



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

---

---

INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS  
APLICADAS Y EN SISTEMAS

PUNTO DE CAMBIO EN  
MODELOS LINEALES GENERALIZADOS

T E S I S

QUE PARA OPTAR POR EL GRADO DE:  
MAESTRO EN CIENCIAS

PRESENTA:  
ACT. BENJAMÍN FIGUEROA SOLANO

DIRECTORA DE TESIS:  
DRA. SILVIA RUIZ VELASCO ACOSTA IIMAS

Ciudad de México, 2016  
Diciembre



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



## Dedicatoria

*Dedico este trabajo a la memoria de mis abuelos Margarita Bahena Aguilar, Daniel Figueroa Estrada y Erasmo Solano García, que en paz descansen, por haber sido ejemplo de rectitud, entrega y amor por la vida.*

*También quiero dar un agradecimiento especial a mis padres, Esther Solano Terán y Felipe Figueroa Bahena, pues sin su amor, motivación y sabios consejos, jamás habría alcanzado mis metas en la vida.*

*¡A ustedes dedico este trabajo!*



## Agradecimientos

*Agradezco,*

*a mi Directora, la Dra. Silvia Ruiz Velasco Acosta, por su invaluable tiempo y asesoría brindados para la elaboración de este trabajo,*

*a mis profesores de la Maestría en Ciencias Matemáticas, por haberme compartido su conocimiento y experiencia en el amplio campo de la estadística,*

*a la Universidad Nacional Autónoma de México, por la educación que recibí en sus aulas y por el honor de formar parte de sus filas como profesor de asignatura,*

*al Consejo Nacional de Ciencia y Tecnología, que aportó los recursos económicos para la beca con que se financiaron mis estudios,*

*al Dr. Luis G. Gorostiza Ortega, quién también me otorgó una beca,*

*a la Dra. Eliane R. Rodrigues, por ser mi tutora en la maestría y enseñarme muchas cosas interesantes en su curso,*

*a mis amigos: Adriana Roldán, Adriana Meneses, Angélica Ascensión, Laura Ramírez, Arturo Hernández, Alejandro Nava, Fernando Baltazar, Galileo Ramírez, Miguel Ángel Chong, Julio Soto, David Luna,*

*a Sarai Estrada Vera, por estar siempre conmigo*

*a todas las personas que por falta de espacio no menciono aquí pero que de una u otra forma han contribuido a mi vida,*

*a Dios Todopoderoso, por la oportunidad de finalizar este trabajo.*



## Índice general

Dedicatoria	3
Agradecimientos	5
Capítulo 1. Introducción	9
Capítulo 2. Modelos lineales generalizados	13
2.1. Introducción	13
2.2. El modelo de regresión lineal clásico	13
2.3. Los modelos lineales generalizados	15
2.4. Ecuaciones de verosimilitud para modelos lineales generalizados	22
2.5. La cuasi-verosimilitud	27
2.6. Bondad de ajuste	32
2.7. Residuos	34
2.8. Métodos gráficos de verificación de supuestos	37
2.9. El modelo de regresión logística	39
2.10. Modelo Poisson	41
Capítulo 3. Puntos de cambio en modelos lineales	47
3.1. Introducción	47
3.2. Cociente de log-verosimilitudes	48
3.3. El enfoque basado en la información	49
3.4. Prueba de Chow	58
3.5. Pruebas basadas en la suma acumulada de los residuos	61
3.6. Algunas pruebas más	66



Capítulo 4. Punto de cambio en modelos lineales generalizados	69
4.1. Introducción	69
4.2. Punto de cambio en el modelo de regresión logístico	70
4.3. Monitoreo de cambios estructurales en modelos lineales generalizados	75
4.4. Método SIC para la detección del cambio paramétrico	80
4.5. Método SIC para detectar punto cambio en covariables	88
4.6. Método SIC con el modelo logístico	89
4.7. Simulación de datos	97
Capítulo 5. Conclusiones	103
Apéndice A. Análisis de un modelo de regresión con R	105
Modelo de regresión lineal	105
Modelo de regresión lineal generalizado	107
Apéndice B. Estimación de los parámetros de un MLG	113
Apéndice C. Códigos de R	117
Apéndice. Bibliografía	121
Bibliografía	121

## Introducción

El análisis de regresión es una de las técnicas estadísticas más importantes con que se cuenta para el análisis de información y su principal objetivo es facilitar el entendimiento de la relación que pudiera existir entre una variable de interés para nosotros (variable dependiente) y una o varias variables explicativas (variables independientes) para las cuales hemos recolectado información de las unidades muestrales de estudio.

En el afán de lograr resultados más precisos derivados de modelos con niveles de bondad de ajuste razonables, en las décadas de 1950 y 1960 se propusieron modelos para analizar datos de tipo categórico o de conteos, dando lugar a modelos como la regresión logística o la regresión Poisson, entre otros. Años después, en la década de los setenta, todas las teorías que surgieron aisladamente fueron unificadas bajo el nombre de modelos lineales generalizados (MLG) con lo cual fue posible relacionar variables no sólo de tipo normal sino cualquier variable que pertenezca a la *familia de distribuciones exponencial*. Algunas de las distribuciones más conocidas de la familia exponencial son la binomial, Poisson y gamma, entre otras.

La principal característica de esta familia de distribuciones es que su varianza se puede escribir como una función de la media. Como consecuencia, las propiedades de segundo orden de los parámetros estimados no dependen del supuesto distribucional sino únicamente de la relación entre media y varianza y de la no correlación entre observaciones, lo cual resulta bastante conveniente ya que pocas veces se

puede confiar en que el supuesto distribucional sea el correcto (McCullagh, 1989, [17]).

No obstante que la aparición de la clase de los modelos lineales generalizados abrió nuevas posibilidades al análisis de ciertos conjuntos de datos, en las últimas décadas se ha ido desarrollando un nuevo campo de investigación, que tiene que ver con el problema de que los datos presenten comportamientos diferenciados en el tiempo (o en alguna otra variable explicativa que guarde algún orden) que originan un rompimiento en la tendencia original de los mismos. En la literatura estadística a este problema se le ha dado el nombre de cambio estructural, o simplemente, *problema del punto de cambio*.

Cuando en un modelo lineal generalizado se sospecha que existe un cambio estructural o punto de cambio, ocurre con frecuencia que la bondad del ajuste del modelo es deficiente. En el presente trabajo se hizo una revisión de la literatura existente para este tema llegando a conocer los métodos con los que se cuenta actualmente para realizar una crítica de lo ya existente y, de ser posible, hacer nuevos aportes a la metodología. Así pues, el objetivo de este trabajo es revisar algunos métodos que permitan la identificación de un punto de cambio en un modelo lineal generalizado, cuando se sospecha que éste ocurrió.

Una razón por la que me interesé en investigar este tema es porque cada vez es más común que en distintas áreas de la ciencia se utilicen modelos lineales generalizados para tratar de ajustar una variable respuesta cuya distribución no es normal. No obstante, en ocasiones el ajuste que se obtiene no resulta satisfactorio. Lo anterior puede deberse, entre otras cosas, a la presencia de un punto de cambio. Dado que se trata de un tema relativamente nuevo, me interesó aprender y sobre todo me llamó la atención la posibilidad de hacer nuevas aportaciones.

A continuación explico el esquema sobre el cual desarrollé mi investigación describiendo a grandes rasgos el contenido de cada capítulo.

En el capítulo uno, se presenta la introducción de la tesis.

En el capítulo dos, se define el modelo lineal generalizado, sus propiedades, la manera de estimar los parámetros y las pruebas para verificar la validez de dichos modelos. Lo anterior con el objeto de establecer la base para el estudio de los métodos que se expondrán dos capítulos más adelante. También se presentan dos ejemplos de conjuntos de datos para los cuales se aplica una regresión usando un modelo lineal generalizado.

En el capítulo tres, se introduce el concepto de punto de cambio para el modelo lineal clásico y se presenta una revisión de algunos de los métodos disponibles para identificar y localizar un punto de cambio en un modelo lineal clásico. El objetivo de esta revisión es conocer cómo abordaron el problema los investigadores para posteriormente intentar generalizar esas metodologías al caso de los modelos lineales generalizados. Se presenta un ejemplo numérico de aplicación de los métodos descritos.

En el capítulo cuatro, se presenta una revisión de los métodos que se conocen para resolver el problema de punto de cambio en un modelo lineal generalizado. También se propone un método simple para identificar un punto de cambio en un modelo lineal generalizado. En este capítulo también se presentan algunos ejemplos con conjuntos de datos para los cuales se tenía la sospecha de que había ocurrido un punto de cambio.

Por último, en el capítulo cinco, se dan las conclusiones a las que se llegó como resultado de esta investigación; en éstas se expone el punto de vista del autor sobre las ventajas y desventajas de los métodos que aquí se presentaron.

Tengo la esperanza de que este trabajo sea de utilidad para que quienes estén interesados en el tema de punto de cambio en modelos lineales generalizados puedan introducirse al tema que es por demás vasto.

## Modelos lineales generalizados

### 2.1. Introducción

En este capítulo se hace una revisión de la teoría de los modelos lineales generalizados. Con tal propósito, en primer lugar se aborda el modelo lineal clásico para posteriormente estudiar los modelos lineales generalizados. Se comienza por definir qué es un modelo lineal generalizado, se establecen sus propiedades, se presenta un algoritmo para estimar sus parámetros y, por último, se presentan algunas pruebas para verificar la validez de estos modelos. Se presentan dos ejemplos para los cuales se puede ajustar una regresión con un modelo lineal generalizado. Todo lo anterior con el objetivo de establecer las bases para el estudio de los métodos que serán expuestos más adelante.

### 2.2. El modelo de regresión lineal clásico

¿Qué es y para qué sirve un modelo? Un modelo es una representación abstracta de la realidad que percibimos. Los modelos sirven para explicar el entorno. Un buen modelo es aquel que reproduce con relativa exactitud lo que ocurre en la realidad. El modelo lineal clásico ha sido utilizado en gran cantidad de aplicaciones en diferentes ramas de la ciencia. A continuación un poco de su historia.

El modelo lineal clásico y los mínimos cuadrados surgen con los trabajos de Carl Friedrich Gauss (30/4/1777-23/2/1855) y Adrien-Marie Legendre (18/9/1752-10/1/1833) (Stigler, 1981) quienes aplicaron esta última técnica a datos astronómicos. En el

siglo XVIII, por primera vez, se empezó a estudiar problemas tipo regresión, buscando ayudar a los navegantes quienes utilizaban las estrellas para guiarse en el mar. En 1805, Adrien-Marie Legendre desarrolló el método de mínimos cuadrados. Casi simultáneamente, Carl Friedrich Gauss desarrolló el mismo método y en 1809, probó que los mínimos cuadrados son solución óptima cuando la distribución de los errores de experimentos es Normal. Gauss introdujo la distribución Normal como mecanismo para describir la variabilidad de los errores de experimentos y probó que las propiedades importantes de los estimadores por mínimos cuadrados dependen de la suposición de varianza constante, más no del supuesto de normalidad. Hasta finales del siglo XIX, esta metodología se utilizó de manera exclusiva en la física y la astronomía, pero en 1875, Francis Galton (16/2/1822-17/1/1911) la utilizó para explicar el fenómeno de que los hijos de padres altos tienden a ser altos pero no tanto como sus padres, y asimismo, los hijos de padres bajos tienden a ser bajos pero no tanto como sus padres. A este fenómeno Galton lo llamó: *regresión a la media o efecto regresivo*.

Actualmente los modelos de regresión lineal se siguen empleando en la ciencia y la ingeniería para resolver una gran cantidad de problemas, por varias razones. La primera, porque son buenas aproximaciones para relaciones funcionales más complejas. La segunda, porque una vez que se tiene el modelo resulta bastante sencillo estimar sus parámetros. La tercera y última razón, es porque la teoría estadística de los modelos lineales está bien construida, es elegante y por si fuera poco, es bella.

Ahora se definirá el modelo de regresión lineal. Considere la ecuación  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$  donde:

- $\mathbf{Y}$  es un vector aleatorio observable de  $n \times 1$ ,

- $\mathbf{X}$  es una matriz de  $n \times p$  de números fijos observables (los elementos de  $\mathbf{X}$  no son variables aleatorias),
- $\boldsymbol{\beta}$  es un vector de parámetros no observables de  $p \times 1$  definidos en el espacio paramétrico  $\Omega_{\boldsymbol{\beta}}$ ,
- $\boldsymbol{\epsilon}$  es un vector aleatorio no observable de  $n \times 1$  tal que  $\mathbb{E}[\boldsymbol{\epsilon}] = 0$  y  $\text{Cov}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma} = \sigma^2 I$ .

Estas cuatro especificaciones definen por completo al modelo lineal clásico.

La forma usual de ajustar un modelo lineal clásico es mediante mínimos cuadrados, minimizando la forma cuadrática:

$$(2.2.1) \quad \min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \min \sum_i (y_i - \hat{y}_i)^2 = \min \sum_i (y_i - x_i \hat{\boldsymbol{\beta}})$$

correspondiente a la suma de cuadrados de los errores, SCE ( $x_i$ , vector transpuesto de los valores de las variables explicativas para el individuo  $i$ ). Bajo el supuesto de que los errores siguen una distribución  $N(0, \sigma^2)$ , entonces  $\mathbf{Y}$  también se distribuye Normal y los estimadores por mínimos cuadrados coinciden con los estimadores máximo verosímiles que se obtienen al maximizar la log-verosimilitud

$$(2.2.2) \quad \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}, \mathbf{X}) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

En el caso de que  $\mathbf{X}$  sea de rango completo, sin multicolinealidad, el estimador de  $\boldsymbol{\beta}$  resulta:  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$ . En el Apéndice A se muestra el ajuste y código para ajustar un modelo lineal usando R.

### 2.3. Los modelos lineales generalizados

**2.3.1. Antecedentes.** La distribución Normal juega un papel central en la teoría de los modelos de regresión clásicos, sin embargo, hay una gran cantidad de situaciones prácticas en las que ni remotamente van a satisfacerse las hipótesis



de normalidad. Por ejemplo, cuando la variable respuesta es discreta, como en el caso de los conteos. A menudo nos encontramos con conteos de piezas defectuosas o cualquier otro clase de eventos no consistentes, como son pacientes con traumatismo u otro determinado padecimiento e incluso ciertos fenómenos naturales como terremotos o huracanes. Los modelos lineales generalizados permiten desarrollar modelos para el análisis de conteos análogamente al modelo lineal clásico para variables continuas. Otra posibilidad son las variables de respuesta binaria, éxito o fracaso, las cuales son bastante comunes en muchas áreas de la ciencia y la ingeniería. Existen también situaciones en las que la variable respuesta es continua pero que no es razonable suponer normalidad. Un ejemplo es el tiempo para que ocurra la primera falla en un sistema o un componente. Estos tipos de variables de respuesta son no negativas y típicamente tienen un comportamiento altamente asimétrico. Los modelos lineales generalizados (MLG) son a menudo mejores modelos para esas situaciones que los modelos de regresión lineal ordinarios.

La clase de los modelos lineales generalizados incluye como casos especiales la regresión lineal, el análisis de varianza (ANOVA), los modelos logit y probit para el análisis de datos binarios, la regresión Poisson, el modelo log-lineal para conteos, entre otros. Se ha probado que todos estos modelos comparten propiedades en común y que es posible definir un método general para estimar sus parámetros que los considere a todos.

En 1972, J. A. Nelder y R. W. Wedderburn publicaron el artículo *Generalized linear models* [4] en el que unifican la teoría de varios modelos estadísticos, específicamente aquellos que relacionan un conjunto de variables explicativas con una variable respuesta. En 1987, Jørgensen [16] probó que todos los modelos enunciados tienen una distribución perteneciente a la *familia exponencial de dispersión*, que es una generalización de la familia exponencial con alguna transformación de

la media. Debido a su importancia para la teoría de los modelos lineales generalizados, a continuación se presenta un resumen de la familia exponencial.

**2.3.2. La familia exponencial de distribuciones.** Un importante concepto unificador dentro de los modelos lineales generalizados es el de familia exponencial de distribuciones. Se tiene un conjunto de  $n$  variables aleatorias independientes  $Z_i$ , cuya función de densidad de probabilidad se puede escribir como:

$$(2.3.1) \quad f(z_i; \xi_i) = r(z_i)s(\xi_i) \exp(t(z_i)u(\xi_i))$$

Si en la ecuación anterior se hacen los cambios de variable  $v(z_i) = \log(r(z_i))$  y  $w(\xi_i) = \log(\omega(\xi_i))$ , entonces

$$(2.3.2) \quad f(z_i; \xi_i) = \exp[t(z_i)u(\xi_i) + v(z_i) + w(\xi_i)]$$

donde  $\xi_i$  es parámetro de localización. Si además aplicamos 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 función de densidad dada por:

$$(2.3.3) \quad f(y_i; \theta_i) = \exp(y_i\theta_i - b(\theta_i) + c(y_i)),$$

(dentro de esta familia se encuentran las distribuciones Poisson y binomial, entre otras).

La familia exponencial se puede generalizar al incluir un parámetro de escala  $\phi$  de la siguiente manera:

$$(2.3.4) \quad f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right].$$

A la familia que resulta se le da el nombre de *familia exponencial de dispersión*, en la cual se encuentran las distribuciones Normal y Gamma, entre otras. En el caso

$\phi$  conocido, se tiene a la familia exponencial de un parámetro cuya forma analítica está dada por la ecuación (2.3.3). Para la familia exponencial de dispersión existe una importante relación entre media y varianza (Lyndsey, 1997 [21]), misma que se desarrollará a continuación. Para obtener la relación media-varianza se usará la función de puntajes.

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

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

la función de puntajes. Usando propiedades de esperanza y varianza se puede probar que si la distribución en cuestión cumple con ciertas condiciones de regularidad, entonces

$$(2.3.5) \quad \mathbb{E}[U_i] = 0$$

y

$$(2.3.6) \quad \text{Var}[U_i] = \mathbb{E}[U_i^2] = \mathbb{E} \left[ -\frac{\partial U_i}{\partial \theta_i} \right]$$

Puesto que la familia de dispersión exponencial cumple las condiciones de regularidad, las ecuaciones (2.3.5) y (2.3.6) 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,

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

por lo que usando la ecuación (2.3.5) se tiene:

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

y,

$$\text{Var}[U_i] = \frac{\text{Var}[Y_i]}{a_i^2(\phi)}.$$

Utilizando la ecuación (2.3.6) y el resultado anterior

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

Sea  $V(\mu_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$ , la cual se conoce con el nombre de función varianza, entonces

$$\text{Var}[Y_i] = V(\mu_i) a_i(\phi)$$

donde  $\theta_i$  es el parámetro de interés,  $\phi$  es el parámetro de dispersión que comúnmente se considera como parámetro de ruido y  $a_i(\phi)$  es parte de la función de distribución.

El Cuadro (1) presenta las distribuciones univariadas más importantes de la familia exponencial.

	Normal	Poisson	Binomial	Gamma
Notación	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \pi)/m$	$G(\mu, \nu)$
Rango de $y$	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0, \infty)$
Parámetro de dispersión: $\phi$	$\phi = \sigma^2$	$\mu$	$1/m\phi$	$= \nu^{-1}$
Función cumulante: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$
$c(y; \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log y!$	$\log\left(\frac{m}{my}\right)$	$\nu \log(\nu y) - \log(y) - \log\Gamma(\nu)$
$\mu(\theta) = \mathbb{E}(Y; \theta)$	$\theta$	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$
Liga canónica: $\theta(\mu)$	identidad	log	logit	recíproco
Función varianza: $V(\mu)$	1	$\mu$	$\mu(1 - \mu)$	$\mu^2$

CUADRO 1. Características de algunas distribuciones univariadas comunes de la familia exponencial.

**2.3.3. Los modelos lineales.** Un modelo lineal está compuesto de una parte sistemática y otra aleatoria. La parte sistemática del modelo es una combinación lineal de las variables explicativas en términos de un conjunto de parámetros desconocidos  $\beta_1, \dots, \beta_p$ . La parte aleatoria se refiere a la respuesta promedio  $\mathbb{E}(Y_i) = \mu_i$ . Adicionalmente, para la parte aleatoria del modelo debemos suponer independencia y varianza constante de los residuales (estas suposiciones son fuertes y necesitan validarse hasta donde sea posible).

La relación entre la parte sistemática y la parte aleatoria del modelo se escribe como:

$$(2.3.7) \quad \mathbb{E}(Y_i) = \mu_i = \sum_{j=1}^p x_{ij}\beta_j \quad i = 1, \dots, n$$

donde  $x_{ij}$  es el valor de la  $j$ -ésima covariable para la observación  $i$ .

En notación matricial

$$(2.3.8) \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

donde  $\boldsymbol{\mu}$  es de  $n \times 1$ ;  $\mathbf{X}$  de  $n \times p$ , es la matriz de diseño, y  $\boldsymbol{\beta}$ , de  $p \times 1$ , es el vector de parámetros.

En general, un modelo lineal generalizado consiste de:

1. Un *componente aleatorio* que consta de observaciones independientes de  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  donde  $Y_i$  es una variable aleatoria con distribución perteneciente a la familia exponencial. En el caso del modelo lineal clásico, ésta corresponde a la Normal con media  $\mu_i$  y varianza  $\sigma^2$  constante para  $i = 1, \dots, n$ .
2. Un *componente sistemático* que se basa en suponer que la respuesta media varía según una función suave  $\eta(\cdot)$  que se conoce como *predictor lineal*, es

decir,

$$(2.3.9) \quad \eta(\cdot) = \mathbf{X}\boldsymbol{\beta}$$

3. Una *función liga* que describe la relación entre la media de la  $i$ -ésima observación y su predictor lineal. Para  $g(\cdot)$  monótona y diferenciable

$$(2.3.10) \quad \eta_i = g(\mu_i) = \sum_{j=1}^p \beta_j x_{ij} \quad i = 1, \dots, n$$

La función  $g(\cdot)$  se conoce como *liga canónica* si se cumple que  $\theta = \eta$ , donde  $\theta$  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. En el Cuadro (1) se muestra la liga canónica para algunas de las distribuciones más importantes. En el Apéndice A se muestra el ajuste y código para ajustar un modelo lineal generalizado usando R.

