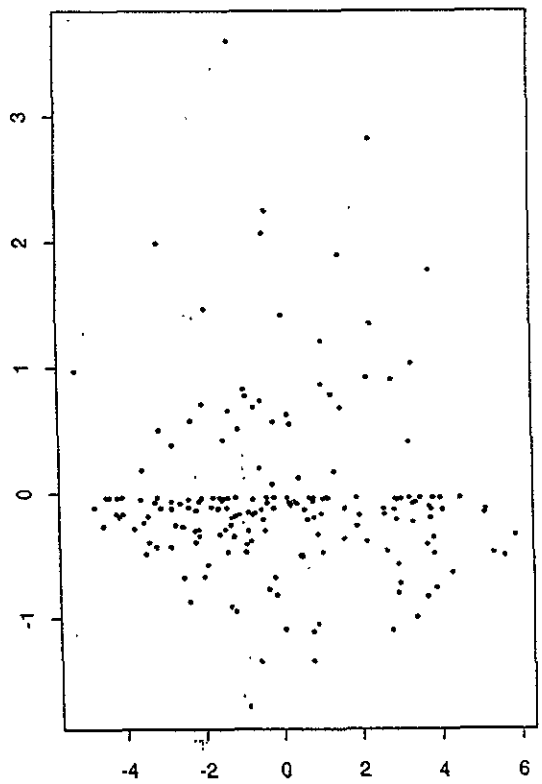
GRAFICA No.3



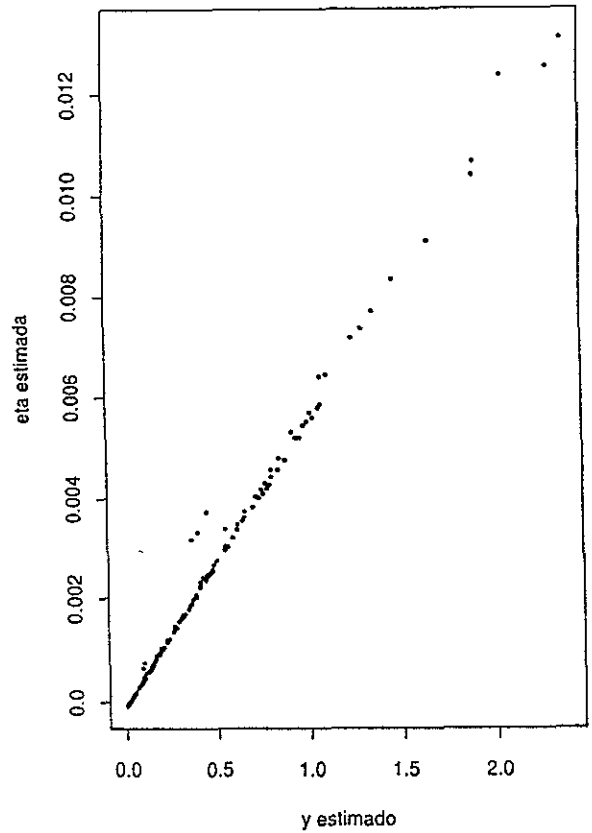
GRAFICA No.4



GRAFICA No.5



GRAFICA No.6



Capítulo 4

Conclusiones

En este trabajo se ha desarrollado el modelo lineal generalizado junto con el modelo aditivo generalizado, se dio una breve explicación de sus componentes y de la teoría que existe alrededor de ellos.

Se pudo observar dentro del capítulo uno, en el que sólo se habló de los GLM, como el algoritmo de ajuste, puntajes de Fisher, resulta ser el mismo para cualquier GLM, sin importar la elección de la distribución para el componente aleatorio o la elección de la función liga. Existen varios paquetes computacionales que pueden realizar el ajuste tanto de los modelos aditivos generalizados como de los GLM, sólo por citar algunos, tenemos: Splus, Glim, SAS.

El enfoque que se le dio en este tabajo a los modelos aditivos generalizados, fue en un sentido exploratorio, es decir, se usó como una herramienta para sugerir transformaciones paramétricas sobre la relación funcional entre la variable de respuesta y las covariables contenidas en el modelo. Una vez que se encontraron las transformaciones, se procedió a realizar un ajuste usando el GLM con la reparametrización sugerida

Este modelo no se puede utilizar para hacer inferencias, porque no proporciona el valor del coeficiente estimado asociado a una covariable, por tal motivo, no podemos evaluar el impacto de dicha covariable sobre la respuesta. No obstante, el modelo permite hacer predicciones, como se mencionó en el capítulo 2; pero en esta aplicación no era el objetivo hacer predicciones, sino evaluar el impacto de las variables

ambientales sobre el ausentismo de los niños.

