57

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ALGUNOS CRITERIOS BASADOS EN LA DISTRIBUCION PREDICTIVA PARA LA SELECCION DE MODELOS.

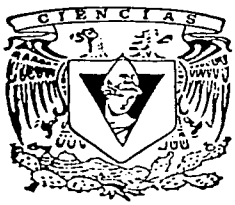
T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

JESICA HERNANDEZ ROJANO



FACULTAD DE CIENCIAS
UNAM

DIRECTOR DE TESIS: M. en C. ALEJANDRO ALEGRIA HERNANDEZ

A mis abuelitos, tíos, primos y amigos, por su apoyo y cariño; a mis profesores, por los conocimientos que me brindaron y en especial a mis padres y hermano, porque sin ustedes nada de esto hubiera sido posible.

INDICE

INTRODUCCIÓN

Para realizar inferencias acerca de cierto proceso o sistema, es conveniente contar con un modelo que represente apropiadamente la información contenida en los datos de los que se dispone. Sin embargo, encontrarlo no es fácil, ya que primero se debe tener un conjunto de modelos que sean candidatos y entre estos elegir al "mejor". Este es un problema muy importante en Estadística, y por muchos años se han propuesto infinidad de criterios para seleccionar modelos.

Al elegir cuál será el conjunto potencial de modelos se debe prestar atención en su tamaño, ya que lo ideal es que sea lo suficientemente grande para no omitir ningún modelo importante, pero no tanto como para incluir modelos que no contengan información acerca del sistema estudiado. Al seleccionar el "mejor" entre estos, debe tomarse en cuenta el propósito para el cual se requiere; por ejemplo, un modelo que pueda considerarse bueno para describir los datos puede no serlo para hacer predicciones. Así, los métodos de selección deben señalar cuáles son los mejores modelos, es decir, los más aceptables, para que, entre ellos, se elija el más conveniente. Además, estos deben ser fáciles de utilizar y ampliamente aplicables en la práctica. El utilizar un buen procedimiento de selección es primordial, ya que si se elige un modelo inapropiado, la inferencia basada en los datos y en ese modelo será poco eficiente.

En la presente tesis se exponen cuatro criterios que se basan en la distribución predictiva final y que se considera tienen algunas ventajas sobre otros métodos de selección de modelos. Son criterios más flexibles, ya que no llevan a elegir un modelo en específico, sino que, al tomar en cuenta la variabilidad de la información, dan la posibilidad de elegir entre varios modelos que se consideran buenos. También son adaptables y se les pueden dar diversos usos.

En el capítulo 1 se habla en general de los modelos, de cuál es el "mejor", de la inferencia basada en éste, de la incertidumbre en la selección y del principio de parsimonia. En el capítulo 2 se da una introducción a la Estadística Bayesiana, con los conceptos básicos necesarios para entender el desarrollo de los criterios. La metodología predictiva se presenta en el capítulo 3 y se explica cómo obtener las distribuciones predictivas finales que se utilizarán para obtener las expresiones analíticas de los criterios.

En el capítulo 4 se definen los criterios L, M y K, se explica su uso y se expone cómo se pueden utilizar las probabilidades iniciales y finales como otro criterio de selección. Para finalizar, en el capítulo 5 se describen brevemente algunos métodos de selección de modelos y se aplican los criterios a diferentes problemas, con la finalidad de ilustrar su uso y ventajas sobre los otros métodos.

1 MODELOS

1.1 SELECCIÓN DE MODELOS

Los datos, y los modelos estocásticos de éstos, se utilizan en las ciencias empíricas para hacer inferencias en relación a procesos y parámetros de interés. Un conjunto simple y único de datos puede ser el sujeto del análisis, pero, con mayor frecuencia, se utilizan datos recolectados en diferentes campos de trabajo o laboratorios, con el objeto de realizar un análisis más amplio. Los datos comúnmente pueden ser extensos y estar particionados por alguna característica en especial. En modelos de regresión lineal, por ejemplo, hay muchas variables explicativas. Frecuentemente, hay variables que tienen efectos muy grandes en esos conjuntos de datos; los parámetros en el modelo representan los efectos de estas variables.

Antes de analizar los datos, se supone que el investigador se ha planteado cuidadosamente una pregunta científica y que se han recolectado datos relevantes utilizando algún tipo de muestreo. Se debe de prestar mucha atención a estos aspectos, ya que se puede tener mala información si la colección de datos es defectuosa o si la pregunta ha sido planteada erróneamente. Se hace hincapié en inferencias que sirvan para entender la estructura y función del sistema estudiado, obtener estimadores de los parámetros relevantes y medidas válidas de precisión o para hacer predicciones.

Frecuentemente, un análisis de este tipo necesitará basarse en un modelo que represente la información contenida en los datos. Los modelos son importantes debido a la posible interpetación de los parámetros involucrados y a que hacen explícitas las posibles relaciones entre las variables. Los parámetros tienen interpretaciones útiles y relevantes, aún cuando se refieran a cantidades que no son directamente observables, pero sin las cuales la ciencia estaría muy

limitada. Por medio del estudio de los datos, se realiza inferencia estadística acerca de una población o proceso real o conceptual, en base a los modelos que involucran tales parámetros.

Al analizar datos empíricos se utiliza la inducción para hacer inferencias estadísticas acerca de una población o proceso definidos, dada una muestra o un conjunto experimental de datos. Entonces, el análisis de datos que nos lleva a una inferencia válida es el proceso integrado por: formulación cuidadosa de una familia de modelos, selección del modelo óptimo, estimación de parámetros y medición de la precisión (incluyendo un componente de variación debido a la incertidumbre en la selección del modelo). La selección del modelo usualmente es un aspecto crítico e integral del análisis científico de datos.

1.1.1 Inferencia, dado un Modelo.

R. A. Fisher (1922) discutió tres aspectos del problema general de una inferencia válida: (i) Especificación del modelo, (ii) estimación de los parámetros del modelo y (iii) estimación de la precisión. La especificación del modelo se puede dividir en 2 componentes: formulación de un conjunto potencial de modelos y selección de uno, o un pequeño número de ellos, para hacer las inferencias. Desde principios del siglo XX, han estado disponibles varios métodos para la objetiva y eficiente estimación de los parámetros de los modelos y su precisión. La *teoría de la verosimilitud* de Fisher ha sido el principal enfoque a estos temas, pero supone que se conoce la estructura del modelo y únicamente los parámetros en tal modelo estructural deben de ser estimados. Tales parámetros pueden estimarse utilizando el método de *máxima verosimilitud* (MV), que, en general, provee una teoría objetiva para la estimación de los parámetros, dado un modelo apropiado.

1.1.2 ¿Qué modelo utilizar?

Fisher creía que la especificación del modelo estaba fuera del campo de la estadística matemática, y esta actitud prevaleció dentro de la comunidad estadística por varias décadas, hasta al menos, principios de los 70's. "¿Qué modelo utilizar?" es la pregunta crítica al hacer una inferencia válida basada en los datos.

Es importante seleccionar un modelo apropiado para el análisis de un conjunto de datos específico; sin embargo, no es lo mismo que tratar de encontrar el "modelo verdadero", es decir, el que supuestamente los generó. Se requieren métodos de selección de modelos con un nivel profundo de soporte teórico y, particularmente, que sean fáciles de utilizar y ampliamente aplicables en la práctica. "Aplicabilidad" se refiere, en parte, a que los métodos tengan buenas características operativas aplicables para tamaños de muestra realistas.

1.1.3 El conjunto potencial de modelos.

La especificación o formulación de un modelo es conceptualmente más difícil que estimar sus parámetros y su precisión. Es en este punto donde la información científica se incorpora en la investigación. La construcción del conjunto potencial de modelos es subjetiva. La literatura publicada y la experiencia en el proceso estudiado pueden ser utilizados como ayuda para formular un conjunto a *priori* de candidatos. La parte más original e innovativa del trabajo científico es la fase relacionada a la formulación de una pregunta apropiada. Modelos que aproximen bien, en conjunción con un buen conjunto de datos relevantes, pueden proveer cierta idea de proceso o sistema estudiado y su estructura.

Los modelos surgen de preguntas teóricas y prácticas que el investigador se hace acerca del proceso o sistema estudiado. Tradicionalmente, las preguntas

teóricas surgen de diversas fuentes como: la literatura científica, resultados de experimentos manipulados, experiencia personal o debates entre la comunidad científica. Las preguntas más prácticas se derivan de programas monitoreados, experimentos, etc.

El desarrollo de un conjunto *a priori* de modelos frecuentemente deberá incluir un modelo global, que tenga muchos parámetros, incluya todos los efectos potencialmente relevantes y refleje los mecanismos causales que se piensen probables. Este modelo deberá también reflejar el diseño del estudio y los atributos del sistema o proceso estudiado. La especificación del modelo global no se basará en un examen de los datos que serán analizados; se deberá investigar primero si el modelo global se ajusta a los datos (examinando residuos y medidas de ajuste, tales como R^2, devianza o pruebas formales de bondad de ajuste). Se procederá al análisis en caso de que se juzgue que el modelo se ajusta aceptablemente a los datos. Se pueden derivar modelos con menos parámetros, como casos especiales del modelo global. Este conjunto de modelos reducidos representa alternativas plausibles basadas en hipótesis o en lo que se conoce acerca del proceso o sistema estudiado. Generalmente, los modelos alternativos diferirán en el número de parámetros en, al menos, un grado en orden de magnitud. Es importante considerar la teoría aceptada, tener un conocimiento profundo del problema y contar con información inicial para saber cuáles son las restricciones sobre los parámetros y las variables de los modelos.

Mientras más parámetros se utilicen, mejor se ajustará el modelo a los datos. Los conjuntos de datos grandes y extensos probablemente soportarán más complejidad. Si un modelo en particular no tiene sentido, no se deberá incluir en el conjunto potencial de modelos. Al formar este conjunto, se debe reconocer cierto balance entre mantenerlo pequeño y enfocado a hipótesis plausibles y hacerlo lo suficientemente grande para evitar omitir algún buen modelo *a priori*. Para mantener este balance se recomienda incluir todos los modelos que parezcan tener una justificación razonable, antes del análisis de los datos.

Freedman (1983) notó que, cuando hay muchas variables que se utilizan para predecir una variable respuesta (Y), los métodos de selección de variables darán ecuaciones de regresión con valores altos de R^2, valores F y muchos coeficientes de regresión "significantes", reflejados en valores grandes de t, aún si las variables explicativas son independientes de Y. Esta indeseable situación ocurre más frecuentemente cuando el número de variables es del mismo orden que el número de observaciones. Una solución parcial a este problema se logra manteniendo pequeño el número de modelos potenciales y logrando muestras de tamaño grande, con respecto al número de parámetros que deberán estimarse. También debe considerarse un conjunto bien fundamentado de modelos para minimizar la inclusión de variables y relaciones falsas.

1.1.4 Los modelos y la realidad.

Algo fundamental es que ninguno de los modelos considerados como base para el análisis de datos es el "modelo real" que genera los datos. La "verdad" (completa realidad) en las ciencias biológicas, por ejemplo, tiene una dimensión infinita, de ahí que la completa realidad no se pueda revelar sólo con muestras finitas de datos y un modelo de ellos. Generalmente, es un error el creer que existe un modelo verdadero simple y que durante el análisis de datos este modelo será descubierto. Los sistemas estudiados pueden ser complejos, con muchos efectos pequeños, interacciones, heterogeneidad individual y covariaciones individuales (la mayoría desconocidas); lo más a lo que se puede aspirar es a identificar un modelo que sea una buena aproximación de la realidad a partir de los datos disponibles.

Un modelo es una simplificación o aproximación de la realidad y, por lo tanto, no la refleja en su totalidad. Como un modelo nunca puede ser la "verdad", puede ser clasificado en muy útil, útil, hasta cierto punto útil e inútil. Los métodos de selección de modelos intentan ordenar los modelos dentro del conjunto de candidatos, de acuerdo a esa clasificación. Que alguno de los

modelos sea realmente "bueno" depende principalmente de la calidad de la metodología aplicada y del pensamiento *a priori* que hubo dentro de la modelación. La completa realidad es evasiva. Una modelación y un análisis de datos apropiados dicen qué inferencias soportan los datos, no qué tan reales son. Un tamaño de muestra grande permite perseguir la completa realidad, pero nunca alcanzarla.

Al utilizar algunos métodos de selección de modelos, se supone que el conjunto de modelos candidato contiene al "modelo real", es decir, el que generó los datos. Esta suposición no debe hacerse, a menos que los datos realmente hayan sido generados por métodos Monte Carlo, por ejemplo. En el análisis de datos reales, no está garantizado que el "modelo verdadero" siquiera exista. Aún si se encontrara la forma del verdadero modelo, no sería bueno para realizar la inferencia general (i.e. entendimiento o predicción) de algún sistema, porque sus numerosos parámetros tendrían que estimarse a partir de una cantidad finita de datos y la precisión sería muy baja.

Frecuentemente, el investigador desea simplificar alguna representación de la realidad para entender los aspectos dominantes del sistema en estudio. Si se tiene una fórmula no lineal con 200 valores de parámetros, se pueden hacer predicciones correctas, pero sería difícil entender las dinámicas principales del sistema sin una posterior simplificación o análisis. En ese caso, se debe de tolerar cierta inexactitud para facilitar el entendimiento del fenómeno.

La aproximación de modelos debe de estar relacionada a la cantidad de datos e información disponible. Los conjuntos pequeños de datos sólo soportarán apropiadamente modelos con pocos parámetros; mientras que conjuntos de datos más grandes soportarán, si es necesario, modelos más complejos.

En los datos existe información sobre la población, proceso o sistema bajo estudio y la meta es expresarla en una manera más compacta y entendible

utilizando un "modelo". Los datos tienen una cantidad finita y fija de información. La meta de la selección de modelos es conseguir una perfecta traducción uno-a-uno, tal que no se pierda ninguna información; sin embargo, tal ideal no se puede conseguir. Los datos deben de particionarse en *información* y *ruido*. El *ruido* es la parte de los datos que no es informativa. Conceptualmente, el papel de un buen modelo es filtrar los datos con el objetivo de separar la información del ruido.

1.1.5 El mejor modelo.

Para que la inferencia sea válida al analizar los datos, un buen modelo debe de tener algunas propiedades. Idealmente, el proceso por medio del cual se ha elegido al "mejor" modelo debe de ser objetivo y repetible; estos son principios fundamentales en la ciencia. Sería ideal que los estimadores de los parámetros fueran insesgados y exactos. El mejor modelo tendría intervalos de confianza de ancho mínimo para los estimadores de los parámetros y su precisión. Finalmente, se quiere que la aproximación de la estructura del sistema sea tan buena como lo permita la información. Si la meta es la predicción, tener todo lo anterior en cuenta podrá garantizar cierta confianza en las predicciones basadas en el modelo. Hay muchos casos en que 2 o más modelos son los "mejores" y esto deberá tomarse en cuenta en posteriores análisis e inferencias. En otros casos podrá haber más modelos que tengan algún soporte, y esos también merecerán escrutinio al sacar conclusiones de los datos, basándose en las inferencias de más de un modelo.

1.1.6 Incertidumbre en la selección de modelos.

Si de un conjunto razonable de modelos se ha seleccionado el mejor, el sesgo en los estimadores de los parámetros puede ser pequeño. Sin embargo, existe incertidumbre en cuál es el mejor modelo a utilizar y esta incertidumbre es una componente de la varianza en los estimadores. La incertidumbre en la selección de un modelo es la componente de la varianza que refleja que esa

selección sólo estima cuál es el mejor, basándose en un conjunto único de datos; para otro conjunto de datos podría escogerse un modelo diferente (dentro del conjunto fijo de modelos considerados).

Cuando no se toma en cuenta la incertidumbre en la selección de modelos, frecuentemente se tienen varianzas y covarianzas muestrales estimadas muy pequeñas, y en consecuencia, los intervalos de confianza encontrados están debajo del nivel de confianza requerido. Los métodos óptimos para hacer frente a la incertidumbre en la selección de modelos están a la vanguardia en la investigación científica: se espera que con los años se encuentren mejores métodos, especialmente con los incrementos en la potencia de las computadoras. La incertidumbre en la selección de modelos es problemática al hacer inferencias estadísticas; si el objetivo sólo es la descripción de datos, entonces tal vez la incertidumbre en la selección sea una cuestión de menor importancia.

1.2 Inferencia y el Principio de Parsimonia.

El éxito del análisis de datos reales y de la inferencia a partir de ellos, frecuentemente depende de la elección del mejor modelo. El análisis de datos debe basarse en un modelo parsimonioso que provea una aproximación precisa a la información estructural contenida en los datos; pero esto no debe verse como la búsqueda del "modelo verdadero".

1.2.1 El principio de Parsimonia.

Si el ajuste se mejora con un modelo con más parámetros, entonces ¿dónde se debe parar? Box y Jenkins (1970) sugirieron que el *principio de parsimonia* debe de llevar a un modelo con "... el menor número posible de parámetros para la adecuada representación de los datos." El principio de parsimonia puede

verse como un balance entre el sesgo y la varianza. En general, el sesgo decrece y la varianza crece conforme aumenta la dimensión del modelo. Frecuentemente, se utiliza al número de parámetros en el modelo como una medida del grado de estructura inferido de los datos. El ajuste de cualquier modelo puede mejorarse incrementando el número de parámetros; sin embargo, al seleccionar un modelo de inferencia, debe considerarse un balance con la creciente varianza. Los modelos parsimoniosos logran un balance adecuado entre sesgo y varianza. Todos los métodos de selección de modelos se basan, hasta cierto punto, en este principio.

Para entender la utilidad de un modelo de aproximación para un conjunto dado de datos, es conveniente considerar dos posibilidades indeseables: el sobre-ajuste y el sub-ajuste en el modelo. Los términos sub- y sobre-ajuste se refieren a la relación entre el modelo elegido y el "mejor modelo". Un modelo sub-ajustado puede ignorar alguna estructura conceptualmente repetible en otras muestras y, así, fallar al identificar efectos que están realmente sustentados por los datos. En este caso, el sesgo en los estimadores de los parámetros es sustancial y la varianza muestral se subestima, resultando ambos factores en malos intervalos de confianza. Estos modelos tienden a ignorar efectos importantes de los tratamientos en los experimentos. Por otro lado, los modelos sobre-ajustados, comparados con el mejor modelo, frecuentemente tienen estimadores de los parámetros libres de sesgo, pero tienen varianzas muestrales estimadas que son innecesariamente grandes (la precisión de los estimadores es menor, en comparación con lo que puede obtenerse con un modelo más parsimonioso). Este tipo de modelos tiende a identificar efectos y variables falsas.

El objetivo de la recolección de datos y su análisis es el hacer inferencias, a partir de una muestra, que se puedan aplicar apropiadamente a la población. A partir de los modelos considerados y los parámetros estimados en cada uno, las inferencias se relacionan con la información acerca de la estructura del sistema estudiado. Una consideración importante es la repetibilidad, con buena

precisión, de cualquier inferencia alcanzada. Si se tienen varias muestras de una misma población, hay características reconocibles que son comunes a casi todas las muestras. De tales características es de las que se busca hacer inferencias más fuertes (a partir de una única muestra obtenida). Otras características que aparecen en, por ejemplo, 60% de las muestras, todavía pueden reflejar algo real acerca de la población o proceso estudiado y se pueden hacer inferencias, más débiles, a partir de ellas. Las características que aparecen sólo en pocas muestras deben de ser incluidas, al modelar, en el término de error del modelo. Si, a partir de este último tipo de características, que aparecen casi de manera única en la muestra estudiada, se hicieran inferencias como si se aplicaran a todas (o casi todas) las muestras (y, en consecuencia, a la población), entonces se dice que el modelo sobreajusta a la muestra. De manera inversa, al fallar al identificar características presentes que son fuertemente repetibles en muchas muestras, se está sub-ajustando. Los datos no son los que están siendo aproximados; lo que se aproxima es la información estructural en los datos que se repite en varias muestras. La cuantificación de tal estructura, con un modelo y estimadores de parámetros, está sujeta a alguna "variación muestral", que también debe de estimarse (inferirse) a partir de los datos.

Las repeticiones son muy ventajosas, pero generalmente se cuenta con ellas sólo en caso de experimentos estrictos basados en la repetición y la aleatoriedad. Tales réplicas experimentales permiten una estimación válida de la variación en los errores. Entender estas cuestiones permite comprender lo que se pierde cuando la única posibilidad es un estudio observacional.

El mejor modelo se logra por medio de un balance apropiado entre los errores de sub-ajuste y sobre-ajuste. Tal balance se logra controlando el sesgo y la varianza. Una apropiada selección de modelos rechaza aquellos que estén lejos de la realidad e intenta identificar un modelo en el cual el error de aproximación y el error debido a fluctuaciones aleatorias estén bien equilibrados. Algunos métodos de selección de modelos son "parsimoniosos" pero tienden a

seleccionar modelos que son muy simples; así, las estimaciones tienen un sesgo grande, una precisión sobrestimada y los intervalos construídos están por debajo del nivel de confianza requerido. Tales instancias no son satisfactorias para la inferencia. Hay que concentrarse en los intervalos de confianza para los estimadores de los parámetros en el modelo seleccionado, incluyendo la componente de varianza para la incertidumbre al seleccionar el modelo.

2 ELEMENTOS DE ESTADÍSTICA BAYESIANA

2.1 INTRODUCCIÓN

El objetivo de la Estadística Bayesiana, en su forma más general, es la revisión de la creencia y la elección de una acción a la luz de nuevos datos. El planteamiento general incorpora tanto inferencia bayesiana como teoría de decisiones. La primera se enfoca en la modificación de las suposiciones por medio de nuevos datos; implica obtener una conclusión probabilística apropiada acerca del fenómeno de interés. La teoría de decisiones implica ir más allá de la inferencia combinando la información que dan las probabilidades de los eventos con la proporcionada por los datos, con el objeto de elegir entre diversas acciones.

Los objetivos y métodos de la Estadística Bayesiana se basan en la interpretación que de la probabilidad tiene la llamada "escuela subjetiva". En ésta, la probabilidad se ve como un grado de creencia y, por lo tanto, no es única, puede variar de individuo a individuo. Es la opinión que tiene cada persona de la ocurrencia de un evento.

Muchas veces las personas pueden estar de acuerdo en la probabilidad de cierto evento y, cuando no es así, la acumulación de datos puede servir para que individuos con diferentes creencias iniciales lleguen a un acuerdo. El punto principal es que la probabilidad no se ve como un hecho objetivo de la naturaleza, sino como un juicio subjetivo que cada persona hace, basándose en su información y experiencia. Así, lo que se cree acerca de los valores de cantidades desconocidas se expresa directamente en términos de probabilidades. El Teorema de Bayes cambia las suposiciones iniciales acerca del evento incorporando la información que proporcionan los nuevos datos. Las

creencias así modificadas, y expresadas en términos de probabilidades subjetivas, constituyen el resultado inferencial.

Se conoce como *Inferencia Estadística* al conjunto de métodos cuyo propósito es extraer conclusiones que van más allá de la mera descripción de los datos. Su problema principal es proporcionar una metodología que permita obtener información acerca de cantidades desconocidas, que no son observables directamente, por medio de variables aleatorias observables, cuyo *comportamiento es influenciado por las cantidades desconocidas.* A estas últimas se les llama *parámetros* de la población o de los procesos aleatorios de los cuales se obtienen las observaciones.

2.2 INFERENCIA.

Ahora, en base a lo anteriormente expuesto, si se está interesado en el valor del *parámetro* θ, se deberá expresar la información disponible sobre su valor mediante una medida de probabilidad que describa el grado de creencia que se tiene acerca de los posibles valores que puede tomar. Se trata de una cantidad aleatoria cuya distribución de probabilidad, $p(\theta)$, describe la información que inicialmente se tiene del parámetro. Esta distribución recibe el nombre de *distribución inicial de* θ. (En realidad, dado que $p(\theta)$ es una medida de la creencia, θ *es considerada una cantidad incierta más que una variable* aleatoria).

Con el objeto de mejorar la información sobre θ, se realiza un experimento ε cuyo resultado $x = \{x_1, x_2, \ldots, x_n\}$ será una muestra aleatoria con una distribución $p(x \mid \theta)$ que depende de θ; en tal caso, la observación x proporcionará, indirectamente, información sobre el valor del parámetro. A $p(x \mid \theta)$, considerada como función de θ, se le llama *función de verosimilitud* y

se denota con $l_x(\theta)$. Esta función expresa la información que proveen las observaciones acerca del parámetro.

Ejemplo.

La función de verosimilitud correspondiente a una muestra aleatoria $\{x_1, x_2, \ldots, x_n\}$ de una población $N(x \mid \mu, \sigma)$, con media μ y desviación estándar σ, será de la forma

$$p(x_1, \ldots, x_n \mid \mu, \sigma) = \prod_{i=1}^{n} N(x_i \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}$$

y

$\hat{\mu} = \overline{x} = \dfrac{\sum x_i}{n}$ y $\hat{\sigma}^2 = \dfrac{\sum (x_i - \overline{x})^2}{n}$ serán los estimadores máximo-verosímiles.

Dada la distribución inicial, $p(\theta)$, y la verosimilitud, $p(x \mid \theta)$, y utilizando el Teorema de la Probabilidad Total, se obtiene para el caso continuo

$$p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta,$$

que recibe el nombre de *distribución predictiva*, ya que se puede utilizar, entre otras cosas, para hacer predicciones sobre los valores de x a que dará lugar el experimento ε.

Ejemplo.

Suponiendo que el parámetro desconocido μ tiene distribución inicial $p(\mu) = N(\mu \mid \mu_0, \sigma_0)$ y que x, el resultado del experimento ε, tiene función de verosimilitud $p(x \mid \mu) = N(x \mid \mu, \sigma)$, con σ conocida, la distribución predictiva está dada por

$$p(x) = \int_{-\infty}^{\infty} N(x \mid \mu, \sigma) N(\mu \mid \mu_0, \sigma_0)\, d\mu = \frac{1}{\sqrt{(\sigma_0^2 + \sigma^2)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\mu_0)^2}{(\sigma_0^2 + \sigma^2)}\right\}$$

que es una densidad $N\left(x \mid \mu_0, \sqrt{(\sigma_0^2 + \sigma^2)}\right)$.

Después de observar el resultado x del experimento, la información de que disponemos sobre el valor de θ estará descrita por su *distribución final* $p(\theta \mid x)$, la cual se obtiene a partir de la distribución inicial $p(\theta)$ y la función de verosimilitud $p(x \mid \theta)$ por medio del *Teorema de Bayes.*

Teorema A (*de Bayes*).- Sea x el resultado del experimento ε, definido mediante el modelo $p(x \mid \theta)$, y sea $p(\theta)$ la distribución inicial de θ. La *distribución final de* θ *es entonces*

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

donde $p(x)$ es la distribución predictiva de x.

Como $p(\theta \mid x)$ es función de θ, $p(x)$ se considera como constante y se puede quitar de la expresión, quedando el Teorema de Bayes expresado, sin pérdida de generalidad, en la forma

$$p(\theta \mid x) \propto p(x \mid \theta)\, p(\theta) = l_x(\theta)\, p(\theta)$$

donde el símbolo $\propto$ significa "es proporcional a".

La distribución final $p(\theta \mid x)$ combina la información inicial sobre θ, contenida en $p(\theta)$, con la información sobre θ proporcionada por el resultado experimental x.

Si se tiene una muestra aleatoria $\{x_1, x_2, \ldots, x_n\}$, resultado del experimento ε, con una distribución $p(x_i \mid \theta)$, y $p(\theta)$ describe la información inicial sobre el valor de θ, la información acumulada sobre θ vendrá dada, en virtud del Teorema de Bayes, por

$$p(\theta \mid x_1, \ldots, x_n) \propto \prod_{i=1}^{n} p(x_i \mid \theta)\, p(\theta)$$

Esta información puede combinarse con la distribución $p(z \mid \theta)$ de una nueva observación z para describir la información que se posee sobre sus posibles valores

$$p(z \mid x_1, \ldots, x_n) = \int p(z \mid \theta)\, p(\theta \mid x_1, \ldots, x_n)\, d\theta$$

Esta distribución recibe el nombre de *distribución predictiva final*. Con la distribución predictiva inicial $p(x)$ pueden hacerse predicciones sobre el resultado experimental x antes de realizar observación alguna. En la distribución predictiva final $p(z \mid x_1, \ldots, x_n)$ se recoge la información proporcionada por las

observaciones $\{x_1, x_2, \ldots, x_n\}$ y se utiliza para predecir el resultado de la próxima observación z.

Ejemplo.

Sea $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una población $N(x_i \mid \mu, \sigma)$, σ conocida. Si la probabilidad inicial es $p(\mu) = N(\mu \mid \mu_0, \sigma_0)$, la correspondiente distribución final $p(\mu \mid x_1, x_2, \ldots, x_n)$ será

$$p(\mu \mid x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n \mid \mu)\, p(\mu)$$
$$\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right\}$$

por lo tanto,

$$p(\mu \mid x_1, \ldots, x_n) = N(\mu \mid \mu_n, \sigma_n)$$

con

$$\sigma_n^2 = \left\{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right\}^{-1} = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + n\sigma_0^2} \quad \text{y} \quad \mu_n = \sigma_n^2\left\{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right\},$$

Sean $h_n = \frac{1}{\sigma_n^2}$, $h_0 = \frac{1}{\sigma_0^2}$, $h = \frac{1}{\sigma^2}$, las precisiones respectivas, se obtiene finalmente que

$$\sigma_n = \frac{1}{\sqrt{h_n}}, \quad h_n = h_0 + nh \quad \text{y} \quad \mu_n = \frac{h_0\mu_0 + n\bar{x}h}{h_0 + nh}.$$

Puede observarse que la precisión final h_n es la suma de la precisión inicial h_0 y la proporcionada por los datos nh. La moda μ_n de la distribución final es la media ponderada de la moda μ_0 de la distribución inicial y la media muestral $\overline{x}$, con pesos proporcionales a sus respectivas precisiones. Así, la distribución final combina la información inicial y la del experimento.