#### 2.4. Ecuaciones de verosimilitud para modelos lineales generalizados

El método de máxima verosimilitud es la base teórica para la estimación de parámetros en los modelos lineales generalizados. No obstante, la implementación real de la máxima verosimilitud resulta en un algoritmo basado en mínimos cuadrados iterativamente ponderados.

Considere el método de máxima verosimilitud aplicado a los modelos lineales generalizados y suponga que usamos la liga canónica. La función de log-verosimilitud es

$$L = \sum_{i=1}^n \{ [y_i \theta_i - b(\theta_i)] / a(\phi) + c(y_i, \phi) \}$$

Para la liga canónica tenemos  $\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ , por lo tanto

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\beta}} &= \frac{\partial L}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^n \frac{1}{a(\phi)} \left[ y_i - \frac{\partial b(\theta_i)}{\partial \theta_i} \right] \mathbf{x}_i \\ &= \sum_{i=1}^n \frac{1}{a(\phi)} (y_i - \mu_i) \mathbf{x}_i \end{aligned}$$

Usar la liga canónica simplifica la dificultad matemática. Podemos encontrar los estimadores máximo verosímiles de los parámetros resolviendo el siguiente sistema de ecuaciones para  $\boldsymbol{\beta}$

$$\sum_{i=1}^n \frac{1}{a(\phi)} (y_i - \mu_i) \mathbf{x}_i = 0$$

en la mayoría de casos,  $a(\phi)$  es constante, de manera que estas ecuaciones se transforman en

$$\sum_{i=1}^n (y_i - \mu_i) \mathbf{x}_i = 0$$

Esto es en realidad un sistema de  $p$  ecuaciones, una por cada parámetro del modelo.

En forma matricial estas ecuaciones

$$\mathbf{X}^T (\mathbf{y} - \boldsymbol{\mu}) = 0$$

donde  $\boldsymbol{\mu}^T = [\mu_1, \dots, \mu_n]$ . Estas son llamadas ecuaciones de puntaje. Entonces la función de puntaje, la cual era operativa para la regresión lineal múltiple (errores Normales), regresión logística, y regresión Poisson, es relevante para una clase más grande de modelos, a saber, para los modelos lineales generalizados en los cuales se utiliza la liga canónica.

Sea  $\mathbf{b}$  el valor final que el algoritmo anterior produce como estimador  $\boldsymbol{\beta}$ . Si las suposiciones del modelo, incluyendo la elección de la función liga, son correctos,



entonces se puede probar que asintóticamente

$$\mathbb{E}(\mathbf{b}) = \boldsymbol{\beta}$$

dado que  $\mathbf{b}$  es la solución de las ecuaciones. La matriz de información  $\mathbf{I}(\mathbf{b})$  de estimadores dados por la varianza del puntaje es

$$\mathbf{I}(\mathbf{b}) = \text{Var} \left\{ \frac{1}{a(\phi)} [\mathbf{X}^T (\mathbf{y} - \boldsymbol{\mu})] \right\} = \frac{\mathbf{X}^T \mathbf{V} \mathbf{X}}{[a(\phi)]^2}$$

donde  $\mathbf{V} = \text{diag} \{ \sigma_i^2 \}$  la matriz de varianzas y covarianzas asociada al vector aleatorio  $\mathbf{Y}$  (por la independencia es diagonal) y  $\sigma_i^2$ , la cual es una función de  $\mu_i$ , depende de la distribución en cuestión. Entonces la matriz de varianza-covarianza asintótica de  $\mathbf{b}$  está dado por

$$\text{Var}(\mathbf{b}) = \mathbf{I}^{-1}(\mathbf{b}) = [\mathbf{X}^T \mathbf{V} \mathbf{X}]^{-1} [a(\phi)]^2$$

Los errores estándar estimados de los coeficientes vienen de las raíces cuadradas de los elementos de la diagonal de la matriz anterior con elementos de  $\mathbf{V}$  reemplazados por sus estimadores apropiados.

Para el caso de la Normal todo  $\sigma_i^2 = \sigma^2$ ,  $a(\phi) = \sigma$ , y por lo tanto tenemos la familiar

$$\text{Var}(\mathbf{b}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$$

Para los modelos logístico y Poisson  $a(\phi) = 1$ , y recordando

$$\text{Var}(\mathbf{b}) = (\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1}$$

donde para el caso Poisson  $\sigma_i^2 = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$  y para la binomial

$$\sigma_i^2 = n_i \pi_i (1 - \pi_i)$$

Es importante señalar que mientras la liga canónica es la liga más natural a considerar, dada la distribución involucrada, esto por ningún medio implica que una función liga no canónica no debería ser considerada. Elegir una liga es equivalente al ejercicio de selección de modelos en regresión lineal múltiple estándar, uno intenta que el modelo sea el que mejor describe el mecanismo de los datos. Por ejemplo, si los rangos sobre los regresores son suficientemente estrechos y un modelo lineal es suficiente, entonces podría ser considerada la liga identidad. Entonces es importante revisar qué cambios son relevantes en lo anterior cuando uno utiliza una liga no canónica.

Si no utilizamos la liga canónica, entonces  $\eta_i \neq \theta_i$ , y la derivada apropiada de la log-verosimilitud es

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{dL}{d\theta_i} \frac{d\theta_i}{d\eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}$$

Observamos que

$$\frac{dL}{d\theta_i} = \sum_{i=1}^n \frac{1}{a(\phi)} \left[ y_i - \frac{db(\theta_i)}{d\theta_i} \right] = \frac{1}{a(\phi)} \sum_{i=1}^n (y_i - \mu_i)$$

y

$$\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \mathbf{x}_i$$

( $x_i$ , vector transpuesto de los valores de las variables explicativas para el individuo  $i$ ) poniendo todo esto junto produce

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi)} \frac{d\theta_i}{d\eta_i} \mathbf{x}_i$$

Típicamente  $a(\phi)$  es una constante. Podemos expresar las ecuaciones de puntajes como

$$(2.4.1) \quad \mathbf{X}^T \Delta (\mathbf{y} - \boldsymbol{\mu}) = 0$$

donde  $\Delta = \text{diag} \{d\theta_i/d\eta_i\}$ . La matriz de información para el estimador máximo-verosímil MLE resultante de los coeficientes,  $\boldsymbol{\beta}$ , es

$$\mathbf{I}(\mathbf{b}) = \frac{\mathbf{X}^T \Delta \mathbf{V} \Delta \mathbf{X}}{[a(\phi)]^2}$$

Entonces la matriz varianza-covarianza asintótica de  $\mathbf{b}$  es

$$\text{Var}(\mathbf{b}) = (\mathbf{X}^T \Delta \mathbf{V} \Delta \mathbf{X})^{-1} [a(\phi)]^2$$

La ecuación para la liga canónica es más simple que para el caso general, además es claro que  $\boldsymbol{\mu}$  y  $d\theta_i/d\eta_i$  son ambas no lineales en general. Para comprender la complejidad que se involucra, considere un ejemplo con la distribución Poisson. Aún en el caso de la liga canónica tenemos

$$\sum_{i=1}^n (y_i - e^{\mathbf{x}_i \boldsymbol{\beta}}) \mathbf{x}_i = 0$$

la cual, por supuesto, no puede ser resuelta en forma cerrada puesto que no es lineal en  $\boldsymbol{\beta}$ . Ahora suponga que se decide que puede justificarse la liga identidad, por lo tanto resulta en un modelo que es lineal aunque la liga no es canónica. Tenemos que  $\mu = \mathbf{x}\boldsymbol{\beta}$ , y  $\partial\theta_i/\partial\eta_i = (1/\mathbf{x}_i\boldsymbol{\beta})$ , y entonces usando la ecuación (2.4.1), se llega a la ecuación de puntajes dada por

$$\sum_{i=1}^n \left[ \frac{(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \mathbf{x}_i}{\mathbf{x}'_i \boldsymbol{\beta}} \right] = 0$$

la cual debe ser resuelta iterativamente aun cuando el modelo es lineal. Si se usa el método de puntajes, la forma de estimar los parámetros del MLG es mediante modelos lineales con pesos ponderados. En el caso de la liga canónica, el método de puntajes es equivalente a Newton-Raphson.

## 2.5. La cuasi-verosimilitud

El propósito de esta sección es mostrar cómo se puede hacer inferencia con experimentos en los cuales no hay suficiente información para construir una función de verosimilitud. Se analizará el caso en el que se tienen observaciones independientes y donde los efectos de interés pueden ser descritos por un modelo para  $\mathbb{E}(\mathbf{Y})$ . Si se relajan las suposiciones sobre la media y la varianza este procedimiento es equivalente a mínimos cuadrados.

**2.5.1. Construcción de la función de cuasi-verosimilitud.** Considere una sola componente del vector de respuestas  $\mathbf{Y}$ , la cual escribimos como  $Y$ , sin subíndice. Considere la función

$$U = u(\mu; Y) = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

la cual tiene las siguientes propiedades en común con la log-verosimilitud:

$$\begin{aligned} \mathbb{E}(U) &= 0 \\ \text{Var}(U) &= \frac{1}{\sigma^2 V(\mu)} \\ -\mathbb{E}\left(\frac{\partial U}{\partial \mu}\right) &= \frac{1}{\sigma^2 V(\mu)} \end{aligned}$$

Puesto que la mayor parte de la teoría asintótica de primer orden conectada con las funciones de verosimilitud está fundada en estas tres propiedades, hasta cierto punto no es sorprendente que la integral

$$(2.5.1) \quad Q(\mu; y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt$$

cuando existe, se comparta como una función de log-verosimilitud para  $\mu$ . Ejemplos de tales cuasi-verosimilitudes para varias de las funciones de varianza comunes están dados en el Cuadro (2.5.1). Para algunas funciones varianza, estas cuasi-verosimilitudes corresponden a verdaderas log-verosimilitudes.

Nos referimos a  $Q(\mu; y)$  como la cuasi-verosimilitud o como log-cuasi-verosimilitud para  $\mu$  basada en los datos  $y$ . Puesto que por suposición las componentes de  $\mathbf{Y}$  son independientes, la cuasi-verosimilitud de los datos completos es la suma de las contribuciones individuales

$$Q(\mu; y) = \sum_i Q_i(\mu_i; y_i).$$

Función varianza	Cuasi-verosimilitud	Parámetro canónico	Nombre	Rango
1	$-(y - \mu)^2/2$	$\mu$	Normal	-
$\mu$	$y \log \mu - \mu$	$\log \mu$	Poisson	$\mu > 0, y \geq 0$
$\mu^2$	$-y/\mu - \log \mu$	$-1/\mu$	Gama	$\mu > 0, y \geq 0$
$\mu^3$	$-y/(2\mu^2) + 1/\mu$	$-1/(2\mu^2)$	Gaussiana inversa	$\mu > 0, y \geq 0$
$\mu^\zeta$	$\mu^{-\zeta} \left( \frac{\mu y}{1-\zeta} - \frac{\mu^2}{2-\zeta} \right)$	$\frac{1}{(1-\zeta)\mu^{\zeta-1}}$	-	$\mu > 0, \zeta \neq 0, 1, 2$
$\mu(1 - \mu)$	$y \log \left( \frac{\mu}{1-\mu} \right) + \log(1 - \mu)$	$\log \left( \frac{\mu}{1-\mu} \right)$	Binomial/ $m$	$0 < \mu < 1, 0 \leq y \leq 1$
$\mu^2(1 - \mu)^2$	$(2y - 1) \log \left( \frac{\mu}{1-\mu} \right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$	-	-	$0 < \mu < 1, 0 \leq y \leq 1$
$\mu + \mu^2/k$	$y \log \left( \frac{\mu}{k+\mu} \right) + k \log \left( \frac{k}{k+\mu} \right)$	$\log \left( \frac{\mu}{k+\mu} \right)$	Binomial Negativa	$\mu > 0, y \geq 0$

CUADRO 2. Cuasi-verosimilitudes asociadas con algunas funciones varianza simples.

**2.5.2. Funciones de covarianza.** Suponga que los componentes del vector de respuestas  $\mathbf{Y}$  son independientes con vector de medias  $\boldsymbol{\mu}$  y matriz de covarianza  $\sigma^2\mathbf{V}(\boldsymbol{\mu})$ , donde  $\sigma^2$  puede ser desconocida y  $\mathbf{V}(\boldsymbol{\mu})$  es una matriz de funciones conocidas. En esta sección se asumirá que los parámetros de interés  $\boldsymbol{\beta}$ , están relacionados con la dependencia de  $\boldsymbol{\mu}$  sobre las covariables  $\mathbf{x}$ . La naturaleza de esa relación no debe preocuparnos por el momento, así que escribimos  $\boldsymbol{\mu}(\boldsymbol{\beta})$ , absorbiendo de ese modo las covariables dentro de la función de regresión. Un punto importante es que  $\sigma^2$  se supone constante -en particular, que  $\sigma^2$  no depende de  $\boldsymbol{\beta}$ . Bajo el supuesto de que las componentes de  $\mathbf{Y}$  son independientes, la matriz  $\mathbf{V}(\boldsymbol{\mu})$  debe ser diagonal. Entonces escribimos

$$\mathbf{V}(\boldsymbol{\mu}) = \text{diag}\{V_1(\mu), \dots, V_n(\mu)\}.$$

Se requiere una suposición adicional respecto de las funciones  $V_i(\mu)$ , a saber que  $V_i(\mu)$  debe depender únicamente de la  $i$ -ésima componente de  $\boldsymbol{\mu}$ . En principio, incluso bajo independencia, es posible para  $V_i(\mu)$  depender de varias componentes de  $\boldsymbol{\mu}$ . Sin embargo, es difícil imaginar un mecanismo físico plausible que produjera tal dependencia en la función de varianza y que al mismo tiempo mantuviera las variables aleatorias estadísticamente independientes. En la mayoría de las aplicaciones las funciones  $V_1(\cdot), \dots, V_n(\cdot)$  pueden tomarse idénticas, aunque sus argumentos y por lo tanto sus valores, sean diferentes.

**2.5.3. Estimación de los parámetros.** Las ecuaciones de estimación de la cuasi-verosimilitud para los parámetros de la regresión  $\boldsymbol{\beta}$  obtenidas al derivar  $Q(\boldsymbol{\mu}; \mathbf{y})$  pueden ser escritas en la forma  $U(\hat{\boldsymbol{\beta}}) = 0$ , donde

$$(2.5.2) \quad U(\boldsymbol{\beta}) = \mathbf{D}^T \mathbf{V}^{-1} \left( \frac{Y - \boldsymbol{\mu}}{\sigma^2} \right)$$

se conoce como función cuasi-puntajes. En esta expresión las componentes de  $\mathbf{D}$ , de orden  $n \times p$ , son  $D_{ir} = \frac{\partial \mu_i}{\partial \beta_r}$ , las derivadas de  $\mu(\boldsymbol{\beta})$  con respecto a los parámetros.

La matriz de covarianza de  $U(\boldsymbol{\beta})$ , la cual es también el valor esperado negativo de  $\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ , es

$$(2.5.3) \quad i_{\boldsymbol{\beta}} = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} / \sigma^2$$

Para las funciones de cuasi-verosimilitud esta matriz juega el mismo papel que la información de Fisher para las funciones de verosimilitud ordinarias. En particular, bajo las condiciones limitantes usuales sobre los eigenvalores de  $i_{\boldsymbol{\beta}}$ , la matriz de covarianzas asintótica de  $\hat{\boldsymbol{\beta}}$ , son

$$\text{cov}(\hat{\boldsymbol{\beta}}) \simeq i_{\boldsymbol{\beta}}^{-1} = \sigma^2 (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1}.$$

Comenzando con un valor arbitrario  $\hat{\boldsymbol{\beta}}_0$  suficientemente cercano a  $\hat{\boldsymbol{\beta}}$ , la sucesión de parámetros estimados generados por el método Newton-Raphson con el puntaje de Fisher estima a  $\hat{\boldsymbol{\beta}}$  iterando hasta alcanzar convergencia. Una propiedad importante de la sucesión es que no depende del valor de  $\sigma^2$ , similar a como ocurre en máxima verosimilitud.

En todos los aspectos de arriba la cuasi-verosimilitud se comporta justo como una verosimilitud ordinaria. No obstante, para la estimación de  $\sigma^2$ ,  $Q(\cdot; y)$  no se comporta como una log-verosimilitud. El estimador convencional de  $\sigma^2$  es un estimador por momentos basado en el vector de residuos  $\mathbf{Y} - \hat{\boldsymbol{\mu}}$ , a saber

$$\tilde{\sigma}^2 = \frac{1}{n-p} \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{V_i(\hat{\mu}_i)} = \frac{X^2}{n-p}$$

donde  $X^2$  es el estadístico de Pearson generalizado.



## 2.6. Bondad de ajuste

Dentro de la modelación estadística una de las etapas más importantes es la verificación de los supuestos que se hicieron al construir el modelo. Los modelos lineales generalizados no son la excepción. A continuación se presenta una forma de verificar los supuestos mediante un resumen de los residuos y algunas gráficas que ayudan a verificar supuestos tales como la función liga, la forma paramétrica de las variables, entre otras.

Uno de los criterios de bondad de ajuste más utilizados en los modelos lineales generalizados es la devianza residual, que proviene del cociente de verosimilitudes entre el modelo propuesto y el modelo saturado, es decir, el modelo cuyo estimador para la variable respuesta son las mismas observaciones de la variable respuesta sin ajuste alguno.

Sea  $l(\mu, \phi; y)$  la log-verosimilitud maximizada para  $\beta$  para algún valor fijo del parámetro de dispersión  $\phi$ . Sean  $\hat{\theta} = \theta(\hat{\mu})$  y  $\tilde{\theta} = \theta(y)$  estimadores del parámetro canónico del modelo saturado y del modelo propuesto, respectivamente. Suponiendo que  $a_i(\phi) = \phi/w_i$ , con  $w_i$  el peso inicial que podría variar de observación en observación, entonces, la devianza de la familia exponencial de dispersión se puede escribir como

$$(2.6.1) \quad D(y; \mu)\phi = \sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i)\}\phi$$

que corresponde a la devianza del modelo que se desea ajustar. En la siguiente tabla se muestra la forma de la devianza para algunas distribuciones importantes de la familia exponencial.

Normal	$\sum (y - \hat{\mu})^2$
Poisson	$2 \sum \left\{ y \log \left( \frac{y}{\hat{\mu}} \right) - (y - \hat{\mu}) \right\}$
Binomial	$2 \sum \left\{ y \log \left( \frac{y}{\hat{\mu}} \right) - (m - y) \log \left( \frac{m - y}{m(1 - \pi)} \right) \right\}$
(si $m = 1$ )	$-\sqrt{2}  \log(1 - \hat{\pi}) $
(si $m = 1$ y $y = 4$ )	$\sqrt{2}  \log(\hat{\pi}) $
Gamma	$2 \sum \left\{ y \log \left( \frac{y}{\hat{\mu}} \right) + \frac{y - \hat{\mu}}{\hat{\mu}} \right\}$
Gaussiana Inversa	$\sum \frac{(y - \hat{\mu})^2}{\hat{\mu}^2 y}$

CUADRO 3. Forma de la devianza para algunas distribuciones importantes de la familia exponencial.

Nos referimos a  $Q(\mu; y)$  como la cuasi-verosimilitud, o más correctamente, como la log cuasi-verosimilitud para  $\mu$  basada en los datos  $y$ . Puesto que las componentes de  $\mathbf{Y}$  son independientes por suposición, la cuasi-verosimilitud para los datos completos es la suma de las contribuciones individuales

$$Q(\mu; y) = \sum Q_i(\mu_i; y_i).$$

Por analogía, la función cuasi-devianza correspondiente a una sola observación es

$$(2.6.2) \quad D(y; \mu) = -2\sigma^2 Q(\mu; y) = 2 \int_{\mu}^y \frac{y-t}{V(t)} dt$$

la cual es estrictamente positiva excepto en  $y = \mu$ . La cuasi-devianza total  $D(y; \mu)$ , obtenida por sumar sobre los componentes, es una función calculable que depende de  $y$  y  $\mu$  únicamente: no depende de  $\sigma^2$ . Cabe mencionar que si  $\mu$  y  $V(\mu)$  corresponden a los de un miembro de la familia exponencial la cuasi-verosimilitud corresponde a máxima verosimilitud.

Otra estadística para conocer qué tan adecuado es el modelo que se propone respecto a lo observado es la  $\chi^2$  de Pearson generalizada

$$(2.6.3) \quad X^2 = \sum \frac{(y - \hat{\mu})^2}{V(\hat{\mu})}$$

donde  $V(\hat{\mu})$  es la varianza estimada para la distribución que se está tratando. Si la distribución del error es Normal, la  $X^2$  y la devianza tienen distribución  $\chi^2$  exacta; para el resto de las distribuciones se tiene un resultado asintótico (McCullagh y Nelder, [17]).

## 2.7. Residuos

En el caso de los modelos lineales generalizados se requiere extender la definición de residuo de modo que se pueda aplicar a todas las distribuciones que pertenecen a la familia exponencial de dispersión. Los residuos sirven para explorar si el modelo es adecuado, si la función liga es correcta, si los términos del predictor lineal son significativos, la elección de la función varianza, etc. A continuación se definen diferentes residuos y se muestra el uso gráfico de algunos de ellos.

Un elemento muy importante para el cálculo de residuos en el modelo de regresión clásico es la matriz de proyección  $\mathbf{H}$ , también conocida como matriz sombrero. En el caso de los modelos lineales generalizados también existe una matriz que cumple las propiedades de la matriz sombrero de los modelos lineales y viene dada por

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

donde

$$(2.7.1) \quad \mathbf{V} = \text{diag} \left[ V_i(\mu) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 a_i(\phi) \right]$$

Residuo puro. El residuo puro se define simplemente como

$$r_i = y_i - \hat{y}_i.$$

Si bien es el tipo de residuo más simple que hay, no resulta de mucha utilidad a la hora de verificar la validez de un modelo lineal generalizado y por tal motivo no son muy utilizados.

