UNIVERSIDAD NACIONAL
AUTÓNOMA DE
MÉXICO

LA TRANSFORMACION DE BOX Y COX

Y SUS APLICACIONES

TESIS PROFESIONAL

VERONICA E. ROHEN ORTEGA

1977

PROLOGO

Al analizar datos reales, de naturaleza física, biológica, antropológica, es común considerar la posibilidad de aplicar una transformación a una o mas de las variables bajo estudio; sobre todo en situaciones donde la variable dependiente no es lineal en relación a la variable independiente y, existe alguna razón para pensar que dicha relación es un polinomio de un grado particular. Como se sabe, los polinomios son usados con frecuencia y, en lo que cabe, fácilmente manejables; sin embargo, polinomios de alto grado llevan a serias complicaciones numéricas y es difícil su interpretación.

Por tanto, se desea que un modelo complejo pueda ser simplificado por medio de una transformación que pertenezca a una familia particular.

En este trabajo, se presenta una familia particular de transformaciones propuesta por Box & Cox.

El primer capítulo presenta la teoría de dicha familia; análisis del parámetro de transformación: estimación, intervalos de confianza y pruebas. El capítulo II, habla de las ventajas y desventajas de la estadística presentada en el primer capítulo, y la comparación con otras pruebas. El último capítulo analiza detalladamente un ejemplo que servirá para ilustrar la teoría presentada en los dos primeros capítulos; se darán comparaciones (numéricamente) de los diversos criterios expuestos y las conclusiones que se derivan de tal ilustración.

Agradezco al Dr. Enrique de Alba la dirección de este trabajo

INDICE

CAPITULO I

"Un Análisis de Transformaciones"

1. INTRODUCCION

Con frecuencia nos encontramos con trabajos experimentales que están relacionados con funciones de respuesta

$$Y = E(Y) + \varepsilon$$

$$E(Y) = X\theta$$

Y es un vector columna de n observaciones; X una matriz n x k de elementos fijos; θ un vector columna de k coeficientes de regresión y ε un vector columna de n errores. Nuestro objetivo será por lo general checar si el modelo que estamos suponiendo es adecuado; estimar los valores de los parámetros (y por tanto las funciones de respuesta) y finalmente obtener medidas de precisión respecto a los estimadores.

Para todo ésto, se suponen tres hipótesis que idealmente deben cumplir nuestros datos a analizar:

i) aditividad en el modelo, i,e; simplicidad en la estructura de E(Y)
ii) varianza constante del error
iii) independencia y normalidad en los errores (y por tanto en las observaciones).

En la práctica estas hipótesis se violan al menos en parte y se complica nuestro análisis. En tales casos se considera prudente utilizar alguna transformación en los datos con el propósito de aumentar el grado de aproximación a las tres pro-

piedades para el análisis estadístico.

Estas transformaciones pueden aplicarse a las variables dependientes, independientes o ambas. Aquí daremos más atención al caso en que la transformación es aplicada a las variables dependientes.

2. TRANSFORMACION DE LA VARIABLE DEPENDIENTE

Box & Cox propusieron una familia paramétrica de transformaciones de Y a $Y^{(\lambda)}$ donde λ es un parámetro que define una transformación particular.

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases} \qquad Y > 0 ; \qquad 1.1$$

una variación también analizada por ellos es

$$Y^{(\lambda)} = \begin{cases} \dfrac{(Y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 \\ \log (Y + \lambda_2) & \lambda = 0 \end{cases} \qquad Y > -\lambda_2 \qquad 1.2$$

La transformación 1.1 es equivalente a la propuesta por Tukey

$$Y_i^{(\lambda)} = \begin{cases} Y^{\lambda} & \lambda \neq 0 \\ \log Y & \lambda = 0 \end{cases}$$

y con la ventaja de que es contínua en $\lambda = 0$ (esto se verifica fácilmente obteniendo el límite de $(Y^{\lambda} - 1)/\lambda$ cuando $\lambda \to 0$). Box & Cox suponen que $Y^{(\lambda)}$ es una función monótona de Y para toda λ, sobre un rango adecuado.

Trabajando bajo el supuesto de que para algún λ desconocido las observaciones transformadas $Y_i^{(\lambda)}$ (i=1,...,n) satisfacen las hipótesis antes mencionadas (i,e; son independientes y normalmente distribuidas con varianza constante σ^2 y esperanza $E(Y^{(\lambda)}) = a\theta$; donde a es una matriz de elementos conocidos y θ un vector de parámetros desconocidos asociados con las observaciones transformadas), se hace el análisis bajo dos criterios: el enfoque

clásico, o sea, la teoría de máxima verosimilitud en muestras grandes, y el Bayesiano. El primero lleva a estimar los parámetros, aproxima pruebas y obtiene intervalos de confianza en base a una χ^2. El segundo propone una distribución prior localmente uniforme de los parámetros θ y $\log \sigma$, para obtener una distribución posterior de λ.

Consideremos primero el enfoque clásico. Para obtener una función de densidad de las observaciones originales, se multiplica la densidad normal multivariada por el jacobiano de la transformación, de tal manera que en relación a las observaciones originales la función de verosimilitud es

$$(2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left\{-\frac{(Y^{(\lambda)}-a\theta)'(Y^{(\lambda)}-a\theta)}{2\sigma^2}\right\} J(\lambda;Y) \qquad 1.3$$

donde

$$J(\lambda;Y) = \prod_i \left|\frac{dY_i^{(\lambda)}}{dY_i}\right| \qquad i = 1,\ldots,n$$

De aquí se prosigue a encontrar los estimadores; notándose que los estimadores máximo verosímiles de las θ's son los obtenidos por el método de mínimos cuadrados de la variable dependiente $Y^{(\lambda)}$ y que el estimador de σ^2 para una λ fija está dado por

$$\hat{\sigma}^2(\lambda) = \frac{Y^{(\lambda)\prime} a_r Y^{(\lambda)}}{n} = \frac{S(\lambda)}{n}$$

donde si a es de rango completo

$$a_r = I - a(a'a)^{-1}a'$$

y $S(\lambda)$ es la suma de cuadrados de los residuales en el análisis de varianza de $Y^{(\lambda)}$.

Se sigue de ésto que el logaritmo de la función de verosimilitud es, a excepción de alguna constante,

$$\text{Lmax}(\lambda) = -\tfrac{1}{2} n \log \hat{\sigma}^2(\lambda) + \log J(\lambda;Y), \qquad 1.4$$

A través de la gráfica de Lmax (λ) contra λ se puede encontrar el valor $\hat{\lambda}$ que maximiza el logaritmo de la función de verosimilitud y obtener una región de confianza para λ del $100(1-\alpha)$ % a partir de la estadística.

$$T_L = \text{Lmax}\ (\hat{\lambda}) - \text{Lmax}\ (\lambda) < \frac{1}{2}\chi^2_{\nu_\lambda}(\alpha) \qquad 1.5$$

donde ν_λ es el número de componentes en λ . Este resultado podemos explicarlo de la siguiente manera: Una hipótesis nula sobre λ (i,e; $H_o: \lambda = \lambda_o$) se puede probar formando el logaritmo de la razón de verosimilitud valuada en los estimadores máximo verosímiles de θ y σ^2 condicionada sobre $\lambda = \lambda_o$. Para la hipótesis nula, este logaritmo es asintóticamente distribuido como $1/2\ \chi^2$ con ν_λ grados de libertad.

Si trabajamos con la variable estandarizada $Z^{(\lambda)} = Y^{(\lambda)} J^{-\frac{1}{n}}$ (con $J = J(\lambda;Y)$), obtenemos expresiones semejantes. La ecuación - 1.4 se reduce a

$$\begin{aligned} \text{Lmax}\ (\lambda) &= -\frac{1}{2} n \log \hat{\sigma}^2(\lambda;Z) + \log J(\lambda;Z) \\ &= -\frac{1}{2} n \log \hat{\sigma}^2(\lambda;Z) \end{aligned} \qquad 1.6$$

y

$$\hat{\sigma}^2(\lambda;Z) = \frac{Z^{(\lambda)\prime} a Z^{(\lambda)}}{n} = \frac{S(\lambda;Z)}{n}$$

con $S(\lambda;Z)$ la suma de cuadrados de los residuales de $Z^{(\lambda)}$. Concluyéndose que dado que la función de verosimilitud es $\left[\frac{S(\lambda;Z)}{n}\right]^{-\frac{n}{2}}$, el estimador máximo verosímil de λ es obtenido minimizando - $S(\lambda;Z)$ con respecto a λ .

Analicemos ahora el enfoque Bayesiano. La función de densidad de las observaciones originales $P(Y|\lambda,\theta,\sigma^2)$ está dada por 1.3.

Si escribimos

$$S(\lambda) = (Y^{(\lambda)} - a\hat{\theta}_\lambda)'(Y^{(\lambda)} - a\hat{\theta}_\lambda) = \nu_r s^2(\lambda)$$

donde $\hat{\theta}_\lambda = (a'a)^{-1} a'Y^{(\lambda)}$ es el estimador por mínimos cuadrados de θ para λ fija; ν_r = n - rango de a; $s^2(\lambda)$ el estimador usual de σ^2 ; entonces la distribución posterior conjunta es

$$P(\theta,\log\sigma,\lambda|Y) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{S(\lambda) + (\theta-\hat{\theta}_\lambda)'a'a(\theta-\hat{\theta}_\lambda)}{2\sigma^2}\right\} J(\lambda,Y) P(\theta,\log\sigma,\lambda) \qquad 1.7$$

donde $P(\theta, \log\sigma, \lambda)$ es la distribución prior conjunta tomada como el producto de la densidad marginal prior $P_o(\lambda)$ de λ y la distribución prior condicional de θ y $\log\sigma$ dado λ : $P(\theta, \log\sigma|\lambda)$. Suponiéndola localmente uniforme, i,e; efectivamente uniforme sobre el rango para el cual la verosimilitud es apreciable

$$P(\theta, \log\sigma, \lambda) = P_o(\lambda)P(\theta, \log\sigma|\lambda) \propto P_o(\lambda)g(\lambda).$$

La forma de la función $g(\lambda)$ se decide suponiendo que sobre el rango de los datos

$$Y_i^{(\lambda)} = \text{kte} + \ell_\lambda Y_i^{(\lambda)} \qquad i = 1,\ldots,n \qquad \lambda \text{ fija}, \qquad 1.8$$

donde ℓ_λ es un valor representativo del gradiente $\frac{dY^{(\lambda)}}{dY}$.

De la expresión 1.8 se deduce que

$$\log\sigma_\lambda = \log|\ell_\lambda| + \log\sigma_{\lambda_1}$$

pero como $\log\sigma_{\lambda_1}$ (independiente de λ) y θ son localmente uniformes e independientemente a priori, entonces

$$P(\theta, \log\sigma|\lambda) \propto |\ell_\lambda|^{-h}$$

donde ℓ_λ se escogió como

$$\left[\prod_i \left|\frac{dY_i^{(\lambda)}}{dY_i}\right|\right]^{\frac{1}{n}} = J(\lambda;Y)$$

entonces $P(\theta, \log\sigma|\lambda) \propto g(\lambda) \propto J(\lambda;Y)^{-(n-\nu_r)/n}$ y por tanto

$$P(\theta, \log\sigma, \lambda) \propto P_o(\lambda)J(\lambda;Y)^{-(n-\nu_r)/n} \qquad 1.9$$

Para encontrar la posterior de λ, se sustituye 1.9 en 1.7 y se integra con respecto a θ y $\log\sigma$, obteniendo

$$P(\lambda|Y) = K_Y \frac{J(\lambda;Y)^{\nu_r/n}}{\{s^2(\lambda)\}^{\nu_r/2}} P_o(\lambda)$$

donde K_Y es una constante independiente de λ, que normaliza.

