CRITERIO DE AKAIKE PARA LA SELECCIÓN DE MODELOS CON TRANSFORMACIONES

Leonel Amaya Jiménez

Universidad Santo Tomás
Facultad de estadística
Bogotá D.C., Colombia
2018

CRITERIO DE AKAIKE PARA LA SELECCIÓN DE MODELOS CON TRANSFORMACIONES

Leonel Amaya Jiménez

Trabajo de grado presentado como requisito para optar al título de:
Profesional en estadística

Director(a):
Profesor: Andres Felipe Ortiz Rico

Universidad Santo Tomás
Facultad de Estadística
Bogotá D.C., Colombia
2018

Dedicado a

A Dios por ser siempre ese sentimiento de alegría, tranquilidad y serenidad en cada momento de esta etapa de vida que esta próxima a culminar espero ser digno por tan valioso esfuerzo.

A mis padres, mi mamá Amalia Jiménez y mi papá Leopoldo Amaya, no hay un día en el que no le agradezca a Dios el haberme colocado dado a los padres que tengo y el tesoro mas valioso son todos y cada unos de los valores que me inculcaron, dándome cada día las fuerzas necesarias para continuar en la búsqueda de la meta y que se ve cada vez mas próxima.

Por ultimo agradecer al profesor Andres Felipe Ortíz Rico que gracias a su dedicación y esfuerzo le logro llegar a buen término, por lo cuál, deseo expresar toda mi gratitud hacia usted deseándole éxito y el mejor de los augurios en su trayectoria profesional.

Leonel Amaya Jiménez

La preocupación por el hombre y su destino siempre debe ser el interés primordial de todo esfuerzo técnico. Nunca olvides esto entre tus diagramas y ecuaciones.

Albert Einstein

Agradecimientos

A mi familia, amigos y a todas aquellas personas que directa e indirectamente contribuyeron en la realización de este trabajo de grado con el cuál se culmina una nueva meta llena de retos y aprendizajes.

A la Facultad de Estadística de la Universidad Santo Tomás dado que pone en todo momento sus recursos físicos, tecnológicos y humanos, especialmente a cada uno de los profesores que hicieron de está experiencia algo gratificante al sobrepasar cada uno de los retos que me ponían en sus clases y que hacían poner a prueba todo lo aprendido y la resistencia frente al fracaso.

En particular al profesor Andres Felipe Ortiz Rico por su ayuda dedicación, paciencia, comentarios y apoyo en la realización de este proyecto, que al principio era un simple reto y que poco a poco se fue formando en un gran aprendizaje y se logro llevar a fin con sus comentarios y correcciones.

Índice general

Introducción

En el trabajo de modelación estadística, es de primordial importancia la selección del modelo, es decir, elegir dentro de un conjunto de modelos alternativos el modelo más apropiado para el conjunto de datos. Por ejemplo, en teoría de valores extremos algunas veces se desea elegir entre la distribución generalizada de valores extremos con un parámetro de forma muy pequeño o una distribución Gumbel, donde ésta última se toma como un caso límite de la primera cuando el parámetro de forma tiende a cero. En tal caso es deseable un estadístico que permita seleccionar entre un modelo u otro. Los índices AIC y BIC (Criterio de información de Akaike y criterio de información bayesiano, respectivamente) son dos criterios de uso frecuente para la selección de modelos. El AIC fue propuesto por Akaike (1974) como un estimador insesgado asintótico de la información de Kullback-Leibler esperada, entre un modelo candidato ajustado y el verdadero modelo. El BIC fue derivado por Schwarz en 1978 como una aproximación a una transformación de la probabilidad posterior de un modelo candidato.

A través del tiempo el uso de ambos criterios para la selección de modelos ha crecido significativamente. Entre algunas de las primeras aplicaciones del AIC sugeridas por el autor se encuentran: el análisis factorial, análisis de componentes principales, regresión múltiple y series de tiempo. Otras aplicaciones recientes de ambos criterios también se tienen en ecología (Anderson et al., 1994;

Johnson y Omland, 2004; Dennis et al., 2006; Ponciano et al., 2009) y bio-informática (Edwards et al., 2010; Abreu et al., 2010), por mencionar algunas.

Por otra parte, al realizar un análisis de las aplicaciones y el uso de estos dos criterios se encuentra que son aplicados sin tener en cuenta algunas características de los modelos como son las posibles transformaciones que puedan tener la variable respuesta y las implicaciones que estas puedan tener a la hora de realizar el cálculo del criterio para seleccionar el modelo adecuado, es por este motivo que se va a analizar este problema y dar una posible solución al mismo y lograr mejorar en la práctica la selección de los modelos cuando se presenta alguna transformación en la variable respuesta.

El trabajo está organizado de la siguiente manera: En el Capítulo 1 se presenta una descripción teórica sobre el criterio de Akaike (AIC) a partir de la información de Kullback-Leibler y su importancia en la selección del modelo que mejor se ajuste a los datos. Comenzando por definir los conceptos de verosimilitud y la entropía de Boltzmann que son necesarios para aplicar la distancia de Kullback-Leibler, con este recorrido llegar a las ideas fundamentales del autor. Por último, se definen a partir del criterio AIC los elementos necesarios para llegar a la expresión del criterio BIC.

En el Capítulo 2, se plantea la modificación de los criterios de AIC y BIC, para los modelos que presenten alguna transformación en la variable respuesta, dado que es el objetivo principal del trabajo, para lograr este objetivo se hace uso de conceptos como el teorema de la transformación y el Jacobiano asociado a la transformación.

En el capítulo 3, se realiza la construcción de la función en R que nos permita calcular los criterios de AIC_J y BIC_J para los modelos con la variable respuesta transformada y así poder brin-

dar una herramientas a las personas que diariamente trabajan en la selección de modelos en diferentes contextos.

En el capitulo 4, en este capítulo se presenta una aplicación de la función y la comparación entre los diferentes criterios de AIC, BIC y los criterios modificados por el Jacobiano $AICJ$ y $BICJ$, donde se tiene en cuenta las transformaciones a la variable respuesta.

Justificación

Una de las motivaciones para plantear este tema es la falta de información de la aplicación del criterio de Akaike (AIC) a los modelos con transformaciones en la variable respuesta y las implicaciones que pueda tener estas transformaciones a la hora de decidir cuál es el mejor modelo que se ajusta a los datos y como se puede llegar a solucionar este inconveniente tanto teóricamente como en la práctica, para darle más herramientas a las personas que realizan esta selección.

Se debe tener en cuenta que para la selección de modelos por el criterio de Akaike (AIC), esté criterio se basa en conceptos como la función de verosimilitud, la entropía asociada y la información contenida en el modelo, que son fundamentales para su construcción, a pesar que su expresión es bastante sencilla los conceptos que se tienen en cuenta son bastante importantes y permiten llegar a su expresión, para que sea aplicado correctamente. Por lo anterior, se tiene que la selección del mejor modelo se realiza por el menor valor obtenido de este criterio y esté sería el que mejor se ajuste a los datos. Esté criterio se puede aplicar a todo tipo de modelos lineales y series de tiempo, teniendo en cuenta las características de cada uno, que son los que al final vamos a ver en la aplicación de este trabajo.

Lo anterior nos indica el camino hacia el criterio AIC guiados por el concepto de verosimilitud, sobre todo si tenemos en cuenta

que la función de verosimilitud se caracteriza por ser muy sensible a pequeños cambios en los parámetros lo que la hace adecuada para medir la bondad del ajuste, en donde radica la importancia de contemplar los modelos con la variable respuesta transformada y que cambios se obtienen al comparar con modelos sin la variable transformada.

De igual manera, por la importancia del criterio de Akaike (AIC) para la selección de modelos en diferentes contextos, es necesario contemplar todas las posibles variaciones y transformaciones para los mismos, donde la dificultad de la selección de un modelo reside en obtener la "verosimilitud"del mismo (Akaike, 1978a), al tener en cuenta estas dificultades se realiza la exploración y adecuación teórica del criterio donde se tienen en cuenta estas transformaciones de la variable respuesta.

Para terminar, se realiza la construcción de una función en el software R en donde se contemplen dichas transformaciones en la variable respuesta, dado que las funciones encontradas en R no las contemplan y por lo que en la práctica las personas que hacen uso de ellas, no contemplan el error que se puede cometer a la hora de seleccionar modelos con alguna transformación en su variable respuesta.

Objetivos

Objetivo general

Determinar las diferencias que se pueden encontrar en la selección de modelos por el criterio de Akaike (AIC) cuando la variable respuesta tiene transformaciones y cuando no las tiene.

Objetivos Específicos

1. Evidenciar las malas prácticas generadas al utilizar el criterio de Akaike (AIC) cuando la variable respuesta ha sido transformada.

2. Proponer una modificacióon al criterio de Akaike (AIC) que permita comparar modelos con transformaciones en la variable respuesta..

3. Elaborar una función en R la cual permita realizar el cálculo del criterio del Akaike de diferentes modelos transformados.

Capítulo 1

Marco Teórico

En este Capítulo, se realiza un recuento del surgimiento de los criterios AIC y BIC, los cuales son los mas utilizados a la hora de seleccionar el modelo mas apropiado entre un conjunto de modelos propuestos para los datos que se estén estudiando. Primero se hace mención del surgimiento del modelo AIC con su construcción teórica hasta llegar al modelo BIC.

1.1. Construcción del criterio AIC

La definición del AIC se relaciona con conceptos estadísticos tan importantes como son la función de verosimilitud, la entropia asociada y la información contenida en el modelo. por lo tanto cuando hablamos del criterio de Akaike AIC, debemos tener en cuenta los diferentes conceptos mencionados y que han sido utilizados por los diferentes autores que lo han planteado como lo realizó Akaike (1974) y las relaciones entre ellos, como se muestra en la siguiente gráfica.

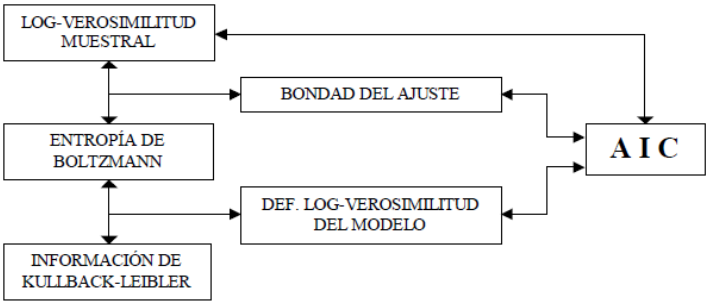

Figura 1.1: Interacción AIC y otros conceptos.