Podemos decir que el suavizamiento no paramétrico puede ser de utilidad en epidemiología del medio ambiente, ya que en los estudios propios de esta área, en general, no existe una relación lineal entre el riesgo y la respuesta ni tampoco una transformación que sea fácil de identificar. En este sentido, los modelos aditivos generalizados son una herramienta útil para evidenciar qué tipo de relaciones se presentan entre los factores de riesgo(covariables) y la respuesta.

En general, podemos decir que la poca teoría que existe de los modelos aditivos generalizados, como lo es existencia, unicidad y convergencia del modelo, estimación de los grados de libertad de los términos del modelo, y estimación de curvas de desviación estándar para las funciones ajustadas, impiden la utilización del modelo en su totalidad, sin embargo, esta teoría se sigue desarrollando con una gran cantidad de preguntas que restan por resolverse. Una, por ejemplo, es el efecto de la dependencia de las covariables sobre el algoritmo de ajuste, desviaciones estándar y grados de libertad del ajuste. Un progreso parcial ha sido desarrollado en esta dirección pero aún resta mucho por hacer.

Así se puede concluir lo siguiente:

Gracias al modelo aditivo generalizado que se ajustó se pudo observar que el efecto de la variable $pm254$ no era lineal en el modelo de interés, pero de mayor importancia fue que sugirió los puntos de corte para posteriormente categorizar la variable, de esta manera evitó que se categorizará en forma arbitraria y sin tomar en cuenta el real comportamiento de la variable.

Al ajustar el modelo lineal generalizado posteriormente a la categorización realizada, se observó que existía un efecto diferenciado de $pm254$ en las ausencias de los escolares por enfermedades respiratorias para cada categoría de misma. Con esto es posible decir que, las partículas suspendidas son un factor importante en las enfermedades respiratorias y por ende en en el ausentismo de los niños de la Escuela Alexander Bein. Sin embargo, la variable de mayor impacto resultó ser el periodo (calor o frío) que se presentaba en el momento de estudio, de esto podemos decir que posiblemente el ausentismo que se observó en la muestra estaba más ligado al cambio de temperatura que a la posible afectación que pudieran representar las partículas

suspendidas en la salud de los niños.

Desafortunadamente los datos con los que se trabajó no permitieron mostrar qué tan valioso resulta ser el modelo aditivo generalizado, ya que sólo una variable resultó ser no-lineal.

Finalmente es de importancia mencionar que el presente trabajo no es una revisión exhaustiva tanto para los modelos lineales generalizados como para los modelos aditivos generalizados, debido a que temas como por ejemplo: sobredispersión para los MLG y el uso de los modelos aditivos generalizados en el modelo de Cox no son tratados en el presente trabajo.

Apéndice 1

Base de datos utilizada en la aplicación del capítulo 4

ofset	y	pm254	periodo	diad	tempmin
181	0	19.0925	0	7	13.3
181	0	22.3375	0	1	13.1
180	0	24.52	0	2	10.6
180	0	25.3225	0	3	13.3
179	0	24.5975	0	4	13.5
178	3	23.360001	0	5	11
170	0	24.985001	0	6	12.8
170	0	24.68	0	7	13
170	2	22.700001	0	1	13.6
173	0	20.174999	0	2	12.4
173	1	16.8325	0	3	12
176	0	17.637501	0	4	12.9
178	0	21.122499	0	5	7.5
178	0	23.1175	0	6	7.7
178	0	26.567499	0	7	10.3
178	0	27.275	0	1	13.4
181	0	24.172501	0	2	12.4
181	0	22.629999	0	3	7.8
180	0	20.535	0	4	6.2
178	0	18.059999	0	5	7.8
182	0	25.584999	0	6	8.6
182	0	31.737499	0	7	13
182	0	34.669998	0	1	11.2

