2ej
21

Universidad Nacional Autónoma de México

FACULTAD DE CIENCIAS

SELECCION DE VARIABLES EN ANALISIS DE REGRESION BAYESIANO.

T E S I S

Que para obtener el título de

M A T E M A T I C O

P R E S E N T A

HORTENSIA REYES CERVANTES

México, D. F. 1987

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

INTRODUCCIÓN

El origen de esta tesis surge del interés de conocer el material referente al Análisis de Regresión Bayesiano, orientado a resolver el problema de selección de variables, "Ya que el Análisis de Regresión es una de las técnicas estadísticas más empleadas para analizar múltiples factores de datos"(Montgomery,1982,p.v)

Es importante tener en mente, que no siempre es deseable efectuar una selección de variables ya que de realizarse pueden cometerse errores que llevan a perder información y/o en algunas ocaciones se complique innecesariamente el problema y esto, al final de cuentas, repercutirá en lo bueno que sean los resultados finales. Por ejemplo, tomando como referencia el procedimiento de selección dado por Lindley,(1968), se tendrían que tomar en cuenta: costos, captación y riesgos de haber elegido a las variables explicativas. Además, el peligro de una mala inferencia se implementaría por la selección de variables.

El no haber hecho una selección de variables, implica el resolver el problema de regresión con todas las variables explicativas que se tengan y esto, actualmente ya no resulta un trabajo difícil en algunos casos, pues en gran medida han ayudado los avances tecnológicos de las nuevas computadoras.

En general, la elección del mejor modelo de regresión no es un mecánismo estadístico único (Weisberg,1980,p.174), la razón se debe a la diversidad de las situaciones en que se desarrollan los problemas de regresión, y a la manera distinta que cada individuo tiene de concebir y actuar ante un problema.

La elección del mejor modelo de regresión debe establecer un compromiso entre dos criterios (opuestos) que son presentados por Draper-Smith (1981, p.294), las cuales son:

1. Incluir en el modelo, todas las posibles variables explicativas para predecir mejor a la variable dependiente.

2. Dado que cada variable explicativa requiere un costo de obtención y captación de información, se pretende elegir un menor número de variables explicativas.

La justificación para llevar a cabo una selección de variables es ocasionada por alguna o todas las siguientes causas, que se pueden presentar en un problema de regresión.

1º. El problema de regresión cuenta con "Muchas Variables", por lo que debido a las consideradiones hechas por el interesado y/o estadístico, se llega a la conclusión que hay información repetitiva en las observaciones.

2º. Las características propias del problema de regresión hacen que el número de variables explicativas sea menor que el número de observaciones.

3º. Debido al costo desmedido en el manejo y captación de las variables explicativas, se desea que para metas futuras de mantenimiento en el modelo de regresión, éste cuente con el menor número de variables explicativas.

A continuación se presenta de manera breve el contenido del trabajo.

En el Capítulo 1, se plantea el Modelo de Regresión Bayesiana se dan algunas indicaciones en donde el Modelo de Regresión Clásico y el Bayesiano coinciden. En el Capítulo 2, se presentan algunos de los procedimientos de Selección Bayesiano y en el Capítulo 3 se presentan las conclusiones, así como comentarios respecto a cual podría ser la mejor técnica de selección.

I

ANALISIS DE REGRESION

CAPITULO I

ANALISIS DE REGRESION.

El Análisis de Regresión es una de las técnicas estadísticas más empleadas para analizar situaciones en donde el objetivo esencial es establecer una función que permita al estadístico relacionar un conjunto de variables llamadas independientes ($X_1, \ldots, X_n$) con otra variable llamada dependiente (y).

Los objetivos principales de los modelos de regresión son a grandes rasgos: 1) Descripción de la relación funcional entre la variables independiente y la variable dependiente, y 2) Predicción de la variable dependiente.

Existen dos enfoques estadísticos alternativos que solventan los problemas de regresión, estos son: el punto de vista bayesiano y el punto de vista clásico. El enfoque que se utilizará a lo largo de este trabajo, es el bayesiano, luego entonces el enfoque clásico (Draper-Smith, 1981) no se desarrollará y únicamente se referirá a él, en casos especiales para comparar algunos resultados.

En el Análisis de Regresión Bayesiano, el estadístico puede introducir el conocimiento disponible sobre el problema, a la información producida por las observaciones, mediante el Teorema de Bayes, y además, se puede proporcionar un criterio de decisión basándose en una función de pérdida o ganancia. Si desde un enfoque bayesiano se utiliza una distribución inicial de referencia, entonces las similitudes entre los resultados obtenidos por ambos enfoques son gran

des.

En lo referente a problemas de inferencia, pueden presentarse situaciones en que no haya necesidad de elegir formalmente un es-- timador porque no se tiene o no se quiere utilizar una función de pérdida, ya que lo único que se pretende es hacer la Inferencia Estadística con respecto a los parámetros involucrados. Por ejemplo, expresar el conocimiento disponible sobre el parámetro de localización y/o el parámetro de escala, así como otras inferencias acerca de las distribuciones de las observaciones.

En este capítulo, se pretende dar un panorama general del Análisis de Regresión Bayesiano desde dos puntos de vista o necesidades del problema: Inferencia y Decisión. Por último se tocará muy sucintamente el tema del Modelo Lineal General.

1. MODELO DE REGRESIÓN LINEAL.

La formulación del problema de Regresión Bayesiano es necesariamente más complicada y requiere de más cuidado que la descripción teórica muestral clásica. Esta diferencia se debe a 1) que todas las cantidades $(y, x_1, \ldots, x_n)$ y los parámetros son variables aleatorias, y 2) todas las afirmaciones probabilísticas son condicionales a los eventos de interés (Lindley, 1968).

El problema de Regresión Lineal Múltiple está contemplado en el área que involucra los modelos lineales, porque se supone que la media de cada una de las observaciones es una combinación lineal de un conjunto de parámetros fijos. Por lo cual, se puede escribir de la siguiente forma

$$\underline{y} = X\underline{\theta} + \underline{\varepsilon} \qquad (1.1)$$

donde: $\underline{\theta}$ es un vector de $(k+1)$ parámetros desconocidos cuyos valores se encuentran en $\mathbb{R}$, $\underline{\varepsilon}$ es un vector de $n\times 1$ con distribución Normal Multivariada con media cero y varianza común σ^2 e independientes $(\varepsilon_i \sim N(0, \sigma^2)\ \forall i = \overline{1,n})$, por lo tanto $\underline{y}$ es un vector $(n\times 1)$ que tiene distribución Normal para toda $i, i=\overline{1,n}$ $(\underline{y} = (y_1, \ldots, y_n)^t)$; cuya media es una combinación de k componentes de $\underline{\theta}$ $(E(\underline{y}/\underline{\theta}) = X\underline{\theta})$ y cuya varianza es σ^2 $(V(\underline{y}/X) = \sigma^2 I)$. Por último, X matriz $(n\times k)$ de valores x_{ij} (con $i=\overline{1,n}$; $j=\overline{1,k}$) cuyo rango es igual a k; que pueden representar niveles o funciones de una variable de un experimento, o quizá se representen valores de variables indicadoras (representados por unos ó ceros).

Se denota por σ^2 a la varianza cuyo valor debe ser positivo, y

se puede demostrar que sus estadísticas respectivas $\underline{\theta}$ y s^2 son independientes. Estas estadísticas más adelante se presentarán.

La información expresada en la muestra, es decir, las observaciones $X_1, \ldots, X_n$ se representan mediante la función de $\underline{y}$, que escrita en forma matricial es

$$p(\underline{y}/\underline{\theta},\sigma^2,X) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}\, e^{-\frac{1}{2\sigma^2}(\underline{y}-X\underline{\theta})^t(\underline{y}-X\underline{\theta})} \qquad (1.2)$$

Haciendo una breve observación sobre esta densidad, se tiene que tomando $w = \frac{1}{\sigma^2}$, la precisión, se sigue conservando la normalidad en la densidad de $\underline{y}$. Solo se obtiene una expresión que para algunas personas proporciona un manejo más cómodo en cuanto a su interpretación.

$$p(\underline{y}/\underline{\theta},\omega,X) = \left(\frac{1}{\sqrt{2\pi}}\right)^n w^{n/2}\, e^{-\frac{w}{2}(\underline{y}-X\underline{\theta})^t(\underline{y}-X\underline{\theta})} \qquad (1.2')$$

Retomando la expresión (1.2), está se puede expresar como

$$p(\underline{y}/\underline{\theta},\sigma^2,X) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}\, e^{-\frac{1}{2\sigma^2}\left[(\underline{y}-\hat{\underline{y}})^t(\underline{y}-\hat{\underline{y}}) + (\underline{\theta}-\underline{\theta})^t X^t X (\underline{\theta}-\hat{\underline{\theta}})\right]} \qquad (1.3)$$

que consiste en sumar y restar las estadísticas $\hat{\underline{\theta}}$, s^2, agrupando adecuadamente. Este procedimiento es el mismo que Zellner hizó (1968, p.66) y así obtuvo la densidad de $\underline{y}$ como

$$p(\underline{y}/\underline{\theta},\sigma^2,X) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}\, e^{-\frac{1}{2\sigma^2}\left[\nu s^2 + (\underline{\theta}-\hat{\underline{\theta}})^t X^t X (\underline{\theta}-\hat{\underline{\theta}})\right]} \qquad (1.4)$$

con

$$\hat{\underline{\theta}} = (X^t X)^{-1} X^t \underline{y} \,, \qquad (1.5)$$

$$s^2 = \frac{1}{\nu} (\underline{y} - \hat{\underline{y}})^t (\underline{y} - \hat{\underline{y}}) \,, \quad \hat{\underline{y}} = X\hat{\underline{\theta}} \quad y \quad \nu = n - K \qquad (1.6)$$

A partir de estas expresiones y apoyados en otros resultados se puede demostrar lo siguiente: 1) ($\hat{\underline{\theta}}, s^2$) es una estadística suficiente conjunta de ($\underline{\theta}, \sigma^2$), 2) que el estimador de $\underline{\theta}$, $\hat{\underline{\theta}}$, tiene una distribución Normal Multivariada con varianza $\sigma^2 (X^t X)^{-1}$ (esto es, $\hat{\underline{\theta}} \sim N_K (\underline{\theta}, \sigma^2 (X^t X)^{-1})$ y 3) νs^2 tiene una distribución independiente de $\hat{\underline{\theta}}$ con distribución ji-cuadrada con ν-grados de libertad (se denota por $\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu$).

Después de haber perdido normalidad en las observaciones y haber obtenido su densidad, $p (\underline{y} / \underline{\theta}, \sigma^2)$ es necesario establecer en seguida que tipo de distribución inicial (referente al parámetros desconocido) es la contemplada por el interesado y/o estadístico. Para lo cual se detallarán a continuación los casos de 1.1) cuando el interesado expresa su conocimiento acerca del (os) parámetro(s) desconocido(s) y aplica la regla de Bayes. Concluyendo con una distribución posterior acerca de(los) parámetro(s), ó cuando 1.2) existen situaciones en que el interesado no expresa su conocimiento por no poder o querer, ocasionándo así que no se establezca una distribución inicial informativa acerca de los parámetros desconocidos.

1.1. DISTRIBUCIÓN INICIAL INFORMATIVA,

Esgrimiendo el hecho de que la información inicial se necesita

incorporar mediante una distribución adecuada, DeGroot (1970, p.183) propone utilizar el Modelo Normal Gamma, que trae como consecuencia el establecimiento de que la distribución condicional de $\underline{\theta}$ cuando $W=w$ es una distribución normal multivariada con vector de medias $\underline{\mu}$ y matriz de precisión $w\underline{\tau}$ (esto es, $\underline{\theta}/W=w \sim N_k(\underline{\mu}, w\underline{\tau})$), y que la distribución marginal de W es una Gamma con parámetros α y β (ambos positivos). Entonces la distribución inicial conjunta de $\underline{\theta}$ y w es:

$$\xi(\underline{\theta},w) \propto w^{k/2} e^{-\frac{w}{2}(\underline{\theta}-\underline{\mu})^t \underline{\tau}(\underline{\theta}-\underline{\mu})} w^{\alpha-1} e^{-\beta w} \qquad (1.7)$$

Por lo tanto, la distribución posterior $p(\underline{\theta},w/\underline{y},X)$ se obtiene haciendo el producto de (1.2') y (1.7), concluyendo con

$$p(\underline{\theta},w/\underline{y},X) \propto w^{k/2} e^{-\frac{w}{2}(\underline{\theta}-\underline{\mu})^t \underline{\tau}(\underline{\theta}-\underline{\mu})} w^{\alpha-1} e^{-\beta w} w^{\frac{n}{2}} e^{-\frac{w}{2}(\underline{y}-X\underline{\theta})^t(\underline{y}-X\underline{\theta})}, \qquad (1.8)$$

tomando a

$$\underline{\mu}_1 = (\underline{\tau}+X^tX)^{-1}(\underline{\tau}\underline{\mu}+X^t\underline{y}) \quad \text{y} \qquad (1.9)$$

$$\beta_1 = \beta + \tfrac{1}{2}\left[(\underline{y}-X\underline{\mu}_1)^t\underline{y} + (\underline{\mu}-\underline{\mu}_0)^t\underline{\tau}\underline{\mu}\right] \quad \text{en (1.8)}$$

se obtiene la distribución posterior $p(\underline{\theta},w/\underline{y},X)$, así:

$$p(\underline{\theta},w/\underline{y},X) \propto w^{k/2+\alpha+\frac{n}{2}-1} e^{-\frac{w}{2}(\underline{\theta}-\underline{\mu}_1)^t(\underline{\tau}+X^tX)(\underline{\theta}-\underline{\mu}_1)-\beta_1 w} \qquad (1.10)$$

Las marginales de esta distribución involucradas son: La distribución condicional de $\underline{\theta}$ dado $W=w$ que es una distribución normal multivariada con media $\underline{\mu}_1$ y matriz de precisión $w(\underline{\tau}+\underline{X}^t\underline{X})$ (es decir, $\underline{\theta}/W=w \sim N_n(\underline{\mu}_1, w(\underline{\tau}+\underline{X}^t\underline{X}))$, la distribución marginal de w es una Gamma con parámetros $\alpha+\frac{n}{2}$ y β_1 ($w \sim Gamma(\alpha+\frac{n}{2}, \beta_1)$) y la distribución marginal de $\underline{\theta}$ tiene una distribución t-multivariada con $2\alpha+n$ grados de libertad, vector de localización $\underline{\mu}_1$ y matriz de precisión $\frac{2\alpha+n}{2\beta_1}(\underline{\tau}+\underline{X}^t\underline{X})$.

1.1.1. INFERENCIA.

Las inferencias acerca de los parámetros desconocidos son el resultado del análisis y estudio de la distribución posterior conjunta o marginal. Para realizar estos objetivos es de gran utilidad el obtener regiones de alta densidad.