De aquí se observa que la contribución que hacen las observaciones a la distribución posterior de λ es la que representa el cociente

$$\frac{J(\lambda;Y)^{\nu_e/n}}{\left[s^2(\lambda)\right]^{\nu_e/2}}$$

o equivalentemente

$$L_b(\lambda) = -\frac{1}{2}\nu_e \log s^2(\lambda) + (\nu_e/n)\log J(\lambda;Y) \qquad 1.10$$

volviendo a la transformación normalizada, 1.10 se expresa como

$$L_b(\lambda) = -\frac{1}{2}\nu_e \log s^2(\lambda;Z) \qquad 1.11$$

y la densidad posterior de λ es

$$P(\lambda|Y) = \text{kte}\ P_o(\lambda)\ S(\lambda;Z)^{-\frac{\nu_e}{2}}$$

y finalmente se selecciona el valor de λ graficando $S(\lambda;Z)^{-\frac{\nu_e}{2}}$ contra λ y combinándolo con información a priori acerca de λ

Comparemos ahora las expresiones 1.6 y 1.11; el logaritmo de la función de verosimilitud maximizada y el logaritmo de la contribución a la distribución posterior de λ respectivamente

$$L_{max}(\lambda) = -\frac{1}{2}\, n \log\left\{S(\lambda;Z)/n\right\}$$

$$L_b(\lambda) = -\frac{1}{2}\nu_e \log\left\{S(\lambda;Z)/\nu_e\right\}$$

Ambas expresiones son funciones monótonas de la suma de cuadrados $S(\lambda;Z)$ y alcanzan su máximo cuando ésta es minimizada. Claramente se ve que la diferencia básica es que el número de observaciones en la primera es sustituido por los grados de libertad en la segunda; sin embargo en la práctica, esta pequeña diferencia no simpre puede ser despreciada ya que

$$\frac{\nu_e}{n} = \frac{n - \text{rango de a}}{n} = 1 - \frac{\text{rango de a}}{n}$$

es bastante menor que 1 aún cuando n sea lo suficientemente grande, siendo ésta, una de las varias razones que llevan a preferir $L_t(\lambda)$. (Para mayores detalles ver: Bartlett: "Propiedades de suficiencia y pruebas estadísticas". Proy. Roy. Soc. A, 160).

Box & Cox analizan un par de ejemplos, donde al hacer la transformación se cumplen simultáneamente las hipótesis ideales; construyen intervalos de confianza, pruebas de hipótesis, densidades posteriores, gráficas y finalmente comparan el mejor estimador de λ . Uno de estos ejemplos será ilustrado en el capítulo III.

3. ANALISIS ADICIONAL DE LA TRANSFORMACION

Es necesario tener especial cuidado al aplicar una transformación, pues hipótesis que ya habían sido logradas pueden ser violadas al transformar los datos; en general, puede suceder que no se cumplan simultáneamente las hipótesis ideales.

Box & Cox analizan la contribución que por separado da cada hipótesis a la transformación, y es lo que se expondrá a continuación.

Este análisis tiene por objetivo indicarnos tres factores importantes:

i) qué tan simple se puede usar el modelo?

ii) qué ponderación puede darse a las consideraciones: aditividad, homoscedasticidad y normalidad para elegir el valor de λ ?

iii) si son necesarias diferentes transformaciones para lograr los objetivos y por tanto si el valor de λ escogido usando el proceso completo, es o no compatible.

En este análisis también se consideran dos puntos de vista, uno vía máxima verosimilitud y otro vía Bayes.

Se considerará un modelo general, al cual se podrá aplicar o eliminar una restricción C.

Sean Lmax (λ) el logaritmo de la verosimilitud maximizada para el modelo general y Lmax $(\lambda|C)$ el logaritmo de la verosimilitud maximizada para el modelo restringido, entonces

$$\text{Lmax}\ (\lambda|C) = \text{Lmax}\ (\lambda) + \left[\text{Lmax}\ (\lambda|C) - \text{Lmax}\ (\lambda)\right] \qquad 1.12$$

el término dentro de los corchetes es una estadística para probar la presencia de la restricción. La expresión 1.12 puede ser generalizada

$$\text{Lmax}\ (\lambda|C_1,C_2) = \text{Lmax}\ (\lambda) + \{\text{Lmax}\ (\lambda|C_1) - \text{Lmax}\ (\lambda)\} + \{\text{Lmax}\ (\lambda|C_1,C_2) - \text{Lmax}\ (\lambda|C_1)\} \qquad 1.13$$

Equivalentemente, desde el punto de vista Bayesiano, la densidad posterior de λ será

$$P(\lambda|C) = P(\lambda)\,\frac{P(C|\lambda)}{P(C)} \qquad 1.14$$

$P(C) = E_\lambda\left[P(C|\lambda)\right]$ es una constante independiente de λ ; generalizando

$$P(\lambda|C_1,C_2) = P(\lambda)\,\frac{P(C|Y)}{P(C_1)}\,\frac{P(C_2|\lambda,C_1)}{P(C_2,C_1)} \qquad 1.15$$

$P(C_2,C_1) = E_\lambda\left[P(C_2|\lambda,C_1)\right]$ es una constante independiente de λ .

$P(C_1|\lambda)$ y $P(C_2|\lambda,C_1)$ se pueden interpretar como las contribuciones de las restricciones a la función $P(\lambda|C_1,C_2)$.

Etiquetando nuestras restricciones tenemos

A - modelo lineal más simple; i,e; no interacción, no hay términos de orden mayor a 1,

H - homogeneidad de varianza $(\sigma_1^2 = \ldots = \sigma_k^2)$,

N - normalidad.

Veamos primero cuál es la contribución que da la restricción H, suponiendo que se cumple normalidad. Para ello suponen Box & Cox que en k grupos de datos, la varianza y la esperanza son

constantes dentro de cada grupo. Si se aplica una transformación a las variables de tal forma que se dé la normalidad en todos los grupos de manera simultánea, entonces, trabajando con la variable estandarizada

$$\text{Lmax } (\lambda|N) = -\frac{1}{2}\sum n_i \log \left\{S^{(i)}(\lambda;Z)/n_i\right\}$$

donde $\sum n_i = n$; $S^{(i)}(\lambda;Z)$ es la suma de cuadrados de las desviaciones, que tiene $\nu_i = n_i - 1$ grados de libertad ($\sum \nu_i = n - k$).

Considérese la restricción H y supóngase que la transformación también induce homoscedasticidad, entonces si $\sum S^{(i)} = S_v$ es la suma combinada de cuadrados entre grupos

$$\text{Lmax } (\lambda|H,N) = -\frac{1}{2} n \log \left\{S_v(\lambda;Z)/n\right\}$$

y por tanto

$$\text{Lmax } (\lambda|H,N) - \text{Lmax } (\lambda|N) = \log\left(\frac{\prod\left|S^{(i)}(\lambda;Z)/n_i\right|^{\frac{1}{2}n_i}}{S_v(\lambda;Z)/n\ ^{\frac{1}{2}n}}\right) \qquad 1.16$$

éste es el logaritmo del criterio de Neyman-Pearson para probar homoscedasticidad ($\sigma_1^2 = \ldots = \sigma_k^2$), ($\sigma_i^2$ = varianza dentro del i-ésimo grupo).

Por tanto

$$\text{Lmax } (\lambda|H,N) = \text{Lmax } (\lambda|N) + \log\left[\frac{\prod\left|S^{(i)}(\lambda;Z)/n_i\right|^{\frac{n_i}{2}}}{\left\{S_v(\lambda;Z)/n\right\}^{\frac{n}{2}}}\right]$$

Pasemos ahora al análisis Bayesiano

$$P(\lambda|H,N) = \text{kte } P(\lambda|N)\ P(\sigma_1^2 = \ldots = \sigma_k^2|\lambda,N)$$

y

$$P(\lambda|N) = C\ P_o(\lambda) \prod\left[S^{(i)}(\lambda;Z)\right]^{-\nu_i/2} \qquad 1.17$$

donde $(\text{kte})^{-1} = \underset{\lambda|N}{E}\left[P(\sigma_1^2 = \ldots = \sigma_k^2|\lambda,N)\right]$; C una constante de normalización y $P(\lambda|N)$ la distribución prior para el modelo general (no

restringido). La distribución prior para el modelo restrigido será

$$P(\lambda|H,N) = P_o(\lambda)\ C_\nu \left\{S_\nu(\lambda;Z)\right\}^{-\nu/2} \qquad 1.18$$

C_ν una constante de normalización.

Dividiendo 1.18 por 1.17 se puede saber cual es la expresión para

$$P(\sigma_1^2 = \ldots = \sigma_k^2 | \lambda, N) = \frac{C_\nu \Pi \left\{S^{(i)}(\lambda;Z)\right\}^{\nu_i/2}}{C\left\{S_\nu(\ ;Z)\right\}^{\nu/2}} \qquad 1.19$$

haciendo un poco de álgebra, se ve que 1.19 puede expresarse como

$$\frac{C_\nu \Pi\, \nu_i^{\frac{1}{2}\nu_i}}{C\ \nu^{\frac{1}{2}\nu}} \exp\left\{-\frac{1}{2}\left[\nu \log\left(\frac{S_\nu(\lambda;Z)}{\nu}\right) - \sum_i \nu_i \log\left(\frac{S^{(i)}(\lambda;Z)}{\nu_i}\right)\right]\right\}.$$

La expresión

$$\nu \log\left(\frac{S_\nu(\lambda;Z)}{\nu}\right) - \sum \nu_i \log\left(\frac{S^{(i)}(\lambda;Z)}{\nu_i}\right)$$

es la estadística para probar homogeneidad de varianza, esto es, el criterio de Bartlett y le etiquetaremos como $M(\lambda;Z)$

Vayamos ahora a la cuestión más importante ¿Qué tan simple podemos usar la forma para $E(Y^{(\lambda)})$?.

Usando 1.13

$$\text{Lmax}\,(\lambda|A,H,N) = \text{Lmax}(\lambda|H,N) + \left\{\text{Lmax}\,(\lambda|A,H,N) - \text{Lmax}\,(\lambda|H,N)\right\}$$

Se puede particionar el parámetro $\theta = (\theta_1, \theta_2)$ tal que $\theta_2 = 0$ sea la restricción A.

En términos de la variable estandarizada

$$\text{Lmax}\,(\lambda|\theta_2 = 0,H,N) = -\frac{1}{2}\, n \log\left[S_{\nu_1+\nu_2}(\lambda;Z)/n\right]$$

y

$$\text{Lmax } (\lambda|H,N) = -\frac{1}{2}\, n \log\left[S_{\nu_r}(\lambda;Z)/n\right]$$

donde S_{ν_r} es la suma de cuadrados de los residuales en el modelo de segundo grado con ν_r grados de libertad. $S_{\nu_r+\nu_2}$ es la suma de cuadrados de los residuales de un modelo de primer orden, - aquí ν_2 representa los grados de libertad asociados con θ_2 y ν_1 los asociados con θ_1.