Ahora, la distribución predictiva final de una nueva observación z será

$$p(z \mid x_1, x_2, \ldots, x_n) = \int N(z \mid \mu, \sigma)\, N(\mu \mid \mu_n, \sigma_n)\, d\mu$$
$$= N\left(z \mid \mu_n, \sqrt{\sigma^2 + \sigma_n^2}\right)$$

con σ_n^2 y μ_n como se definieron anteriormente.

Como se esperaba, la distribución predictiva final es de la misma forma que la distribución predictiva inicial, pero usando los parámetros de la distribución final de μ en lugar de los de su distribución inicial.

Comparando la expresión de la distribución predictiva final con la distribución final de μ, $N(\mu \mid \mu_n, \sigma_n)$, se observa que tiene la misma media y una desviación estándar necesariamente mayor. Cuando el tamaño muestral n crece, la desviación estándar de la distribución final, σ_n, tiende a cero, lo que nos permite conocer el valor de μ con una precisión arbitrariamente grande; sin embargo, cuando n crece, la desviación estándar de la distribución predictiva final sólo tiende a σ, por lo que no podemos aspirar a predecir con mayor precisión el resultado de una observación futura.

2.3 MÉTODOS APROXIMADOS DE INFERENCIA.

2.3.1 Comportamiento asintótico de la distribución final.

Cuando se tiene acceso a un gran número de datos, se puede obtener de estos mayor información; en este caso, la distribución final admite una aproximación asintótica que es más precisa mientras mayor es la cantidad de datos de que se dispone. Más concretamente, para casi todos los modelos probabilísticos $p(x \mid \theta)$ y cualquier distribución inicial $p(\theta)$, la distribución final $p(\theta \mid x)$, después de observar los resultados experimentales $x = \{x_1, \ldots, x_n\}$, se aproxima a una distribución Normal a medida que crece el tamaño n de la muestra. Este es el contenido fundamental del siguiente resultado.

Teorema B.- Sean x_i los resultados de un experimento ε cuya distribución es $p(x \mid \theta)$ y supongamos que el conjunto X de los posibles valores de x_i no depende de θ. Sea $p(\theta)$ la distribución inicial de θ y $x = \{x_1, \ldots, x_n\}$ el resultado de n realizaciones independientes de tal experimento. Entonces, para valores de n suficientemente grandes, la distribución final de θ, $p(\theta \mid x_1, \ldots, x_n)$, es aproximadamente Normal con media y desviación estándar dadas, respectivamente, por

$$E(\theta \mid x) = \{\theta_0 h_0 + \hat{\theta}\, h(\hat{\theta})\} / \{h_0 + h(\hat{\theta})\}$$

$$D(\theta \mid x) = 1 \Big/ \left\{\sqrt{h_0 + h(\hat{\theta})}\right\}$$

donde

$\hat{\theta}$ es el valor que maximiza la función de verosimilitud $p(x \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$,

θ_0 es la moda de la distribución inicial de θ,

$$h(\hat{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p(x_i \mid \theta)\Big|_{\theta=\hat{\theta}}, \text{ y}$$

$$h_0 = -\frac{\partial^2}{\partial \theta^2} \log p(\theta)\Big|_{\theta=\theta_0}$$

Ejemplo.

En el modelo Normal: Suponiendo que $x = \{x_1, \ldots, x_n\}$ es el resultado de n observaciones independientes con distribución $N(x_i \mid \mu, \sigma)$, σ conocida, y $p(\mu) = N(\mu \mid \mu_0, \sigma_0)$; la moda inicial de μ será μ_0 y $h_0 = 1/\sigma_0^2$. Por otra parte, $\hat{\mu} = \bar{x}$, $h(\hat{\mu}) = n/\sigma^2$, $h_0 = 1/\sigma_0^2$ y $h = 1/\sigma$. En consecuencia, aplicando el Teorema B, la distribución asintótica de μ es Normal con

$$E(\mu \mid x) = \{\mu_0 h_0 + n\bar{x}h\} / \{h_0 + nh\} = \mu_n$$
$$D(\mu \mid x) = 1 / \left\{\sqrt{h_0 + nh}\right\} = \sigma_n$$

que son los parámetros obtenidos al calcular exactamente la distribución final de μ. En consecuencia, si el modelo es $N(x \mid \mu, \sigma)$, σ conocida, y la distribución inicial es normal $N(x \mid \mu_0, \sigma_0)$, el teorema da lugar al resultado exacto.

Teorema C.- Bajo las condiciones del teorema anterior, si el tamaño n de la muestra es muy grande, entonces la distribución final $p(\theta \mid x)$ puede ser aproximada por

$$p^*(\theta \mid \hat{\theta}) = N\{\theta \mid \hat{\theta}, D(\hat{\theta})\}$$

donde $\hat{\theta}$ es el estimador máximo verosímil de θ, $D(\hat{\theta}) = \{h(\hat{\theta})\}^{-\frac{1}{n}}$ y

$$h(\hat{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p(x_i \mid \theta)\Big|_{\theta=\hat{\theta}}$$

La distribución final $p^*(\theta \mid \hat{\theta})$ así obtenida es la que generalmente se conoce como *distribución asintótica* de θ correspondiente al modelo $p(x \mid \theta)$. Cualquiera que sea la distribución inicial $p(\theta)$, la verdadera distribución final $p^*(\theta \mid x)$ se aproxima a $p^*(\theta \mid \hat{\theta})$ cuando n tiende a infinito.

El Teorema C puede generalizarse en el caso de que el modelo $p(x \mid \theta)$ dependa de varios parámetros, de manera que $\theta = \{\theta_1, \ldots, \theta_k\}$ es un vector.

Teorema D.- La distribución asintótica del vector $\theta = \{\theta_1, \ldots, \theta_k\}$ de dimensión k es la distribución normal k-variada $p^*(\theta \mid \theta) = N\{\theta \mid \theta, H(\theta)^{-1}\}$ donde $\hat{\theta}$ es el vector que maximiza la función de verosimilitud $p(x \mid \theta)$ y la matriz de precisión $H(\hat{\theta})$ tiene como elemento genérico

$$h_{ij}(\hat{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x_i \mid \theta)\Big|_{\theta=\hat{\theta}}$$

Ejemplo.

En la distribución Normal

$$\hat{\theta} = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \begin{pmatrix} \bar{x} \\ s \end{pmatrix}$$

donde $\bar{x} = \frac{\sum x_i}{n}$ y $s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$, con matriz de precisión

$$H(\hat{\theta}) = H(\mu, \sigma) = \begin{pmatrix} n/s & 0 \\ 0 & 4n/s^2 \end{pmatrix}$$

Los parámetros μ y σ resultan asintóticamente independientes y, como consecuencia, sus distribuciones asintóticas marginales y condicionales coinciden y son de la forma

$$p^*(\mu \mid \bar{x}, s, \sigma) = p^*(\mu \mid \bar{x}, s) \cong N\{\mu \mid \bar{x}, s/\sqrt{n}\}$$
$$p^*(\sigma \mid \bar{x}, s, \sigma) = p^*(\sigma \mid \bar{x}, s) \cong N\{\sigma \mid s, s/(2\sqrt{n})\}$$

2.3.2 Información proporcionada por la muestra.

Por medio de la teoría de la información se puede dar una medida que permite cuantificar la información inicial sobre θ en términos de $p(\theta)$. Dicha medida debe ser una función decreciente de $p(\theta)$. Se elige la función (Fernández (1978))

$$\log \frac{1}{p(\theta)}$$

Dicha medida decrece en el intervalo [0,1] y es única, bajo ciertas condiciones.

Suponiendo que se tiene la información de que θ toma el valor θ_i, la cantidad de información recibida se mide con $\log(1/p(\theta_i))$. Sin embargo, antes de recibir tal información, solamente se conoce la probabilidad de que θ tome

cierto valor, no el valor en sí. No obstante, es posible calcular la información esperada

$$H\{p(\theta_i)\} = \sum_{i=1}^{n} p(\theta_i) \log \frac{1}{p(\theta_i)} = -\sum_{i=1}^{n} p(\theta_i) \log\left(p(\theta_i)\right)$$

si θ es una variable aleatoria discreta, o

$$H\{p(\theta)\} = -\int p(\theta) \log\left(p(\theta)\right) d\theta$$

si θ es una variable aleatoria continua. Esta expresión también recibe el nombre de *entropía* de la distribución $p(\theta)$.

Supóngase ahora que se ha realizado el experimento ε. Si se informa que θ toma el valor θ_i se tiene una información $\log\left(1/p(\theta_i)\right)$, si se utiliza la probabilidad inicial; y una información $\log\left(1/p(\theta_i \mid x)\right)$, si se utiliza la información final. De forma natural se puede definir la información asociada a x, que transforma la probabilidad inicial $p(\theta_i)$ en la final $p(\theta_i \mid x)$, como

$$\log \frac{1}{p(\theta_i)} - \log \frac{1}{p(\theta_i \mid x)} = \log \frac{p(\theta_i \mid x)}{p(\theta_i)}$$

Se define entonces a

$$I^{\theta}\{\varepsilon, p(\theta_i) \mid x\} = \sum p(\theta_i \mid x) \log \frac{p(\theta_i \mid x)}{p(\theta_i)}$$

para θ discreta, o

$$I^{\theta}\{\varepsilon, p(\theta) \mid x\} = \int p(\theta \mid x) \log \frac{p(\theta \mid x)}{p(\theta)} d\theta$$

para θ continua, como la *información esperada* proporcionada por x sobre el valor de θ.

Ejemplo.

Sea $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una población $N(x_i \mid \mu, \sigma)$, σ conocida. Si la probabilidad inicial es $p(\mu) = N(\mu \mid \mu_0, \sigma_0)$, entonces $p(\mu \mid x_1, \ldots, x_n) = N(\mu \mid \mu_n, \sigma_n)$, así la *cantidad de información* proporcionada por la muestra será

$$I^{\mu}\{\varepsilon, N(\mu \mid \mu_0, \sigma_0) \mid x_1, \ldots, x_n\} = \int N(\mu \mid \mu_0, \sigma_0) \log \frac{N(\mu \mid \mu_n, \sigma_n)}{N(\mu \mid \mu_0, \sigma_0)} \, d\mu$$

cuyo valor esperado resulta ser

$$I^{\mu}\{\varepsilon, N(\mu \mid \mu_0, \sigma_0) \mid n\} = \log\left(\frac{\sigma_0}{\sigma_n}\right) = \log \sigma_0 - \log \sigma_n .$$

esto es, la reducción en la incertidumbre que se tiene sobre el valor de μ, en escala logarítmica. Esta expresión también puede escribirse como $(1/2)\log(1 + nh/h_0)$, que es una función cóncava y creciente del tamaño muestral n, y una función creciente del cociente h/h_0 de la precisión de las observaciones y la precisión inicial.

2.3.3 Aproximación normal a la distribución final.

La mayor parte de las densidades de probabilidad que aparecen como resultado de problemas reales de inferencia son difíciles de integrar para obtener las probabilidades deseadas. Por lo tanto, resulta muy conveniente disponer de métodos que nos permitan reducir el cálculo de probabilidades asociadas a una distribución cualquiera, al cálculo de probabilidades utilizando una distribución Normal.

Definición.- La *discrepancia* entre una densidad de probabilidad $p(\theta)$ y su aproximación $q(\theta)$ es el valor de la integral

$$\delta\{p(\theta), q(\theta)\} = \int p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta ,$$

que es la cantidad de información sobre θ que resulta necesaria para pasar de la aproximación $q(\theta)$ a la densidad de probabilidad verdadera $p(\theta)$. La aproximación $q(\theta)$ será mejor cuanto menor sea la discrepancia $\delta\{p(\theta), q(\theta)\}$.

Para empezar, se sabe que la mejor aproximación normal (con menos discrepancia) a una distribución $p(\theta)$ es la distribución $N(\theta \mid E(\theta), D(\theta))$, esto es, aquella distribución Normal con la misma media y desviación estándar que la distribución original (**Teorema E**).

No siempre se puede aproximar bien, es decir, con una discrepancia pequeña, una distribución cualquiera por medio de una distribución normal; no obstante, frecuentemente es posible encontrar una transformación monótona de θ, $\xi = \xi(\theta)$, razonablemente sencilla, cuya densidad de probabilidad, $p(\xi)$, sea aproximadamente normal.

En efecto, si se desea calcular $p[a < \theta < b]$ y se sabe que la función $\xi = \xi(\theta)$ tiene una distribución aproximadamente normal $p(\xi) = N(\xi \mid E(\xi), D(\xi))$, tenemos que

$$p[a < \theta < b] = p[\xi(a) < \xi < \xi(b)]$$

que resulta ser

$$\Phi\left[\frac{\xi(b) - E(\xi)}{D(\xi)}\right] - \Phi\left[\frac{\xi(a) - E(\xi)}{D(\xi)}\right]$$

donde Φ es la función de distribución normal estándar.

Para una determinada distribución inicial $p(\theta)$, el problema de encontrar la mejor transformación normalizadora posible, $\xi = \xi(\theta)$, que minimice la discrepancia entre $p(\theta)$ y alguna distribución normal, equivale, en virtud del Teorema de Cambio de Variable y el Teorema E, a encontrar la función monótona de θ, $\xi = \xi(\theta)$, que minimiza la integral

$$\int p(\xi) \log\left[\frac{p(\xi)}{N(\xi \mid E(\xi), D(\xi))}\right] d\xi$$

donde $p(\xi) = p(\theta)/|\partial\xi/\partial\theta|$ y $E(\xi)$, $D(\xi)$ son respectivamente la media y desviación estándar de ξ. Este es un problema difícil de cálculo de variaciones para el que sólo se conocen resultados parciales.

2.3.4 Distribuciones finales de referencia.

Cuando no se dispone de información inicial o cuando tal información no se puede o no se quiere utilizar, se puede determinar una *distribución inicial de referencia*, que describe la situación en que los datos experimentales contienen toda la información relevante, en lugar de proporcionar sólo una parte de ella, como sucede cuando se dispone de información inicial.

Sea C la clase de distribuciones iniciales admisibles. Sea ε un experimento cuyo resultado $z = \{z_1, z_2, \ldots, z_k\}$ tiene una distribución $p(z \mid \theta)$, y sea $I^{\theta}\{\varepsilon(k), p(\theta)\}$ la información que podría esperarse de k repeticiones independientes de ε cuando la distribución inicial es $p(\theta)$. Repitiendo indefinidamente el experimento, se llegaría a conocer exactamente el valor de θ; en consecuencia,

$$\lim_{k \to \infty} \quad I^{\theta}\{\varepsilon(k), p(\theta)\}$$

mide, para cada $p(\theta)$, la *cantidad de información desconocida* sobre θ cuando su distribución inicial es $p(\theta)$. Sea $\pi_k(\theta)$ la distribución de θ que *maximiza* en C el valor de $I^{\theta}\{\varepsilon(k), p(\theta)\}$; la *distribución inicial de referencia* es, entonces,

$$\pi(\theta) = \lim_{k \to \infty} \pi_k(\theta)$$

La *distribución final de referencia*, una vez observado el resultado z del experimento ε, es entonces,

$$\pi(\theta \mid z) \propto p(z \mid \theta)\, \pi(\theta)$$

Teorema F.- Consideremos un experimento ε, cuyo resultado z tiene una distribución que depende de un parámetro continuo θ y supongamos que su

correspondiente distribución asintótica es $p^*(\theta \mid \hat{\theta})$. La distribución inicial de referencia para θ, correspondiente a ε, viene entonces dada por

$$\pi(\theta) = \lim_{k \to \infty} \exp\left\{ \int \dots \int p(z_1, \dots, z_k \mid \theta) \log p^*(\theta \mid \hat{\theta}) dz_1 \dots dz_k \right\}$$

donde $\hat{\theta}$ es el valor de θ que maximiza $p(z_1, \dots, z_k \mid \theta)$.

Teorema G.- Supongamos que la distribución asintótica del parámetro θ de un experimento ε es $N\{\theta \mid \hat{\theta}, D(\hat{\theta})\}$, donde $\hat{\theta}$ es el estimador máximo verosímil de θ. Entonces, si C es la clase de todas las distribuciones de θ que pueden definirse sobre el conjunto Θ, y no existen parámetros marginales, la distribución inicial de referencia para θ correspondiente al experimento ε es

$$\pi(\theta) \propto \begin{cases} 1/D(\theta), & \theta \in \Theta \\ 0 & e.o.c. \end{cases}$$

Bajo las condiciones usuales de regularidad para la normalidad asintótica, la distribución inicial de referencia será

$$\pi(\theta) \propto I_\theta^{1/2}$$

donde

$$I_\theta = -E\left\{ \frac{\partial^2}{\partial \theta^2} \log p(x \mid \theta) \right\} = -\int p(x \mid \theta) \frac{\partial^2}{\partial \theta^2} \log p(x \mid \theta) \, d\theta$$

En el caso multivariado con $\underline{\theta} \in \Theta \subset \Re^p$, y siempre y cuando se deseen realizar inferencias conjuntas,

$$\pi(\underline{\theta}) \propto \left| I_{\underline{\theta}} \right|^{1/2}$$

donde

$$I_{\underline{\theta}} = \{a_{ij}\},$$

$$a_{ij} = -E\left\{\frac{\partial^2}{\partial\theta_i\theta_j}\log p(x \mid \theta)\right\} = -\int p(x \mid \theta)\frac{\partial^2}{\partial\theta^2}\log p(x \mid \theta)\, dx$$

I_θ es la conocida matriz de información de Fisher.

Ejemplo.

Considerando el modelo Normal. La distribución asintótica del parámetro μ de una distribución normal $N(x \mid \mu, \sigma)$, σ conocida, es $N\left(\mu \mid \overline{x}, \sigma/\sqrt{n}\right)$. En consecuencia, la distribución inicial de referencia para μ debe cumplir

$$\pi(\mu) \propto \frac{1}{D(\mu)} = \frac{\sqrt{n}}{\sigma}$$

Suponiendo que $\mu \in (a_0, a_1)$, la distribución inicial de referencia para μ será

$$\pi(\mu) = \frac{1}{a_1 - a_0},$$

la distribución uniforme sobre el conjunto de posibles valores.

Si se han realizado n observaciones independientes $\{x_1, x_2, \ldots, x_n\}$ de una distribución $N(x \mid \mu, \theta)$, σ conocida, la distribución final de referencia será, usando el Teorema de Bayes con $\pi(\mu) = 1/(a_1 - a_0)$,

$$\pi(\mu \mid \bar{x}) = \begin{cases} \dfrac{N\left(\mu \mid \bar{x}, \sigma/\sqrt{n}\right)}{\Phi\left(\dfrac{a_1 - \bar{x}}{\sigma/\sqrt{n}}\right) - \Phi\left(\dfrac{a_0 - \bar{x}}{\sigma/\sqrt{n}}\right)}, & para \quad \mu \in (a_0, a_1) \\ 0, & e.o.c. \end{cases}$$

. Si $a_0 \to -\infty$ y $a_1 \to \infty$ el resultado anterior se reduce a $\pi(\mu \mid \bar{x}) = N\left(\mu \mid \bar{x}, \sigma/\sqrt{n}\right)$.

Existen varios procedimientos para obtener distribuciones iniciales, como el de Jeffreys, que coincide con el presentado aquí, que es el de Bernardo.

3 METODOLOGÍA PREDICTIVA

3.1 INTRODUCCIÓN.

Como ya se mencionó anteriormente, una pregunta fundamental en Estadística es: ¿Cuál de los modelos $M_1, M_2, \ldots, M_m$ explica mejor a un conjunto de datos? Sin embargo, en la mayoría de las circunstancias la pregunta más pertinente sería: ¿Cuál de los modelos $M_1, M_2, \ldots, M_m$ proporciona las mejores predicciones para futuras observaciones del mismo proceso que generó el conjunto de datos? Esta pregunta es más difícil de responder que la primera, sin embargo, se relaciona más directamente con el uso que se les da a los modelos.

Si el proceso que genera las observaciones es aleatorio y se especifica una función de distribución $F(\cdot \mid M_k)$, dado el modelo M_k, entonces se calculan las verosimilitudes bajo los diferentes modelos y se puede escoger al más probable, dados los datos. Si están disponibles las probabilidades iniciales que reflejan la incertidumbre sobre varios modelos, entonces, dados los datos, se puede calcular al más probable entre aquellos en consideración.

Pero en la realidad los problemas no son tan fáciles de resolver. En el mejor de los casos, los modelos incluyen distribuciones que están especificadas, salvo ciertos parámetros desconocidos. En otras circunstancias, los modelos pueden implicar funciones de distribución y parámetros completamente distintos, de manera que ningún modelo puede considerarse dentro de otro. Alternativamente, cuando un modelo puede considerarse como un caso especial de otro, o dentro de éste, se dice que están anidados.

Existen técnicas Bayesianas y no Bayesianas que tratan la selección de modelos. Estas técnicas pueden aplicarse, por ejemplo, en la selección de

variables o la selección de transformaciones adecuadas para la variable predictiva y/o la variable respuesta; o también en la selección de la función de varianza apropiada, en el modelo lineal heteroscedástico. Los problemas anteriores surgen en los modelos lineales generalizados, en particular en el caso de regresión lineal. También pueden aplicarse esas técnicas en el análisis de series de tiempo o en modelos no-lineales.

Se han propuesto algunos criterios para seleccionar modelos, entre los cuales los más aceptados son el criterio de información de Akaike (AIC) y el criterio Bayesiano de Información de Schwarz (BIC). El problema con estos criterios es que no permiten la incorporación de información inicial al realizar la selección; además, sus definiciones y/o calibraciones se basan de manera sustantiva en consideraciones asintóticas.

En el presente trabajo se expondrán tres criterios, propuestos por Laud & Ibrahim (Laud, 1995) que pueden utilizarse en la selección de modelos. Tales criterios se enfocan en los datos observados y se basan en cierta densidad predictiva Bayesiana. Tienen una base única que es simple e interpretable, están libres de definiciones asintóticas y permiten la incorporación de información inicial. Además, dos de ellos se calculan fácilmente. El caso de selección de variables, también se tratará desde un punto de vista Bayesiano, en el cual el investigador necesita especificar la probabilidad inicial para cada modelo, una distribución inicial para todos los parámetros y calcular la probabilidad final para cada modelo dados los datos, para así elegir al más probable.

Sólo se considerará el problema de selección de modelos en regresión lineal.

3.2 MODELO DE REGRESIÓN LINEAL.

3.2.1 Distribución predictiva.

Se comienza con el modelo completo,

$$Y = X\beta + \varepsilon, \tag{3.1}$$

donde $Y_{n\times 1}$ es el vector de respuestas, $\beta_{(k+1)\times 1}$ es el vector de los coeficientes de regresión, $\varepsilon_{n\times 1}$ es el vector de errores aleatorios y $X_{n\times(k+1)}$ es la matriz de variables independientes de rango completo, que consiste de una columna de unos, para el término de intersección, seguida por k columnas, cada una de las cuales representa una variable (la primera columna de esta matriz será la número 0).

$$Y_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \beta_{(k+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ y } X_{n\times(k+1)} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}$$

La distribución de ε usualmente se supone normal multivariada de dimensión n, con media 0 y matriz de *precisión* τI, donde τ es un escalar positivo e $I_{n\times n}$ es la matriz identidad, es decir, $\varepsilon \mid \tau \sim N_n\left(\varepsilon \mid 0, (\tau I)^{-1}\right)$.

Para seleccionar variables se deben considerar los 2^k posibles modelos que se obtienen del modelo (3.1) al conservar varios subconjuntos de las últimas k columnas de la matriz X y modificando β de acuerdo al subconjunto. Se le llamará m al modelo relacionado con el conjunto de enteros $\{0,1,\ldots,k\}$, y k_m al número de elementos de este conjunto. Así, m identificará al modelo con la

intersección (representada por el 0) y las $(k_m - 1)$ variables explicativas elegidas. Sea M el conjunto de los 2^k modelos considerados, podemos expresarlos como:

$$Y = X_m \beta^{(m)} + \varepsilon, \qquad m \in M \tag{3.2}$$

donde X_m denota la matriz de variables explicativas de rango completo de $(n \times k_m)$ entradas bajo el modelo m y $\beta^{(m)}$ es el correspondiente vector de coeficientes. El objetivo de los métodos de selección de variables es elegir uno de los modelos representados por la ecuación (3.2).

Para el problema de selección de modelos en general, en lugar de utilizar la ecuación (3.2) y las descripciones de las distribuciones de las diferentes cantidades, se pueden considerar modelos probabilísticos para el vector Y observado, condicionado a cada modelo m, y su correspondiente vector de parámetros $\theta^{(m)}$. Entonces se escribe

$$p\left(y \mid m, \theta^{(m)}\right), \qquad m \in M, \quad \theta^{(m)} \in \Theta^{(m)},$$

donde M es el espacio de modelos y $\Theta^{(m)}$ es el espacio parametral para el modelo m. Como en la selección de modelos el parámetro $\theta^{(m)}$ no tiene porqué tener, en principio, un significado físico, se puede elegir un modelo m adoptando un punto de vista Bayesiano, ya que éste permite dar mayor importancia a las observaciones que a los parámetros.

En la estadística Bayesiana se utiliza la llamada densidad predictiva para pronosticar observaciones futuras. Hay muchas formas de predecir observaciones, sin embargo, el enfoque Bayesiano es natural ya que la predicción se basa en la distribución condicional del futuro dado el pasado. Para hacer esto es conveniente tratar a los parámetros del modelo como aleatorios.

Aquí se implementará la filosofía predictiva de dos maneras. Primero, cuando es posible, las distribuciones iniciales de $\theta^{(m)} \mid m$ se construyen a partir de una predicción inicial de Y y un número que cuantifica la creencia que se tiene en relación a la información contenida en el experimento. Segundo, no se utiliza una distribución inicial en el espacio de modelos M; en su lugar, para cada modelo $m \in M$, se calcula un criterio y se elige el modelo que lo optimice.

Supóngase que se ha especificado una distribución inicial de referencia $\pi\left(\theta^{(m)} \mid m\right)$ para cada $\theta^{(m)}$, $m \in M$, entonces la distribución final de referencia de $\theta^{(m)}$ bajo cada modelo m, dados los datos $Y = y$, está dada por

$$\pi\left(\theta^{(m)} \mid y, m\right) = \frac{p\left(y \mid \theta^{(m)}, m\right)\pi\left(\theta^{(m)} \mid m\right)}{\int p\left(y \mid \theta^{(m)}, m\right)\pi\left(\theta^{(m)} \mid m\right) d\theta^{(m)}}$$
$$\propto p\left(y \mid \theta^{(m)}, m\right)\pi\left(\theta^{(m)} \mid m\right)$$

Ahora, imaginando que se repite el experimento completo, se denotará por Z al vector de respuestas resultante. Tal experimento tiene la misma matriz X que el primero. En regresión lineal, tenemos que $\theta^{(m)} = \left(\beta^{(m)}, \tau\right)$, donde cada m especifica a una matriz X_m. Más aún, bajo cualquier modelo $m \in M$, de nuevo se tiene la misma matriz X_m. Podemos expresar los modelos resultantes del experimento repetido como

$$Z = X_m \beta^{(m)} + e, \qquad m \in M \tag{3.3}$$

donde e es el vector de errores aleatorios. Z se puede ver como el vector de "predicciones".

La densidad predictiva final para Z bajo el modelo m es

$$p(z \mid m, y) = \int p\left(z \mid m, \theta^{(m)}\right) \pi\left(\theta^{(m)} \mid y, m\right) d\theta^{(m)}$$

a la que se llamará *densidad predictiva final del experimento replicado* (DPER). En el caso de la ecuación (3.3), esta densidad depende de X_m para el modelo m, quedando entonces expresada como

$$p(z \mid X_m, y) = \iint p\left(z \mid X_m, \beta^{(m)}, \tau\right) p\left(\beta^{(m)}, \tau \mid X_m, y\right) d\beta^{(m)} d\tau$$

Para facilitar la notación, se denotará a la DPER como f_m.

Aunque la DPER es primordial para calcular los criterios, no se espera que realmente se repita el experimento ni se centra el interés en ninguna matriz de variables independientes. El experimento repetido es un elemento imaginario que hace que y y Z sean comparables. Además, los parámetros en el modelo no tienen mucha importancia en la repetición. Es claro que entre todos los modelos que se consideran, los buenos son los que producen predicciones cercanas a lo que se ha observado en un experimento idéntico.

3.2.2 Distribución predictiva con distribución inicial no informativa.

Primero se propondrá una distribución inicial de referencia no informativa para el modelo lineal (3.2). Sea entonces

$$\pi\left(\beta^{(m)}, \tau\right) \propto \tau^{-1}, \qquad \beta^{(m)} \in \Re^{k_m}, \tau > 0 \tag{3.4}$$

que es la llamada distribución inicial no informativa para $\beta^{(m)}$ y τ y que fue desarrollada por Jeffreys. Esta distribución inicial satisface ciertas reglas de invarianza y transmite muy poca información acerca de los parámetros. Aunque

esta densidad inicial es impropia, produce una densidad final normal-gamma para $\beta^{(m)}$ y τ, que es propia.