Las regiones de alta densidad de tamaños $(1-\alpha)$ son comúnes en la Estadística Bayesiana (Box-Tiao, 1971, pp. 121-136) y se construyen sobre el espacio parámetrico de interés. La caracterización de estas regiones (R), está dada por cualquiera de las siguientes condiciones: 1) que la densidad de probabilidad de todo punto interior a R sea mayor a la densidad correspondiente a todo punto exterior (esto es, si $\underline{\theta}_1 \in R, \underline{\theta}_2 \notin R \Rightarrow p(\underline{\theta}_1|\underline{y}) \geq p(\underline{\theta}_2|\underline{y})$); 2) la región R, para una probabilidad contenida $1-\alpha$, debe ser tal que tenga el volúmen más pequeño entre todas las regiones en el espacio paramétrico con contenido de probabilidad $1-\alpha$, esta propiedad se expresa así: si A y R son conjuntos tales que $P_r(\underline{\theta} \in A/\underline{y}) = P_r(\underline{\theta} \in R/\underline{y}) = 1-\alpha$, entonces el volúmen $(A) \geq$ volúmen (R).

Además, se puede demostrar que al tomar una condición trae como consecuencia la otra.

Si definimos una transformación $\underline{\Phi} = f(\underline{\theta})$, de los parámetros de $\underline{\theta}$ a $\underline{\Phi}$, para alguna región contenida en el espacio de $\underline{\Phi}$ y si la región de $\underline{\theta}$ es de alta densidad, su región transformada en $\underline{\Phi}$ no será de alta densidad, a menos que la transformación utilizada sea lineal.

Para determinar si un parámetro puntual $\underline{\theta}_0$, se encuentra en una región, R_α, de alta densidad que no está determinada. Se puede proceder de dos formas: 1) determinar analíticamente la región ó 2) calcular $P_r\{p(\underline{\theta}/\underline{y}) > p(\underline{\theta}_0/\underline{y})\} = \rho$ si $\rho \leq 1-\alpha$ $\underline{\theta}_0 \in R_\alpha$, en caso contrario $\underline{\theta}_0 \notin R_\alpha$.

En esta expresión, la función de densidad $p(\underline{\theta}/\underline{y})$ es tratada como una variable aleatoria, pues $p(\underline{\theta}_0/\underline{y})$ es constante.

1.1.2 DECISION.

Cuando se quiere estimar un parámetro desconocido, digamos $\underline{\omega} = (\omega_1, \ldots, \omega_k)^t$ *que por ejemplo puede ser el vector de medias o de presición, a través de la teoría de decisión (Bernardo, 1980, pp 4.2a 4.26) es necesario especificar (de acuerdo a los axiomas de coherencia) los siguientes conjuntos: El conjunto de decisiones y que en este caso coincide con el espacio paramétrico* ($\omega \in \Omega$), *el conjunto de consecuencias (C) y construir una función de utilidad que describa las preferencias del decisior, entre las posibles consecuencias* ($U(d, \omega)$).

Partiendo de la definición de utilidad de Degroot (1971, pp. 91-92) pueden definirse fácilmente las funciones de ganancia o de pérdida, que son resultados de una transformación apropiada de la función de utilidad. Esto es, si U es una función de utilidad, la transformación $aU + b$ *define una función de ganancia si* $a > 0$; *y en el caso de que* $a < 0$ *se obtiene una función de pérdida.*

Retomando lo anterior, se ha definido el problema de decisión sobre el espacio paramétrico ω **, en donde la finalidad es seleccionar una decisión** d^* (con $d^* \in D$) *que minimice o maximice la función de utilidad.**

Si se cuenta con los axiomas de coherencia (Bernardo, 1980, pp. 4.2 4.26) el único criterio para elegir alternativas consiste en maximizar la utilidad esperada sobre el espacio de decisiones. [1] *Equivalen-*

(1) Tomando $p \in \mathcal{P}$ (la clase de distribuciones de probabilidad sobre el espacio paramétrico Θ, y que g es una función de utilidad de d, el criterio consiste en

$$\max_{d \in D} E(g|P) = \int_{\Omega} g(r)\, dP(r), \quad \text{si} \quad E(g|P) \text{ existe}$$

temente se puede expresar el criterio de decisión como maximizar la ganancia esperada o minimizar la pérdida esperada.

Los ejemplos más comunes de funciones de pérdidas son:

La función de pérdida lineal

$$L(d,\omega) = a|\omega - d| \quad , \quad a>0 \tag{1.14}$$

y la cuadrática

$$L(d,\omega) = b(\omega - d)^2, \quad b>0. \tag{1.15}$$

Tomando al parámetro $\underline{\omega} \in \mathbb{R}^k$ y $\underline{d} \in \mathbb{R}^k$, la función de pérdida se define como $L(\underline{d},\underline{\omega}) = (\underline{\omega}-\underline{d})^t A (\underline{\omega}-\underline{d})$ donde A es una matriz simétrica. El objetivo en este caso es maximizar $E((\omega-d)^t A (\omega-d))$ en d.

Algunas veces, como se verá más adelante, se puede presentar el problema de regresión como un planteamiento de Teoría de Decisiones.

1.2 DISTRIBUCIÓN INICIAL DE REFERENCIA.

Como ya se había comentado si no se cuenta con una distribución inicial no es posible obtener la distribución posterior del parámetro desconocido. De aquí, que es necesario un mecanismo para substituirlo matemáticamente en el Teorema de Bayes y a estas funciones se les llama Distribuciones Iniciales de Referencia.

Existen diversos métodos para encontrar Iniciales de Referencia, de las cuales sólo se referirán a las conluídas por Degroot (1970, pp.155 - 222), Box y Tiao (1971, pp. 1-14) y Bernardo (1980). Estos métodos en general no coinciden, sin embargo suponiéndo normalidad en las observaciones y siendo de interés solo uno de los parámetros desconocidos (θ ó $w = 1/\sigma^2$) se tienen Distribuciones Iniciales de Referencia iguales.

Esta situación se plantea de la siguiente manera [1] *: si el vector de medias ($\underline{\theta}$) es desconocido y la precisión es conocida, la inicial de referencia que se propone es:*

$$p(\underline{\theta}, w = 1/\sigma^2) \propto \text{constante}. \qquad (1.16)$$

Ahora si el vector de medias ($\underline{\theta}$) y la precisión (w) son desconocidos, y únicamente es de interés un parámetro, Bernardo (1979) propone la inicial de referencia dada por

$$p(w) \propto \frac{1}{w} \text{ y } p(\theta/w) \propto \text{constante} \qquad (1.17)$$

(1) Los resultados coinciden con las reglas de Jeffreys (Zellner, pp.42-44), que son distribuciones iniciales que representan poco conocimiento o ignorancia.

A continuación se distinguen dos situaciones prácticas de incertidumbre acerca de los parámetros involucrados en el desarrollo del Análisis de Regresión: 1.2.1 varianza conocida y media desconocida y 1.2.2 varianza y media desconocidas.

1.2.1 $\underline{\theta}$ DESCONOCIDA Y σ^2 CONOCIDA.

Caso poco habitual en el práctica, pero sirve de sustento teórico para el caso precedente (Sección 1.2.2).

Usando la regla de Bayes y expresando el resultado en términos de las estadísticas suficientes, para obtener la distribución final del parámetros desconocido, se tiene

$$p(\underline{\theta}/\underline{y},x) = p(\hat{\underline{\theta}}/\underline{y},x,\underline{\theta})\, p(\underline{\theta}/\hat{\underline{\theta}},\underline{y},x)$$

$$p(\underline{\theta}/\underline{y},x) \propto p(\hat{\underline{\theta}}/\underline{\theta})\, p(\underline{\theta}), \qquad (1.18)$$

se substituyen los términos (1.4) y (1.16) en (1.18) y quitando lo que no es desconocido, se tiene

$$p(\underline{\theta}/\underline{y},x) \propto e^{-\frac{1}{2\sigma^2}[(\underline{\theta}-\hat{\underline{\theta}})^t x^t x (\underline{\theta}-\hat{\underline{\theta}})]}, \qquad (1.19)$$

encontrando la constante de integración sobre $\underline{\theta}$

$$\int_{\theta_1} \cdots \int_{\theta_k} e^{-\frac{1}{2\sigma^2}[(\underline{\theta}-\hat{\underline{\theta}})^t x^t x (\underline{\theta}-\hat{\underline{\theta}})}\, d\theta_1 \ldots d\theta_k = \frac{\sigma^k (\sqrt{2\pi})^k}{|x^t x|^{1/2}}, \qquad (1.20)$$

por lo que sustituyendo (1.20) en (1.19), se encuentra

$$p(\underline{\theta}/\underline{y},X) = \frac{|X^tX|^{1/2}}{(\sqrt{2\pi}\,\sigma)^k}\; e^{-\frac{1}{2\sigma^2}[(\underline{\theta}-\hat{\underline{\theta}})^t X^tX(\underline{\theta}-\hat{\underline{\theta}})]} \qquad -\infty < \theta_i < \infty \;\; \forall i=\overline{1,k}, \tag{1.21}$$

de la expresión (1.21), se puede notar que lo único relevante acerca de $\underline{\theta}$ en la distribución, fué la información presentada en la muestra. El vector de medias se distribuye como una Normal Multivariada con media $\hat{\underline{\theta}}$ y varianza $\sigma^2(X^tX)^{-1}$; es decir

$$\underline{\theta} \sim N_k(\hat{\underline{\theta}},\ \sigma^2(X^tX)^{-1}). \tag{1.22}$$

Para propósitos de inferencia o decisión acerca del parámetro pasar a la sección 1.1.1 ó 1.1.2, respectivamente

1.2.2 $\underline{\theta}$ Y σ^2 DESCONOCIDOS.

Como ambos parámetros son desconocidos la distribución final consistirá de

$$p(\underline{\theta},\sigma^2/\underline{y},X) \propto p(\underline{\theta},\sigma^2)\,p(\underline{\theta},\sigma^2/\underline{y},X), \tag{1.23}$$

y usando la dependencia de las estadísticas $\hat{\underline{\theta}}$ y S^2, dados $\underline{\theta}$ y σ^2, se puede escribir (1.23) de otra forma, en términos de las estadísticas suficientes

$$p(\underline{\theta},\sigma^2/\underline{y},X) \propto p(\underline{\theta},\sigma^2)\,p(\hat{\underline{\theta}}/\underline{\theta},\sigma^2)\,p(S^2/\sigma^2). \tag{1.24}$$

Por lo que ahora utilizando la distribución inicial de referen-

$$p(\underline{\theta}/\underline{y},X) = \frac{|X^tX|^{1/2}}{(2\pi)^k\sigma^k}\; e^{-\frac{1}{2\sigma^2}[(\underline{\theta}-\hat{\underline{\theta}})^t X^tX(\underline{\theta}-\hat{\underline{\theta}})]} \qquad -\infty<\theta_i<\infty \;\; \forall i=\overline{1,k}, \tag{1.21}$$

de la expresión (1.21), se puede notar que lo único relevante acerca de $\underline{\theta}$ en la distribución, fué la información presentada en la muestra. El vector de medias se distribuye como una Normal Multivariada con media $\hat{\underline{\theta}}$ y varianza $\sigma^2(X^tX)^{-1}$; es decir

$$\underline{\theta} \sim N_k(\hat{\underline{\theta}},\ \sigma^2(X^tX)^{-1}). \tag{1.22}$$

Para propósitos de inferencia o decisión acerca del parámetro pasar a la sección 1.1.1 ó 1.1.2, respectivamente

1.2.2 $\underline{\theta}$ Y σ^2 DESCONOCIDOS.

Como ambos parámetros son desconocidos la distribución final consistirá de

$$p(\underline{\theta},\sigma^2/\underline{y},X) \propto p(\underline{\theta},\sigma^2)\,p(\underline{\theta},\sigma^2/\underline{y},X), \tag{1.23}$$

y usando la dependencia de las estadísticas $\hat{\underline{\theta}}$ y S^2 , dados $\underline{\theta}$ y σ^2, se puede escribir (1.23) de otra forma, en términos de las estadísticas suficientes

$$p(\underline{\theta},\sigma^2/\underline{y},X) \propto p(\underline{\theta},\sigma^2)\,p(\hat{\underline{\theta}}/\underline{\theta},\sigma^2)\,p(s^2/\sigma^2). \tag{1.24}$$

Por lo que ahora utilizando la distribución inicial de referen-

cia para los parámetros desconocidos en (1.17), tenemos que también se puede escribir como

$$p(\underline{\theta},\sigma^2) \propto \sigma^{-1}. \qquad \begin{matrix} -\infty<\theta_i<\infty \\ 0<\sigma<\infty \end{matrix} \qquad (1.25)$$

Box y Tiao (1971, pp.62) encuentran una expresión para $p(s^2/\sigma^2)$ y obtienen la constante de proporcionalidad para expresar (1.23), es decir, reexpresar a $p(\underline{\theta},\sigma^2/\underline{y},X)$ de la siguiente manera

$$p(\underline{\theta},\sigma^2/\underline{y},X) = p(\sigma^2/s^2)\, p(\underline{\theta}/\hat{\underline{\theta}},s^2), \qquad (1.26)$$

además, se determinaron cada uno de los términos en (1.26); la densidad $p(\sigma^2/s^2)$ le corresponde una ji-cuadrada invertida (Box y Tiao, pp.63) y la densidad $p(\underline{\theta}/\hat{\underline{\theta}},s^2)$, se notó que es la misma que en $p(\underline{\theta}/\underline{y},X)$ (1.21). Entonces, se puede expresar la posterior de ambos parámetros desconocidos mediante

$$p(\underline{\theta},\sigma^2/\underline{y},X) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n-1}{2\sigma^2}\right)^{\frac{n-1}{2}} (s^2)^{\frac{n-1}{2}-1} e^{-\frac{(n-1)s^2}{2\sigma^2}} \frac{|X^tX|}{(\sqrt{2\pi}\sigma)^k} e^{-\frac{1}{2\sigma^2}(\underline{\theta}-\hat{\underline{\theta}})^tX^tX(\underline{\theta}-\hat{\underline{\theta}})}, \qquad (1.27)$$

y reagrupando (1.27), se tiene

$$p(\underline{\theta},\sigma^2/\underline{y},X) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n-1}{2\sigma^2}\right)^{\frac{n-1}{2}} (s^2)^{\frac{n-1}{2}-1} \frac{|X^tX|}{(\sqrt{2\pi}\sigma)^k} e^{-\frac{(n-1)s^2 + (\underline{\theta}-\hat{\underline{\theta}})X^tX(\underline{\theta}-\hat{\underline{\theta}})}{2\sigma^2}} \qquad (1.28)$$

Se concluye que la densidad de $\underline{\theta}$, posee a $\hat{\underline{\theta}}$ el llamado estimador de "Mínimos Cuadrados". Este valor minimiza la forma cuadrática $Q(\theta) = (\underline{y}-X\hat{\underline{\theta}})^tX^tX(\underline{y}-X\hat{\underline{\theta}})$, y se expresa como en (1.5) y (1.6).