En general

$$S_{\nu_r+\nu_2}(\lambda;Z) = S_{\nu_r}(\lambda;Z) + S_{\nu_2\cdot\nu_1}(\lambda;Z) \qquad 1.20$$

donde $S_{\nu_2\cdot\nu_1}(\lambda;Z)$ denota la suma de cuadrados extra de $Z^{(\lambda)}$ para θ_2 corregida por θ_1 y tiene ν_2 grados de libertad.

Entonces, si

$$F(\lambda;Z) = \frac{S_{\nu_2\cdot\nu_1}(\lambda;Z)/\nu_2}{S_{\nu_r}(\lambda;Z)/\nu_r}$$

es la estadística para probar la restricción para el modelo más simple, y si se usa 1.20

$$\text{Lmax } (\lambda|\theta_2=0,H,N) = \text{Lmax } (\lambda|H,N) + \\ - \frac{1}{2}\, n \log\left[1 + \frac{\nu_2}{\nu_r} F(\lambda;Z)\right]. \qquad 1.21$$

Puede apreciarse que esta expresión proporciona el análisis del criterio completo en dos partes; una tomando el hecho de que el modelo cumple solo homoscedasticidad y normalidad y la otra considerando la restricción adicional de aditividad dadas homoscedasticidad y normalidad.

Consideremos ahora el enfoque Bayesiano. De 1.15

$$P(\lambda|\theta_2=0,H,N) = \text{kte } P(\lambda|H,N)\, P(\theta_2=0|\lambda,H,N) \qquad 1.22$$

donde, como ya habíamos visto, $(\text{kte})^{-1} = \underset{\lambda|H,N}{E}\left[P(\theta_2=0|\lambda,H,N)\right]$.

Dado que la restricción A se dá, no existen componentes para θ_2 en la distribución prior, así que 1.22 viene siendo la distribución posterior de bajo la suposición de A. Por tanto

$$P(\lambda|\theta_2=0,H,N) = P_0(\lambda)\ C_{\nu_1+\nu_2}\left[S_{\nu_1+\nu_2}(\lambda;Z)\right]^{-\frac{1}{2}(\nu_1+\nu_2)}$$

y

$$P(\lambda|H,N) = P_0(\lambda)\ C_{\nu_1}\left[S_{\nu_1}(\lambda;Z)\right]^{-\frac{1}{2}\nu_1}$$

donde $C_{\nu_1+\nu_2}$ y C_{ν_1} son constantes que normalizan.

Usando las expresiones anteriores, se tiene que $P(\theta_2=0|\lambda,H,N)$ puede ser expresada como

$$\text{kte}\ \frac{C_{\nu_1}}{C_{\nu_1+\nu_2}}\ \frac{\left[S_{\nu_1}(\lambda;Z)\right]^{\frac{1}{2}\nu_1}}{\left[S_{\nu_1+\nu_2}(\lambda;Z)\right]^{\frac{1}{2}(\nu_1+\nu_2)}} \qquad 1.23$$

Por tanto

$$P(\lambda|\theta_2=0,H,N) \propto P(\lambda|H,N)\ \frac{C_{\nu_1}}{C_{\nu_1+\nu_2}}\ \frac{\left[S_{\nu_1}(\lambda;Z)\right]^{\frac{1}{2}\nu_1}}{\left[S_{\nu_1+\nu_2}(\lambda;Z)\right]^{\frac{1}{2}(\nu_1+\nu_2)}}$$

De aquí se deduce que esta expresión proporciona el análisis de la densidad conjunta en dos partes; una tomando en cuenta solo la homoscedasticidad y normalidad y la otra expresada por 1.23, en la cual se mide la influencia de la restricción aditividad.

Podemos finalmente comparar los dos enfoques escribiendo 1.23 como

$$\text{cte}\ \left\{S_{\nu_1}(\lambda;Z)\right\}^{-\frac{\nu_1}{2}}\left\{1+\frac{\nu_2}{\nu_1}F(\lambda;Z)\right\}^{-\frac{1}{2}(\nu_1+\nu_2)} \qquad 1.24$$

expresión similar a la correspondiente en verosimilitud

$$-\frac{1}{2}\,n\,\log\left\{1+\frac{\nu_2}{\nu_1}F(\lambda;Z)\right\} \qquad 1.25$$

Vemos que una diferencia básica es el término $S_{\nu_1}(\lambda;Z)$ en 1.24

Desde el punto de vista para muestras grandes, ν_2/ν_1 es muy pequeño y la variación debida al término adicional es despreciable en el límite (en la práctica no siempre sucede ésto).

Graficando la razón estandar F como una función de λ , podemos representar el efecto de 1.24 y 1.25.

Las expresiones 1.24 y 1.25, no siempre llevan a conclusiones compatibles, i,e; puede sucedero que dada una λ , ésta de un valor grande para $S_{\nu_1}(\lambda;Z)$ pero pequeño para $F(\lambda;Z)$.

Desde el punto de vista Bayesiano la diferencia estriba en que la razón F determinada por 1.25 es una función monótona de la probabilidad fuera del contorno de la distribución posterior que pasa por $\theta_2=0$.

Esta teoría es ilustrada con ejemplos y se observa si aporta mucho o poco el criterio completo a la transformación, y se compara con la aportación que hace el criterio sin la restricción aditividad.

4. ANALISIS DE EFECTOS DESPUES DE LA TRANSFORMACION

Podemos analizar el efecto que produce la elección de una λ adecuada, i,e; considerando conocida $\hat{\lambda}$ se puede justificar el comportamiento de un análisis estandar con la variable estandar $Z^{(\hat{\lambda})}$como si la transformación hubiera sido dada de antemano; y se demostrará que la única modificación será la reducción de los grados de libertad de los residuales.

Considérese de nuevo la densidad prior localmente uniforme de θ y λ , entonces la densidad posterior de θ se obtendrá integrando la conjunta respecto a λ

$$P(\theta|Y) = \int P(\theta,\lambda|Y)d\lambda = \int P(\theta|\lambda,Y)\ P(\lambda|Y)d\lambda$$

i,e;

$$P(\theta|Y) = \frac{\int \left[(Z^{(\lambda)}-a\theta)'(Z^{(\lambda)}-a\theta)\right]^{-\frac{n}{2}} d\lambda}{\int \left[\nu_r s^2(\lambda;Z)\right]^{-\frac{\nu_r}{2}} d\lambda}$$

Resolviendo las integrales expandiendo alrededor del máximo de los integrandos, se obtiene

$$\frac{\left[(Z^{(\hat{\lambda})}-a\theta)'(Z^{(\hat{\lambda})}-a\theta)\right]^{-\frac{1}{2}(n-\nu_\lambda)}}{\left[\nu_r s^2(\hat{\lambda};Z)\right]^{-\frac{1}{2}(\nu_r-\nu_\lambda)}} \qquad 1.26$$

donde el máximo del integrando en el numerador está cercano a $\hat{\lambda}$ tanto como θ lo esté a su valor máximo verosímil, y el denominador lo alcanza en el estimador máximoverosímil de $\hat{\lambda}$.

Siendo 1.26 la densidad posterior de θ para alguna λ fija conocida con grados de libertad reducidos en ν_λ (número de parámetros de transformación).

Expandiendo $(Z^{(\hat{\lambda})}-a\theta)'(Z^{(\hat{\lambda})}-a\theta)$ en series de Taylor alrededor de λ:

$$S(\lambda,Z) + Q(\theta,\lambda) = S(\hat{\lambda};Z) + (\lambda - \hat{\lambda})'b(\lambda - \hat{\lambda}) + Q(\theta,\hat{\lambda}) + g(\theta,\lambda)$$

donde, b es la matriz de segundas derivadas de $S(\lambda;Z)$ con respecto a λ evaluadas en $\hat{\lambda}$. $Q(\theta,\hat{\lambda}) = J_\lambda^{-2/n}(\theta - \hat{\theta}_\lambda)'a'a(\theta - \hat{\theta}_\lambda)$ y $g(\theta,\lambda) \doteq \sum_i \frac{\partial Q}{\partial \lambda_i}\Big|_{\hat{\lambda}} (\lambda_i - \hat{\lambda}_i) + \frac{1}{2} \sum_i \sum_j \frac{\partial^2 Q}{\partial \lambda_i \partial \lambda_j}\Big|_{\hat{\lambda}} (\lambda_i - \hat{\lambda}_i)(\lambda_j - \hat{\lambda}_j)$; para una n moderada, la influencia de $g(\theta,\lambda)$ es tan pequeña que puede despreciarse.

Obteniendo por fin una distribución posterior de. λ

$$P(\lambda|Y) \cong \text{kte} \left\{1 + \frac{(\lambda - \hat{\lambda})'b(\lambda - \hat{\lambda})}{\nu_r s^2(\hat{\lambda};Z)}\right\}^{-\frac{1}{2}\nu_\lambda}$$

de aquí se deduce que

$$\frac{\lambda_j - \hat{\lambda}_j}{s(\hat{\lambda};Z)\sqrt{b^{jj}}}$$

tiene aproximadamente una distribución posterior t-multivariada, y

$$\frac{(\lambda - \hat{\lambda})' b (\lambda - \hat{\lambda})}{\nu_r s^2(\hat{\lambda}; Z)}$$

una distribución posterior F.

Hasta aquí hemos visto la teoría de Box & Cox acerca de la transformación de variables para obtener las tres hipótesis básicas de las que hablamos antes. Podría generalizarse este proceso y considerar en lugar de un solo parámetro, un vector de parámetros de transformación y hacer un análisis multivariado de lo anterior.

También pueden hacerse algunas extensiones a series de tiempo y modelos econométricos, como se verá en el próximo capítulo.

CAPITULO II

"ALGUNAS EXTENSIONES AL ANALISIS DE TRANSFORMACIONES"

1. INTRODUCCION

Muchas han sido las notas que se han escrito acerca del artículo de Box & Cox: "Un Análisis de Transformaciones". En el presente capítulo se expondrán algunas de ellas, que en su mayoría consistirán de pruebas de hipótesis alternativas referentes al parámetro de transformación; ventajas y desventajas de ellas comparadas con la expuesta por Box & Cox; se darán algoritmos para calcular dicho parámetro; se hablará también de consistencia, robustés, invarianza, etcétera.

2. UNA PRUEBA ALTERNATIVA

Andrews, propone una prueba para el valor del parámetro que es "exacta" en el sentido de que la distribución nula de la estadística es conocida, de esta prueba se obtienen regiones de confianza, y se estima la potencia de la prueba para predecir la simetría de las inferencias que se derivan de todo este análisis.

A pesar de que el modelo que se ajusta a Y es no lineal, una prueba F puede ser derivada para probar la hipótesis $\lambda = \lambda_0$ un valor dado. Esta deducción está basada en expansiones localmente lineales de aquellos términos del modelo que son no lineales (aquí los términos de la expansión, dependerán del valor -

de Y).