ofset	y	pm254	periodo	diad	tempmin
177	0	38.406666	0	2	12.5
177	0	32.043331	0	3	14.3
179	0	22.333332	0	4	14.5
179	0	21.184999	0	5	11.2
180	0	18.283333	0	6	9.8
182	0	19.063334	0	7	10
182	0	30.366667	0	3	12.6
182	0	42.73	0	4	11.5
182	0	41.282501	0	5	11
182	0	40.735001	0	6	12.8
182	0	33.0075	0	7	13.3
182	1	20.200001	0	1	11
178	0	15.88	0	2	10.4
178	0	16.0375	0	3	9.4
181	0	24.0725	0	4	10
179	0	27.465	0	5	11.8
182	1	28.9	0	6	11.5
181	0	31.967501	0	7	10
181	0	25.2225	0	1	8.6
177	0	22.82	0	2	8
177	0	21.959999	0	3	7
180	0	18.033333	0	4	7.4
176	0	17.620001	0	5	7.6
177	0	26.443333	0	6	12
175	0	25.620001	0	7	7.5
175	0	30.129999	0	1	8.8
160	0	32.122501	0	2	8.5
160	0	27.922501	0	3	12.2
160	0	29.032499	0	4	7.4
181	0	27.9025	0	5	9.5
181	1	31.174999	0	6	8.2
180	0	36.630001	0	7	9.3
180	1	35.452499	0	1	9.8
180	0	30.182501	0	2	12.2

ofset	y	pm254	periodo	diad	tempmin
180	0	25.23	0	3	7.9
175	0	20.065001	0	4	4.5
179	0	19.0375	0	5	4.3
179	0	21.8825	0	6	4.8
182	0	26.605	0	7	9.4
182	0	30.4175	0	1	9
182	0	34.092499	0	2	6.9
182	1	35	0	3	8.3
181	0	32.1325	0	4	8.4
181	1	29.7875	0	5	6.5
180	1	29.6875	0	6	5.5
180	0	28.209999	0	7	4.3
180	1	30.4375	0	1	5.6
180	0	29.41	0	2	6
180	0	27.525	0	3	5
181	0	29.172501	0	4	7.5
181	0	23.52	0	5	7.5
181	0	21.405001	0	6	5.5
181	0	26.59	0	7	5.2
181	0	26.209999	0	1	5.4
178	0	28.059999	0	2	6.7
178	0	28.34	0	3	7.9
175	0	24.775	0	4	12.3
178	0	22.2075	0	5	10.6
178	0	23.4375	0	6	9.8
117	0	25.627501	0	7	7.8
117	0	28.6775	0	1	8.2
117	0	32.529999	0	2	7.2
117	0	33.557499	0	3	6.4
117	0	34.517502	0	4	7
112	0	30.74	0	5	5.8
117	1	30.147499	0	6	6
181	0	31.370001	0	7	5
181	2	29.8825	0	1	10.2

ofset	y	pm254	periodo	diad	tempmin
174	0	34.669998	0	2	5.7
174	0	35.157501	0	3	5.2
180	0	39.915001	0	4	6
174	0	45.1875	0	5	6.1
170	0	43.744999	0	6	6
180	1	36.052502	1	2	9.1
170	1	36.477501	1	3	7.1
159	2	40.083332	1	4	7.2
165	3	43.386665	1	5	8
169	2	41.546665	1	6	7.8
171	0	39.046665	1	7	4
171	1	38.700001	1	1	5
179	0	34.612499	1	2	4.3
179	0	32.817501	1	3	5.9
179	1	35.807499	1	4	3.7
181	0	36.567501	1	5	5
180	0	35.567501	1	6	4
179	0	35.426666	1	7	2.8
179	2	34.490002	1	1	2
178	0	31.533333	1	2	3.6
178	0	32.056667	1	3	3.4
178	3	29.5975	1	4	3.2
174	0	27.0875	1	5	5
174	2	27.584999	1	6	4.8
165	0	29.122499	1	7	5
165	1	29.7875	1	1	6.8
173	0	29.8575	1	2	5.9
173	1	29.1675	1	3	6.8
176	0	30.2125	1	4	7.2
177	0	31.625	1	5	4
175	1	34.139999	1	6	5.8
180	0	35.0075	1	7	7
180	6	30.5525	1	1	5.5
172	0	28.487499	1	2	8

ofset	y	pm254	periodo	diad	tempmin
172	1	28.514999	1	3	7.7
170	1	23.309999	1	4	5
174	1	27.7075	1	5	4.8
177	4	30.865	1	6	5.5
170	0	28.610001	1	7	6
170	0	33.576668	1	1	6
174	0	27.153334	1	2	5
174	0	25.290001	1	3	5
176	1	26.83	1	4	5.8
173	2	31.15	1	5	8.3
174	3	32.642502	1	6	6
171	0	30.23	1	7	4.8
171	0	25.995001	1	1	5
179	0	18.639999	1	2	5.4
179	1	16.82	1	3	5.9
175	0	18.5875	1	4	4.1
177	1	20.950001	1	5	6
175	0	21.352501	1	6	5.1
176	0	24.540001	1	7	9
176	0	25.1775	1	1	7.3
177	0	25.965	1	2	7.5
177	0	28.122499	1	3	10
181	0	26.393333	1	4	5
181	0	35.653332	1	1	10
175	0	35.814999	1	2	7.3
175	0	36.7925	1	3	7.6
182	0	35.517502	1	4	12.4
181	0	35.689999	1	5	9
182	0	36.5625	1	6	10
181	0	34.5975	1	7	8
181	0	35.945	1	1	8
181	0	31.115	1	2	7
181	0	25.182501	1	3	4.9
180	0	21.969999	1	4	5.5