De necesitar alguna decisión o inferencia iniciar con la sección 1.1.2 ó 1.1.1, respectivamente y seguir de nuevo aquí (en caso de inferencia).

Box y Tiao (1971, pp. 125 - 127) hacen una justificación bayesiana del empleo de la tabla de Análisis de Varianza, que es comunmente utilizada en la Estadística Clásica, por medio de los siguientes argumentos; la forma cuadrática de $p(\underline{\theta}/\underline{y})$ definida por Ks^2 se distribuye como una F con ciertos parámetros conocidos, (esto es, $Q(\underline{\theta})/_{s^2 K} \sim F_{(K,v,\alpha)}$). Para saber si un punto θ_0 se encuentra en la región de alta densidad $1-\alpha$ es suficiente y necesario que $(\underline{\theta}_0 - \hat{\underline{\theta}})\, X^t X\, (\underline{\theta}_0 - \hat{\underline{\theta}}) < Ks^2 F_{(K,v,\alpha)}$ (Este resultado esta relacionado con la expresión (1.13)). O equivalentemente calcular la cantidad

$$P_r \left\{ F_{(K,v)} < \frac{(\underline{\theta}_0 - \underline{\theta})^t X^t X (\underline{\theta}_0 - \underline{\theta})^t}{Ks^2} \right\} \qquad (1.29)$$

donde $F_{(K,v)}$ es una variable con (K,v) grados de libertad, da la probabilidad contenida de la región de alta densidad que incluye a $\underline{\theta}_0$.

La región de alta densidad de contenido $1-\alpha$ es numéricamente idéntica a la región de confianza de tamaño mínimo, y el complemento de la probabilidad en (1.29) da el nivel de significancia asociado con la hipótesis nula $H_0: \theta = \theta_0$ contra la alternativa $H_a: \theta \neq \theta_0$.

Así clásicamente en la tabla de Análisis de Varianza, se determina si un parámetro está incluido en la región de alta densidad por medio del cálculo de los Cuadrados Medios y el cociente de la $F_{(K,v,\alpha)}$ de tablas.

A continuación se presenta la tabla de Análisis de Varianza

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	COCIENTE DE CUADRADOS MEDIOS.
Discrepancia de parámetros	$(\underline{\theta}_0-\hat{\underline{\theta}})^t X^t X (\underline{\theta}_0-\hat{\underline{\theta}}) = A$	K	$A/K = C$	C/D
Residual	$(\underline{y}-\hat{\underline{y}})^t (\underline{y}-\hat{\underline{y}}) = B$	$n-K$	$B/n-K = D$	
Total	$(\underline{y}-X\underline{\theta}_0)^t(\underline{y}-X\underline{\theta}_0)$	n	1	

En ocasiones, se desea inferir sobre combinaciones de los parámetros y para tales casos, se determina una función $\underline{\Phi}$, con m parámetros definidos de tal manera que $m \leq K$ y

$$\underline{\Phi} = (\phi_1, \cdots, \phi_m)^t \quad , \quad \Phi_i = f_i(\underline{\theta}) \qquad i = \overline{1, K-1}, \qquad (1.30)$$

donde $\underline{\Phi}$ es una transformación lineal de $\underline{\theta}$ ($\underline{\theta} \in R^K$), mediante la obtención de densidad correspondiente, por lo que pueden realizarse las inferencias deseadas.

Las comparaciones de parámetros que se pueden efectuar es mediante las regiones de alta densidad y surgen a partir de problemas estadísticos importantes como: comparaciones de medias o de varianzas en K distribuciones normales.

2. GENERALIZACIÓN DEL MODELO LINEAL.

Una generalización del modelo lineal (Box y Tiao, pp. 176-178) consiste en la aplicación del modelo de regresión lineal a observaciones que no tienen una distribución normal en los errores, pero que pertenecen a la familia exponencial potencia.

El modelo lineal general es entonces

$$\underline{y} = X\underline{\theta} + \underline{\varepsilon}, \qquad (1.31)$$

donde:

> $\underline{y}$ *es un vector de observaciones,* X *una matriz de rango completo con elementos fijos,* $\underline{\theta}$ *vector de coeficientes de regresión desconocidos y* $\underline{\varepsilon}$ *vector de errores aleatorios.*

Suponiendo que los elementos de los errores son independientes y que tienen una distribución exponencial potencia definida por

$$p(\varepsilon/\sigma,\beta) = \omega(\beta)\,\sigma^{-1}\, e^{-c(\beta)\left(\frac{\varepsilon}{\sigma}\right)^{2/(1+\beta)}} \qquad -\infty < \varepsilon < \infty \qquad (1.32)$$

con

$$\omega(\beta) = \frac{\left[\Gamma\left[\frac{3}{2}(1+\beta)\right]\right]^{1/2}}{(1+\beta)\left[\Gamma\left[\frac{1}{2}(1+\beta)\right]\right]^{3/2}} \qquad \sigma > 0\ ,\ -\infty < \theta < \infty\ ,\ -1 < \beta < 1$$

$$c(\beta) = \left[\frac{\Gamma\left[\frac{3}{2}(1+\beta)\right]}{\Gamma\left[\frac{1}{2}(1+\beta)\right]}\right]^{\frac{1}{1+\beta}}.$$

Como última suposición se pide que el modelo no tenga problemas

de multicolinealidad.

A los parámetros σ y β son conocidos como la desviación stándar y medida de la curtosis, respectivamente. El último parámetro β, es una medida de la no normalidad en la distribución orígen. Por ejemplo, si $\beta = 0$ resulta ser la conocida normal y con $\beta = 1$, la doble exponencial.

Las complicaciones en las expresiones de la función de verosimilitud son consecuencias del manejo de tres parámetros $(\underline{\theta}, \sigma, \beta)$ y su representación es como sigue:

$$l(\underline{\theta}, \sigma, \beta | \underline{y}) \propto (\omega(\beta))^n \sigma^{-n} e^{-c(\beta) \sum_{i=1}^{n} \left| \frac{y_i - x_{(i)}^t(\underline{\theta})}{\sigma} \right|^{2/1+\beta}}, \qquad (1.33)$$

con

$$-\infty < \theta < \infty \quad , \quad \sigma > 0 \quad , \quad -1 < \beta < 1$$

$x_{(i)}^t$ es el $i$ésimo renglón de X.

Suponiendo independencia entre los parámetros se puede expresar la densidad inicial conjunta como

$$p(\underline{\theta}, \sigma, \beta) = p(\beta)\, p(\underline{\theta}, \sigma) , \qquad (1.34)$$

que es una manera de manejar separadamente el parámetro de la curtosis y los otros parámetros $(\underline{\theta}, \sigma)$. Utilizando en $p(\underline{\theta}, \sigma)$ una inicial de referencia, la densidad resulta como:

$$p(\underline{\theta}, \sigma, \beta) \propto (\omega(\beta))^n \sigma^{-(n+1)} e^{-c(\beta) \sum_{i=1}^{n} \left| \frac{y_i - x_{(i)}(\underline{\theta})}{\sigma} \right|^{2/1+\beta}} p(\beta) . \qquad (1.35)$$

Integrando con respecto a θ, la expresión (1.35), Box y Tiao la expresa en terminos de

$$p(\underline{\theta},\beta|\underline{y}) = p(\underline{\theta}/\beta,\underline{y})\, p(\beta|\underline{y}), \qquad (1.36)$$

y tomando únicamente la distribución condicional $p(\underline{\theta}/\beta,\underline{y})$ en (1.36) se obtiene

$$p(\underline{\theta}|\beta,\underline{y}) = \left[\int_R [M(\underline{\theta})]^{-\frac{n}{2}(1+\beta)} d\underline{\theta}\right]^{-1} [M(\underline{\theta})]^{-\frac{n}{2}(1+\beta)}, \qquad -\infty < \theta_j < \infty \quad j=1,\ldots,k \qquad (1.37)$$

donde

$$M(\underline{\theta}) = \sum_{i=1}^{n} |y_i - x'_{(i)}(\theta)|^{2/1+\beta}.$$

El estudio de la distribución $p(\underline{\theta},\beta,\underline{y})$ como una función de β, puede determinar el tipo de desviación de la normalidad. Por ejemplo, en el caso de tener $\beta=0$, $M(\underline{\theta}) = (\underline{y}-X\underline{\theta})'(\underline{y}-X\underline{\theta})$ y la distribución en $p(\underline{\theta}/\beta,\underline{y})$ es una $t_K(\hat{\underline{\theta}}, S^2(X'X)^{-1}, \nu)$. Para otros casos cuando $\underline{\beta}$ es pequeña ($K=2$ ó $K=3$), los contornos de $M(\underline{\theta})$ únicamente se pueden graficar.

Como se ve, en el caso de no contar con normalidad en los errores, puede resultar muy difícil el manejo práctico de la distribución final, por las expresiones tan grandes. Sería conveniente efectuar pruebas de hipótesis en $p(\beta/y)$ (1.36) para inferir si hay o no separación de normalidad (es decir, $H_0: \beta = 0$ vs $H_a: \beta \neq 0$) y si no hay mucho problema de la normalidad se podrá hacer una transformación o cambiar el modelo a otro más adecuado, pero si la evidencia apunta a la no normalidad se tendrá que trabajar con el modelo lineal general.

II

METODOS DE SELECCION DE VARIABLES

CAPITULO II

METODOS DE SELECCION DE VARIABLES.

La idea de esta capítulo es la de exhibir algunos de los métodos que existen en la Estadística Bayesiana y cuyo objetivo es la selección de variables en problemas de regresión.

El problema relacionado con Selección de Variables resulta ser un caso particular de la selección de modelos, es decir, cuando en la estadística se presenta el problema de escoger un conjunto de variables explicativas en el Análisis de Regresión, se está seleccionando a la vez el modelo que esta constituído por aquellas variables.

Debido a las dificultades que produce la Multicolinealidad en los modelos (Draper y Smith(1981,pp.258) y Chatterjee(1977,pp.143-192)) de regresión para inferir y predecir, es conveniente hacer una división en la discusión de los conjuntos de variables que presentan este problema, para aplicarles "Un trato especial" previo a los procedimientos de Selección de Variables.

Los conjuntos de variables que no tengan problemas de Multicolinealidad serán directamente sometidos a los métodos de selección de variables que se expondran más adelante.

Existen diversas técnicas que indican y permiten corregir la Mul

ticolinealidad, aquí únicamente se comentará el método de componentes principales (para una discusión más detallada ver Draper y Smith o Chatterjee).

Habiendo eliminado de alguna manera el problema anterior, es necesario validar si en realidad el modelo con el conjunto de todas las variables (modelo completo) explica o predice la variable dependiente, de no ser así, no tendrá sentido efectuar una selección de variables que no satisfacen ésta propiedad, pues cualquier subconjunto de ellas lo hará.

Se comenzará suponiendo que el modelo completo esta formado por p parámetros (o variables explicativas). De aquí, que el número total de posibles modelos a analizar con respecto al modelo completo son $2^p - 1$. Para lo que si no es muy grande esto crece mucho, inclusive para valores relativamente pequeños de p, esto se agranda, por ejemplo si p es igual a 10 resultan 1022 modelos diferentes para analizar.

Para algunos estadísticos la elección del "Mejor modelo" es consecuencia de la bondad (aproximada) de su predicción o explicación; de lo fácil y sencillo del manejo de datos, así como también de su costo de captación de datos. Aunque estas características en general pueden cambiar dependiendo del problema particular y de sus intereses.

Los procedimientos en la literatura bayesiana que se encontraron se calsificaron en dos secciones: 2.2.1 la sección de inferencia, donde se agrupa a los procedimientos que no tienen objetivos predictivos implícitamente, y la segunda sección, 2.2.2 donde se involu-

cran a los métodos que especifican tener metas predictivas. Esta última sección se divide en dos subsecciones: 2.2.2.1 artículos que no se desarrollan en base a la Teoría de Decisiones y 2.2.2.2. artículos involucrados en la teoría de decisiones (porque involucran una función de pérdida).

2.1 MULTICOLINEALIDAD.

El surgimiento del problema de Multicolinealidad ó también llamado datos colineales se debe principalmente a 1) un inadecuado diseño de muestreo, 2) una sobredefinición del modelo, pues existen más variables que observaciones y 3) las características inherentes al procedo de investigación, que ocasionan interrelaciones entre las variables explicativas.

Los síntomas para sospechar Multicolinealidad principalmente son: 1) grandes cambios en los coeficientes estimados cuando una variable es sumada o eliminada, 2) grandes cambios en los coeficientes cuando se excluye o incluye en el análisis un dato, 3) resultados no esperados tanto en los signos como en los coeficientes estimados, y 4) los coeficientes de las variables que se esperaría fueran importantes tienen grandes errores estándar.

Otra indicación para detectar Multicolinealidad se presenta en la matriz de correlación de las variables explicativas, ya que aparecen fuera de la diagonal correlaciones proximas a 1, se dirá que hay Multicolinealidad. Esta indicación no es muy buena, ya que puede existir Multicolinealidad producida por alguna agrupación de las variables explicativas y no indicarse en la matriz de correlación.

El procedimiento de componentes principales es una herramienta estadística del Análisis Multivariado (Chatfield y Collins(1980,pp. 57-62)) que consiste en construir un conjunto de nuevas variables explicativas ortogonales $(z_1, z_2, \ldots, z_n)$, que son formadas con base en combinaciones lineales de las variables explicativas originales $(x_1, \ldots, x_n)$.

Componentes principales se aplica tanto sobre la matriz P de correlaciones como sobre la matriz de varianzas -covarianzas, Σ (se obtienen resultados distintos al usar estas matrices). La matriz de varianzas y covarianzas de las nuevas variables, Λ , es una matriz diagonal cuyos elementos son iguales a las raíces características λ_i ($i = \overline{1,n}$) de la matriz P ó Σ (dependiendo de la matriz que se usó). Estos valores propios son las varianzas respectivas de las nuevas componentes. Por construcción la suma de las varianzas de las variables originales y las suma de las componentes principales son iguales ($\sum_{i=1}^{n} Var(z_i) = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} Var(x_i)$).

En el caso de utilizar inicialmente la matriz P al aplicar componentes principales, la matriz Λ será tal, que en su diagonal puede haber solo unos. Si esto ocurre resulta que las variables $X'\Lambda$ son ortogonales. Pero si este no es el caso, hay al menos una $\lambda_i = 0$ ó muy próxima a cero, entonces se dice que existe Multicolinealidad en las variables explicativas originales.