Supóngase que la transformación se puede aproximar por expansiones lineales alrededor del verdadero valor de λ (λ_o), entonces

$$Y^{(\lambda_o)} = a\theta + \mathbb{J}(\lambda_o - \lambda) + \varepsilon_i$$

donde la matriz jacobiana J de elementos $\left.\frac{\partial [Y_i^{(\lambda)}]}{\partial Y_j}\right|_{\lambda=\lambda_o}$ depende solo de Y y debe ser de alguna manera modificada para lograr una prueba simple, de tal manera que los términos de la matriz $\mathbb{J}$ quedarán

$$\left.\frac{\partial [Y_i^{(\lambda)}]}{\partial \lambda_j}\right|_{\lambda=\lambda_o,\ y=\hat{y}}$$

donde λ_j es uno de los q elementos del vector de parámetros λ.

Por tanto, del modelo

$$Y^{(\lambda_o)} = a\theta + \hat{\mathbb{J}}(\lambda_o - \lambda) + \varepsilon$$

puede ser construida la prueba para $\lambda = \lambda_o$.

Esta estadística de prueba se desarrolla a partir de la suma de cuadrados de los residuales de las $Y^{(\lambda)}$; si se diferencía esta suma de cuadrados respecto a λ, se obtendrá la suma de productos de los residuales de Y y de las derivadas $\partial Y_i^{(\lambda)}/\partial\lambda_j$.

Para formar la prueba "exacta", los valores observados de Y en $\mathbb{J}$ son reemplazados por los estimadores por mínimos cuadrados de

$$\hat{Y}^{(\lambda)} = a(a'a)^{-1}a'Y^{(\lambda)}$$

dando como resultado un vector de derivadas estimadas $\hat{\mathbb{J}}$.

De aquí que la estadística de prueba resultante es

$$T_A = -\frac{[Y^{(\lambda_o)}a_r\hat{\mathbb{J}}]^2}{s_y^2\hat{\mathbb{J}}'a_r\hat{\mathbb{J}}} \qquad 2.1$$

donde s_y^2 es un estimador de la varianza σ^2.

Se puede ver que el numerador de 2.1 es el resultado de la regresión de los residuales de un modelo lineal, sobre una cantidad $(a, \hat{J})$ que depende de Y solo a través de $\hat{\theta}$, pues los residuales y los parámetros $\hat{\theta}$ son independientes; por tanto T_A tendrá una distribución exacta F.

Andrews hace notar que la precisión en las aproximaciones puede quizá, afectar la eficiencia de la prueba, pero no su exactitud.

A partir de esta prueba Andrews construye intervalos de confianza para λ de la forma

$$C(Y) = \left(\lambda \middle| \alpha(\lambda) \geq \alpha_0\right)$$

donde α_0 es el tamaño exacto de la prueba, con un coeficiente de confianza de $1 - \alpha_0$.

Andrews compara su prueba "exacta" con la prueba para no aditividad (similar a la que propone Tukey) con p grados de libertad, que son seleccionados de tal manera que hagan sensible la prueba a pequeños cambios en la transformación; un grado de libertad para cada parámetro de transformación independiente.

Esta última prueba se basa en la estadística F, que se ajusta a una variable adicional: un vector cuyos elementos son de la forma $\hat{Y}_i \log \hat{Y}_i$.

Dado que F es invariante bajo la multiplicación de $\mathbb{J}$ por un escalar y bajo la suma de alguna combinación lineal de las columnas de a y $\mathbb{J}$; por medio de algún artificio es posible escribir $\hat{Y}_i \log \hat{Y}_i$ como $(1 + d_i)\log(1 + d_i)$. Si $d_i < 1$, entonces es de esperar que las dos pruebas lleven a resultados similares, mientras que si $d_i > 1$, llevarán a resultados bastante diferentes; de aquí que se concluya que mucha de la información en la prueba "exacta" se derive de la suposición de aditividad.

Una buena comparación entre pruebas, sería la de ver si son o no sensibles a datos erráticos (outliers), y Andrews la hace con su prueba "exacta" y la propuesta por Box & Cox (máxima verosimilitud).

La prueba la hace de una manera ilustrativa (se verá en el capítulo III) y concluye que el método de máxima verosimilitud es más sensible a las fallas de normalidad. Estas fallas son difíciles de identificar en los datos transformados (usando este método), pero su efecto será pequeño.

2.1 POTENCIA DE LA PRUEBA

Analicemos ahora la potencia de la prueba. Dado que el numerador de la estadística T_A es una χ^2 no central, la distribución no nula de la estadística T_A será una F no central; de aquí que la potencia de la prueba será una función del parámetro de no centralidad; así que este parámetro está dado por

$$\hat{\gamma}^2 = \hat{\sigma}^{-2}(\lambda - \lambda_o)'\hat{\theta}'a'a\hat{\theta}(\lambda - \lambda_o) \qquad 2.2$$

donde $\hat{\sigma}^2$ es un estimador insesgado de σ^2.

Entonces el tamaño del intervalo de confianza podrá ser encontrado buscando primero el parámetro de no centralidad γ^2, asociado a una potencia; después se reemplaza $\hat{\gamma}^2$ por γ en 2.2 y se resuelve para $(\lambda - \lambda_o)$ para obtener una elipsoide.

Por tanto la potencia de la prueba está condicionada a una variable independiente dada, $\hat{\theta}$, cuya distribución es conocida.

La desventaja principal del método propuesto por Andrews, es que no lleva a un claro resúmen gráfico de las conclusiones; y una gran ventaja, que se puede observar durante este desarrollo, es que el estimador de λ no se necesita; otra ventaja, ya mencionada, es que se conoce la distribución nula de la estadística de prueba.

3. OTRA PRUEBA DE NORMALIDAD

Atkinson declara que la prueba de Andrews fué construida con todas sus ventajas, pero ignorando el jacobiano de la transformación, y ello llevará a algunas consecuencias, que se darán al comparar esta prueba con la que propone Atkinson; algunas de estas comparaciones las hace sobre unos datos, que se verán en el capítulo III.

La prueba de Atkinson, como se verá, tiene también la propiedad de que λ no necesita ser estimada. Esta prueba, $C(\alpha)$, se desarrolla a partir de la función de verosimilitud, diferenciando respecto a λ y dividiendo por el error estandar estimado de la derivada de la matriz de información.

Esta resulta ser una prueba asintóticamente equivalente a la de razón de verosimilitud, pero localmente más potente; sin embargo los elementos de la matriz de información son muy difíciles de calcular, y por esta razón Atkinson se aproxima a $C(\alpha)$ de la siguiente manera:

Podemos derivar la suma de cuadrados de los residuales de la variable estandarizada $Z^{(\lambda)}$ respecto a λ y reemplazar la matriz de información correspondiente para esta variable; pues dado que λ es fija, la función de verosimilitud, como se vió en el primer capítulo, es proporcional a dicha suma de cuadrados, y se obtiene

$$T_D = - \frac{\left(Z^{(\lambda)'} a_{\gamma} w^{(z)} \right)^2}{s^2 w^{(z)'} a_{\gamma} w^{(z)}} \qquad 2.3$$

donde $w^{(z)} = \partial Z^{(\lambda)} / \partial \lambda$; y un estimador de $\sigma^2(\lambda;Z)$ es s_z^2.

A pesar de que el grado de dificultad para calcular esta prueba es análogo al de la prueba "exacta" de Andrews, que es menor que el requerido para calcular la prueba razón de verosimilitud, tiene la desventaja de que es poco exacto el conocimiento de la

distribución nula, pues tanto $Z^{(\lambda)}$ como $w^{(\lambda)}$ aparte de no ser independientes, tienen distribuciones no centrales. Para eludir el problema, Atkinson normaliza el numerador de la estadística T_b , sin embargo no llegó a una χ^2 central.

Finalmente, Atkinson hace una comparación entre las pruebas T_L, T_A y T_b con un ejemplo, y concluyó que las pruebas derivadas de la función de verosimilitud, resultan ser uniformemente más potentes que la de Andrews; y da dos razones de esto; una es refiriéndose al ejemplo numérico (ver capítulo III), y la otra, debido a que no toma en cuenta el jacobiano de la transformación, de tal manera que impide el cambio de escala de las observaciones sobre la transformación.

Una desventaja de la estadística T_b es que no se acerca la distribución nula a la χ^2 central; pero por otro lado, toda la información está contenida en la derivada de la función de verosimilitud, lo cual nos ahorra la estimación del parámetro λ , la cual, no aumenta nuestra información.

4. PRUEBA DE SIMETRIA

Supongamos que observaciones transformadas (usando 1.1) tienen esperanza dada por el bien conocido modelo

$$E(Y^{(\lambda)}) = a\theta \qquad 2.4$$

con errores distribuidos como una muestra aleatoria de alguna distribución no normal. Nos preguntamos ahora si no existe alguna λ para la cual el modelo (con sus hipótesis ideales) no se cumple exactamente; cuál es la precisión aparente con la cual λ es estimada; qué tan útil será el valor de λ que se obtiene del método.

Draper & Cox muestran que la familia de transformaciones 1.1 - puede ser útil aún en situaciones donde la transformación de po

tencia no lleva a normalidad exactamente; esto es, el proceso de Box & Cox es robusto para no normalidad, siempre que el término error sea suficientemente simétrico.

Pensemos en que alguna de las dos cosas que siguen pasa: el verdadero valor de λ es $\lambda_0 \neq 0$ o si no se satisface el modelo exactamente para alguna λ, el estimador máximo verosímil de λ converge a λ_0.

Para cualquiera de estos casos, la condición para que el estimador máximoverosímil de λ; cuando los errores siguen una distribución no normal, con varianza constante; sea consistente es que

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} = 0 \qquad 2.5$$

donde L es el logaritmo de la función de verosimilitud bajo el modelo; sea en este caso la función de verosimilitud de la forma

$$-n \log \sigma - (2\sigma^2)^{-1} \sum_i (Y_i^{(\lambda)} - a\theta_i)^2 + (\lambda - 1)\sum_i \log Y_i$$

nótese que no se trata de $L_{max}(\lambda)$.

Fijémos ahora las condiciones bajo las cuales se satisface la ecuación 2.5, sin perder de vista las hipótesis arriba mencionadas

$$\frac{\partial L}{\partial \lambda} = \frac{1}{\sigma^2}\sum_i (Y_i^{(\lambda)} - a\theta_i)\,(Y_i^{\lambda} \log Y_i^{\lambda} - \lambda Y_i^{(\lambda)}) + \sum_i \log Y_i$$

Tomando esperanza, podemos escoger, sin pérdida de generalidad, las unidades de medida de Y_i tales que

$$E \sum_i \log Y_i^{\lambda} = 0 \qquad 2.6$$

y además como el modelo postulado por Box & Cox supone que para algúna λ

$$Y_i^{(\lambda)} = a_i \theta_i + \epsilon_i$$

entonces

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} = -\frac{1}{\lambda_0^3 \sigma^2} E\left\{\sum_i \lambda\varepsilon_i \left[(\lambda a\theta_i + \lambda\varepsilon_i + 1)\log(\lambda a\theta_i + \lambda\varepsilon_i + 1) - \lambda a\theta_i - \lambda\varepsilon_i\right]\right\} \qquad 2.27$$

como por hipótesis $E(\lambda_0 \varepsilon_i) = 0$; $v(\lambda_0 \varepsilon_i) = \lambda_0^2 \sigma^2$ y $E(\lambda_0 \varepsilon_i)^r = \lambda_0^r \mu_{(r)}$, se ve que estos momentos ($\mu_{(r)}$ r-ésimo momento central de ε_i) son independientes de i. Como se vió en el capítulo I, $Y_i > 0$, así que $(a\theta_i + \lambda_0^{-1})\sigma^{-1}$ es mucho mayor que uno; por tanto, expandiendo en series de potencia el logaritmo y haciendo algo de álgebra, la expresión 2.7 quedará

$$-\frac{1}{\lambda_0^3 \sigma^2} \sum_i \left\{\sum_{r>2} \frac{(-1)^{r+1} \lambda_0^r \mu_{(r)}}{(r-1)(r-2)(1+\lambda_0 a\theta_i)^{r-2}} + \lambda^2 \sigma^2 \log(1+\lambda_0 a\theta_i)\right\}$$

$i = 1, \ldots, n$; $(r > 2) \in N$.