Residuo de Pearson. El residuo de Pearson, definido como

$$r_i^P = \frac{y_i - \hat{y}_i}{\sqrt{V(\mu_i)}}$$

se utiliza como medida de la variación residual en los modelos lineales generalizados. Se trata del residuo puro escalado por la desviación estándar estimada de  $Y$ . Su nombre se deriva del hecho que

$$\sum r_P^2 = X^2$$

donde  $X^2$  es el estadístico de prueba para bondad de ajuste de Pearson, también conocido como ji-cuadrada de Pearson.

Residuo de la devianza estandarizada. Estos residuos pueden indicar cuál de las observaciones contribuye 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  $\eta$  que maximiza la verosimilitud del modelo saturado,  $h_{ii}$  es el elemento  $i$ -ésimo de la diagonal de la matriz sombrero y  $\hat{\eta}_i$  pertenece al modelo propuesto.

En el caso de que los errores tengan distribución Normal, los residuos de la devianza estandarizada estarán dados por:

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

En el caso de la regresión logística, para la cual los errores tienen distribución Binomial, el residuo de la devianza estandarizada es

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

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

Residuo estudentizado. Se define como

$$r_i^e = \frac{\text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}\sqrt{\hat{\text{Var}}[Y_i]}}$$

(que también se conoce como residuo de Pearson estandarizado, debido a que  $\frac{(y_i - \hat{y}_i)^2}{\hat{\text{Var}}[Y_i]}$  es la contribución de la  $i$ -ésima observación a la estadística de Pearson).

Residuo de Anscombe. El residuo de Anscombe tiene la forma general

$$r_i^A = \frac{A(Y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}}$$

donde la función  $A(\cdot)$  está dada por  $A(z) = \int V^{-1/3}(z)dz$ . Para el modelo Poisson por ejemplo,

$$r_i^A = 3 \left( Y_i^{2/3} - \hat{\mu}_i^{2/3} \right) / \left( 2\hat{\mu}_i^{1/6} \right).$$

La función  $A(\cdot)$  es tal que los residuos se aproximan mejor a la normalidad.

## 2.8. Métodos gráficos de verificación de supuestos

Residuos no estandarizados contra residuos para  $\mathbf{u}$  (una variable omitida en el modelo). Esta gráfica conocida como variable-agregada se usa para ver si una variable  $\mathbf{u}$  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  $\mathbf{u}$  tomándola como respuesta en un modelo lineal usando los mismos predictores lineales y pesos cuadráticos como si se tratara de  $y$ . De esta forma, para poder decir que la omisión de  $\mathbf{u}$  fue correcta, los residuos no estandarizados de  $y$  graficados contra de los residuos de  $\mathbf{u}$  no deben mostrar ningún patrón identificable.

Variable dependiente ajustada  $\hat{y}$  contra el predictor lineal ajustado,  $\hat{\eta}$ . Si la gráfica resulta una línea recta significa que no hay patrón de comportamiento alguno, es decir, que la función liga es correcta, sin embargo, para datos binarios esta gráfica no es informativa por lo que se necesita usar métodos formales. Uno de estos métodos consiste en agregar  $\hat{\eta}^2$  como covariable extra y evaluar cuánto disminuye la devianza. Las validaciones sobre la conveniencia de la función liga elegida son afectadas inevitablemente por los errores que pudieran haberse cometido en la forma paramétrica de las variables explicativas que son parte del predictor lineal. Entonces, si esta prueba presenta desviaciones puede significar que la función liga es incorrecta o que la forma paramétrica de las variables explicativas es incorrecta, o, ambas. Para descifrar el problema es necesario apoyarse de las diferentes gráficas de residuos.

La gráfica del residuo parcial. Esta gráfica es una herramienta importante para determinar cuándo un término  $\beta x$  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

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

con  $\hat{\mathbf{y}}$  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 ser aproximadamente lineal. Si no, su forma podría servir como una alternativa recomendable.

Si se llegara a presentar incertidumbre en el patrón de comportamiento de la gráfica de residuos, un suavizamiento podría ayudar a descifrar tal comportamiento. En particular, la gráfica de residuos parciales suavizada puede ser de mucha utilidad para datos binarios. Sin embargo, pueden ocurrir distorsiones en la gráfica si la forma paramétrica de las otras variables explicativas es incorrecta.

El anterior es un breve resumen del amplio campo de los residuos y sus gráficas. El lector interesado en conocer más del tema puede revisar McCullagh, 1989 [17].

Puntos influyentes o puntos palanca. La distancia de Cook estima la influencia de un punto sobre los parámetros del predictor lineal. Por tal motivo se ha utilizado para indicar puntos particularmente importantes de revisar y validar (los puntos con una distancia de Cook grande deben ser revisados a detalle), para indicar regiones donde sería recomendable obtener más observaciones, o bien, para medir el efecto de borrar una observación dada. De tal manera que la distancia de Cook es útil para examinar la forma en que cada observación afecta al conjunto de parámetros estimados completo.

La siguiente ecuación permite la comparación entre los valores ajustados con la observación y sin ésta

$$C_i = \frac{1}{p} \left( \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} \right)^T \mathbf{X}^T \mathbf{V} \mathbf{X} \left( \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} \right)$$

donde  $\hat{\beta}_{(i)}$  es el parámetro estimado sin la  $i$ -ésima observación.

Para evitar hacer el ajuste en cada observación esta distancia se puede aproximar por:

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

donde  $p$  es el número de variables explicativas; esta última expresión se usa de manera más común en la gráfica contra índices (el orden en que fue tomada cada observación).

Las razones por las cuales el ajuste no es adecuado pueden ser diversas y una de ellas podría ser la existencia de un punto de cambio; problemática que se abordará más adelante.

## 2.9. El modelo de regresión logística

El modelo de regresión logística es, posiblemente, el MLG más utilizado en la práctica. En la década de los cincuenta se dió a conocer por sus aplicaciones a la bioestadística. Por aquellos años aún no se acuñaba el término *modelo lineal generalizado*. A continuación se desarrolla el modelo de regresión logística.

Sea  $Y$  una variable aleatoria discreta con soporte en  $0, 1, 2, \dots, n$ . Se dice que  $Y$  tiene distribución binomial si su función de densidad está dada por

$$f(y; n, p) = \binom{n}{y} p^y (1 - p)^{n-y},$$



se usará  $Y \sim B(n, p)$  para abreviar lo anterior. Cuando  $n = 1$ , la distribución binomial describe un *ensayo Bernoulli*. La función de densidad

$$f(y; p) = p^y(1 - p)^{1-y} = \begin{cases} 1 - p & \text{para } y = 0 \\ p & \text{para } y = 1 \end{cases}$$

se conoce como distribución Bernoulli, generalmente se abrevia como  $Y \sim Ber(p)$  o  $Y \sim B(1, p)$ . Si se tienen  $n$  ensayos Bernoulli independientes con probabilidad de éxito  $p$  en cada ensayo, la distribución binomial sirve para modelar el número total de éxitos en  $n$  ensayos Bernoulli. Si  $Y \sim B(n, p)$ , entonces

$$E[Y] = np \quad \text{y} \quad \text{Var}[Y] = np(1 - p).$$

Suponga que se realizan  $n$  experimentos independientes  $y_i$  cuya respuesta es cero o uno y depende de un conjunto de covariables  $x_1, x_2, \dots, x_p$ . Si  $y_i = 1$  indica éxito mientras que  $y_i = 0$  fracaso; estamos modelando la respuesta media  $\mathbb{P}(\mathbf{x}_i)$  donde  $\mathbb{P}(\mathbf{x}_i)$  es la probabilidad de éxito y  $\mathbf{x}_i$  son las covariables para la  $i$ -ésima observación. Entonces, para  $\mathbb{P}(\mathbf{x}_i)$  el modelo logístico está dado por

$$\mathbb{P}(\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i' \boldsymbol{\beta}}}$$

donde  $\mathbf{x}_i' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$  es el predictor lineal. Note que  $0 \leq \mathbb{P}(\mathbf{x}_i) \leq 1$ .

**2.9.1. La distribución binomial como modelo de dispersión exponencial.** Para cualquier  $n \in \mathbb{N}$  fija, la familia de distribuciones  $\{B(n, p)\}_{p \in (0,1)}$  es una familia exponencial con función liga canónica la función logit  $\theta = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{\mu}{1-\mu}\right)$ , función generadora de cumulantes  $b(\theta) = n \log(1 + \exp(\theta))$  y espacio parametral canónico  $\Omega_\beta = \mathbb{R}$ .

La función varianza unitaria es

$$V(\mu) = \mu(1-\mu)$$

que coincide con la varianza de una  $B(1, \mu)$ . Por otro lado, el valor medio de mapeo es

$$\mu(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = p$$

y por tanto el parámetro  $p$  es el valor medio de una unidad de observación. La devianza unitaria es

$$d(z; \mu) = 2 \left\{ z \log \left( \frac{z}{\mu} \right) + (1 - z) \log \left( \frac{1 - z}{1 - \mu} \right) \right\},$$

donde  $z = y/n$ , es decir, es el número de éxitos por el logaritmo del cociente del número de éxitos entre su valor esperado ( $np$ ) sumado con otro término similar pero para el número de fracasos (en cuyo caso el número esperado de fracasos es  $n(1 - p)$ ).

## 2.10. Modelo Poisson

Sea  $Y$  una variable aleatoria discreta que toma los valores  $0, 1, 2, \dots$ , si además su función de densidad es

$$f(y; \lambda) = \frac{\lambda^y}{y!} e^{-\lambda}$$

se dice que  $Y$  sigue una distribución Poisson de parámetro  $\lambda$ , con  $\lambda > 0$ . Para abreviar lo anterior escribimos  $Y \sim P(\lambda)$ . La media y la varianza de una  $P(\lambda)$  son

$$E[Y] = \lambda \quad \text{y} \quad \text{Var}[Y] = \lambda.$$

Sean  $y_1, y_2, \dots, y_n$  variables aleatorias independientes con distribución Poisson. Puesto que para la distribución Poisson la media y la varianza son iguales, entonces  $\mathbb{E}(y_i) = \mu_i$  y  $\text{Var}(y_i) = \mu_i$ . Un conjunto de covariables  $x_1, x_2, \dots, x_p$  tienen influencia sobre  $\mu$  via el modelo

$$\mu_i = e^{\mathbf{x}'_i \boldsymbol{\beta}}, \quad i = 1, 2, \dots, n$$

donde  $\mathbf{x}'_i = [1, x_{i1}, x_{i2}, \dots, x_{ik}]$ .

**2.10.1. La distribución Poisson como modelo de dispersión exponencial.** La función de densidad de la Poisson se puede escribir como

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} = \exp [y \log(\mu) - \mu - \log(y!)]$$

entonces se dice que la familia de distribuciones  $P(\mu)$  para  $\mu \in \mathbb{R}^+$  es una familia exponencial con función liga canónica la función logaritmo  $\theta = \log(\mu)$ , espacio parametral canónico  $\Omega_\beta = \mathbb{R}$ , función generadora de cumulantes  $b(\theta) = \exp(\theta)$ , valor medio de mapeo  $\mu(\theta) = \exp(\theta)$  y  $c(y, \phi) = -\log(y!)$ .

La devianza unitaria correspondiente a la distribución Poisson es

$$d(y; \mu) = 2 \left\{ y \log \left( \frac{y}{\mu} \right) \right\}$$

observe que no se incluye ningún parámetro índice aquí. La familia de distribuciones Poisson es el único modelo de dispersión exponencial con esta propiedad. Este fenómeno puede explicarse por el hecho de que la función varianza  $V(\mu)$  para la distribución Poisson es la identidad.

A continuación se muestra un caso práctico para el cual se utilizó el modelo Poisson.

**2.10.2. Ejemplo: modelación de tasas de mortalidad.** Dentro de una población grande la probabilidad de elegir aleatoriamente a un individuo agonizante en un momento particular es pequeña. Si suponemos que las muertes por una enfermedad no infecciosa son eventos independientes entonces  $Y$ , el número de muertes en una población, puede ser modelado por una distribución Poisson

$$f(y; \mu) = \frac{\mu^y e^{-\mu}}{y!}$$

donde  $y$  puede tomar los valores  $0, 1, 2, \dots$  y  $\mu = \mathbb{E}(Y)$  es el número esperado de muertes en un período de tiempo específico, por ejemplo, un año. El parámetro  $\mu$  depende del tamaño de la población, el período de observación y características de la población como edad, sexo e historial clínico del paciente. Puede ser modelado de la siguiente manera

$$\mathbb{E}(Y) = \mu = n\lambda(\mathbf{x}'\boldsymbol{\beta})$$

donde  $n$  es el tamaño de la población y  $\lambda(\mathbf{x}'\boldsymbol{\beta})$  es la tasa por cada 100,000 habitantes en un año, la cual depende de las características de la población descritas por el componente lineal  $\mathbf{x}'\boldsymbol{\beta}$ .

Los cambios en la mortalidad de acuerdo a la edad pueden modelarse tomando las variables aleatorias independientes  $Y_1, \dots, Y_n$  como el número de muertes ocurridas en los grupos de edad. El Cuadro (2.10.2) muestra los datos de muerte por enfermedad de coronaria por grupo quinquenal.

La Figura (2.10.1) muestra cómo la tasa de mortalidad  $y_i/n_i \times 100,000$  incrementa con respecto a la edad. Se utilizó la escala logarítmica en el eje vertical. En esta escala la gráfica de dispersión es aproximadamente lineal sugiriendo que la relación

Grupo de edad (años)	Número de muertes, $y_i$	Tamaño de población, $n_i$	Tasa anual por cada 100,000 hombres, $y_i/n_i \times 100,000$
30-34	1	17,742	5.6
35-39	5	16,554	30.2
40-44	5	16,059	31.1
45-49	12	13,083	91.7
50-54	25	10,784	231.8
55-59	38	9,645	394.0
60-64	54	10,706	504.4
65-69	65	9,933	654.4

CUADRO 4. Número de muertes en hombres por enfermedad coronaria y tamaño de población por grupo quinquenal del Valle Hunter, en Nueva Gales del Sur, Australia, 1991.

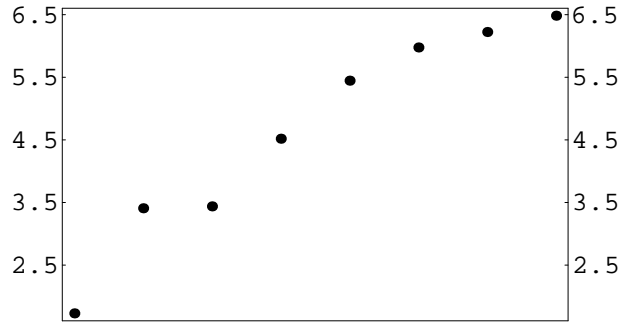


FIGURA 2.10.1. Tasa de mortalidad

entre  $y_i/n_i$  y el grupo de edad  $i$  es aproximadamente exponencial. Por tanto, un posible modelo es

$$\mathbb{E}(Y_i) = \mu_i = n_i e^{\theta i}, \quad Y_i \sim \text{Poisson}(\mu_i)$$

donde  $i = 1$  para el grupo de edad 30-34 años,  $i = 2$  para 35-39, ...,  $i = 8$  para 65-69 años.

El modelo anterior puede escribirse como un modelo lineal generalizado usando la función liga logarítmica

$$g(\mu_i) = \log(\mu_i) = \log(n_i) + \theta$$

la cual utiliza el componente lineal  $\mathbf{x}^T \boldsymbol{\beta}$  con  $\mathbf{x}_i^T = [\log(n_i), i]$  y  $\boldsymbol{\beta} = \begin{bmatrix} 1 \\ \theta \end{bmatrix}$ , y en este caso  $\log(n_i)$  se conoce como *offset*, una variable que se encuentra en el componente sistemático con coeficiente 1.

De esta manera concluimos nuestra revisión de la teoría de los modelos lineales generalizados, la cual espero brinde un panorama amplio de lo que son estos modelos, cómo funcionan, pero también de cómo pueden aplicarse a la solución de problemas reales en diferentes áreas de estudio.



## Puntos de cambio en modelos lineales

### 3.1. Introducción

El análisis de regresión es una de las técnicas estadísticas más empleadas por otras disciplinas. En ciertas aplicaciones los estadísticos se enfrentaron con algunos problemas al intentar establecer modelos de regresión sobre ciertos conjuntos de datos. En ocasiones utilizar un modelo lineal clásico para explicar datos cuya estructura se sospecha cambió en determinado punto del tiempo, los deja desajustados o pobremente explicados. Debido a esta dificultad se originó la hipótesis de punto de cambio.

El problema del punto de cambio consiste en decidir si ocurrió un cambio estructural en el período de observación y en caso que así fuera, determinar dónde ocurrió. Entonces, surge la pregunta: ¿cómo detectar el punto de cambio?

Desde hace algunas décadas, diversos autores han estudiado el problema de punto de cambio. Por ejemplo Quandt (1958 [1], 1960 [3]), que propuso una prueba basada en un cociente de log-verosimilitudes para validar y estimar un modelo lineal que obedece dos regímenes separados. Sen y Srivastava (1973) estudiaron el problema de punto de cambio simple para el vector de medias de una sucesión de observaciones Gaussianas multivariadas. Por su parte, Brown, Durbin y Evans (1975), introdujeron un método de residuales recursivos para probar puntos



de cambio en modelos de regresión múltiple. Años después, Srivastava y Worsley (1986) propusieron una estadística de prueba para múltiples puntos de cambio y obtuvieron su distribución. Hawkins (1989) utilizó el enfoque de unión e intersección para probar cambios en un modelo de regresión lineal. Kim (1994) propuso una prueba para un punto de cambio en una regresión lineal por medio de la estadística de cociente de verosimilitudes y estudió el comportamiento asintótico de la misma.

### 3.2. Cociente de log-verosimilitudes

Esta técnica descrita por Quandt en dos artículos, el primero de 1958 [1] y el segundo de 1960 [3], puede utilizarse cuando se sospecha que ocurrió un cambio abrupto en la regresión en un punto desconocido del tiempo  $t = i$  pasando de una regresión especificada por  $\beta_1$  y  $\sigma_1^2$ , a otra especificada por  $\beta_2$  y  $\sigma_2^2$ . Para cada  $i$  desde  $i = p + 1$  hasta  $i = n - p - 1$  se debe calcular y graficar

$$(3.2.1) \quad \lambda_i = \log \left( \frac{\max \text{verosimilitud de las observaciones dado } H_0}{\max \text{verosimilitud de las observaciones dado } H_1} \right)$$

donde  $H_1$  es la hipótesis de que las observaciones en los segmentos de tiempo  $(1, \dots, i)$  y  $(i + 1, \dots, n)$  se pueden separar en dos regresiones diferentes. La estadística  $\lambda_i$  es la estadística estándar de cociente de verosimilitudes para decidir entre dos hipótesis  $H_0$  y  $H_1$ , y se puede probar que bajo normalidad

$$\lambda_i = \frac{i}{2} \log \hat{\sigma}_1^2 + \frac{n-i}{2} \log \hat{\sigma}_2^2 - \frac{n}{2} \log \hat{\sigma}^2,$$

donde  $\hat{\sigma}_1^2$ ,  $\hat{\sigma}_2^2$  y  $\hat{\sigma}^2$  son el cociente de la suma de cuadrados de los residuos entre el número de observaciones cuando se ajusta una regresión con las primeras  $i$  observaciones, con las  $n - i$  observaciones restantes y con la totalidad de las observaciones, respectivamente. La estimación del punto en el cual ha ocurrido el cambio de una relación a otra es el valor de  $i$  para el cual  $\lambda_i$  alcanza el mínimo.

Por otro lado, puesto que se desconoce la distribución de  $\min \lambda_i$  bajo  $H_0$ , no se cuenta con una prueba para  $\min \lambda_i$ . Sin embargo, el comportamiento de la gráfica de  $\lambda_i$  contra  $i$  permite obtener información sobre la estabilidad de la regresión e indica si han ocurrido cambios tales como una transición abrupta.

### 3.3. El enfoque basado en la información

En 1973, Hirotugu Akaike introdujo el Criterio de Información de Akaike, también conocido como AIC por sus siglas en inglés, para la selección de modelos en estadística (Akaike, 1973 [5]). Desde entonces, este criterio ha influenciado profundamente el desarrollo de la estadística, particularmente en series de tiempo, análisis de observaciones atípicas (Kitagawa, 1979 [11]), robustez, análisis de regresión, análisis multivariado (Bozdogan, Sclove y Gupta, 1994 [19]), etcétera. Más tarde, sobre la base del trabajo de Akaike, otros autores introdujeron criterios de información que utilizaron en campos tales como la econometría, la psicometría, la teoría de control y la teoría de las decisiones.

Akaike (1973), propuso el criterio de información

$$AIC(i) = -2 \log L(\hat{\Theta}_i) + 2i, \quad i = 1, 2, \dots, p,$$

como medida de evaluación de un modelo, con  $L(\hat{\Theta}_i)$  la función de verosimilitud del modelo con  $i = 1, \dots, p$  parámetros. Un modelo que minimiza el AIC (mínimo AIC estimado) se considera el modelo más apropiado. Se han hecho otras propuestas de criterios de información y una de ellas es el Criterio de Información de Schwarz, también conocido como SIC por sus siglas en inglés, el cual fue propuesto por Schwarz en 1978.

El SIC se expresa de la siguiente manera:

$$SIC(i) = -2 \log L(\hat{\Theta}_i) + i \log(n), \quad i = 1, 2, \dots, p$$

donde  $L(\hat{\Theta}_i)$  es la función de verosimilitud del modelo,  $i$  es el número de parámetros del modelo y  $n$ , el tamaño de muestra. Entonces, se considera más apropiado el modelo que minimiza el SIC.

Una forma de abordar el problema del punto de cambio es mediante el Criterio de Información de Schwarz, o SIC. Bajo este enfoque se reducen mucho los cálculos en comparación con otros enfoques, como el clásico de cociente de verosimilitudes.

Considere el siguiente modelo:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n$$

donde  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , es un vector no aleatorio de dimensión  $p + 1$  con  $\mathbf{x}_i^T = (1, x_{1i}, x_{2i}, \dots, x_{pi})$ ,  $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$  un vector de parámetros de regresión desconocido de dimensión  $p + 1$ , y por último,  $\epsilon_i$ ,  $i = 1, \dots, n$  es un error aleatorio con distribución  $N(0, \sigma^2)$  donde  $\sigma^2$  desconocida y las  $\epsilon_i$  son no correlacionadas observación a observación. Es decir,  $y_i$  es una variable aleatoria independiente con distribución  $N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$  para  $i = 1, \dots, n$ .