Por lo tanto en esta sección se discute el planteamiento del criterio de AIC partiendo de algunos de los conceptos anteriores, como la información de $Kullback - Leibler(K - L)$ el cuál es un criterio para la evaluación estadística de modelos que aproximan a la verdadera distribución de probabilidad que genera los datos, se dan algunas de sus propiedades teóricas más importantes que están relacionadas al criterio AIC.

Se parte del criterio de información estadística de $(K-L)$ hasta llegar a la presentación del criterio de Akaike (AIC). También se describen las ideas de Akaike (1974), para la derivación del criterio AIC quedando como punto de partida para definir el criterio de información Bayesiana (BIC) con todos sus elementos.

1.2. Comentarios del criterio AIC

Siempre se ha buscado resolver el problema de la elección del modelo más apropiado para un conjunto de datos, por lo que algu-

nos autores han desarrollado diferentes criterios basados en la teoría de la información de $Kullback - Leibler(K - L)$ la cuál permite realizar la identificación del modelo de una manera más sencilla y automática, entre estos criterios se destacan el AIC ("Akaike's Information Criterion") y el criterio BIC ("Bayesian Information Criterion"), estos dos criterios son lo que vamos a tener en cuenta a lo largo del trabajo, dado que son los más utilizados según la teoría.

Dado que el criterio más utilizado es el conocido como AIC (Akaike, 1974), se parte de que esté criterio presenta una formulación simple y una fácil aplicación: donde una vez calculado el criterio AIC para cada modelo se elige aquel cuyo AIC es mínimo. Por otra parte, las implicaciones que de él se derivan han abierto una nueva perspectiva en el área de la identificación no solo aplicable a las series temporales sino a cualquier otra técnica de modelización, como se puede observar en los campos de: modelos en análisis factorial, análisis de la varianza y regresión múltiple, entre otras técnicas.

Una vez revisadas las implicaciones del criterio de información de Akaike (AIC), vamos a remarcar algunas de sus particularidades que se han logrado encontrar a lo largo de la historia y por los diferentes autores que lo han aplicado:

- El AIC mide, como ya hemos dicho, el desajuste entre una distribución hipotética y una distribución teórica.
- La minimización del criterio AIC supone realizar de manera simultánea la selección del modelo y la estimación de los parámetros.
- El AIC sigue el principio de parsimonia: Cuando el número de parámetros de un modelo k aumenta el AIC también, por tanto escoger el modelo que tiene el mínimo AIC supone elegir el modelo con el menor número de parámetros posible.

- La minimización del AIC esta de acuerdo con el principio de maximización de la entropía: [1]
- Algunos resultados obtenidos al utilizar el criterio de AIC como criterio de identificación demuestran que este no siempre consigue minimizar el número de parámetros del modelo y lleva en algunas ocasiones a sobrestimar el modelo.
- El cálculo del AIC supone una aproximación válida cuando el tamaño de la muestra es grande, en muestras finitas el valor del AIC es sólo aproximado.
- Para la aplicación del AIC no es necesario que los diferentes modelos que compiten para su selección estén anidados entre sí.
- Destacar que el AIC es una medida global de la bondad del ajuste del modelo, su cálculo se realiza desde un punto de vista predictivo lo que supone que los modelos identificados a partir de este criterio tienen un buen comportamiento respecto a la predicción.

1.2.1. Verosimilitud y Entropía de Boltzmann

El estudio de la función de verosimilitud de la distribución lleva a Akaike a relacionar este concepto con el de entropía de Boltzmann. Este último, partiendo del estudio de la distribución de la energía de las moléculas de los gases y propuesto en 1877 para definir la entropía en términos estadísticos haciendo uso de las densidades de las distribuciones, tomando como densidad secundaria

[1] Principio de maximización de la entropía (Akaike, 1977): "Todas las actividades estadísticas tienen como fin maximizar la entropía esperada de la distribución de predicción"

f respecto a una densidad primaria g. El razonamiento que Boltzmann utilizó se muestra a continuación:

Sea x una variable aleatoria con valores $x_1, x_2, ..., x_k$ y probabilidades asociadas $q_1, q_2, ..., q_k$ y con las siguientes condiciones $q_i > 0$ y $q_1 + q_2 + ... + q_k = 1$. A partir de la observación de las variables obtienen las frecuencias $n_1, n_2, ..., n_k$ tal que $n_1 + n_2 + ... + n_k = n$, de tal manera que la verosimilitud esta dada por:

$$\Omega = \frac{n!}{n_1!n_2!...n_k!} * q_1^{n_1} * q_2^{n_2} * \cdots * q_k^{n_k} \tag{1.1}$$

Si tomamos logaritmo natural obtendríamos la expresión de la log-verosimilitud $log\ \Omega$, y utilizando la siguiente igualdad

$$logn! = n * log(n) - n$$

se demuestra que[2]:

$$log\ \Omega = -n\ \sum_{i=1}^{k} \frac{n_i}{n}\ log\left(\frac{n_i}{nq_i}\right) \tag{1.2}$$

en esta expresión, si suponemos $pi = \frac{n_i}{n}$ se puede reescribir como:

$$log\ \Omega = -n\ \sum_{i=1}^{k} p_i\ log\left(\frac{p_i}{q_i}\right) \tag{1.3}$$

Boltzmann define la entropía de la distribución secundaria p respecto a la primaria q como:

$$B(p;q) = -\sum_{i=1}^{k} p_i\ log\ \left(\frac{p_i}{q_i}\right)$$

[2]Ver Tong, H. (1990). Non-linear time series: a dynamical system approach. Oxford: Oxford University Press, pg. 282

después al considerar las distribuciones en términos de densidades f y g se llega a la siguiente definición:

$$B(g;f) = -\int g(x)\ log\left(\frac{g(x)}{f(x)}\right) dx$$

Donde $log\frac{g(x)}{f(x)}$, se considera como una medida entre la diferencia de las dos densidades f y g, donde si son iguales la expresión anterior es cero y entre mas diferentes sean el logaritmo natural va creciendo hasta no tener convergencia.

De esta manera queda establecido que log-verosimilitud y entropía se relacionan a partir de la siguiente expresión: [3]

$$log\ \Omega = n\ B(g;f) \tag{1.4}$$

Desde el punto de vista de la teoría de la información, $-B(g;f)$ se puede interpretar como una medida de la variación de información al pasar de la información inicial o a priori, a la final o a posteriori, caracterizadas cada una de ellas por las funciones de densidad $f(x)$ y $g(x)$, respectivamente. Kullback (1951), define la información de Kullback-Leibler $I(g;f)$ a partir de la siguiente expresión:

$$I(g;f) = \int g(x)\ log\left(\frac{g(x)}{f(x)}\right) dx \tag{1.5}$$

lo que nos permite establecer la siguiente igualdad:

$$I(g;f) = -B(g;f) \tag{1.6}$$

Si consideramos g la distribución teórica y f la función de densidad obtenida para una muestra, podemos entender la variación de la información entre ambas distribuciones como una medida de la bondad del ajuste de f respecto a g, lo que nos permite relacionar el concepto de entropía de Boltzmann con el de bondad del ajuste.

[3] Ver Raymond, Jose luis y Uriel, Esequiel., 1987. investigación econométrica aplicada un caso de estudio. alfa centauro S.A., pg. 224-227.

1.2.2. Distancia de Kullback-Leibler

Nuestro punto de partida para llegar a la formulación del criterio AIC, el cual ha sido ampliamente usado para la selección de modelos estadísticos es la estimación de la información de Kullback-Leibler. Está información es definida en la ecuación (1.7) presentada a continuación, la cuál es considerada como una medida de bondad de ajuste del modelo propuesto $f(x)$ hacia el modelo verdadero $g(x)$[4].

$$I(g,f) = \int \left[log\left(\frac{g(x)}{f(x)}\right)\right] g(x)dx = E_x\left[log\left(\frac{g(x)}{f(x)}\right)\right] \qquad (1.7)$$

donde E_x denota que la esperanza es tomada con respecto a la variable aleatoria x.

Algunas de las propiedades de la información de $K-L$ son:

- $I(g,f) \geq 0$
- $I(g,f) = 0 \Leftrightarrow g(x) = f(x)$,

es decir, la información de $K-L$ siempre es positiva, excepto cuando las dos distribuciones son iguales [5]. De aquí, está información puede interpretarse directamente como una "distancia" entre dos modelos, en este caso $f(x)$ y $g(x)$, aunque estrictamente no lo sea, ya que la medida de f a g no necesariamente es la misma que de g a f.

Aunque la información $K-L$ es bastante razonable para evaluar qué tan adecuado es un modelo dado, en la práctica es bastante limitada ya que casi siempre se desconoce la verdadera distribución que genera los datos, lo cuál impide calcular (1.7), donde se hace

[4]ver Shibata (1995)

[5]Burnham y Anderson, 2002, página 430

necesario el conocimiento de está distribución que en algunos casos simplemente no es posible obtener por métodos sencillos.

Por lo tanto si le hacemos una modificación a la ecuación (1.7) está se puede re-expresar como:

$$I(g, f) = E_X[log(g(x))] - E_x[log(f(x))] \tag{1.8}$$

de donde se tiene que, para la comparación de diferentes modelos es suficiente considerar $E_x[log(f(x))]$, ya que $E_X[log(g(x)]$ es un término común que puede ser ignorado. El segundo término de (1.8) se conoce como log-verosimilitud esperada para el modelo f. Así, de un conjunto de modelos candidatos el modelo que tenga mayor log-verosimilitud esperada es el que corresponde al que tiene menor información de $K-L$, y en consecuencia es el mejor modelo.

Si el modelo $f(x)$ está completamente especificado, entonces obsérvese que un estimador natural y consistente para $E_x[log(f(x))]$ es:

$$\frac{1}{n}\sum_{i=1}^{n} log(f(x_i)) \tag{1.9}$$

donde $x_i = (x_1, \ldots, x_n)'$ es una muestra aleatoria de la verdadera distribución $g(x)$. Teniendo que (1.9) es un estimador insesgado para $E_x[log(f(x))]$, y cuando n tiende a infinito converge a $E_x[log(f(x))]$ con probabilidad 1, asumiendo que $|E_x[log(f(x))]| < \infty$, esta convergencia se da bajo estas condiciones, para ver mas detalles de esta convergencia, (P. Billingsley, 1999).