ofset	y	pm254	periodo	diad	tempmin
179	0	23.924999	1	5	2.8
180	0	28.74	1	6	5
182	0	32.4175	1	7	7.5
182	1	34.82	1	1	8
176	0	33.59	1	2	6.5
176	1	26.993334	1	3	6.5
177	0	25.799999	1	4	5.5
176	2	26.27	1	5	4.5
177	0	26.549999	1	6	4.8
181	0	26.01	1	7	6
181	2	27.4575	1	1	6.5
179	0	27.372499	1	2	6.2
179	0	25.004999	1	3	6.7
179	0	25.67775	1	4	8
179	0	23.2900001	1	5	7.3
179	0	26.075001	1	6	9.6
178	0	29.5	1	7	9.6
178	2	33.032501	1	1	10.5
180	0	34.977501	1	2	10.2
180	0	35.990002	1	3	9.5
177	0	34.904999	1	4	12.1
179	0	33.482498	1	5	9
179	0	32.4575	1	6	10.2
182	0	29.5075	1	7	10.2
182	1	29.967501	1	1	7.5
178	0	28.360001	1	2	9
178	1	26.975	1	3	8.7
177	0	26.2225	1	4	13
181	0	24.125	1	5	6.8
181	0	22.1875	1	6	7

Apéndice 2

4.1 Rutinas de Splus utilizadas en el capítulo 3.

- *read.table("nombre del archivo", header=T):* lee los datos así como los nombres de las variables inscritos en el primer renglón de un archivo externo en formato ascii.

- *factor(variable)*: convierte una variable en variable categórica

- *gam(resp ~suavizador(variable1)+..+offset(log(var_ofset)), family=Poisson, data=archivo de datos):* ajusta un modelo aditivo, la función suavizador puede ser; *lo*(variable, span) correspondiente a un suavizador de regresión, *s*(variable, df) correspondiente a un suavizador de regresión si df=1 es una regresión lineal.

- *summary(modelo_gam)*: despliega el resultado del ajuste de un modelo aditivo generalizado

- *plot.gam(modelo1,residuals=T,rug=F)*: grafica la contribución de cada variable al predictor aditivo ajustado.

- *plot.gam(modelo1)*: grafica la contribución de cada variable al predictor aditivo ajustado, de manera interactiva.

- *glm(resp ~suavizador(variable1)+..+offset(log(var_ofset)), family=Poisson, data=archivo de datos)*: ajusta un modelo lineal generalizado.

- *fitted(modelo_lin):* despliega los valores ajustados del modelo lineal generalizado
- *predict(modelo_lin)*: despliega los valores del predictor lineal del modelo lineal generalizado
- La siguiente rutina ayuda a calcular lo p-valores de significancia de las covariables del modelo y despliega el resultado junto con el valor de los valores estimados, t-valor y el p-valor.

  estimados <- modelo$coefficients

  df <- modelo$df.residual

  tvalor <- summary(modelo)$coefficients[,3]

  pvalor <- 2(1-pt(abs(tvalor),df))*

  round(cbind(estimados,tvalor,pvalor),4)

  deviance(modelo_lin): despliega el valor de la devianza del modelo lineal generalizado
- *modelo_lin$resid*: despliega los valores de los residuos de trabajo del modelo lineal generalizado los cuales se usan para calcular los residuos parciales de la siguiente manera:

  residuos_parciales = modelo_lin$resid + fitted(modelo_lin)

Bibliografía

[1] Barndorff-Nielsen, O. (1978) *Information and Exponential Families in Statistical Theory.* New York; John Wiley.

[2] Bliss, C. I.(1935) The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* **22**, 134-167.

[3] Bradley, E. (1986). Double Exponential Families and Their Use in GLR. *J. Am. Statist. Assoc.* **81**, 709-721

[4] Breiman, L. y Friedman, J. H. (1982), *Estimating Optimal Transformations for Multiple Regression and Correlation,* Dept. of Statistics Tech. Rept, Orion 16, Stanford.