Se busca probar si existe un punto de cambio en el modelo, para ello se contrastarán las hipótesis nula:

$$H_0 : \mu_{y_i} = \mathbf{x}_i^T \boldsymbol{\beta} \quad i = 1, \dots, n$$

contra la alternativa:

$$H_1 : \mu_{y_i} = \mathbf{x}_i^T \boldsymbol{\beta}_1 \quad i = 1, \dots, k$$

$$\text{y } \mu_{y_i} = \mathbf{x}_i^T \boldsymbol{\beta}_2 \quad i = k + 1, \dots, n$$

donde  $k$  es la localización del punto de cambio y  $\boldsymbol{\beta}$ ,  $\beta_1$  y  $\beta_2$  son parámetros de regresión desconocidos.

Sean

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

y

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}$$

entonces la hipótesis nula  $H_0$  corresponde al modelo  $\mu_y = \mathbf{X}\boldsymbol{\beta}$ , donde

$$\mu_y = \begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \vdots \\ \mu_{y_n} \end{pmatrix}.$$

Bajo  $H_0$  los estimadores máximo verosímiles para  $\boldsymbol{\beta}$  y  $\sigma^2$  son

$$\begin{aligned} \mathbf{b} = \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ \hat{\sigma}^2 &= \frac{1}{n} (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) \end{aligned}$$

respectivamente. Por lo tanto, bajo  $H_0$  el Criterio de Información de Schwarz, denotado por  $SIC(n)$ , se obtiene como:

$$\begin{aligned} SIC(n) &= -2 \log L_0(\mathbf{b}, \hat{\sigma}^2) + (p+2) \log n \\ &= n \log [(\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})] + n(\log(2\pi) + 1) + (p+2-n) \log n, \end{aligned}$$

donde  $L_0(\mathbf{b}, \hat{\sigma}^2)$  es la función de máxima verosimilitud bajo  $H_0$ .

Sean

$$\mathbf{y}_1 = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} y_{k+1} \\ y_{k+2} \\ \vdots \\ y_n \end{pmatrix}$$

$$\mathbf{X}_1 = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1k} & \cdots & x_{pk} \end{pmatrix}$$

$$\mathbf{X}_2 = \begin{pmatrix} 1 & x_{1(k+1)} & \cdots & x_{p(k+1)} \\ 1 & x_{1(k+2)} & \cdots & x_{p(k+2)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}$$

$$\boldsymbol{\beta}_1 = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\beta}_2 = \begin{pmatrix} \beta_0^* \\ \beta_1^* \\ \vdots \\ \beta_p^* \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

entonces la hipótesis alternativa  $H_1$  corresponde a los modelos  $\mu_{y_1} = \mathbf{X}_1\boldsymbol{\beta}_1$  y  $\mu_{y_2} = \mathbf{X}_2\boldsymbol{\beta}_2$  para  $k = p + 1, \dots, n - p$  donde

$$\mu_{y_1} = \begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \vdots \\ \mu_{y_k} \end{pmatrix}, \quad y \quad \mu_{y_2} = \begin{pmatrix} \mu_{y_{k+1}} \\ \mu_{y_{k+2}} \\ \vdots \\ \mu_{y_n} \end{pmatrix}$$

En este caso, los estimadores máximo verosímiles de los parámetros son

$$(3.3.1) \quad \mathbf{b}_1 = \hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{y}_1$$

$$(3.3.2) \quad \mathbf{b}_2 = \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T \mathbf{y}_2$$

$$(3.3.3) \quad \hat{\sigma}^2 = \frac{1}{n} [(\mathbf{y}_1 - \mathbf{X}_1 \mathbf{b}_1)^T (\mathbf{y}_1 - \mathbf{X}_1 \mathbf{b}_1) + (\mathbf{y}_2 - \mathbf{X}_2 \mathbf{b}_2)^T (\mathbf{y}_2 - \mathbf{X}_2 \mathbf{b}_2)]$$

respectivamente (suponemos que los modelos son independientes). Por lo tanto, bajo  $H_1$  el criterio de información de Schwarz, denotado por  $SIC(k)$ , se obtiene como

$$\begin{aligned} SIC(k) &= -\log L_1(\mathbf{b}_1, \mathbf{b}_2, \hat{\sigma}^2) + (2p + 3) \log n \\ &= n \log [(\mathbf{y}_1 - \mathbf{X}_1 \mathbf{b}_1)^T (\mathbf{y}_1 - \mathbf{X}_1 \mathbf{b}_1) + (\mathbf{y}_2 - \mathbf{X}_2 \mathbf{b}_2)^T (\mathbf{y}_2 - \mathbf{X}_2 \mathbf{b}_2)] \\ &\quad + n(\log 2\pi + 1) + (2p + 3 - n) \log n, \end{aligned}$$

donde  $L_1(\mathbf{b}_1, \mathbf{b}_2, \hat{\sigma}^2)$  es la función de máxima verosimilitud bajo  $H_1$ .

De acuerdo con el principio de información para la selección del modelo, no se rechaza  $H_0$  si

$$SIC(n) \leq \min\{SIC(k) : 2 \leq k \leq n - 2\}$$

y se rechaza  $H_0$  si

$$SIC(n) > \min\{SIC(k) : 2 \leq k \leq n - 2\}.$$

Cuando se rechaza  $H_0$ , la posición estimada  $\hat{k}$  será tal que

$$SIC(\hat{k}) = \min\{SIC(k) : 2 \leq k \leq n - 2\}.$$

**3.3.1. Ejemplo: volumen de ventas en mercado de valores.** Considere el siguiente conjunto de datos sobre volumen de ventas del mercado de valores (Holbert, 1982 [13]). Se utilizarán estos datos para ilustrar el método SIC para localizar el punto de cambio en una regresión lineal. El volumen de ventas de dólar mensual (cifras en millones) de la Bolsa de Valores de Boston (BSE, por sus siglas en inglés) se considera la variable de respuesta, y la Bolsa de Valores de Nueva York, el predictor (NYAMSE, por sus siglas en inglés). Los valores SIC calculados se muestran en el Cuadro 1 junto con los valores originales del BSE y el NYAMSE dados por Holbert. El valor SIC con el asterisco (ver Cuadro 1) es el menor valor SIC y corresponde al punto 23, por lo tanto el cambio en el modelo comienza a partir del punto 24 que corresponde a diciembre de 1968. Este resultado coincide con el obtenido por Holbert [13].

CUADRO 1. Valores de la NYAMSE y del BSE y valores SIC calculados

Punto	Fecha	NYAMSE	BSE	SIC
1	Ene. 1967	10581.6	78.8	-
2	Feb. 1967	10234.3	69.1	368.5736
3	Mar. 1967	13299.5	87.6	368.0028
4	Abr. 1967	10746.5	72.8	367.9975
5	May. 1967	13310.7	79.4	366.8166
6	Jun. 1967	12835.5	85.6	366.1827
7	Jul. 1967	12194.2	75.0	365.3197
8	Ago. 1967	12860.4	85.3	364.4143
9	Sep. 1967	11955.6	86.9	364.0418
10	Oct. 1967	13351.5	107.8	364.0670
11	Nov. 1967	13285.9	128.7	365.1320
12	Dic. 1967	13784.4	134.5	365.8783
13	Ene. 1968	16336.7	148.7	365.7791
14	Feb. 1968	11040.5	94.2	366.0318
15	Mar. 1968	11525.3	128.1	367.1252
16	Abr. 1968	16056.4	154.1	367.2805
17	May. 1968	18464.3	191.3	367.4632
18	Jun. 1968	17092.2	191.9	367.6615
19	Jul. 1968	15178.8	159.6	367.8082
20	Ago. 1968	12774.8	185.5	368.8873
21	Sep. 1968	12377.8	178.0	368.9790
22	Oct. 1968	16856.3	271.8	364.2126
23	Nov. 1968	14635.3	212.3	359.3774*
24	Dic. 1968	17436.9	139.4	362.7803
25	Ene. 1969	16482.2	106.0	366.7591
26	Feb. 1969	13905.4	112.1	367.4118
27	Mar. 1969	11973.7	103.5	367.6757
28	Abr. 1969	12573.6	92.5	368.4138
29	May. 1969	16566.8	116.9	370.6948
30	Jun. 1969	13558.7	78.9	372.0000
31	Jul. 1969	11530.9	57.4	372.1517
32	Ago. 1969	11278.0	75.9	371.8513
33	Sep. 1969	11263.7	109.8	372.1726
34	Oct. 1969	15649.5	129.2	-
35	Nov. 1969	12197.1	115.1	361.4956

La gráfica de la Figura 3.3.1 muestra la dispersión del BSE contra el NYAMSE.

Los círculos indican las parejas de (BSE, NYAMSE) antes de diciembre de 1968



y las estrellas indican las parejas de (BSE, NYAMSE) en y después de diciembre de 1968. En la gráfica, las dos líneas de regresión (antes y después del punto de cambio) también son graficadas; la línea intermitente indica antes del cambio mientras que la línea punteada indica la regresión después del punto de cambio.

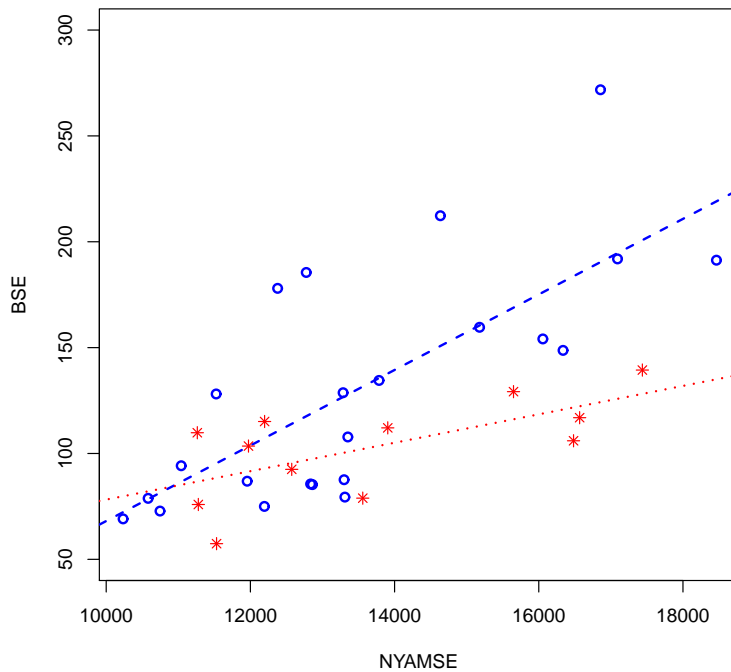


FIGURA 3.3.1. Punto de cambio con datos de la Bolsa de Valores

Lo anterior nos indica que cuando la inferencia de un punto de cambio es vista como un problema de selección de modelo, el SIC, o cualquier otro criterio de selección de modelo bueno, provee una manera efectiva de localizar el punto de cambio.

Sobre este mismo conjunto de datos se aplicó la prueba del cociente de verosimilitudes de Quandt y se obtuvo el mismo resultado que con la prueba SIC. En

[28], Zeileis *et al.* presentan el paquete de R `strucchange` el cual contiene diversas rutinas para detectar cambio estructural en modelos de regresión lineal. El paquete `strucchange`, contiene la función `Fstats`, la cual calcula el cociente de las log-verosimilitudes bajo  $H_0$  y  $H_1$  y devuelve la gráfica de las  $\lambda_i$  (ver fórmula 3.2.1); con esta misma gráfica es posible identificar el punto del tiempo en el cual ocurrió un punto de cambio. A diferencia de lo descrito en la Sección 3.2 sobre esta prueba, con la función `Fstats` de `strucchange` el punto de cambio se identifica como el punto donde se alcanza el máximo. Usando los datos del ejemplo anterior, el máximo se alcanza en el período 23, que corresponde a la fecha noviembre de 1969, dato que es consistente con el método que se revisó anteriormente.

```
# Se carga el paquete strucchange
require(strucchange)

# Se cargan los datos
datos <- read.csv("datos_nyamse_bse.csv")
datos <- ts(datos)

# Se realiza la prueba de Quandt
qlr <- Fstats(BSE ~ NYAMSE, data=datos)

# Se grafican los datos
plot(qlr, alpha=0.05, xlab="Fecha", lwd=3, lty=3)
text(6, 1.6, "may._1967")
text(23, 11.1, "nov._1968")
text(28, 0.35, "jun._1969")
text(23, 0.5, "23")
abline(v=23, lty=3, lwd=1.5)
```

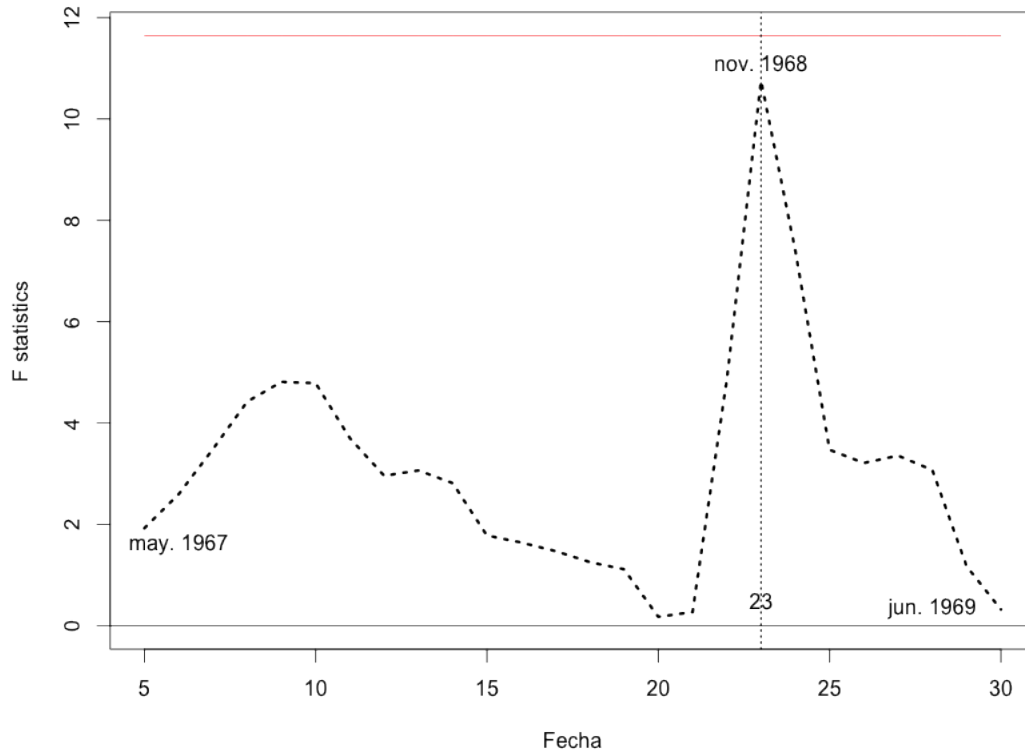


FIGURA 3.3.2. Resultado de la prueba de cociente de verosimilitudes de Quandt

### 3.4. Prueba de Chow

La prueba de Chow [2] es otro método muy utilizado para determinar si ha ocurrido un punto de cambio en un conjunto de datos. Esta prueba se utiliza cuando existe la sospecha de que el modelo al que corresponde una parte es distinto del resto de la muestra. Este método supone que la muestra se puede dividir en dos submuestras independientes una de la otra de tamaños  $k$  y  $n - k$ , respectivamente.

El modelo de regresión asociado a la existencia del cambio estructural es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon \quad i = 1, \dots, k$$

$$y_i = \beta_0^* + \beta_1^* x_{1i} + \beta_2^* x_{2i} + \cdots + \beta_p^* x_{pi} + \epsilon \quad i = k + 1, \dots, n$$

con  $\beta_i \neq \beta_i^*$  para alguna  $i = 1, \dots, p$ ,  $y \epsilon \sim N(0, \sigma^2)$ , por lo que las hipótesis de prueba del modelo son

$$H_0 : (\beta_0, \beta_1, \dots, \beta_p) = (\beta_0^*, \beta_1^*, \dots, \beta_p^*)$$

$$H_1 : (\beta_0, \beta_1, \dots, \beta_p) \neq (\beta_0^*, \beta_1^*, \dots, \beta_p^*)$$

La estadística de prueba bajo normalidad es la siguiente:

$$F_c = \frac{(SCE - (SCE_{1k} + SCE_{2k})) / (p + 1)}{(SCE_{1k} + SCE_{2k}) / (n - 2(p + 1))} \sim F_{p+1, n-2(p+1)}$$

y la región de rechazo para  $H_0$  es  $F_c > F_{\alpha, p+1, n-2(p+1)}$ .

La característica principal de la prueba de Chow es que se conoce el momento en que ocurre el posible cambio, pero, ¿qué pasa cuando no se conoce  $k$ ? Incluso podría pasar que se desconoce si  $k$  está en la región de observación, es decir  $1 \leq k \leq n$ . En este caso las hipótesis de prueba son:

$$H_0 : k \geq n \quad \text{vs.} \quad H_1 : k < n$$

Bajo el supuesto de que la hipótesis nula es cierta, el modelo seguido por las observaciones es

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

para toda  $i = 1, 2, \dots, n$ , y entonces no existe punto de cambio en la región de observación. Bajo el supuesto de que la hipótesis nula es falsa el modelo resulta

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \epsilon_i & \text{si } i \leq k \\ \beta_0^* + \beta_1^* x_{1i} + \beta_2^* x_{2i} + \dots + \beta_p^* x_{pi} + \epsilon_i & \text{si } i > k \end{cases}$$

con  $k < n$  y  $\beta_i \neq \beta_i^*$  para alguna  $0 \leq i \leq p$ , lo cual significa que se manifestó un punto de cambio en la región de observación.

En la Sección 3.3, se utilizó información del precio de venta del dolar en dos mercados de valores diferentes para estimar el punto de cambio. Con estos datos se llevó a cabo la prueba de Chow para el período 23 (noviembre de 1968), que de acuerdo con el criterio SIC ahí ocurrió el punto de cambio.

Dentro de las pruebas que están contenidas en el paquete de R `strucchange` (Zeileis *et al.* [28]) se encuentra la de Chow.

*# Se carga el paquete strucchange*

`require(strucchange)`

*# Se cargan los datos*

`datos = read.csv("datos_nyamse_bse.csv")`

*# Se realiza la prueba de Chow*

`sctest(BSE ~ NYAMSE, data=datos, type="Chow", point=23)`

Chow test

`data: BSE ~ NYAMSE`

F = 5.376, p-value = 0.0099

Con base en el resultado de la prueba, con un p-value de 0.0099 se rechaza la hipótesis nula y se acepta la de que en el período seleccionado (nov. 1968) ocurrió un cambio estructural o punto de cambio en la relación entre las variables BSE (respuesta) y NYAMSE (explicativa).

### 3.5. Pruebas basadas en la suma acumulada de los residuos

Suponga un modelo con  $n$  observaciones en total. El residuo recursivo  $i$ -ésimo es el error de predicción ex-post (para instantes temporales pasados) de  $y_i$  cuando la regresión se estima utilizando solamente las  $i - 1$  primeras observaciones

$$e_i = y_i - \mathbf{x}_i^T \mathbf{b}_{i-1}$$

donde  $\mathbf{x}_i$  es el vector de regresores asociado con la observación  $y_i$  y  $\mathbf{b}_{i-1}$  es el coeficiente de mínimos cuadrados calculado utilizando  $i - 1$  observaciones. La varianza predicha de este residuo es

$$\sigma_{fi}^2 = \sigma^2 \left[ 1 + \mathbf{x}_i^T (\mathbf{X}_{i-1}^T \mathbf{X}_{i-1})^{-1} \mathbf{x}_i \right]$$

Sea el residuo  $r$ -ésimo

$$w_r = \frac{e_r}{\sqrt{1 + \mathbf{x}_r^T (\mathbf{X}_{r-1}^T \mathbf{X}_{r-1})^{-1} \mathbf{x}_r}},$$

donde  $\mathbf{X}_{r-1}$  es la matriz particionada formada por las primeras  $r - 1$  observaciones. Bajo la hipótesis de que los coeficientes permanecen constantes durante todo el período muestral resulta que  $w_r \sim N(0, \sigma^2)$  y es independiente de  $w_s$  para todo  $s \neq r$ . En este caso la evidencia de que la distribución de  $w_r$  cambia con el tiempo se contrapone a la hipótesis de estabilidad del modelo.

**3.5.1. Prueba CUSUM.** Brown *et al.* [9], sugirieron dos contrastes basados en  $w_r$ . Al primero lo llamaron CUSUM y está basado en la suma acumulada de los residuos

$$W_i = \sum_{r=k+1}^i \frac{w_r}{\hat{\sigma}^2},$$

donde

$$\hat{\sigma}^2 = \frac{\sum_{r=k+1}^n (w_r - \bar{w})^2}{n - k - 1}$$

y

$$\bar{w} = \frac{\sum_{r=k+1}^n w_r}{n - k}$$

Si  $\beta_i$  es constante hasta el tiempo  $i = i_0$  y difiere de ahí en adelante,  $w_r$  tendrá media cero desde  $r = 1$  hasta  $i_0$ , pero en general no tendrá media cero. Lo anterior sugiere revisar algunas gráficas para determinar posibles desviaciones de las  $w_r$  cuando éstas se mueven junto con la serie a través del tiempo. La primera gráfica a considerar es la de CUSUM

$$W_i = \sum_{r=k+1}^i \frac{w_r}{\hat{\sigma}^2}$$

contra  $r$  para  $r = k + 1, \dots, n$ , donde  $\hat{\sigma}^2 = \frac{S_T}{n - k}$ . Es necesario un método para probar la significancia de la desviación de la trayectoria  $W_i$  de su línea de valor medio  $\mathbb{E}(W_i) = 0$ . Un procedimiento suficientemente adecuado es encontrar un par de líneas simétricas arriba y abajo de la línea  $W_i = 0$  tales que la probabilidad de cruzar una o ambas líneas sea  $\alpha$ , el nivel de significancia requerido.