Por simplicidad etiquetaremos algunas medidas: $\gamma = \mu_{(3)}/\sigma^3$; $\alpha_i = 1 + \lambda_0 a\theta_i$; $\beta_i = \lambda_0 \sigma / \alpha_i$ y $K = \mu_{(4)}/\sigma^4 - 3$; donde podemos reconocer que β_i es el coeficiente de variación de Y_i^{λ}; efectivamente, α_i es la esperanza de Y_i^{λ}; γ es la medida de asimetría de $Y_i^{(\lambda)}$ y K la medida de kurtosis de $Y_i^{(\lambda)}$ para toda i. Tomando los dos primeros términos de la expansión y simplificando, se tiene

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} = -\frac{1}{\lambda_0 \sigma^2} \sum_i \sigma_i^2 \left[\beta_i\left(\tfrac{1}{2}\gamma - \tfrac{1}{6}\beta_i K\right) + \lambda_0 E(\log Y_i)\right] \qquad 2.8$$

como por hipótesis $\sigma_i^2 = \sigma^2$ y como se cumple 2.6 entonces 2.8 se reduce a

$$-\frac{1}{\lambda} \sum_i \beta_i \left[\frac{1}{2}\gamma + \frac{1}{6} K\beta_i\right] = 0 \qquad 2.9$$

para el primer orden en β_i, γ debe ser cero, i,e; el parámetro de transformación λ es consistentemente estimado, siempre que el término error sea simétrico ($\gamma = 0$); para el segundo orden en β_i, la condición es $\gamma = \frac{1}{3} K\beta$ si las β_i se parecen mucho a β, que será siempre que las $a\theta_i$ sean parecidas. Ahora bien, como $\beta_i^{-1} = (\lambda a\theta_i + 1)/(\lambda\sigma)$ es mucho mayor que uno, de tal forma que sea pequeño, la expansión de segundo orden también tiene un ni-

vel de simetría bastante bajo, además que el coeficiente de kurtosis también es bajo.

De aquí se concluye que aún cuando no se logre exactamente la normalidad, la aplicación de la familia 1.1 ayuda al arreglo regular de los datos.

Draper & Cox analizan por último la precisión, por lo menos aparente, del estimador de λ.

Expresando en series, hasta el segundo orden en β, el valor esperado de la segunda derivada de L, se obtiene

$$\text{Var}(\hat{\lambda}) = \frac{2}{3n\beta^2}\left(1 - \frac{1}{3}\gamma^2 + \frac{7}{8}K\right)^{-1}$$

donde el término en el paréntesis resulta el efecto de no normalidad.

Por lo tanto la precisión con la cual podemos estimar λ, depende en gran parte del coeficiente de variación de $Y^{(\lambda)}$.

5. HETEROSCEDASTICIDAD EN EL ERROR

Zarembka completa la prueba de Draper & Cox para cuando no se cumple homoscedasticidad en los errores. Lo que él prueba, es que el parámetro λ no es robusto bajo heteroscedasticidad.

Supongamos ahora que la normalidad sí se obtiene al aplicar la transformación 1.1 pero que la varianza del error no es constante a través de las observaciones, entonces como $\gamma = K = 0$, la expresión 2.8 será

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} = -\frac{1}{\sigma^2}\sum_i \sigma_i^2 E(\log Y_i) \qquad 2.10$$

y

$$\text{Var}\,(Y_i^{(\lambda)}) = \left[E(Y_i)\right]^{2\lambda-2}\text{Var}(Y_i) = \sigma_i^2,$$

sustituyendo en 2.10

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} = -\frac{1}{\sigma^2}\sum_i \left\{\left[E(Y_i)\right]^{2\lambda_0-2} \mathrm{Var}(Y_i)\right\} E\log Y_i \qquad 2.11$$

Si σ^2 se obtiene con las observaciones Y_i, entonces el estimador de λ es sesgado en aquella dirección en la cual la varianza del error tiende a estabilizarse.

Es posible obtener la magnitud de dicho sesgo, a partir de la expresión 2.11, si tan solo se da alguna suposición acerca de la relación que hay entre la varianza de las Y_i y su esperanza.

Para ésto, sea h un parámetro tal que

$$\left[\mathrm{Var}(Y_i)\right]^{1/2} \propto \left[E(Y_i)\right]^h$$

entonces la ecuación 2.11 será escrita como

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} \propto -\frac{1}{\sigma^2}\sum_i \left\{\left[E(Y_i)\right]^{2(h+\lambda_0-1)} E(\log Y_i)\right\}$$

y ahora se tendrá que

$$\sigma^2 = \frac{1}{n}\sum_i \sigma_i^2 \propto \frac{1}{n}\sum_i \left[E(Y_i)\right]^{2(h+\lambda-1)}$$

por tanto

$$E\left(\frac{\partial L}{\partial \lambda}\right)_{\lambda=\lambda_0} = -\frac{n\sum_i\left\{\left[E(Y_i)\right]^{2(h+\lambda-1)} E(\log Y_i)\right\}}{\sum_i\left[E(Y_i)\right]^{2(h+\lambda-1)}} \qquad 2.12$$

expandiendo en series $\left[E(Y_i)\right]^{2(h+\lambda-1)}$; tomando $\log E(Y_i) \doteq E(\log Y_i)$ y usando 2.6; 2.12 quedará

$$-\frac{n\sum_i\left[2(h+\lambda-1)E^2(\log Y_i) + 2(h+\lambda-1)^2 E^3(\log y_i)\right]}{\sum_i\left[1 + 2(h+\lambda-1)^2 E^2(\log Y_i)\right]}$$

donde $\sum_i E^2(\log Y_i)/n$ es la varianza muestral de $E(\log Y)$ y $-\sum_i E^3(\log Y_i)/n$ es el tercer momento que estará muy cercano a cero si $E(\log Y)$ es lo suficientemente simétrica; entonces 2.12 quedará

$$-(h+\lambda-1)\,n\,\frac{\mathrm{Var}[E(\log Y)]}{\frac{1}{2}+(h+\lambda-1)^2\,\mathrm{Var}[E(\log Y)]}\,.$$

Ahora, como $\frac{\partial L}{\partial \lambda}$ converge a $\frac{d\ Lmax(\lambda)}{d\lambda}$ y como $\mathrm{Var}[E(\log Y)]$ se puede aproximar a $\mathrm{Var}(\log Y)$ si la varianza de Y es mucho mayor que la varianza del error; entonces, un estimador consistente para λ bajo heteroscedasticidad puede darse, si se encuentra una λ tal que

$$\frac{d\ Lmax(\lambda)}{d\lambda} = (1-\lambda-h)\,n\,\frac{\mathrm{Var}\ (\log Y)}{\frac{1}{2}+(1-\lambda-h)\ \mathrm{Var}(\log Y)} \qquad 2.13$$

Esta λ puede obtenerse, igualando la expresión 2.13, para h fija, con

$$\frac{d\ Lmax(\lambda)}{d\lambda} \propto -\,n\,\frac{Y^{(\lambda)'}a_r Q^{(\lambda)}}{Y^{(\lambda)'}a_r Y^{(\lambda)}}+\frac{n}{\lambda}$$

obtenida a partir de 1.4, y usando 1.1; donde $Q^{(\lambda)}$ es el vector con elementos de la forma : $\lambda^{-1}Y_i^{\lambda}\log Y_i$.

6. MAS SOBRE SIMETRIA

Hinkley propone algunas medidas de asimetría; y escoge aquél valor del parámetro que haga óptimas dichas medidas.

Sabemos que si $Y^{(\lambda)}$ tiene una distribución simétrica, entonces los quantiles p y (1 - p) equidistarán de la mediana de la población, i,e;

$$\xi_{0.5}^{\lambda} - \xi_{p}^{\lambda} = \xi_{1-p}^{\lambda} - \xi_{0.5}^{\lambda}$$

donde, como es evidente $\xi_{0.5}^{\lambda}$ representa la mediana de los datos transformados y ξ_{p}^{λ} y ξ_{1-p}^{λ} el p-ésimo y (1 - p)-ésimo quantil.

6.1 UNA RAPIDA ELECCION DE LA TRANSFORMACION DE POTENCIA

Otra medida del grado de asimetría propuesta por Hinkley es

$$\leq = \frac{\text{media muestral - mediana}}{\text{escala muestral}} \qquad 2.14$$

la escala muestral puede ser cualquiera que sea representativa; las que propone Hinkley son la desviación estandar muestral y el rango interquartil; aclara que esta última es mas robusta en el sentido de eficiencia.

Se puede de esta manera medir $\leq$ para diferentes valores de λ que están en el rango (-1, 0, 0.5, 1, 2) y se toma aquella λ para - la cual $\leq$ se acerca a cero.

Este método es eficiente solo para aquellas muestras que son homogéneas.

Nótese que este método no pide normalidad en los datos transformados, y a pesar de que el método de muestras grandes sí pide normalidad, da una medida mucho más clara de cuánto se desvían los datos de las hipótesis ideales.

7. INVARIANZA

Hay modelos en los cuales no es posible eliminar la constante aditiva y entonces al usar la familia 1.1, llegaremos a procesos que no son invariantes bajo escalas, para $\lambda \neq 0$.

Schlesselman hace notar esta falla de la familia 1.1 en el siguiente desarrollo.