La distribución inicial de Jeffreys implica que, a priori, $\beta^{(m)}$ y τ son independientes, $\beta^{(m)}$ tiene una densidad constante sobre $\Re^{k_m}$ y la densidad marginal inicial de τ es $\pi(\tau) \propto \tau^{-1}, \tau > 0$.

Por otro lado, la función de verosimilitud es

$$p\left(y \mid \beta^{(m)}, \tau\right) \propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left(y - X_m\beta^{(m)}\right)'\left(y - X_m\beta^{(m)}\right)\right) \qquad (3.5)$$

entonces, la distribución final para $\beta^{(m)}$ y τ, será

$$\begin{aligned} p\left(\beta^{(m)}, \tau \mid X_m, y\right) &= p\left(y \mid \beta^{(m)}, \tau\right) \cdot \pi\left(\beta^{(m)}, \tau\right) \\ &\propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left(y - X_m\beta^{(m)}\right)'\left(y - X_m\beta^{(m)}\right)\right) \cdot \tau^{-1} \\ &= \tau^{\frac{n}{2}-1} \exp\left(-\frac{\tau}{2}\left(y - X_m\beta^{(m)}\right)'\left(y - X_m\beta^{(m)}\right)\right) \\ &= \tau^{\frac{n-k_m}{2}-1} \exp\left(-\frac{\tau}{2}A\right) \cdot \tau^{\frac{k_m}{2}} \exp\left(-\frac{\tau}{2}B\right) \end{aligned} \qquad (3.6)$$

que es una normal-gamma, con

$$A = y'(I - P_m)y$$
$$B = \left[\beta^{(m)} - (X_m'X_m)^{-1}X_m'y\right]'(X_m'X_m)\left[\beta^{(m)} - (X_m'X_m)^{-1}X_m'y\right]$$

y donde $P_m = X_m(X_m'X_m)^{-1}X_m'$ es la matriz de proyección ortogonal sobre $C(X_m)$, el espacio generado por las columnas de X_m.

Como se mencionó anteriormente, el vector Z tiene la misma matriz de diseño X_m, bajo el modelo m, que el vector Y, entonces

$$p\left(z \mid X_m, \beta^{(m)}, \tau\right) \propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left(Z - X_m\beta^{(m)}\right)'\left(Z - X_m\beta^{(m)}\right)\right) \qquad (3.7)$$

La DPER entonces estará dada por

$$\begin{aligned} p(z \mid X_m, y) &= \int_0^\infty \int_{\Re^{k_m}} p\left(z \mid X_m, \beta^{(m)}, \tau\right) p\left(\beta^{(m)}, \tau \mid X_m, m\right) d\beta^{(m)} d\tau \\ &\propto \int_0^\infty \int_{\Re^{k_m}} \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}C\right) \cdot \tau^{\frac{n}{2}-1} \exp\left(-\frac{\tau}{2}D\right) d\beta^{(m)} d\tau \\ &= \int_0^\infty \int_{\Re^{k_m}} \tau^{n-1} \exp\left(-\frac{\tau}{2}(C+D)\right) d\beta^{(m)} d\tau \end{aligned}$$

con

$$\begin{aligned} C &= \left(Z - X_m\beta^{(m)}\right)'\left(Z - X_m\beta^{(m)}\right) \\ D &= \left(y - X_m\beta^{(m)}\right)'\left(y - X_m\beta^{(m)}\right) \end{aligned}$$

Ahora, desarrollando $C + D$ para completar cuadrados en $\beta^{(m)}$, queda

$$\begin{aligned} C + D &= y'y + Z'Z - \tfrac{1}{2}\left(X_m'y + X_m'Z\right)'\left(X_m'X_m\right)\left(X_m'y + X_m'Z\right) \\ &\quad + 2\left(\beta^{(m)} - \tfrac{1}{2}\left(X_m'X_m\right)'\left(X_m'y + X_m'Z\right)\right)'\left(X_m'X_m\right)\left(\beta^{(m)} - \tfrac{1}{2}\left(X_m'X_m\right)'\left(X_m'y + X_m'Z\right)\right) \end{aligned}$$

Así

$$p(z \mid X_m, y) = \int_0^\infty \tau^{\frac{2n-k_m}{2}-1} \exp\left(-\frac{\tau}{2}C^*\right)\left[\int_{\Re^{k_m}} \tau^{\frac{k_m}{2}} \exp\left(-\tau D^*\right) d\beta^{(m)}\right] d\tau$$

con

$$C^* = y'y + Z'Z - \tfrac{1}{2}(X_m'y + X_m'Z)'(X_m'X_m)(X_m'y + X_m'Z)$$

$$D^* = \left(\beta^{(m)} - \tfrac{1}{2}(X_m'X_m)'(X_m'y + X_m'Z)\right)'(X_m'X_m)\left(\beta^{(m)} - \tfrac{1}{2}(X_m'X_m)'(X_m'y + X_m'Z)\right)$$

Después de integrar, la distribución predictiva final de Z es

$$p(z \mid X_m, y) \propto \left[y'y + Z'Z - \tfrac{1}{2}(X_m'y + X_m'Z)'(X_m'X_m)(X_m'y + X_m'Z)\right]^{-\frac{2n-k_m}{2}}$$

$$= \left[(Z - P_m y)'(I - \tfrac{1}{2}P_m)(Z - P_m y) + y'(I - P_m)y\right]^{-\frac{2n-k_m}{2}}$$

que puede expresarse como

$$Z \sim S_n\left(n - k_m,\ P_m y,\ s_m^2(I + P_m)\right) \qquad (3.8)$$

donde

$$s_m^2 = (n - k_m)^{-1} y'(I - P_m)y,$$

y $S_n(\nu, \mu, \Sigma)$ denota una distribución t-multivariada de dimensión n con ν grados de libertad, parámetro de localización μ y matriz de dispersión Σ.

3.2.3 Distribución predictiva con distribución inicial informativa.

Seleccionar distribuciones iniciales informativas para $\beta^{(m)}$ que tengan sentido no es una tarea fácil, aún para un modelo fijo m. Para una colección de modelos, generalmente no es factible interpretar cada componente de $\beta^{(m)}$ para cada modelo posible. Sólo se puede aspirar a utilizar distribuciones iniciales que lleven a resultados útiles, en vez de desear cuantificar con precisión cualquier

información real subjetiva. Viendo el modelo principalmente como un instrumento predictivo, el elemento principal al especificar una distribución inicial es la variable respuesta.

Supóngase que, a partir de cierto conocimiento previo, se tiene una creencia acerca del valor del vector respuesta $Y_{n\times 1}$, observado con la matriz de variables explicativas X, y se denota tal creencia por η_0. Este será un vector fijo a pesar del modelo considerado. Puede especificarse de diferentes maneras. Por ejemplo, en Regresión Lineal, si en análisis previos se ha utilizado en particular algún submodelo m del modelo (3.1), con matriz X_m y se estimó al vector de coeficientes como $\tilde{\beta}^{(m)}$, entonces, se escoge a η_0 como $\tilde{Y}$, es decir, $X_m\tilde{\beta}^{(m)}$.

Bajo el modelo m con matriz de variables independientes X_m, se recomienda que la media de la distribución inicial de $\beta^{(m)} \mid \tau$ sea

$$\mu^{(m)} = (X_m'X_m)^{-1}X_m'\eta_0 \qquad (3.9)$$

que, claramente, es la solución por mínimos cuadrados a las ecuaciones normales escritas con la matriz de diseño del modelo considerado y la creencia inicial para Y. Si X_m es de rango incompleto, entonces $\mu^{(m)}$ es la proyección ortogonal de la solución por mínimos cuadrados sobre el espacio generado por las columnas de $X_m'X_m$.

Como siguiente paso, se escoge a la matriz de precisión de $\beta^{(m)} \mid \tau$ de la forma τT_m, donde

$$T_m = c(X_m'X_m) \qquad (3.10)$$

con $c \geq 0$, que cuantifica, en términos del presente experimento, la importancia que se le quiere dar a la creencia inicial η_0. Así, bajo el modelo m, T_m es un

múltiplo de la matriz de información de Fisher para $\beta^{(m)}$ y tiene la ventaja de que lleva a soluciones analíticamente tratables y computacionalmente factibles.

Ahora, se toma a $\beta^{(m)} \mid \tau$ como una distribución normal, es decir,

$$\beta^{(m)} \mid \tau \sim N_{k_m}\left(\mu^{(m)}, \tau T_m\right) \quad , \qquad (3.11)$$

Como resultado de enfocarse en las observaciones, sólo fueron necesarias pocas cantidades fácilmente interpretables para especificar la distribución inicial. En particular, la predicción η_0 se utiliza para determinar una distribución inicial de referencia para $\beta^{(m)} \mid \tau$ para cada m.

Finalmente, la distribución inicial para τ se toma como una gamma con parámetros $(\delta_0/2, \gamma_0/2)$, es decir, con densidad

$$\pi(\tau) \propto \tau^{\frac{\delta_0}{2}-1} \exp(-\gamma_0\tau/2) \qquad (3.12)$$

Se eligen las densidades normal y gamma para que la distribución inicial conjunta para $\beta^{(m)}$ y τ sea una normal-gamma, que es un miembro de una familia conjugada y, por lo tanto, la distribución conjunta final sea también normal-gamma, para un modelo fijo m. Así,

$$\begin{aligned}\pi\left(\beta^{(m)}, \tau\right) &= p\left(\beta^{(m)} \mid \tau\right) \cdot \pi(\tau) \\ &\propto \tau^{\frac{k_m+\delta_0}{2}-1} \exp\left(-\frac{\tau}{2}\left(\gamma_0 + \left(\beta^{(m)} - \mu^{(m)}\right)' T_m \left(\beta^{(m)} - \mu^{(m)}\right)\right)\right)\end{aligned}$$

Con esta distribución inicial y la verosimilitud (3.5), la distribución final para los parámetros es

$$p\left(\beta^{(m)}, \tau \mid X_m, y\right) = p\left(y \mid \beta^{(m)}, \tau\right) \cdot \pi\left(\beta^{(m)}, \tau\right)$$
$$\propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2} D\right) \cdot \tau^{\frac{k_m+\delta_0}{2}-1} \exp\left(-\frac{\tau}{2}(\gamma_0 + E)\right)$$
$$= \tau^{\frac{n+k_m+\delta_0}{2}-1} \exp\left(-\frac{\tau}{2}(\gamma_0 + D + E)\right) \tag{3.13}$$

con

$$D = \left(y - X_m \beta^{(m)}\right)'\left(y - X_m \beta^{(m)}\right)$$
$$E = \left(\beta^{(m)} - \mu^{(m)}\right)' T_m \left(\beta^{(m)} - \mu^{(m)}\right)$$

Con esta distribución y la (3.6) se llega a la DPER, que será

$$p(z \mid X_m, y) = \int_0^\infty \int_{\Re^{k_m}} p\left(z \mid X_m, \beta^{(m)}, \tau\right) p\left(\beta^{(m)}, \tau \mid X_m, m\right) d\beta^{(m)} d\tau$$
$$\propto \int_0^\infty \int_{\Re^{k_m}} \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2} C\right) \cdot \tau^{\frac{n+k_m+\delta_0}{2}-1} \exp\left(-\frac{\tau}{2}(\gamma_0 + D + E)\right) d\beta^{(m)} d\tau$$
$$= \int_0^\infty \int_{\Re^{k_m}} \tau^{\frac{2n+k_m+\delta_0}{2}-1} \exp\left(-\frac{\tau}{2}(\gamma_0 + C + D + E)\right) d\beta^{(m)} d\tau$$

con C, D y E como se definieron anteriormente.

Ahora, desarrollando $C + D + E$ para completar cuadrados en $\beta^{(m)}$, queda

$$C + D + E = y'y + Z'Z - \mu^* + c\eta_0' P_m \eta_0 + (c+2)\left(\beta^{(m)} - \mu^*\right)'\left(X_m' X_m\right)\left(\beta^{(m)} - \mu^*\right)$$

con

$$\mu^* = \tfrac{1}{2+c}(X_m'Z + X_m'y + cX_m'\eta_0)'(X_m'X_m)^{-1}(X_m'Z + X_m'y + cX_m'\eta_0)$$

Así

$$p(z \mid X_m, y) = \int_0^\infty \tau^{\frac{2n+\delta_0}{2}-1} \exp\left(-\frac{\tau}{2}E^*\right)\left[\int_{\mathfrak{R}^{k_m}} \tau^{\frac{k_m}{2}} \exp\left(-\frac{\tau}{2}F^*\right) d\beta^{(m)}\right] d\tau$$

con

$$E^* = \gamma_0 + y'y + Z'Z - \mu^* + c\eta_0'P_m\eta_0$$
$$F^* = (c+2)\left(\beta^{(m)} - \mu^*\right)'(X_m'X_m)\left(\beta^{(m)} - \mu^*\right)$$

Después de integrar, la distribución predictiva final de Z es

$$p(z \mid X_m, y) \propto \left[\gamma_0 + y'y + Z'Z - \mu^* + c\eta_0'P_m\eta_0\right]^{-\frac{2n-\delta_0}{2}}$$
$$= \left[(Z-\eta_m)'(I - \tfrac{1}{2+c}P_m)(Z - P_m y) + q_m + \gamma p_m + \gamma_0\right]^{-\frac{2n-\delta_0}{2}}$$

que puede expresarse como

$$Z \sim S_n\left(n+\delta_0,\ \eta_m,\ s_m^2\{I + (1-\gamma)P_m\}\right), \qquad (3.14)$$

donde $\gamma = c/(1+c)$, $\eta_m = P_m\{\gamma\eta_0 + (1-\gamma)y\}$, $s_m^2 = (n+\delta_0)^{-1}(q_m + \gamma p_m + \gamma_0)$, $q_m = y'(I - P_m)y$ y $p_m = (y-\eta_0)'P_m(y-\eta_0)$.

La DPER en la distribución (3.8) para distribuciones iniciales no informativas puede obtenerse a partir de la distribución (3.14) haciendo $\gamma = 0$, $\delta_0 = -k_m$ y $\gamma_0 = 0$. Más aún, si X_m tiene rango $r_m < k_m$, se reemplaza k_m por r_m en la distribución (3.14).

4 CRITERIOS

Como ya se mencionó anteriormente, y es el vector de los valores observados y Z será el vector de las predicciones obtenidas. Es claro que los buenos modelos serán aquellos que produzcan predicciones cercanas a los valores observados en un experimento idéntico. Los criterios siguientes se definen con esta motivación.

4.1 CRITERIO L_2

Para un modelo m dado, considérese la cantidad

$$L_m^2 = E\left\{(Z-y)'(Z-y)\right\}, \tag{4.1}$$

donde la esperanza se toma con respecto a la DPER f_m, es decir, con respecto a Z. Desarrollando L_m^2:

$$\begin{aligned} L_m^2 &= E\left\{(Z-y)'(Z-y)\right\} \\ &= E\left\{\sum_1^n (Z_i - y_i)^2\right\} \\ &= \sum_{i=1}^n E\left\{Z_i^2 - 2Z_i y_i + y_i^2\right\} \\ &= \sum_{i=1}^n \left\{E\left(Z_i^2\right) - 2y_i E(Z_i) + y_i^2 + E^2(Z_i) - E^2(Z_i)\right\} \\ &= \sum_{i=1}^n \left\{[E(Z_i) - y_i]^2 + Var(Z_i)\right\} \end{aligned}$$

Entonces, L_m^2 se expresa como la suma de dos componentes, una que involucra las medias de la distribución predictiva y otra que involucra las varianzas. Así, el desempeño de un modelo se mide por la distancia que existe entre las predicciones y los datos observados, más la varianza de las predicciones. Claramente, los buenos modelos tendrán valores pequeños de L_m^2.

Es más conveniente utilizar la medida

$$L_m = \sqrt{L_m^2}$$

ya que es la distancia medida en las mismas unidades que la variable respuesta. A L_m^2 se le conoce como *criterio L*.

Para obtener este criterio se necesitarán los siguientes resultados (Seber, 1977):

Teorema H.- Sea $X_{n\times 1}$ una matriz aleatoria y $A_{n\times n}$ una matriz de constantes, $E(X)=\theta$ y $D(X)=\Sigma$, con $D(X)$ la matriz de dispersión. Entonces

$$E(X'AX) = tr(A\Sigma) + \theta'A\theta$$

Corolario 1.- $E\left[(X-b)'A(X-b)\right] = tr(A\Sigma) + (\theta-b)'A(\theta-b)$

4.1.1 Criterio L para el caso en el que se tiene una distribución inicial informativa.

En este caso, como se vio en el capítulo anterior, se llega a la distribución predictiva (3.14)

$$Z \sim S_n\left(n+\delta_0,\ \eta_m,\ s_m^2\{I+(1-\gamma)P_m\}\right),$$

que tiene $E(Z)=\eta_m$ y $D(Z)=\frac{n-\delta_0}{n-\delta_0-2}s_m^2\{I+(1-\gamma)P_m\}$.

Para obtener la expresión del criterio L, se desarrolla (4.1) y se utiliza el Corolario 1.

$$\begin{aligned}
L_m^2 &= E\left\{(Z-y)'(Z-y)\right\} \\
&= tr\left(\frac{n-\delta_0}{n-\delta_0-2}s_m^2\{I+(1-\gamma)P_m\}\right)+(\eta_m-y)'(\eta_m-y) \\
&= \frac{n-\delta_0}{n-\delta_0-2}s_m^2(n+(1-\gamma)k_m)+\gamma^2(y-\eta_0)'P_m(y-\eta_0)+y'(I-P_m)y \\
&= \frac{n+(1-\gamma)k_m}{n+\delta_0-2}(q_m+\gamma p_m+\gamma_0)+\gamma^2 p_m+q_m \\
&= (\lambda_m+1)q_m+\gamma(\lambda_m+\gamma)p_m+\gamma_0\lambda_m
\end{aligned}$$

Entonces,

$$L_m=\{(\lambda_m+1)q_m+\gamma(\lambda_m+\gamma)p_m+\gamma_0\lambda_m\}^{\frac{1}{2}}$$

con $\lambda_m=\frac{n+(1-\gamma)k_m}{n+\delta_0-2}$, $\gamma=\frac{c}{(1+c)}$, $q_m=y'(I-P_m)y$ y $p_m=(y-\eta_0)'P_m(y-\eta_0)$.

L_m^2 es una función lineal de q_m y p_m. La cantidad q_m es la longitud al cuadrado de la proyección de los datos sobre el espacio de errores del modelo m, es decir, la suma de cuadrados del error del modelo m. La cantidad p_m representa una penalización por una mala creencia del vector Y; es la longitud al cuadrado de la proyección del "error de creencia" sobre el espacio columna del modelo.

4.1.2 Criterio L para el caso en el que se tiene una distribución inicial no informativa.

En este caso, como se vio en el capítulo anterior, se llega a la distribución predictiva (3.8)

$$Z \sim S_n\left(n-k_m,\ P_m y,\ s_m^2(I+P_m)\right)$$

que tiene $E(Z)=P_m y$ y $D(Z)=\dfrac{n-k_m}{n-k_m-2}s_m^2(I+P_m)$.

Para obtener la expresión del criterio L, se desarrolla (4.1) y se utiliza el Corolario 1.

$$\begin{aligned}
L_m^2 &= E\left\{(Z-y)'(Z-y)\right\} \\
&= tr\left(\frac{n-k_m}{n-k_m-2}s_m^2(I+P_m)\right)+(P_m-y)'(P_m-y) \\
&= \frac{n-k_m}{n-k_m-2}s_m^2(n+k_m)+y'(I-P_m)y \\
&= q_m\left(\frac{n+k_m}{n-k_m-2}+1\right) \\
&= \frac{2(n-1)}{n-k_m-2}q_m
\end{aligned}$$

Entonces,

$$L_m=\left\{2(n-1)(n-k_m-2)^{-1}q_m\right\}^{\frac{1}{2}}$$

4.2 Criterio M

Para definir el segundo criterio, considérese la cantidad

$$M_m^* = f_m(y)$$

que es la DPER bajo el modelo m evaluada en la respuesta observada y. Un buen modelo tendrá un valor grande de M_m^*. Otra vez, para facilitar la interpretación, se utilizará la medida

$$M_m = \left(M_m^*\right)^{-\frac{1}{n}}$$

que está en las mismas unidades que la variable respuesta. Los valores pequeños de esta medida indican buenos modelos. A M_m se le conoce como *criterio* M.

4.2.1 Criterio M para el caso en el que se tiene una distribución inicial informativa.

Utilizando la distribución (3.14) se obtiene que

$$M_m^* = f_m(y) = \frac{\Gamma\{(n+\delta_0)/2\}}{\Gamma\{n+\delta_0/2\}}(n+\delta_0)^{-\frac{n}{2}}\pi^{-\frac{n}{2}}\left|\frac{(n+\delta_0)}{a_m}\left[I-\tfrac{1}{2+c}P_m\right]\right|^{\frac{1}{2}}(G)^{-\frac{2n+\delta_0}{2}}$$

con

$$G=\left[(Z-\eta_m)'\left(I-\tfrac{1}{2+c}P_m\right)(Z-P_m y)+(q_m+\gamma p_m+\gamma_0)\right]^{-\frac{2n-\delta_0}{2}}$$

Desarrollando G y el determinante queda que

$$M_m^* = f_m(y) = \frac{\Gamma\{(n+\delta_0)/2\}}{\Gamma\{n+\delta_0/2\}} \pi^{-\frac{n}{2}} a_m^{-\frac{n}{2}} (2-\gamma)^{-k_m} \left[1+\frac{b_m}{a_m}\right]^{-\frac{2n+\delta_0}{2}}$$

donde

$$a_m = q_m + \gamma p_m + \gamma_0$$
$$b_m = q_m + \frac{\gamma^2}{2-\gamma} p_m$$

que también son combinaciones lineales de la suma de cuadrados de los residuos y el error de creencia.

Entonces

$$M_m = \left(M_m^*\right)^{-\frac{1}{n}} = \pi^{\frac{1}{2}} \left(\frac{\Gamma\{(n+\delta_0)/2\}}{\Gamma\{n+\delta_0/2\}} (2-\gamma)^{\frac{k_m}{2}} \right)^{\frac{1}{n}} a_m^{\frac{1}{2}} \left(1+\frac{b_m}{a_m}\right)^{1+\frac{\delta_0}{2n}} \qquad (4.2)$$

4.2.2 Criterio M para el caso en el que se tiene una distribución inicial no informativa.

La expresión para este criterio bajo una distribución inicial no informativa es similar a la de la distribución inicial normal-gamma, obtenida en la sección anterior, pero utilizando la distribución (3.8). También puede obtenerse haciendo $\gamma = \gamma_0 = 0$ y $\delta_0 = -k_m$ en (4.2), quedando así

$$M_m = 2\pi^{\frac{1}{2}} \left(\frac{\Gamma\{(n-k_m)/2\}}{\Gamma\{n+\delta_0/2\}} \right)^{\frac{1}{n}} q_m^{\frac{1}{2}}$$

4.3 Criterio K

El tercer criterio es la divergencia Kullback-Leibler (*KL*) entre dos densidades predictivas. Como ya se mencionó en el capítulo 2, por medio de la teoría de información se puede dar una medida que cuantifique la información contenida en una función $f(x)$. Tal medida es

$$\log \frac{1}{f(x)}$$

Supóngase que $f_1(x)$ y $f_2(x)$ son dos densidades predictivas, entonces la diferencia

$$\log \frac{1}{f_2(x)} - \log \frac{1}{f_1(x)} = \log \frac{f_1(x)}{f_2(x)}$$

nos dice qué tanta información se pierde al cambiar de una función a otra o la "distancia" que existe entre las dos. Ahora, se define a la divergencia *KL* entre f_1 y f_2 por

$$K(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx$$

En general, $K(f_1, f_2) \neq K(f_2, f_1)$ y $K(a, b) \geq 0$, dándose la igualdad sólo si $a = b$. Esta divergencia se ha utilizado para una gran variedad de problemas estadísticos y en particular con la distribución predictiva en estadística Bayesiana.

Para la selección de modelos, supóngase que m_0 es un modelo fijo en M a partir del cuál se van a comparar otros modelos. Una selección natural de m_0 sería el modelo completo (3.1), que contiene las k variables independientes. Utilizando las DPER de m_0 y m, se define

$$K_m = K(m_0, m) + K(m, m_0)$$

Los criterios L_m y M_m miden qué tan cerca está el vector de datos de la DPER f_m para cada m, mientras que K_m mide qué tan lejos están las DPER de los dos modelos. Si K_m es pequeño, entonces m y m_0 proporcionan casi la misma información. Este criterio se utiliza cuando se quieren comparar dos modelos específicos; sin embargo, también puede utilizarse para comparar todos los posibles modelos fijando m_0 y calculando K_m para todo $m \in M$. En el problema de selección de variables, al tomar al modelo completo como m_0, K_m puede interpretarse como la cantidad de información que se pierde al quitar algunas variables del modelo m_0 para obtener m. Así, los valores pequeños de K_m indican que el modelo m es casi tan bueno como el completo, lo que implica que las variables que fueron removidas no eran muy importantes. También se puede elegir a m_0 como el modelo que contiene sólo a la intersección. En este caso, K_m se interpretaría como la información que se gana al ir agregando variables al modelo que tiene sólo la intersección. Aquí se elegirá al modelo m con el que se obtenga el valor más grande de K_m. Se puede notar que $K_{m_0} = 0$, mientras que los criterios L y M no tienen un valor mínimo fijo. Otra diferencia es que los valores deseables de K_m dependen de qué modelo se escoja como m_0. En este sentido, la selección de modelos basada en K_m no es tan sencilla como con L_m y M_m. A K_m se le llama *criterio K*.

4.3.1 Criterio K para el caso en el que se tiene una distribución inicial informativa.

No se dispone de una expresión exacta para este criterio, ya que la integral necesaria no es tratable. Sin embargo, para n lo suficientemente grande, se puede aproximar la distribución (3.14) por una distribución

$$N_n\left(\eta_m, \left(\frac{n+\delta_0}{n+\delta_o-2}\right)s_m^2\{I+(1-\gamma)P_m\}\right)$$

donde $N_n(\mu, \Sigma)$ denota una distribución normal multivariada de dimensión n, con vector de medias μ y matriz de varianzas-covarianzas Σ.

Tomando m_0 como el modelo completo, se define

$$\upsilon = \frac{(n+\delta_m)(n+\delta_{m_0}-2)}{(n+\delta_{m_0})(n+\delta_m-2)}$$

donde $\delta_m = \delta_{m_0} = \delta_0$, ya que independientemente del modelo, la distribución inicial para τ es una gamma con parámetros $(\delta_0/2, \gamma_0/2)$. Con

$$\zeta = \frac{n+\delta_{m_0}-2}{2s_{m_0}^2(n+\delta_{m_0})(2-\gamma)} + \frac{n+\delta_m-2}{2s_m^2(n+\delta_m)}$$

y

$$\eta_{m,m_o} = (P_{m_0}-P_m)\{\gamma\eta_0 + (1-\gamma)y\}$$

se obtiene

$$K_m \approx \zeta\eta'_{m,m_0}\eta_{m,m_0} + \frac{n}{2}\left(\frac{\upsilon s_m^2}{s_{m_0}^2} + \frac{s_{m_0}^2}{\upsilon s_m^2} - 2\right) + \frac{k_{m_0}-k_m}{2}\left\{\frac{(1-\gamma)s_{m_0}^2}{\upsilon s_m^2} - \frac{(1-\gamma)\upsilon s_m^2}{(2-\gamma)s_{m_0}^2}\right\}$$

Se sigue una expresión similar si se toma m_0 como el modelo con sólo la intersección.

4.3.2 Criterio K para el caso en el que se tiene una distribución inicial no informativa.

La expresión es similar a la del caso anterior, pero haciendo $\delta_0 = \delta_m = -k_m$ y $\gamma = \gamma_0 = 0$. Tomando m_0 como el modelo completo, queda

$$\upsilon = \frac{(n-k_m)(n-k_{m_0}-2)}{(n-k_{m_0})(n-k_m-2)},$$

$$\zeta = \frac{n-k_{m_0}-2}{4s_{m_0}^2(n-k_{m_0})} + \frac{n-k_m-2}{2s_m^2(n-k_m)}$$

con $s_m^2 = (n-k_m)^{-1}q_m$ y $\eta_{m,m_0} = (P_{m_0} - P_m)y$, obteniendo así

$$K_m \approx \zeta \eta'_{m,m_0}\eta_{m,m_0} + \frac{n}{2}\left(\frac{\upsilon s_m^2}{s_{m_0}^2} + \frac{s_{m_0}^2}{\upsilon s_m^2} - 2\right) + \frac{k_{m_0}-k_m}{2}\left\{\frac{s_{m_0}^2}{\upsilon s_m^2} - \frac{\upsilon s_m^2}{2s_{m_0}^2}\right\}$$

Se sigue una expresión similar si se toma m_0 como el modelo con sólo la intersección.