Cuando los resultados apoyen la existencia de datos colineales hay ciertos métodos en estadística que permiten elegir en su conjunto de componentes principales. Uno de ellos, consiste en elegir las primeras componentes principales[1], digamos m , de tal manera que en conjunto expliquen cierto porcentaje de variación total en los datos originales. Por ejemplo, se aceptan las m primeras λ's si $\sum_{j=1}^{m} \lambda_j / \sum_{i=1}^{n} \lambda_i$ rebasa cierto porcentaje establecido. Otros criterios analizan a la matriz Λ , de la siguiente forma: si λ es aproximadamente cero (generalmente se empiezan analizando la última λ_n , λ_{n-1}) entonces la correspondiente componente principal es aproximadamente constante, de donde se concluye que no se está aportando ninguna variabilidad y esto permite estudiar a la Multicolinealidad. De ésta manera, se puede expresar a una de las variables originales como combinación lineal de las otras.

(1) Esto se debe a la construcción de componentes principales, $\lambda_i \geq \lambda_{i+1}$, $\forall i = \overline{1,n}$.

El paso siguiente consiste en pedir información (extra-matemática) al interesado sobre la situación inherente del problema, para que apoyados en las información[1] evalúen la posibilidad de eliminar alguna combinación lineal de las otras. Todo esto con el objetivo de reducir el número de variables originales al rango de la matriz $\Sigma(\delta\ \rho)$. En el caso de no poder eliminar alguna variable original, se expresa, cada una de las componentes principales diferentes de cero, en términos de las variables X's para plantear el modelo de regresión obtenido. Por ejemplo, aplicando componentes principales se obtienen diferentes λ's, de las cuales, hay K diferentes de cero $(1 \leq K < n)$ y las restantes $n-K$, se determinaron cero. En base a lo anterior, se expresa al modelo de regresión que explica δ predice a Y, en términos de las K componentes principales diferentes de cero, es decir, a las componentes que se determinaron como cero, se estandarizan con respecto a su media y se les suman a término constante del modelo, lo cual permite plantear al modelo en términos de las K componentes principales. Substituyendo en cada una de las i's por las combinaciones respectivas de las Y's, queda resuelto el problema de Multicolinealidad sobre un subespacio (o dimensión menor) del problema original.

Un aspecto importante sobre el procedimiento de componentes principales, es en el siguiente sentido, no se asegura que los modelos obtenidos por este procedimiento sean "buenos" para los objetivos específicos en que fueron creados, solo se sabe que estos nuevos datos no presentan el problema de multicolinealidad y que sus componentes son combinaciones lineales de los datos originales. Por lo que estos modelos pueden ser malos si se trata de explicar a la variable dependiente.

(1) Esta situación va ir de acuerdo con las correlaciones de las variables explicativas.

2.2 MÉTODOS DE SELECCIÓN DE MODELOS.

Antes de iniciar la presentación de los métodos de selección de variables se establecerá una terminología general.

Tomando en cuenta el anterior capítulo, se recordará que el problema de regresión involucra en esencia tres términos: 1) los parámetros $\underline{\theta}$ y ω, donde $\underline{\theta}$ es un vector en $\mathbb{R}^p$ ($\underline{\theta} = (\theta_1, \ldots, \theta_p)^t$) y ω un parámetro real positivo, 2) un vector de observaciones comunmente conocido como $\underline{y}$ de tamaño $1 \times p$ ($\underline{y} = (y_1, \ldots, y_n)^t$) y por último 3) una matriz de datos conocidos de tamaño $n \times p$, que se puede reexpresar como p vectores columna de tamaño $1 \times p$ ($X = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p)^t$).

Suponiendo normalidad en los errores, se tiene que la verosimilitud de la muestra de los datos, es una distribución normal n-variada con media $X\underline{\theta}$ y precisión ωI_n ($p(\underline{y}/\underline{\theta}, \omega, X) \sim N_n(X\underline{\theta}, \omega I_n)$).

Por comodidad se tomará la siguiente notación para hablar de modelos que contengan j variables explicativas con $j = \overline{1, p}$. Se le asignará el nombre de modelo completo a $M_c : f(x_1, \ldots, x_p)$, a todos los demás subconjuntos de variables z_i que tengan menos de p variables explicativas se le asociará un modelo M_i que representa un modelo de regresión sobre las variables explicativas en z_i con $i = \overline{1, 2^p - 2}$.

Luego entonces, el modelo $i$ésimo tendrá la siguiente forma

$$M_i : f(z_i) = z_i \underline{\theta}_i + \underline{\varepsilon} \qquad (2.1)$$

donde $\underline{z}_i$ representa las variables explicativas incluidas en el modelo M_i y $\underline{\theta}_i$ representa el vector de parámetros correspondientes.

Cabe aclarar un punto importante acerca de los procedimientos de que se presentaran. Los eventos de interés relacionados con los modelos no han sido establecidos en todos los métodos de selección, ya que algunos de éstos (por ejemplo: Zellner y Atkinson) no especifican a que evento relacionado con el modelo se refieren. No se sabe si hablan de la probabilidad de que el modelo sea el verdadero ó a la probabilidad de que el modelo contenga un conjunto determinado de variables ó tal vez se ha dejado abierta la especificación de este evento para que dependiendo del caso y las necesidades del interesado se establezcan. Pero como se verá más adelante, existen otros procedimientos de selección (por ejemplo: Davis ó Geisser y Eddy) en donde sí establecen el evento asociado a los modelos, es la probabilidad de que el modelo sea correcto.

2.2.1 INFERENCIA.

Los procedimientos que se presentaran aquí, estan desarrollados por Zellner, Atkinson, Davis, Broemeling y Perichi.

PROBABILIDADES POSTERIORES.

Este procedimiento estriba en elegir un modelo con mayor probabilidad posterior, o sea, el método consiste en la comparación de hipótesis, en donde la información inicial del modelo puede ser involucrada en las probabilidades posteriores (Zellner, 1971, pp.291-298).

Las hipótesis involucradas en este método se relacionan con los modelos de regresión M_i, donde estos pueden ser de $i = \overline{1,m}$ y $m \leq 2^k - 2$, considerando una posible eliminación de los modelos por razones no estadísticas.

Otra característica de estos procedimientos, es que presentan cualidades de exclusividad y exhaustividad, esto es, deben incluir a todos los posibles modelos aceptables y entre ellos, debe de estar incluido el "Modelo Satisfactorio", o como algunos autores lo llaman "Modelo Verdadero".

La probabilidad posterior del modelo i ($i = \overline{1,m}$) con respecto a los datos, es decir $p(M_i / X, \underline{\theta}, \underline{z}_i, \underline{Y})$ se encuentra por medio del producto de la probabilidad asignada al modelo i por la verosimilitud del modelo i fijando a los datos, entre todos los modelos con iguales restricciones, es decir,

$$P(M_i/datos) = P(M_i/X,\underline{\theta},\omega,z_i,\underline{Y}) = \frac{p(M_i)\,L(M_i/datos)}{\sum_{i=1}^{2} p(M_i)\,L(M_i/datos)} \quad . \qquad (2.2)$$

Cada verosimilitud de M_i con respecto a los datos, es el resultado del producto de n funciones de distribución, $p(\underline{y}/M_i, z_i, X_i)$ que involucra a la vez las funciones de densidad, $f_Y(\underline{y}/z_i, \theta_i, \omega)$, para Y, además la función de densidad inicial $f(\underline{\theta}_i,\omega)$, se expresa como

$$L(M_i/datos) = \prod_{j=1}^{n} p(y_j/M_i,z_i) = \prod_{j=1}^{n} \iint_{R^k\,\Omega} f_Y(Y/z_i,\underline{\theta}_i,\omega)\, f(\theta_i,\omega)\, d\underline{\theta}_i\, d\omega.$$

Uno de los pioneros que trabajó este procedimiento de probabilidades posteriores fué Zellner(1971,pp.291-298) que desarrolló su trabajo en el caso más simple.

Zellner considerá únicamente la existencia de dos modelos (M_1, M_2) que explican la variación de la variable dependiente y y donde sólo uno de ellos es el verdadero (aquí, M_1: H_o y M_2: H_o^c).

El vector de observaciones $\underline{y}^t = (y_1, \ldots, y_n)$ está dado en la forma general (1.1), que en esta situación se expresa como:

$$M_1 : \underline{y} = X_1 \underline{\beta}_1 + \underline{u}_1$$

$$M_2 : \underline{y} = X_2 \underline{\beta}_2 + \underline{u}_2 \ ,$$

donde X_1 y X_2 son matrices de $n \times K$ con cantidades conocidas y cada una

con rango K , β_1 y β_2 vectores de coeficientes ($K\times1$) con elementos no comunes, $\underline{u}_1$ y $\underline{u}_2$ vectores de errores ($n\times1$), por último, para el modelo verdadero M_1 ó M_2 se asume que existe independencia en los errores y se distribuye como una normal con media cero y varianza σ_i^2 , $i = 1$ ó 2 .

Se hacen ciertas suposiciones distribucionales para los parámetros que consisten en $p(\underline{\beta}_i, \sigma_i) = p(\beta_i/\sigma_i)\,p(\sigma_i)$, se pide que $p(\beta_i, \sigma_i)$ se distribuya como una normal con media $\bar{\beta}_i$ y varianza $\sigma_i^2 C_i^{-1}$ (C_i es una matriz simétrica definida positivamente de $K\times K$) y $p(\sigma_i)$ se distribuye como una Gamma invertida (Anexo 1) con parámetros q_i y $\bar{s}_i^2$ que son asignados por el investigador ($q_i>0$ y $\bar{s}_i>0, i=1,2$).

Se concluye llegando al "Momio" posterior dado por

$$K_{12} = \frac{P(M_1)}{P(M_2)}\left[\frac{|C_1|/|A_1|}{|C_2|/|A_2|}\right]^{1/2}\left(\frac{\delta_1}{\delta_2}\right)^{-n/2}\frac{\bar{s}_1^2/\delta_1 + q_{1,n}\left(\bar{s}_1^2/\delta_1\right)}{\bar{s}_2^2/\delta_2 + q_{2,n}\left(\bar{s}_2^2/\delta_2\right)} \qquad (2.3)$$

con

$$\delta_i = \frac{1}{n}\left[\upsilon s_i^2 + (\hat{\beta}_i - \bar{\beta}_i)^t C_i (\beta_i - \bar{\beta}_i)^* + (\hat{\beta}_i - \bar{\beta}_i)^t X^t X (\hat{\beta}_i - \bar{\beta}_i)\right], \quad i = 1,2$$

$$A_i = C_i + X_i^t X, \qquad \bar{\beta}_i = A_i^{-1}(C_i\bar{\beta}_i + X_i^t X \hat{\beta}_i)$$

$$\hat{\beta}_i = (X_i^t X_i)^{-1} X_i^t \underline{y}, \qquad \bar{\beta}_i \text{ es el vector de medias inicial para } \beta_i$$

$$\upsilon s_i^2 = (\underline{y} - X_i\hat{\beta}_i)^t(\underline{y} - X_i\hat{\beta}_i), \qquad \upsilon = n-K$$

$$q_{i,n}\left(\bar{s}_i^2/\delta_i\right)$$

$F_{q_i,n}(\bar{s}_i^2/s_i)$ denota la ordenada de la distribución F con q_i y n grados de libertad.

Esta expresión resulta ser un caso particular del cociente (2.2) de probabilidades posteriores.

La expresión (2.3) cambia bajo algunas de estas condiciones; si n es grande se asume $\frac{|C_1|/|A_1|}{|C_2|/|A_2|} \to 1$, y $K_{12} = \left(\frac{S_1^2}{S_2^2}\right)^{-(n+q)/2} e^{-\frac{q\bar{s}^2}{2\sigma_1^2} + \frac{q\bar{s}^2}{2\sigma_2^2}}$, pero si además se cuenta con información inicial "difusa" K_{12}, se reduce a

$$K_{12} = \left(\frac{S_1^2}{S_2^2}\right)^{-\frac{n}{2}} \qquad (2.4)$$

La importancia de estos resultados radica, en que si se cuenta con una función de pérdida simétrica (ver anexo) se elegirá aquel modelo con mayor probabilidad posterior y bajo las condiciones de Zellner, es elegir el modelo con mayor coeficiente de determinación o que es lo mismo que elegir el modelo con menor S^2.

Como último comentario al trabajo de Zellner, éste no fué ubicado en los procedimientos de Teoría de Decisiones porque la función de pérdida que se menciona es muy particular y probablemente no muy común. Esto se debe a que únicamente se tiene dos estados de la naturaleza (θ_1, θ_2) y ambos tienen asociados la misma pérdida en el caso de haber tomado malas decisiones $(L(d_1, \theta_2) = L(d_2, \theta_1) = c)$ y pérdida cero si la decisión es correcta. Además, si se toman ahora tres estados de la naturaleza $(\theta_1, \theta_2, \theta_3)$ con tres decisiones (d_1, d_2, d_3) se concluye que el maximizar la probabilidad no coincide siempre con la decisión de minimizar la pérdida esperada. Por ejemplo, si se tiene la siguiente función de pérdidas, en donde la decisión correc

ta trae una pérdida cero; es decir, $L(d_1, \theta_1) = L(d_2, \theta_2) = L(d_3, \theta_3) = 0$, la pérdida asociada en el caso de haber tomado una mala decisión es la misma en todos los casos $L(d_1, \theta_1) = L(d_1, \theta_3) = L(d_2, \theta_1) = L(d_2, \theta_3) = L(d_3, \theta_1) = L(d_3, \theta_2) = c$, y la probabilidad asignada a los estados de la naturaleza es la siguiente $P(\theta_1) = .4$ y $P(\theta_2) = .3 = P(\theta_3)$. En el caso de maximizar probabilidades se elige como óptimo a d_1 y en el caso de minimizar la pérdida esperada se elige como decisión óptima también a d_1. Por lo que se concluye que en este caso coinciden ambos criterios. Otro ejemplo, basado en el anterior enunciado se mantienen fijas las probabilidades asignadas a los estados de la naturaleza y se establece la función de pérdidas, como a continuación se especifica:

$$L(d_1, \theta_1) = 0 \ , L(d_1, \theta_2 \cup \theta_3) = .6c, L(d_2, \theta_1) = L(d_3, \theta_1) = .4c \ ,$$

$$L(d_2, \theta_2 \cup \theta_3) = L(d_3, \theta_2 \cup \theta_3) = .3c,$$

por lo cual con el criterio de maximizar las probabilidades se elige a d_2 ó d_3, y en minimizar la pérdida esperada se elige a d_1, por lo tanto, hay contradicción y ésta se debe a que el criterio de maximizar probabilidades posteriores viola los axiomas de coherencia, por lo tanto es una mala decisión el usarlo.