En términos de la variable estandarizada, se había visto que

$$\text{Lmax}\ (\lambda) = -\frac{1}{2}\, n \log\left(Z^{(\lambda)'} a_r Z^{(\lambda)}/n\right) .$$

El problema reside en que en aquellos modelos de los que hablamos hace un momento, puede alterarse la suma de cuadrados $S(\lambda;Z)$ al reescalar las observaciones originales; esto es, bajo la reescala $Y \xrightarrow{w} wY$; $Y^{(\lambda)}$ no va a dar a un término $wY^{(\lambda)}$, sino que nos

sobra un elemento:

$$Y^{(\lambda)} \xrightarrow{w} \frac{w^\lambda Y^\lambda - 1}{\lambda} = \frac{w^\lambda Y^\lambda}{\lambda} - \frac{w^\lambda}{\lambda} + \frac{w^\lambda}{\lambda} - \frac{1}{\lambda}$$

$$= w^\lambda \left[\frac{Y^\lambda - 1}{\lambda}\right] + \frac{1}{\lambda}(w^\lambda - 1)$$

$$= w^\lambda Y^{(\lambda)} + \lambda^{-1}(w^\lambda - 1)$$

siguendo el mismo desarrollo vemos que

$$Z^{(\lambda)'} a_r Z^{(\lambda)} \xrightarrow{w} Z^{(\lambda)'} a_r Z^{(\lambda)} + (2Z^{(\lambda)'} a_r \mathcal{G} + \mathcal{G}' a_r \mathcal{G}) \qquad 2.15$$

donde $\mathcal{G}$ es el vector con elementos $\lambda^{-1}(1 - w^\lambda)/ J^{1/n}$, considerando que $J^{1/n} \xrightarrow{w} w^\lambda J^{1/n}$, y por tanto

$$Lmax(\lambda) \xrightarrow{w} Lmax(\lambda) - \frac{1}{2} n \log \left\{ 1 + \frac{2Z^{(\lambda)'} a_r \mathcal{G} + \mathcal{G}' a_r \mathcal{G}}{nZ^{(\lambda)'} a_r Z^{(\lambda)}} \right\} \qquad 2.16$$

La familia de transformaciones 1.1 puede ser definida de muchas maneras, con tal que se den simultáneamente la continuidad en $\lambda = 0$ y la invarianza bajo escalas. De hecho, la familia 1.1 puede ser considerada como un caso particular de la familia

$$Y_c^{(\lambda)} \begin{cases} = \dfrac{Y^\lambda - C^\lambda}{\lambda} & \lambda \neq 0 \\ \log (Y/C) & \lambda = 0 \end{cases} \qquad 2.17$$

aquí C es una constante positiva que está medida en las mismas unidades de Y y $Z_c^{(\lambda)} = Y_c^{(\lambda)}/J_c^{1/n}$.

Bajo la misma reescala: $C \xrightarrow{w} wC$, tal que $Y_c^{(\lambda)} \xrightarrow{w} w^\lambda Y_c^{(\lambda)}$, entonces $S(\lambda;Z)$ es invariante bajo la escala; y como $Y^{(\lambda)} = Y_c^{(\lambda)} + C^{(\lambda)}$, entonces

$$Z^{(\lambda)'} a_r Z^{(\lambda)} \xrightarrow{w} Z_c^{(\lambda)'} a_r Z_c^{(\lambda)} + (2Z_c^{(\lambda)'} a_r C_*^{(\lambda)} + C_*^{(\lambda)'} a_r C_*^{(\lambda)}) \qquad 2.18$$

donde el vector cuyos elementos son $C^{(\lambda)}/J^{1/n}$ se denota como $C_*^{(\lambda)}$.

Nótese además que $J_c = J$. El término en el paréntesis tanto en la expresión 2.16 cono en la 2.17 es cero si la condición siguiente se cumple: el vector de parámetros θ tiene un parámetro que representa una media general, con la correspondiente columna en la matriz a que consiste de unos; entonces todo renglón o columna de a suma cero ($1'a_r = 0 = a_r 1$); y por tanto se logrará la invarianza bajo escala.

Schlesselman, finalmente expone varias familias de transformaciones que son invariantes bajo escalas y contínuas en $\lambda = 0$, pero tienen la desventaja de ser relativamente dependientes de los valores de la muestra. Las expresiones se obtienen sustituyendo C en 2.17 por la esperanza muestral ($\overline{Y}$); la media geométrica muestral ($\dot{Y}$) o la media muestral (Y').

Se puede concluir que si se incluye un término constante al modelo, se logrará la invarianza de escala

8. TRANSFORMACION DE VARIABLES EN SISTEMAS DE ECUACIONES SIMULTANEAS

Una vez más es usada la familia 1.1 por Tintner & Kadekodi; en el sistema

$$AY + \mathcal{E} = 0 \qquad 2.19$$

donde A es una matriz de orden g y Y tendrá dimensión $n = g + h$.

Si se particiona de una manera conveniente el vector Y;

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

donde Y_1 es un vector de g variables endógenas y Y_2 de variables predeterminadas, y se particiona también A, de forma congruente a Y

$$A = [A_1, A_2]$$

se tendrá que 2.19 se puede expresar como

$$A_1 Y_1 + A_2 Y_2 + \varepsilon = 0 \qquad 2.20$$

Aplícando la transformación a 2.20 se tiene

$$A_1^* Y_1^{(\lambda)} + A_2^* Y_2^{(\lambda)} + \varepsilon^* = 0 \qquad 2.21$$

ahora A_1^* y A_2^* van a ser las matrices asociadas con las variables transformadas y ε^* el vector de errores después de la transformación.

Del sistema 2.21 se obtiene la forma reducida

$$Y_1^{(\lambda)} = -A_1^{*-1} A_2^* Y_2^{(\lambda)} - A_1^{*-1} \varepsilon^* \qquad 2.22$$

entonces Lmax(λ) correspondiente a este sistema es

$$-\frac{n}{2} \log \sigma(\lambda; Y) + (\lambda - 1) \sum_i \sum_q \log Y_{iq} + g\, n \log|\lambda| \; ;$$

maximizando por el método de mínimos cuadrados, se pueden establecer intervalos de confianza, y se continúa con el proceso de prueba.

Si se usa un valor de λ en las variables endógenas distinto al de las predeterminadas, se dificulta el cálculo del proceso, sobre todo si el modelo es no lineal.

9. UN ANALISIS DE TRANSFORMACIONES PARA SERIES DE TIEMPO

Ansley, Spivey y Wrobleski usan la familia paramétrica de transformaciones propuesta por Box & Cox para lograr un proceso homogeneo y estacionario en series de tiempo; desarrollan después de haber llegado a este proceso la función de verosimilitud, y un algoritmo del parámetro λ.

En esta sección pensaremos que $Y = Y_t$ son observaciones de una serie de tiempo. Supóngase que $Y^{(\lambda)} = Y_t^{(\lambda)}$, definida por 1.1, si

gue un proceso ARIMA (Integrated Autorregresive - Moving - Average) de orden (p,d,q) se supone que ε_t sigue las hipótesis ideales. Los parámetros de este proceso serán los autorregresivos: $\phi_1,\ldots,\phi_p$ y los de promedio movible: $\theta_1,\ldots,\theta_q$; δ es el término constante del modelo

$$U_t = \phi_1 U_{t-1} + \ldots + \phi_p U_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \qquad 2.23$$

donde $U_t = \nabla^d Y_t^{(\lambda)}$ es la d-ésima diferencia de $Y_t^{(\lambda)}$. Es obvio que U_t forme un proceso ARMA (Autorregresive - Moving - Average) de orden (p,q), pues, como ya se mencionó antes, $Y_t^{(\lambda)}$ formaba un proceso ARIMA; entonces tomando d observaciones transformadas ($Y_t^{(\lambda)}$) fijas y multiplicando por el jacobiano de la transformación para la d-ésima diferencia; la densidad conjunta de $Y_t^{(\lambda)}$ es

$$p(Y^{(\lambda)} | Y_t^{(\lambda)},\ldots, Y_{t-d+1}^{(\lambda)}, \phi,\theta,\delta,\sigma^2;\lambda) = (2\pi\sigma^2)^{-n/2} |M|^{1/2} \exp\left\{-\frac{S(\lambda,U)}{2\sigma^2}\right\} J \qquad 2.24$$

donde M es el inverso de la matriz de correlación de la nueva serie U_t y $S(\lambda;U)$ la suma de cuadrados de los residuales de U_t. Como se puede notar, 2.24 es la función de verosimilitud, pues $|M|$ está cercano a uno y puede despreciarse. Por tanto tomando logaritmo se tiene

$$L \propto -\frac{n}{2}\log\sigma^2 - \frac{S(\lambda;U)}{2\sigma^2} + \log J.$$

Maximizando respecto a los parámetros

$$L\max(\lambda) \propto -\frac{n}{2}\log\frac{S(\lambda;U)}{n} + \log J \qquad 2.25$$

si estandarizamos U_t : $Z_t^{(\lambda)} = U_t / J^{1/n}$; 2.25 quedará expresada como

$$L\max(\lambda;Z_t) = K - \frac{n}{2}\log\frac{S(\lambda;Z_t^{(\lambda)})}{n}$$

y por tanto el estimador máximo verosímil de λ será obtenido minimizando la suma de cuadrados de los residuales para las variables estandarizadas.

10. OTRA EXTENSION DEL ANALISIS DE TRANSFORMACIONES

Es importante el poder aplicar más de una transformación a un conjunto de datos, para obtener simultáneamente las hipótesis ideales.

Wood, supone un parámetro λ que define una transformación; supone además que $E(Y^{(\lambda)})$ no logra una forma lineal, así que propone otra transformación λ_1, aplicada simultáneamente con λ, para inducir normalidad. Para ésto, definimos

$$\psi = f^{-1}\left[E(Y^{(\lambda)})\right] \qquad 2.26$$

la transformación de la esperanza de $Y^{(\lambda)}$ a la escala original. Entonces, para λ y λ_1 fijas, y para la variable normalizada $Z^{(\lambda)}$.

$$\begin{aligned} Lmax\ (\lambda,\lambda_1) &= -\frac{1}{2}\, n\, \log\, \hat{\sigma}^2(\lambda,\lambda_1;Z) \\ &= -\frac{1}{2}\, n\, \log\, S(\lambda,\lambda_1;Z)/n \end{aligned}$$

Wood supone que λ y λ_1 pertenecen a la familia 1.1, así $Y^{(\lambda)}$ y $\psi^{(\lambda_1)}$ tendrán la misma forma.

Si $\lambda=\lambda_1$, el caso se reduce al de Box & Cox, así pues se considera $\lambda \neq \lambda_1$. Para hacer pruebas de hipótesis, simplemente se aplica 1.13

$$\begin{aligned} Lmax(\lambda,\lambda_1) = Lmax(\lambda,\lambda_1) &+ Lmax(\lambda,\hat{\lambda}_1) - Lmax(\lambda,\lambda_1) \\ &+ Lmax(\hat{\lambda},\hat{\lambda}_1) - Lmax(\lambda,\hat{\lambda}_1) \end{aligned} \qquad 2.27$$

los dos primeros términos de 2.27 son asintóticamente e independientemente distribuidos como una χ^2.

11. TRANSFORMACION DE VARIABLES INDEPENDIENTES

Box & Tidwell suponen que los errores en las observaciones son independientes y se acercan a una normal con varianza constante; así que sólo se preocupan por encontrar una transformación, para reducir a su forma más simple, la función de las variables independientes.

11.1 PROCESO DE TRANSFORMACION

Supóngase que se analizan n observaciones $Y_1,\ldots,Y_n$ y n conjuntos de condiciones (variables independientes) $a_1,\ldots,a_n$ donde a_i es un vector columna de dimensión k.

Supóngase que la respuesta $E(Y_i)$ es una función de las variables ya transformadas $\alpha_i = \alpha_i(a_i;\lambda_i)$ $(\alpha_i = \alpha_{i1},\ldots,\alpha_{ik}$ y $\lambda_i = \lambda_{i1},\ldots,\lambda_{ir_i})$ y de los parámetros θ asociados a las variables transformadas:

$$E(Y_i) = f(\alpha_i,\theta)$$

Supóngase que $\lambda_i^{(o)}$ es la primera conjetura acerca del parámetro λ_i; entonces $\alpha_i^{(o)}$ es el vector con elementos de la forma $\alpha_i(a_{iu},\lambda_i^{(o)})$.