De las propiedades de las  $w_i$  bajo  $H_0$ , la sucesión  $W_{k+1}, \dots, W_i$  es una sucesión de variables aleatorias aproximadamente normales tales que

$$\mathbb{E}(W_i) = 0, \quad \text{Var}(W_i) = i - k \quad \text{y} \quad \text{Cov}(W_i, W_j) = \text{mín}(i, j) - k,$$

para una buena aproximación. Para derivar la prueba,  $W_i$  es aproximada por un proceso Gaussiano continuo  $\{Z_t, k \leq t \leq n\}$  con estas funciones media y covarianza. De hecho, se trata de un movimiento browniano que comienza en 0 al tiempo  $t = k$ . La desviación estándar de  $Z_t$  es  $\sqrt{t - k}$ . En consecuencia, si se quiere encontrar una curva tal que bajo  $H_0$  la probabilidad de que la trayectoria esté por encima de la curva en cualquier punto entre  $i = k$  e  $i = n$  sea constante, se deben elegir curvas de la forma  $\pm\lambda\sqrt{i - k}$  con  $\lambda$  constante. No obstante, dado que se desea limitar a líneas rectas, se adopta una familia de rectas tangentes a las curvas  $\pm\lambda\sqrt{i - k}$  en los puntos medios entre  $i = k$  e  $i = n$ . Lo anterior conduce a la familia de parejas de líneas rectas a través de los puntos  $\{k, \pm a\sqrt{n - k}\}$ ,  $\{n, \pm 3a\sqrt{n - k}\}$ , donde  $a$  es un parámetro. De manera que para cualquier línea dada en esta familia, la probabilidad de que el punto  $(i, W_i)$  caiga fuera de la línea es un máximo para  $i$ , punto medio entre  $i = k$  e  $i = n$ .

Se quiere encontrar un miembro de esta familia tal que la probabilidad de que la trayectoria  $Z_t$  lo cruce es  $\alpha/2$ . Ciertos resultados de la teoría del movimiento browniano dan el valor de

$$Q\left(\frac{d + c(n - k)}{\sqrt{n - k}}\right) + \exp(-2dc) Q\left(\frac{d - c(n - k)}{\sqrt{n - k}}\right),$$

para la probabilidad de que la trayectoria  $Z_t$  cruce la línea  $y = d + c(i - k)$  para alguna  $i$  en  $(k, n)$ , donde

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} \exp\left(-\frac{u^2}{2}\right) du$$

(ver Durbin, 1971, Lema 3). Si se substituyen los valores  $d = a\sqrt{n - k}$  y  $c = \frac{2a}{\sqrt{n - k}}$  se obtiene la ecuación

$$Q(3a) + \exp(-4a^2) (1 - Q(a)) = \frac{\alpha}{2},$$



que debe ser resuelta para  $\alpha$ . Parejas de valores útiles para  $a$  y  $\alpha$  son

$$\alpha = 0.01, \quad a = 1.143$$

$$\alpha = 0.05, \quad a = 0.948$$

$$\alpha = 0.10, \quad a = 0.850.$$

La hipótesis nula se rechaza si  $W_i$  sale fuera de los límites.

**3.5.2. Prueba de cuadrados de CUSUM.** Una prueba similar se basa en los cuadrados de los residuos recursivos. El contraste de cuadrados CUSUMQ de CUSUM utiliza

$$S_i = \frac{\sum_{r=k+1}^i w_r^2}{\sum_{r=k+1}^n w_r^2} = \frac{S_r}{S_T}, \quad r = k + 1, \dots, T.$$

Como los residuos son independientes, cada uno de los dos términos es aproximadamente una suma de variables chi-cuadrado con un grado de libertad. Por lo tanto,  $\mathbb{E}[S_i]$  es aproximadamente  $\frac{i-k}{n-k}$ . El contraste se lleva a cabo mediante la construcción de bandas para  $\mathbb{E}[S_i]$  para valores de  $i$  y trazando gráficamente  $S_i$  y las bandas en función de  $i$ .

Las bandas apropiadas son  $\mathbb{E}[S] \pm c_0$ , donde  $c_0$  depende de  $(n-k)$  y del nivel de significancia deseado. Al igual que antes, si la suma acumulada se extravía fuera de las bandas de confianza, surge la duda sobre la hipótesis de estabilidad de los parámetros. Esta prueba resulta un complemento útil para la prueba CUSUM, particularmente cuando la desviación de la  $\beta$  es más casual que sistemática.

Se puede probar, aunque aquí no lo haremos, que bajo  $H_0$  las  $s_i$  siguen una distribución beta con media  $\frac{i-k}{n-k}$ . Esto sugiere dibujar un par de líneas  $s_i = \frac{i-k}{n-k} \pm c_0$

sobre el diagrama paralelo a la línea de valor medio tal que la probabilidad de que la trayectoria cruce una o ambas líneas sea  $\alpha$ , el nivel de significancia requerido.

Para encontrar los niveles de significancia  $c_0$ , resulta conveniente considerar primero el caso cuando  $n - k$  es par. Entonces, la distribución conjunta de las  $\frac{n - k}{2} - 1$  estadísticas  $s_{k+2}, s_{k+4}, \dots, s_{n-2}$ , es la misma que la de una muestra ordenada del mismo tamaño de observaciones independientes de una uniforme  $U(0, 1)$ . Lo anterior puede probarse escribiendo

$$n = \frac{T - k}{2} - 1 \quad \text{y} \quad z_j = \frac{w_{k+2j}^2 + w_{k+2j-1}^2}{2\sigma^2}, \quad j = 1, \dots, n + 1.$$

De este modo, resulta que las  $z_j$  son variables aleatorias independientes exponenciales con media uno. Si llamamos  $Z$  a la suma de las  $z_j$ , entonces

$$s_{k+2j} = \frac{z_1 + \dots + z_j}{Z}, \quad j = 1, \dots, n.$$

El resultado se sigue de la transformación de las variables  $z_1, \dots, z_{n+1}$  para dar la distribución conjunta de  $s_{k+2}, \dots, s_{T-2}$ ,  $Z$  y entonces integrando  $Z$ .

La distribución de una muestra ordenada de observaciones independientes de la distribución uniforme  $U(0, 1)$  juega un papel importante en la teoría de la estadística no paramétrica, asimismo son importantes las estadísticas  $c^+$  y  $c^-$ , las cuales se definen como

$$c^+ = \max_{j=1, \dots, m-1} \left( s_{k+2j} - \frac{j}{m} \right), \quad \text{y} \quad c^- = \max_{j=1, \dots, m-1} \left( \frac{j}{m} - s_{k+2j} \right)$$

con  $m = \frac{T - k}{2}$ , las cuales se consideran equivalentes a la estadística Kolmogorov-Smirnov modificada  $C_n^+$ , con  $n = m - 1$  (Pyke, 1959). Las estadísticas  $c^+$  y  $c^-$  son la máxima desviación positiva y negativa respectivamente, del conjunto de estadísticas  $(s_{k+2}, \dots, s_{T-2})$  respecto de su línea de valor medio.

Durbin (1969) presenta una tabla de valores de significancia para la cantidad  $C_n^+$  para  $n = m - 1$ . El procedimiento sugerido para la prueba de cuadrados de las sumas acumuladas consiste en tomar estos valores como aproximaciones a los valores significativos de

$$c_1^+ = \max_{i=1, \dots, T-k-1} \left( s_{k+i} - \frac{i}{T-k} \right) \quad \text{y} \quad c_1^- = \max_{i=1, \dots, T-k-1} \left( \frac{i}{T-k} - s_{k+i} \right),$$

las cuales son la máxima desviación positiva y negativa respectivamente del conjunto de las  $s_r$  respecto de la línea de valor medio.

Para el nivel de significancia  $\lambda$ , normalmente 0.1 o menor, la probabilidad de cruzar ambas líneas es insignificante, de modo que dado un nivel de significancia  $\lambda$ , para encontrar el valor correspondiente para  $c_0$  podemos tomar el valor de la tabla para  $n = \frac{T-k}{2} - 1$  y  $\lambda/2$ . Si  $T-k$  es impar se sugiere hacer una interpolación lineal entre los valores de  $n = \frac{T-k}{2} - \frac{3}{2}$  y  $n = \frac{T-k}{2} - \frac{1}{2}$ . Hutchison llevó a cabo simulaciones Monte Carlo y comprobó que esta prueba arroja resultados significativos con mayor frecuencia que la prueba exacta, pero que la discrepancia es muy pequeña cuando  $T-k$  es superior a 30.

### 3.6. Algunas pruebas más

Beckman y Cook (1979, [10]) propusieron el estadístico de prueba  $F = \max(F_k)$  para  $p+1 \leq k \leq n-p-1$  para el modelo lineal  $Y = \beta_0 + \beta_1 X + \epsilon$ , donde

$$F_k = \frac{(SCE - (SCE_{1k} + SCE_{2k})) / (p+1)}{(SCE_{1k} + SCE_{2k}) / (n-2(p+1))} \sim F_{p+1, n-2(p+1)}.$$

Worsley (1983, [14]) propuso la estadística,

$$F = \max_{1 \leq i \leq p} \frac{U_i/p}{(Q - U_i) / (n - 2p)},$$

para un modelo de regresión lineal con  $p$  variables explicativas bajo las hipótesis  $H_0 : \delta = 0$  contra  $H_1 : \delta \neq 0$ , donde  $\delta = \beta - \beta^*$  con

$$\begin{aligned} U_i &= \hat{\delta}_i^T \Sigma_i^{-1} \hat{\delta}_i \\ \hat{\delta}_i &= \hat{\beta}_i - \hat{\beta}_i^* \\ \Sigma_i^{-1} &= \text{Varianza bajo } H_0 \\ Q &= \text{Suma de cuadrados del error bajo } H_0 \end{aligned}$$

Horvath y Shao (1993, [18]) consideraron un modelo de regresión lineal con una variable explicativa y utilizaron la prueba  $H_0 : k \geq n$  contra  $H_1 : 1 \leq k < n$  y la estadística de prueba

$$T_n = \max_{p \leq i \leq n-p} \left\{ \frac{1}{\hat{\sigma}^2} (\hat{\beta}_i - \hat{\beta}_i^*) H_i^{-1} (\hat{\beta}_i - \hat{\beta}_i^*) \right\}$$

donde  $H_i = (X_i^T X_i)^{-1} + (X_i^{T*} X_i^*)^{-1}$ ,  $H^{-1}$  es la varianza total del modelo.

Antoch y Huscova (2001, [27]) consideraron el modelo de localización y propusieron la estadística de prueba

$$T_n = \max_{1 < i < n} \left\{ \sqrt{\frac{n}{i(n-i)s^2}} \left| \sum_{j=1}^i (\bar{X}_{Rj} - \bar{X}_n) \right| \right\}$$

y rechazaron la hipótesis nula si  $T_n$  se halla en el 5% más alto de los valores de

$$T_k(R) = \max_{1 < i < n} \left\{ \sqrt{\frac{n}{i(n-i)s^2}} \left| \sum_{j=1}^i (\bar{X}_{Rj} - \bar{X}_n) \right| \right\}$$

donde  $R$  indica que el cálculo se hace sobre todas las posibles permutaciones de los datos muestrales, estos autores encontraron que

$$\begin{aligned} P(\sqrt{2 \log \log n T_k(R)} \leq y + 2 \log \log n - \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi | X_1, \dots, X_n) \\ \longrightarrow \exp(-2 \exp(-y)) \end{aligned}$$

casi seguramente. Antoch, Gregoire y Jaruscova (2004) generalizaron la propuesta de Antoch y Huscova al modelo lineal con  $p$  variables independientes.

Vito Muggeo (2003) abordó el problema de estimación del punto de cambio estructural desde un punto de vista distinto al de los autores anteriores pues menciona que la técnica de máxima verosimilitud es inaplicable por el hecho de que el logaritmo de la función de verosimilitud no es diferenciable en el punto de cambio.

Muggeo consideró un modelo no lineal para modelar la tendencia en todo el intervalo de observación y propuso estimar a  $n_1$  aplicando una simple técnica de linealización.

De esta manera se han revisado algunos de los métodos que existen hoy día en la literatura para probar la existencia de un punto de cambio así como para determinar la localización del mismo. En el siguiente capítulo abordaremos finalmente el problema del punto de cambio en modelos lineales generalizados.

## Punto de cambio en modelos lineales generalizados

### 4.1. Introducción

En este capítulo se revisan algunos métodos para detectar y localizar un punto de cambio en un modelo lineal generalizado. Uno de los principales problemas que aborda la literatura de punto de cambio tiene que ver con probar hipótesis, para lo cual se analiza una sucesión de observaciones independientes obtenidas previamente con el objetivo de determinar si todas comparten la misma distribución. Es común suponer que si ocurrió un cambio en la distribución, entonces éste es único y las observaciones después del cambio tienen todas la misma distribución, la cual difiere de la distribución de las observaciones antes del cambio (por ejemplo, Page, 1954; Sen y Srivastava, 1975; Pettitt, 1980; Worsley, 1983; Gombay and Horvath, 1994).

Por otro lado, numerosos estudios han tratado con observaciones independientes las cuales no son idénticamente distribuidas (por ejemplo, Quandt, 1960; Brown *et al.*, 1975; Kim and Siegmund, 1989; Ulm, 1991; Pastor y Guallar, 1998). Por ejemplo, Ulm (1991) propuso una prueba para detectar un cambio en los parámetros de la regresión logística basándose en una cuasi distribución  $\chi_1^2$ .

Muchos estudios epidemiológicos se basan en un modelo de regresión logística bisegmentado, en el cual el término lineal asociado a la exposición continua en la

regresión logística estándar es reemplazado por un polinomio bisegmentado con un punto de cambio desconocido. En esos estudios, los parámetros del modelo (incluyendo el punto de cambio) a menudo se estiman sin previamente probar su existencia (e.g. Pastor y Guallar, 1998; Gossl y Kuchenhoff, 2001).

#### 4.2. Punto de cambio en el modelo de regresión logístico

En esta sección se revisa una prueba de máxima verosimilitud generalizada potente y asintótica cuyo propósito es detectar un punto de cambio en una regresión logística cuando la hipótesis alternativa específica que ocurrió un cambio en los parámetros del modelo. Asimismo se presenta una cota superior no asintótica garantizada para el nivel de significancia de la prueba. En los casos en los que la prueba apoya la conclusión de que hubo un punto de cambio, existe un estimador máximo verosímil del punto de cambio para el cual se presentan resultados relacionados con las propiedades asintóticas del estimador (Gurevich, G. y Vexler, A. (2004) [32]).

Considere una sucesión finita de observaciones independientes  $(Y_i, x_i)$ ,  $1 \leq i \leq n$  donde  $x_i = [x_{i1}, \dots, x_{ip}]^T$ , son  $p$  variables explicativas fijas y  $Y_i = [Y_{i1}, \dots, Y_{im_i}]^T$ , donde las  $Y_{ij}$  son observaciones Bernoulli cuya distribución exacta depende del predictor  $x_i$ . Suponga que

$$(4.2.1) \quad P\{Y_{ij} = 1|x_i\} = (1 + \exp(-x_i^T \beta_1))^{-1} I\{i < k\} \\ + (1 + \exp(-x_i^T \beta_2))^{-1} I\{i \geq k\},$$

$$\beta_l = [\beta_{l1}, \dots, \beta_{lp}]^T, \quad l = 1, 2, \quad j = 1, \dots, m_i, \quad i = 1, \dots, n, \beta_1 \neq \beta_2,$$

donde  $I\{\cdot\}$  es la función indicadora y  $\beta_1$ ,  $\beta_2$  y  $k$  son parámetros del modelo. El modelo corresponde a una situación donde hasta un punto de cambio desconocido  $k > 0$ , las observaciones satisfacen el modelo de regresión logística con parámetro  $\beta_1$  y más allá de ahí (siempre y cuando  $k \leq n$ ), el parámetro cambia a  $\beta_2$ .

Formalmente, se trata de una prueba de hipótesis para un problema de punto de cambio, donde

$$(4.2.2) \quad H_0 : k \notin [1, n], \text{ vs. } H_1 : 1 \leq k \leq n; \quad k > 0 \text{ desconocida.}$$

En el siguiente apartado se presenta la prueba para el caso simple, es decir cuando  $\beta_1$  y  $\beta_2$  son conocidos y el EMV del punto de cambio está definido.

**4.2.1. Prueba con parámetros  $\beta_1$  y  $\beta_2$  conocidos.** Considere el modelo (4.2.1) junto con la prueba de hipótesis (4.2.2), donde  $\beta_1$  y  $\beta_2$  son conocidos. El análisis que a continuación se presenta es relativamente claro y contiene los ingredientes básicos de casos más generales. Mediante el análisis del caso más simple en el siguiente ejemplo, se pretende que pueda asociarse de manera práctica con el caso más general: una herramienta manual indicará  $\beta_1$  o  $\beta_2$  si el fabricante espera un cambio funcional en el trabajo corriente de la herramienta; si la muestra presenta colas grandes, i.e., se hicieron numerosas observaciones antes y después de un posible cambio en la distribución.

Sea  $P_{H_0}$  la distribución de la sucesión  $Y_1, \dots, Y_n$  bajo  $H_0$  y  $P_{k=v}$  denota la distribución de la sucesión  $Y_1, \dots, Y_n$  bajo  $H_1$  con  $k = v$ ;  $v = 1, \dots, n$ . De modo similar, sean  $E_{H_0}$  y  $E_{k=v}$  la esperanza bajo  $P_{H_0}$  y  $P_{k=v}$ , respectivamente. Denotando

$$\Lambda_v^n(\beta_1, \beta_2 | x_k, \dots, x_n) = \frac{\prod_{i=v}^n \prod_{j=1}^{m_i} P_{k=v} \{Y_{ij} | x_i\}}{\prod_{i=v}^n \prod_{j=1}^{m_i} P_{H_0} \{Y_{ij} | x_i\}} = \frac{\prod_{i=v}^n \prod_{j=1}^{m_i} P_{k=i} \{Y_{ij} | x_i\}}{\prod_{i=v}^n \prod_{j=1}^{m_i} P_{H_0} \{Y_{ij} | x_i\}}$$



$$= \exp \left( \sum_{i=v}^n \sum_{j=1}^{m_i} (x_i^T \beta_2 - x_i^T \beta_1) Y_{ij} \right) \prod_{i=v}^n \left( \frac{1 + \exp(x_i^T \beta_1)}{1 + \exp(x_i^T \beta_2)} \right)^{m_i},$$

donde

$$P_{k=v} \{Y_{ij}|x_i\} = P_{k=v} \{Y_{ij} = 1|x_i\}^{Y_{ij}} (1 - P_{k=v} \{Y_{ij} = 1|x_i\})^{1-Y_{ij}},$$

$$P_{H_0} \{Y_{ij}|x_i\} = P_{H_0} \{Y_{ij} = 1|x_i\}^{Y_{ij}} (1 - P_{H_0} \{Y_{ij} = 1|x_i\})^{1-Y_{ij}}.$$

Se propone la siguiente prueba: rechazar  $H_0$  sí y sólo sí

$$(4.2.3) \quad \max_{1 \leq v \leq n} \Lambda_v^n(\beta_1, \beta_2 | x_k, \dots, x_n) > C$$

donde  $C$  es una constante.

*Nivel de significancia de la prueba:* en (4.2.3) se propone una prueba de cociente de máxima verosimilitud generalizada para el problema (4.2.2). Es ampliamente conocido en la literatura del punto de cambio que tales pruebas tienen potencia alta. Por lo tanto, la evaluación de su nivel de significancia es un asunto principal.

La mayoría de los resultados que tienen que ver con el nivel de significancia de la prueba de cociente de máxima verosimilitud generalizada, aún en los casos más simples de observaciones i.i.d. antes y después del cambio, son asintóticas ( $n \rightarrow \infty$ ), e.g. Gombay y Horvath (1994).

Gurevich y Vexler (2004, [32]) encontraron que una cota superior no asintótica garantizada para el nivel de significancia de la prueba,  $\alpha$ , satisface

$$\alpha = P_{H_0} \left\{ \max_{1 \leq v \leq n} \Lambda_v^n(\beta_0, \beta_1 | x_v, \dots, x_n) > C \right\} \leq \frac{1}{C}$$

De lo anterior se tiene una cota superior (que no involucra a  $n$  y no se considera ninguna condición sobre las variables explicativas  $x_1, \dots, x_n$ ) para el nivel de

significancia de la prueba (4.2.3):  $\alpha \leq 1/C$ . Entonces,  $C = 1/\alpha$  determina una prueba con un nivel de significancia que no excede  $\alpha$ .

*Potencia de la prueba:* para toda  $1 \leq i \leq n$ , definimos

$$(4.2.4) \quad a_i \equiv E_k \left[ \ln \left( \prod_{j=1}^{m_i} \frac{P_k \{Y_{ij} | x_i\}}{P_{H_0} \{Y_{ij} | x_i\}} \right) \right]$$

$$= m_i \left( (x_i^T \beta_2 - x_i^T \beta_1) \frac{\exp(x_i^T \beta_2)}{1 + \exp(x_i^T \beta_2)} + \ln \left( \frac{1 + \exp(x_i^T \beta_1)}{1 + \exp(x_i^T \beta_2)} \right) \right)$$

*Estimación máximo verosímil de un punto de cambio:* si se rechaza  $H_0$ , entonces, siguiendo a Borovkov (1998), se considera el siguiente estimador para la  $k$

$$(4.2.5) \quad \hat{k} = \arg \max_{1 \leq k \leq n} \Lambda_k^n(\beta_1, \beta_2 | x_k, \dots, x_n \leq C).$$

En el caso en que uno o ambos parámetros sean desconocidos se remite al lector al artículo de Gurevich y Vexler (2004, [32]) para mayor referencia.

**4.2.2. Ejemplo: distribución del número de hijos.** La Oficina Central de Estadísticas de Israel realizó una investigación para determinar si hubo un cambio en la distribución del número de hijos por cada mujer Israelí en décadas pasadas, y de ser así, en que año ocurrió esto. La encuesta, basada en una muestra aleatoria del censo de población de 1995 en Israel, supone que el posible cambio en la distribución puede ser resultado de la inmigración hacia Israel. Por lo tanto, todas las mujeres que llegaron a Israel después del año en el cual ocurrió el cambio en la distribución serían designadas como nuevos inmigrantes (NI). Esta designación es relevante para ciertas aplicaciones sociológicas. Claramente, si no hay cambio en la distribución no hay necesidad de definir el estatus de NI.