4.4 NÚMEROS DE CALIBRACIÓN

Aunque muchos métodos de selección de modelos basados en criterios no cuantifican la incertidumbre inherente en los valores del criterio, es deseable hacerlo. Utilizando el modelo m^*, que tiene el mínimo valor del criterio, se puede calcular la desviación estándar del mismo. Para esto último se puede ver al criterio como una función del vector de observaciones Y, con respecto a su distribución marginal. En particular, para el criterio L se calcula

$$S_L = \left[Var\left\{L_{m^*}(Y)\right\}\right]^{\frac{1}{2}}$$

A S_L se le llama *número de calibración* para el criterio L. De igual manera se calcula S_M, el número de calibración para el criterio M.

El número de calibración es importante en la selección de modelos, ya que con él se determina un intervalo, dentro del cual se encuentran valores del criterio asociados a modelos que también se consideran buenos; pudiéndose así escoger un modelo más parsimonioso. Por ejemplo, para el criterio L el intervalo es

$$\left(L_{m^*},\ L_{m^*} + S_L\right)$$

A continuación se darán las expresiones para calcular los números de calibración para los criterios L y M. En el capítulo 5 se ilustrará su uso.

4.4.1 Números de calibración para el criterio L.

Distribución inicial normal-gamma.

Para calcular el número de calibración S_L, se generan l muestras de la distribución marginal

$$Y \sim S_n\left(\delta_0,\ \eta_{m^*},\ \gamma_0\delta_0^{-1}\left\{I + \gamma^{-1}(1-\gamma)P_{m^*}\right\}\right)$$

y se calcula L_{m^*} con cada muestra, con m^* el modelo que minimiza L_m. La desviación estándar de esos valores dan una aproximación Monte Carlo a S_L. Para que tal aproximación sea buena el número l debe de ser grande.

Distribución inicial no informativa de Jeffreys.

En este caso, la distribución marginal de Y es impropia, pero se pueden obtener muestras de la distribución condicional $Y \mid \tau \sim N_n\left(0, \left(\tau\left(I - P_{m^*}\right)\right)^{-1}\right)$, con τ reemplazada por $\tilde{\tau}$, la moda de la distribución de marginal final de τ utilizando m^*. El problema con esta distribución es que la matriz de varianzas-covarianzas es singular, así que no se pueden generar buenas muestras. Por lo tanto, se busca una aproximación al número de calibración.

La desviación estándar de las muestras de L_{m^*} resultantes puede verse como una aproximación a $D_L = \left[Var\left(L_{m^*} \mid \tau = \tilde{\tau}\right)\right]^{\frac{1}{2}}$. Para n grande puede obtenerse la aproximación analítica

$$D_L = \frac{\tilde{\tau}^{-\frac{1}{2}}}{2}\left(1 - \frac{k_{m^*}}{n}\right)^{\frac{1}{2}}\left[1 - \frac{1}{n}\left\{1 + \frac{1}{32}\left(1 - \frac{k_{m^*}}{n}\right)\left(1 - \frac{2}{n}\right)\right\}\right]^{\frac{1}{2}} \qquad (4.3)$$

que se obtiene calculando $E\left[L_{m^*}^2 \mid \tau\right]$ y utilizando la aproximación por series de Taylor para $E\left[L_{m^*} \mid \tau\right]$.

Como ya se ha visto, si la distribución inicial conjunta de $\beta^{(m)}$ y τ es la de Jeffreys, entonces, la distribución final para $\beta^{(m)}$ y τ para el modelo m^* será

$$p\left(\beta^{(m^*)}, \tau \mid X_{m^*}, y\right) \propto \tau^{\frac{n}{2}-1} \exp\left(-\frac{\tau}{2}\left(Y - X_{m^*}\beta^{(m^*)}\right)'\left(Y - X_{m^*}\beta^{(m^*)}\right)\right)$$

$$= \tau^{\frac{n-k_{m^*}}{2}-1} \exp\left\{-\frac{\tau}{2} q_{m^*}\right\} \tau^{\frac{k_{m^*}}{2}} \exp\left\{-\frac{\tau}{2} B^*\right\}$$

que es una normal-gamma con

$$B^{\cdot} = \left[\beta^{(m^{\cdot})} - (X'_{m^{\cdot}} X_{m^{\cdot}})^{-1} X'_{m^{\cdot}} y\right]' (X'_{m^{\cdot}} X_{m^{\cdot}}) \left[\beta^{(m^{\cdot})} - (X'_{m^{\cdot}} X_{m^{\cdot}})^{-1} X'_{m^{\cdot}} y\right]$$

La densidad marginal final de τ para el modelo $m^{\cdot}$ es

$$\begin{aligned} p(\tau \mid X_{m^{\cdot}}, y) &= \int_{\Re^{k_{m^{\cdot}}}} p\left(\beta^{(m^{\cdot})}, \tau \mid X_{m^{\cdot}}, y\right) d\beta^{(m^{\cdot})} \\ &\propto \tau^{\frac{n-k_{m^{\cdot}}}{2}-1} \exp\left\{-\frac{\tau}{2} q_{m^{\cdot}}\right\} \int_{\Re^{k_{m^{\cdot}}}} \tau^{\frac{k_{m^{\cdot}}}{2}} \exp\left\{-\frac{\tau}{2} B^{\cdot}\right\} d\beta^{(m^{\cdot})} \\ &\propto \tau^{\frac{n-k_{m^{\cdot}}}{2}-1} \exp\left\{-\frac{\tau}{2} q_{m^{\cdot}}\right\} \end{aligned}$$

que es una distribución gamma con parámetros $\frac{n-k_{m^{\cdot}}}{2}$ y $\frac{q_{m^{\cdot}}}{2}$.

La moda de una distribución Gamma(a,b) es $\frac{a-1}{b}$, entonces, la moda para la distribución marginal final de τ para el modelo $m^{\cdot}$ es

$$\tilde{\tau} = \frac{n - k_{m^{\cdot}} - 2}{q_{m^{\cdot}}}$$

4.4.2 Número de calibración para el criterio *M*.

Distribución inicial normal-gamma.

Para calcular el número de calibración S_M, se generan l muestras de la distribución marginal

$$Y \sim S_n\left(\delta_0, \eta_{m^{\cdot}}, \gamma_0 \delta_0^{-1}\left\{I + \gamma^{-1}(1-\gamma) P_{m^{\cdot}}\right\}\right)$$

y se calcula M_{m^*} con cada muestra, con m^* el modelo que minimiza M_m. La desviación estándar de esos valores dan una aproximación Monte Carlo a S_M. Para que tal aproximación sea buena el número l debe de ser grande.

Distribución inicial no informativa de Jeffreys.

Análogamente a la aproximación (4.3), para el criterio M se tiene

$$D_M = 2^{\frac{k_{m^*}}{2n}}(2\pi)^{\frac{1}{2}}\tau^{-\frac{1}{2}}\left\{\left(\frac{n}{n-2}\right)^{\frac{n-k_{m^*}}{2}} - \left(\frac{n}{n-1}\right)^{n-k_{m^*}}\right\}^{\frac{1}{2}}$$

4.5 Otro Criterio

Desde el punto de vista Bayesiano y tomando en cuenta el modelo de regresión lineal desarrollado en el capítulo 3, se pueden calcular las probabilidades inicial y final para cada modelo m, para así, escoger al más probable dados los datos. Para esto, se necesita especificar, para cada modelo, su probabilidad inicial, una distribución inicial conjunta para los parámetros y calcular la distribución final para cada uno, dados los datos. La distribución inicial debe de especificar un vector sobre M de tamaño 2^k, que dé la probabilidad inicial para cada modelo y, dado el modelo m, una distribución inicial para $\left(\beta^{(m)}, \tau\right)$. A continuación se desarrollarán tales probabilidades, en base a los datos observados y en algunos parámetros iniciales que deben ser especificados.

4.5.1 Distribución inicial en el espacio de modelos.

Ahora, se recordarán algunos conceptos ya mencionados en la sección 3.2.3 y que serán de utilidad en la presente sección.

A partir de cierto conocimiento previo, se puede tener una creencia acerca del valor del vector respuesta $Y_{n\times 1}$, observado con la matriz de variables independientes X, que se denotará en este caso por η. Este será un vector fijo a pesar del modelo considerado. Una vez más, se dará mayor importancia a las observaciones que a los parámetros.

Empleando la familia conjugada normal-gamma, se especifica la distribución inicial para $\left(\beta^{(m)}, \tau\right)$, para cada modelo $m \in M$, utilizando η y un escalar positivo c, que cuantifica la importancia que se le da a la predicción inicial η en relación con la información proporcionada por los datos; obteniendo así

$$\beta^{(m)} \mid \eta, \tau \sim N_{k_m}\left(\mu^{(m)}, \tau T_m\right) \tag{4.4}$$

con

$$\mu^{(m)} = \left(X_m' X_m\right)^{-1} X_m' \eta_0 \tag{4.5}$$

$$T_m = c\left(X_m' X_m\right) \tag{4.6}$$

y

$$\pi(\tau) \propto \tau^{\frac{\delta}{2}-1} \exp(-\lambda\tau/2) \tag{4.7}$$

donde δ y λ son parámetros iniciales adicionales.

Si se quiere escribir la distribución para el vector de datos observados Y sin tomar en cuenta a los modelos de M, se puede considerar a las ecuaciones (3.1) y (3.2) sin relación con la regresión, reemplazando (3.1) por $Y = \mu + \varepsilon$. Utilizando la distribución inicial $\mu \mid \tau, \eta \sim N_n(\eta, c\tau I)$, que se obtiene de la misma manera que (4.4), pero considerando $X_m = I$, se deduce que

$$E(Y) = E(\mu) + E(\varepsilon) = \eta$$
$$Var(Y) = Var(\mu) + Var(\varepsilon) = \left(\frac{1}{c\tau} + \frac{1}{\tau}\right) I = \frac{1}{\tau}\left(\frac{c+1}{c}\right) I$$

entonces

$$Y \mid \tau, \eta \sim N_n\left(\eta, (\gamma\tau)^{-1} I\right) \tag{4.8}$$

con $\gamma = c/(1+c)$. Por otro lado, vista a través de un modelo m y utilizando la distribución inicial (4.4), se obtiene que

$$Y \mid \tau, \eta \sim N_n\left(X_m \mu^{(m)}, \tfrac{1}{\tau}\left(I + \tfrac{\gamma}{1-\gamma} P_m\right)\right) \tag{4.9}$$

ya que

$$\begin{aligned} E(Y) &= E\left(X_m \beta^{(m)}\right) + E(\varepsilon) = X_m \mu^{(m)} \\ Var(Y) &= Var\left(X_m \beta^{(m)}\right) + Var(\varepsilon) \\ &= X_m (\tau T_m)^{-1} X_m' + (\tau)^{-1} I \\ &= \frac{1}{\tau}\left(I + \frac{\gamma}{1-\gamma} P_m\right) \end{aligned}$$

Imagínese ahora que antes de obtener y se realizó una repetición del experimento, que tuvo como resultado la respuesta Y_0. Considérese, además, que se comenzó con una distribución uniforme en M, es decir, con $p_0(m) = 2^{-k}$ para toda $m \in M$, y que la distribución inicial de $\left(\beta^{(m)}, \tau\right)$ se tomó como

$$\pi_0\left(\beta^{(m)},\tau\right) \propto \left|X_m' X_m\right|^{\frac{1}{2}} e^{-\frac{k_m}{2}} (2\pi)^{-\frac{k_m}{2}} \tau^{\frac{k_m}{2}-1} \tag{4.10}$$

Utilizando la verosimilitud (3.5), con Y_0 en lugar de y, se tiene que

$$\begin{aligned} p(Y_0 \mid m) &= \int_0^\infty \int_{\Re^{k_m}} p\left(Y_0 \mid \beta^{(m)},\tau\right)\cdot \pi_0\left(\beta^{(m)},\tau\right) d\beta^{(m)} d\tau \\ &\propto \left|X_m' X_m\right|^{\frac{1}{2}} e^{-\frac{k_m}{2}} (2\pi)^{-\frac{k_m}{2}} \int_0^\infty \int_{\Re^{k_m}} \tau^{\frac{n+k_m}{2}-1} \exp\left(-\frac{\tau}{2} D^{**}\right) d\beta^{(m)} d\tau \\ &= e^{-\frac{k_m}{2}} \int_0^\infty \tau^{\frac{n}{2}-1} \exp\left(-\frac{\tau}{2} A^{**}\right) \left[\int_{\Re^{k_m}} \left|X_m' X_m\right|^{\frac{1}{2}} (2\pi)^{-\frac{k_m}{2}} \tau^{\frac{k_m}{2}} \exp\left(-\frac{\tau}{2} B^{**}\right) d\beta^{(m)}\right] d\tau \\ &\propto e^{-\frac{k_m}{2}} \left[Y_0'(I-P_m)Y_0\right]^{-\frac{n}{2}} \end{aligned} \tag{4.11}$$

donde

$$\begin{aligned} D^{**} &= \left(Y_0 - X_m \beta^{(m)}\right)'\left(Y_0 - X_m \beta^{(m)}\right) \\ B^{**} &= \left[\beta^{(m)} - (X_m' X_m)^{-1} X_m' Y_0\right]'(X_m' X_m)\left[\beta^{(m)} - (X_m' X_m)^{-1} X_m' Y_0\right] \\ A^{**} &= Y_0'(I - P_m) Y_0 \end{aligned}$$

Utilizando el Teorema de Bayes y la ecuación (4.11) se llega a que

$$\begin{aligned} p(m \mid Y_0) &= \frac{p(Y_0 \mid m)\, p_0(m)}{\sum_{m \in M} p(Y_0 \mid m)\, p_0(m)} \\ &= \frac{\left[Y_0'(I-P_m)Y_0\right]^{-\frac{n}{2}} e^{-\frac{k_m}{2}}}{\sum_{m \in M} \left[Y_0'(I-P_m)Y_0\right]^{-\frac{n}{2}} e^{-\frac{k_m}{2}}} \end{aligned} \tag{4.12}$$

Como el vector Y_0 realmente no fue observado, las probabilidades anteriores se ven como cantidades aleatorias que deben ser estimadas o

predichas. Una elección natural es promediar $p(m \mid Y_0)$ con respecto a la distribución de Y_0 para obtener las probabilidades deseadas $p(m)$; entonces, se calcula $p(m) = E[p(m \mid Y_0)]$ donde $Y_0 \sim N_n(\eta, (\gamma\tau)^{-1}I)$. Sin embargo, esta expresión no tiene una forma analítica, por lo que lo más conveniente es reemplazar $Y_0'(I - P_m)Y_0$ por su esperanza, que es

$$\begin{aligned} E(Y_0'(I-P_m)Y_0) &= tr\big((I-P_m)(\gamma\tau)^{-1}I\big) + \eta'(I-P_m)\eta \\ &= (\gamma\tau)^{-1}(n-k_m) + \eta'(I-P_m)\eta \\ &\approx \tau^{-1}(n-k_m) + \gamma\eta'(I-P_m)\eta \end{aligned}$$

Entonces

$$p(m \mid \tau) \approx \frac{\left[\gamma\,\eta'(I-P_m)\eta + \tau^{-1}(n-k_m)\right]^{-\frac{n}{2}} e^{-\frac{k_m}{2}}}{\sum_{m\in M}\left[\gamma\,\eta'(I-P_m)\eta + \tau^{-1}(n-k_m)\right]^{-\frac{n}{2}} e^{-\frac{k_m}{2}}} \qquad (4.13)$$

Finalmente, reemplazando τ por su moda $\lambda^{-1}(\delta-2)$ bajo la distribución inicial (4.7) y permitiendo que λ y γ dependan de m queda

$$p(m) = \frac{\left[\gamma_m\,\eta'(I-P_m)\eta + (\delta-2)^{-1}\lambda_m(n-k_m)\right]^{-\frac{n}{2}} e^{-\frac{k_m}{2}}}{\sum_{m\in M}\left[\gamma_m\,\eta'(I-P_m)\eta + (\delta-2)^{-1}\lambda_m(n-k_m)\right]^{-\frac{n}{2}} e^{-\frac{k_m}{2}}} \qquad (4.14)$$

Es conveniente elegir a λ_m y γ_m como

$$\lambda_m = l(n-k_m)^{-1}, \quad l > 0 \qquad (4.15)$$

y

$$\gamma_m = b\,\alpha^{\frac{1}{k_m}}, \quad 0 \le b,\ \alpha \le 1 \qquad (4.16)$$

Se observa que, cuando $\alpha = 0$, las probabilidades iniciales para cada k_m fija son iguales, esto es, se obtienen distribuciones uniformes para modelos con el mismo tamaño. Conforme $\alpha \to 1$, $p(m)$ puede ser dominada por $\eta'(I - P_m)\eta$, dependiendo de los valores de b, δ y l. En la práctica, se puede elegir $\eta \in C(X_{m^*})$, es decir, η puede ser generado por el espacio de las columnas de X_{m^*}, para alguna m^*, de acuerdo al contexto del experimento. Tal especificación da como resultado que $\eta'(I - P_m)\eta = 0$ cada vez que $\eta \in C(X_m)$. Esto significa que las probabilidades relacionadas con aquellos modelos cuyos espacios columna contienen a η dependen sólo de λ y δ. Si se elige a λ y δ como se mencionó anteriormente, se tienen las siguientes propiedades de $p(m)$ para tales modelos: (i) Todos los modelos con el mismo número de predictores tendrán la misma probabilidad inicial; (ii) Para dos modelos m y m', $k_{m'} > k_m$ implica que $p(m') < p(m)$, dando así mayor probabilidad a modelos más chicos. También se nota que con la elección anterior de λ y δ, tanto la media inicial como la varianza inicial de τ disminuyen conforme k_m aumenta. Así, los modelos más grandes llevan a precisiones iniciales esperadas más pequeñas. En general, tal elección de λ y δ favorece a modelos más pequeños cuando sus espacios columna contienen a η.

Ahora, dadas las probabilidades iniciales $p(m)$, y dados los datos y, las probabilidades finales se calculan de manera directa como

$$p(m \mid y) \propto p(y \mid m)\, p(m)$$

Utilizando las ecuaciones (4.4) – (4.7), y la verosimilitud (3.5), se tiene que

$$p(y \mid m) = \int_0^{\infty} \int_{\Re^{k_m}} p\left(y \mid \beta^{(m)}, \tau\right) \cdot \pi_0\left(\beta^{(m)}, \tau\right) d\beta^{(m)} d\tau$$
$$= \int_0^{\infty} \int_{\Re^{k_m}} (n-k_m)^{\frac{\delta}{2}} \left(\frac{1}{c}\right)^{\frac{k_m}{2}} \tau^{\frac{n+k_m+\delta}{2}-1} \exp\left(-\frac{\tau}{2}(\lambda + D + E)\right) d\beta^{(m)} d\tau$$

con

$$D = \left(y - X_m \beta^{(m)}\right)' \left(y - X_m \beta^{(m)}\right)$$
$$E = \left(\beta^{(m)} - \mu^{(m)}\right)' T_m \left(\beta^{(m)} - \mu^{(m)}\right)$$

Después de integrar se obtiene

$$p(y \mid m) \propto b^{\frac{k_m}{2}} (n-k_m)^{-\frac{\delta}{2}} \left[(y - P_m \eta)' (I - (1-\gamma) P_m)(y - P_m \eta) \right]^{-\frac{n+\delta}{2}}$$

Entonces

$$p(m \mid y) \propto p(m) * b^{\frac{k_m}{2}} (n-k_m)^{-\frac{\delta}{2}} *$$
$$\left[l(n-k_m)^{-1} + (y - P_m \eta)' (I - (1-\gamma) P_m)(y - P_m \eta) \right]^{-\frac{n+\delta}{2}} \qquad (4.17)$$

En el caso en que $\alpha = 0$, $b = 1$ la expresión queda libre de la predicción inicial η, reduciéndose a

$$p(m \mid y) \propto e^{-\frac{k_m}{2}} (n-k_m)^{-\frac{\delta}{2}} \left[l(n-k_m)^{-1} + y'(I - P_m) y \right]^{-\frac{n+\delta}{2}} \qquad (4.18)$$

Ahora, haciendo $l = \delta = 0$ queda

$$p(m \mid y) \propto e^{-\frac{k_m}{2}} [y'(I - P_m)y]^{-\frac{n}{2}} \tag{4.19}$$

Esta última expresión es (4.11) escrita con los datos observados y en lugar del vector imaginario Y_0. En otras palabras, al hacer $\alpha = l = \delta = 0$ y $b = 1$ se obtienen las probabilidades finales calculadas utilizando la distribución inicial (4.10) y una distribución uniforme sobre M.

5 APLICACIONES

5.1 SELECCIÓN DE VARIABLES

En muchas aplicaciones del análisis de regresión, el conjunto de variables que se incluyen en el modelo no está predeterminado; en esos casos, la primera parte del análisis es seleccionar esas variables. Hay ocasiones en las cuales existen consideraciones teóricas o de otro tipo que determinan las variables que se incluirán en la ecuación, sin embargo, en otras situaciones no existen tales consideraciones, así que el problema de selección de variables para la ecuación de regresión es muy importante.

Al formular un modelo de regresión surgen preguntas que deben de responderse, como: ¿Cuáles variables deben incluirse y en qué forma; es decir, deben de entrar a la ecuación en su forma original x ó como una transformación de ésta, por ejemplo x^2, $\log x$ ó una combinación de ambas? Aunque lo ideal es que ambos problemas se resuelvan de manera simultánea, por simplicidad se resuelven secuencialmente. Primero se determinan las variables que se incluirán en el modelo y después se investigará en qué forma entran a éste. De esta manera se simplifica el problema de selección de variables y se hace más tratable.

Hay dos puntos importantes que deben tomarse en cuenta en la selección de variables. Primero, no hay un "mejor conjunto" de variables para incluirse en la ecuación de regresión. Como ésta última puede utilizarse para varios propósitos, el conjunto de variables que puede ser el mejor para un propósito puede no serlo para otro; por ejemplo, un modelo que es bueno para realizar predicciones, puede no serlo para describir el fenómeno. Segundo, considerando lo anterior, puede haber varios subconjuntos de variables que sean adecuados y puedan utilizarse al formar la ecuación. Un buen procedimiento de selección de variables debe señalar cuáles

son los mejores modelos para así poder escoger entre ellos el mejor para cada propósito.

Para ejemplificar el uso y ventajas de los criterios tratados en la presente tesis, estos se aplicarán a un problema de selección de variables y se compararán con otros criterios, de los cuales, a continuación, se dará una breve descripción.

En lo siguente se considerará a $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$ como el modelo completo, y a $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ como el modelo reducido, con $p < k$.

Stepwise. Es un método secuencial en el cual, en cada paso, una variable es agregada o eliminada del modelo de regresión. Para esto se utiliza la estadística $F = \dfrac{(SEC_p - SEC)/(k-p)}{SEC/(n-k-1)}$; donde SEC_p es la suma de los cuadrados del error para el modelo reducido; SEC es la suma de los cuadrados del error para el modelo completo; p es el número de parámetros a estimar en el modelo considerado y n es el número de observaciones.

Al iniciar se fijan dos niveles de la estadística F, uno de entrada y uno de salida. Se comienza con el modelo que incluye solamente la ordenada al origen; en cada paso una variable, X_j, se introduce al modelo si el valor de F para el modelo considerado, agregando X_j, es mayor o igual a la F de entrada. Si existen varias variables que pueden entrar al modelo, se elige la que tiene el mayor valor de F. Si no se puede agregar ninguna variable de esta manera, entonces otra variable, X_i, es eliminada si el valor de F para el modelo de regresión considerado es menor o igual a la F de salida. Si existen varias variables que pueden salir del modelo, la elegida es la que tiene el menor valor de F. El procedimiento termina cuando ninguna variable puede ser agregada o eliminada del modelo.

Métodos Forward y Backward. Son variaciones del método Stepwise. En el primero, se comienza con el modelo que contiene solamente la ordenada al origen, y sólo se van agregando variables utilizando el criterio de la F de entrada. En el segundo, se comienza con el modelo completo y se van eliminando variables utilizando el criterio de la F de salida. Sin embargo, estos dos métodos no necesariamente llegan al mismo modelo.

C_p de Mallows. Si se utiliza a $\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$ para predecir a Y, lo que se busca es que la diferencia entre $E(\hat{Y}_p) - E(Y)$ sea mínima. Así, un criterio sugerido es minimizar la suma de cuadrados escalada

$$\Delta_p = \frac{1}{\sigma^2} E\left(\sum_{i=1}^{n} \left(\hat{Y}_{pi} - E(Y) \right)^2 \right) = p + \frac{SCS_p}{\sigma^2}$$

con SCS_p la suma de cuadrados del sesgo, dada por $SCS_p = \left(E(\hat{Y}_p) - E(Y)\right)' \left(E(\hat{Y}_p) - E(Y)\right)$. Un estimador insesgado de esta cantidad es la estadística C_p de Mallows, definida como

$$C_p = \frac{SEC_p}{\hat{\sigma}^2} + (2p - n),$$

con $\hat{\sigma}^2 = \frac{SEC}{n-p}$, así que cuando el sesgo es insignificante, el valor esperado de C_p es, aproximadamente, p. Entonces, además de encontrar los modelos con valores pequeños de la estadística, se sugiere graficar los valores de C_p contra p, para cada modelo, y dibujar la recta $C_p = p$. Los conjuntos de variables correspondientes a puntos cercanos a la recta son los modelos que tienen menor sesgo.

***AIC* (Criterio de información de Akaike).** Este criterio de información busca medir la cercanía entre el modelo estimado y el proceso real que generó los datos utilizando la distancia de Kullback-Leibler, que mide la divergencia entre dos densidades. Akaike propuso seleccionar el modelo que minimice tal distancia, y al encontrar una relación entre ésta y el máximo del logaritmo de la verosimilitud, definió el criterio como

$$AIC = -2\log\left(L\left(\hat{\theta} \mid y\right)\right) + 2K$$

donde $L\left(\hat{\theta} \mid y\right)$ es la verosimilitud evaluada en el estimador máximo verosímil del parámetro θ y K es el número de parámetros a estimar en el modelo. Se elige al modelo con el menor valor del criterio, porque se considera que, dentro de todos los modelos considerados, éste es el que está más cerca de la "realidad" desconocida que generó los datos.

***BIC* (Criterio Bayesiano de información).** Es una modificación del criterio *AIC*. Se basa en la relaciòn que existe entre los estimadores máximos verosímiles y la distribución asintótica final. Para seleccionar el modelo se utiliza una función de pérdida de la forma 0,1, que se minimiza cuando se toma la moda de la distribución; esto último lleva a definir el criterio como

$$BIC = -2\log\left(L\left(\hat{\theta} \mid y\right)\right) + K\log(n)$$

Al igual que en el criterio *AIC*, se elige al modelo con el menor valor *BIC*.

5.1.1 Descripción de los datos y análisis.

Estudio del desempeño de un supervisor (Chatterjee, 1991).

Una encuesta reciente, que se aplicó a los empleados administrativos de una gran organización financiera, incluía preguntas relacionadas con la satisfacción del

empleado con sus supervisores. Había una pregunta que medía el desempeño general del supervisor, así como preguntas relacionadas a actividades específicas que involucran interacciones entre el supervisor y el empleado. Se realizó un estudio exploratorio para tratar de explicar la relación entre características específicas del supervisor y la satisfacción general con los supervisores, percibida por los empleados. Inicialmente, se seleccionaron 6 preguntas del cuestionario como posibles variables explicativas. La descripción de las variables es la siguiente

Variable	Descripción
Y	Indice general del trabajo realizado por el supervisor.
X_1	Manejo de las quejas de los empleados.
X_2	No permite privilegios especiales.
X_3	Oportunidad de aprender nuevas cosas.
X_4	Aumentos de sueldo basados en el desempeño.
X_5	Muy crítico con pobre desempeño.
X_6	Ritmo de avance a mejores puestos.

Las variables X_1, X_2 y X_5 se refieren a relaciones interpersonales directas entre el empleado y el supervisor, mientras que las variables X_3 y X_4 son de una naturaleza menos personal y se refieren al trabajo en general. La variable X_6 no es una evaluación directa del supervisor, pero sirve como una medida general de cómo el empleado percibe su propio progreso dentro de la compañía. Los datos para el análisis fueron generados de la respuesta individual a las preguntas del cuestionario. La respuesta a cada pregunta tenía una escala del 1 al 5, indicando las respuestas muy satisfactorio hasta muy insatisfactorio, respectivamente. Se creó un índice dicotómico para cada pregunta agrupando la escala de respuesta en dos categorías: $\{1,2\}$, que se interpreta como respuesta favorable y $\{3,4,5\}$, que representa una respuesta desfavorable. Los datos se colectaron en 30

departamentos seleccionados al azar de la organización. Cada departamento tenía aproximadamente 35 empleados y un supervisor. Los datos utilizados en el análisis se obtuvieron agrupando las respuestas para obtener la proporción de respuestas favorables para cada pregunta para cada departamento. Los datos resultantes consisten, entonces, de 30 observaciones de 7 variables, una observación por cada departamento.

Análisis.

Los criterios L_m, M_m y K_m se van a comparar con los criterios de selección de variables más comúnmente utilizados, como son: Stepwise, Forward, Backward, C_p de Mallows, AIC y BIC, descritos anteriormente. Los primeros tres se calcularán considerando que se tiene una distribución inicial no informativa (de Jeffreys), para que tenga sentido la comparación.

Se denotará al modelo $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ como $(X_1, X_2, \ldots, X_p)$, con $p = 1, 2, \ldots, k$. En este ejemplo $k = 7$.