Otro trabajo basado en probabilidades posteriores, es el de Atkinson (1978), en donde considera el caso general con m modelos con las características ya antes mencionadas, teniendo entonces modelos en competencia, es decir, todas son modelos que predicen o explican y sólo una de estos puede ser elegido. Una condición que se pedirá es que todos los modelos contengan el mismo número de parámetros, porque de lo contrario acarrearía inferencias arbi-

trarias como por ejemplo, el favorecer a modelos con pocos parámetros.

Otra condición que se pide es que la suma sobre las probabilidades de los modelos condicionados a los datos sume uno.

Atkinson indica que cuando σ^2 es desconocida en los modelos de Regresión, no existe ninguna difultad en este procedimiento, por lo que resulta una expresión análoga a (2.3) y que únicamente se modificará su distribución por una t-student.

Mediante ejemplos de dos modelos en diferentes circunstancias se exponen las fallas y alcances que se presentan a utilizar este procedimiento de probabilidades posteriores, las cuales son: 1) que si los modelos tienen igual número de parámetros y la información inicial es no informativa, entonces en este procedimiento se favorece al modelo incorrecto, 2) para el caso de tener modelos incorrectos, el procedimiento elige de todos los modelos, un modelo cualquiera. De aquí, que se requiera checar los modelos adecuados por medio de un estimador de la varianza; para revisar posible carencia de ajuste, 3) si se trabaja con modelo "separados" (modelos que no están anidados[1], y tienen variables explicativas distintas), entonces las probabilidades posteriores tienden a favorecer asintóticamente al modelo "correcto" ó al modelo con menor número de parámetros, y 4) con modelos anidados correctos se obtienen resultados que apoyan al modelo con menor número de parámetros.

Atkinson hace la suposición de la existencia de un modelo verda-

(1) Se dice que el modelo A está contenido en el modelo B, si todas las variables explicativas de A son variables explicativas en B.

dero, que en la práctica no necesariamente es único, como en el caso de regresión con modelos anidados. Aclarando ésto, en caso de tener un problema de Selección de Variables en Regresión para algunos subconjuntos de variables se tienen modelos anidados, por lo que al asignarle una probabilidad a uno de estos modelos (si el evento relacionado con el modelo es obtener el mejor modelo), se le tendría que asignar una probabilidad mayor o igual a todos aquellos modelos que tengan esas variables explicativas. Esto ocasionaría que el modelo con mayor probabilidad resulte ser el modelo completo, ocasionando que el criterio de probabilidades posteriores siempre eliga el modelo completo.

También existe otro problema que no es propio de este procedimiento de selección, pues se presenta en todos los métodos clásicos o bayesianos. En este caso se ve muy claro cuando el investigador asigna probabilidades a los modelos considerados y no ha podido establecer a "todos" los modelos posibles. Como en este método no es factible tener la opción "otros modelos" resulta que "otros modelos" tienen probabilidades cero, y puede suceder que entre los no considerados existan algunos mejores modelos. Aunque siempre exista un modelo que no se tome en cuenta, se espera que a pesar de ello, se acerquen los resultados a los "verdaderas" soluciones.

Siguiendo con este procedimiento de Probabilidades Posteriores se tiene el trabajo realizado por Davis (1979), que se proporciona una cota para el error obtenido al calcular la probabilidad productiva cuando una densidad inicial aproximada[1] es usada. Es decir, partiendo del principio de estimación estable[2], Davis encuentra

(1) Algunos estadísticos creen en la existencia de densidades iniciales verdaderas y por lo tanto en densidades posteriores verdaderas. Con esta idea se establecen posibles aproximaciones a ellas.

(2) Este principio supone la existencia de un conjunto sobre el cual para una muestra lo suficientemente grande, la distribución inicial es dominada por los datos. Trabajo realizado por Lindeman, Savage y Edwards, en 1963.

una cota para el porcentaje de error en las aproximaciones de las probabilidades posteriores al modelo verdadero y los modelos aproximados.

TEORÍA DE INFORMACIÓN.

A continuación se presenta la investigación de Perichi (1984), trabajo que considera el planteamiento y desarrollo del procedimiento bayesiano utilizado por Zellner (1971) y Atkinson (1978). Renombra este procedimiento por el título de Procedimiento Bayesiano Estándar y lo maneja en términos de cantidad de información esperada acerca de un vector aleatorio.

Debido al análisis de diversas situaciones del procedimiento estándar en que se concluyen los resultados arbitrarios, Perichi encuentra explicaciones y elabora una modificación en su estudio por medio de la cantidad relativa a la ganancia esperada en información.

Con las suposiciones habituales de este procedimiento estándar, se establecen; la existencia de un único modelo verdadero en los modelos alternativos, $M_j : Y = X_{nj}\,\theta_j + \epsilon$ $(j = \overline{1,J})$, que los errores en los modelos son no correlacionados.

Para el caso de contemplar modelos con σ^2 conocida, se asume tener una densidad inicial Normal Multivariada para el parámetro, $p(\theta_j / h)$ tiene media θ_{oj} y precisión $V_{oj}^{-1} h$ donde $h = \sigma^{-2}$ (es decir, $p(\theta_j / h) \sim N_m(\theta_{oj};\ V_{oj}^{-1} h)$, k_j es la dimensión de θ_j y la probabilidad inicial del j-ésimo modelo, es $\frac{1}{m}$ $(P_r(M_j) = w_j = 1/m)$.

Perichi plantea el procedimiento estándar ya presentado anteriormente:

$$S_{mi} = \left(\frac{R_{mi} + R_{nm} + Q_{ni} + Q_{nm}}{2\theta^2} \right) + \frac{1}{2} \log \left(\frac{|V_{om}^{-1}|\,|V_{oi}^{-1} + V_{ii}^{-1}|}{|V_{om}^{-1} + V_{im}^{-1}|\,|V_{oi}^{-1}|} \right), \tag{2.5}$$

donde

$$\theta_{nj} = (\hat{\theta}_{nj} - \theta_{oj})^t (V_{oj} - V_{ij})^{-1} (\hat{\theta}_{nj} - \theta_{oj}), \qquad j = \overline{m,i}$$

$V_{ij}^{-1} = X_{nj}^t X_{nj}$, R_{nj} , y

$\hat{\theta}_{nj}$ *son las sumas de cuadrados de los residuales estandarizados de*

$\hat{\theta}_j$ *es el estimador de mínimos cuadrados de*

El segundo término de (2.5) es igual a la diferencia entre la ganancia esperada en información acerca de los parámetros entre los modelos, esto es

$$\frac{1}{2} \log \left(\frac{|V_{om}^{-1}|\,|V_{oi}^{-1} + V_{ii}^{-1}|}{|V_{om}^{-1} + V_{im}^{-1}|\,|V_{oi}^{-1}|} \right) = I^{\theta_i}\{\varepsilon, p(\theta_i)\} - I^{\theta_m}\{\varepsilon, p(\theta_m)\}, \tag{2.6}$$

donde análogamente para:

$$I^{\theta_i}\{\varepsilon, p(\theta_i)\} = H\{p(\theta_i)\} - \int p(y)\, H\{p(\theta_i|y)\}\, dy$$

$$H\{p(\theta_i)\} = -\int p(\theta_i) \log\{p(\theta_i)\}\, d\theta_i$$

$$p(y) = \int p(y|\theta_i)\, p(\theta_i)\, d\theta_i$$

Considerando a σ^2 desconocida, se hace la suposición que la conjugada inicial bajo M_j , tiene densidad Normal Gamma y se escribe como :

$$p(\theta_j, h_j) = p(\theta_j/h_j)\, p(h_j) \tag{2.7}$$

Esta última expresión, Perichi la plantea en términos de información esperada, como sigue:

$$I^{\theta,h}\{\epsilon, p(\theta,h)\} = I^{\theta}\{\epsilon, p(\theta/h_j)\} + I^{h}\{\epsilon, p(\theta;h)\}. \tag{2.8}$$

Y además, la densidad de h_j ($h_j = \sigma_j^{-2}$) es una Gamma con parámetros d_j y β_j. En base a esto, se obtiene un procedimiento estándar para σ^2 desconocida.

$$S_{m_i} = \log\left\{\frac{\beta_m^{d_m}\,\beta_{ni}^{d_{ni}}\,\Gamma(d_{nm})\,\Gamma(d_i)}{\beta_i^{d_i}\,\beta_{nm}^{d_{nm}}\,\Gamma(d_m)\,\Gamma(d_{ni})}\right\} + \frac{1}{2}\log\left\{\frac{|V_{oi}^{-1}+V_{ii}^{-1}|\,|V_{om}^{-1}|}{|V_{oi}^{-1}|\,|V_{om}^{-1}+V_{im}^{-1}|}\right\}, \tag{2.9}$$

donde

$$\beta_{nj} = \tfrac{1}{2}(2\beta_j + \beta_{nj} + \theta_{nj}), \quad j = \overline{m,i} \quad \text{y}$$
$$d_{nj} = d_j + \frac{n}{2}.$$

Perichi dice que bajo la suposición (2.7) y (2.8) se obtiene

$$I^{\theta,h}\{\theta, p(\theta,h)\} = \log\left\{\frac{|V_0^{-1}+V_1^{-1}|^{1/2}\,\Gamma(\alpha)}{|V_0^{-1}|^{1/2}\,\Gamma(d_n)}\right\} + \alpha_n\{\psi(\alpha_n)-1\} - \alpha\{\psi(\alpha)-1\}, \tag{2.10}$$

donde

$$\psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha) \quad \text{es la función digamma}$$

y usando (2.10), *se expresa nuevamente* (2.9)

$$S_{mi} = \ln \log \left(\frac{\theta_{mi}}{\theta_{nm}} \right) + I^{\theta_i, h}\{\varepsilon, p(\theta_i, h)\} - I^{\theta_m, h}\{\varepsilon, p(\theta_m, h)\} \qquad (2.11)$$

El autor concluye que estos resultados ocasionan por su definición diversos problemas, como por ejemplo, la expresión (2.6) *crea engañosas regiones en el espacio paramétrico*[1] *y* (2.11) *plantea resultados inconsistentes.*

Perichi propone una solución a este problema, que consiste en asignar pesos iniciales a los modelos sin discriminar a ningún modelo ($P_r(\omega_j) = 1/3$). *Define al vector de parámetros desconocidos como* $\underline{\eta}$ *y* $a M$, *una variable aleatoria que toma J modelos ajenos y simples* (exclusivos). *El objetivo principal es maximizar la pérdida de información acerca de* M *y maximizar lo que se espera ser aprendido acerca de los parámetros de los modelos.*

Estableciendo simbólicamente lo que se espera ser aprendido con respecto a $\underline{\eta}$ *sobre la distribución posterior, se establece para* $\underline{\omega} = (\omega_1, \ldots, \omega_j)$

(1) Esto se refiere a que el procedimiento estándar infla la probabilidad posterior de los modelos, que tienen un incremento de información esperada pequeño acerca de sus parámetros.

$$R\{\varepsilon, p(\eta,M)\} = \sum_j w_j \left[I^{\eta_j}\{\varepsilon, p(\eta_j)\} - \log w_j \right]. \qquad (2.12)$$

Para la maximización de (2.12) se toma $\vec{w} = (\bar{w}_1, \ldots, \bar{w}_j)$, que es,

$$\max_{w} \; R\{\varepsilon, p(\eta,M)\}. \qquad (2.13)$$

Esta solución, se expresa como:

$$\log \bar{w}_i - \log \bar{w}_k = I^{\eta_i}\{\varepsilon, p(\eta_i)\} - I^{\eta_k}\{\varepsilon, p(\eta_k)\}, \qquad (2.14)$$

para modelos ajenos y simples

Algunas observaciones sobre este criterio son: si $I^{\eta_i}\{\varepsilon, p(\eta_i)\} < I^{\eta_k}\{\varepsilon, p(\eta_k)\}$, en base a (2.14) se penaliza a M_i, si la ganancia esperada en información acerca de los parámetros es igual para todos los modelos en competencia, el criterio (2.14) coincide con el procedimiento estándar, ya sea (2.5) ó (2.11) dependiendo del caso, y más aún, cuando todos los parámetros son exactamente conocidos, la razón de los momios posteriores es igual a la razón de verosimilitudes.

Procedimientos de Broemeling (para el caso de inferir sobre los parámetros).

Cuando se tiene el objetivo de inferir sobre los parámetros, un criterio de selección es el propuesto por Broemeling (1985, pp. 97-99), quien indica seleccionar el modelo con mayor determinante de los coeficientes de regresión en la matriz de precisión de la distribución posterior marginal.

La matriz de precisión de la distribución marginal de $\underline{\theta}$ es

$$H = (n-\rho)\, X^t X \left(\underline{y}^t \underline{y} - \underline{y}^t X (X^t X)^{-1} X^t \underline{y} \right)^{-1}, \qquad (2.15)$$

donde ρ es el número de parámetros en el modelo.

En la siguiente sección se presentarán otros procedimientos ubicados en la elección de modelos con objetivos predictivos.

2.2.2 MÉTODOS QUE ESPECÍFICAN LA PREDICCIÓN.

2.2.2.1 PROCEDIMIENTOS QUE NO UTILIZAN FUNCIONES DE PÉRDIDA.

El siguiente trabajo que se describe lo desarrollaron conjuntamente Geisser y Eddy (1979), en donde al parecer los resultados tienen un amplio alcance operacional y resulta muy sencillo de aplicarse con la ayuda de las computadoras. Los antecedentes de este método están en los artículos de Stone (1974,1976) y Geisser (1975), quienes al mismo tiempo y de manera independiente trabajaron sobre la misma idea de Selección de Variables, por lo cual ambos nombran a sus métodos de diferente manera. Stone a su procedimiento lo nombra como validación cruzada (CROSS-VALIDATION) y Geisser lo llama como reuso de la muestra predictiva (Predictive sample reuse, PSR) aunque ambas significan lo mismo.

Como el problema de regresión se tiene que las densidades predictivas de $\underline{y}$ *dado* X *son sensibles a los datos, se introduce el procedimiento de validación cruzada (o PSR) para eliminar la observación* (x_i, y_i) *para juzgar la potencia predictiva del modelo con la densidad predictiva de* $\underline{y}$ *dado* X *en el valor* x_i $(S_{(i)} = S - \{(x_i, y_i)\}$, *con* S *igual a* $\{(x_i, y_i)\ \forall\ i = \overline{1,n}\} = S)$. *Y se concluye con una función que depende del modelo* α $(\alpha = \overline{1,m})$, *que se plantea así:*

$$A(\alpha) = \sum_{i=1}^{n} \log f(y_i \mid x_i, \alpha, S_{(i)}) \qquad (2.16)$$

En el artículo de Stone 1976, se muestra que el criterio de validación cruzada es asintóticamente equivalente bajo ciertas condiciones al criterio de Akaike (esto es, un criterio para seleccionar modelos).