Considere el siguiente modelo. Sea

$$x_i = \left[ 1, i, \frac{1}{m_i} \sum_{j=1}^{m_i} a_{ij} \right]^T$$

y defina las variables aleatorias

$$Y_{ij} = \begin{cases} 1 & \text{si el número de hijos por mujer es } \leq 2 \text{ (para 1995),} \\ 0 & \text{si el número de hijos por mujer es } > 2 \text{ (para 1995),} \end{cases}$$

donde  $i = 80, \dots, 95$  es el año de arribo a Israel,  $a_{ij}$  la edad de la mujer  $ij$  en el año  $i$ ,  $1 \leq j \leq m_i$ ,  $m_i$  representa el número de mujeres que llegaron a Israel en el año  $i$  y aparece en los datos. Se sabe que el número promedio de hijos por mujer en Israel es de alrededor 2.

Entonces, de acuerdo con el modelo anterior, en el cual  $p = 3$ ,  $j = 1, \dots, m_i$ ,  $i = 80, \dots, 95$ , se desea probar las hipótesis

$$H_0 : k \notin [80, 95], \text{ vs. } H_1 : 80 \leq k \leq 95; k > 79 \text{ es desconocido}$$

Hubo una decisión previa de que la definición del estado específico NI para mujeres que llegaron a Israel hacia 1980 es irrelevante. Por lo tanto utilizamos el modelo de regresión que es comúnmente aceptado por la Oficina Central de Estadísticas en Jerusalén, basado en observaciones relacionadas a los años previos a 1980. El modelo supone  $\beta_1 = [-0.8302, -0.0081, 0.0501]^T$ .

Aplicando el criterio anterior tenemos

$$\ln \left( \max_{80 \leq k \leq 95} \Lambda_k^{95} \left( \beta_1, \hat{\beta} | x_k, \dots, x_{95} \right) \right) = \ln \left( \Lambda_{89}^{95} \left( \beta_1, \hat{\beta} | x_{89}, \dots, x_{95} \right) \right) = 3631.2,$$

donde  $\hat{\beta}$  es el estimador de máxima verosimilitud de  $\beta_2$  ( $\hat{\beta}^{(l,n)} = \beta_1, n - l \leq 7$ ). Por lo tanto, basados en la prueba de hipótesis y la proposición anteriores se rechaza  $H_0$  con un p-valor  $\leq \exp(-3631) \approx 0$ .

Aplicando la fórmula anterior obtienen  $\hat{k} = 90$  como estimador del punto de cambio y definen el estado NI para las mujeres que llegaron a Israel comenzando en 1990.

### 4.3. Monitoreo de cambios estructurales en modelos lineales generalizados

**4.3.1. El modelo y los problemas de la prueba.** Sean  $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$  un conjunto de parejas aleatorias independientes donde las  $Y_i$  son una respuesta escalar y las  $X_i$  vectores  $p$ -dimensionales de variables explicativas.

Suponga que la densidad condicional de  $Y_i$  dada  $X_i = x_i$  tiene la forma

$$f(y_i | X_i = x_i) = \exp \left\{ \frac{y_i \theta(x_i) - b(\theta(x_i))}{a(\phi)} + c(y_i, \phi) \right\}$$

para algunas funciones conocidas  $a(\cdot)$ ,  $b(\cdot)$ , y  $c(\cdot)$ . Suponga que las covariables y la variable de respuesta están relacionadas de la siguiente manera:

$$\mathbb{E}(Y_i | X_i = x_i) = \mu_i = g(\eta_i) = g(x_i^T \beta_i), \quad i = 1, 2, \dots, n$$

donde  $g^{-1}(\cdot)$  es una función liga conocida apropiada,  $\beta$  es un vector  $p$ -dimensional de coeficientes de regresión, y  $\eta_i$  es el predictor lineal. Tomando la función liga canónica, i. e.,  $g^{-1}(\cdot) = b'^{-1}(\cdot)$  obtenemos

$$\text{Var}(Y_i | X_i = x_i) = \sigma_i^2 = a(\phi) g'(x_i^T \beta_i), \quad i = 1, 2, \dots, n$$

Las suposiciones necesarias para los resultados teóricos se listan abajo.

La primera suposición que hacemos es la siguiente

$$\beta_i = \beta_1, \quad i = 1, 2, \dots, m \leq n$$

la cual se conoce como supuesto de no contaminación, propuesto por Chu *et al.* (1996).

Bajo la suposición anterior, existe un vector  $p$ -dimensional  $d_1$  y una matriz  $D$  positiva definida  $p \times p$  tal que

$$\left\| \frac{1}{m} \sum_{i=1}^m g(x_i^T \beta_1) x_i - d_1 \right\| = O(m^{-\alpha}) \quad a.s. \quad (a > 0)$$

$$\frac{1}{mt} \sum_{i=1}^{[mt]} \sigma_i^2 x_i x_i^T \xrightarrow{p} D, \quad \text{para toda } t \in (0, +\infty).$$

Se quiere probar la hipótesis nula

$$H_0 : \beta_i = \beta_1, \quad i = m, m+1, \dots$$

contra la alternativa

$$H_1 : \beta_i = \beta_1, \quad i = m, m+1, \dots, m+k^*;$$

$$\beta_i = \beta_*, \quad i = m+k^*+1, \dots, \quad \text{con } \beta_0 \neq \beta_*.$$

Los valores de los parámetros  $\beta_1$ ,  $\beta_*$ , y  $k^*$  no son especificados. Además, se necesita imponer la siguiente condición sobre  $k^*$  :  $k^* = [m\lambda]$ ,  $\lambda > 0$ .

**4.3.2. Procesos de puntajes.** Una función muy importante en el procedimiento anterior es la suma parcial del vector de residuos ponderados

$$S_m(k; \hat{\beta}_1) = \sum_{i=1}^k \left( y_i - g(x_i^T \hat{\beta}_1) \right) x_i, \quad k = m, m+1, \dots,$$

donde  $\hat{\beta}_1$  es una solución del siguiente sistema de ecuaciones

$$\sum_{i=1}^m \left( y_i - g \left( x_i^T \hat{\beta}_1 \right) \right) x_i = 0,$$

i. e., en la definición de los residuos usamos únicamente los datos históricos para estimar  $\beta$ .

Bajo condiciones de regularidad y el supuesto de no contaminación,  $\hat{\sigma}_i^2 = a(\phi) g' \left( x_i^T \hat{\beta}_1 \right) \xrightarrow{p} \sigma_i^2$  y  $\hat{\beta}_1 \xrightarrow{p} \beta_1$ . Para más información, consulte Fahrmeir y Kaufmann (1985).

Sea  $L([1, +\infty)^p)$  el espacio de funciones que son continuas por la derecha con límites por la izquierda sobre  $[1, +\infty)^p$ ; el espacio es dotado con la topología de Skorohod. Obviamente,  $S_m([mt]; \hat{\beta}_0)$ ,  $t \geq 1$  está en  $L([1, +\infty)^p)$ .

De tal manera que se construye el proceso empírico de puntajes

$$\begin{aligned} W_m(t; \hat{\beta}_1) &= \hat{J}_m^{-\frac{1}{2}} \frac{1}{\sqrt{m}} \sum_{i=1}^{[mt]} \left( y_i - g \left( x_i^T \hat{\beta}_1 \right) \right) x_i, \\ &= \hat{J}_m^{-\frac{1}{2}} \frac{1}{\sqrt{m}} S_m([mt]; \hat{\beta}_1), \quad t \geq 1, \end{aligned}$$

donde

$$\hat{J}_m = \frac{1}{m} \sum_{i=1}^m \hat{\sigma}_i^2 x_i x_i^T.$$

**TEOREMA 1.** *Dada una función de densidad de probabilidad de la familia exponencial, suponiendo que se cumplen  $H_0$  y los supuestos enunciados anteriormente, resulta que*

$$(W_m(t; \hat{\beta}_0), t \geq 1) \implies (W^0(t), t \geq 1)$$

donde  $(W^0(t), t \geq 1)$  es un puente Browniano  $p$ -dimensional.

Los métodos de prueba CUSUM y MOSUM que se exponen a continuación están basados en el proceso de puntajes empírico anterior, el cual refleja la sintomatología del cambio estructural desde un ángulo ligeramente distinto. Bajo  $H_0$ , los

comportamientos de estos dos procesos están bien representados por  $(W^0(t), t \geq 1)$  y  $(W^0(t) - W^0(t-h), t \geq 1)$ . Si  $H_0$  es falsa, los dos procesos flucturán en el sentido de que sus comportamientos no estarán completamente caracterizados por los procesos anteriores. Por lo tanto, las pruebas pueden construirse detectando las fluctuaciones anormales de estos procesos empíricos.

**4.3.3. CUSUM basado en los residuos ponderados.** Desde que Brown *et al.* (1975) utilizaron los residuos recursivos para detectar cambios estructurales, los residuos recursivos han jugado un rol importante en la estadística (Gombay y Horvath, 1994; Horvath *et al.*, 2004). Sin embargo, no funcionan para el modelo lineal generalizado porque en este caso tanto la media como la varianza de las observaciones son funciones de los parámetros de la regresión y si hay un cambio en estos parámetros, se afectan tanto la media como la varianza.

Para superar esta dificultad Xia *et al.* (2009, [36]) definen el residuo ponderado  $\hat{e}_i = \left( y_i - g \left( x_i^T \hat{\beta}_0 \right) \right) x_i$ . Estos autores construyen una función a la que llaman detector

$$\Gamma_{m+k} = W_m \left( \frac{m+k}{m}; \hat{\beta}_0 \right), \quad k = 0, 1, 2, \dots$$

y otra función frontera  $f(t)$  apropiada la cual a menudo eligen más por conveniencia matemática que por su optimalidad para la ley del logaritmo iterado.

Basados en el proceso empírico, estos autores construyen la siguiente estadística CUSUM

$$T_m = \max_{t \geq 1} \frac{\|W_m(t; \hat{\beta}_0)\|_\infty}{f(t)},$$

donde la notación  $\|\cdot\|_\infty$  indica la norma máx, es decir,  $\|x\|_\infty = \max_{1 \leq i \leq p} \{|x_i|\}$  para  $x \in \mathbb{R}^p$  y  $f(t)$  es positiva y continua en  $[1, \infty)$ .

El procedimiento se detiene y se rechaza  $H_0$  cuando

$$HT_m = \begin{cases} \inf \{k \geq 1 : \Gamma_{m+k} \geq g(m, k)\} \\ \infty, \text{ si } \Gamma_{m+k} < g(m, k), \text{ para toda } k = 1, 2, \dots \end{cases}$$

el cual a menudo se conoce como primer tiempo de paro. El estadístico  $\Gamma_{m+k}$  y la frontera deben ser elegidos de manera que

$$\begin{aligned} \lim_{m \rightarrow \infty} \mathbb{P} \{HT_m < \infty | H_0\} &= \alpha, \alpha \in (0, 1), \\ \lim_{m \rightarrow \infty} \mathbb{P} \{HT_m < \infty | H_1\} &= 1 \end{aligned}$$

El primer requisito significa que la probabilidad de parar bajo la hipótesis nula de no cambio es  $\alpha$  mientras que de acuerdo con el segundo requerimiento paramos con probabilidad uno bajo la alternativa.

**4.3.4. MOSUM basado en los residuos ponderados.** Bauer y Hackl (1978) encontraron que las sumas acumuladas de residuos recursivos no son muy sensibles para ciertos tipos de cambios en los parámetros. En algunos casos, para mejorar la potencia de las pruebas propusieron una prueba MOSUM para las sumas móviles de los residuos recursivos. Xia *et al.* (2009) probaron que la prueba móvil también funciona en los modelos lineales generalizados usando residuos ponderados en lugar de residuos recursivos.

Construyendo el detector

$$\Gamma_{m+k} = MW_m \left( \frac{m+k}{m}; \hat{\beta}_1, h \right), \quad k = 0, 1, 2, \dots; \quad h \in (0, 1]$$



y una función frontera adecuada  $f(t)$ . Por conveniencia matemática, toman la función (Leisch et al., 2000)

$$f(t) = \sqrt{2 \log_+ t} = \begin{cases} \sqrt{2} & \text{si } t \leq e \\ \sqrt{2 \log t} & \text{si } t > e \end{cases}$$

donde  $e = 2.71828\dots$

Basados en el proceso de puntajes empíricos móviles, construyen la siguiente estadística MOSUM:

$$MT_m(h) = \max_{t \geq 1} \frac{\|MW_m(t; \hat{\beta}_1, h)\|_\infty}{\sqrt{2 \log_+ t}}.$$

#### 4.4. Método SIC para la detección del cambio paramétrico

Esta metodología se conoce también como enfoque basado en información ([26]). Una ventaja de este enfoque es que el probar una hipótesis de punto de cambio puede verse como un problema de selección de modelo. El criterio de información es una excelente herramienta para la selección de modelos que además resulta computacionalmente efectiva.

A continuación se obtendrá el SIC para el problema de punto de cambio paramétrico, es decir, sin considerar variables explicativas. A continuación se desarrollan los modelos para datos discretos binomial y Poisson, así como el modelo exponencial para datos continuos.

**4.4.1. Cambio paramétrico en el modelo binomial.** Suponga que se tienen  $n$  variables aleatorias binomiales y que  $x_i = m_i$  para  $i = 1, \dots, n$ , con  $m_i$  el número de casos exitosos. Se desea probar el siguiente juego de hipótesis

$$H_0 : p_1 = p_2 = \dots = p_n = p$$

versus la alternativa

$$H_1 : p_1 = \cdots = p_k = p \neq p_{k+1} = \cdots = p_n = p'.$$

Es decir, se desea probar si hubo cambio en la distribución de los datos sin tomar en cuenta variables explicativas en el modelo. Lo anterior equivale a utilizar una función lisa constante,  $g(\mu_i) = c$ .

Se denota por  $M_k = \sum_{i=1}^k m_i$ , y  $N_k = \sum_{i=1}^k n_i$ , para  $k = 1, \dots, n$  y  $M \equiv M_n$ ,  $N \equiv N_n$ ,  $M'_k = M - M_k$ , y por último,  $N'_k = N - N_k$ .

Bajo  $H_0$ , la función de verosimilitud evaluada en el estimador máximo verosímil se obtiene como

$$L_0(\hat{p}) = \prod_{i=1}^n \binom{n_i}{m_i} \hat{p}^{m_i} (1 - \hat{p})^{n_i - m_i},$$

donde  $\hat{p} = M/N$ . Por lo tanto, el Criterio de Información de Schwarz bajo  $H_0$ , denotado por  $SIC(n)$  se obtiene como

$$\begin{aligned} SIC(n) &= -2 \log L_0(\hat{p}) + \log(n) \\ &= -2 \sum_{i=1}^n \log \binom{n_i}{m_i} - 2M \log \left( \frac{M}{N} \right) \\ &\quad - 2(N - M) \log \left( \frac{N - M}{N} \right) + \log(n). \end{aligned}$$

Bajo  $H_1$ , la función de máxima verosimilitud evaluada en los estimadores máximo verosímiles se obtiene como

$$L_1(\hat{p}, \hat{p}') = \prod_{i=1}^k \binom{n_i}{m_i} \hat{p}^{m_i} (1 - \hat{p})^{n_i - m_i} \cdot \prod_{j=k+1}^n \binom{n_j}{m_j} \hat{p}'^{m_j} (1 - \hat{p}')^{n_j - m_j}$$

donde  $\hat{p} = \frac{M_k}{N_k}$  y  $\hat{p}' = \frac{M'_k}{N'_k}$ . Por lo tanto, el criterio de Información de Schwarz bajo  $H_1$ , denotado por  $SIC(k)$ , para  $k = 1, \dots, n-1$ , se obtiene como

$$\begin{aligned} SIC(k) &= -2 \log L_1(\hat{p}, \hat{p}') + 2 \log(n) \\ &= -2 \sum_{i=1}^n \log \begin{pmatrix} n_i \\ m_i \end{pmatrix} - 2M_k \log \left( \frac{M_k}{N_k} \right) - 2(N_k - M_k) \log \left( \frac{N_k - M_k}{N_k} \right) \\ &\quad 2M'_k \log \left( \frac{M'_k}{N'_k} \right) - 2(N'_k - M'_k) \log \left( \frac{N'_k - M'_k}{N'_k} \right) + 2 \log(n) \end{aligned}$$

De acuerdo al principio del criterio de mínima información, no se rechaza  $H_0$  si,

$$SIC(n) \leq \min_{1 \leq k \leq n-1} SIC(k),$$

o bien, no se rechaza  $H_0$  si,

$$\min_{1 \leq k \leq n-1} \Delta(k) \geq 0,$$

donde  $\Delta$  representa el cociente de momios y se calcula de la siguiente manera:

$$\begin{aligned} \Delta(k) &= M \log(M) + (N - M) \log(N - M) - N \log(N) \\ &\quad - M_k \log(M_k) - (N_k - M_k) \log(N_k - M_k) + N_k \log(N_k) \\ &\quad - M'_k \log(M'_k) - (N'_k - M'_k) \log(N'_k - M'_k) + N'_k \log(N'_k) \\ &\quad - \frac{1}{2} \log \frac{1}{n} \end{aligned}$$

Estos dos criterios son equivalentes pero éste último, a diferencia del primero, es más sencillo de obtener.

Cuando se acepta  $H_1$ , la posición del punto de cambio  $k$  es estimada por  $\hat{k}$ , de tal suerte que

$$\Delta(\hat{k}) = \min_{1 \leq k \leq n-1} \Delta(k).$$

**4.4.2. Ejemplo: pie equinvaro congénito y herbicida 2,3,5-T.** Hanify *et al.* (1981), estudiaron el número de casos de pie equinvaro congénito en el primer mes de embarazo en relación al número total de nacimientos entre los años 1960 y 1976 en la región norte de Nueva Zelanda. Estos investigadores sospechaban que podría haber una alta correlación entre la incidencia de pie equinvaro y la cantidad de herbicida 2,4,5-T utilizado en la región durante esos años. Worsley (1983) analizó estos datos suponiendo un modelo binomial y aplicó tanto la prueba del cociente de verosimilitudes como la prueba CUSUM Q para localizar el punto de cambio, el cual se localizó en la sexta observación correspondiente a 1965, año en que el herbicida 2,4,5-T fue utilizado por primera vez en la región.

Se utilizaron los mismos datos para aplicar el criterio de selección de modelo, SIC. Los cálculos muestran que  $\Delta(\hat{k}) = \min_{1 \leq k \leq n-1} \Delta(k)$  también se cumple para  $k = 6$ . Por lo tanto, el punto de cambio es exitosamente localizado y coincide con el reportado por Worsley (1983). Los valores calculados para  $\Delta(k)$  se muestran en la Tabla 1 junto con el conjunto de datos original tal como aparece en Worsley (1983). El valor de  $\Delta(k)$  marcado con asterisco es el mínimo de los  $\Delta(k)$  y corresponde al año 1965 o bien el punto  $k = 6$ . El código en R para este ejemplo se presenta en el apéndice C.

**4.4.3. Cambio paramétrico en el modelo Poisson.** Sea  $x_1, \dots, x_n$  una sucesión de variables aleatorias independientes Poisson con media  $\lambda_i, i = 1, 2, \dots, n$ . Se desea probar el juego de hipótesis  $H_0$  vs.  $H_1$  por posible cambio en la media o

Año	$i$	$m_i$	$n_i$	$\Delta(k)$
1960	1	8	2409	1.0399
1961	2	3	2453	-2.0785
1962	3	9	2290	-1.5289
1963	4	12	2171	-0.0234
1964	5	6	2084	-0.9095
1965	6	4	1993	-2.7511*
1966	7	14	2157	-0.6740
1967	8	12	2091	0.1898
1968	9	7	2152	-0.5198
1969	10	5	2007	-2.0086
1970	11	13	2027	-0.6181
1971	12	11	1963	-0.0351
1972	13	12	1982	0.6741
1973	14	9	1974	0.5833
1974	15	6	1932	-0.7964
1975	16	13	1807	0.7130
1976	17	12	1919	-

CUADRO 1. Conjunto de datos y valores  $\Delta(k)$ 

en la varianza,  $\lambda_i$ . Es decir, probamos

$$H_0 : \lambda_1 = \lambda_2 = \dots = \lambda_n = \lambda$$

contra la alternativa

$$H_1 : \lambda_1 = \dots = \lambda_k = \lambda \neq \lambda_{k+1} = \dots = \lambda_n = \lambda'$$

Bajo  $H_0$ , la función de máxima verosimilitud está dada por

$$L_0(\hat{\lambda}) = \frac{e^{-n\hat{\lambda}} \hat{\lambda}^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

donde  $\hat{\lambda} = M/n$  con  $M = \sum_{i=1}^n x_i$ , el conteo total. Por lo tanto, el Criterio de Información de Schwarz bajo  $H_0$ , denotado por  $SIC(n)$ , se obtiene como

$$\begin{aligned} SIC(n) &= -2 \log L_0(\hat{\lambda}) + \log(n) \\ &= 2M - 2M \log \left( \frac{M}{n} \right) + 2 \log \left( \prod_{i=1}^n x_i! \right) + \log(n). \end{aligned}$$

Bajo  $H_1$ , la función de máxima verosimilitud es

$$L_1(\hat{\lambda}, \hat{\lambda}') = \frac{e^{-k\hat{\lambda}\hat{\lambda}'\sum_{i=1}^k x_i}}{\prod_{i=1}^k x_i!} \cdot \frac{e^{-(n-k)\hat{\lambda}'\hat{\lambda}'\sum_{i=k+1}^n x_i}}{\prod_{i=k+1}^n x_i!}$$

donde  $\hat{\lambda} = \frac{M_k}{k}$ , y  $\hat{\lambda}' = \frac{M'_k}{n-k}$  con  $M_k = \sum_{i=1}^k x_i$  y  $M'_k = \sum_{i=k+1}^n x_i$ . Por lo tanto, el Criterio de Información de Schwarz bajo  $H_1$ , denotado por  $SIC(k)$ , para  $k = 1, \dots, n-1$ , se obtiene como

$$\begin{aligned} SIC(k) &= -2 \log L_1(\hat{\lambda}, \hat{\lambda}') + 2 \log(n) \\ &= 2M_k - 2M_k \log \frac{M_k}{k} + 2M'_k - 2M'_k \log \frac{M'_k}{n-k} \\ &\quad - 2 \log \left( \prod_{i=1}^n x_i! \right) + 2 \log(n) \end{aligned}$$

De acuerdo con el principio del criterio de mínima información, se rechaza  $H_0$  si

$$SIC(n) > \min_{1 \leq k \leq n-1} SIC(k),$$

o

$$\min_{1 \leq k \leq n-1} \left[ M \log \frac{M}{n} - M_k \log \frac{M_k}{k} - M'_k \log \frac{M'_k}{n-k} \right] < \frac{1}{2} \log \frac{1}{n},$$

y se estima la posición del punto de cambio por  $\hat{k}$ , de tal suerte que

$$SIC(\hat{k}) = \min_{1 \leq k \leq n-1} SIC(k).$$

**4.4.4. Cambio paramétrico en el modelo exponencial.** Un modelo exponencial es útil y apropiado en muchas ciencias experimentales; por lo tanto, es deseable hacer inferencia sobre un punto de cambio en un modelo exponencial. En la literatura, varios autores han estudiado algunos aspectos de un punto de cambio ocurriendo en un modelo exponencial. Por ejemplo, Worsley (1986) utilizó el método de cociente de máxima verosimilitud para probar un punto de cambio en una sucesión de variables aleatorias independientes de la familia exponencial, con énfasis en la distribución exponencial. Haccou, Meelis, and Geer (1988) investigaron el problema de punto de cambio para una sucesión de variables aleatorias exponenciales usando la prueba de cociente de verosimilitudes, y obtuvieron la distribución asintótica nula de la estadística de prueba. Haccou y Meelis (1988) dieron un procedimiento para probar el número de puntos de cambio en una sucesión de variables aleatorias exponenciales basados en una partición de la verosimilitud de acuerdo con una jerarquía de sub-hipótesis. Ramanayake y Gupta (1997) estudiaron el cambio epidémico usando el método de cociente de verosimilitudes.