1.2.3. Estimador de Máxima de Verosimilitud (EMV)

Teniendo en cuenta la definición de $K - L$ y dado que en la práctica es complejo tener el modelo completamente especificados

o definido para los datos y sus parámetros son desconocidos. Lo habitual es asumir un modelo paramétrico $\{f(x|\theta); \theta \in \Theta \subset \mathbb{R}^p\}$ y luego estimar los parámetros θ por el "método de máxima verosimilitud". Aún más, debido a que muchas veces no se tiene bien identificado un modelo, lo usual es proponer varios modelos para el mismo problema y posteriormente hacer el cálculo del AIC y del BIC, para seleccionar el que menor valor del criterio tenga y sera el que mejor se ajuste a los datos.

Se sabe que bajo ciertas condiciones de regularidad el estimador de máxima verosimilitud (EMV) de $\theta, \widehat{\theta} = \widehat{\theta}_n(X_n)$, converge en probabilidad a $\theta_0 = \underset{\theta\in\Theta}{arg\,mín}\ I[g(\cdot), f(\cdot|\theta)]$

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_0 \tag{1.10}$$

$f(x|\theta_0)$ es llamada la mejor aproximación a $g(x)$ (Claeskens y Hjort, 2008). Así, como en la práctica θ_0 es imposible de calcular ya que no se conoce $g(x)$, el EMV de θ proporciona la mejor aproximación paramétrica a la verdadera distribución g dentro de las distribución $f(x|\theta)$. Para cuando θ es un escalar y $g(x) = f(x|\theta^*)$, para algún $\theta^* \in \Theta$, para ver mayores detalles sobre el resultado de la ecuación (1.10) se puede consultar, Wasserman (2004). Una forma alternativa de relacionar los conceptos de log-verosimilitud e información de $K - L$ es la siguiente:

Maximizar la log-verosimilitud $l_n(\theta) =: \sum_{i=1}^n\ log(f(x_i|\theta))$ es equivalente a maximizar:

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ log\left(\frac{f(x_i|\theta)}{g(X_i)}\right) \tag{1.11}$$

y además por ley débil de los grandes números, se obtiene la

siguiente expresión:[6].

$$M_n(\theta) \xrightarrow[n\to\infty]{P} E_g\left[log\left(\frac{f(x|\theta)}{g(X)}\right)\right] = -E_g\left[log\left(\frac{g(x)}{f(x|\theta)}\right)\right] = -I[g(\cdot), f(\cdot|\theta)]. \tag{1.12}$$

De aquí, $M_n(\theta) \approx -I[g(\cdot), f(\cdot|\theta)]$. Pero como se mencionó anteriormente, $I[g(\cdot), f(\cdot|\theta)]$ se minimiza en θ_0, así que $-I[g(\cdot), f(\cdot|\theta)]$ es maximizada en θ_0. De esta forma, se espera que el θ que maximiza $M_n(\theta), \hat{\theta}_n$, tienda a θ_0.

1.2.4. Criterio de Akaike (AIC)

Este criterio, definido por Akaike (1974) como "An Information Criterion", se basa en la medida de información de Kullback-Leibler (1951), la cuál permite interpretar la distancia entre dos distribuciones a partir de la log-verosimilitud de un modelo, por lo que anteriormente se realizó la descripción de la medida de información de $K - L$.

En la derivación del AIC por Akaike (1974), de entrada se considera la situación donde $g(x) = f(x|\theta_0)$, es decir, la densidad de probabilidad $g(x)$ verdadera se encuentra incluida en la familia dada, $\{f(x|\theta) : \theta \in \Theta \subset \mathbb{R}^p\}$. Si $K(\theta_0, \theta)$ denota $I(g(\cdot), f(\cdot|\theta)$ y además si θ está suficientemente cercano a θ_0, $K(\theta_0, \theta)$ se puede aproximar a la siguiente expresión (Kullback (1977)): [7].

$$K(\theta_0, \theta_0 + \Delta\theta) \approx \frac{1}{2}\Delta\theta' I(\theta_0)\Delta\theta \tag{1.13}$$

[6]Ver ley de los grandes números y teorema central del limite en $http : //www.depfe.unam.mx/ramirez - cruz/publicaciones/ramirez\_2016\_articulo - stata.pdf$

[7]ver Kullback, S. (1997). Information Theory and Statistics, Dover Publications, New York. página 28

donde:

$$I(\theta_0) = \int g(x) \frac{\partial log[f(x|\theta_0)]}{\partial \theta} \frac{\partial log[f(x|\theta_0)]}{\partial \theta'} dx \tag{1.14}$$

y

$$\frac{\partial log[f(x|\theta_0)]}{\partial \theta} := \left. \frac{\partial log[f(x|\theta_0)]}{\partial \theta} \right|_{\theta=\theta_0} \tag{1.15}$$

Una forma alternativa de obtener la aproximación en la ecuación (1.13) es primero expandir por serie de Taylor de segundo grado la función $ln[f(x|\theta_0 + \Delta\theta)]$ alrededor de θ_0, restar $log[f(x;\theta_0)]$ y calcular las esperanzas de estos términos, donde para este cálculo es necesario tener en cuenta el siguiente resultado:

$$E\left[\left. \frac{\partial log[f(x|\theta_0)]}{\partial \theta} \right|_{\theta=\theta_0} \right] = 0 \tag{1.16}$$

Por la ecuación (1.13), obsérvese que cuando el EMV $\hat{\theta_n}$ de θ se encuentra cerca a θ_0, la desviación de la distribución ajustada $f(x|\hat{\theta}_n)$ a la verdadera $f(x|\theta_0)$ puede medirse por $(\hat{\theta}_n - \theta_0)' I(\theta_0)(\hat{\theta}_n - \theta_0)/2$.

Ahora, considérese el caso en que θ es restringido a un subespacio de menor dimensión $\Theta_k \subset \Theta_p$ que no incluye a θ_0. Si $\hat{\theta}_{n,k}$ denota el EMV de θ en Θ_k, y $\Theta_k = \underset{\theta \in \Theta_k}{arg\,mín} I[g(\cdot), f(\cdot|\theta)]$ está suficientemente cerca a θ_0, entonces bajo condiciones de regularidad la distribución asintótica de $\sqrt{n}(\hat{\theta}_{n,k} - \theta_k)$ es aproximadamente $MVN[0, I(\theta_0) - 1]$, y de aquí la distribución de $n(\hat{\theta}_{n,k} - \theta_0)' I(\theta_0)(\hat{\theta}_{n,k} - \theta_0)$ es aproximadamente Ji-cuadrada no central $\chi^2_{k,\lambda}$, con parámetro de no centralidad $\lambda = n(\theta_k - \theta_0)' I(\theta_0)(\theta_k - \theta_0)$. Así de esto y de la ecuación (1.13)

$$E\left[2nK(\theta_0, \hat{\theta}_{n,k})\right] \approx n(\theta_k - \theta_0)' I(\theta_0)(\theta_k - \theta_0) + k \tag{1.17}$$

donde k es la dimensión de Θ_k o el número de parámetros. De esto, si se cuenta con una estimación de $n(\theta_k - \theta_0)'I(\theta_0)(\theta_k - \theta_0)$ y se tienen varios modelos, es natural adoptar como mejor el que tenga un valor $E[2nK(\theta_0, \hat{\theta}_{n,k})]$ más pequeño.

Para obtener una estimación de $n(\theta_k - \theta_0)'I(\theta_0)(\theta_k - \theta_0)$ nótese que como asintóticamente $2[l_n(\hat{\theta}_n) - l_n(\hat{\theta}_n, k)] \sim \chi^2_{p-k,\lambda}$ (donde $\lambda = n(\theta_k - \theta_0)'I(\theta_0)(\theta_k - \theta_0)$), y $2[l_n(\hat{\theta}_n) - l_n(\theta_0)] \sim \chi^2_p$, entonces:

$$\begin{aligned} E\left[2[l_n(\theta_0) - l_n(\hat{\theta}_{n,k})]\right] &= E\left[2[l_n(\hat{\theta}_n) - l_n(\hat{\theta}_{n,k})] - 2[l_n(\hat{\theta}_n) - l_n(\theta_0)]\right] \\ &\approx E\left[2[l_n(\hat{\theta}_n) - l_n(\hat{\theta}_{n,k})]\right] - E\left[2[l_n(\hat{\theta}_n) - l_n(\theta_0)]\right] \\ &= (p - k + \lambda) - p \\ &= n(\theta_k - \theta_0)'I(\theta_0)(\theta_k - \theta_0) - k \end{aligned} \tag{1.18}$$

De aquí, un estimador aproximadamente insesgado para

$$n(\theta_k - \theta_0)'I(\theta_0)(\theta_k - \theta_0)$$

está dado por:

$$2[l_n(\theta_0) - l_n(\hat{\theta}_{n,k})] + k$$

y en consecuencia, por la ecuación (1.17), un estimador aproximadamente insesgado para $E[2nK(\theta_0, \hat{\theta}_k)]$ es:

$$\begin{aligned} \hat{E}[2nK(\theta_0, \hat{\theta}_k)] &=: 2[l_n(\theta_0) - l_n(\hat{\theta}_{n,k})] + k + k \\ &= 2[l_n(\theta_0) - l_n(\hat{\theta}_{n,k})] + 2k \end{aligned} \tag{1.19}$$

Así, por lo mencionado anteriormente, la comparación de varios modelos candidatos se puede llevar a cabo con la ecuación (1.19), pero como $l_n(\theta_0)$ es un término común puede ignorarse, y tal comparación resulta equivalente a realizarla con:

$$AIC = -2l(\hat{\theta}_{n,k}) + 2k \tag{1.20}$$

el llamado criterio de información de Akaike.

1.3. Criterio Bayesiano de Schwarz (BIC)

Dentro del conjunto de los criterios utilizados en la práctica para la selección de modelos, el AIC es sin duda el mas conocido, pero existen diferentes criterios construidos con el mismo fin, pero con características diferentes. Uno de ellos y que se relaciona directamente con el criterio AIC es el criterio BIC ("Bayesian Information Criterion") derivado del criterio AIC al introducir una modificación bayesiana en la construcción del criterio AIC revisada anteriormente.

El criterio BIC fue propuesto por Schwarz (1978) y ha sido uno de los métodos más populares usado para la selección de modelos. Este es un criterio de evaluación de modelos en términos de sus probabilidades posteriores. Una motivación detrás del criterio BIC junto con un bosquejo de la derivación de este se presenta en seguida.

Se tiene el problema de seleccionar, dentro de un conjunto de r modelos el que mejor describa a un conjunto de datos $x_n = (x_1, \ldots, x_n)'$, donde la densidad condicional de estos dado por el i-iésimo modelo candidato (M_i) y su correspondiente vector de parámetros (θ^i), está dada por $f_i(x_n|\theta^i) =: f(x_n|M_i, \theta^i)(\theta^i \in \Theta_i \subset \mathbb{R}^p)$.