El criterio de Akaike estriba en seleccionar al modelo M que maximize a

$$L(M_\alpha, \hat{\theta}_{M_\alpha}) - P_{M_\alpha} \quad , \alpha = \overline{1,m}. \tag{2.17}$$

donde $L(M_\alpha, \hat{\theta}_{M_\alpha})$ es el logaritmo de la función de verosimilitud, $\hat{\theta}_{M_\alpha}$ es el estimador máximo verosímil del parámetro θ_{M_α} en el modelo M_α y P_{M_α}, el número de parámetros. Por otro lado, se tiene que el criterio de Akaike es equivalente si los modelos tienen varianza conocida, a otro procedimiento de selección en modelos de regresión lineal múltiple, que consiste en minimizar la C'p de Mallow's (Draper-Smith, 1981, pp.299), esta es:

$$C_p = \frac{SCR_{M_\alpha}}{\sigma^2} - (n - 2P_{M_\alpha}) \qquad \alpha = \overline{1,m} \tag{2.18}$$

donde SCR_{M_α} es la suma de cuadrados residuales para el modelo M_i, n el tamaño de la muestra y σ^2 la varianza conocida.

Continuando con el trabajo de Geisser y Eddy (1979), estos también omiten la posible influencia de la i-ésima observación de la muestra cuando se calcula la densidad predictiva de X_i. Además, se denomina la validación cruzada mediante el nombre de predictive sample reuse (PSR), y se deriva un criterio que estriba en maximizar el producto de predictivas condicionales.

Se define el vector de datos $X_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ con la -ésima observación eliminada, con base en ésto, la densidad predic

tiva de x_i cuando M_α es el modelo verdadero, se representa como $f_i(x_i / X_{(i)}, S_i, M_\alpha)$. En base a esto, se dá el criterio de selección que consiste en elegir aquella α^* que minimice L_α, eligiendo entonces al modelo M_{α^*}, donde

$$L_\alpha = \prod_{i=1}^{n} f_i(x_i / X_{(i)}, S_i, M_\alpha) \quad , \alpha = \overline{1,m} . \tag{2.19}$$

Se presentan dos maneras distintas para obtener la densidad predictiva de observaciones futuras, cuando M_α es verdadero, que se sustentan en: 1) como $f(x / \theta_\alpha, S_i, M_\alpha)$ es la densidad de las observaciones y $\hat{\theta}_\alpha$ es el estimador máximo verosímil, calcular $f(x / \hat{\theta}_\alpha, S_i, M_\alpha)$, esta densidad es llamada "densidad predictiva cuasi-verosímil" y 2) calcular la densidad predictiva bayesiana obtenida de una distribución inicial vaga.

El siguiente resultado, es un caso particular de este procedimiento, pues sólo es válido para el anidamiento de modelos, esto es, si M_α está anidado en $M_{\alpha'}$ ($M_\alpha \subset M_{\alpha'}$), se tiene que la densidad predictiva cuasi-verosímil bajo M_t es $f_i(x_i / S_i, \hat{\theta}_{t(i)}, M_t)$, donde $\hat{\theta}_{t(i)}$ es el estimador máximo verosímil de θ_t omitiendo a x_i, entonces si además M_α es verdadero, el cociente $\hat{L}_\alpha / \hat{L}_{\alpha'}$ converge a

$$\ell^{-p(M_\alpha) + p(M_{\alpha'}) + \log \lambda} \quad , \text{ si } n \to \infty . \tag{2.20}$$

donde $p(M_\alpha)$ y $p(M_{\alpha'})$ es el número de parámetros desconocidos en θ_α y $\theta_{\alpha'}$, respectivamente

$$\lambda = \prod_{i=1}^{n} \frac{f_i(x_i / S_i, \hat{\theta}_\alpha, M_\alpha)}{f_i(x_i / S_i, \hat{\theta}_{\alpha'}, M_{\alpha'})}$$

Para este caso, los autores concluyen diciendo que (2.20) es esencialmente el criterio de Akaike. Si $\lambda<1$ se elige al modelo M_1, y en caso contrario ($\lambda>1$) se elige al modelo M_2.

Como última contribución, los autores trabajaron otras situaciones, donde no se cuenta con una función de verosimilitud para al menos uno o más de los modelos involucrados, por lo que se introduce una medida de la discrepancia puntual entre x_i y $\hat{x}_i$, con $i=\overline{1,n}$, dada por $D_k=\sum_{i=1}^{n} d(x_i,\hat{x}_i)$, donde $d(x_i,\hat{x}_i)$ es alguna medida de la distancia entre x_i y $\hat{x}_i$. Este resultado, se generaliza para los modelos de regresión múltiple, y se concluye eligiendo al modelo que tenga una discrepancia menor.

Broemeling (1985,pp.97-99) también aporta esta sección de métodos predictivos un procedimiento que involucra a la distribución predictiva de las observaciones futuras ($P_k(y_0/S)$), donde su limentación depende en conocer previamente el valor a predecir, x_0; S es la muestra y y_0 la observación futura, κ, el k-ésimo modelo y $P_k(y_0/S)$ se distribuye como una T-Student, para los m modelos en general. El criterio consiste en elegir el modelo que haga máxima la distribución predictiva en las observaciones futuras. O sea, elegir M_L si

$$P_L(y_0|S) > \max_{m\neq L} P_m(y_0|S) \quad , \quad 1\leq \kappa \leq m. \tag{2.21}$$

Los procedimientos que toman en cuenta explícitamente el problema de predicción en el Análisis de Regresión son los procedimientos que a continuación se presentan.

2.2.2.2 TEORÍA DE DECISIONES.

En esta sección se presentan cuatro trabajos desarrollados en una misma línea bayesiana, que radica en plantear el problema de selección de variables con objetivos predictivos en base a un problema de decisión. Estos trabajos son realizados por Lindley (1968), Brooks (1972,1974) y Goldeisten (1976). Los tres primeros tienen un desarrollo común dentro de la teoría de decisiones y son consecutivos en su realización. En el último artículo se emplea distribuciones iniciales no paramétricas.

El trabajo de Lindley (1968) plantea las bases para tratar el problema de selección de variables por medio de la teoría de decisiones. Se denotará por H, a todo el conocimiento disponible para tomar la decisión de cuál es el mejor modelo incluyendo a los datos, tanto de las variables independientes como de la variable dependiente.

Enseguida se presentan las suposiciones estructurales, con las que Lindley desarrolla el trabajo. Para algunas de ellas su formulación resulta natural, debido a que van de acuerdo con las suposiciones usuales del Análisis de Regresión.

Suposición 1: $\underline{\theta}$ y X son variables aleatorias en $\mathbb{R}^r$ (espacio euclídeano de r dimensiones) las cuales son independientes condicionalmente a H. Esto es, su densidad conjunta se puede factorizar de la siguiente manera:

$$p(\underline{\theta}, X \mid H) = p(\underline{\theta} \mid H)\, p(X \mid H) \qquad (2.22)$$

Suposición 2: $\underline{y}$ *es una variable aleatoria en* $\mathbb{R}^n$ *con densidad* $\rho(\underline{y}/\underline{\theta},X,H)$, *la cual no depende de* H. *Es decir* $\rho(\underline{y}/\underline{\theta},X,H)=\rho(\underline{y}/\underline{\theta},X)$ *y su densidad tiene la propiedad de*

$$E(\underline{y}/\underline{\theta},X)=\underline{\theta}^t X$$
$$Var(\underline{y}/\underline{\theta},X)=\sigma^2 \qquad (2.23)$$

Las siguientes suposiciones presentan restricciones fuertes en su planteamiento.

Suposición 3: σ^2 *es conocida. Se denotará por* I *a un subconjunto (puede ser vacío) de los enteros* $1,2,\ldots,r$ *tal que contenga* S *elementos y a su respectivo complemento, que tiene* $r-s$ *elementos, se le nombrará mediante* J.

Sea X *tal que* $X=(X_1,\ldots,X_r)$ *y se define a* X_I *de tal forma que sea el vector cuya componente* X_i *es tal que* $i \in I$, *análogamente se define* X_J. *O sea, mediante* X *se denota a las variables explicativas y con* I *se determina a un subconj-nto* X_I *de éstas.*

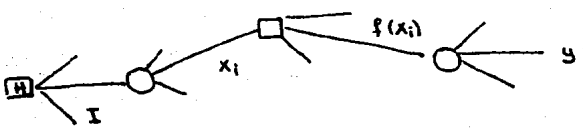
Suposición 4: El *espacio de decisión consiste de elementos* $d=(I,\delta(\cdot))$; *donde* I *ya se describió anteriormente y representa los subconjuntos de variables explicativas que se van a considerar y por* $\delta(\cdot)$, *a una función de* $\mathbb{R}^s$ *a* $\mathbb{R}$ *que es llamado el "predictor de* Y*" porque evalúa la predicción* $\delta(x_I)$, *en todos los subconjuntos de variables explicativas.*

Suposición 5: La *función de pérdida es:*

$$\left\{\underline{y} - f(x_I)\right\}^2 + c_I \quad \text{con} \quad c_I \geq 0, \qquad (2.24)$$

donde $\underline{y}$ es el valor por predecir de la variable independiente[1].

Para comprender esta última suposición se presenta la anatomía o estructura cualitativa de un problema de selección de variables, como un arreglo cronológico en un árbol de decisión



Este árbol cuenta en su diagrama con nodos cuadrados y circulares. Los primeros nodos denotan situaciones de elección o decisión, y los segundos simbolizan situaciones que están influídas bajo un proceso aleatorio. El árbol se lee de izquierda a derecha. La figura anterior representa un problema de selección de variables con objetivos predictivos, el primer nodo representa el conocimiento particular del marco de referencia en que se desenvuelve el problema, así como la decisión de cuales serán las variables explicativas que se van a observar. Ya teniendo claro cuales son las variables explicativas involucradas (en I) para el problema, estas se observarán en condiciones generales y se obtendrán los valores, x_I. Para estos valores se calcularían las predicciones, $f(x_I)$, y se llega al problema de decidir cual es la mejor predicción de los subconjuntos de variables explicativas que predicen a $\underline{y}$. Finalmente considerando al verdadero valor de $\underline{y}$, se evalúa la función de pérdida descri

(1) Entendiéndose por esto, que $\underline{y}$ es un valor que va a ocurrir (ó ya paso).

ta en (2.24) y con la función de pérdida se describe la exactitud de la predicción realizada.

Teóricamente esta situación se resuelve de derecha a izquierda para analizar su desarrollo y planteamiento, esto es, si se minimizó la pérdida esperada en (2.24) quiere decir, que la diferencia entre y y $f(x_I)$ fué muy pequeña de donde, $f(x_I)$ es un buen predictor que está siendo obtenido por un buen subconjunto de variables explicativas y a las que además, se les tomó en cuenta su costo de obtención.

Las suposiciones 4 y 5 sustentan el desarrollo del trabajo de Lindley en cuanto a la teoría de decisiones. A continuación se presentarán otras tres suposiciones utilizadas en el desarrollo de Lindley.

Suposición 6: La distribución $p(X/H)$ (que cumple la suposición 1) tiene la siguiente característica

$$E(X_J / X_I) = B_{JI} X_I \qquad (2.25)$$

con

B_{JI} matriz de $(r-s) \times s$

Suposición 7: Ahora H contiene n observaciones estocásticamente independientes $y_1, \ldots, y_n$ con valores en x_{ij} $(i = \overline{1,n}; j = \overline{1,n})$ de las variables independientes con densidad $p(y_i / \underline{\theta}, \underline{x}_i)$ dadas por la suposi-

ción 2 y se tienen las siguientes relaciones

$$\mathcal{E}(\underline{y}/x,\underline{\theta}) = x\,\underline{\theta} \quad y \quad \mathcal{V}(\underline{y}/x,\underline{\theta}) = \sigma^2 I_n \tag{2.26}$$

Suposición 8: Las $\underline{X}_i$ (de la suposición 7) son variables aleatorias idénticamente distribuidas según una normal multivariada y además, se supone independencia de la variable aleatoria X, vector en el cual se realiza la predicción con los valores $\underline{X}_i$. Los parámetros de la distribución de X son independientes de $\underline{\theta}$.

El siguiente supuesto establece el manejo de distribuciones iniciales de referencia.

Suposición 9: Se denotará por H_0 a la información inicial que plantea el interesado. Dado este conocimiento, H_0, la distribución de $\underline{\theta}$ es uniforme sobre $\mathbb{R}^n$, la distribución de las medias, de la distribución normal de X es similarmente uniforme sobre $\mathbb{R}^r$, la distribución de la matriz de dispersión de la distribución normal de $\underline{X}$ es inversamente proporcional a su determinante. Y además se pide que estas tres distribuciones sean independientes.

Bajo las anteriores suposiciones, Lindley establece un teorema en donde se obtiene la solución óptima del problema de predicción, que consiste en elegir el subconjunto de variables independientes I que satisfaga

$$\min_{I} \left\{ \frac{R(I:I)}{n} + C_I \right\} + \sigma^2\left(1+\frac{r}{n}\right) \tag{2.27}$$

y se predice a y mediante $E(\underline{\theta})^t \; E(X/X_I)$

donde

$$R(J:I) = E(\underline{\theta}_j)^t V(X_j) \; E(\underline{\theta}_j),$$

$R(J:I)$ es la reducción en suma de cuadrados debida a $\underline{\theta}_j$ fijando $\underline{\theta}_I$, $V(X_j)$ es la partición de $V(W/H) \sim \frac{W^t W}{n}$ (1) y está, es la matriz de varianzas-covarianzas.

Eliminando los términos constantes en (2.27), se encuentra que la solución óptima para el problema de predicción radica en seleccionar a la I que satisfaga

$$\min_I \left\{ \frac{R(J:I)}{n} + C_I \right\}. \qquad (2.28)$$

Cuando este resultado se compara con otros criterios ó técnicas, por ejemplo con la estadística C_p de Mallows[2], se encuentra una relación muy cercana entre ambas técnicas bajo una transformación lineal si se toman costos iguales. O sea, si por R_I se denota a la suma de cuadrados residual fijando las variables x_i, $i \in I$, se define como

$$C_p = \frac{R_I}{\hat{\sigma}^2} - (n-2p), \qquad (2.29)$$

(1) $V(W/H) = \frac{W^t W}{n}\frac{(n+1)}{(n-3)}$ donde W es una matriz cuyo elemento típico es $x_{ij} - x_{\cdot j}$. La matriz aproximadamente es igual a la matriz de dispersión muestral.

(2) Esta estadística ya fué presentada en el artículo de Geisser y Eddy (1979).

donde σ^2 es la varianza estimada, $R_I = R(J:I) + R$, R es la estimación residual de todas las variables.