Tal como se vió antes, el enfoque basado en la información para la hipótesis de punto de cambio en el modelo exponencial es una alternativa moderna para el análisis estadístico del punto de cambio. Se quiere probar el juego de hipótesis

$$H_0 : x_1, \dots, x_n \sim i.i.d. \quad f(x; \lambda) = \lambda e^{-\lambda x}$$

contra

$$H_1 : \begin{aligned} &x_1, \dots, x_k \sim i.i.d. \quad f(x; \lambda_1) = \lambda_1 e^{-\lambda_1 x} \\ &x_{k+1}, \dots, x_n \sim i.i.d. \quad f(x; \lambda_2) = \lambda_2 e^{-\lambda_2 x}, \quad 1 \leq k \leq n-1 \end{aligned}$$

Tenemos que las funciones de verosimilitud bajo la hipótesis nula y alternativa evaluadas en el estimador máximo verosímil son

$$L_0(\hat{\lambda}) = \left( \frac{n}{\sum_{i=1}^n x_i} \right)^n e^{-n},$$

$$L_1(\hat{\lambda}_1, \hat{\lambda}_2) = \left( \frac{k}{\sum_{i=1}^k x_i} \right)^k \left( \frac{n-k}{\sum_{i=k+1}^n x_i} \right)^{n-k} e^{-n}$$

Entonces

$$\begin{aligned} SIC(n) &= -2 \log L_0(\hat{\lambda}) + \log(n) \\ &= 2n \log \sum_{i=1}^n x_i + 2n - 2n \log(n), \end{aligned}$$

y

$$\begin{aligned} SIC(k) &= -2 \log L_1(\hat{\lambda}_1, \hat{\lambda}_2) + 2 \log(n) \\ &= 2k \log \sum_{i=1}^k x_i + 2(n-k) \log \sum_{i=k+1}^n x_i + 2n \\ &\quad - 2k \log k - 2(n-k) \log(n-k) + 2 \log(n) \end{aligned}$$

para  $k = 1, 2, \dots, n-1$ . Entonces, si

$$\text{mín } SIC = \text{mín } \{SIC(n), SIC(k), k = 1, 2, \dots, n-1\}$$

se alcanza en  $SIC(n)$ , no se rechaza  $H_0$ ; pero si  $\text{mín } SIC$  se alcanza en  $SIC(k)$ , para  $k = 1, 2, \dots, n-1$ , se rechaza  $H_0$  y se dice que hay un punto de cambio en  $\hat{k}$  cuando  $\hat{k}$  es tal que  $SIC(\hat{k}) = \text{mín}_{1 \leq k \leq n-1} SIC(k)$ .



#### 4.5. Método SIC para detectar punto cambio en covariables

Como se vió en el capítulo anterior, una forma de atacar el problema del punto de cambio es a través de un criterio de información. En el capítulo anterior se utilizó el Criterio de Información de Schwarz, SIC, para decidir en qué momento había ocurrido un punto de cambio en el parámetro de una distribución de la familia exponencial. El método que se describe a continuación tiene como objetivo detectar el cambio a partir de las covariables del modelo.

Sean  $(y_1, x_1), \dots, (y_n, x_n)$  con  $x_i = (x_{i1}, \dots, x_{ip})$  vector de  $p$  variables explicativas y  $y_i$  observación de la familia de distribuciones exponencial. Suponga que la densidad de  $Y_i$  dada  $X_i = x_i$  tiene la forma

$$f_{Y_i}(y_i|X_i = x_i) = \exp \left\{ \frac{y_i \theta(x_i) - b(\theta(x_i))}{a(\phi)} + c(y_i, \phi) \right\}$$

para algunas funciones conocidas  $a(\cdot)$ ,  $b(\cdot)$ , y  $c(\cdot)$ .

Se ajusta un modelo lineal generalizado con estas observaciones y se quiere contrastar el siguiente juego de hipótesis

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 \quad \text{v.s.} \quad H_1 : \hat{\beta}_1 \neq \hat{\beta}_2$$

Bajo  $H_0$ , la log-verosimilitud resulta

$$\log L_0(\hat{\beta}) = \sum_{i=1}^n \log f(y_i; \theta_i, \phi) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi)$$

Por tanto, bajo  $H_0$ , el Criterio de Información de Schwarz, al que se denotará por  $SIC(n)$ , queda de la siguiente manera

$$SIC(n) = -2 \log L_0(\hat{\beta}) + p \log(n)$$

Por otra parte, bajo  $H_1$ , la función de log-verosimilitud se obtiene como

$$\log L_1(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^k \frac{y_i \theta_{i1} - b(\theta_{i1})}{a(\phi)} + \sum_{i=k+1}^n \frac{y_i \theta_{i2} - b(\theta_{i2})}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi)$$

con  $\theta_{i1} = g(\mu_{i1}) = \eta_{i1} = X \hat{\beta}_1$  y  $\theta_{i2} = g(\mu_{i2}) = \eta_{i2} = X \hat{\beta}_2$ .

Por tanto, el criterio de Información de Schwarz bajo  $H_1$ , denotado por  $SIC(k)$ , para  $k = 1, \dots, n-1$ , se obtiene como

$$SIC(k) = -2 \log L_1(\hat{\beta}_1, \hat{\beta}_2) + 2 \log(n)$$

Entonces de acuerdo al principio de mínima información, no se rechaza  $H_0$  si

$$SIC(n) \leq \min_{1 \leq k \leq n-1} SIC(k)$$

#### 4.6. Método SIC con el modelo logístico

Considere una sucesión finita de observaciones independientes  $(Y_1, x_1), \dots, (Y_n, x_n)$  con  $x_i = (x_{i1}, \dots, x_{ip})$  son  $p$  variables explicativas y  $Y_i$  son observaciones Bernoulli. Se quieren probar las hipótesis

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2$$

versus la alternativa

$$H_1 : \hat{\beta}_1 \neq \hat{\beta}_2$$

Bajo  $H_0$ , la función de verosimilitud resulta

$$L_0(\hat{\beta}) = \prod_{i=1}^n \frac{1}{1 + \exp(-x_i^T \hat{\beta})}$$

Por tanto, el Criterio de Información de Schwarz bajo  $H_0$ , al que se denotará por  $SIC(n)$  queda de la siguiente manera:

$$\begin{aligned}
 SIC(n) &= -2 \log L_0(\hat{\beta}) + p \log(n) \\
 &= -2 \log \left\{ \prod_{i=1}^n \frac{1}{1 + \exp(-x_i^T \hat{\beta})} \right\} + p \log(n) \\
 &= -2 \sum_{i=1}^n \log \left\{ \frac{1}{1 + \exp(-x_i^T \hat{\beta})} \right\} + p \log(n) \\
 &= 2 \sum_{i=1}^n \log \left\{ 1 + \exp(-x_i^T \hat{\beta}) \right\} + p \log(n)
 \end{aligned}$$

Bajo  $H_1$ , la función de verosimilitud se obtiene como

$$L_1(\hat{\beta}_1, \hat{\beta}_2) = \prod_{i=1}^k \frac{1}{1 + \exp(-x_i^T \hat{\beta}_1)} \cdot \prod_{i=k+1}^n \frac{1}{1 + \exp(-x_i^T \hat{\beta}_2)}$$

Por tanto, el Criterio de Información de Schwarz bajo  $H_1$ , al que se denotará por  $SIC(k)$ , para  $k = p, p+1, \dots, n-p$  se obtiene como

$$\begin{aligned}
 SIC(k) &= -2 \log L_1(\hat{\beta}_1, \hat{\beta}_2) + 2p \log(n) \\
 &= 2 \sum_{i=1}^k \log \left\{ 1 + \exp(-x_i^T \hat{\beta}_1) \right\} \\
 &\quad + 2 \sum_{i=k+1}^n \log \left\{ 1 + \exp(-x_i^T \hat{\beta}_2) \right\} + 2p \log(n)
 \end{aligned}$$

De acuerdo al principio del criterio de mínima información, no se rechaza  $H_0$  si

$$SIC(n) \leq \min_{p \leq k \leq n-p} SIC(k)$$

Cuando no se acepta  $H_0$ , la posición del punto de cambio  $k$  es estimada por  $\hat{k}$ , de tal suerte que

$$SIC(\hat{k}) = \min_{p \leq k \leq n-p} SIC(k).$$

Para ilustrar la aplicación del método que acaba de ser descrito se utilizarán los datos de Gurevich y Vexler (2004, [32]) mismos que se muestran a continuación.

El *Central Bureau of Statistics* de Israel llevó a cabo una investigación encaminada a determinar si hubo cambio en la distribución del número de hijos por mujer en el transcurso de unas décadas, y si fue así, en qué año ocurrió. La encuesta, basada en una muestra aleatoria del censo de población de 1995 en Israel, supone que el posible cambio en la distribución puede ser resultado de la inmigración a Israel. Por lo tanto, todas las mujeres que llegaron a Israel después del año en que ocurrió el cambio en la distribución serían designadas como nuevos inmigrantes (NI). Esta designación es relevante para ciertas aplicaciones sociológicas. Claramente, si no hubo un cambio en la distribución, entonces no es necesario definir el estatus NI.

El código de R que se utilizó para analizar estos datos y la gráfica de los resultados obtenidos se muestran a continuación. Se asume un modelo en el cual el año y número de hijos son las variables explicativas y una liga logística. Como los datos son agrupados hay que calcular el número de éxitos (mujeres con dos hijos o menos) y el número de fracasos.

LISTING 4.1. Código R método SIC

```
rm(i , m, a , Y)
# Año de llegada a Israel
i = 80:95
# Numero de mujeres que llegaron a Israel durante ese
# año
m = c(1357, 783, 776, 856, 1168, 768, 598, 829, 834,
      1678, 14235, 11956, 5578, 5300, 5642, 4133)
# Edad promedio de las mujeres llegadas en ese año
a = c(38.44, 38.51, 38.61, 38.95, 39.01, 39.06, 39.73,
```

$i$	$m_i$	$\frac{1}{m_i} \sum_{j=1}^{m_i} a_{ij}$	$\frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij}$
80	1357	38.44	0.59
81	783	38.51	0.58
82	776	38.61	0.59
83	856	38.95	0.59
84	1168	39.01	0.62
85	768	39.06	0.61
86	598	39.73	0.65
87	829	39.8	0.72
88	834	39.85	0.71
89	1678	39.92	0.8
90	14235	39.98	0.82
91	11956	39.99	0.81
92	5578	40.01	0.82
93	5300	40	0.81
94	5642	40.08	0.83
95	4133	40.43	0.82

CUADRO 2. Muestra aleatoria extraída del censo de población de 1995 en Israel

39.80, 39.85, 39.92, 39.98, 39.99, 40.01, 40.00,  
40.08, 40.43)

*# Porcentaje de mujeres llegadas con 2 hijos o menos*

$Y = \mathbf{c}(0.59, 0.58, 0.59, 0.59, 0.62, 0.61, 0.65, 0.72,$   
 $0.71, 0.80, 0.82, 0.81, 0.82, 0.81, 0.83, 0.82)$

*# Numero de mujeres llegadas con 2 hijos o menos*

$y = \mathbf{round}(m*Y, 0)$

$\mathbf{rm}(\mathbf{datos})$

```

datos = data.frame(i=i, m=m, a=a, Y=Y, y=y, yy=m-y)
rm(i, m, a, Y, y)
attach(datos)

modelo = glm(cbind(y, yy) ~ i + a,
             family = binomial(link="logit"))

model = function(b0, b1, b2){1/(1+exp(-(b0 + b1*i +
                                         b2*a)))}

par = list(b0 = modelo$coefficients[1],
           b1 = modelo$coefficients[2],
           b2 = modelo$coefficients[3])

var = list(x = "y",
           size = m,
           prob = "predicted",
           log = TRUE)

logver = likeli(model, par, var, datos, dbinom)
SIC0 = -2*logver + 3*log(n)
SIC0
detach(datos)

# *****

```

```

SICk = function(k, datos){
  n = nrow(datos)

  # primera parte
  datos1 = datos[1:k,]
  attach(datos1)
  modell = function(b0, b1, b2){1/(1+exp(-(b0 + b1*i +
  b2*a)))}

  modelo1 = glm(cbind(y, yy) ~ i + a,
    family = binomial(link="logit"),
    data = datos1)

  par = list(b0 = modelo1$coefficients[1],
    b1 = modelo1$coefficients[2],
    b2 = modelo1$coefficients[3])

  var = list(x = "y",
    size = m,
    prob = "predicted",
    log = TRUE)

  logver1 = likeli(modell, par, var, datos1, dbinom)
  detach(datos1)

  # segunda parte

```

```

datos2 = datos [(k+1):n,]
attach(datos2)
model2 = function(b0_, b1_, b2_){1/(1+exp(-(b0_ + b1_*i +
                                             b2_*a)))}

modelo2 = glm(cbind(y, yy) ~ i + a,
              family = binomial(link="logit"),
              data = datos2)

par = list(b0_ = modelo2$coefficients[1],
           b1_ = modelo2$coefficients[2],
           b2_ = modelo2$coefficients[3])

var = list(x = "y",
           size = m,
           prob = "predicted",
           log = TRUE)

logver2 = likeli(model2, par, var, datos2, dbinom)
detach(datos2)

SICk = -2*(logver1+logver2) + 6*log(n)
return (SICk)}

SIC_k = NULL
for(z in 6:13) SIC_k = c(SIC_k, SICk(z, datos))

```



```

plot(1985:1992, SIC_k, type = 'l',
     col = 'blue', lwd = 2, lty = 2,
     ylim=c(min(SIC_k)-1,SIC0+1),
     xlab="Años",
     ylab="SIC(k)", las = 2)
abline(h = SIC0, col = 'red', lwd = 2, lty = 3)

```

La gráfica que resulta de calcular el  $SIC(k)$  para  $k = 83, \dots, 93$  con el método antes descrito queda de la siguiente manera.

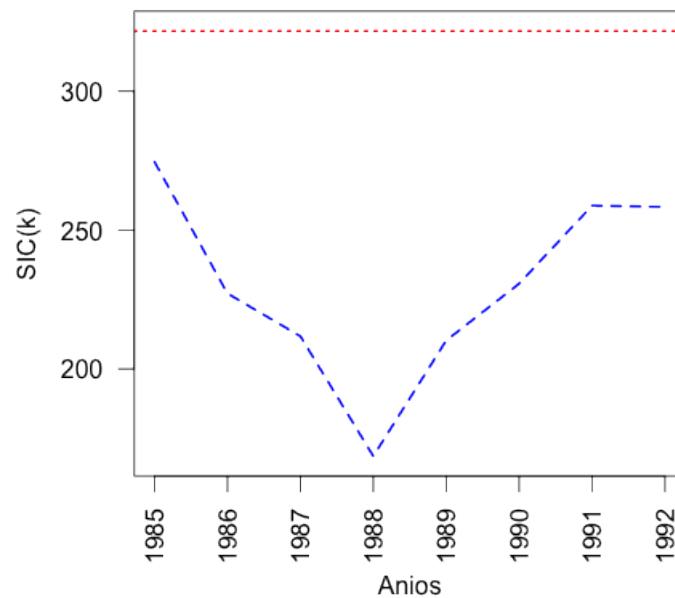


FIGURA 4.6.1. Valores de  $SIC(k)$  para  $k = 83, \dots, 93$

#### 4.7. Simulación de datos

En este apartado se realizarán simulaciones de conjuntos de datos con distribución Poisson(30) en los cuales estará plenamente identificado el punto de cambio. Posteriormente, se llevará a cabo la prueba de hipótesis con el objetivo de determinar si el método SIC es capaz de identificar el punto de cambio que se determinó en la simulación.

Se simularon 100 repeticiones de conjuntos de 50 datos y se contabilizó el número de veces que se detectó correctamente el punto de cambio bajo el modelo SIC. Se analizaron cuatro posibles escenarios: suponiendo que no existe un punto de cambio, suponiendo que el punto de cambio se ubica al inicio (observaciones 4 y 5), en la región media (observaciones 24, 25 y 26) y al final (observaciones 45 y 46) de cada conjunto de datos. La siguiente tabla resume los resultados obtenidos en las simulaciones.

Modelo SIC	Repeticiones simulación	Detecciones correctas	Tasa de acierto
Punto de cambio en obs. 4	100	35	0.35
Punto de cambio en obs. 5	100	30	0.30
Punto de cambio en obs. 24	100	41	0.41
Punto de cambio en obs. 25	100	32	0.32
Punto de cambio en obs. 26	100	36	0.36
Punto de cambio en obs. 45	100	29	0.29
Punto de cambio en obs. 46	100	24	0.24
No existe punto de cambio	100	0	0

A continuación se muestra el código R para la simulación de los datos.

LISTING 4.2. Código simulaciones

```

set .seed(781247)

# SIC bajo H0, no hay cambio en la probabilidad
sic0 = function(n, m){
  c = length(n)
  M = sum(m)
  N = sum(n)
  SIC0 = sum(lchoose(n,m)) + M*log(M/N) +
    (N-M)*log((N-M)/N)
  SIC0 = log(c) - 2*SIC0
  return (SIC0)}

# SIC bajo H1, hay cambio en la probabilidad
sic1 = function(n, m, k){
  c = length(n)
  Mk = sum(m[1:k])
  Nk = sum(n[1:k])
  Mk_ = sum(m[(k+1):c])
  Nk_ = sum(n[(k+1):c])
  SIC1 = sum(lchoose(n, m)) + Mk*log(Mk/Nk) +
    (Nk - Mk)*log((Nk - Mk)/Nk) +
    Mk_*log(Mk_/Nk_) +
    (Nk_ - Mk_)*log((Nk_ - Mk_)/Nk_)
  SIC1 = 2*log(c) - 2*SIC1
  return (SIC1)}

```

```

delta = function(n, m, k){
  c = length(n)
  M = sum(m)
  N = sum(n)
  Mk = sum(m[1:k])
  Nk = sum(n[1:k])
  Mk_ = sum(m[(k+1):c])
  Nk_ = sum(n[(k+1):c])
  DELTA = M*log(M) + (N - M)*log(N - M) - N*log(N) -
    Mk*log(Mk) - (Nk - Mk)*log(Nk - Mk) +
    Nk*log(Nk) - Mk_*log(Mk_) -
    (Nk_ - Mk_)*log(Nk_ - Mk_) + Nk_*log(Nk_) -
    (0.5)*log(1/c)
  return (DELTA)}

```

```

year = 1951:2000
i     = 1:50
n1    = 25
n2    = 25
KK    = n1 + n2 - 1
b1    = c( 7, 2.5)
b2    = c(10, 1.5)

```

```

aciertos = NULL
for(kkk in 1:100){

```

```

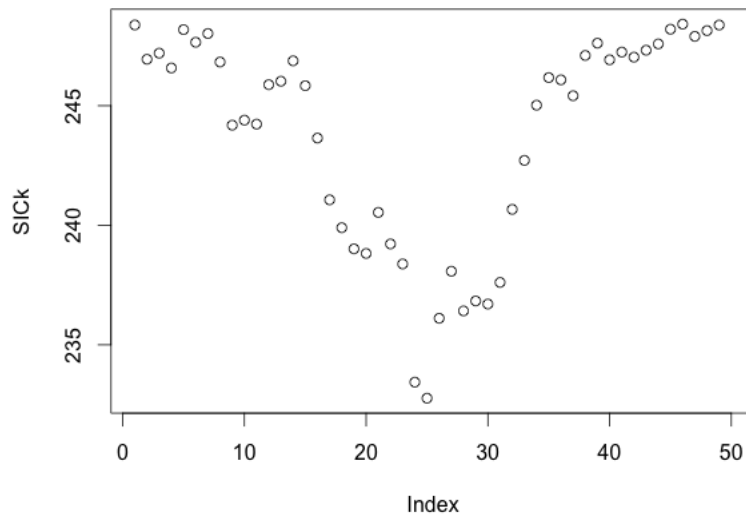
X1 = cbind(rep(1, n1), rnorm(n1, b1[1], b1[2]))
X2 = cbind(rep(1, n2), rnorm(n2, b2[1], b2[2]))
y1 = rpois(n1, 30)
y2 = rpois(n2, 30)
n = c(y1, y2)
m = c(round(X1[, 2], 0), round(X2[, 2], 0))
datos=data.frame(year, i, m, n)
deltak = NULL
SICk = NULL
for(k in 1:KK){
  deltak = c(deltak, delta(n, m, k))
  SICk = c(SICk, sic1(n, m, k))
}
aciertos = c(aciertos, which(deltak==min(deltak))==n1)
}
mean(aciertos)

SICk = c(SICk, NA)
plot(SICk)

datos = transform(datos, SICc = round(sic0(n,m), 4))
datos = transform(datos, SICk = round(SICk, 4))
datos = transform(datos, dk = round(deltak, 4))
View(datos)

```

A continuación se muestra la gráfica de los valores SIC para el caso en que el punto de cambio ocurre en la región media, específicamente en la observación número 25; se puede observar que efectivamente el algoritmo SIC detecta el valor mínimo en el dato número 25.