Sea $\pi_i(\theta)$ la densidad a priori para el vector θ^i dado el modelo M_i, y $p(M_i)$ una densidad de probabilidad discreta a priori que asigna probabilidad positiva a cada uno de los modelos $M_1, \ldots, M_r$. Dado estos supuestos, por el teorema de Bayes de probabilidad total, la probabilidad a posteriori del i-ésimo modelo está dada por:

$$\begin{aligned} P(M_i|x_n) &= \frac{P(M_i)f(x_n|M_i)}{f(x_n)} \\ &= \frac{P(M_i)\int_{\Theta_i} f(x_n|M_i,\theta)\pi_i(\theta)d\theta}{f(x_n)} \\ &= \frac{P(M_i)\int_{\Theta_i} f_i(x_n|\theta)\pi_i(\theta)d\theta}{f(x_n)} \\ &= \frac{(M_i)f_i(x_n)}{f(x_n)} \end{aligned} \tag{1.21}$$

donde:

$$f_i(x_n) = \int_{\Theta_i} f_i(x_n|\theta)\pi_i(\theta)d\theta \tag{1.22}$$

La probabilidad condicional $P(M|x_n)$ dada en (1.21) se interpreta como la probabilidad de que los datos sean generados por el modelo M_i dado que se ha observado x_n, entonces desde un punto de vista Bayesiano es natural adoptar como mejor modelo el que tenga mayor probabilidad a posteriori.

Para comparar diferentes modelos a través de sus probabilidades a posteriori, $f(x_n)$ no es importante ya que es un término común para todos los modelos, y al ignorarse, tal comparación resulta equivalente a realizarla con solamente el numerador de (1.21), es decir, usando $P(M_i)f_i(x_n)$. Además, si asume que las probabilidades a priori $P(M_i)$ son iguales para todos los modelos, entonces el modelo que maximice (1.22) es el que debe seleccionarse como el mejor.

En la práctica los valores de (1.22) son difíciles de calcular, además de que para ello se requiere de la especificación de las densidades a priori $\pi(\theta)$. Una aproximación del logaritmo de (1.22)) está dada por

$$\begin{aligned} log[f_i(x_n)] &\approx log[f_i(x_n|\hat{\theta}_{n,k})] - \frac{k_i}{2} log(n) \\ &= l_{n,i}(\hat{\theta}_{n,k}) - \frac{k_i}{2} log(n) \end{aligned} \tag{1.23}$$

donde $\hat{\theta}_{n,k}$ es el EMV de θ^k y $l_{n,k}(\theta_k) = log[f_k(x_n|\theta_k)]$ es la log-verosimilitud correspondiente al modelo M_i (Konishi y Kitagawa, 2008; Claeskens y Hjort, 2008; Burnham y Anderson,2002). Así, ya que la función logaritmo es monótona creciente, seleccionar como mejor modelo el que maximice (1.22) equivale aproximadamente a seleccionar al que maximice (1.23), lo que a su vez equivale a seleccionar el modelo que minimice

$$BIC = -2l(\hat{\theta}_{n,k}) + k\ log(n) \tag{1.24}$$

el llamado criterio de información Bayesiana.

Para finalizar, si comparamos el AIC y el BIC vemos que la diferencia básica entre ambos criterios radica en que este último penaliza más los modelos con un número mayor de parámetros estimados (debido a la sustitución del 2 por $log(n)$), obteniéndose así modelos de orden inferior a los obtenidos a partir del AIC y corrigiendo, por tanto, la tendencia a la sobrestimación observada en éste último.

Capítulo 2

Modificación de los criterios AIC y BIC

En esté capítulo se presentan los cálculos necesarios para realizados para modificar los criterios AIC y BIC, donde se busca incluir las transformaciones que se le realizan a la variable respuesta, para nuestro caso vamos analizar las transformaciones $z = log(y)$ y $z = y^{\lambda}$ con λ conocido y positivo que son las transformaciones más usuales en la práctica, dado que estas también hacen que los valores obtenidos de los criterios varíen de acuerdo a la transformación que se utilice y en la revisión realizada de las propuestas teóricas no se tienen en cuenta dichas transformaciones, llegando a tener errores en la selección del modelo más apropiado para los datos.

2.1. Modificación de la Log-verosimilitud

Para realizar la modificación de la log-verosimilitud se deben tener en cuenta algunos conceptos y propiedades sobre la función de densidad, mostradas a continuación.

Sea x una variable aleatoria con función de densidad $f(x)$ y

$g(x)$ una transformación de x. Se supone que tanto g como su inversa son continuas y crecientes dado por que las transformaciones escogidas cumplen con estas características.

Se tiene como supuesto sobre a función de densidad, que una función de densidad de probabilidad es una función f cuyo dominio es un intervalo (a, b) y tiene las siguientes propiedades:

- $f(x) \geq 0$ para toda x
- $\int_a^b f(x)dx = 1$

Permitimos que sea infinita a, b o los dos de modo que la integral sería impropia.

Otro concepto importante y necesario para la modificación de la log-verosimilitud se enuncia a continuación:

Teorema de inversión: Sea X una variable aleatoria con función de densidad de probabilidad acumulada F, continua e invertible, y sea F^{-1} su función inversa. Entonces, la variable aleatoria $U = F(X)$ tiene distribución uniforme en $(0; 1)$. Como consecuencia, si U es una variable aleatoria uniforme en $(0; 1)$ entonces la variable aleatoria $X = F^{-1}(U)$ satisface la distribución F y conserva sus propiedades.

Sea $z = g(y) \implies g^{-1}(z) = y$

Por definición de función de densidad se tiene:

$$f_z(z) = P(Z \leq z) = P(g(y) \leq z) = (y \leq g^{-1}(z))f_y(g^{-1}(z))$$

Y por el teorema de la transformación [1] se obtiene la densidad de z aplicando la transformación, se tiene la siguiente expresión:

[1] $f_z(z) = f_y(h(z))|J(z)|$, Blanco, L (2004)

$$f_z(z) = f(y)(g^{-1}(z))\frac{\partial g^{-1}(z)}{\partial z} \tag{2.1}$$

Donde: $J(z) = \frac{\partial g^{-1}(z)}{\partial z}$ es el Jacobiano de la transformación.

Ahora se necesita la densidad conjunta de $z_1, z_2, \ldots, z_n$ con $z_i = g(y_i)$ con las observaciones $z_1, z_2, \ldots, z_n$ independientes e igualmente distribuidas, a partir de la función de densidad conjunta la verosimilitud esta dada por la siguiente expresión:

$$f(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} f_z(z_i) = \prod_{i=1}^{n} f_y(g^{-1}(z_i))\frac{\partial g^{-1}(z_i)}{\partial z_i} \tag{2.2}$$

Ahora se obtiene la log-verosimilitud

$$l(z_1, z_2, \ldots, z_n) = \sum_{i=1}^{n} \left\{ log(f_y(g^{-1}(z_i)) + log\frac{\partial g^{-1}(z_i)}{\partial z_i} \right\} \tag{2.3}$$

Dado que $g^{-1}(z_i) = y_i$, la log-verosimilitud es la siguiente:

$$\begin{aligned} l(z_1, z_2, \ldots, z_n) &= \sum_{i=1}^{n} \left\{ log(f_y(y_i)) + log\left(\frac{\partial y_i}{\partial z_i}\right) \right\} \\ &= \sum_{i=1}^{n} log(f_y(y_i)) + \sum_{i=1}^{n} log\left(\frac{\partial y_i}{\partial z_i}\right) \\ &= log(f(y_1, y_2, \ldots, y_n)) + \sum_{i=1}^{n} log\left(\frac{\partial y_i}{\partial z_i}\right) \end{aligned} \tag{2.4}$$

Donde, el primer término de la la expresión anterior es la log-verosimilitud de los datos originales y el segundo término es el Jacobiano de la transformación (penalización por la transformación).

2.1.1. Transformación $z_i = log(y_i)$

Lo anterior fue el desarrollo en forma general aplicado para cualquier trasformación de la variable y, ahora se vera para las transformaciones de $z = log(y)$ y $z = y^{\lambda}$.

Sea $z = log(y)$ entonces, $g^{-1}(y) = e^z$. La función de densidad para una sola observación esta dada por:

$$f_z(z) = f_y(e^z) = e^z f_y(e^z)$$

Si tenemos una muestra $y_1, y_2, \ldots, y_n$, la transformación aplicada $z_i = log(y_i) \rightarrow g^{-1}(y_i) = e^{z_i}$, calculamos su verosimilitud y la log-verosimilitud de la transformación obteniendo lo siguiente:

$$f(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} e^{z_i} f_y(e^{z_i}) \tag{2.5}$$

$$l(z_1, z_2, \ldots, z_n) = \sum_{i=1}^{n} \{z_i + log(f_y(e^{z_i}))\} \tag{2.6}$$

Por ejemplo, si $y \sim N(\mu, \sigma^2)$, entonces su función de densidad está dada por:

$$f_y(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \tag{2.7}$$

Y la función de distribución de $g^{-1}(y_i)$ está dada por:

$$f_y(e^{z_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(e^{z_i} - \mu)^2\right\}$$

la log-verosimilitud queda de la siguiente forma:

$$log(f_y(e^{z_i})) = -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(e^{z_i} - \mu)^2 \tag{2.8}$$

Reemplazando $log(f_y(e^{z_i}))$ en la ecuación (2.6), obtenemos la log-verosimilitud para la transformación $z_i = log(y_i)$.

$$l(z_1, z_2, \ldots, z_n) = \sum_{i=1}^{n} \left\{ z_i + -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(e^{z_i} - \mu)^2 \right\} \quad (2.9)$$

para los datos originales se tiene:

$$l(y_1, y_2, \ldots, y_n) = -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 + n\overline{log(y)} \quad (2.10)$$

Tenemos que los dos primeros términos de la ecuación (2.10) son la log-verosimilitud de los datos originales y_i y el término $n\overline{log(y)}$ es la penalización por la trasformación realizada a la variable respuesta (y), donde $\overline{log(y)}$ es el promedio de los logaritmos de los datos originales, se puede observar que en la ecuación se hace uso de los datos originales y.