Si se expande la función de respuesta en series, alrededor del parámetro $\lambda_i^{(o)}$ se tendrá aproximadamente

$$E(Y_i) = f(\alpha_i^{(o)},\theta) + \sum_i\sum_j (\lambda_{ij}-\lambda_{ij}^{(o)}) \left\{\frac{\partial f(\alpha_i,\theta)}{\partial \lambda_{ij}}\right\}_{\lambda_i=\lambda_i^{(o)},\ \alpha=\alpha_i^{(o)}} \qquad 2.28$$

donde

$$\left\{\frac{\partial f(\alpha,\theta)}{\partial \lambda_{ij}}\right\}_{\lambda_i=\lambda_i^{(o)},\alpha=\alpha_i^{(o)}} = \left\{\frac{\partial f(\alpha,\theta)}{\partial \alpha_i}\right\}_{\alpha=\alpha_i^{(o)}} \left\{\frac{\partial \alpha_{iu}}{\partial \lambda_{ij}}\right\}_{\lambda_i=\lambda_i^{(o)}}$$

$\left\{\frac{\partial \alpha_{iu}}{\partial \lambda_{ij}}\right\}$ puede fácilmente calcularse, conociendo la forma de la transformación α_i mientras que $\left\{\frac{\partial f(\alpha,\theta)}{\partial \alpha_i}\right\}$ necesita de un ajuste preliminar de las observaciones por el método de mínimos cuadrados.

Digamos que el ajuste es

$$Y_i = f(\alpha_i^{(o)},\hat{\theta}) \qquad 2.29$$

donde $\hat{\theta}$ es el estimador de θ por mínimos cuadrados.

Diferenciando la expresión 2.29 respecto a $\alpha_i^{(o)}$ obtendremos valores que se aproximan a $\left\{\frac{\partial f(\lambda,\theta)}{\partial \alpha_i}\right\}$; se sustituyen estos valores en la ecuación 2.28 y se resuelve. Una vez obtenido el nuevo valor de λ_{ij} se sustituye en el valor de $\lambda_{ij}^{(o)}$ y se vuelve un proceso

iterativo, hasta obtener un valor del parámetro de transformación cercano al real.

Proponen Box & Tidwell que un método con frecuencia mas conveniente para encontrar valores de $\left\{\frac{\partial f(d,\theta)}{\partial d_i}\right\}$, es calculando las partes ortogonales de $\left\{\frac{\partial f(d,\theta)}{\partial \lambda_i}\right\}$, que no son más que los residuales en el modelo; y al ir iterando, se puede observar, que el proceso converge en la parte ortogonal.

Pueden construirse, otra vez, intervalos de confianza para los parámetros de la transformación, en base a dichas funciones ortogonales y compararse con los obtenidos por el método de muestras grandes. Pueden usarse también pruebas de bondad de ajuste y tener así un análisis completo.

12. LA EXTENSION DE ZELLNER

Hasta aquí hemos visto la transformación de la variable dependiente, o de la independiente, pero en esta sección, veremos la transformación de ambas variables simultáneamente.

Zellner considera el modelo

$$Y = a\theta + \varepsilon$$

donde a no necesariamente es lineal, al igual que Y.

Supóngase que se usa una λ tal que $Y^{(\lambda)}$ y $a^{(\lambda)}$ son de la forma 1.1, entonces la función de verosimilitud 1.3 puede escribirse como

$$(2\pi)^{n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}(Y^{(\lambda)}-a^{(\lambda)}\theta)'(Y^{(\lambda)}-a^{(\lambda)}\theta)\right\} J(\lambda;Y) \qquad 2.30$$

maximizando el logaritmo de la expresión 2.3o respecto a θ, σ^2 y λ tenemos

$$Lmax(\lambda) = -\frac{n}{2}\log\hat{\sigma}^2(\lambda) + (\lambda-1)\sum_i \log Y_i$$

donde

$$\hat{\sigma}^2 = \frac{1}{n}\left[Y^{(1)} - a^{(1)}\hat{\theta}\right]'\left[Y^{(1)} - a^{(1)}\hat{\theta}\right]$$

y

$$\hat{\theta} = \left[a^{(1)\prime} a^{(1)}\right]^{-1} a^{(1)\prime} Y^{(1)}$$

y se sigue el proceso visto en el capítulo I.

Como puede verse son muchas las aplicaciones y variaciones que se pueden hacer acerca de la transformación de variables y es además una teoría muy útil en la práctica, aunque a veces difícil de usar.

CAPITULO III

"ILUSTRACION DE LA TEORIA"

1. DESCRIPCION DEL EJEMPLO

Se intenta ajustar un modelo lineal a las respuestas obtenidas de un experimento, que consistió en lo siguiente:

Cuarenta y ocho animales fueron agrupados en doce celdas (cuatro animales en cada celda), se les aplicó un veneno y un tratamiento distinto por celda; en total tres venenos y cuatro tratamientos, y se midió el tiempo de supervivencia (la unidad fué 10 horas) después de las aplicaciones. La tabla 3.1 muestra las respuestas.

2. ANALISIS DE LOS DATOS

Una gráfica de la desviación estandar contra la media, muestra una marcada tendencia a aumentar, dentro de las celdas, la varianza, en la medida que las medias, también dentro de las celdas, crece. La figura 3.1 indica dicha tendencia.

Un análisis de varianza de los datos originales, muestran una posible interacción; nótese el valor de los cuadrados medios de P x T.

Es deseable que los efectos de venenos y tratamientos, sea aditivo, así, si $\theta_{t.} - \theta_{..}$ representa el cambio medio en el tiempo de supervivencia producido por el t-ésimo tratamiento y $\theta_{.p} - \theta_{..}$ el cambio medio producido por el p-ésimo veneno, se tendría un modelo de la forma

TABLA 3.1

Veneno	Tratamiento A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Tiempo de supervivencia de animales. Unidad 10 horas

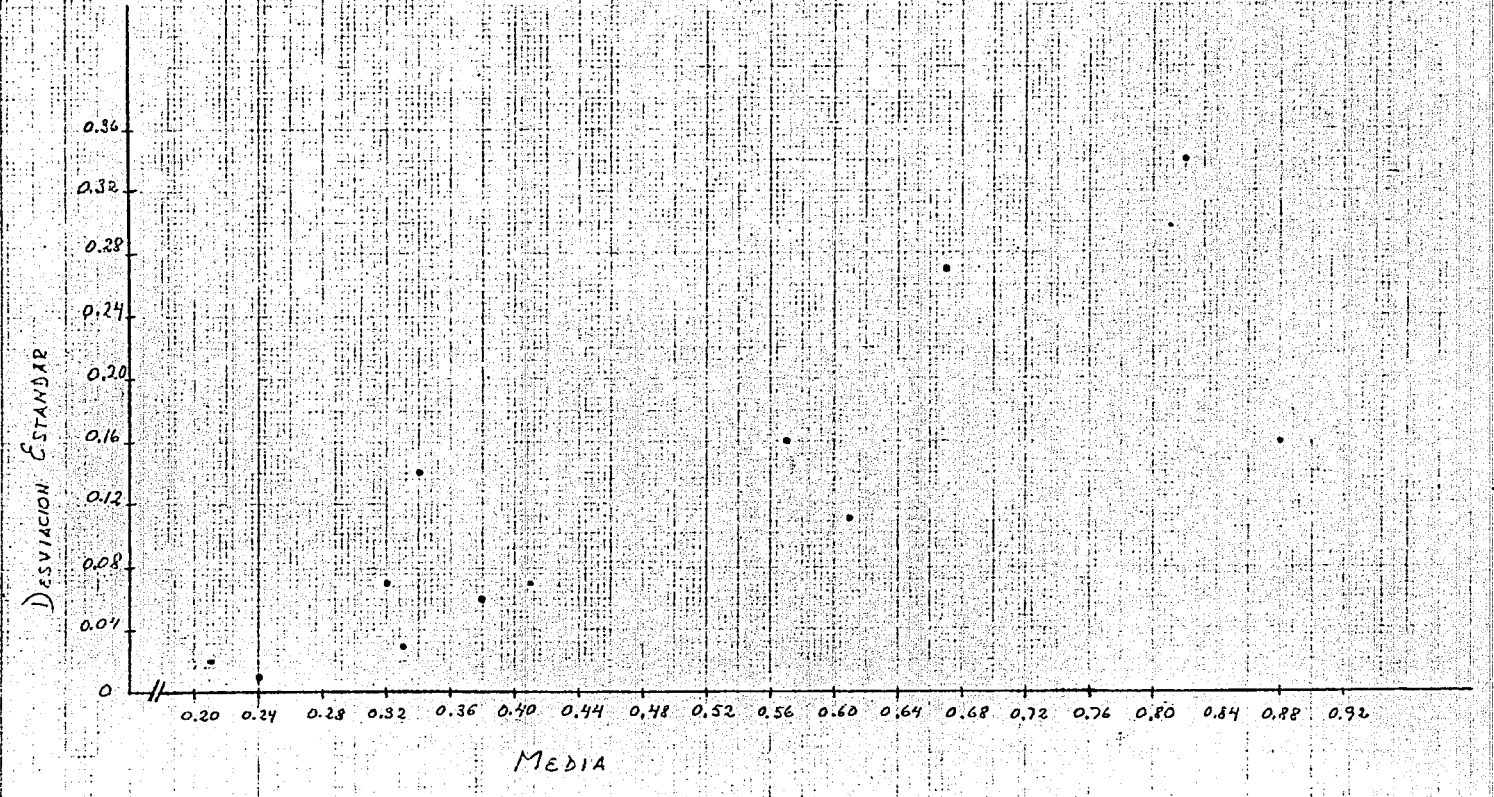
TABLA 3.2

Tabla de Análisis de Varianza para Datos no Transformados

Fuente de Variación	Grados de Libertad	Cuadrados Medios x 1000
Veneno	2	516.5
Tratamiento	3	307.1
P x T	6	41.7
Dentro de Grupos	36	22.2

FIGURA 3.1

MEDIA Y DESVIACIÓN ESTANDAR DENTRO DE LAS CELDAS



$$E(Y_i) = \theta_\bullet + (\theta_{t\cdot} - \theta_\bullet)a_{ti} + \ldots + (\theta_{4\cdot} - \theta_\bullet)a_{4i} +$$

$$+ (\theta_{i_j} - \theta_\bullet)a_{ji} + \ldots + (\theta_{\cdot 3} - \theta_\bullet)a_{3i} + \varepsilon_i \qquad 3.1$$

donde a_{ti} toma los valores 1 o 0 si el i-ésimo animal recibió o no el t-ésimo tratamiento y a_{pi} toma los valores 1 o 0 si el i-ésimo animal recibió o no el p-ésimo veneno.

Es fácil ver que la matriz indicacor a, construida a partir del modelo 3.1, no es de rango completo, por lo que proponen Box & Cox, omitir los términos $(\theta_{4\cdot} - \theta_\bullet)a_{4i}$ y $(\theta_{\cdot 3} - \theta_\bullet)a_{3i}$.