Primero, se realiza la regresión con el modelo completo y se obtiene la tabla siguiente

R= .85592172 R^2= .73260199 Adjusted R^2= .66284599
F(6,23)=10.502 p<.00001 Std.Error of estimate: 7.0680

	B	St. Err. of B	t(23)	p-level
Intercpt	10.7871	11.5893	0.9308	0.3616
X1	0.6132	0.1610	3.8090	0.0009
X2	-0.0731	0.1357	-0.5382	0.5956
X3	0.3203	0.1685	1.9009	0.0699
X4	0.0817	0.2215	0.3690	0.7155
X5	0.0384	0.1470	0.2611	0.7963
X6	-0.2171	0.1782	-1.2180	0.2356

Tabla 5.1.1 Estimadores de los coeficientes de regresión para las 6 variables y coeficientes de determinación. (Tabla obtenida utilizando Statistica)

Puede observarse que los estimadores de los coeficientes que son significativamente diferentes de cero son los correspondientes a las variables X_1, X_3 y X_6. Entonces, se puede suponer que el modelo seleccionado contendrá al menos una de estas variables. Con un nivel de significancia $\alpha = 0.05$, la única variable que se seleccionaría sería X_1. El coeficiente de determinación ajustado R_a^2 indica que alrededor del 66% de la variación total del índice general del trabajo realizado por el supervisor es explicado por las 6 variables.

Método Forward.

R= .85181864 R^2= .72559500 Adjusted R^2= .69393288 F(3,26)=22.917 p<.00000 Std.Error of estimate: 6.7343				
	B	St. Err. of B	t(26)	p-level
Intercpt	13.5777	7.5439	1.7998	0.0835
X1	0.6227	0.1181	5.2708	0.0000
X3	0.3124	0.1542	2.0259	0.0532
X6	-0.1870	0.1449	-1.2906	0.2082

Tabla 5.1.2 Estimadores de los coeficientes de regresión para las variables elegidas por el método forward y coeficientes de determinación. (Tabla obtenida utilizando Statistica)

Este método incluye primero a la variable X_1, luego a X_3 y por último a X_6, con un valor de F de entrada de 1. El coeficiente de determinación ajustado es mayor al obtenido con el modelo completo e indica que si se toma el modelo con las tres variables, éstas explicarían alrededor del 69% de la variación total de Y. Con un nivel de significancia $\alpha = 0.05$ se seleccionaría solamente a X_1, y tal vez a X_3, ya que su p-valor es un poco mayor a 0.05.

Con una F de entrada de 4, se elige al modelo (X_1) que, según el coeficiente de determinación ajustado explica alrededor del 67% de la variación total de Y (Tabla 5.1.3).

Método backward.

R= .82541757 R^2= .68131416 Adjusted R^2= .66993253 F(1,28)=59.861 p<.00000 Std.Error of estimate: 6.9933				
	B	St. Err. of B	t(28)	p-level
Intercpt	14.3763	6.6200	2.1717	0.0385
X1	0.7546	0.0975	7.7370	0.0000

Tabla 5.1.3 Estimadores de los coeficientes de regresión para las variables elegidas por el método backward y coeficientes de determinación. (Tabla obtenida utilizando Statistica)

Este método elimina a todas las variables excepto a X_1, es decir, elige al modelo (X_1) como el mejor. El coeficiente de determinación ajustado indica que este modelo explica alrededor del 67% de la variabilidad del índice general del trabajo realizado por el supervisor.

Stepwise.

$rss:
[1] 1,369.38 1,254.64

$size:
[1] 1 2

$which:

	X_1	X_2	X_3	X_4	X_5	X_6
1(+1)	T	F	F	F	F	F
2(+3)	T	F	T	F	F	F

$f.stat:
[1] 59.8608 2.46906

Tabla 5.1.4 Resultados de aplicar el método stepwise utilizando SPlus, donde: $rss es la suma de los cuadrados de los errores para los 2 modelos considerados; $size es el número de variables consideradas en cada modelo; $which son las variables consideradas dentro del modelo a cada paso (T corresponde a la variable considerada y F a la no considerada); $f.stat es el valor de F para cada modelo.

Este método elige, en 2 pasos, al modelo (X_1, X_3) como el mejor, con un valor de entrada de $F = 2$. De acuerdo al coeficiente de determinación ajustado este modelo explica alrededor del 69% de la variabilidad del índice general del trabajo realizado por el supervisor.

R= .84143639 R^2= .70801520 Adjusted R^2= .68638669 F(2,27)=32.735 p<.00000 Std.Error of estimate: 6.8168				
	B	St. Err. of B	t(27)	p-level
Intercpt	9.8709	7.0612	1.3979	0.1735
X1	0.6435	0.1185	5.4316	0.0000
X3	0.2111918	0.1344037	1.5713241	0.1277539

Tabla 5.1.5 Estimadores de los coeficientes de regresión para el modelo con 2 variables y sus coeficientes de determinación. (Tabla obtenida utilizando Statistica)

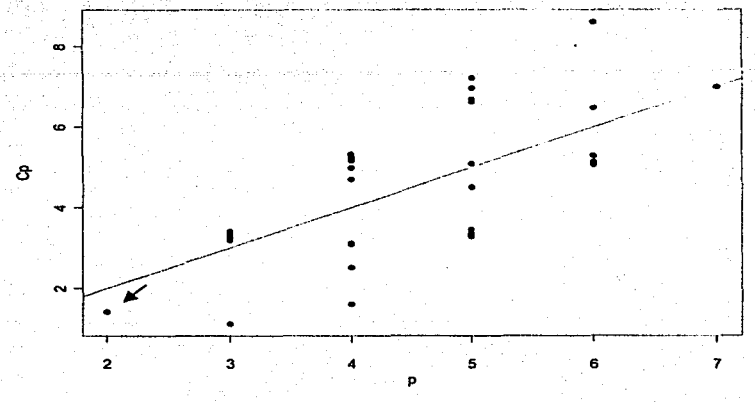
C_p de Mallows.

X_1	X_2	X_3	X_4	X_5	X_6	C_p	p	$\lvert C_p\text{-}p \rvert$
1	0	1	0	0	0	1.1148	3	1.8852
1	0	0	0	0	0	1.4115	2	0.5885
1	0	1	0	0	1	1.6027	4	2.3973
1	1	1	0	0	0	2.5136	4	1.4864
1	0	1	1	0	0	3.0910	4	0.9090
1	0	1	0	1	0	3.1148	4	0.8852
1	0	0	1	0	0	3.1892	3	0.1892
1	1	0	0	0	0	3.2610	3	0.2610
1	1	1	0	0	1	3.2805	5	1.7195
1	0	0	0	0	1	3.3284	3	0.3284

Tabla 5.1.6 Valores de la estadística C_p de Mallows, para los 10 modelos que tienen valores menores y distancias entre C_p y p. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye. (Los valores de esta estadística se obtuvieron utilizando Splus).

Los modelos que tienen valores menores de la estadística, es decir, tienen menor error, son (X_1), (X_1, X_3) y (X_1, X_3, X_6). Entre estos el que tiene una distancia menor con la recta $C_p = p$ (menor sesgo) es (X_1); por lo tanto, según este criterio,

este seria el modelo elegido. En la gráfica 5.1.1. el punto correspondiente a este modelo es el señalado con una flecha.



Gráfica 5.1.1 Valores de la estadística C_p para 32 de los modelos posibles y la recta $C_p = p$. (Gráfica obtenida en Splus).

Criterio *AIC*.

	X_1	X_2	X_3	X_4	X_5	X_6	*AIC*
1	1	0	1	0	0	0	203.1387
2	1	0	1	0	0	1	203.2758
3	1	0	0	0	0	0	203.7638
4	1	1	1	0	0	0	204.4119
5	1	1	1	0	0	1	204.8634
6	1	0	1	1	0	1	204.9550
7	1	0	1	0	1	1	205.0927
8	1	0	1	1	0	0	205.1103
9	1	0	1	0	1	0	205.1387
10	1	0	0	1	0	0	205.5196

Tabla 5.1.7 Los 10 modelos que tienen los menores valores del criterio AIC. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye. (Los valores de esta estadística se obtuvieron utilizando Splus).

Este criterio escoge al modelo (X_1, X_3) como el mejor. Un modelo más pequeño sería (X_1), y según este criterio podría escogerse, ya que la distancia entre éste y el primero es muy pequeña.

Criterio *BIC*.

	X_1	X_2	X_3	X_4	X_5	X_6	*BIC*
1	1	0	0	0	0	0	206.5662
2	1	0	1	0	0	0	207.3423
3	1	0	1	0	0	1	208.8806
4	1	0	0	1	0	0	209.7232
5	1	1	0	0	0	0	209.8023
6	1	0	0	0	0	1	209.8764
7	1	0	0	0	1	0	209.9672
8	1	1	1	0	0	0	210.0167
9	1	0	1	1	0	0	210.7151
10	1	0	1	0	1	0	210.7435

Tabla 5.1.8 Los 10 modelos que tienen los menores valores del criterio BIC. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye.

Para este criterio el mejor modelo es (X_1), ya que tiene el menor valor. Según el número de variables que se deseen en el modelo, también podrían escogerse (X_1, X_3) o (X_1, X_3, X_6).

Criterio L_m.

El modelo (X_1, X_3, X_6) es el que tiene el menor valor del criterio. Con una desviación estándar (una unidad de calibración) pueden elegirse los modelos con valores hasta de 58.8871. Un modelo más parsimonioso es (X_1, X_3), que tiene un valor de L_m = 53.9517, que difiere del primero en 0.5029, menos de un décimo de unidad de calibración. Sin embargo, también podría elegirse el modelo (X_1), cuyo valor del criterio es 55.27, que difiere del "mejor" modelo en 1.3183, que es poco menos de un quinto de unidad de calibración.

	X_1	X_2	X_3	X_4	X_5	X_6	L_m
1	1	0	1	0	0	1	53.3808
2	1	0	1	0	0	0	53.9517
3	1	1	1	0	0	1	54.1554
4	1	0	1	1	0	1	54.2381
5	1	0	1	0	1	1	54.3628
6	1	1	1	0	0	0	54.4012
7	1	0	1	1	0	0	55.0382
8	1	0	1	0	1	0	55.0642
9	1	1	1	1	0	1	55.1195
10	1	1	1	0	1	1	55.2007
11	1	0	0	0	0	0	55.2700
12	1	0	1	1	1	1	55.3835
13	1	1	1	1	0	0	55.5499
14	1	1	1	0	1	0	55.5687
15	1	0	0	1	0	0	56.1356
16	1	1	0	0	0	0	56.2096
17	1	0	1	1	1	0	56.2165
18	1	0	0	0	0	1	56.2791
19	1	1	1	1	1	1	56.3332
20	1	0	0	0	1	0	56.3644
21	1	0	0	1	0	1	56.7806
22	1	1	1	1	1	0	56.7852
23	1	1	0	1	0	0	57.0833
24	1	0	0	1	1	0	57.2669
25	1	1	0	0	0	1	57.3304
26	1	1	0	0	1	0	57.3675
27	1	0	0	0	1	1	57.4312
28	1	1	0	1	0	1	57.9161
29	1	0	0	1	1	1	57.9882
30	1	1	0	1	1	0	58.2860
31	1	1	0	0	1	1	58.5559

Núm. de calibración D_L	5.5063

	X_1	X_2	X_3	X_4	X_5	X_6	L_m
32	1	1	0	1	1	1	59.2035
33	0	0	1	1	0	1	68.1783
34	0	1	1	1	0	1	68.8010
35	0	0	1	1	1	1	69.5733
36	0	1	1	1	1	1	70.2853
37	0	0	1	1	0	0	74.0017
38	0	1	1	1	0	0	74.9726
39	0	1	1	0	0	1	75.0220
40	0	0	1	0	0	1	75.2231
41	0	0	1	0	1	1	75.3448
42	0	1	1	0	1	1	75.3830
43	0	0	1	1	1	0	75.5120
44	0	1	0	1	0	1	76.0531
45	0	0	1	0	0	0	76.5311
46	0	1	1	1	1	0	76.5627
47	0	1	1	0	0	0	76.8536
48	0	0	0	1	0	1	77.4268
49	0	1	0	1	1	1	77.5531
50	0	0	1	0	1	0	77.5874
51	0	1	1	0	1	0	78.1051
52	0	1	0	1	0	0	78.5225
53	0	0	0	1	1	1	78.8482
54	0	0	0	1	0	0	79.0395
55	0	1	0	1	1	0	79.8529
56	0	0	0	1	1	0	80.2905
57	0	1	0	0	0	0	88.5722
58	0	1	0	0	1	0	89.8297
59	0	1	0	0	0	1	90.3215
60	0	1	0	0	1	1	91.6696
61	0	0	0	0	0	0	96.0756
62	0	0	0	0	1	0	96.7003
63	0	0	0	0	0	1	96.7212
64	0	0	0	0	1	1	97.9388

Tabla 5.1.9 Valores del criterio L_m para todos los modelos y el número de calibración D_L. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye. En la primera parte de la tabla (lado izquierdo) están los modelos que se consideran buenos, por una desviación (D_L).

Criterio M_m.

Para este criterio también el modelo (X_1, X_3, X_6) es el que tiene un valor menor, por lo tanto, sería el "mejor" modelo. El número de calibración nos indica que cualquier modelo con un valor menor o igual a 33.6824 también puede ser elegido. Un modelo más parsimonioso sería (X_1, X_3) con un valor del criterio igual a 28.0554, que dista del anterior en 0.5029, que es menos de un décimo de unidad de calibración. El modelo (X_1) tiene un valor de 28.9427, que difiere del primero

en menos de un cuarto de unidad de calibración, por lo que este modelo también podría ser elegido.

	X_1	X_2	X_3	X_4	X_5	X_6	M_m
1	1	0	1	0	0	1	27.5525
2	1	1	1	0	0	1	27.7307
3	1	0	1	1	0	1	27.7731
4	1	0	1	0	1	1	27.8369
5	1	1	1	1	0	1	27.9850
6	1	1	1	0	1	1	28.0262
7	1	0	1	0	0	0	28.0554
8	1	1	1	0	0	0	28.0791
9	1	0	1	1	1	1	28.1191
10	1	1	1	1	1	1	28.3409
11	1	0	1	1	0	0	28.4079
12	1	0	1	0	1	0	28.4213
13	1	1	1	1	0	0	28.4448
14	1	1	1	0	1	0	28.4544
15	1	0	1	1	1	0	28.7861
16	1	1	1	1	1	0	28.8307
17	1	0	0	0	0	0	28.9427
18	1	0	0	1	0	0	29.1911
19	1	1	0	0	0	0	29.2296
20	1	0	0	0	0	1	29.2657
21	1	0	0	1	0	1	29.3073
22	1	0	0	0	1	0	29.3100
23	1	1	0	1	0	0	29.4635
24	1	0	0	1	1	0	29.5583
25	1	1	0	0	0	1	29.5910
26	1	1	0	0	1	0	29.6102
27	1	0	0	0	1	1	29.6430
28	1	1	0	1	0	1	29.6564
29	1	0	0	1	1	1	29.6933
30	1	1	0	1	1	0	29.8458
31	1	1	0	0	1	1	29.9840
32	1	1	0	1	1	1	30.0585

Núm. de calibración D_M	6.1299

	X_1	X_2	X_3	X_4	X_5	X_6	M_m
33	0	0	1	1	0	1	35.1901
34	0	1	1	1	0	1	35.2301
35	0	0	1	1	1	1	35.6256
36	0	1	1	1	1	1	35.6849
37	0	0	1	1	0	0	38.4816
38	0	1	1	0	1	1	38.6005
39	0	1	1	1	0	0	38.6970
40	0	1	1	0	0	1	38.7225
41	0	0	1	0	1	1	38.8891
42	0	0	1	1	1	0	38.9754
43	0	0	1	0	0	1	39.1167
44	0	1	1	1	1	0	39.2046
45	0	1	0	1	0	1	39.2547
46	0	1	0	1	1	1	39.7117
47	0	1	1	0	0	0	39.9646
48	0	0	1	0	0	0	40.0762
49	0	0	0	1	0	1	40.2627
50	0	1	1	0	1	0	40.3138
51	0	0	1	0	1	0	40.3462
52	0	0	0	1	1	1	40.6974
53	0	1	0	1	0	0	40.8325
54	0	1	0	1	1	0	41.2160
55	0	0	0	1	0	0	41.3898
56	0	0	0	1	1	0	41.7518
57	0	1	0	0	0	0	46.3817
58	0	1	0	0	1	0	46.7123
59	0	1	0	0	0	1	46.9680
60	0	1	0	0	1	1	47.3152
61	0	0	0	0	1	0	50.6380
62	0	0	0	0	0	0	50.6430
63	0	0	0	0	0	1	50.6490
64	0	0	0	0	1	1	50.9291

Tabla 5.1.10 Valores del criterio M_m para todos los modelos y el número de calibración D_M. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye. En la primera parte de la tabla (lado izquierdo) están los modelos que se consideran buenos, por una desviación (D_M).

Criterio K_m.

Este criterio, al tomar como referencia al modelo completo, elige como el mejor al modelo que tiene el menor valor, ya que éste se interpreta como la cantidad de información que se pierde al quitar algunas variables del modelo. Así, el elegido es $(X_1, X_2, X_3, X_4, X_6)$. Como puede notarse, el criterio tiende a elegir modelos con un número de variables cercano al total, en este caso 6, así

que si se desean modelos con menos variables es mejor tomar como referencia al modelo sin variables.

	X_1	X_2	X_3	X_4	X_5	X_6	K_m
1	1	1	1	1	1	1	0.0000
2	1	1	1	1	0	1	0.2995
3	1	1	1	0	1	1	0.3458
4	1	0	1	1	1	1	0.4505
5	1	1	1	0	0	1	0.7017
6	1	0	1	1	0	1	0.7499
7	1	0	1	0	1	1	0.8229
8	1	0	1	0	0	1	1.1818
9	1	1	1	1	1	0	1.2778
10	1	1	1	1	0	0	1.5363
11	1	1	1	0	1	0	1.5479
12	1	1	1	0	0	0	1.8055
13	1	0	1	1	1	0	1.9508
14	1	0	1	1	0	0	2.2073
15	1	0	1	0	1	0	2.2239
16	1	0	1	0	0	0	2.4820
17	1	1	0	1	1	1	2.8025
18	1	1	0	1	0	1	3.0512
19	1	0	0	1	1	1	3.0992
20	1	1	0	1	1	0	3.2988
21	1	0	0	1	0	1	3.3534
22	1	1	0	0	1	1	3.4812
23	1	1	0	1	0	0	3.5594
24	1	0	0	1	1	0	3.6853
25	1	1	0	0	0	1	3.7291
26	1	1	0	0	1	0	3.7546
27	1	0	0	0	1	1	3.7986
28	1	0	0	1	0	0	3.9518
29	1	1	0	0	0	0	4.0035
30	1	0	0	0	0	1	4.0522
31	1	0	0	0	1	0	4.1121
32	1	0	0	0	0	0	4.3646

	X_1	X_2	X_3	X_4	X_5	X_6	K_m
33	0	1	1	1	1	1	11.2122
34	0	1	1	1	0	1	11.4547
35	0	0	1	1	1	1	12.1334
36	0	0	1	1	0	1	12.3917
37	0	1	1	0	1	1	17.5646
38	0	1	1	1	1	0	18.7358
39	0	1	1	1	0	0	18.9237
40	0	1	1	0	0	1	18.9741
41	0	0	1	0	1	1	19.3045
42	0	0	1	1	1	0	19.4763
43	0	0	1	1	0	0	19.6660
44	0	1	0	1	1	1	19.7365
45	0	1	0	1	0	1	20.0356
46	0	0	1	0	0	1	20.9571
47	0	1	1	0	1	0	22.2008
48	0	1	1	0	0	0	22.7207
49	0	0	0	1	1	1	23.0022
50	0	0	0	1	0	1	23.3517
51	0	0	1	0	1	0	23.5294
52	0	1	0	1	1	0	24.1001
53	0	0	1	0	0	0	24.2073
54	0	1	0	1	0	0	24.5732
55	0	0	0	1	1	0	26.5872
56	0	0	0	1	0	0	27.1076
57	0	1	0	0	1	1	38.2288
58	0	1	0	0	1	0	38.3494
59	0	1	0	0	0	1	38.9960
60	0	1	0	0	0	0	39.1189
61	0	0	0	0	1	1	49.5067
62	0	0	0	0	1	0	50.5648
63	0	0	0	0	0	1	50.5956
64	0	0	0	0	0	0	52.4450

Tabla 5.1.11 Los 10 modelos que tienen los valores menores del criterio K_m, tomando al modelo completo como m_0. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye.

Como se puede observar en la Tabla 5.1.12, tomando como referencia al modelo sin variables, el criterio elige como el mejor al modelo con el mayor valor, ya que éste se interpreta como la cantidad de información que se gana al agregar variables al modelo. Así, se elige a (X_1, X_3, X_6) con un valor de 59.2881. Si se desearan modelos con menos variables se podría elegir a (X_1, X_3) o a (X_1), ya que son los modelos con 2 y 1 variables, respectivamente, que tienen los mayores valores del criterio.

	X_1	X_2	X_3	X_4	X_5	X_6	K_{ml}
1	1	0	1	0	0	1	59.2581
2	1	1	1	0	0	1	57.0771
3	1	0	1	0	0	0	56.9953
4	1	0	1	1	0	1	56.8117
5	1	0	1	0	1	1	56.4139
6	1	1	1	0	0	0	55.9246
7	1	1	1	1	0	1	54.4194
8	1	1	1	0	1	1	54.1726
9	1	0	1	1	0	0	53.9409
10	1	0	1	0	1	0	53.8614
11	1	0	1	1	1	1	53.6210
12	1	1	1	1	0	0	52.7643
13	1	1	1	0	1	0	52.7083
14	1	0	0	0	0	0	52.4847
15	1	1	1	1	1	1	51.1980
16	1	0	1	1	1	0	50.8200
17	1	0	0	1	0	0	50.3152
18	1	1	0	0	0	0	50.1030
19	1	0	0	0	0	1	49.9046
20	1	0	0	0	1	0	49.6621
21	1	1	1	1	1	0	49.5754
22	1	0	0	1	0	1	48.8665
23	1	1	0	1	0	0	48.0344
24	1	0	0	1	1	0	47.5362
25	1	1	0	0	0	1	47.3652
26	1	1	0	0	1	0	47.2656
27	1	0	0	0	1	1	47.0950
28	1	1	0	1	0	1	46.1783
29	1	0	0	1	1	1	45.9908
30	1	1	0	1	1	0	45.2243
31	1	1	0	0	1	1	44.5405
32	1	1	0	1	1	1	43.2931

	X_1	X_2	X_3	X_4	X_5	X_6	K_{ml}
33	0	0	1	1	0	1	25.1634
34	0	1	1	1	0	1	24.5816
35	0	0	1	1	1	1	23.4534
36	0	1	1	1	1	1	22.7949
37	0	0	1	1	0	0	16.9946
38	0	1	1	1	0	0	16.2716
39	0	1	1	0	0	1	16.2169
40	0	1	1	0	1	1	16.1827
41	0	0	1	0	1	1	15.8624
42	0	0	1	1	1	0	15.6809
43	0	0	1	0	0	1	15.6152
44	0	1	0	1	0	1	15.1031
45	0	1	1	1	1	0	14.9347
46	0	1	0	1	1	1	13.9387
47	0	1	1	0	0	0	13.8923
48	0	0	1	0	0	0	13.8214
49	0	0	0	1	0	1	13.3169
50	0	0	1	0	1	0	13.1585
51	0	1	1	0	1	0	13.0395
52	0	0	0	1	1	1	12.3393
53	0	1	0	1	0	0	12.2587
54	0	1	0	1	1	0	11.4301
55	0	0	0	1	0	0	11.3712
56	0	0	0	1	1	0	10.6600
57	0	1	0	0	0	0	4.2517
58	0	1	0	0	1	0	3.9566
59	0	1	0	0	0	1	3.6848
60	0	1	0	0	1	1	3.3786
61	0	0	0	0	1	0	0.2263
62	0	0	0	0	0	1	0.2179
63	0	0	0	0	1	1	0.1954
64	0	0	0	0	0	0	0.0000

Tabla 5.1.12 Los 10 modelos que tienen los valores menores del criterio K_m, tomando al modelo sin variables como m_0. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye.

Cálculo de probabilidades iniciales y finales.

El modelo (X_1, X_3) es el que tiene una probabilidad final mayor, por lo tanto, según este criterio, es el que se elige como el mejor. Si se desea un modelo más parsimonioso puede elegirse a (X_1), ya que su probabilidad final también es alta y muy cercana al del primer modelo, y además su probabilidad inicial es la mayor de todos los modelos, lo que indica que si no se tuviera la información que proporcionan los datos Y, sería el elegido.

	X_1	X_2	X_3	X_4	X_5	X_6	$p(m)$	$p(m\|y)$
1	1	0	1	0	0	0	0.02	0.12
2	1	0	0	0	0	0	0.04	0.11
3	1	0	1	0	0	1	0.01	0.10
4	1	0	1	1	0	1	0.01	0.07
5	1	1	1	0	0	1	0.01	0.07
6	1	0	1	0	1	1	0.01	0.06
7	1	1	1	0	0	0	0.01	0.05
8	1	0	1	0	1	0	0.01	0.04
9	1	0	1	1	0	0	0.01	0.04
10	1	0	0	1	0	0	0.02	0.04
11	1	1	0	0	0	0	0.02	0.04
12	1	1	1	0	1	0	0.01	0.03
13	1	1	1	1	0	0	0.01	0.03
14	1	0	0	0	0	1	0.02	0.03
15	1	0	0	0	1	0	0 02	0.03
16	1	0	1	1	1	0	0.01	0.02
17	1	0	0	0	1	1	0.01	0.01
18	1	0	0	1	0	1	0.01	0.01
19	1	0	0	1	1	0	0.01	0.01
20	1	0	0	1	1	1	0.01	0.01
21	1	1	0	0	0	1	0.01	0.01
22	1	1	0	0	1	0	0.01	0.01
23	1	1	0	0	1	1	0.01	0.01
24	1	1	0	1	0	0	0.01	0.01
25	1	1	0	1	0	1	0.01	0.01
26	1	1	0	1	1	0	0.01	0.01

Tabla 5.1.13 Modelos cuya probabilidad final es mayor que cero. El valor 1 indica que la variable X_i se incluye en el modelo, 0 que no se incluye.

5.1.2 Conclusiones.

En general, los métodos eligen a los modelos (X_1, X_3, X_6), (X_1, X_3) y (X_1) como los "mejores". Dependiendo del uso que se les vaya a dar y de qué tan parsimonioso se quiere que sea el modelo elegido, se podría utilizar cualquiera de estos tres y se tendría una buena explicación de los datos.

5.2 SELECCIÓN DE TRANSFORMACIONES.

En regresión lineal, las transformaciones de las variables explicativas pueden llevar a predicciones más exactas y a un modelo que ajuste mejor los datos. Box y Cox consideraron unas transformaciones que, aunque enfatizan en trasformar la variable respuesta, pueden ser utilizadas para las variables independientes.

En la presente sección se mostrará cómo dos de los criterios predictivos, L y M, pueden utilizarse para seleccionar a un miembro específico de una familia paramétrica de transformaciones.

Considérese la ecuación (3.2) donde un modelo $m \in M$ consiste de un miembro específico de una familia de transformaciones dada y es indizado por un vector de parámetros $\alpha = (\alpha_1, \ldots, \alpha_k)'$. Así, X_m en la ecuación (3.2) denota una matriz de variables independientes transformadas y $\beta^{(m)}$ es el vector de los coeficientes de regresión correspondientes a X_m. La tarea es seleccionar o estimar el parámetro α utilizando los criterios L_m y M_m, que ahora son funciones de α y pueden escribirse alternativamente como $L(\alpha)$ y $M(\alpha)$. También X_m puede escribirse como X_α.

Se trabajará con la ampliamente utilizada familia de transformaciones potencia Box-Cox, que está dada por

$$g(x;\alpha) = \begin{cases} (x^\alpha - 1)/\alpha & \alpha \neq 0, \\ \log(x) & \alpha = 0. \end{cases}$$

Con esta familia, para cada una de las variables explicativas, la i-ésima fila de X_α sería $(g(x_{i1};\alpha_1), \ldots, g(x_{ik};\alpha_k))$. Otra opción es elegir diferentes familias de

transformaciones para diferentes variables o transformar dos variables independientes con un valor del parámetro común, para la misma o diferentes familias. También puede considerarse el utilizar dos o más valores del parámetro de una familia de transformaciones sobre una sola variable; por ejemplo, suponiendo que se tiene una sola variable independiente X cuyo valor para la i-ésima observación es x_i, y se toma $\alpha = (\alpha_1, \alpha_2)'$. Así la i-ésima fila de X_α es de la forma $(g(x_{i1};\alpha_1), g(x_{i2};\alpha_2))$. Este tipo de transformaciones puede ser conveniente cuando se cree que la función de regresión es lineal en dos diferentes potencias de una variable. Finalmente, se puede optar por no transformar algunas de las variables. Los métodos aquí presentados permiten cierta flexibilidad, ya que α puede tener una dimensión diferente al número de columnas de X_α, es decir, puede ser diferente del número de variables que se tienen en el problema.