2.1.2. Transformación $z_i = (y_i)^\lambda$

Sea $z_1, z_2, \ldots, z_n$ una muestra y la transformación aplicada $z_i = (y_i)^\lambda \rightarrow g^{-1}(y_i) = {z_i}^{1/\lambda}$, calculamos su verosimilitud y la log-verosimilitud obteniendo lo siguiente:

$$f(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} f_y({z_i}^{1/\lambda}) \frac{1}{\lambda} {z_i}^{1/\lambda - 1} \quad (2.11)$$

$$l(z_1, z_2, \ldots, z_n) = \sum_{i=1}^{n} \left\{ log\left(\frac{1}{\lambda} {z_i}^{1/\lambda - 1} \right) + log(f_y({z_i}^{1/\lambda})) \right\} \quad (2.12)$$

Por ejemplo, si $y \sim N(\mu, \sigma^2)$, entonces su función de densidad está dada por:

$$f_y(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \tag{2.13}$$

Y la función de distribución de $g^{-1}(y_i)$ está dada por:

$$f_y({z_i}^{1/\lambda}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}({z_i}^{1/\lambda}-\mu)^2\right\} \tag{2.14}$$

la log-verosimilitud queda de la siguiente forma:

$$log(f_y({z_i}^{1/\lambda})) = -\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}({z_i}^{1/\lambda}-\mu)^2 \tag{2.15}$$

Reemplazando $log(f_y({z_i}^{1/\lambda}))$ en la ecuación (2.12), obtenemos la log-verosimilitud para la transformación $z_i = (y_i)^\lambda$.

$$\begin{aligned} l(z_1, z_2, \ldots, z_n) &= \sum_{i=1}^{n}\left\{log\left(\frac{1}{\lambda}{z_i}^{1/\lambda-1}\right) - \frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}({z_i}^{1/\lambda}-\mu)^2\right\} \\ &= -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i^{1/\lambda}-\mu)^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda}{z_i}^{1/\lambda-1}\right) \end{aligned} \tag{2.16}$$

Donde

$$\begin{aligned} z_i^{1/\lambda} z_i^{-1} &= (y_i)^{1/\lambda}(y_i^\lambda)^{-1} \\ &= y_i * y_i^{-\lambda} \\ &= y^{1-\lambda} \end{aligned} \tag{2.17}$$

Reemplazando $y_i = z_i^{1/\lambda}$ y $z_i^{1/\lambda-1} = y^{1-\lambda}$ en la ecuación (2.16) se tiene:

$$l(z_1, z_2, \ldots, z_n) = -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i^{1/\lambda}-\mu)^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda}{z_i}^{1/\lambda-1}\right) \tag{2.18}$$

para los datos originales se tiene la log-verosimilitud.

$$\begin{aligned} l(y_1, y_2, \ldots, y_n) &= -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda}{y_i}^{1-\lambda}\right) \\ l(y_1, y_2, \ldots, y_n) &= -\frac{n}{2}(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - nlog(\lambda) + n(1-\lambda)\overline{log(y)} \end{aligned} \tag{2.19}$$

.

Se tiene que los dos primeros términos de la ecuación (2.19) es la log-verosimilitud de los datos originales (y) y los términos $nlog(\lambda) + n(1-\lambda)\overline{log(y)}$ son la penalización por la trasformación realizada a la variable respuesta y $\overline{log(y)}$ es el promedio de los logaritmos de la variable y, en está ultima ecuación se puede observar esta en términos de los datos originales y.

2.2. Modelo lineal

Ahora se inicia con el modelo de regresión lineal simple, que consiste en expresar la dependencia lineal de la variable objetivo o dependiente, y, respecto a otras dos variables: la variable independiente, explicativa o covariable, x, y el término error o perturbación del modelo, e así:

$$y_i = \beta_o + \beta_1 x_i + e_i, \text{ con } (x_i, y_i) \text{ variables numéricas.}$$

donde y_i y e_i son variables aleatorias, x_i es una variable conocida y β_0 y β_1 son los parámetros desconocidos del modelo.

Las hipótesis del modelo se pueden formular en términos de la variable e_i, o de forma equivalente en términos de la variable dependiente, y. La variable e_i sigue una distribución normal, por tanto la distribución de y para cada x también sigue una distribución normal, expresadas de la siguiente forma:

- $e_i \sim N(0, \sigma^2)$
- $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- Los errores, e_i, son independientes entre sí y las variables y_i son independientes entre sí.

2.2.1. Transformación $z_i = log(y_i)$

Sea $z_i = log(y_i)$ la transformación, se tiene $y_i = e^{z_i}$.

Se encuentra la función de densidad conjunta de $z_1, z_2, \ldots, z_n$ y aplicando el teorema de la transformación se obtiene:

$$\begin{aligned} f(z_1, z_2, \ldots, z_n) &= \prod_{i=1}^{n} f_z(z_i) \\ &= \prod_{i=1}^{n} f_y(e^{z_i}) e^{z_i} \end{aligned} \tag{2.20}$$

Como se tiene que $y_i \sim N(\beta_o + \beta_1 x_i, \sigma^2)$, entonces la función de densidad esta dada por:

$$f(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ \frac{1}{2\sigma^2}(e^{z_i} - \beta_0 - \beta_1 x_i)^2 \right\} (e^{z_i}) \right\} \tag{2.21}$$

$$\begin{aligned} l(\beta_0, \beta_1) &= \sum_{i=1}^{n} \left\{ -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(e^{z_i} - \beta_0 - \beta_1 x_i)^2 + log\left(e^{z_i}\right) \right\} \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(e^{z_i} - \beta_0 - \beta_1 x_i\right)^2 + \sum_{i=1}^{n} z_i \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 + n\overline{log(y)} \\ &. \end{aligned} \tag{2.22}$$

En la anterior expresión se obtiene, que

$$-\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

es la log-verosimilitud del modelo con los datos originales y $n\overline{log(y)}$ es la penalización por la transformación aplicada a la variable respuesta del modelo.

2.2.2. Transformación $z_i = (y_i)^\lambda$

Sea $z_i = (y_i)^\lambda$ la transformación, con λ conocido y positivo se tiene que $y_i = z_i^{1/\lambda}$.

Se encuentra la función de densidad conjunta de $z_1, z_2, \ldots, z_n$ y aplicando el teorema de la transformación se obtiene:

$$\begin{aligned} f(z_1, z_2, \ldots, z_n) &= \prod_{i=1}^{n} f_z(z_i) \\ &= \prod_{i=1}^{n} f_{y_i}(z_i^{1/\lambda}) \frac{1}{\lambda} z_i^{1/\lambda - 1} \end{aligned} \tag{2.23}$$

Como se tiene que $y_i \sim N(\beta_o + \beta_1 x_i, \sigma^2)$, entonces la función de densidad del modelo esta dada por:

$$f(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ \frac{1}{2\sigma^2}(z_i^{1/\lambda} - \beta_0 - \beta_1 x_i)^2 \right\} \left(\frac{1}{\lambda} z_i^{1/\lambda - 1} \right) \right\}$$

$$\begin{aligned} l(\beta_0, \beta_1) &= \sum_{i=1}^{n} \left\{ -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z_i^{1/\lambda} - \beta_0 - \beta_1 x_i)^2 + log\left(\frac{1}{\lambda} z_i^{1/\lambda - 1} \right) \right\} \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(z_i^{1/\lambda} - \beta_0 - \beta_1 x_i \right)^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda} z_i^{1/\lambda - 1} \right) \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda} y_i^{1-\lambda} \right) \end{aligned} \tag{2.24}$$

En la anterior expresión se obtiene el resultado donde:

$$-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

es la log-verosimilitud del modelo y $\sum_{i=1}^{n} log\left(\frac{1}{\lambda}y_i^{1-\lambda}\right)$ es la penalización por la transformación aplicada. En este ultimo término debemos analizar su comportamiento, teniendo en cuenta que para que el $log\left(\frac{1}{\lambda}y_i^{1-\lambda}\right)$ exista $\frac{1}{\lambda}y_i^{1-\lambda} > 0$.

2.3. Regresión Múltiple

El modelo de regresión lineal múltiple es la extensión del modelo de regresión lineal simple cuando consideramos k variables explicativas. En general, la variable objetivo Y depende de muchas otras variables $x_1, ..., x_k$, aunque algunas de éstas pueden no ser observables o desconocidas. El modelo de regresión incluye las que más efecto tienen y las restantes las representa como una variable aleatoria que se denomina error del modelo, por tanto, tenemos:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e \tag{2.25}$$

para cada observación, sería:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i, \ con\ i = 1, ..., n \tag{2.26}$$

Del modelo se tienen las siguientes propiedades:

- $E[e] = 0$ y $E[y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $Var(e) = \sigma^2$ y $Var(y) = \sigma^2$
- Los errores, e_i, son independientes entre sí y las variables y_i son independientes entre sí.
- $e \sim N(0, \sigma^2)$ y $y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$.

2.3.1. Transformación $z_i = log(y_i)$

Sea $z_i = log(y_i)$ la transformación, se tiene $y_i = e^{z_i}$.

Se encuentra la función de densidad conjunta de $z_1, z_2, \ldots, z_n$ y aplicando el teorema de la transformación se obtiene:

$$\begin{aligned} f(z_1, z_2, \ldots, z_n) &= \prod_{i=1}^{n} f_z(z_i) \\ &= \prod_{i=1}^{n} f_y(e^{z_i}) e^{z_i} \end{aligned} \tag{2.27}$$

Como se tiene que $y_i \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$, entonces la función de densidad esta dada por:

$$f(z_1, \ldots, z_n) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ \frac{1}{2\sigma^2} (e^{z_i} - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_k)^2 \right\} (e^{z_i}) \right\}$$

$$\begin{aligned} l(\beta_0, \beta_1 x_1, \cdots, \beta_k x_k) &= \sum_{i=1}^{n} \left\{ -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (e^{z_i} - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_{ki})^2 + log\left(e^{z_i}\right) \right\} \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (e^{z_i} - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_{ki})^2 + \sum_{i=1}^{n} z_i \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_{ki})^2 + n\overline{log(y)} \\ &. \end{aligned} \tag{2.28}$$

En la anterior expresión se obtiene, que

$$-\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \cdots + \beta_k x_{ki})^2$$

es la log-verosimilitud del modelo con los datos originales y $n\overline{log(y)}$ es la penalización por la transformación aplicada a la variable respuesta.

2.3.2. Transformación $z_i = (y_i)^\lambda$

Sea $z_i = (y_i)^\lambda$ la transformación, con λ conocido se tiene que $y_i = z_i^{1/\lambda}$.