Por lo que basándose en la suposición (que es muy restrictiva), en donde cada una de las variables tienen el mismo costo ($C = 2\sigma^2/n$) se obtiene.

$$C_p \frac{\sigma^2}{n} = \left\{ \frac{R(J:I)}{n} + C_I \right\} \rightarrow \frac{R}{n} - \sigma^2. \qquad (2.30)$$

Como ejemplos, se presenta el procedimiento aplicado a los datos de Hald que tienen cuatro variables explicativas. Se les trabajó con la suma de cuadrados residuales y suponiendo aditividad en sus costos, se elige en cada caso el mínimo de ellas. Es decir utilizando la expresión más simple de la solución del problema de predicción (2.28). La decisión óptima es consecuencia de elegir el modelo que minimice la siguiente expresión:

$$\min_I \left\{ \frac{R(J:I)}{n} + C_I \right\} = n \min_I \left\{ R_I - R + nC_I \right\}$$

$$= \min_I \left\{ R_I + nC_I \right\}.$$

TABLA No.1 *ANALISIS DE PREDICCION EN LOS DATOS DE HALD*

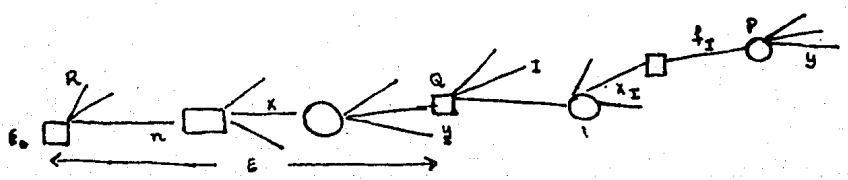
Variables en I	R_I	R_I+nC_I	R_I+nC_I	R_I+nC_I
ningúno	2716	2716	2716	2716
1	1266	1766	1366	2266
2	906	1406	1006	1006
3	1939	2039	2439	2039
4	884	984	1384	2884
12	58	1058	258	1158
13	1227	1827	1827	2327
14	75	675	675	3075
23	415	1015	1015	615
24	869	1469	1469	2969
34	176	376	1176	2276
123	48+	1148	748	1248
124	48-	1148	748	3148
134	51	751	1151	3151
234	74	774	1174	2274
todos	48	1248	1248	3248
costos	nC_1	500	100	1000
	nC_2	500	100	100
	nC_3	100	500	100
	nC_4	100	500	2000

Considerando ahora los trabajos de Brooks (1972 y 1974) se encuentra que si se parte del procedimiento de Lindley, planteando ahora el problema de encontrar un diseño óptimo de regresión, para encontrar las mejores variables para poder predecir. Esto es, realizar un diseño experimental que además de lo que encontró Lindley, se tome en cuenta que variables se van a observar, en que rango de variabilidad

se van a medir, en que momento se van a observar, costos de observación, costos de obtención, costos de predicción y otros más.

Un aspecto importante del trabajo de Brooks, se debe a la ampliación del trabajo de Lindley para cuando σ^2 es desconocida, que prácticamente consiste en estimar a σ^2.

El arreglo cronológico del árbol de decisión, se plantea así:



En general, el problema de encontrar el diseño óptimo de regresión, es un planteamiento más completo que el discutido por Lindley, resultando que las expresiones involucradas son más complicadas. Sin embargo, sus expresiones finales involucran el criterio de decisión de Lindley.

La extensión de σ^2 desconocida, Brooks la soluciona en base a la suposición 1, pues ahí se asumió que el vector de parámetros $(\underline{\theta}, h = 1/\sigma^2)$ y X son variables aleatorias, dado el conocimiento inicial, ξ_0. Mediante lo anterior dice que lo importante es reemplazar a σ^2 por $E(\sigma^2/\xi_0)$, donde $E(\sigma^2/\xi_0)$ representa el valor esperado de la distribución inicial de σ^2.

El criterio de decisión consiste en elegir el modelo que haga mínima la siguiente expresión:

$$\min_{I} \left\{ \omega + tr\left[\omega M (Z'Z)^{-1}\right] + C_L + C \right\} \qquad (2.31)$$

donde

$$\omega = E(\sigma^2/\varepsilon_0), \; M = E(ZZ'/\varepsilon_0), \; Z' = (\underline{1}, X')$$

C denota el costo del experimento por unidad de error en la predicción $C(n,X)$ y C_I denota el costo de las observaciones de las variables X_I por costo de unidad de error en la predicción.

Goldstein (1976) plantea su expresión sobre el problema de regresión bayesiano usando una distribución no paramétrica para el parámetro desconocido θ. El manejo y planteamiento de las distribuciones iniciales no paramétricas se encuentran en un trabajo de 1975 de Goldstein.

Tomando en cuenta todas las consideraciones dadas en el problema de regresión (Capítulo 1), el autor señala la no justificación que se hace siempre al pedir entre las variables explicativas y la variable dependiente una relación lineal (esto es, $E(\underline{y}/X=x) = X\underline{\beta}, \forall x$), por lo que propone realizar una correcta selección de $\underline{\theta}$, considerando un buen predictor de Y, de esta manera para el caso de tener a σ^2 conocida se establece un criterio razonable para obtener un estimador bayesiano de $\underline{\theta}$, que consiste en seleccionar el valor que minimice el riesgo $R(\beta)$, donde

$$R(\beta) = E(y - \Sigma \beta_i x_i)^2 \qquad (2.32)$$
$$R(\beta) = E((y-\Sigma\beta_i x_i)^2 / y, x, \beta).$$

Este criterio da el mejor predictor lineal de Y dado X. Los valores óptimos, β^* *, que minimizan* $R(\beta)$ *y* $R(\beta^*)$*, están dados por:*

$$\beta^* = D^{-1}b \qquad y \qquad R(\beta^*) = \frac{\begin{vmatrix} D & b \\ b' & c \end{vmatrix}}{|D|} \qquad (2.33)$$

con

$$D = (d_{ij}) \qquad d_{ij} = E(x_i x_j / y, x, \beta), \quad i,j = \overline{1,p},$$

$$b = (b_1, \ldots, b_p)^t, \; b_i = E(y x_i / y, x, \beta) \; y \; c = E(y^2).$$

Goldstein prueba que para n grande, el estimador de mínimos cuadrados (capítulo 1) es una buena aproximación para β^* *sobre un rango amplio de distribuciones iniciales del parámetro desconocido y se dice que este resultado es de gran ayuda cuando el cálculo de (2.33) sea muy difícil.*

Considerando el problema de seleccionar el óptimo subconjunto de variables de regresión se propone realizar la comparación del riesgo expresado en $R(\beta^*)$*. Este procedimiento no especifica el tamaño a partir de cual, el riesgo es inaceptable, por lo que se espera que el interesado ó estadístico lo determine en base a sus necesidades.*

El planteamiento para saber si un subconjunto de variables puede ser eliminado es el siguiente: Se define S_k *un conjunto de elementos* $\{i_1, i_2, \ldots, i_k\}$ *y la matriz* $D(S_k)$ *de tamaño* $(p-k)\times(p-k)$ *que es el resultado de la reducción de renglones y columnas de la matriz* D *y* $b(S_k)$*,*

vector de $(p-k)$ que proviene de la reducción de $i_1, \ldots, i_k$ entradas en $\underline{b}$, entonces

$$R_{S_k}(\beta^*) = \frac{\begin{vmatrix} D(S_k) & \underline{b}(S_k) \\ \underline{b}^T(S_k) & g_k \end{vmatrix}}{|D(S_k)|}, \qquad (2.34)$$

$$\Delta = R_{S_k}(\beta^*) - R(\beta^*). \qquad (2.35)$$

Por Δ se denota al riesgo incurrido si omitimos las variables $(x_{i_1}, \ldots, x_{i_k})$ que significa el incremento en el valor mínimo de (2.32), si se omiten las variables y también se pueden eliminar subconjuntos de variables para las cuales Δ no sea mayor que un valor determinado anticipadamente. Este valor fijo establecerá un compromiso entre las demandas de un buen modelo para predecir y la simplicidad para predecir.

III

CONCLUSIONES

CONCLUSIONES

A continuación se presentan las conclusiones que se desprenden del anterior capítulo referentes a los procedimientos de selección de variables, las cuales son: procedimientos de probabilidades posteriores, procedimiento de Broemeling, procedimiento de Geisser y Eddy, procedimiento de Lindley y procedimiento de Goldstein.

Se comenzará comentando una dificultad global referente a la funcionalidad de los procedimientos de selección de variables, ya sea tomando el punto de vista Clásico o Bayesiano. Esta dificultad se presenta en todos los procedimientos de selección que se basan en el cálculo de "todas" las regresiones; como el número de posibles modelos a calcular depende del número de variables explicativas (a saber p), el total de posibles modelos a calcular dependerá de la cantidad de variables explicativas con las que se cuenten, que en general serán $2^p - 1$ modelos distintos. En la práctica puede ocurrir que el número de variables explicativas sea grande, por lo que no es recomendable aplicar de inmediato alguno de los procedimientos antes mencionados de selección, el adecuado a juicio del interesado o el estadístico, sino que se sugiere realizar una "selección" de variables previa, por medio de un criterio extramatemático o un criterio de selección (por ejemplo backware o stepwise) que al no presentar este problema, permite eliminar algunos de los modelos menos importantes (dependiendo del caso a analizar) y permite obtener una idea sobre el conjunto de variables "óptimo". Con esta información en mente se aplicará alguno de los procedimientos de selección estudiados en este trabajo. De no seguir ésta recomendación, se tendrá que usar una cantidad impresionante de tiempo de máquina, pues los resultados requerirán de tal vez años de uso de computadora.

Tomando como referencia los resultados que Atkinson obtuvo, se vé que este criterio no es posible de aplicar tal cual a los modelos de regresión, ya que se tienen graves consecuencias al asignar las probabilidades a la mayoría de los modelos de regresión, porque éstos en general son modelo anidados y al darle "un peso" a un subconjunto de variables explicativas, se les tendrá que asignar el mismo o mayor peso dependiendo del caso a todos los modelos que involucren ese conjunto de variables, ocasionando así, que la suma de todos los modelos pudiera ser mayor que uno, lo cual es una contradicción a los supuestos de la teoría de probabilidades. Otro resultado desfavorable, se tiene en la asignación de las probabilidades de los modelos pues modelos formados por ciertos subconjuntos de variables explicativas se les asignará un peso, que va hacer menor que el peso de cualquier otro modelo formado por más de esas variables explicativas y esta probabilidad va hacer menor que la probabilidad asignada al modelo completo. Esto ocasionaría que el modelo completo tendrá mayor oportunidad de ser elegido, que otros modelos. Por lo cual este criterio siempre eligirá al modelo completo y de aquí, que sea un mal procedimiento de selección.

Otra dificultad que se tiene en este procedimiento de probabilidades posteriores, es la restricción de aplicarlo solo a modelos con el mismo número de parámetros, por lo que en caso de utilizarse en los modelos de regresión se tiene que hacer una clasificación en niveles, de tal forma que cada uno de ellos se agrupe a los modelos con igual número de parámetros. Entonces se presenta otro problema al tratar de asignar la probabilidad de los modelos, pues en cada uno de ellos se va asignar la probabilidad condicionada al nivel y a la variables explicativas que se tengan, ésto hace más difícil la aplicación de este modelo.

Se concluye lo siguiente, este procedimiento no es recomendable

Ejemplificando lo anterior, si el número de variables explicativas es 10 se tendrán que calcular $2^{10}-1 = 1023$ modelos diferentes; y si por cada modelo, la máquina tarda un promedio de 1.5 minutos, se tendran que esperar 4534.5 minutos, que corresponden a 75.6 horas y en días serían, 3,15 días. Este resultado muestra que si el número de variables es mayor que 10 y considerando que el número de modelos diferentes crece de manera exponencial, entonces se necesitaran más de 3 días de "uso exclusivo" de la computadora y se concluye , que la solución dada por el procedimiento es casi imposible de obtener por requerir mantener tantos recursos ocupados en un solo problema.

A continuación se comentará cada uno de los diferentes procedimientos de selección estudiados.

PROBABILIDADES POSTERIORES;

Es importante notar que este procedimiento viola los axiomas de coherencia, por lo cual, la decisión de elegir un modelo determinado no es necesariamente buena.

El procedimiento de probabilidades posteriores inicialmente tiene problemas en lo referente a la asignación de la probabilidad a los modelos, pues la interpretación de la probabilidad no se indica explícitamente en los artículos referidos (Zellner y Atkinson), aunque en otro artículo posterior a estos (Geisser y Eddy,1979) se indica que se refieren a la probabilidad asignada de que el modelo sea el modelo verdadero. También el problema de asignar probabilidades se agudiza para el caso de trabajar con muchas variables, pues cada uno de los modelos deberá tener asignada una probabilidad y la suma de todas debe valer uno.

para selección de variables en regresión. Quizá resulte válido para modelos que no son de regresión o fuera de un contexto de la selección de variables y por individuos que no toman en cuenta los axiomas de coherencia.

PROCEDIMIENTO DE BROEMELING.

Este procedimiento está diseñado para inferir sobre los parámetros desconocidos. Es un método basado en el cálculo de la matriz de presición de la distribución posterior marginal de cada uno de los modelos. Además, es el único método encontrado cuyo objetivo explícitamente es inferir sobre los parámetros y es además, sencillo de aplicarse.

A continuación se presentan los procedimientos que tienen objetivos predictivos, estos son: los métodos de Broemeling, Geisser y Eddy, Goldeistein y Lindley. En los primeros tres métodos, no importa que los parámetros de regresión sean o no desconocidos. El método de Lindley, sólo se plantea en el caso de tener varianza conocida, sin embargo se puede aplicar considerando la estimación de la varianza. Así lo aplico Lindley en los ejemplos que presenta y así, lo trabajo posteriormente Brooks.

PROCEDIMIENTO DE BROEMELING,

Este método se aplica si se desea predecir un valor de la variable dependiente, cuando el vector de las variables explicativas toma un valor en particular. Esta situación por lo general no se presenta en los problemas de regresión y de aquí, que es un factor limitante para la aplicabilidad de este procedimiento.

El procedimiento de Broemeling, consiste en calcular la precisión de la distribución predictiva en la observación futura, para cada uno de los modelos que se tengan y elegir al modelo que maximize la predicción futura o elegir al modelo que minimice la varianza.

Para el caso de tener interés en predecir el valor de la variable dependiente en varios vectores de las variables explicativas, es necesario efectuar el mismo procedimiento todas las veces que sea necesario porque el método calcula la precisión de la distribución predictiva en base en cada una de las observaciones futuras. De donde puede concluirse que si se tiene ésta situación se recomienda no usar este procedimiento.

PROCEDIMIENTO DE GEISSER Y EDDY.

Este es un procedimiento basado en el criterio de validación cruzada, que consiste en eliminar la posible influencia de la i-ésima observación de la muestra, al calcular la distribución predictiva tantas veces como observaciones se tengan. Estas densidades a su vez se multiplican entre si y el cálculo se repite para cada uno de los 2^p-1 modelos de regresión que se van a comparar. De aquí, que sea un método que requiere de un mayor manejo de cómputo que los procedimientos presentados aquí. Para algunos problemas de selección de variables en regresión en que p es "relativamente pequeño", se tiene un proceso de cómputo relativamente sencillo. Más sin embargo, existen otros problemas en donde es imposible esperar la respuesta del procedimiento, ya que el tiempo requerido es muy grande.