5.2.1 Experimento generado.

Se generaron observaciones Y de cierto experimento ficticio, utilizando el modelo

$$Y = \beta_0 + \beta_1\left(\frac{X_1^{\alpha_1'} - 1}{\alpha_1}\right) + \beta_2\left(\frac{X_2^{\alpha_2'} - 1}{\alpha_2}\right) + \varepsilon$$

donde X_1, X_2 y ε son variables aleatorias que fueron generadas utilizando S-PLUS, con distribuciones $X_1 \sim N(x_1 \mid 6, 2.5)$, $X_2 \sim N(x_2 \mid 30, 10)$. Se realizaron dos ejercicios de simulación, en donde $\beta = (15, 4, -8)$, con $\varepsilon \sim N(\varepsilon \mid 0, 15)$ para el Ejercicio 1 y $\varepsilon \sim N(\varepsilon \mid 0, 5)$ para el Ejercicio 2. Después, utilizando Y, X_1 y X_2 se busca $\alpha = (\alpha_1, \alpha_2)'$ que minimice los valores de los criterios. Claramente, lo que se desea es que con este método se escojan los valores de α_1' y α_2' que se sabe fueron utilizados en un principio para crear Y. Todo lo anterior se hizo con la finalidad de probar la funcionalidad de este método de selección de transformaciones.

Ejemplo 1.

Se generaron 100 observaciones del experimento, con $\alpha_1' = 2$ y $\alpha_2' = 0.5$. Utilizando los criterios L_m y M_m para seleccionar α se obtuvieron los siguientes valores:

α_1	1.7020
α_2	0.3480
L_m	217.5664
D_L	7.5747
M_m	63.1634
D_M	4.4910

Tabla 5.2.1 Valores de α y de los criterios, para el Ejemplo 1.

Según la tabla, los valores de que minimizan los criterios son $\alpha_1 = 1.7020$ y $\alpha_2 = 0.3480$. Con $\alpha_1' = 2$ y $\alpha_2' = 0.5$ los valores de los criterios son

L'_m	222.6302
$\lvert L_m - L'_m \rvert$	5.0638
M'_m	64.6335
$\lvert M_m - M'_m \rvert$	1.4701

Tabla 5.2.2 Valores de los criterios para α y diferencias entre los valores de los criterios para α y α', para el Ejemplo 1.

Como $|L_m - L'_m| < D_L$, es decir, la diferencia entre los valores del criterio L para α y α' es menor al número de calibración D_L, la transformación realizada con los valores $\alpha_1' = 2$ y $\alpha_2' = 0.5$ es tan buena como la que se obtiene utilizando $\alpha_1 = 1.7020$ y $\alpha_2 = 0.3480$. Con el criterio M se llega a la misma conclusión, ya que también $|M_m - M'_m| < D_M$.

R= .96992697 R^2= .94075832 Adjusted R^2= .93953684 F(2,97)=770.18 p<0.0000 Std.Error of estimate: 15.302				
	B	St. Err. of B	t(97)	p-level
Intercpt	17.1670877	8.25848136	2.0787221	0.04028244
X1T2	6.93224514	0.19719043	35.1550785	0
X2T2	-12.5611606	1.162026	-10.809707	2.4169E-18

Tabla 5.2.3 Estimadores de los coeficientes de regresión y coeficientes de determinación, para la transformación con $\alpha_1 = 1.7020$ y $\alpha_2 = 0.348$, para el Ejemplo 1. (Tabla obtenida utilizando Statistica)

R= .96848776 R^2= .93796855 Adjusted R^2= .93668955 F(2,97)=733.36 p<0.0000 Std.Error of estimate: 15.658				
	B	St. Err. of B	t(97)	p-level
Intercpt	16.1538343	7.2455413	2.2294862	0.02808853
X1T	3.90533413	0.11388781	34.2910637	0
X2T	-7.73863048	0.73282915	-10.5599382	8.3204E-18

Tabla 5.2.4 Estimadores de los coeficientes de regresión y coeficientes de determinación, para la transformación con $\alpha'_1 = 2$ y $\alpha'_2 = 0.5$, para el Ejemplo 1. (Tabla obtenida utilizando Statistica)

En las tablas anteriores puede observarse que $R_a^2 = 0.9395$ para la transformación con $\alpha_1 = 1.7020$ y $\alpha_2 = 0.3480$, lo que significa que este modelo explica alrededor del 93.9% de la variación total de Y. Utilizando α' el valor del coeficiente de determinación ajustado es un poco menor (0.9367) y los valores de $\hat{\beta}$ son más cercanos a los valores reales.

Ejemplo 2.

Se generaron 100 observaciones del experimento, con $\alpha'_1 = 1$ y $\alpha'_2 = 0$. Utilizando los criterios L_m y M_m para seleccionar α se obtuvieron los siguientes valores:

α_1	0.8420
α_2	0.1670
L_m	50.4316
D_L	1.7558
M_m	14.6412
D_M	1.0410

Tabla 5.2.5 Valores de α y de los criterios, para el Ejemplo 2.

Según la tabla, los valores de L que minimizan los criterios son $\alpha_1 = 0.8240$ y $\alpha_2 = 0.1670$. Utilizando $\alpha'_1 = 1$ y $\alpha'_2 = 0$ los valores de los criterios son

L'_m	50.8100
$\lvert L_m - L'_m \rvert$	0.3784
M'_m	14.7511
$\lvert M_m - M'_m \rvert$	0.1099

Tabla 5.2.6 Valores de los criterios para α y diferencias entre los valores de los criterios para α y α', para el Ejemplo 2.

Como $|L_m - L'_m| < D_L$, es decir, la diferencia entre los valores del criterio L para α y α' es menor al número de calibración D_L, la transformación realizada con los valores de α' es tan buena como la que se obtiene utilizando α. Con el criterio M se llega a la misma conclusión, ya que también $|M_m - M'_m| < D_M$. En las tablas 5.2.7 y 5.2.8 puede observarse que, además, ambas transformaciones explican prácticamente el mismo porcentaje de la variación total de Y, alrededor del 89%.

R= .94380783 R^2= .89077321 Adjusted R^2= .88852112 F(2,97)=395.53 p<0.0000 Std.Error of estimate: 3.5469				
	B	St. Err. of B	t(97)	p-level
Intercpt	8.4239749	2.44701508	3.44255129	0.00085204
X1T2Y2	5.25802261	0.21965361	23.9377924	1.7138E-42
X2T2Y2	-4.92455177	0.47469536	-10.3741307	2.0906E-17

Tabla 5.2.7 Estimadores de los coeficientes de regresión y coeficientes de determinación, para la transformación con $\alpha_1 = 0.842$ y $\alpha_2 = 0.167$, para el Ejemplo 2. (Tabla obtenida utilizando Statistica)

R= .94293573 R^2= .88912780 Adjusted R^2= .88684177 F(2,97)=388.94 p<0.0000 Std.Error of estimate: 3.5735				
	B	St. Err. of B	t(97)	p-level
Intercpt	15.4772078	2.91986041	5.30066702	7.2384E-07
X1TY2	3.98469541	0.16763754	23.7697089	3.0759E-42
X2TY2	-8.20274776	0.80061782	-10.2455223	3.9584E-17

Tabla 5.2.8 Estimadores de los coeficientes de regresión y coeficientes de determinación, para la transformación con $\alpha'_1 = 1$ y $\alpha'_2 = 0$, para el Ejemplo 2. (Tabla obtenida utilizando Statistica)

Con los anteriores ejemplos se ha observado que, aunque con el método de selección de transformaciones basado en los criterios L y M no se elige el valor exacto de α utilizado para crear Y, éste queda dentro del intervalo (determinado por los números de calibración) de los valores asociados a modelos que también se consideran buenos. Es decir, se puede tener la certeza de que el valor de α elegido utilizando este método da lugar a una transformación suficientemente buena.

5.2.2 Aplicación del método de selección de transformaciones a datos reales.

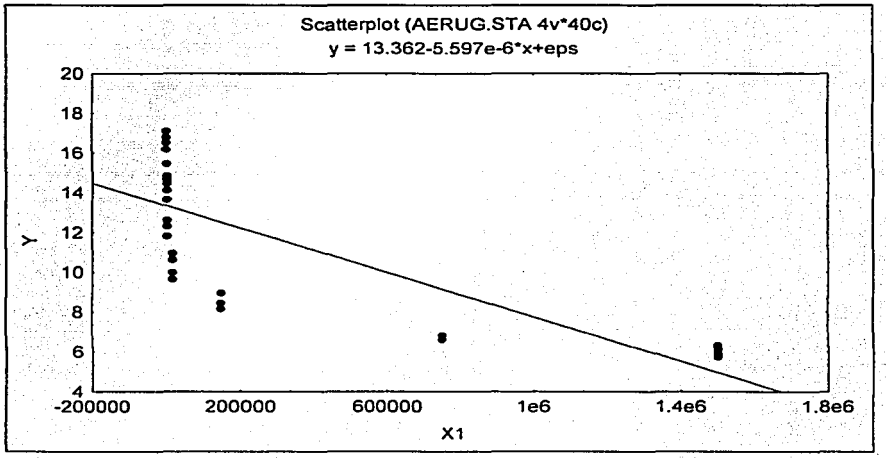
Tiempo de crecimiento de una colonia de bacterias.

En un estudio reciente de Cancerología se observó que el catéter que se coloca a los pacientes puede provocar infecciones a diferentes niveles, desde piel

hasta sangre. Para conocer este nivel se puede contar el número de bacterias presentes, ya que, mientras más sean, mayor será el nivel de infección. Sin embargo, en la realidad esto no se puede hacer, así que se cultivaron colonias de bacterias (Pseudomonas Aeruginosa) de diferentes tamaños, se les colocó cierto reactivo y se registró el tiempo (en horas) que tardó en presentarse la infección. Con estos datos, lo que se pretendía era encontrar un modelo de regresión adecuado con el que, dada cierta observación del tiempo que tardó en presentarse la infección en un paciente, se pudiera saber aproximadamente el número de bacterias presentes y, así, el nivel de infección.

A continuación, para ilustrar la selección de transformaciones con datos reales, se tomará como variable explicativa (X) al tamaño de la colonia y como variable respuesta (Y) al tiempo, y se buscará la mejor transformación que explique la relación entre estas variables.

Análisis.



Gráfica 5.2.1 X_1 vs. Y (Gráfica obtenida utilizando Statistica)

En la gráfica anterior se observa que los datos no tienen una tendencia lineal, por lo que es necesario realizar una transformación para linealizarlos y así poder realizar predicciones de los valores de Y. Además, el coeficiente de determinación ajustado indica que con este modelo sólo se explica alrededor del 61% de la variación total de Y (Tabla 5.2.9).

Regression Summary for Dependent Variable: Y R= .78533627 R^2= .61675305 Adjusted R^2= .60666760				
	B	St. Err. of B	t(38)	p-level
Intercpt	13.3620609	0.45042684	29.6653299	6.9103E-28
X1	-5.5972E-06	7.1575E-07	-7.82002421	1.9666E-09

Tabla 5.2.9 Estimadores de los coeficientes de regresión y coeficientes de determinación, para el modelo $Y = \beta_0 + \beta_1 X_1$. (Tabla obtenida utilizando Statistica)

α	0.0010
L_m	4.5618
D_L	0.2485
M_m	2.0799
D_M	0.2390

Tabla 5.2.10 Valores de α y de los criterios.

Utilizando los criterios L y M para seleccionar una transformación se obtienen los valores de la Tabla 5.2.10, según los cuales la mejor transformación es con $\alpha = 0.001$. Como α es muy cercana a cero, se calcularán los valores de los criterios para la transformación $\log X_1$ $(\alpha' = 0)$, para ver si ésta se puede utilizar en lugar de la ya obtenida.

L'_m	4.6266
$\lvert L_m - L'_m \rvert$	0.0648
M'_m	2.0993
$\lvert M_m - M'_m \rvert$	0.0194

Tabla 5.2.11 Valores de los criterios para α y diferencias entre los valores de los criterios para α y α´, para el Ejemplo 2.

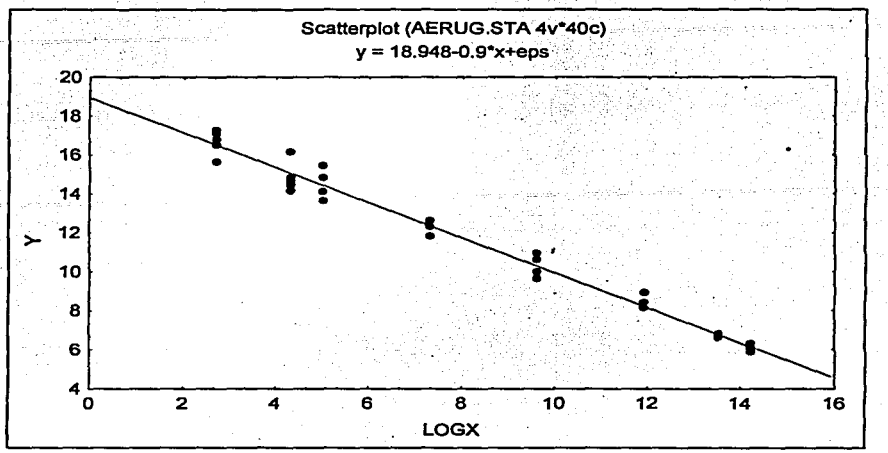
La diferencia entre L_m y L'_m es del orden de un cuarto de unidad de calibración, por lo tanto, la transformación $\log X_1$ $(\alpha' = 0)$ se considera buena, según este criterio. Con el criterio M sucede lo mismo, ya que la diferencia entre M_m y M'_m es de poco menos de un décimo de unidad de calibración.

Regression Summary for Dependent Variable: Y
R= .99197530 R^2= .98401499 Adjusted R^2= .98359433

	BETA	St. Err. of B	t(38)	p-level
Intercpt		0.17166154	110.38081	0
LOGX	-0.9919753	0.01860867	-48.3655614	9.6174E-36

Tabla 5.2.12 Estimadores de los coeficientes de regresión y coeficientes de determinación, para el modelo $Y = \beta_0 + \beta_1 \log X_1$. (Tabla obtenida utilizando Statistica)

Esta transformación también ayuda a que aumente el coeficiente de determinación, lo que significa que el modelo se ajusta mejor a los datos, lo que se observa en la Gráfica 5.2.2, explicando así alrededor del 98.4% de la variación total de Y.



Gráfica 5.2.2 log X_1 vs. Y (Gráfica obtenida utilizando Statistica)

5.2.3 Otra aplicación del método de selección de transformaciones.

Como ya se mencionó anteriormente, también puede repetirse la misma familia de transformaciones con una variable, pero con diferentes valores del parámetro. Este tipo de transformaciones puede ser conveniente cuando se cree que la función de regresión es lineal en dos o más potencias de una variable.

A continuación se aplicará este método a los datos de la sección 5.2.2, para obtener un modelo de la forma

$$Y = \beta_0 + \beta_1\left(\frac{X_1^{\alpha_1} - 1}{\alpha_1}\right) + \beta_2\left(\frac{X_1^{\alpha_2} - 1}{\alpha_2}\right) + \varepsilon$$

α_1	-0.1740
α_2	0.1050
L_m	4.5319
D_L	0.2436
M_m	2.0563
D_M	0.2333

Tabla 5.2.12 Valores de α_1, α_2 y de los criterios, para la transformación con una sola variable.

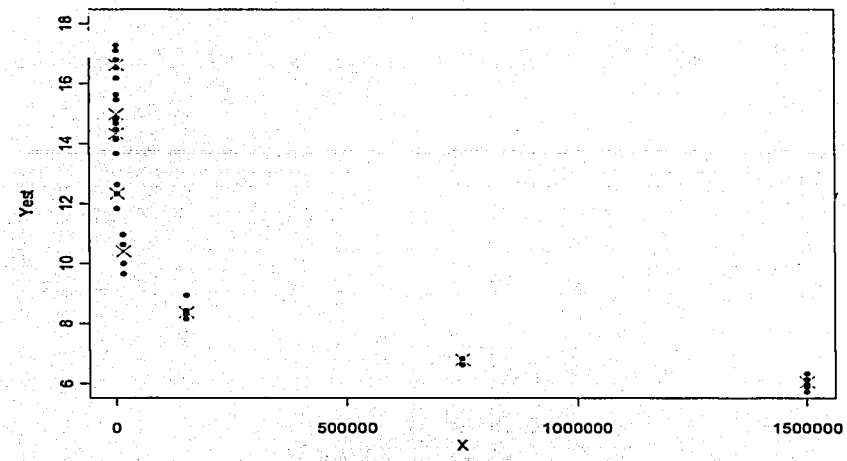
Según los criterios L y M con los valores $\alpha_1 = -0.174$ y $\alpha_2 = 0.105$ se tiene la mejor transformación en 2 potencias de una variable para este problema, obteniéndose la tabla siguiente.

Regression Summary for Dependent Variable: Y
R= .99230201 R^2= .98466327 Adjusted R^2= .98383426

	B	St. Err. of B	t(37)	p-level
Intercpt	20.1179946	0.51145077	39.3351534	8.6048E-32
X1POLIA1	-1.2958958	0.19381657	-6.68619708	7.4502E-08
X2POLIA1	-0.22034201	0.01936502	-11.378354	1.2121E-13

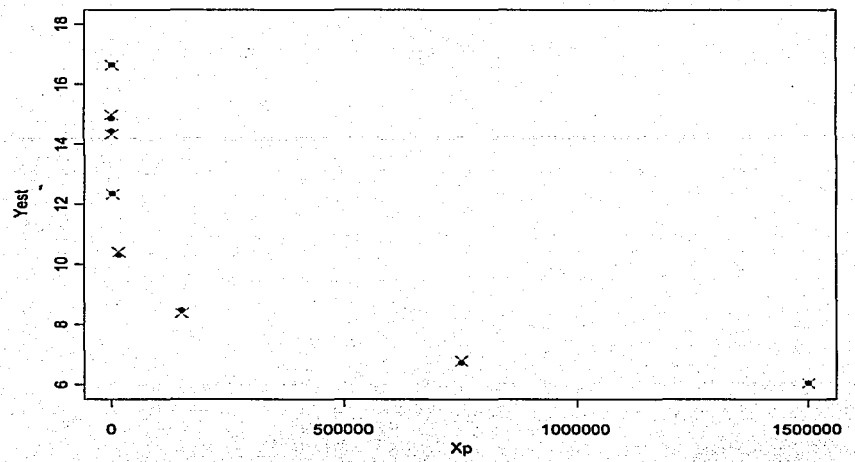
Tabla 5.2.14 Estimadores de los coeficientes de regresión y coeficientes de determinación, para el modelo con 2 potencias de una variable. (Tabla obtenida utilizando Statistica)

Con esta transformación se logra un ajuste mucho mejor que con $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, lo que puede observarse en la Gráfica 5.2.3. Según el coeficiente de determinación ajustado, el nuevo modelo explica alrededor del 98.4% de la variación total de Y.



Gráfica 5.2.3 Gráfica de dispersión, donde • corresponde a los valores observados del vector Y y × corresponde a los valores de Y obtenidos utilizando el nuevo modelo. (Gráfica obtenida utilizando S-Plus)

En la Gráfica 5.2.3 se observa que los valores obtenidos utilizando el modelo $Y = \beta_0 + \beta_1\left(\frac{X_1^{\alpha_1} - 1}{\alpha_1}\right) + \beta_2\left(\frac{X_1^{\alpha_2} - 1}{\alpha_2}\right) + \varepsilon$, con $\alpha_1 = -0.174$ y $\alpha_2 = 0.105$, se acercan mucho a los valores originales, de hecho son muy cercanos al valor del promedio de los valores de Y asociados a cada X (Gráfica 5.2.4), por lo tanto, es un buen modelo para realizar predicciones.



Gráfica 5.2.4 Gráfica de dispersión, donde • corresponde a los promedios de los valores del vector Y asociados a cada valor de X y × corresponde a los valores de Y obtenidos utilizando el nuevo modelo. (Gráfica obtenida utilizando S-Plus)

Si se quisieran utilizar valores más interpretables, podría transformarse con $\alpha'_1 = -0.2$ y $\alpha'_2 = 0.1$, ya que las diferencias entre L_m y L'_m, y entre M_m y M'_m son insignificantes, por lo tanto, con α' la transformación es tan buena como con α.

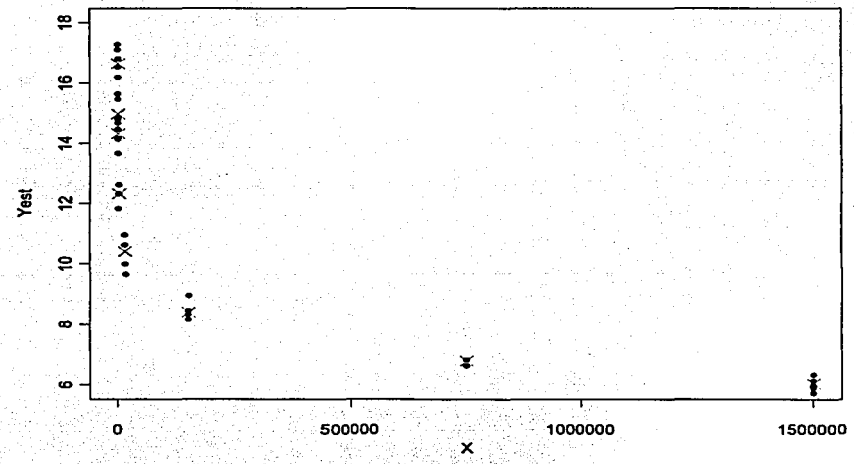
L'_m	4.5327
$\lvert L_m - L'_m \rvert$	0.0008
M'_m	2.0566
$\lvert M_m - M'_m \rvert$	0.0004

Tabla 5.2.15 Valores de los criterios para α y diferencias entre los valores de los criterios para α y α'.

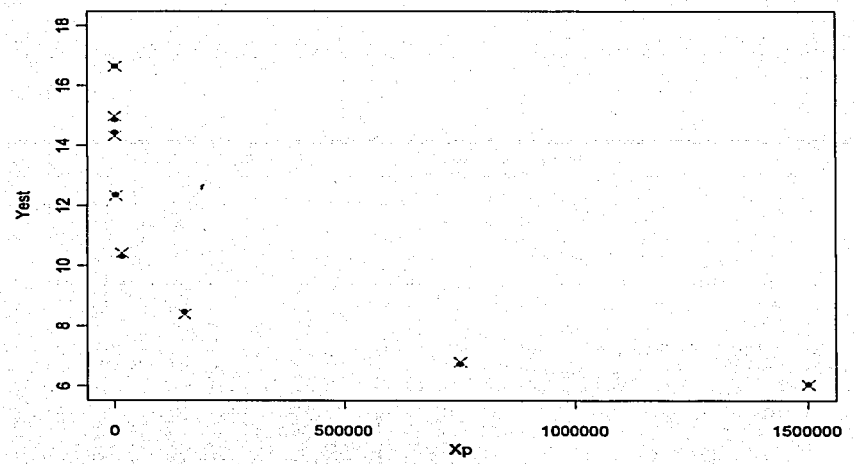
Regression Summary for Dependent Variable: Y R= .99229927 R^2= .98465784 Adjusted R^2= .98382853				
	B	St. Err. of B	t(37)	p-level
Intercpt	20.2950329	0.5484455	37.0046482	7.8061E-31
X1POLI	-1.38200733	0.21624033	-6.39107114	1.8589E-07
X2POLI	-0.24601422	0.01903568	-12.9238482	2.7446E-15

Tabla 5.2.16 Estimadores de los coeficientes de regresión y coeficientes de determinación, para el ' modelo con α'. (Tabla obtenida utilizando Statistica)

Además, según el coeficiente de determinación, con esta nueva transformación se explica prácticamente el mismo porcentaje de la variación total de Y que con la que utiliza α. En la gráfica de dispersión se observa que se sigue conservando la cercanía entre los valores estimados de Y y los originales. Lo mismo sucede con los valores promedio de la variable respuesta. (Gráficas 5.2.5 y 5.2.6)



Gráfica 5.2.5 Gráfica de dispersión, donde • corresponde a los valores del vector Y y × corresponde a los valores de Y obtenidos utilizando el nuevo modelo. (Gráfica obtenida utilizando S-Plus)



Gráfica 5.2.6 Gráfica de dispersión, donde • corresponde a los promedios de los valores del vector Y asociados a cada valor de X y × corresponde a los valores de Y obtenidos utilizando el nuevo modelo. (Gráfica obtenida utilizando S-Plus)

5.2.4 Conclusiones.

El método de selección de transformaciones basado en los criterios L y M es muy útil y versátil, ya que puede modificarse para utilizarse para diferentes fines. La ventaja de utilizar este método es que pueden escogerse valores de α, diferentes al óptimo elegido con los criterios, que se adapten a las necesidades del investigador, teniéndose la seguridad de que la nueva transformación será tan buena como la original, gracias al número de calibración.

Como se mencionó al inicio de la sección 5.1, un buen procedimiento de selección de variables debe señalar cuáles son los mejores modelos para así poder escoger entre ellos el mejor para cada propósito. En el capítulo de Aplicaciones se pudo notar que la ventaja de los criterios L_m y M_m sobre otros es que cumplen con esa propiedad ya que, gracias al número de calibración, se tienen varios modelos, que se sabe son tan buenos como el "mejor", entre los cuales se puede elegir el que más convenga.

La ventaja del criterio K_m es que se puede utilizar de diferentes maneras, dependiendo del tipo de modelos que se desee obtener. En los ejemplos se observó que al utilizar el modelo completo como referencia, los mejores modelos son distintos a los elegidos si se utiliza como referencia al modelo que contiene sólo la intersección. Además, dado que este criterio de alguna forma mide la distancia que existe entre 2 modelos y se interpreta como la información que se pierde o se gana al utilizar uno u otro, también cumple con la propiedad mencionada en el párrafo anterior.

Estos criterios surgen de la filosofía predictiva y la noción simple de un experimento repetido. Permiten la introducción de información inicial, aunque ese caso no se ilustró en las aplicaciones, debido a que se requiere de un mayor conocimiento del problema particular. Los criterios tratados aquí son bastante generales y, en principio, pueden aplicarse en varias situaciones donde se requiera de selección de modelos. Aquí se habló de selección de variables y de transformaciones, sin embargo, también podrían utilizarse, por ejemplo para seleccionar una función varianza, para elegir una transformación que aproxime el modelo a cierta distribución; y no solamente en Regresión Lineal, sino para la clase de modelos lineales generalizados. Tal vez podría cambiarse el método de obtención de los criterios para así obtener otros diferentes, basados en los mismos

principios. Por otro lado, para poder aplicar estos criterios no se necesitan grandes conocimientos de Estadística, ya que se puede explicar de forma sencilla en qué se basa cada uno y cuál es la interpretación del resultado.

Los algoritmos utilizados para obtener los resultados se programaron en S-Plus y los problemas que se tuvieron fueron principalmente de eficiencia, por ejemplo, al aplicar los criterios a una cantidad grande de datos o al querer transformar más de 2 variables explicativas.

Por último, se recomienda un mayor estudio sobre la robustez de estos criterios ante los supuestos de regresión, tales como la normalidad de los errores y la homocedasticidad y ante la presencia de outliers o puntos influyentes, u otros problemas similares.

APÉNDICE A

Datos

1. Estudio del desempeño de un supervisor.

Y	X_1	X_2	X_3	X_4	X_5	X_6
43	51	30	39	61	92	45
63	64	51	54	63	73	47
71	70	68	69	76	86	48
61	63	45	47	54	84	35
81	78	56	66	71	83	47
43	55	49	44	54	49	34
58	67	42	56	66	68	35
71	75	50	55	70	66	41
72	82	72	67	71	83	31
67	61	45	47	62	80	41
64	53	53	58	58	67	34
67	60	47	39	59	74	41
69	62	57	42	55	63	25
68	83	83	45	59	77	35
77	77	54	72	79	77	46
81	90	50	72	60	54	36
74	85	64	69	79	79	63
65	60	65	75	55	80	60
65	70	46	57	75	85	46
50	58	68	54	64	78	52
50	40	33	34	43	64	33
64	61	52	62	66	80	41
53	66	52	50	63	80	37
40	37	42	58	50	57	49
63	54	42	48	66	75	33
66	77	66	63	88	76	72
78	75	58	74	80	78	49
48	57	44	45	51	83	38
85	85	71	71	77	74	55
82	82	39	59	64	78	39

2. Datos generados.

Para el Ejercicio 1.