Se encuentra la función de densidad conjunta de $z_1, z_2, \ldots, z_n$ y aplicando el teorema de la transformación se obtiene:

$$\begin{aligned} f(z_1, z_2, \ldots, z_n) &= \prod_{i=1}^{n} f_z(z_i) \\ &= \prod_{i=1}^{n} f_{y_i}(z_i^{1/\lambda}) \frac{1}{\lambda} z_i^{1/\lambda - 1} \end{aligned} \tag{2.29}$$

Como se tiene que $y_i \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$, entonces la función de densidad esta dada por:

$$f(z_1, z_2, ..., z_n) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ \frac{1}{2\sigma^2} (z_i^{1/\lambda} - \beta_0 - \beta_1 x_1 - ... - \beta_k x_k)^2 \right\} \left(\frac{1}{\lambda} z_i^{1/\lambda - 1} \right) \right\}$$

$$\begin{aligned} l(\beta_0, \beta_1 x_1, ..., \beta_k x_k) &= \sum_{i=1}^{n} \left\{ -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (z_i^{1/\lambda} - \beta_0 - \beta_1 x_1 - ... - \beta_k x_{ki})^2 + log\left(\frac{1}{\lambda} z_i^{1/\lambda - 1} \right) \right\} \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(z_i^{1/\lambda} - \beta_0 - \beta_1 x_1 - ... - \beta_k x_{ki} \right)^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda} z_i^{1/\lambda - 1} \right) \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - ... - \beta_k x_{ki})^2 + \sum_{i=1}^{n} log\left(\frac{1}{\lambda} y_i^{1-\lambda} \right) \\ &. \end{aligned} \tag{2.30}$$

En la anterior expresión, se obtiene el resultado donde:

$$-\tfrac{n}{2} log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_k x_{ki})^2$$

es la log-verosimilitud del modelo y $\sum_{i=1}^{n} log\left(\frac{1}{\lambda} y_i^{1-\lambda} \right)$ es la penalización por la transformación aplicada.

2.4. Expresiones para el AIC_J y BIC_J

Con las expresiones encontradas anteriormente para las transformaciones mencionadas para el modelo lineal y la regresión múltiple, a partir de la las expresiones de los criterios de selección AIC y BIC, haciendo el cambio de la log-verosimilitud encontrada con la trasformación se definen las expresiones para AIC_J y BIC_J. Se parte de las expresiones de AIC y BIC siguientes:

$$AIC = -2l(\theta) + 2k \tag{2.31}$$
$$BIC = -2l(\theta) + klog(n) \tag{2.32}$$

Donde $l(\theta)$ es la log-verosimilitud, y está es la que se modifica teniendo en cuenta la transformación realizada y quedan las siguientes expresiones.

2.4.1. AIC_J y BIC_J para un modelo lineal

Para el modelo lineal se tiene la log-verosimilitud para cada una de las transformaciones mencionadas, las cuales se muestran a continuación:

Para la transformación $z_i = log(y_i)$

$$l(\beta_0, \beta_1) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + n\overline{log(y)} \tag{2.33}$$

Para la transformación $z_i = y_i^\lambda$ con λ conocido y positivo.

$$l(\beta_0, \beta_1) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - nlog(\lambda) + n(1-\lambda)\overline{log(y)} \tag{2.34}$$

Expresiones para los criterios AIC_J y BIC_J con la transformación $z_i = log(y_i)$

$$\begin{aligned} AIC_J &= -2\left(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + n\overline{log(y_i)}\right) + 2k \\ &= -2\left(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) - 2n\overline{log(y)} + 2k \\ &= AIC - 2n\overline{log(y)} \end{aligned} \tag{2.35}$$

$$\begin{aligned} BIC_J &= -2\left(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) - 2n\overline{log(y_i)} + klog(n) \\ &= BIC - 2n\overline{log(y)} \end{aligned} \tag{2.36}$$

Expresiones para los criterios AIC_J y BIC_J con la transformación $z_i = y_i^{\lambda}$

$$\begin{aligned} AIC_J &= -2\left(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - nlog(\lambda) + n(1-\lambda)\overline{log(y)}\right) + 2k \\ &= nlog(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y)} + 2k \\ &= AIC + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y)} \end{aligned} \tag{2.37}$$

$$\begin{aligned} BIC_J &= -2\left(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - nlog(\lambda) + n(1-\lambda)\overline{log(y_i)}\right) + klog(n) \\ &= nlog(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y)} + klog(n) \\ &= BIC + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y)} \end{aligned} \tag{2.38}$$

2.4.2. AIC_J y BIC_J para una regresión múltiple

En la regresión múltiple se tienen la log-verosimilitud para cada una de las transformaciones.

$$l(\beta_0, \beta_1 x_1, ..., \beta_k x_k) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 + n\overline{log(y)} \tag{2.39}$$

$$\begin{aligned} l(\beta_0, \beta_1 x_1, ..., \beta_k x_k) &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &- n log(\lambda) + n(1-\lambda)\overline{log(y)} \end{aligned} \tag{2.40}$$

Expresiones para los criterios AIC_J y BIC_J con la transformación $z_i = log(y_i)$

$$\begin{aligned} AIC_J &= -2(-\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &+ n\overline{log(y)} + 2k \\ &= n log(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &- 2n\overline{log(y)} + 2k \\ &= AIC - 2n\overline{log(y)} \end{aligned} \tag{2.41}$$

$$\begin{aligned} BIC_J &= n log(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &- 2n\overline{log(y_i)} + k log(n) \\ &= BIC - 2n\overline{log(y)} \end{aligned} \tag{2.42}$$

Expresiones para los criterios AIC_J y BIC_J con la transformación $z_i = y_i^\lambda$

$$\begin{aligned} AIC_J &= -2(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &- nlog(\lambda) + n(1-\lambda)\overline{log(y)} + 2k \\ &= nlog(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &+ 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y_i)} + 2k \\ &= AIC + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y)} \end{aligned} \tag{2.43}$$

$$\begin{aligned} BIC_J &= -2(-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &- nlog(y_i) + n(1-\lambda)\overline{log(y)}) + klog(n) \\ &= nlog(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - ... - \beta_k x_{ki})^2 \\ &+ 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y_i)}) + klog(n) \\ &= BIC + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y)} \end{aligned} \tag{2.44}$$

Capítulo 3

Funciones en R para el AIC_J y BIC_J

Antes de mostrar las funciones que nos permiten realizar el cálculo de los criterios AIC_J y BIC_J, para cada una de las transformaciones expuestas en el trabajo, se realiza una comparación entre los diferentes criterios y analizar cuál de ellos es mejor y bajo que condiciones.

3.1. Comparación de criterios

En cuanto a los criterios AIC y BIC, bajo ciertas condiciones, son similares. Suponiendo normalidad, tenemos que el BIC está definido por:

$$BIC = n\ l(\hat{\theta}) + k\ log(n)$$

Por lo tanto, podemos dividir ambos términos por n y considerar:

$$log(\hat{\theta}) + k\ \frac{log(n)}{n}$$

Supongamos ahora que tenemos dos modelos con p_1 y p_2 variables, con $k \leq p_1 < p_2$, entonces:

$$BIC(p_1) = l(\hat{\theta}_{p_1}) + k_1 \ \frac{log(n)}{n}$$
$$BIC(p_2) = l(\hat{\theta}_{p_2}) + k_2 \ \frac{log(n)}{n}$$

Definamos ahora $\nabla BIC = BIC(p_2) - BIC(p_2)$, de donde obtenemos los siguiente:

$$\nabla BIC = l(\hat{\theta}_{p_2}) - l(\hat{\theta}_{p_1}) + (k_2 - k_1)\frac{log(n)}{n}$$

Con lo que elegiremos el modelo con p_1 variables cuando $\nabla BIC \geq 0$. De manera análoga se puede proceder con el AIC, obteniendo:

$$\nabla AIC = l(\hat{\theta}_{p_2}) - l(\hat{\theta}_{p_1}) + (k_2 - k_1)\frac{2}{n}$$

Luego, cuando $log(n) = 2$ ambos criterios serán similares, esto es cuando $n = 8$. Cuando $n > 8$ el BIC penaliza más la inclusión de de variables irrelevantes que el AIC, por lo tanto es mejor.

Ahora cuando tenemos los criterios AIC_J y BIC_J, al realizar la comparación con los criterios clásicos se obtiene que $AIC_J < AIC$ y $BIC_J < BIC$, cuando se aplica la transformación $z = log(y)$ dado que en esta se le resta el siguiente valor $-2n\overline{log(y_i)}$ a los criterios clásicos para obtener los criterios AIC_J y BIC_J, donde este termino es negativo cuando los valores de la variable $y > 1$ y va ser positivo o se va incrementar cuando la variable respuesta es $0 < y < 1$ dado que en este intervalo el logaritmo toma valores negativos, por lo que se considera como una penalización por la transformación.

Por otra parte al comparar los criterios con la transformación $z = y^{\lambda}$, se tiene que van a obtener valores mayores, es decir, una penalización dada por la transformación de la variable respuesta y su valor va depender de los valores dados a λ y los valores que tome

la variable respuesta y, en algunos casos e va presentar que va ser menor y en otros que va ser mayor, entonces también nos permite identificar si la transformación aplicada es la mejor para la variable respuesta o se puede encontrar una que sea mejor para los datos.

3.2. Funciones en R para el AIC_J y BIC_J

En el capítulo anterior se realizó la construcción de las expresiones de los criterios AIC_J y BIC_J en donde se aplicaron las transformaciones $z = log(y)$ y $z = y^{\lambda}$ a las variables respuesta, de donde se obtienen las siguientes expresiones que son el punto de partida para construir la función que nos permite realizar el cálculo del valor de cada uno de los criterios.

Para la construcción de las funciones en R se parte de las siguientes expresiones, tanto para el criterio AIC_J como para el criterio BIC_J.

Para la transformación $z_i = log(y_i)$ se tiene:

$$AIC_J = AIC - 2n\overline{log(y_i)} \tag{3.1}$$

$$BIC_J = BIC - 2n\overline{log(y_i)} \tag{3.2}$$

Para la transformación $z_i = (y_i)^{\lambda}$ se tiene:

$$AIC_J = AIC + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y_i)} \tag{3.3}$$

$$BIC_J = BIC + 2nlog(\lambda) - 2n(1-\lambda)\overline{log(y_i)} \tag{3.4}$$

Observando las expresiones obtenidas para el calculo de cada uno de los criterios y las transformaciones $z = log(y)$ y $z = y^{\lambda}$, la construcción de estas es muy sencilla, dado que la primera parte de la expresión son los criterios AIC y BIC clásicos respectivamente y solo se agrega la parte obtenida para la transformación realizada, que dependiendo de los valores de la variable y puede ser una penalización o una reducción por la transformación.