De esta manera concluye esta breve revisión de algunos métodos estadísticos para determinar un punto de cambio en los modelos lineales generalizados.



## Conclusiones

En este trabajo se realizó una investigación bibliográfica para conocer el estado que guarda la investigación sobre el problema de punto de cambio para modelos lineales generalizados, también conocido como cambio estructural.

Con el objetivo de entender la estructura de los modelos lineales generalizados se comenzó haciendo una exposición de dicha teoría. Posteriormente, se expusieron los métodos comunes para identificar cuándo ha ocurrido un punto de cambio en un modelo lineal, así como las pruebas de hipótesis comunes para determinar la ubicación del mismo. Finalmente, se presentaron algunos de los métodos que se conocen en la literatura para determinar si ha ocurrido un punto de cambio en un modelo lineal generalizado, así como las pruebas de hipótesis para determinar en qué momento del tiempo ocurrió éste.

Confío en que el presente trabajo pueda servir al lector interesado en mejorar la eficiencia de sus ajustes como texto introductorio a los modelos lineales generalizados, pero también para entender el problema de punto de cambio en modelos lineales generalizados.

Comentarios y dudas respecto a este trabajo son bienvenidos en mi dirección de correo electrónico: [benjaminfig1@gmail.com](mailto:benjaminfig1@gmail.com).





## Análisis de un modelo de regresión con R

### Modelo de regresión lineal

R es un popular paquete de software estadístico que se puede descargar gratuitamente de [www.r-project.org](http://www.r-project.org). Con todo y que se distribuye gratuitamente, R es un software estadístico altamente sofisticado. Entre los contribuidores del proyecto se encuentran muchos de los investigadores más importantes en estadística computacional.

Después de esta breve sinopsis pasamos a la descripción de cómo realizar el ajuste de un modelo lineal en R. El primer paso es crear un marco de datos. Suponga que los datos se encuentran en el archivo `transistor.txt`.

La instrucción para cargar un conjunto de datos es la siguiente:

```
trans <- read.table("transistor.txt", header=TRUE,
                    sep=" ")
```

El objeto `trans` contiene el conjunto de datos ya cargado en R mientras que `transistor.txt` es el archivo original con los datos. La instrucción `header=TRUE` le indica a R que los nombres de las variables se encuentran en el primer renglón del archivo original. La instrucción `sep=" "` sirve para indicar que los datos originales están delimitados por espacio. Con las siguientes líneas de código le pedimos a R que estime el modelo, que imprima la ANOVA, los coeficientes estimados y sus pruebas y por último, que imprima las medidas de influencia del modelo.

Obs.	$x_1$ (tiempo de recorrido, minutos)	$x_2$ (dosis, iones $\times 10^{14}$ )	$y$ (ganancia o $hFE$ )
1	195	4.00	1004
2	255	4.00	1636
3	195	4.60	852
4	255	4.60	1506
5	225	4.20	1272
6	225	4.10	1270
7	225	4.60	1269
8	195	4.30	903
9	255	4.30	1555
10	225	4.00	1260
11	225	4.70	1146
12	225	4.30	1276
13	225	4.72	1225
14	230	4.30	1321

CUADRO 1. Datos de la ganancia del transistor. Myers, Montgomery y Anderson-Cook (2009) describen un estudio para el cual se reporta la ganancia del transistor dentro de un dispositivo de circuito integrado entre el emisor y el colector ( $hFE$ ) junto con otras dos variables que pueden ser controladas en el proceso de deposición -tiempo de recorrido del emisor en minutos y dosis del emisor.

```
trans.model <- lm(y~x1+x2, data=trans)
summary(trans.model)
print(influence.measures(trans.model))
```

Con las siguientes instrucciones:

```
yhat <- trans.model$fit
t <- rstudent(trans.model)
qqnorm(t)
plot(yhat, t)
plot(trans$x1, t)
plot(trans$x2, t)
```

se configuran y crean las gráficas de residuos apropiadas. Generalmente R no produce gráficas satisfactorias para los residuos. Con las instrucciones

```
trans2 <- cbind(trans , yhat , t)
write.table(trans2 , "transistor_output.txt ")
```

se crea el archivo `transistor_output.txt` el cual puede exportarse a otros paquetes para la elaboración de las gráficas de los residuos.

### Modelo de regresión lineal generalizado

La función `glm` de la biblioteca MASS es la función de R más popular para analizar un modelo lineal generalizado. La sintaxis básica de la sentencia es:

```
glm(formula , family , data)
```

La especificación de la fórmula es exactamente igual que en el modelo lineal clásico.

Por ejemplo, la fórmula para el modelo  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$  es

```
y ~ x1 + x2
```

Las opciones de familia y las respectivas funciones liga disponibles son:

- binomial (logit, probit, log, log-log complementaria)
- gaussiana (identidad, log, inversa)
- Gamma (identidad, inversa, log)
- gaussiana inversa ( $1/\mu^2$ , identidad, inversa, log)
- poisson (identidad, log, raíz cuadrada), y
- quasi (logit, probit, log-log complementaria, identidad, inversa, log,  $1/\mu^2$ , raíz cuadrada. Corresponde a cuasi Poisson considerando un parámetro de sobredispersión).

Puesto que R es sensible a mayúsculas y minúsculas, la familia es **Gamma** y no **gamma**, por ejemplo. Por defecto, R utiliza la liga canónica. Para especificar la liga probit con la familia binomial, se debe agregar la opción:

```
binomial(link=probit)
```

R calcula dos clases de valores predichos distintos. Para obtener el vector de predicciones en la escala original se especifica `fit`. Para obtener el vector de predicción para el predictor lineal se debe especificar `linear.predictor`. R puede calcular los residuos puros, los de Pearson y los de la devianza. R también calcula las `influence.measures`, las cuales son las estadísticas de las observaciones borradas. Ahora ilustraremos lo anteriormente descrito mediante un ejemplo.

**Ejemplo de ajuste de un modelo lineal generalizado: regresión logística.** Para ilustrar el ajuste de un modelo lineal generalizado mediante la función `glm` estudiaremos un ejemplo. Ashford (1959) considera la proporción de trabajadores de minas de carbón que presentan síntomas de neumoconiosis severa en relación a la cantidad de años expuesto al material. El conjunto de datos es pequeño así que se presenta dentro del mismo código. A continuación se presenta el código de R.

```
>years <- c(5.8, 15.0, 21.5, 27.5, 33.5, 39.5, 46.0,
           51.5)
>cases <- c(0, 1, 3, 8, 9, 8, 10, 5)
>miners <- c(98, 54, 43, 48, 51, 38, 28, 11)
>yamat <- cbind(cases, miners-cases)
>ashford <- data.frame(yamat, years)
>analf <- glm(yamat ~ years, family=binomial,
             data=ashford)
```

```
>summary( analf)
```

```
Call:
```

```
glm(formula = ymat ~ years, family = binomial,
     data = ashford)
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-1.6625	-0.5746	-0.2802	0.3237	1.4852

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.79648	0.56859	-8.436	< 2e-16 ***
years	0.09346	0.01543	6.059	1.37e-09 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
                 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to
be 1)
```

```
Null deviance: 56.9028 on 7 degrees of freedom
Residual deviance: 6.0508 on 6 degrees of freedom
AIC: 32.877
```

```
Number of Fisher Scoring iterations: 4
```

```

>pred_prob <- analf$fit
>eta_hat <- analf$linear.predictor
>dev_res <- residuals(analf, c="deviance")
>influence.measures(analf)

```

Influence measures of

```

glm(formula = ymat ~ years, family = binomial,
     data = ashford) :

```

	dfb.1	dfb.years	dffit	cov.r	cook.d	hat	inf
1	-2.1637	2.022	-2.167	0.231	0.4736	0.317	*
2	-0.3443	0.306	-0.351	1.541	0.0581	0.214	
3	0.1379	-0.114	0.149	1.674	0.0139	0.175	
4	0.7559	-0.536	0.963	0.539	0.3677	0.186	
5	0.0785	-0.025	0.180	1.741	0.0197	0.212	
6	0.0212	-0.127	-0.348	1.665	0.0660	0.249	
7	0.0944	-0.149	-0.214	2.282	0.0272	0.387	*
8	0.1275	-0.172	-0.209	1.853	0.0258	0.260	

```

df <- dfbetas(analf)
df_int <- df[,1]
df_years <- df[,2]
hat <- hatvalues(analf)
qqnorm(dev_res)
plot(pred_prob, dev_res)

```

```
plot(eta_hat, dev_res)
plot(years, dev_res)
plot(pred_prob, df_years)
plot(hat, df_years)
ashford2 <- cbind(ashford, pred_prob, eta_hat,
  dev_res, df_int, df_years, hat)
write.table(ashford2, "ashford_output.txt")
```





## Apéndice B

### Estimación de los parámetros de un MLG

El método más común para estimar los parámetros de un MLG es el de máxima verosimilitud. Suponga que se tienen  $n$  respuestas independientes. En un modelo lineal generalizado la función de log-verosimilitud está dada por

$$(B.0.1) \quad l(\theta, \phi; y_i) = \frac{\sum_{i=1}^n y_i \theta_i - \sum_{i=1}^n b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi_i)$$

donde

$$(B.0.2) \quad \frac{\partial b(\theta)}{\partial \theta} = \mathbb{E}(Y_i) = \mu_i$$

y

$$(B.0.3) \quad g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$$

Dado que las  $\beta_j$  son los parámetros de interés, se busca resolver:

$$(B.0.4) \quad \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) = U_j \quad j = 1, \dots, p$$

donde  $x_{ij}$  es el  $j$ -ésimo elemento de  $\mathbf{x}_i^T$ .

En general, las ecuaciones  $U_j = 0$  no son lineales y se resuelven usando algún método numérico. Si se usa el método Newton-Raphson, entonces la  $m$ -ésima aproximación está dada por

$$(B.0.5) \quad \boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} - \left[ \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\beta = \beta^{(m-1)}}^{-1} \mathbf{U}^{(m-1)}$$

donde

$$(B.0.6) \quad \left[ \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]_{\beta = \beta^{(m-1)}}^{-1}$$

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

Un método alternativo es el de puntajes, que consiste en reemplazar la matriz de segundas derivadas en la ecuación (B.0.5) por la matriz de valores esperados

$$(B.0.7) \quad \mathbb{E} \left[ \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]$$

Realizando cálculos algebraicos se puede probar que la expresión anterior puede escribirse como el negativo de la matriz de varianzas y covarianzas de las  $U_j$ . Por otra parte, se sabe que

$$(B.0.8) \quad L_{jk} = \mathbb{E}[U_j U_k] = -\mathbb{E} \left[ \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]$$

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

$$(B.0.9) \quad L_{jk} = \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\text{Var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)$$

de aquí que  $\mathbf{L}$  se pueda escribir como

$$\mathbf{L} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

donde  $\mathbf{W}$  es una matriz diagonal de  $n \times n$  cuyos elementos están dados por

$$w_{ij} = \frac{1}{\text{Var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

De esta forma, si en la ecuación se sustituye a  $\mathbf{L}$  por la matriz de segundas derivadas como lo indica el método de puntajes, se tiene

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

donde  $\mathbf{L}^{(m-1)}$  denota la matriz de información evaluada en  $\boldsymbol{\beta}^{(m-1)}$ . Multiplicando ambos lados de la igualdad por  $\mathbf{L}^{(m-1)}$

$$(B.0.10) \quad \mathbf{L}^{(m-1)} \boldsymbol{\beta}^{(m)} = \mathbf{L}^{(m-1)} \boldsymbol{\beta}^{(m-1)} + \mathbf{U}^{(m-1)}$$

si se toma la segunda parte de la igualdad anterior y se reemplazan la ecuación (B.0.4) y la ecuación (B.0.9) resulta que  $\boldsymbol{\beta}^{(m)}$  es un vector con elementos

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

evaluados en  $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}$ , es decir, el lado derecho de la ecuación (B.0.10) puede escribirse como  $\mathbf{XWz}$ , donde los elementos de  $\mathbf{z}$  son

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

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

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta}^{(m-1)} = \mathbf{X}^T \mathbf{W} \mathbf{z}$$

La ecuación resultante tiene la forma de una ecuación normal para un modelo lineal de mínimos cuadrados ponderados. Por ello, si se usa el método de puntajes, la forma de estimar los parámetros de un MLG es mediante una rutina de modelos

lineales con pesos ponderados y la estimación se realizó en forma iterativa pues hay ciertos parámetros que dependen de otros. En el caso de la liga canónica el método de Newton-Raphson es equivalente al de puntajes.

## Apéndice C

### Códigos de R

A continuación se muestra el código que se utilizó para localizar el punto de cambio con los datos de Hanify (1981).

#### LISTING C.1. Código para datos de Hanify

```
# Hanify et al. (1981).  
# Datos de 1960 a 1976.  
# Region norte de Nueva Zelanda.  
# m = numero de casos de pie equinvaro congenito  
# (en el primer mes de embarazo)  
# n = numero total de nacimientos  
# Se sospecha que puede haber una alta correlacion  
# entre la incidencia de pie equinvaro y la cantidad  
# de herbicida 2,4,5-T utilizado en la region durante  
# esas fechas.  
  
year = 1960:1976  
i = 1:17  
m = c(8,3,9,12,6,4,14,12,7,5,13,11,12,9,6,13,12)  
n = c(2409,2453,2290,2171,2084,1993,2157,2091,2152,  
      2007,2027,1963,1982,1974,1932,1807,1919)  
hanify = data.frame(year, i, m, n)
```

```
View(hanify)
```

```
# SIC bajo H0, no hay cambio en la probabilidad
```

```
sic0 = function(n, m){
  c = length(n)
  M = sum(m)
  N = sum(n)
  SIC0 = sum(lchoose(n, m)) + M*log(M/N) +
    (N-M)*log((N-M)/N)
  SIC0 = log(c) - 2*SIC0
  return (SIC0)}
```

```
# SIC bajo H1, hay cambio en la probabilidad
```

```
sic1 = function(n, m, k){
  c = length(n)
  Mk = sum(m[1:k])
  Nk = sum(n[1:k])
  Mk_ = sum(m[(k+1):c])
  Nk_ = sum(n[(k+1):c])
  SIC1 = sum(lchoose(n, m)) + Mk*log(Mk/Nk) +
    (Nk - Mk)*log((Nk - Mk)/Nk) +
    Mk_*log(Mk_/Nk_) +
    (Nk_ - Mk_)*log((Nk_ - Mk_)/Nk_)
  SIC1 = 2*log(c) - 2*SIC1
  return (SIC1)}
```

```

delta = function(n, m, k){
  c = length(n)
  M = sum(m)
  N = sum(n)
  Mk = sum(m[1:k])
  Nk = sum(n[1:k])
  Mk_ = sum(m[(k+1):c])
  Nk_ = sum(n[(k+1):c])
  DELTA = M*log(M) + (N - M)*log(N - M) - N*log(N) -
    Mk*log(Mk) - (Nk - Mk)*log(Nk - Mk) +
    Nk*log(Nk) - Mk_*log(Mk_) -
    (Nk_ - Mk_)*log(Nk_ - Mk_) + Nk_*log(Nk_) -
    (0.5)*log(1/c)
  return (DELTA)}

deltak = NULL
SICk = NULL
for(k in 1:16){
  deltak = c(deltak, delta(n, m, k))
  SICk = c(SICk, sic1(n, m, k))}
deltak = c(deltak, NA)
SICk = c(SICk, NA)

hanify = transform(hanify, SICc=round(sic0(n,m),4))
hanify = transform(hanify, SICk=round(SICk,4))
hanify = transform(hanify, dk=round(deltak,4))

```



View(hanify)

## Bibliografía

- [1] Quandt, R. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, 53, 873-880.
- [2] Chow, Gregory C. (1960). Tests of equality between sets of coefficients in two linear equations. *Econometrica*, 28(3): 591-605.
- [3] Quandt, R. (1960). Tests of the hypothesis that a linear regression system obeys two different regimes. *Journal of American Statistical Association*, 55, 324-330.
- [4] Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. A*. 135, pags. 370–384.
- [5] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium of Information Theory (B. N. Petrov and E. Csaki, eds), Akademiai Kiado, Budapest, 267-81.*
- [6] Akaike, Hirotugu (1974). A new look at the statistical model identification. *IEEE Transactions On Automatic Control, Vol. AC-19, No. 6, Dec. 1974, 716-719.*
- [7] Sen, A. y Srivastava, M. (1975). On test for detecting change in mean. *Annals of Statistics* 3: 98–108.
- [8] Durbin, J. y Murphy, M. J. (1975). Seasonal adjustment based on a mixed additive-multiplicative model. *J. R. Statist. Soc. A* 139, 385-410.
- [9] Brown, R. L., Durbin, J. y Evans J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 37, No. 2: 149-192.
- [10] Beckman y Cook, (1979). Testing for two-phase regression. *Technometrics* 21 (1).
- [11] Kitagawa, G. (1979). On the use of AIC for the detection of outliers. *Technometrics*, 21, 193-199.
- [12] Wetherill, G. B. (1981). *Intermediate statistical methods*. Chapman and Hall Ltd.
- [13] Holbert, D. (1982). A Bayesian analysis of a switching linear model. *J. Econometrics*, 19, 77-87.

- [14] Worsley, (1983). Testing for a two-phase multiple regression. *Technometrics* 25(1): 35–42.
- [15] Andrews, D. F. y Herzberg, A. M. (1985). *Data: a collection of problems from many fields for the student and research worker*. Springer series in statistics. Springer-Verlag New York Inc.
- [16] Jørgensen, B. (1987). Exponential dispersion models. *J. R. Statist. Soc. B*, 49, 127-162.
- [17] McCullagh, P., y Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, segunda edición.
- [18] Horváth, L. y Shao, Q. (1993). Limit theorems for the union–intersection test. *Journal of Statistical Planning and Inference*, 44: 133–148.
- [19] Bozgodan, H., Sclove, S. L. y Gupta, A. K. (1994). AIC-Replacements for some multivariate tests of homogeneity with applications in multisample clustering and variable selection. *Proceedings of the Fisrt US/Japan Conference on the Frontiers of Statistical Modeling: An informational Approach, V. 2. Kluwer Academic Publishers, Dordrecht, 199-232*.
- [20] Csörgó, M. y Horváth, L. (1997). *Limit theorems in change point analysis*. John Wiley & Sons Ltd.
- [21] Lindsey, J. K. (1997). *Applying generalized linear models*. Springer-Verlag, New York.
- [22] Pastor, R. y Guallar, E. (1998). Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *American journal of epidemiology*, Vol. 148, No. 7: 631-642.
- [23] Venables, W. N. y Ripley, B. D. (1999). *Modern applied statistics with S-PLUS*. Springer, Statistics and Computing. Third edition.
- [24] Gill, R. (2000). Maximum likelihood estimation in generalized broken-line regression. *The Canadian Journal of Statistics*, Vol. 32, No. 3, 2004: 227-238.
- [25] Graybill, F. (2000). *Theory and application of the linear model*. Duxbury Press.
- [26] Chen, J. y Gupta, A. K. (2000). *Parametric statistical change point analysis*. Birkhäuser.
- [27] Antoch, J. y Husková, M. (2001). Permutation test in change point analysis. *Statistics and Probability Letters*, 53: 37–46.
- [28] Zeileis et ál. (2002). strucchange: an R package for testing for structural change in linear regression models. *Journal of statistical software*, Vol. 7, No. 2: 1-38.
- [29] Collett, D. (2003). *Modelling binary data*. Chapman and Hall, segunda edición.
- [30] Antoch, J., Gregoire, G. y Jurusková, D. (2004). Detection of structural changes in generalized linear models. *Statistics and Probability Letters*, 69: 315–332.
- [31] Faraway, J. (2004). *Linear models with R*. Chapman and Hall/CRC texts in statistical science series; v. 63.

- [32] Gurevich, G. y Vexler, A. (2004). Change point problems in the model of logistic regression. *Journal of statistical planning and inference*, 131: 313-331.
- [33] Dobson, A. (2008). *An introduction to generalized linear models*. Chapman and Hall, tercera edición.
- [34] Zhou, H. y Liang, K. (2008). On estimating the change point in generalized linear models. *IMS Collections. Beyond parametrics in interdisciplinary research: Festschrift in Honor of Professor Pranab K. Sen*, Vol. 1 (2008): 305-320.
- [35] Aitkin, Murray et ál. (2009). *Statistical Modelling in R*. Oxford Statistical Science Series 35. Oxford University Press Inc., New York.
- [36] Xia, Z., Guo, P. y Zhao, W. (2009). Monitoring structural changes in generalized linear models. *Communications in statistics-theory and methods*, 38: 1927-1947.
- [37] Pérez, B. y García, G. (2010). Structural change analysis in linear regression model. *Revista de matemática: teoría y aplicaciones*, 17 (2), 159-178.
- [38] Myers, R. et ál. (2011). *Generalized linear models with applications in engineering and the sciences*. John Wiley & Sons, Inc., second edition.
- [39] Park, J. H. (2011). Changepoint analysis of binary and ordinal probit models: an application to bank rate policy under the interwar gold standard. *Political analysis advance acces*.