X_1	X_2	ERRORES	Y
9.64	27.2	-11.42614	707.59837
7.77	16.7	22.77051	197.20248
4.75	34.0	-7.87900	829.59992
8.98	35.0	3.80883	764.51200
9.99	25.5	-15.60727	702.18161
10.18	23.6	-6.62934	588.37048
4.34	8.3	8.75636	112.91768
6.07	32.1	9.83045	648.12315
5.31	20.4	-2.64906	479.27420
5.74	61.1	5.89181	1339.17054
5.57	42.6	48.67365	580.66778
8.46	7.4	27.41447	-50.46304
2.82	33.7	-15.97561	884.63198
7.47	52.2	3.30073	1158.65807
5.48	39.0	29.56188	649.18566
5.75	36.7	15.21246	710.14264
7.32	26.1	26.38177	382.34217
4.46	39.8	-3.59390	927.61653
6.01	38.4	5.19323	828.05400
4.49	21.3	-1.13419	487.01207
10.63	23.7	-10.42979	621.04132
7.94	39.0	-4.86254	922.87575
5.2	32.3	6.23316	680.15031
5.09	39.3	-17.72954	1028.23604
6.28	22.3	4.81661	464.68620
11.55	14.4	-4.72041	365.21131
5.49	42.7	21.85954	794.13331
3.35	31.7	-0.76174	719.75080
7.49	18.0	29.97868	169.73093
7.23	47.4	5.70268	1030.19468
7.26	24.4	-17.49420	689.27116
5.76	29.3	26.37664	453.68821
4.48	33.8	22.49159	585.48586
4.55	19.5	6.74876	383.97532
8.49	26.3	-24.78995	791.24838
10.61	48.3	0.66941	1093.72527
5.05	36.5	-2.54643	844.81083
4.52	20.9	0.10843	468.14320
4.64	6.9	-15.06625	269.03742
6.18	9.3	8.16912	142.15507
5.68	37.1	10.14006	759.14626
4.09	31.1	-24.40004	893.07477
4.8	38.7	6.92086	820.06315
9.86	9.6	-13.11312	320.34906
6.58	43.7	5.18482	949.37398
4.2	48.6	-3.52495	1127.19382
4.38	31.4	17.40474	570.81593
4.43	33.5	0.53544	751.59728
4.16	41.8	-2.93301	967.65038
3.94	24.2	-4.75541	581.02815

X_1	X_2	ERRORES	Y
9.18	21.7	-11.91733	585.77061
3.78	45.7	-9.26925	1105.99867
3.83	40.8	-5.76057	966.82829
8.74	19.2	11.70387	342.29287
7.91	19.8	4.78667	409.62583
5.33	21.1	0.98241	466.62122
5.35	21.8	12.15613	394.54285
1.15	26.3	25.47778	387.84884
6.06	18.0	-39.75457	717.72944
9.43	32.3	23.51863	548.18801
8.65	41.4	-18.91058	1088.91956
5.37	30.2	9.02824	610.47979
6.96	50.4	14.19173	1031.35096
5.37	30.9	-8.87720	767.49611
5.94	30.3	-9.54708	759.68182
1.7	53.0	-18.57388	1343.45456
4.64	27.3	12.93536	512.93104
3.38	8.9	7.30421	137.06149
9.2	31.9	-9.10144	795.86344
4.85	31.8	1.72482	703.93624
6.38	32.6	11.92788	643.29273
5.46	31.5	2.47269	691.82264
6.25	21.5	6.62818	432.16642
4.03	27.4	-31.12811	861.77430
3.91	30.6	5.12149	648.90925
5.42	20.9	-18.78269	617.88639
4.88	10.3	-8.27491	293.18820
5.12	13.1	23.75997	104.78268
5.73	33.2	-4.02340	781.98523
2.12	26.5	17.97467	452.48995
4.91	22.1	3.92470	465.78946
1.03	24.7	-27.15974	766.02718
4.62	33.1	26.39942	538.89667
7.33	12.3	16.80609	143.56594
5.4	23.5	-4.34336	563.30220
2.75	38.2	4.39689	826.51424
2.97	34.3	23.30841	588.92540
8.72	38.5	-3.62410	902.51280
8.74	13.8	-21.59684	481.70846
4.91	42.4	-12.28661	1055.75978
3.28	39.6	-13.75007	1001.90291
4.82	28.1	16.71210	501.57055
5.74	44.2	9.37415	926.91147
7.58	13.8	-23.87853	498.52592
8.21	20.6	-9.02502	536.96452
3.31	33.1	-5.43491	788.40947
6.11	29.8	6.58084	621.39462
6.73	32.2	19.49275	574.93088
7.58	24.5	1.38255	543.13830
3.91	30.6	-13.35659	794.49811

Para el ejercicio 2.

X_1	X_2	ERRORES	Y
9.64	27.2	4.41602165	66.8668086
7.77	16.7	0.03268507	40.1408563
4.75	34.0	2.1220471	78.8429243
8.98	35.0	-0.8121699	78.1690122
9.99	25.5	2.24367123	61.0984178
10.18	23.6	-4.8964223	49.9326449
4.34	8.3	-2.0428283	20.1280375
6.07	32.1	-0.2552971	72.4768349
5.31	20.4	-3.9198024	43.9595379
5.74	61.1	-0.4201403	133.840572
5.57	42.6	4.11643422	99.1137183
8.46	7.4	2.89985991	23.2955458
2.82	33.7	-4.6515395	71.3696414
7.47	52.2	-4.0196185	111.411419
5.48	39.0	0.7869755	88.1419484
5.75	36.7	0.99193265	83.4750221
7.32	26.1	3.09987871	63.1405845
4.46	39.8	2.56084105	91.5801128
6.01	38.4	7.63887313	93.7379408
4.49	21.3	4.71989472	54.4822756
10.63	23.7	-4.6585117	50.3974685
7.94	39.0	0.51250511	87.9478832
5.2	32.3	1.81236885	74.9404743
5.09	39.3	0.92694063	88.9057804
6.28	22.3	1.66239779	53.6053321
11.55	14.4	-1.8083011	33.5427114
5.49	42.7	3.78224511	98.9891191
3.35	31.7	-3.7123085	68.0821013
7.49	18.0	3.03387313	45.8915538
7.23	47.4	-0.4522114	104.785156
7.26	24.4	2.55524746	58.986512
5.76	29.3	0.87308483	67.6533526
4.48	33.8	4.96824204	81.2558849
4.55	19.5	-0.7818586	45.1627986
8.49	26.3	-8.6266797	51.876677
10.61	48.3	-2.620791	104.636894
5.05	36.5	-2.1369771	79.8988234
4.52	20.9	-2.7243075	46.1902351
4.64	6.9	-2.8139147	16.3958906
6.18	9.3	1.92797663	26.2810301
5.68	37.1	-3.8887831	79.4408372
4.09	31.1	0.87882539	71.4241939
4.8	38.7	-2.3191114	84.3770215
9.86	9.6	5.42668224	30.5366308
6.58	43.7	2.47273802	99.8372858
4.2	48.6	-2.3528001	105.331988
4.38	31.4	4.34734018	75.5388015
4.43	33.5	5.62978137	81.2791758
4.16	41.8	0.20773085	93.4612913
3.94	24.2	-2.4463694	53.4519713

X_1	X_2	ERRORES	Y
9.18	21.7	2.58626477	7.00228642
3.78	45.7	-7.2857505	-2.8697288
3.83	40.8	5.18568681	9.60170846
8.74	19.2	-1.4643902	2.95163148
7.91	19.8	-0.1009174	4.31510425
5.33	21.1	3.91917783	8.33519948
5.35	21.8	2.65115764	7.0671793
1.15	26.3	8.91180333	13.327825
6.06	18.0	0.66931308	5.08533473
9.43	32.3	-1.7354723	2.68054936
8.65	41.4	-3.5087408	0.90728082
5.37	30.2	-2.4685212	1.94750042
6.96	50.4	-2.8052746	1.61074701
5.37	30.9	-1.4336023	2.98241938
5.94	30.3	3.22492272	7.64094437
1.7	53.0	-5.6924729	-1.2764512
4.64	27.3	-6.431931	-2.0159094
3.38	8.9	-4.8965345	-0.4805128
9.2	31.9	-1.8407241	2.57529757
4.85	31.8	5.55559789	9.97161954
6.38	32.6	-9.0322454	-4.6162237
5.46	31.5	1.02225032	5.43827197
6.25	21.5	-6.6690541	-2.2530324
4.03	27.4	-4.3735824	0.04243924
3.91	30.6	-4.3458473	0.07017433
5.42	20.9	-0.8432393	3.57278235
4.88	10.3	0.78086478	5.19688643
5.12	13.1	2.7718866	7.18790825
5.73	33.2	0.94553244	5.36155409
2.12	26.5	-2.1753705	2.24065113
4.91	22.1	0.18399521	4.60001687
1.03	24.7	-2.5008077	1.91521393
4.62	33.1	-2.3208667	2.09515498
7.33	12.3	1.25806134	5.67408299
5.4	23.5	2.0154088	6.43143045
2.75	38.2	-3.6040941	0.81192754
2.97	34.3	0.18719116	4.60321281
8.72	38.5	1.99012151	6.40614316
8.74	13.8	-2.1327766	2.28324505
4.91	42.4	-0.2949875	4.12103412
3.28	39.6	-0.7607392	3.65528245
4.82	28.1	3.90433539	8.32035705
5.74	44.2	4.22182677	8.63784843
7.58	13.8	-4.589422	-0.1734003
8.21	20.6	-4.3777691	0.03825255
3.31	33.1	-0.800587	3.61543467
6.11	29.8	0.88526088	5.30128253
6.73	32.2	-0.2375142	4.17850748
7.58	24.5	0.08395978	4.49998144
3.91	30.6	-5.1594608	-0.7434391

3. Tiempo de crecimiento de una colonia de bacterias.

X_1	Y
1500000	6.14
1500000	5.90
1500000	6.33
1500000	5.98
1500000	5.73
1500000	6.33
750000	6.63
750000	6.83
750000	6.63
750000	6.83
150000	8.95
150000	8.16
150000	8.43
150000	8.32
15000	10.63
15000	10.00
15000	10.96
15000	9.66
1500	12.63
1500	11.83
1500	12.63
1500	12.33
150	14.86
150	15.47
150	13.67
150	14.86
150	14.14
150	13.67
75	16.19
75	14.47
75	14.67
75	14.86
75	14.80
75	14.17
15	16.53
15	16.80
15	17.11
15	16.53
15	15.64
15	17.28

APÉNDICE B

Programas

Programa 1.
Calcula el criterio L para 2 variables transformadas.

```
calculm_function(oalfa1, oalfa2)
{
X1_matrix(scan(file="c:/sptemp/ox1.out"),ncol=1)
X2_matrix(scan(file="c:/sptemp/ox2.out"),ncol=1)
X_cbind(X1,X2)
Y_matrix(scan(file="c:/sptemp/oy.out"),ncol=1)
n_nrow(Y)
km_3
library(Matrix)
unos_matrix(1,nrow=n,ncol=1)
Xtrans_matrix(0,nrow=n,ncol=2)
if(oalfa2!=0)
   {
   for(k in 1:n)
   Xtrans[k,2]_((X[k,2]^oalfa2)-1)/oalfa2
   }
else
   Xtrans[,2]_log(X[,2])
if(oalfa1!=0)
   {
   for(k in 1:n)
   Xtrans[k,1]_((X[k,1]^oalfa1)-1)/oalfa1
   }
else
   Xtrans[,1]_log(X[,1])
   Xm_cbind(unos,Xtrans)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
   oqm_t(Y)%*%(diag(n)-Pm)%*%Y
   oLm_(2*(n-1)*(1/(n-km-2))*oqm)^(1/2)
   oMm_Mm_2*(pi^(1/2))*((gamma((n-km)/2)/gamma(n-
(km/2)))^(1/n))*(oqm^(1/2))
return(oMm,oLm)
}
```

Programa 2.
Calcula la probabilidad final para todos los modelos (con distribución inicial no informativa).

```
calprobasd_function(k)
{
X_matrix(scan(file="c:/sptemp/atdxi.txt"),ncol=k,byrow=F)
r_2^k
n_nrow(X)
Y_matrix(scan(file="c:/sptemp/asdy.txt"),ncol=1)
dimnames(X)_list(NULL,paste("X",1:k))
n0_matrix(1,ncol=1,nrow=n)

#Aqui voy a obtener la matriz de todos los subconjuntos de un conjunto de
k variables.
source("c:/sptemp/matbin.ssc")
matbin(k)
M_matrix(scan(file="c:/sptemp/mat.out"),nrow=r,byrow=F)
```



```
dimnames(M)_list(NULL,paste("X",1:k))

Mnumi_matrix(0,nrow=r,ncol=1)
Mprobi_matrix(0,nrow=r,ncol=1)
Mprobf_matrix(0,nrow=r,ncol=1)
suml_0
sumprue_0
alfa_1_d_0
b_1
for(i in 1:r)
   {
   km_1
   for(h in 1:k)
      {
      if(M[i,h]!=0)
      km_km+1
      }
   Xm_matrix(0,nrow=n,ncol=km)
   Xm[,1]_1
   aux_1
   for(j in 1:k)
      {
      if(M[i,j]==1)
         {
         aux_aux+1
         Xm[,aux]_X[,j]
         }
      }
   Xmt_t(Xm)
   Pm_Xm%*%(solve(Xmt%*%Xm))%*%Xmt
   q0_t(n0)%*%(diag(n)-Pm)%*%n0
gammam_b*(alfa^(1/km))
lm_1/(n-km)
Mnumi[i]_exp(-km/2)
suml_suml + Mnumi[i]
   }
Mprobi_Mnumi/suml
dimnames(Mprobi)_list(NULL,"p(m)")
Mprobi_round(Mprobi,2)
for(i in 1:r)
   {
   km_1
   for(h in 1:k)
      {
      if(M[i,h]!=0)
      km_km+1
      }
   Xm_matrix(0,nrow=n,ncol=km)
   Xm[,1]_1
   aux_1
   for(j in 1:k)
      {
      if(M[i,j]==1)
         {
         aux_aux+1
         Xm[,aux]_X[,j]
```

```
        }
      }
    Xmt_t(Xm)
    Pm_Xm%*%(solve(Xmt%*%Xm))%*%Xmt
    q0_t(n0)%*%(diag(n)-Pm)%*%n0
gammam_b*(alfa^(1/km))
lm_1/(n-km)
Mprobf[i]_Mprobi[i]*exp(-km/2)*(t(Y)%*%(diag(n)-Pm)%*%Y)^(-n/2)
sumprue_sumprue + Mprobf[i]
}
Mprobf_Mprobf/sumprue

dimnames(Mprobf)_list(NULL,"p(m|y)")
Mprobf_round(Mprobf,2)
Matprob_cbind(M,Mprobi,Mprobf)
Probord_Matprob[order(Matprob[,7]),1:8]
write.table(Probord,"c:/sptemp/probfin.out",sep= )
return(Probord)
}
```

Programa 3.
Calcula criterios y números de calibracion, para distribución inicial normal-gamma.

```
caltodo_function(k)
{
X_matrix(scan(file="c:/sptemp/hald1.txt"),ncol=k,byrow=F)
r_2^k
n_nrow(X)
n0_matrix(c(79,77,104,90,99,108,105,73,93,111,88,115,113),ncol=1)
Y_matrix(scan(file="c:/sptemp/hald2.txt"),ncol=1)
dimnames(X)_list(NULL,paste("X",1:k))

#Aqui voy a obtener la matriz de subconjuntos
source("c:/sptemp/matbin.ssc")
matbin(k)
M_matrix(scan(file="c:/sptemp/mat.out"),nrow=r,byrow=F)
dimnames(M)_list(NULL,paste("X",1:k))

Mlmng_matrix(0,nrow=r,ncol=1)
Mmmng_matrix(0,nrow=r,ncol=1)
Mkmng_matrix(0,nrow=r,ncol=1)
Mcp_matrix(0,nrow=r,ncol=1)
Mparkm_matrix(0,nrow=r,ncol=1)
Maic_matrix(0,nrow=r,ncol=1)
Mbic_matrix(0,nrow=r,ncol=1)
unos_matrix(1,nrow=n,ncol=1)
Xm0_cbind(unos,X)
Xm0t_t(Xm0)
Pm0_Xm0%*%(solve(Xm0t%*%Xm0))%*%Xm0t
qm0_(t(Y)%*%(diag(n)-Pm0))%*%Y
pm0_(t(Y-n0)%*%Pm0)%*%(Y-n0)
km0_5
sse_t(Y-Pm0%*%Y)%*%(Y-(Pm0%*%Y))
sigmaest_sse/(n-km0)
```



```
#Aqui empieza a calcular los criterios
for(i in 1:r)
   {
   km_1
   for(h in 1:k)
      {
      if(M[i,h]!=0)
      km_km+1
      }
   Xm_matrix(0,nrow=n,ncol=km)
   Xm[,1]_1
   aux_1
   for(j in 1:k)
      {
      if(M[i,j]==1)
         {
         aux_aux+1
         Xm[,aux]_X[,j]
         }
      }
   g_0.1
   d0_25
   g0_125
   Xmt_t(Xm)
   Pm_Xm%*%(solve(Xmt%*%Xm))%*%Xmt
   nm_Pm%*%((g*n0)+((1-g)*Y))
   qm_(t(Y)%*%(diag(n)-Pm))%*%Y
   pm_(t(Y-n0)%*%Pm)%*%(Y-n0)
   sm2_((n+d0)^(-1))*(qm+(g*pm)+g0)
   clm_(n+((1-g)*km))/(n+d0-2)
   am_qm+(g*pm)+g0
   bm_qm+(((g^2)/(2-g))*pm)
   Lmng_(((1+clm)*qm)+(g*(g+clm)*pm)+(clm*g0))^(1/2)
   Mlmng[i,1]_Lmng
   dimnames(Mlmng)_list(NULL,paste("Lm"))
   Mmng_(pi^(1/2))*(((gamma((n+d0)/2)/gamma(n+(d0/2)))*((2-
g)^(km/2)))^(1/n))*(am^(1/2))*((1+(bm/am))^(1+(d0/(2*n))))
   Mmmng[i,1]_Mmng
   dimnames(Mmmng)_list(NULL,paste("Mm"))
#m0 es el modelo completo
   dm_dm0_d0
   v_((n+dm)*(n+dm0-2))/((n+dm0)*(n+dm-2))
   sm02_((n+d0)^(-1))*(qm0+(g*pm0)+g0)
   cosa_((n+dm0-2)/(2*sm02*(n+dm0)*(2-g)))+((n+dm-2)/(2*sm2*(n-km)))
   nmm0_(Pm0-Pm)%*%((g*n0)+((1-g)*Y))
   Kmng_(cosa*t(nmm0)%*%nmm0)+((n/2)*(((v*sm2)/(sm02))+((sm02)/(v*sm2
2))+(((km0-km)/2)*((((1-g)*sm02)/(v*sm2))-(((1-g)*v*sm2)/((2-g)*sm02)
   Mkmng[i,1]_Kmng
   dimnames(Mkmng)_list(NULL,paste("Km"))
   #Aqui voy a calcular los otros criterios
#Cp
   ssekm_t(Y-(Pm%*%Y))%*%(Y-(Pm%*%Y))
   sigestkm_ssekm/n
   Ccp_(ssekm/sigmaest)+((2*km)-n)
   Mcp[i,1]_Ccp
   dimnames(Mcp)_list(NULL,paste("Cp"))
```

```
    Mparkm[i,1]_km
    dimnames(Mparkm)_list(NULL,paste("p"))
#Aic
    Lveros_((1/((2*pi)*sigestkm))^(n/2))*(exp(-(1/(2*sigestkm))*ssekm))
    Aic_(-2*log(Lveros))+(2*km)
    Maic[i,1]_Aic
    dimnames(Maic)_list(NULL,paste("AIC"))
#Bic
    Bic_(-2*log(Lveros))+(km*log(n))
    Mbic[i,1]_Bic
    dimnames(Mbic)_list(NULL,paste("BIC"))
    }

#Ahora para calcular los números de calibración
minlmng_min(Mlmng)
pmlmng_match(minlmng,Mlmng)
km_1
for(h in 1:k)
    {
    if(M[pmlmng,h]!=0)
    km_km+1
    }
Xlmminng_matrix(0,nrow=n,ncol=km)
Xlmminng[,1]_1
aux_1
for(j in 1:k)
    {
    if(M[pmlmng,j]==1)
    aux_aux+1
    Xlmminng[,aux]_X[,j]
    }
g_0.1
d0_25
g0_125
Xlmmingt_t(Xlmminng)
Pmmin_Xlmminng%*%(solve(Xlmmingt%*%Xlmminng))%*%Xlmmingt
nmmin_Pmmin%*%((g*n0)+((1-g)*Y))
qmmin_t(Y)%*%(diag(n)-Pmmin)%*%Y
pmmin_t(Y-n0)%*%Pmmin%*%(Y-n0)
clm_(n+((1-g)*km))/(n+d0-2)

#Voy a generar la t que necesito
matvar_g0*(1/d0)*(diag(n)+((1/g)*(1-g)*Pmmin))
source("c:/sptemp/genmult1.ssc")
genmult1(6000,t(nmmin),matvar,d0)
tegen_matrix(scan(file="c:/sptemp/te.out"),nrow=6000,byrow=F)
Mlmminng_matrix(0,nrow=6000,ncol=1)
Mmmminng_matrix(0,nrow=6000,ncol=1)
for(i in 1:6000)
    {
    Y_tegen[i,]
    pmmin_(t(Y-n0)%*%Pmmin)%*%(Y-n0)
    qmmin_t(Y)%*%(diag(n)-Pmmin)%*%Y
    Lmminng_(((1+clm)*qmmin)+(g*(g+clm)*pmmin)+(clm*g0))^(1/2)
    Mlmminng[i,1]_Lmminng
    ammin_qmmin+(g*pmmin)+g0
```



```
    bmmin_qmmin+(((q^2)/(2-g))*pmmin)
    Mmminng_(pi^(1/2))*(((gamma((n+d0)/2)/gamma(n+(d0/2)))*((2-
g)^(km/2)))^(1/n))*(ammin^(1/2))*((1+(bmmin/ammin))^(1+(d0/(2*n))))
    Mmmminng[i,1]_Mmminng
    }
varlmng_var(Mlmminng)
ncalsl_sqrt(varlmng)
varmmng_var(Mmmminng)
ncalsm_sqrt(varmmng)

#Prepara las matrices de los criterios y números de calibración
Matcrit_cbind(M,Mlmng,Mmmng,Mkmng,Mcp,Mparkm,Maic,Mbic)
Critord_Matcrit[order(Matcrit[,5]),1:11]
numscal_cbind(ncalsl,ncalsm)
dimnames(numscal)_list(NULL,c("Sl","Sm"))
write.table(Critord,"c:/sptemp/crits.out",sep= )
write.table(numscal,"c:/sptemp/numscal.out",sep= )
return(Critord,numscal)
}
```

Programa 4.
Calcula criterios y números de calibración, para distribución inicial no informativa.

```
critasd_function(k)
{
X_matrix(scan(file="c:/sptemp/atdxi.txt"),ncol=k,byrow=F)
r_2^k
n_nrow(X)
Y_matrix(scan(file="c:/sptemp/asdy.txt"),ncol=1)
dimnames(X)_list(NULL,paste("X",1:k))

#Aqui voy a obtener la matriz de subconjuntos
source("c:/sptemp/matbin.ssc")
matbin(k)
M_matrix(scan(file="c:/sptemp/mat.out"),nrow=r,byrow=F)
dimnames(M)_list(NULL,paste("X",1:k))

Malm_matrix(0,nrow=r,ncol=1)
Mamm_matrix(0,nrow=r,ncol=1)
Makm_matrix(0,nrow=r,ncol=1)
Makml_matrix(0,nrow=r,ncol=1)
Mcp_matrix(0,nrow=r,ncol=1)
Mparkm_matrix(0,nrow=r,ncol=1)
Maic_matrix(0,nrow=r,ncol=1)
Mbic_matrix(0,nrow=r,ncol=1)
unos_matrix(1,nrow=n,ncol=1)
Xm0_cbind(unos,X)
Xm0t_t(Xm0)
Pm0_Xm0%*%(solve(Xm0t%*%Xm0))%*%Xm0t
qm0_(t(Y)%*%(diag(n)-Pm0))%*%Y
km0_k+1
Xml_unos
Xmlt_t(Xml)
Pml_Xml%*%(solve(Xmlt%*%Xml))%*%Xmlt
qml_(t(Y)%*%(diag(n)-Pml))%*%Y
```

```
kml_1
sse_t(Y-(Pm0%*%Y))%*%(Y-(Pm0%*%Y))
sigmaest_sse/(n-km0)

#Aqui empieza a calcular los criterios
for(i in 1:r)
   {
   km_1
   for(h in 1:k)
      {
      if(M[i,h]!=0)
      km_km+1
      }
   Xm_matrix(0,nrow=n,ncol=km)
   Xm[,1]_1
   aux_1
   for(j in 1:k)
      {
      if(M[i,j]==1)
         {
         aux_aux+1
         Xm[,aux]_X[,j]
         }
      }
   g_0
   d0_-km
   g0_0
   Xmt_t(Xm)
   Pm_Xm%*%(solve(Xmt%*%Xm))%*%Xmt
   qm_(t(Y)%*%(diag(n)-Pm))%*%Y
   sm2_((n+d0)^(-1))*(qm)
#Criterio L
   Crlm_(2*(n-1)*((n-km-2)^(-1))*qm)^(1/2)
   Malm[i,1]_Crlm
   dimnames(Malm)_list(NULL,paste("Lm"))
#Criterio M
   Cmm_(pi^(1/2))*((gamma((n+d0)/2)/gamma(n+(d0/2)))^(1/n))*(qm^(1/2))*2
   Mamm[i,1]_Cmm
   dimnames(Mamm)_list(NULL,paste("Mm"))
#Criterio K
#m0 es el modelo completo
#m1 es el modelo con sólo la intersección
   dm_-km
   dm0_-km0
   v_((n+dm)*(n+dm0-2))/((n+dm0)*(n+dm-2))
   sm02_((n+d0)^(-1))*(qm0)
   cosa_((n+dm0-2)/(2*sm02*(n+dm0)*(2-g)))+((n+dm-2)/(2*sm2*(n+dm)))
   nmm0_(Pm0-Pm)%*%Y
   Ckm_(cosa*t(nmm0)%*%nmm0)+((n/2)*(((v*sm2)/(sm02))+((sm02)/(v*sm2))-
2))+(((km0-km)/2)*((((1-g)*sm02)/(v*sm2))-(((1-g)*v*sm2)/((2-g)*sm02))))
   Makm[i,1]_Ckm
   dimnames(Makm)_list(NULL,paste("Km"))
   dm_-km
   dml_-kml
   vl_((n+dm)*(n+dml-2))/((n+dml)*(n+dm-2))
   sml2_((n+d0)^(-1))*(qml)
```



```
   cosal_((n+dml-2)/(2*sml2*(n+dml)*(2-g)))+((n+dm-2)/(2*sm2*(n+dm)
   nmml_(Pml-Pm)%*%Y
   Ckml_(cosal*t(nmml)%*%nmml)+((n/2)*(((vl*sm2)/(sml2))+((sml2)/(vl*sm2
)-2))+(((kml-km)/2)*((((1-g)*sml2)/(vl*sm2))-(((1-g)*vl*sm2)/((2-
g)*sml2))))
   Makml[i,1]_Ckml
   dimnames(Makml)_list(NULL,paste("Kml"))
   #Aqui voy a calcular los otros criterios
#Cp
   ssekm_t(Y-(Pm%*%Y))%*%(Y-(Pm%*%Y))
   sigestkm_ssekm/n
   Ccp_(ssekm/sigmaest)+((2*km)-n)
   Mcp[i,1]_Ccp
   dimnames(Mcp)_list(NULL,paste("Cp"))
   Mparkm[i,1]_km
   dimnames(Mparkm)_list(NULL,paste("p"))
#Aic
   Lveros_((1/((2*pi)*sigestkm))^(n/2))*(exp(-(1/(2*sigestkm))*ssekm
   Aic_(-2*log(Lveros))+(2*km)
   Maic[i,1]_Aic
   dimnames(Maic)_list(NULL,paste("AIC"))
#Bic
   Bic_(-2*log(Lveros))+(km*log(n))
   Mbic[i,1]_Bic
   dimnames(Mbic)_list(NULL,paste("BIC"))
   }

#Ahora para calcular los números de calibración
minlm_min(Malm)
pmlm_match(minlm,Malm)
km_1
for(h in 1:k)
   {
   if(M[pmlm,h]!=0)
   km_km+1
   }
Xlmmin_matrix(0,nrow=n,ncol=km)
Xlmmin[,1]_1
aux_1
for(j in 1:k)
   {
   if(M[pmlm,j]==1)
   aux_aux+1
   Xlmmin[,aux]_X[,j]
   }
g_0
d0_-km
g0_0
Xlmmint_t(Xlmmin)
Pmmin_Xlmmin%*%(solve(Xlmmint%*%Xlmmin))%*%Xlmmint
qmmin_t(Y)%*%(diag(n)-Pmmin)%*%Y
taomoda_(n-km-2)/qmmin
#Num. calibración D1
numcaldl_((taomoda^(-1/2))/2)*((1-(km/n))^(1/2))*((1-(1/n)*(1+(1/32)*(1-
(km/n))*(1-(2/n))))^(1/2)
minmm_min(Mamm)
```

```
pmmm_match(minmm,Mamm)
km_1
for(h in 1:k)
    {
    if(M[pmmm,h]!=0)
    km_km+1
    }
Xlmmin_matrix(0,nrow=n,ncol=km)
Xlmmin[,1]_1
aux_1
for(j in 1:k)
    {
    if(M[pmmm,j]==1)
    aux_aux+1
    Xlmmin[,aux]_X[,j]
    }
g_0
d0_-km
g0_0
Xlmmint_t(Xlmmin)
Pmmin_Xlmmin%*%(solve(Xlmmint%*%Xlmmin))%*%Xlmmint
qmmin_t(Y)%*%(diag(n)-Pmmin)%*%Y
#Num. calibración Dm
taomoda_(n-km-2)/qmmin
numcaldm_(2^(km/(2*n)))*((2*pi)^(1/2))*(taomoda^(-1/2))*(((n/(n-2))^((n-
km)/2))-((n/(n-1))^(n-km)))^(1/2))

Matcrit_cbind(M,Malm,Mamm,Makm,Makml,Mcp,Mparkm,Maic,Mbic)
Critord_Matcrit[order(Matcrit[,7]),1:14]
numcald_cbind(numcaldl,numcaldm)
dimnames(numcald)_list(NULL,c("Dl","Dm"))
Mparkmc_matrix(0,nrow=32,ncol=1)
Mcpc_matrix(0,nrow=32,ncol=1)
for(i in 1:32)
    {
    Mparkmc[i,1]_Critord[i,12]
    Mcpc[i,1]_Critord[i,11]
    }

#Gráfica Cp
win.graph()
plot(Mparkmc,Mcpc,xlab="p",ylab="Cp",type="p")
abline(0,1)

#Tabla stepwise
efroy_stepwise(X,Y,intercept="T",method="efroymson",plot=F)

write.table(Critord,"c:/sptemp/critasd.out",sep=" ")
write.table(numcald,"c:/sptemp/numcalas.out",sep=" ")
return(numcald,efroy)
}
```