Las funciones quedan de la siguiente manera, teniendo en cuenta que se realiza una para cada uno de los criterios AIC_J y BIC_J sin importar la transformación realizada a la variable respuesta, se tiene para el criterio AIC_J la siguiente función en R, la cuál tiene los siguiente parámetros:

- modelo: es el modelo que se quiere evaluar.
- lambda: es el valor para la transformación $z = y^{\lambda}$, este valor debe ser mayor de cero.
- AIC es el criterio clásico que se encuentra en la librería *stats*.

Al ingresar el modelo y el valor de lambda, la función realiza el cálculo del criterio AIC_J, teniendo en cuenta que si el valor de lambda en negativo va salir un mensaje de error solicitando que este valor debe ser mayor que cero, si el valor de lambda es igual a cero la función calcula el valor del AIC_J con la transformación $z = log(y)$.

```
#####Funcion AICJ
AICJ<-function(modelo,lambda) {
  if (lambda<0){"Debe ingresar un valor para lambda
  mayor que cero"}
  else{
  y=modelo$fitted.values+modelo$residuals
  n=length(y)
```

```
  AIC=AIC(modelo)
  if (lambda != 0) {
    AICJ=AIC+2*n*log(lambda)-2*(1-lambda)*n*mean(log(y))
     }else {
    AICJ=AIC-2n*mean(log(y))
    }
  return(AICJ)
  }
}
```

Para el criterio BIC_J la función en R queda de la siguiente forma y tiene los siguiente parámetros:

- modelo: es el modelo que se quiere evaluar.
- lambda: es el valor para la transformación $z = y^{\lambda}$, este valor debe ser mayor de cero.
- BIC es el criterio clásico que se encuentra en la librería *stats*.

Al ingresar el modelo y el valor de lambda, la función realiza el cálculo del criterio BIC_J, teniendo en cuenta que si el valor de lambda en negativo va salir un mensaje de error solicitando que este valor debe ser mayor que cero, si el valor de lambda es igual a cero la función calcula el valor del BIC_J para la transformación $z = log(y)$.

```
#####Funcion BICJ
BICJ<-function(modelo,lambda) {
  if (lambda<0){"Debe ingresar un valor para lambda
  mayor que cero"}
  else{
  y=modelo$fitted.values+modelo$residuals
  n=length(y)
```

```
  BIC=BIC(modelo)
  if (lambda != 0) {
    BICJ=BIC+2*n*log(lambda)-2*(1-lambda)*n*mean(log(y))
     }else {
    BICJ=BIC-2
    n*mean(log(y))
    }
  return(BICJ)
  }
}
```

Capítulo 4

Ejemplo del AIC_J y BIC_J

En este capítulo se presenta un ejemplo donde se aplica una transformación a la variable respuesta, se hace el calculo de cada uno de los criterios y se comparan con los criterios clásicos AIC y BIC, primero se hace una descripción de los datos que se van a utilizar, un análisis exploratorio, se plantean diferentes modelos para los datos y finalmente se calculan los criterios para realizar la selección del mejor modelo para los datos.

4.1. Ejemplo de los criterios AIC_J y BIC_J

A continuación se presenta un ejemplo en R, donde se comparan los resultados obtenidos de los criterios AIC, BIC, AIC_J y BIC_J, para presentar este ejemplo se utiliza el conjunto de datos que nos proporciona Fahrmeir, L. y que contiene las siguientes variables:

- Age: edad (en años).
- Kilometer: kilómetros (en miles).
- TIA: número de meses hasta la próxima Inspección
- Técnica de Vehículos o ITV.

- Extras1: Si tiene ABS o no.
- Extras2: Si tiene techo solar o no.

En donde las 4 primeras variables son cuantitativas y las 2 últimas son factores, por lo que habrá que tenerlo en cuenta a la hora de analizar los resultados, tomando como variable Y o respuesta a la variable *price*.

Se cargan los datos y luego se realiza un estudio gráfico para observar la relación que hay entre la variable objetivo y las variables explicativas, para ello realizamos gráficos de nube de puntos de la variable Y frente a cada una de ellas.

En la figura (4.1) podemos ver que tanto *age* como *kilometer* tienen una relación lineal inversa con la variable objetivo. Y que la variable explicativa TIA, no tiene una relación lineal con Y.

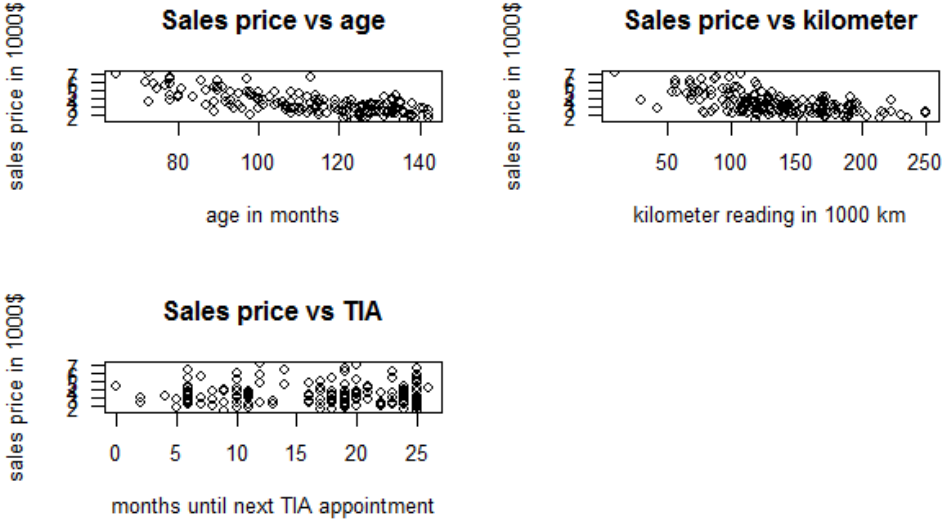

Figura 4.1: Y vs variables cuantitativas

Para tratar con los 2 factores haremos un diagrama de boxplot y asi poder comparar las medianas de llevar o no llevar $extras1$ y $extras2$ para visualizar si puede haber diferencias significativas entre los grupos.

En la figura (4.2) podemos ver que los coches con ABS son un poco más caros que los que no tienen, pero sin una diferencia significativa; podemos apreciar también que los coches con y sin techo solar tienen medianas similares, por lo que tampoco hay una diferencia significativa.

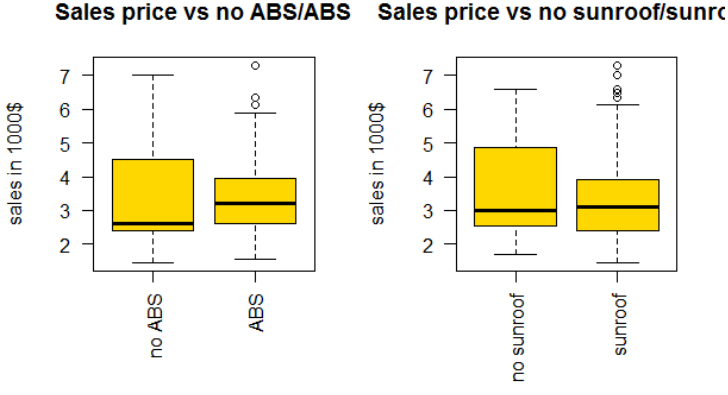

Figura 4.2: Y vs variables cualitativas

Luego de tener este análisis de las variables explicativas y la variable respuesta, se decide que las variables *age* y *kilometer* deben estar en nuestro modelo de regresión lineal. Ahora haremos un estudio de los distintos modelos combinando las restantes variables explicativas, obteniendo los siguientes modelos candidatos.

1. mod1: price = kilometer + age.
2. mod2: price = kilometer + age + extras1.
3. mod3: price = kilometer + age + extras2.
4. mod4: price = kilometer + age + TIA.
5. mod5: price = kilometer + age + extras1 + extras2.
6. mod6: price = kilometer + age + extras1 + TIA.
7. mod7: price = kilometer + age + extras2 + TIA.
8. mod8: price = kilometer + age + extras1 + extras2 + TIA.

A continuación en el cuadro (4.1) se presentan los resultados obtenidos para el AIC, el BIC, el AIC_J y el BIC_J para los modelos seleccionados con la transformacion $z = log(y)$ y poder seleccionar el modelo que menor valor de los criterios tenga.

Cuadro 4.1: Selección de modelos

MODELOS	**AIC**	**BIC**	**AICJ**	**BICJ**
mod1	395,234	420,414	-4,1257442	21,05421
mod2	396,5663	424,8938	-2,7934301	25,53402
mod3	397,119	425,4465	-2,2407419	26,08671
mod4	396,9732	425,3006	-2,3865704	25,94088
mod5	398,4814	429,9564	-0,8782941	30,59665
mod6	398,1434	429,6183	-1,2163702	30,25857
mod7	398,88	430,3549	-0,4797381	30,99521
mod8	400,0852	434,7077	0,7255004	35.34794

En el cuadro (4.2) se presentan los resultados obtenidos para el AIC, el BIC, el AIC_J y el BIC_J para los modelos seleccionados

con la transformacion $z = y^{1/2}$ y poder seleccionar el modelo que menor valor de los criterios tenga.

Cuadro 4.2: Selección de modelos

MODELOS	**AIC**	**BIC**	**AICJ**	**BICJ**
mod1	395,234	420,414	-42,88850	-17,708548
mod2	396,5663	424,8938	-41,55619	-13,228739
mod3	397,119	425,4465	-41,00350	-12,676051
mod4	396,9732	425,3006	-41,14933	-12,821879
mod5	398,4814	429,9564	-39,64105	-8,166109
mod6	398,1434	429,6183	-39,97913	-8,504185
mod7	398,88	430,3549	-39,24250	-7,767553
mod8	400,0852	434,7077	-38,03726	-3,41482

Para los dos ejemplos presentados anteriormente, los resultados obtenidos se muestran en las tablas, se puede observar que los criterios propuestos en el el trabajo realizan una disminución considerable en los valores con respecto a los valores obtenidos con los criterios clásicos, sin embargo la selección del modelo en este caso no varia y se sigue seleccionando el mismo, teniendo en cuenta que este resultado puede variar dado los posibles valores de la variable respuesta y de el valor de λ seleccionado para la transformación.

Capítulo 5

Conclusiones

- Utilizar el criterio de AIC y BIC clásico cuando se tiene la variable respuesta transformada puede llevar a una mala identificación del modelo.
- Este trabajo muestra una construcción detallada de los criterios AIC_J y BIC_J cuando la variable respuesta tiene una transformación, para modelos lineales y regresión múltiple.
- En las expresiones encontradas para los criterios AIC_J y BIC_J, se puede observar que se pueden obtener a partir de los criterios AIC y BIC agregando la penalización dada por cada una de las trasformaciones utilizadas.