[5] Brillinger, D. (1977). Discussion of consistent nonparametric regression, by C. J. Stone. *Ann. Statist.* **5** 622-623

[6] Buja, A., Hastie, T. y Tibshirani, R. (1989), Linear smoothers and additive models. *The annals of statistics.* vol 17, No. 2, 453-555.

[7] Chambers, J. M. (1992) *Statistical Models in S.* Wandsworth and Brooks / Cole Advanced Books and Software Pacific Grove, California.

[8] Collet, D(1991). *Modelling Binary Data.* University of Reeding, UK. Chapman and Hall.

[9] Cook, R.D. y Weisberg. S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, London.

[10] Daryl, Pregibon (1981) Logistic Regression Diagnostics. *The Annals of Statistics.* 9, 705-724

[11] Dobson, J. A.(1990) *An Introduction to Generalized Linear Models.* University of Newcastle. New South Wales, Australia. Chapmand and Hall.

[12] Donald a Pierce y Daniel W. Shafer. *J. Am. Statist. Assoc.* **81**, 977-986.

[13] Fisher, R. A. (1922) On the mathematical foundations of thoretical statistics. Phil. Trans. *R. Soc.* **222**. 309-368.

[14] Fisher, R. A. (1934) Two new properties likelihood. *Proceeding of the Royal Society* A. **144**. 285-307

[15] Friedman, J.H. y Suetzle, W. (1982), *Smoothing of Scatterplots*, Dept. of Statistics Tech. Rept. Orion 3, Stanford University.

[16] Green, P.J. y Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall. New York.

[17] Gu (1989) Penalized likelihood regression: a Bayesian analysis. *Tech rep.* Dept. of Statistics, University of Wisconsin.

[18] Hastie, T. J., y Tibshirani, R. J. (1984). Generalized Additive Models, *Technical report No. 98* October, 1984. Division of Biostatistics Stanford University. Standford California.

[19] Hastie, T. y Tibshirani, R. (1986). Generalized additive models; some applications, *J. Am. Statist. Assoc.* **82**, 371-86.

[20] Hastie, T. J. y Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[21] Her Majesty's Public Health Service (1954). Mortality and Morbity during the London Fog of December 1952, Report No. 95 on Public Health and Medical Subjects. Her Majesty's Stationary Office, London.

[22] Hernan, San Martín (1983). *Salud y Enfermedad 4a edición.* Ediciones científicas la prensa médica S. A., México.

[23] Jørgensen, B.(1987). Exponential Dispersion Models. *J. R. Statist. Soc* B **49** 127-162.

[24] Kleinbaum, Kupper y Muller (1988). *Applied Regression Analysis and Other Multivariate Methods.* 2a. ed. PWS-KENT.

[25] Lindsey, J. K. (1997). *Applying Generalized Linear Models.* Springer-Verlag, New York.

[26] McCullagh, P., y Nelder, J.A. (1989) *Generalized Linear Models 2a, edition.* Chapman and Hall, London.

[27] Nelder, F. A. y Wedderburn, R.W.M. (1972). Generalized linear models. *J. R. Statist. Soc.* A. **135** 370-384.

[28] Owe, A. (1983). The estimation of smooth curves. Unpublished manuscript.

[29] Sander, G.(1995). Dose Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. *Epidemiology Resources Inc.* 6:356-365.

[30] Schwartz, J.(1994). Nonparametric smoothing in the analysis of air pollution and respiratory illnes. Harvard School of public Health. *The Canadian Journal of Statistics.* **22**, 471-487.

[31] Schwartz, J.(1996). Methodological Issues in Studies of Air Pollution and Daily Counts of deaths or hospital Admissions. *J. Epidemiol Community Health*, Apr 50 suppl, 1 53-111.

[32] Schwartz, J.(1994). Nonparametric smoothing in the analysis of air pollution and respiratory disease. *Canadian J. Stat.* **22**, 471-187.

[33] Selvin (1998). *Modern Applied Biostatistical Methods Using S-Plus.* Univ. of California Berkely, Oxford Univ. Press

[34] Stigler, S. M. (1981). Gauss and the invention of least squares. *Ann. Statist.* **9**, 465-474.

[35] Stigler, S. M. (1986). *The History of Statistics*. Belk nap Press, Cambridge, Mass.

[36] Wedderburn, R. W. (1976). On the existence and uniqueness of the maximun likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27-32

[37] Williams, D. A. (1987). GLM Diagnostics Using the Deviance and Single Case Deletions. *Appl Stat.* **36**, 181-191.