El procedimiento de Geisser y Eddy, es un método que tiene ventajas y desventajas, las cuales son: la principal desventaja es ser un método caro porque requiere un fuerte manejo de cómputo aunque p sea

"relativamente pequeño" y las ventajas: 1) no se necesita un valor específico para la predicción y 2) la utilización del criterio de validación cruzada permite poner a prueba a cada distribución predictiva con base a toda la información y esto, trae como consecuencia elegir la mejor distribución predictiva.

Se concluye que es un método recomendable aunque su manejo de cómputo sea grande.

PROCEDIMIENTO DE LINDLEY.

Es el único método encontrado que toma en cuenta explícitamente los costos que se involucran en el proceso de selección de variables de regresión. Este es un criterio planteado con base en la Teoría de Decisiones.

En el procedimiento de Lindley hay planteamientos que no estan muy claros en su manejo, como son: la eliminación de las unidades de los costos para "sumarse" a la suma de cuadrados de los residuales y la supuesta "aditividad" entre los costos. De aquí, que la aplicación de este método no se pueda llevar facilmente a la práctica.

Del ejemplo presentado por Lindley, se puede ver que es posible usar este procedimiento cuando se suponga que el costo implicado en la obtención de las variables explicativas sea cero y también cuando se conozca la relación entre sus costos, por ejemplo: $c_1 = 2c_2$, $c_2 = c_3$, $c_4 = c_1$, sin necesidad de conocer el valor exacto.

Un resultado importante que se desprende del artículo de Lindley

es el que se refiere, a la relación que se guarda de los resultados de Lindley con la Cp de Mallows. Como la C_p es un criterio de selección de variables generalmente utilizada por los estadísticos clásicos, se presenta aquí una posible justificación bayesiana de su utilización. Además, ambos resultados coinciden cuando en el criterio de Lindley se manejan costos ceros.

PROCEDIMIENTO DE GOLDEISTEIN.

Este es un método que teóricamente podría usarse cuando se desconociera la normalidad de los errores de los modelos de selección.

La dificultad principal de este método radica en que no se menciona la manera de calcular el valor con el cual los modelos se van a comparar. Por lo que este procedimiento no es utilizable en la práctica.

Los procedimientos antes estudiados no necesariamente producen los mismos resultados. Para ilustrar este hecho se aplican los procedimientos utilizando los datos de Hald (Draper y Smith, 1982, pp. 629-630).

A continuación se presenta la tabla No.2 basada en los datos de Hald, que ejemplifican algunos de los procedimientos aquí vistos, como son: el procedimiento de PSR (Geisser y Eddy, 1979) ubicado en la columna 1, la Cp de Mallows (Draper y Smith, 1982, pp. 301) colocada en la columna 2, Discrepancias (Geisser y Eddy, 1979) columna 3, y el procedimiento de Lindley (Lindley, 1968) en la columna 4

TABLA No.2 *RESULTADOS DE APLICAR LOS PROCEDIMIENTOS DE GEISSER, EDDY Y LINDLEY (sin costos y con el equivalente a la Cp de Mallows). Datos de Hald.*

VARIABLES EN LA REGRESION	P S R ($\log L_q$)	$\lvert C_p - p \rvert$	D_k	R_I (sin costos)
Ninguna		442.2		2716
1	-65.86	200.5	130.74	1266
2	-64.50	140.5	92.47	906
3	-68.75	313.2	201.26	1939
4	-63.77	136.7	91.86	884
12	-46.23	0.3	7.22	58
13	-66.28	195.1	170.62	1227
14	-48.13	2.5	9.32	15
23	-59.90	59.4	53.98	415
24	-64.67	135.2	112.45	869
34	-54.00	119.4	22.62	176
123	-45.57	1	6.92	48+
124	-45.29	1	6.57	48-
134	-45.71	0.5	7.27	51
234	-48.33	33	11.30	74
1234	-45.80	0	8.49	48

Si se realiza una selección global de todos los modelos involucrados se observa la siguiente (tabla No, 2): en la columna 1, se elige al modelo con las variables X_1, X_2, X_4, pues éste tiene un valor máximo de PSR; si se observa la columna 2, se elige al modelo con las variables X_1, X_2, pues se hace mínima la Cp de Mallows; utilizando la columna 3, se elige al modelo con las variables X_1, X_2, X_4, pues se minimiza la discrepancia y por último en la columna 4, con costo cero, en el procedimiento de Lindley se elige al modelo con las variables X_1, X_2, X_4.

Ahora, viendo el criterio propuesto por Draper y Smith que consiste en calcular todas las regresiones y analizar los modelos agrupados en conjunto de modelos con el mismo número de parámetros. La agrupación resultante se puede analizar de diferentes maneras y en cada una de ellas se obtienen diferentes resultados, a pesar de usar un mismo procedimiento de selección. Esto ocasiona una aparente inconsistencia, que es consecuencia de un manejo inapropiado del criterio de elección y del escaso conocimiento de la estructura del problema que se tiene.

Para ejemplificar lo anterior, si el objetivo es ir construyendo un modelo que vaya introduciendo mayor número de parámetros, por medio del criterio de Lindley, se tiene lo siguiente en la agrupación de los modelos: en el nivel uno (agrupación de modelos con un parámetro) se elige al modelo con la variable explicativa X_4 por tener el valor mínimo en ese nivel, para el siguiente nivel (con dos parámetros) se elige al modelo que tiene a la variable X_4, estos son: los modelos que involucran a las variables X_1X_4, X_2X_4 y X_3X_4. De estos, el valor mínimo lo obtiene el modelo X_1X_4. Siguiendo con este mismo mecanismo en el nivel 3 (con tres parámetros) se elige al modelo $X_1X_2X_4$ por tener un valor mínimo, usando el criterio de Lindley y se finaliza quedándose con él.

Usando este mismo ejemplo, pero de manera diferente como es el construir un modelo que inicie con todas las variables y vaya eliminando a las menos importantes con base al criterio de decisión de Lindley, se tiene lo siguiente: En el nivel 4(cuatro parámetros) se comienza con el modelo completo, se pasa al siguiente nivel inferior y se encuentra con otros posibles modelos a elegir, en donde el modelo elegido en este nivel es el X_1, X_2, X_4, pues minimiza el criterio en este nivel. Si de nuevo se pasa al siguiente nivel inferior, se elige al modelo X_1X_2 . Ahora, al tratar de pasar al nivel 1, el modelo que minimiza el criterio es más de dos veces mayor, que el valor obtenido por el modelo X_1X_2 , por lo cual se elige a este modelo.

Continuando con el mismo ejemplo, si ahora se considera algún tipo de información extramatemática, como es la dada por las correlaciones, que permite ir decidiendo sobre la entrada y salida de las variables explicativas, se tiene que en el caso de estos datos: X_1X_3 y X_4X_2 parecen ser altamente correlacionados, de donde la información de una variable puede ser recogida por la otra, entonces esto permite eliminar: al modelo completo, al modelo $X_1X_2X_4$, a todos los modelos del nivel tres y a los modelos X_1X_3 y X_2X_4 . Por lo cual, se puede concluir en elegir al modelo X_1X_2 . O quiza, elegir a un modelo de una sola variable, si es necesario ó no interesa que el criterio mínimo en el nivel 2 sea aumentado casi dos veces en el nivel 1 al elegir a la variable X_4 .

Estos resultados muestran la diversidad de conclusiones al manejar diferente el criterio propuesto de Draper y Smith. Cabe aclarar que Broemeling, Geisser y Eddy, y Lindley, sugieren no realizar el criterio anterior solamente, si no dan una elección global que permita conocer la estructura global del problema e identificar las causas que ocasionan la aparente inconsistencia.

ANEXO

1. LA DISTRIBUCION UNIVARIADA STUDENT.

Una prueba aleatoria se distribuye como una Student, si se puede expresar de la siguiente manera

$$p(x/\theta, h, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{v}{2}\right)} \left(\frac{h}{v}\right)^{1/2} \left[1+\frac{h}{v}(x-\theta)^2\right]^{-\left(\frac{v+1}{2}\right)} \qquad (A.1)$$

$-\infty < x < \infty$
$-\infty < \theta < \infty$
$0 < h < \infty$
$v > 0$

2. LA DISTRIBUCION GAMMA (α, β)

Una variable aleatoria (v.a.) se distribuye como una Gamma (α, β) si se expresa como:

$$p(X/\alpha, \beta) = \frac{X^{\beta-1} e^{-\frac{x}{\alpha}}}{\Gamma(\beta)\alpha^{\beta}} \qquad (A.2)$$

$0 < x < \infty$
$\alpha, \beta > 0$
α: parámetro de escala

3. LA DISTRIBUCION χ^2 CUADRADA.

Esta distribución es un caso particular de una Gamma (α, β), pues si se substituye en la expresión (A.2) $\alpha = 2$ y $\beta = \frac{v}{2}$, se obtiene la cuadrada.

$$p(X/v) = \frac{X^{v/2-1} e^{-x/2}}{2^{v/2}\Gamma\left(\frac{v}{2}\right)} \qquad (A.3)$$

$0 < x < \infty$
$v > 0$, indica el número de grados de libertad.

4. GAMMA INVERTIDA.

Esta distribución es obtenida de (A.2) haciendo un cambio de variable adecuado, dado por $y^2 = 1/X$, de donde

$$p(y/\alpha, \beta) = \frac{2}{\Gamma(\alpha)\alpha^{\beta} y^{2\beta+1}} e^{-\frac{1}{\alpha y^2}} \qquad (A.4)$$

$0 < y < \infty$
$\alpha, \beta > 0$

Otra expresión, de la Gamma invertida es tomar $\sigma = y$, $\beta = \frac{\nu}{2}$ *y* $\alpha = \frac{\nu}{2\sigma^2}$, *con lo cual*

$$P(\sigma/\nu, S) = \frac{2}{\Gamma\left(\frac{\nu}{2}\right)\left(\frac{\nu}{2\sigma^2}\right)^{\nu/2} y^{\nu+1}} e^{-\frac{\nu S^2}{2y^2}} \qquad \begin{matrix} 0<\sigma<\infty \\ \nu, S>0 \end{matrix} \qquad (A.5)$$

5. *ZELLNER DEFINE UNA FUNCION DE PERDIDA SIMETRICA COMO AQUELLA QUE SATISFAGA LO SIGUIENTE*

$$L(H_0, H_1) = L(H_1, H_0) \quad y \quad L(H_0, H_0) = L(H_1, H_1) = 0 \qquad (A.5)$$

Donde H_0 y H_1 son hipótesis exclusivas y exhaustivas.

Para mayor información ver Zellner (1971, pp.295-297)

6. *LAS REGLAS DE JEFFREYS (ZELLNER, 1971, PP.42-44) SIRVEN PARA SELECCIONAR UNA DISTRIBUCION INICIAL QUE REPRESENTA POCO CONOCIMIENTO O IGNORANCIA. ESTO CONSISTE EN*

a) *Si se cuenta con el parámetro* μ *desconocido, la ignorancia se representa como*

$$P(\mu) \propto \text{constante} \qquad -\infty<\mu<\infty$$

b) *Si* σ *es desconocida, una representación acerca de la ignorancia es:*

$$P(\sigma) \propto \frac{1}{\sigma}$$

BIBLIOGRAFIA

ATKINSON,A.C.(1978), "*Posterior probabilisties for choosing a regression model*". *Biometrika 65, 39-48.*

BERNARDO,J.M. (1979),"*Reference posterior distributions for Bayesian Inference*", *Journal of the Royal Statistical Society,B,41, pp.113-147.*

BERNARDO,J.M.(1980), "*Conceptos,Métodos y Fuentes de la Bioestadística*", *Comunicación Interna No.1.*

BROEMELING,L.D.(1985),"*Bayesian analysis of linear models*", *Marcel Dekker, New York, pp.94-104.*

BROOKS,R.J.(1972),"*A decision theoretic approach to optimal regression designs*", *Biometrika, 59,563-71.*

BROOKS,R.J.(1974),"*On the choise of an experiment for prediction in linear regression*", *Biometrika,61,303-311.*

BOX,G.E.P. & TIAO,G.C.(1973),"*Bayesian inference in Statistical Analysis*", *Addison Wesley.*

CHATFIELD,C & COLLINS,A.J.(1980),*Introduction to Multivariate Analysis*", *Science Paperbacks.*

CHATTERJEE,S & BERTRAM,P.(1977),*Regression analysis by example*", *Wiley.*

DAVIS,W.W.(1979),"*Approximate bayesian predictive distributions and model selection*". *Journal of the American Statistical Association, V.74, pp.312-317.*

DEGROOT,M.h.(1970),"Optimal Statistical Decisions", McGraw-Hill. New York.

DRAPER,N.R.& SMITH,H.(1981),"Applied Regression Analysis". Second Edition,Wiley. New York.

GEISSER,S.(1975),"The predictive sample reuse method with applications", Journal of the American Statistical Association,V.70,pp.320-328.

GEISSER,S.& EDDY,W.F.(1979),"A predictive approach to model selection". Journal of the American Statistical Association, V.74,pp.153-168.

GOLDSTEIN,M.(1976),"Bayesian Analysis of Regression Problems", Biometrika,63,pp.51-58.

HAITOVSKY,Y.(1972),"Regression Estimation From Grouped Observations", No.33, Griffin's Statistical Monograhs & Caurses.Alan Stuart. D;Sc.London.

JEFFREYS,H.(1961),"Theory of Probability (3ra.ed)". Oxford Clarendon.

LINDLEY,D.V.(1968)."The Choice of Variables in Multiple Regression", Journal of the Royal Statistical Society,B.V.30,pp.31-66.

LINDLEY,D.V.(1977),"Principios de la Teoria de la Decisión".Ed. Vicen-Vives. Barcelona.

MONTGOMERY,D.C.& PECK,E.A.(1982),'Introduction to Linear Regression Analysis", John Wiley.

PERICHI,L.R.(1984),"An alternative to the Standar Bayesian Procedure For Discrimination Between Normal Linear Models". Biometrika, V.71,pp.575-586.

STONE,M.(1974),"Cross-Validatory Choise & Assessment of Statistical Predictions",Journal of The Royal Statistical Society,B,36,pp.111-147.

STONE,M.(1976),*An Asympotic Equivalence of Choice of Model by cross-validation and Akaike's criterion, Journal of the Royal Statistical Society,Soc.B,39,pp.44-47.*

WEISBERG,S.(1980),"*Applied Linear Regression*".*John Wiley & Sons, Inc.*

Zellner,A(1971),"*An Introduction to Bayesin Inference in Econometrics*", *Wiley,New York.*