Capítulo 6

Bibliografía

1. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.

2. Beaumont, Adrian N., 2014. "Data transforms with exponential smoothing methods of forecasting,Ïnternational Journal of Forecasting, Elsevier, vol. 30, pages 918-927.

3. Burnham, K. P. and Anderson, D. R. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer, New York.

4. Cholaquidis, A. 2015. Notas para el curso de Probabilidad II. Licenciatura en Estadística.Facultad de Ciencias, Universidad de la República.

5. Claeskens, G. and Hjort, G., L. 2008. Model Selection and Model Averaging, Cambridge: Cambridge University Press.

6. Fahrmeir. L, Kneib. T, Lang. S and Marxregression.B, 2013. Regression: Models, Methods and Applications.Springer

7. De la Horra, J. Introducción a la convergencia de sucesiones de variables aleatorias. Departamento de Matemáticas U.A.M.

8. Inga Santibáñez, Rosa M. (1996) Principio de maximización de la información de Akaike. Venezuela. Universidad Nacional Mayor San Marcos.

9. Konishi, S. and Kitagawa, G. (2008). Information Criteria and Statistical Modeling, Springer Series in Statistics. Springer Verlag, New York.

10. Kullback, S. (1997). Information Theory and Statistics, Dover Publications, New York.

11. Márquez Cebrián, Maria Dolors. 2002. Modelo setar aplicado a la volatilidad de la rentabilidad de las acciones: algoritmos para su identificación. Universidad politécnica de Cataluña.

12. Medel, Carlos A. (2012) Akaike or Schwarz? Which One is a Better Predictor of Chilean GDP? Central Bank of Chile.

13. Montesinos. L,Abelardo. 2011. Estudio del AIC y BIC en la selección de modelos de vida con datos censurados. Guanajuato, Gto, CIMAT.

14. Montgomery, D.; Peck E. y Vining, G. 2006. Introducción al análisis de regresión lineal. México: Limusa Wiley.

15. P. Billingsley. 1999. Convergence of Probability Measures. 2nd ed. John Wiley and Sons.

16. Raymond, Jose luis y Uriel,Esequiel., 1987. investigación econométrica aplicada un caso de estudio. alfa centauro S.A.

17. Salih N. NeftçI (2012) Specification of Economic Time Series Models Using Akaike's Criterion, Journal of the American Statistical Association, 77:379, 537-540.

18. Shibata, R. 1995. Boostrap Estimate of Kullback-Leibler Information for Model Selection, Technical Report No. 424, Department of Statistical University of California Berkeley, California.

19. Tong, H. 1990. Non-linear time series: a dynamical system approach. Oxford: Oxford University Press.

20. Wasserman, L.A. 2004. All of Statistics: A Concise Course in Statistical Inference, Springer, New York.

21. Cholaquidis, A. 2015. Notas para el curso de Probabilidad II. Licenciatura en Estadística.Facultad de Ciencias, Universidad de la República

Apéndice A

Algunos conceptos aplicados

Ley débil de los grandes números Esta ley nos permite encontrar la convergencia en probabilidad y se define de la siguiente manera:

Sea $\{X_n\}_n$ una sucesión de variables aleatorias con esperanza finita. Supongamos que:

$$\frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) \to_n 0 \tag{A.1}$$

entonces

$$\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - E(x_i))\right] \to^p 0 \tag{A.2}$$

en particular si las x_i son independientes dos a dos y existe C tal que $Var(X_i) < C$ para todo i, se cumple (A.1).

Demostración: Denotemos $Z_n = \frac{1}{n}\sum_{i=1^n} x_i$, (A.1) es equivalente a probar que para todo $\epsilon > 0$, $P(|Z_n - E(Z_n)| > \epsilon) \to_n 0$, pero

esto se sigue de la desigualdad de Markov ya que:

$$P|Z_n - E(Z_n)| > \epsilon \leq \frac{1}{n^2\epsilon^2}\left(Var\sum_{i=1}^{n} x_i\right) \rightarrow_n 0. \tag{A.3}$$

Teorema de la transformación

Sea $X = (x_1, x_2, \ldots, x_n)$ un vector aleatorio con densidad conjunta f_x. Sea $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ una aplicación inyectiva. Se supone que tanto g como su inversa $h : A \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ son continuas. Si las derivadas parciales de h existen y son continuas y si su jacobiano, J, es diferente de cero, entonces, el vector aleatorio $Y := g(x)$ tiene función de densidad conjunta f_y dada por:

$$f_y(y) = \begin{cases} f_x(h(y))|J(y)| & \text{si } y \text{ está en el rango de } g \\ 0 & \text{en otro caso} \end{cases} \tag{A.4}$$

Apéndice B

Codigos de R utilizados

```
# cambiar el directorio de trabajo
# change working directory
setwd("C:/Users/leonel_2/Downloads")
library(foreign)

# read data
golf <- read.table(file="base.txt",header = TRUE,
sep = "",dec = ".")
attach(golf)

###grafica 1
par(mfrow=c(2,2), las=1)
plot(age,price,ylab="sales price in 1000$",
xlab="age in months",+ main="Sales price vs age")
plot(kilometer,price,ylab="sales price in 1000$",
+ xlab="kilometer reading in 1000 km",
main="Sales price vs kilometer")
plot(TIA,price,ylab="sales price in 1000$",
+ xlab="months until next TIA appointment",
main="Sales price vs TIA")
```

```
###grafica 2
par(mfrow=c(1,2), las=2)
boxplot(price ~ extras1, main="Sales price vs no ABS/ABS",
+ ylab="sales in 1000$", col="gold", names=c("no ABS", "ABS"))
boxplot(price ~ extras2, main="Sales price vs
no sunroof/sunroof",+ ylab="sales in 1000$",
col="gold", names=c("no sunroof", "sunroof"))

# 8 regression models
mod1 <- lm(price~kilometerop1+kilometerop2+kilometerop3+
ageop1+ageop2+ageop3, data=golf)
mod2 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+extras1, data=golf)
mod3 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+extras2, data=golf)
mod4 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+TIA, data=golf)
mod5 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+extras1+extras2, data=golf)
mod6 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+extras1+TIA, data=golf)
mod7 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+extras2+TIA, data=golf)
mod8 <- lm(price~kilometerop1+kilometerop2+kilometerop3+ageop1+
ageop2+ageop3+extras1+extras2+TIA, data=golf)

#####AIC
nf <- 8
nc <- 1
resAIC <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
rownames(resAIC)<-c("mod1","mod2","mod3", "mod4", "mod5",
"mod6","mod7", "mod8")
colnames(resAIC)<-c("AIC")
resAIC[,1] <- c(AIC(mod1),AIC(mod2),AIC(mod3),AIC(mod4),
```

```
AIC(mod5),AIC(mod6),AIC(mod7),AIC(mod8))
resAIC

## BIC
nf <- 8
nc <- 1
resBIC <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
rownames(resBIC)<-c("mod1","mod2","mod3", "mod4",
"mod5", "mod6","mod7", "mod8")
colnames(resBIC)<-c("BIC")
resBIC[,1] <- c(BIC(mod1),BIC(mod2),BIC(mod3),BIC(mod4),
BIC(mod5),BIC(mod6),BIC(mod7),BIC(mod8))
resBIC

#####Funcion AICJ
AICJ<-function(modelo,lambda) {
  if (lambda<0)
  {"Debe ingresar un valor para lambda mayor que cero"}
  else{
  y=modelo$fitted.values+modelo$residuals
  n=length(y)
  AIC=AIC(modelo)
  if (lambda != 0) {
    AICJ=AIC+2*n*log(lambda)-2*(1-lambda)*n*mean(log(y))
    # Si la condición se cumple (TRUE)
  }else {
    AICJ=AIC-2*n*mean(log(y))
    }
  return(AICJ)
  }
}

#####Funcion BICJ
BICJ<-function(modelo,lambda) {
```

```
  if (lambda<0)
  {"Debe ingresar un valor para lambda mayor que cero"}
  else{
  y=modelo$fitted.values+modelo$residuals
  n=length(y)
  BIC=BIC(modelo)
  if (lambda != 0) {
    BICJ=BIC+2*n*log(lambda)-2*(1-lambda)*n*mean(log(y))
    # Si la condición se cumple (TRUE)
  }else {
    BICJ=BIC-2*n*mean(log(y))
  }
  return(BICJ)
  }
}


## AICJ
nf <- 8
nc <- 1
resAICJ <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
rownames(resAICJ)<-c("mod1","mod2","mod3", "mod4",
"mod5", "mod6","mod7", "mod8")
colnames(resAICJ)<-c("AICJ")
resAICJ[,1] <- c(AICJ(mod1,0),AICJ(mod2,0),AICJ(mod3,0),
AICJ(mod4,0),AICJ(mod5,0),AICJ(mod6,0),AICJ(mod7,0),
AICJ(mod8,0))
resAICJ


## BICJ
nf <- 8
nc <- 1
resBICJ <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
```

```
rownames(resBICJ)<-c("mod1","mod2","mod3", "mod4",
"mod5", "mod6","mod7", "mod8")
colnames(resBICJ)<-c("BICJ")
resBICJ[,1] <- c(BICJ(mod1,1/2),BICJ(mod2,1/2),BICJ(mod3,1/2),
BICJ(mod4,1/2),BICJ(mod5,1/2),BICJ(mod6,1/2),
BICJ(mod7,1/2),BICJ(mod8,1/2))
resBICJ

###todos los criterios
cbind(resAIC,resBIC,resAICJ,resBICJ)
```

Apéndice C

Habeas Data

Bogotá D.C. 07 de Junio de 2018.

Los datos utilizados para el ejemplo de este trabajo fueron propuestos en el libro: Fahrmeir, L.; Kneib, Th.; Lang, S.; Marx, B. Regression: Models, Methods and Applications. New York: Springer. 2013. y se pueden descargar de la pagina http://www.uni-goettingen.de/de/551625.html.

Prices of Used Cars (see Example, 3.19): ASCII

Dicho conjunto posee 5 variables explicativas:

- Age: edad (en años).
- Kilometer: kilómetros (en miles).
- TIA: número de meses hasta la próxima Inspección Técnica de Vehículos o ITV.
- Extras 1: Si tiene ABS o no.
- Extras 2: Si tiene techo solar o no.

Las 3 primeras variables son cuantitativas y las 2 últimas son factores.

Los datos son de consulta libre y pueden ser utilizados en otros trabajo o ejemplos.

El conjunto de datos es de uso libre y se pueden obtener de la pagina anteriormente mencionada sin ninguna restricción.