Programa 5.
Para generar una t multivariada.

```
genmultl_function(nl,mu,lambda,gdel)
{
alfa_beta_gdel/2
a_length(mu)
te_matrix(0,nrow=nl,ncol=a)
#invlamb_solve(lambda)
gengam_rgamma(nl,alfa)
datos_gengam/beta
for(i in 1:nl)
        {
        te[i,]_rmvnorm(1,mean=mu,cov=(1/datos[i])*lambda)
        }
write(te,"c:/sptemp/te.out",ncol=1)
}
```

Programa 6.
Genera las Y's con los datos del programa xbyerror, para alguna transformación de X. (Para ser utilizados en la sección 5.2.1)

```
geny_function(a1,a2)
{
   X1_matrix(scan(file="c:/sptemp/ox1.out"),ncol=1)
   X2_matrix(scan(file="c:/sptemp/ox2.out"),ncol=1)
   B_matrix(scan(file="c:/sptemp/ob.out"),ncol=1)
   error_matrix(scan(file="c:/sptemp/oerror.out"),ncol=1)
   unos_matrix(1,nrow=100,ncol=1)
   if(a1!=0)
      X1a1_((X1^a1)-1)/a1
   else
      X1a1_log(X1)
   if(a2!=0)
      X2a2_((X2^a2)-1)/a2
   else
      X2a2_log(X2)
   Xa_cbind(unos,X1a1,X2a2)
   Y_(Xa%*%B)+error
write(Y,"c:/sptemp/oy.out",ncol=1)
return(Y,a1,a2)
}
```

Programa 7.
Para obtener la matriz de todos los subconjuntos de un conjunto de k variables.

```
matbin_function(k)
{
r_2^k
mat_matrix(0,nrow=r,ncol=k)
for(h in 2:r)
    {
   n_h-1
   i_1
   j_k+1-i
   num_n
   while(num!=1)
      {
      n_n/2
```

```
        num_floor(n)
        if(n!=num)
           mat[h,j]_1
        n_num
        i_i+1
        j_k+1-i
        }
    mat[h,j]_1
    }
write(mat,"c:/sptemp/mat.out",ncol=1)
}
```

Programa 8.
Calcula los valores de los criterios L y M para un modelo con una variable transformada.

```
olmbact_function(oalfa2)
{
X1_matrix(scan(file="c:/sptemp/aerugx.txt"),ncol=1)
Y_matrix(scan(file="c:/sptemp/aerugy.txt"),ncol=1)
n_nrow(Y)
km_3
library(Matrix)
unos_matrix(1,nrow=n,ncol=1)
Xtrans_matrix(0,nrow=n,ncol=1)
if(oalfa2!=0)
    {
    for(k in 1:n)
    Xtrans[k]_((X1[k]^oalfa2)-1)/oalfa2
    }
else
    Xtrans[,1]_log(X1[,1])
    Xm_cbind(unos,Xtrans)
    prod_t(Xm)%*%Xm
    Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
    oqm_t(Y)%*%(diag(n)-Pm)%*%Y
    oLm_(2*(n-1)*(1/(n-km-2))*oqm)^(1/2)
    oMm_2*(pi^(1/2))*((gamma((n-km)/2)/gamma(n-(km/2)))^(1/n))*(oqm^(1/2))
return(oqm,oLm,oMm)
}
```

Programa 9.
Calcula los criterios L y M para un polinomio en 2 potencias de X.

```
olmpoli_function(alfa1,alfa2)
{
X1_matrix(scan(file="c:/sptemp/aerugx.txt"),ncol=1)
X2_X1
X_cbind(X1,X2)
Y_matrix(scan(file="c:/sptemp/aerugy.txt"),ncol=1)
n_nrow(Y)
km_3
Xtrans_matrix(0,nrow=n,ncol=2)
library(Matrix)
unos_matrix(1,nrow=n,ncol=1)
    if(alfa1!=0)
```



```
        {
        for(k in 1:n)
           Xtrans[k,1]_((X[k,1]^alfa1)-1)/alfa1
        }
    else
        Xtrans[,1]_log(X[,1])
    if(alfa2!=0)
        {
        for(k in 1:n)
           Xtrans[k,2]_((X[k,2]^alfa2)-1)/alfa2
        }
    else
        Xtrans[,2]_log(X[,2])
    Xnueva_Xtrans
    Xm_cbind(unos,Xnueva)
    prod_t(Xm)%*%Xm
    Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
    qm_t(Y)%*%(diag(n)-Pm)%*%Y
Lm_(2*(n-1)*(1/(n-km-2))*qm)^(1/2)
Mm_2*(pi^(1/2))*((gamma((n-km)/2)/gamma(n-(km/2)))^(1/n))*(qm^(1/2))
return(qm,Lm,Mm)
}
```

Programa 10.
Grafica el polinomio en 2 potencias de X.

```
polinomio_function(alfa1,alfa2)
{
X1_matrix(scan(file="c:/sptemp/aerugx.txt"),ncol=1)
Y_matrix(scan(file="c:/sptemp/aerugy.txt"),ncol=1)
Xp_matrix(scan(file="c:/sptemp/xlbactp.txt"),ncol=1)
Yp_matrix(scan(file="c:/sptemp/aeruyxlp.txt"),ncol=1)
n_nrow(Y)
np_nrow(Yp)
library(Matrix)
X1t_((X1^alfa1)-1)/alfa1
X2t_((X1^alfa2)-1)/alfa2
unos_matrix(1,nrow=n,ncol=1)
X_cbind(unos,X1t,X2t)
prod_t(X)%*%X
B_solve.Matrix(prod)%*%t(X)%*%Y
Yest_X%*%B
logX_log(X1)
oX_cbind(unos,logX)
oprod_t(oX)%*%oX
oB_solve.Matrix(oprod)%*%t(oX)%*%Y
Yestlog_oX%*%oB

plot(X1,Y,ylim=c(6,18),ylab="Yest",xlab="X")
par(new=T)
plot(X1,Yest,type="p",pch=4,ylim=c(6,18),xlab=" ",ylab=" ",axes=F)
par(new=F)
}
```

Programa 11.
Elige el mejor valor de alfa para transformar una variable explicativa.

```
tranfbact_function(alfa)
{
X_matrix(scan(file="c:/sptemp/aerugx.txt"),ncol=1)
Y_matrix(scan(file="c:/sptemp/aerugy.txt"),ncol=1)
n_nrow(X)
km_ncol(X)+1
Mqm_matrix(0,nrow=41,ncol=1)
Malfa_matrix(0,nrow=41,ncol=1)
unos_matrix(1,nrow=n,ncol=1)
library(Matrix)

alfa_-2.1
i_0
while(alfa<2)
    {
    i_i+1
    alfa_alfa + .1
    alfa_round(alfa,3)
    if(alfa!=0)
        {
        Xtrans_((X^alfa)-1)/alfa
        }
    else
        Xtrans_log(X)
    Xm_cbind(unos,Xtrans)
    prod_t(Xm)%*%Xm
    Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
    qm_t(Y)%*%(diag(n)-Pm)%*%Y
    Mqm[i,1]_qm
    Malfa[i,1]_alfa
    }
MalfaLm_cbind(Malfa,Mqm)
minqm_min(MalfaLm[,2])
posic_match(minqm,MalfaLm[,2])
alfamin_MalfaLm[posic,1]

liminf_alfamin-.1
limsup_alfamin+.1
Mqm_matrix(0,nrow=21,ncol=1)
Malfa_matrix(0,nrow=21,ncol=1)
alfa_liminf-.01
i_0
while(alfa<limsup)
    {
    i_i+1
    alfa_alfa + .01
    alfa_round(alfa,3)
    if(alfa!=0)
        {
        Xtrans_((X^alfa)-1)/alfa
        }
    else
        Xtrans_log(X)
    Xm_cbind(unos,Xtrans)
    prod_t(Xm)%*%Xm
```



```
    Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
    qm_t(Y)%*%(diag(n)-Pm)%*%Y
    Mqm[i,1]_qm
    Malfa[i,1]_alfa
    }
MalfaLm_cbind(Malfa,Mqm)
minqm_min(MalfaLm[,2])
posic_match(minqm,MalfaLm[,2])
alfamin_MalfaLm[posic,1]

liminf_alfamin-.01
limsup_alfamin+.01
Mqm_matrix(0,nrow=21,ncol=1)
Malfa_matrix(0,nrow=21,ncol=1)
alfa_liminf-.001
i_0
while(alfa<limsup)
    {
    i_i+1
    alfa_alfa + .001
    alfa_round(alfa,3)
    if(alfa!=0)
        {
        Xtrans_((X^alfa)-1)/alfa
        }
    else
        Xtrans_log(X)
    Xm_cbind(unos,Xtrans)
    prod_t(Xm)%*%Xm
    Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
    qm_t(Y)%*%(diag(n)-Pm)%*%Y
    Mqm[i,1]_qm
    Malfa[i,1]_alfa
}
MalfaLm_cbind(Malfa,Mqm)
minqm_min(MalfaLm[,2])
posic_match(minqm,MalfaLm[,2])
alfamin_MalfaLm[posic,1]

Lm_(2*(n-1)*(1/(n-km-2))*minqm)^(1/2)
Mm_2*(pi^(1/2))*((gamma((n-km)/2)/gamma(n-(km/2)))^(1/n))*(minqm^(1/2))

taomoda_(n-km-2)/minqm
numcaldl_((taomoda^(-1/2))/2)*((1-(km/n))^(1/2))*((1-(1/n)*(1+(1/32)*(1-
(km/n))*(1-(2/n))))^(1/2)
numcaldm_(2^(km/(2*n)))*((2*pi)^(1/2))*(taomoda^(-1/2))*((((n/(n-2))^((n-
km)/2))-((n/(n-1))^(n-km)))^(1/2))
return(alfamin,minqm,Lm,Mm,numcaldl,numcaldm)
}
```

Programa 12.
Elige el mejor valor de alfa para transformar dos variables explicativas.

```
transfgen_function(alfa1, alfa2)
{
X1_matrix(scan(file="c:/sptemp/ox1.out"),nrow=100,ncol=1)
```

```
X2_matrix(scan(file="c:/sptemp/ox2.out"),nrow=100,ncol=1)
X_cbind(X1,X2)
Y_matrix(scan(file="c:/sptemp/oy.out"),nrow=100,ncol=1)
n_nrow(Y)
km_3
library(Matrix)
unos_matrix(1,nrow=n,ncol=1)
Mqm_matrix(0,nrow=21,ncol=1)
Xtrans_matrix(0,nrow=n,ncol=2)
Malfa_matrix(0,nrow=21,ncol=2)
Malfamin_matrix(0,nrow=21,ncol=1)

Malfminl_matrix(0,nrow=21,ncol=1)
Mqmmin_matrix(0,nrow=21,ncol=1)
library(Matrix)

alfal_alfal-.1
for(i in 1:21)
   {
   alfal_alfal + 0.1
   alfal_round(alfal,3)
   if(alfal!=0)
      {
      Xtrans[,1]_((X[,1]^alfal)-1)/alfal
      }
   else
      Xtrans[,1]_log(X[,1])
      alfal2_alfa2-0.1
      for(j in 1:21)
         {
         alfal2_alfal2 + 0.1
         alfal2_round(alfal2,3)
         if(alfal2!=0)
            {
            Xtrans[,2]_((X[,2]^alfal2)-1)/alfal2
            }
         else
            Xtrans[,2]_log(X[,2])
   Xm_cbind(unos,Xtrans)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve(prod))%*%t(Xm)
   qm_t(Y)%*%(diag(n)-Pm)%*%Y
   #Lm_(2*(n-1)*(1/(n-km-2))*qm)*(1/2)
   Mqm[j,1]_qm
   Malfa[j,1]_alfal
   Malfa[j,2]_alfal2
         }
   MalfaLm_cbind(Malfa,Mqm)
   minqm_min(MalfaLm[,3])
   posic_match(minqm,MalfaLm[,3])
   Malfamin[i]_MalfaLm[posic,1]
   Malfminl[i]_MalfaLm[posic,2]
   Mqmmin[i]_Mqm[posic,1]
   }
Malfaqm_cbind(Malfamin,Malfminl,Mqmmin)
minqm_min(Malfaqm[,3])
```



```
posic_match(minqm,Malfaqm[,3])
alfamin_Malfaqm[posic,1]
alfaminl_Malfaqm[posic,2]

liminf_alfamin-.1
limsup_alfamin+.1
liminfl_alfaminl-.1
limsupl_alfaminl+.1
Mqm_matrix(0,nrow=21,ncol=1)
Malfa_matrix(0,nrow=21,ncol=2)
Xtrans_matrix(0,nrow=n,ncol=2)
Malfamin_matrix(0,nrow=21,ncol=1)
Malfminl_matrix(0,nrow=21,ncol=1)
Mqmmin_matrix(0,nrow=21,ncol=1)
alfa_liminf-.01
i_0
for(i in 1:21)
   {
   alfa_alfa + .01
   alfa_round(alfa,3)
   if(alfa!=0)
      {
      for(k in 1:n)
         Xtrans[k,1]_((X[k,1]^alfa)-1)/alfa
      }
   else
      Xtrans[,1]_log(X[,1])
      alfal_liminfl-.01
      for(j in 1:21)
         {
         alfal_alfal + .01
         alfal_round(alfal,3)
         if(alfal!=0)
            {
            for(k in 1:n)
            Xtrans[k,2]_((X[k,2]^alfal)-1)/alfal
            }
         else
            Xtrans[,2]_log(X[,2])
   Xm_cbind(unos,Xtrans)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve(prod))%*%t(Xm)
   qm_t(Y)%*%(diag(n)-Pm)%*%Y
   Mqm[j]_qm
   Malfa[j,1]_alfa
   Malfa[j,2]_alfal
         }
   MalfaLm_cbind(Malfa,Mqm)
   minqm_min(MalfaLm[,3])
   posic_match(minqm,MalfaLm[,3])
   Malfamin[i]_MalfaLm[posic,1]
   Malfminl[i]_MalfaLm[posic,2]
   Mqmmin[i]_Mqm[posic,1]
   }
Malfaqm_cbind(Malfamin,Malfminl,Mqmmin)
minqm_min(Malfaqm[,3])
```

```
posic_match(minqm,Malfaqm[,3])
alfamin_Malfaqm[posic,1]
alfaminl_Malfaqm[posic,2]

liminf_alfamin-.01
limsup_alfamin+.01
liminfl_alfaminl-.01
limsupl_alfaminl+.01
Mqm_matrix(0,nrow=21,ncol=1)
Malfa_matrix(0,nrow=21,ncol=2)
Xtrans_matrix(0,nrow=n,ncol=2)
Malfamin_matrix(0,nrow=21,ncol=1)
Malfminl_matrix(0,nrow=21,ncol=1)
Mqmmin_matrix(0,nrow=21,ncol=1)
alfa_liminf-.001
for(i in 1:21)
   {
   alfa_alfa + .001
   alfa_round(alfa,3)
   if(alfa!=0)
      {
      for(k in 1:n)
      Xtrans[k,1]_((X[k,1]^alfa)-1)/alfa
      }
   else
      Xtrans[,1]_log(X[,1])
      alfal_liminfl-.001
      for(j in 1:21)
         {
         alfal_alfal + .001
         alfal_round(alfal,3)
         if(alfal!=0)
            {
            for(k in 1:n)
               Xtrans[k,2]_((X[k,2]^alfal)-1)/alfal
            }
         else
            Xtrans[,2]_log(X[,2])
   Xm_cbind(unos,Xtrans)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
   qm_t(Y)%*%(diag(n)-Pm)%*%Y
   Mqm[j,1]_qm
   Malfa[j,1]_alfa
   Malfa[j,2]_alfal
         }
   MalfaLm_cbind(Malfa,Mqm)
   minqm_min(MalfaLm[,3])
   posic_match(minqm,MalfaLm[,3])
   Malfamin[i]_MalfaLm[posic,1]
   Malfminl[i]_MalfaLm[posic,2]
   Mqmmin[i]_Mqm[posic,1]
   }
Malfaqm_cbind(Malfamin,Malfminl,Mqmmin)
minqm_min(Malfaqm[,3])
posic_match(minqm,Malfaqm[,3])
```



```
alfamin_Malfaqm[posic,1]
alfaminl_Malfaqm[posic,2]

Lm_(2*(n-1)*(1/(n-km-2))*minqm)^(1/2)
Mm_2*(pi^(1/2))*((gamma((n-km)/2)/gamma(n-(km/2)))^(1/n))*(minqm^(1/2))

taomoda_(n-km-2)/minqm
numcaldl_((taomoda^(-1/2))/2)*((1-(km/n))^(1/2))*((1-(1/n)*(1+(1/32)*(1-
(km/n))*(1-(2/n))))^(1/2)
numcaldm_(2^(km/(2*n)))*((2*pi)^(1/2))*(taomoda^(-1/2))*((((n/(n-2))^((n-
km)/2))-((n/(n-1))^(n-km)))^(1/2))
return(alfamin, alfaminl, Lm, Mm, numcaldl, numcaldm)
}
```

Programa 13.
Elige los mejores valores de transformación para obtener un polinomio en 2 potencias de X.

```
transpol_function(alfamin,alfaminl)
{
X1_matrix(scan(file="c:/sptemp/aerugx.txt"),ncol=1)
X2_X1
X_cbind(X1,X2)
Y_matrix(scan(file="c:/sptemp/aerugy.txt"),ncol=1)
n_nrow(Y)
km_3
library(Matrix)
unos_matrix(1,nrow=n,ncol=1)
Mqm_matrix(0,nrow=41,ncol=1)
Malfa_matrix(0,nrow=41,ncol=2)
Xtrans_matrix(0,nrow=n,ncol=2)
Malfamin_matrix(0,nrow=41,ncol=1)
Malfminl_matrix(0,nrow=41,ncol=1)
Mqmmin_matrix(0,nrow=41,ncol=1)

alfa_alfamin-.1
for(i in 1:41)
   {
   alfa_alfa + .1
   alfa_round(alfa,3)
   if(alfa!=0)
      {
      for(k in 1:n)
         Xtrans[k,1]_((X[k,1]^alfa)-1)/alfa
      }
   else
      Xtrans[,1]_log(X[,1])
   alfal_alfaminl-.1
   for(j in 1:41)
      {
      alfal_alfal + .1
      alfal_round(alfal,3)
      if(alfal!=alfa)
         {
         if(alfal!=0)
            {
```

```
            for(k in 1:n)
               Xtrans[k,2]_((X[k,2]^alfal)-1)/alfal
            }
         else
            Xtrans[,2]_log(X[,2])
   Xnueva_Xtrans
   Xm_cbind(unos,Xnueva)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
   qm_t(Y)%*%(diag(n)-Pm)%*%Y
   Mqm[j,1]_qm
   Malfa[j,1]_alfa
   Malfa[j,2]_alfal
         }
      }
   MalfaLm_cbind(Malfa,Mqm)
   minqm_min(MalfaLm[,3])
   if (minqm==0)
      {
      Mord_MalfaLm[order(MalfaLm[,3]),1:3]
      minqm_Mord[2,3]
      Malfamin[i]_Mord[2,1]
      Malfminl[i]_Mord[2,2]
      Mqmmin[i]_minqm
      }
   else
      {
      posic_match(minqm,MalfaLm[,3])
      Malfamin[i]_MalfaLm[posic,1]
      Malfminl[i]_MalfaLm[posic,2]
      Mqmmin[i]_Mqm[posic,1]
      }
   }
Malfaqm_cbind(Malfamin,Malfminl,Mqmmin)
minqm_min(Malfaqm[,3])
posic_match(minqm,Malfaqm[,3])
alfamin_Malfaqm[posic,1]
alfaminl_Malfaqm[posic,2]
Mqm_matrix(0,nrow=21,ncol=1)
Malfa_matrix(0,nrow=21,ncol=2)
Xtrans_matrix(0,nrow=n,ncol=2)
Malfamin_matrix(0,nrow=21,ncol=1)
Malfminl_matrix(0,nrow=21,ncol=1)
Mqmmin_matrix(0,nrow=21,ncol=1)

liminf_alfamin-.1
liminfl_alfaminl-.1
alfa_liminf-.01
for(i in 1:21)
   {
   alfa_alfa + .01
   alfa_round(alfa,3)
   if(alfa!=0)
      {
      for(k in 1:n)
         Xtrans[k,1]_((X[k,1]^alfa)-1)/alfa
```



```
      }
   else
      Xtrans[,1]_log(X[,1])
      alfal_liminf1-.01
      h_0
      for(j in 1:21)
         {
         alfal_alfal + .01
         alfal_round(alfal,3)
      if(alfal!=alfa)
         {
         h_h+1
         if(alfal!=0)
            {
            for(k in 1:n)
               Xtrans[k,2]_((X[k,2]^alfal)-1)/alfal
            }
         else
            Xtrans[,2]_log(X[,2])
   Xnueva_Xtrans
   Xm_cbind(unos,Xnueva)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
   qm_t(Y)%*%(diag(n)-Pm)%*%Y
   Mqm[h,1]_qm
   Malfa[h,1]_alfa
   Malfa[h,2]_alfal
         }
      }
   MalfaLm_cbind(Malfa,Mqm)
   minqm_min(MalfaLm[,3])
   if (minqm==0)
      {
      Mord_MalfaLm[order(MalfaLm[,3]),1:3]
      minqm_Mord[2,3]
      Malfamin[i]_Mord[2,1]
      Malfminl[i]_Mord[2,2]
      Mqmmin[i]_minqm
      }
   else
      {
      posic_match(minqm,MalfaLm[,3])
      Malfamin[i]_MalfaLm[posic,1]
      Malfminl[i]_MalfaLm[posic,2]
      Mqmmin[i]_Mqm[posic,1]
      }
   }
Malfaqm_cbind(Malfamin,Malfminl,Mqmmin)
minqm_min(Malfaqm[,3])
posic_match(minqm,Malfaqm[,3])
alfamin_Malfaqm[posic,1]
alfaminl_Malfaqm[posic,2]
Mqm_matrix(0,nrow=21,ncol=1)
Malfa_matrix(0,nrow=21,ncol=2)
Xtrans_matrix(0,nrow=n,ncol=2)
Malfamin_matrix(0,nrow=21,ncol=1)
```

```
Malfminl_matrix(0,nrow=21,ncol=1)
Mqmmin_matrix(0,nrow=21,ncol=1)

liminf_alfamin-.01
liminfl_alfaminl-.01
alfa_liminf-.001
for(i in 1:21)
   {
   alfa_alfa + .001
   alfa_round(alfa,3)
   if(alfa!=0)
      {
      for(k in 1:n)
         Xtrans[k,1]_((X[k,1]^alfa)-1)/alfa
      }
   else
      Xtrans[,1]_log(X[,1])

      alfal_liminfl-.001
      for(j in 1:21)
         {
         alfal_alfal + .001
         alfal_round(alfal,3)
         if(alfal!=0)
            {
            for(k in 1:n)
               Xtrans[k,2]_((X[k,2]^alfal)-1)/alfal
            }
         else
            Xtrans[,2]_log(X[,2])
   Xnueva_Xtrans
   Xm_cbind(unos,Xnueva)
   prod_t(Xm)%*%Xm
   Pm_Xm%*%(solve.Matrix(prod))%*%t(Xm)
   qm_t(Y)%*%(diag(n)-Pm)%*%Y
   Mqm[j,1]_qm
   Malfa[j,1]_alfa
   Malfa[j,2]_alfal
         }
   MalfaLm_cbind(Malfa,Mqm)
   minqm_min(MalfaLm[,3])
   if (minqm==0)
      {
      Mord_MalfaLm[order(MalfaLm[,3]),1:3]
      minqm_Mord[2,3]
      Malfamin[i]_Mord[2,1]
      Malfminl[i]_Mord[2,2]
      Mqmmin[i]_minqm
      }
   else
      {
      posic_match(minqm,MalfaLm[,3])
      Malfamin[i]_MalfaLm[posic,1]
      Malfminl[i]_MalfaLm[posic,2]
      Mqmmin[i]_Mqm[posic,1]
      }
```



```
   }
Malfaqm_cbind(Malfamin,Malfminl,Mqmmin)
minqm_min(Malfaqm[,3])
posic_match(minqm,Malfaqm[,3])
alfamin_Malfaqm[posic,1]
alfaminl_Malfaqm[posic,2]

Lm_(2*(n-1)*(1/(n-km-2))*minqm)^(1/2)
Mm_2*(pi^(1/2))*((gamma((n-km)/2)/gamma(n-(km/2)))^(1/n))*(minqm^(1/2))

taomoda_(n-km-2)/minqm
numcaldl_((taomoda^(-1/2))/2)*((1-(km/n))^(1/2))*((1-(1/n)*(1+(1/32)*(1-
(km/n))*(1-(2/n))))^(1/2)
numcaldm_(2^(km/(2*n)))*((2*pi)^(1/2))*(taomoda^(-1/2))*((((n/(n-2))^((n-
km)/2))-((n/(n-1))^(n-km)))^(1/2))
return(alfamin,alfaminl,Lm,Mm,numcaldl,numcaldm)
}
```

Programa 14.
Genera las Xs, Bs y errores para el experimento simulado.

```
xbyerror_function(n)
{
      X1_rnorm(100,6,2.5)
      X1_round(X1,2)
      X2_rnorm(100,30,10)
      X2_round(X2,1)
      unos_matrix(1,nrow=100,ncol=1)
      B_c(15,4,-8)
      error_rnorm(100,0,4)
write(X1,"c:/sptemp/ox1.out",ncol=1)
write(X2,"c:/sptemp/ox2.out",ncol=1)
write(B,"c:/sptemp/ob.out",ncol=1)
write(error,"c:/sptemp/oerror.out",ncol=1)
}
```

REFERENCIAS

- Atkinson, A. C. (1987), *Plots, Tranformations and Regression (An Introduction to Graphical Methods of Diagnostic Regression Analysis)*, New York: Oxford University Press.
- Bernardo, J. M. & Smith, A. F. M. (1994), *Bayesian Theory*, Chichester: Wiley.
- Box, G. E. P. & Jenkins, G. M. (1970). *Time series analysis: forecasting and control*, Holden-Day, London.
- Box, G. E. P. & Tidwel, P. W. (1962), "Transformation of the Independent Variables", *Technometrics*, 4, 531-550.
- Broemeling, L. D. (1985), *Bayesian Analysis of Linear Models*, New York: Marcel Dekker Inc.
- Burnham, K. & Anderson, D. (1998), *Model Selection and Inference: A Practical Information-Theoretic Approach*, New York: Springer.
- Chatterjee, S. & Price, B. (199.1), *Regression Analysis by Example*, (2a. Ed.), New York: John Wiley & Sons.
- Carrol R. J. & Rupert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman and Hall.
- Casella, G. & Berger R. L. (1990), *Statistical Inference*, California: Duxbury Press.
- Fernández, J. (1978), "Acerca de la teoría de la información y algunas de sus aplicaciones", *Com. Int.* 23, Depto. de Matemáticas, Fac. de Ciencias.
- Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics. Royal Society of London". *Philosophical Transactions*, Ser. A, 222, 309-368.
- Freedman, D. A. (1983), " A note on screening regression equations". *The American Statician*, 37, 152-155.
- Geisser, S. & Eddy, W. F. (1979), "A Predictive Approach to Model Selection", *Journal of the American Statistical Association*, 74, 153-160.
- Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, New York: Springer.

- Intriligator, M., et. al. (1978), *Econometric Models, Techniques and Applications*, (2ª. ed.), New York: Prentice Hall.
- Laud, P. W., & Ibrahim, J.G. (1995), "Predictive model Selection", *Journal of the Royal Statistical Society*, Ser. B, 57, 247-262.
- Laud, P. W., & Ibrahim, J.G. (1995), "Predictive Specification of Prior Model Probabilities in Variable Selection", *Biostat Technical report Series*, Reporte Técnico No. 9, 1-12.
- Maddala, G. S. (1988), *Introduction to Econometrics*, (2ª. ed.), New York: Macmillan Publishing Company.
- McCullagh, P., & Nelder, J. A. (1989), *Generalized Linear Models*, (2ª. ed.), London: Chapman and Hall.
- Pollard, W. E. (1986), *Bayesian Statistics for Evaluation Research. An Introduction*, Sage Publications.
- Schwarz, G. (1978), "Estimating the Dimension of a Model", *Annals of Statistics*, 6, 461-464.
- Seber, G. A. F. (1977), *Linear Regression Analysis*, New York: John Wiley & Sons.
- Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Lineal Models", *Journal of the Royal Statistical Society*, Ser. B., 42, 213-220.