Se puede escribir el modelo en una forma más general

$$E(Y_i) = \theta_\bullet + \sum_t(\theta_{t\cdot} - \theta_\bullet)a_{ti} + \sum_p(\theta_{\cdot p} - \theta_\bullet)a_{pi} + \sum_{t,p}(\theta_{tp} - \theta_{t\cdot} - \theta_{\cdot p} + \theta_\bullet)a_{tpi} \quad 3.2$$

y se puede obtener el modelo 3.1 eliminando la última suma de 3.2 que contiene los seis parámetros interacción independientes (no se olvide que se han eliminado algunos términos para hacer de rango completo la matriz indicador, por lo que t = 1,2,3 y p = 1,2); y tan solo quedan seis parámetros funcionalmente aditivos.

3. ESTIMACION DEL PARAMETRO DE TRANSFORMACION

Como ya se había visto en la tabla de Análisis de Varianza que hay una posible interacción entre veneno y tratamiento, es necesario aplicar una transformación a los datos para poder ajustar el modelo 3.1

Box & Cox propusieron la transformación de potencia 1.1 y trabajaron con la variable estandarizada $Z^{(\lambda)}$ que tendrá para $\lambda \neq 0$ la expresión

$$Z^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda \dot{Y}^{\lambda-1}}$$

Ahora ya eliminados los términos de interacción, la suma de -

$$E(Y_i) = \theta_\cdot + (\theta_{1\cdot} - \theta_\cdot)a_{1i} + \ldots + (\theta_{4\cdot} - \theta_\cdot)a_{4i} +$$

$$+ (\theta_{\cdot 1} - \theta_\cdot)a_{\cdot i} + \ldots + (\theta_{\cdot 3} - \theta_\cdot)a_{3i} + \varepsilon_i \qquad 3.1$$

donde a_{ti} toma los valores 1 o 0 si el i-ésimo animal recibió o no el t-ésimo tratamiento y α_{pi} toma los valores 1 o 0 si el i-ésimo animal recibió o no el p-ésimo veneno.

Es fácil ver que la matriz indicacor a, construida a partir del modelo 3.1, no es de rango completo, por lo que proponen Box & Cox, omitir los términos $(\theta_{4\cdot} - \theta_\cdot)a_{4i}$ y $(\theta_{\cdot 3} - \theta_\cdot)\alpha_{3i}$.

Se puede escribir el modelo en una forma más general

$$E(Y_i) = \theta_\cdot + \sum_t(\theta_{t\cdot} - \theta_\cdot)a_{ti} + \sum_p(\theta_{\cdot p} - \theta_\cdot)\alpha_{pi} + \sum_{t,p}(\theta_{tp} - \theta_{t\cdot} - \theta_{\cdot p} + \theta_\cdot)a_{tpi} \quad 3.2$$

y se puede obtener el modelo 3.1 eliminando la última suma de 3.2 que contiene los seis parámetros interacción independientes (no se olvide que se han eliminado algunos términos para hacer de rango completo la matriz indicador, por lo que t = 1,2,3 y p = 1,2); y tan solo quedan seis parámetros funcionalmente aditivos.

3. ESTIMACION DEL PARAMETRO DE TRANSFORMACION

Como ya se había visto en la tabla de Análisis de Varianza que hay una posible interacción entre veneno y tratamiento, es necesario aplicar una transformación a los datos para poder ajustar el modelo 3.1

Box & Cox propusieron la transformación de potencia 1.1 y trabajaron con la variable estandarizada $Z^{(\lambda)}$ que tendrá para $\lambda \neq 0$ la expresión

$$Z^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda \dot{Y}^{\lambda-1}}$$

Ahora ya eliminados los términos de interacción, la suma de -

cuadrados $S(\lambda;Z)$ tendrá 42 grados de libertad (recuérdese que eran seis los parámetros de interacción).

Para diferentes valores de λ se calcula $Lmax(\lambda)$ y $P(\lambda|Y)$ en un rango aproximado para λ de (-2, -1, 0, 0.5, 1, 2).

La expresión para $Lmax(\lambda)$ y $P(\lambda|Y)$ (tomando en esta última logaritmo, i,e; usando $L_b(\lambda)$) para este ejemplo son respectivamente

$$-\frac{48}{2}\log\left\{S(\lambda;Z)/48\right\} = \log\left\{S(\lambda;Z)\right\}^{-24} + 92.91$$

$$-\frac{42}{2}\log\left\{S(\ ;Z)/42\right\} = 0.866\times10^{-10}\left\{S(\ ;Z)\right\}^{-21}$$

La figura 3.2 muestra las curvas $Lmax(\lambda)$ y $P(\lambda|Y)$. Analizando esta gráfica, se ve que un valor óptimo para λ sería -0.75, que se encuentra en el intervalo (-1.13, -0.37) obtenido a partir de 1.5 a un nivel de significancia del 5%.

También se puede apreciar que la distribución $P(\lambda|Y)$ se aproxima a una normal con media -0.75 y varianza 0.048; y que el 95% de la población se encuentra dentro de los límites -1.18 y -0.32

Box & Cox propusieron un valor para λ igual a -1 ya que era un valor apropiado para medir (podría interpretarse) una tasa de mortandad.

Analizando la tabla de Análisis de Varianza bajo la transformación recíproca, se ve que la hipótesis de normalidad se da más cercana con los datos transformados. Compárese con la tabla de Análisis de Varianza para los datos no transformados.

El cuadrado medio dentro de los grupos, se redujo casi en un tercio y ahora se parece mucho a los cuadrados medios de los venenos y tratamientos P×T.

De aquí podría concluirse que cuando los datos regresan a su métrica original, la dispersión de la distribución posterior

TABLA 3.3

Tabla de Análisis de Varianza para Datos Transformados $\lambda = -1$

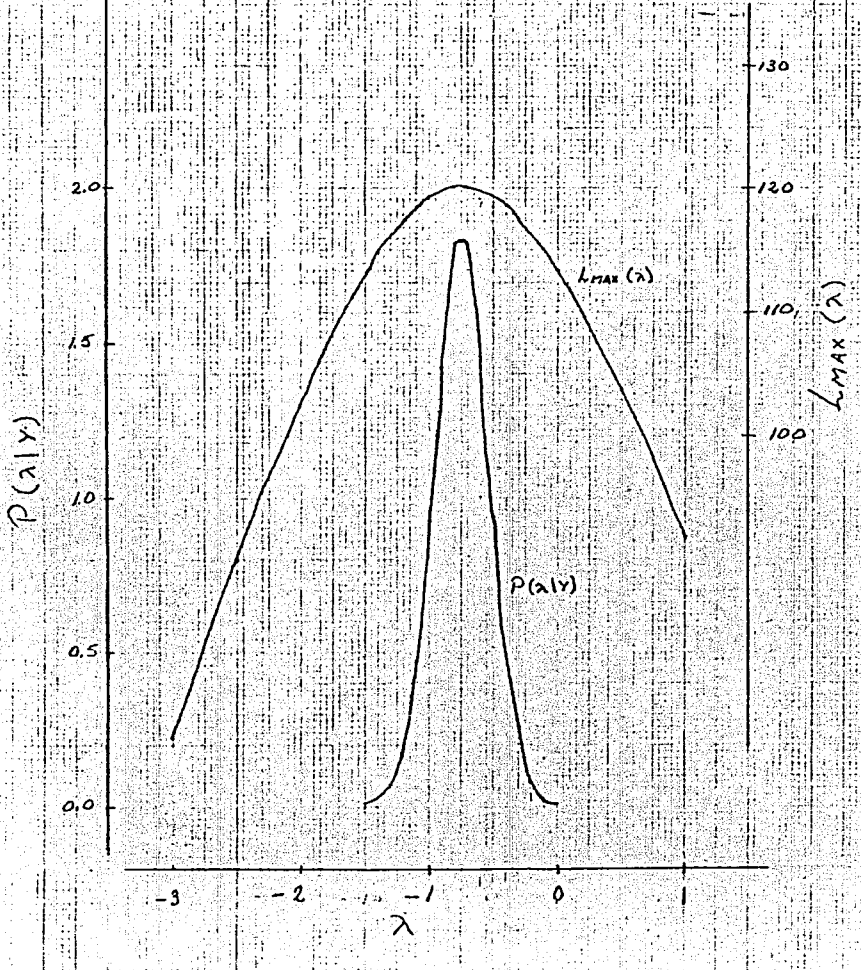
Fuente de Variación	Grados de Libertad	Cuadrados Medios x 1000
Veneno	2	568.7
Tratamiento	3	221.9
P x T	6	8.5
Dentro de Grupos	36	7.8

TABLA 3.4

	Datos Originales	Datos Modificados (1 valor errático)
Método de Verosimilitud		
75% intervalo de conf.	(-0.95, -0.05)	(-0.3. 0.05)
estimador maximoverosímil	$\hat{\lambda} = -0.75$	$\hat{\lambda} = -0.15$
Método de Significancia		
75% intervalo de conf.	(-0.9, 0.05)	(-1.2, 0.0)
mínimo estimador F	$\hat{\lambda} = -0.5$	$\hat{\lambda} = -0.5$

FIGURA 3.2

FUNCIONES: $L_{MAX}(\lambda)$ Y $P(\lambda|Y)$



de los efectos se reduce en un factor de $\sqrt{3}$.

Se podría analizar ahora el ejemplo en tres partes, como se hizo en el capítulo I, de una manera teórica; usando 1.13. Supóngase primero que se ha escogido una λ tal que se da normalidad en todas las celdas pero no varianza constante ni aditividad; entonces, para cada celda

$$Y_j^{(\lambda)} = \theta_j 1 + \varepsilon_j \qquad j = 1,\ldots,12$$

donde para la celda j; ε_j es una normal esférica $N_{n_j}(0,\sigma_j^2 I)$; $n_j = 1,\ldots,4$.

Considérese ahora que se logra la homoscedasticidad; entonces se usa el criterio de Neyman Pearson (visto en el capítulo I); y por último, se supondrá aditividad y se usará el criterio de Bartlett:

$$Lmax(\lambda|A,H,N) = Lmax(\lambda|N) + \log\left[\frac{\prod(S^{(i)}(\lambda;Z)/n)^{\frac{n_i}{2}}}{(S(\lambda;Z)/n)^{\frac{n}{2}}}\right] - \frac{1}{2}n\log\left[1 + \frac{\nu_2}{\nu_r}F(\lambda;Z)\right]$$

Ahora usando 1.15 se hará el mismo proceso y se tendrá

$$P(\lambda|A,H,N) \propto P(\lambda|N)\,P(\sigma_1^2=\ldots=\sigma_{12}^2|\,N).\,P(\theta_2=0|\lambda,H,N)$$

El análisis se hará a partir de las figuras 3.3 y 3.4 que muestran el efecto que tiene el incluir secuencialmente las hipótesis N, H y A

Primero se ve a partir de Lmax(λ|N) y P(λ|N), que es insignificante la información acerca de λ que proporciona la normalidad dentro de las celdas; mientras que, considerando varianza constante, mejora de una manera considerable la información acerca de λ ; sin embargo, esta información no mejora mucho suponiendo un modelo aditivo, lo cual lleva a concluir que la verosimilitud está determinada por la homoscedasticidad.

FIGURA 3.3

FUNCIÓN $L_{MAX}(\lambda)$ BAJO DIFERENTES MODELOS : ADITIVIDAD, HOMOSCEDASTICIDAD Y NORMALIDAD.

LAS FLECHAS MUESTRAN EL 95% DEL INTERVALO DE CONFIANZA PARA λ

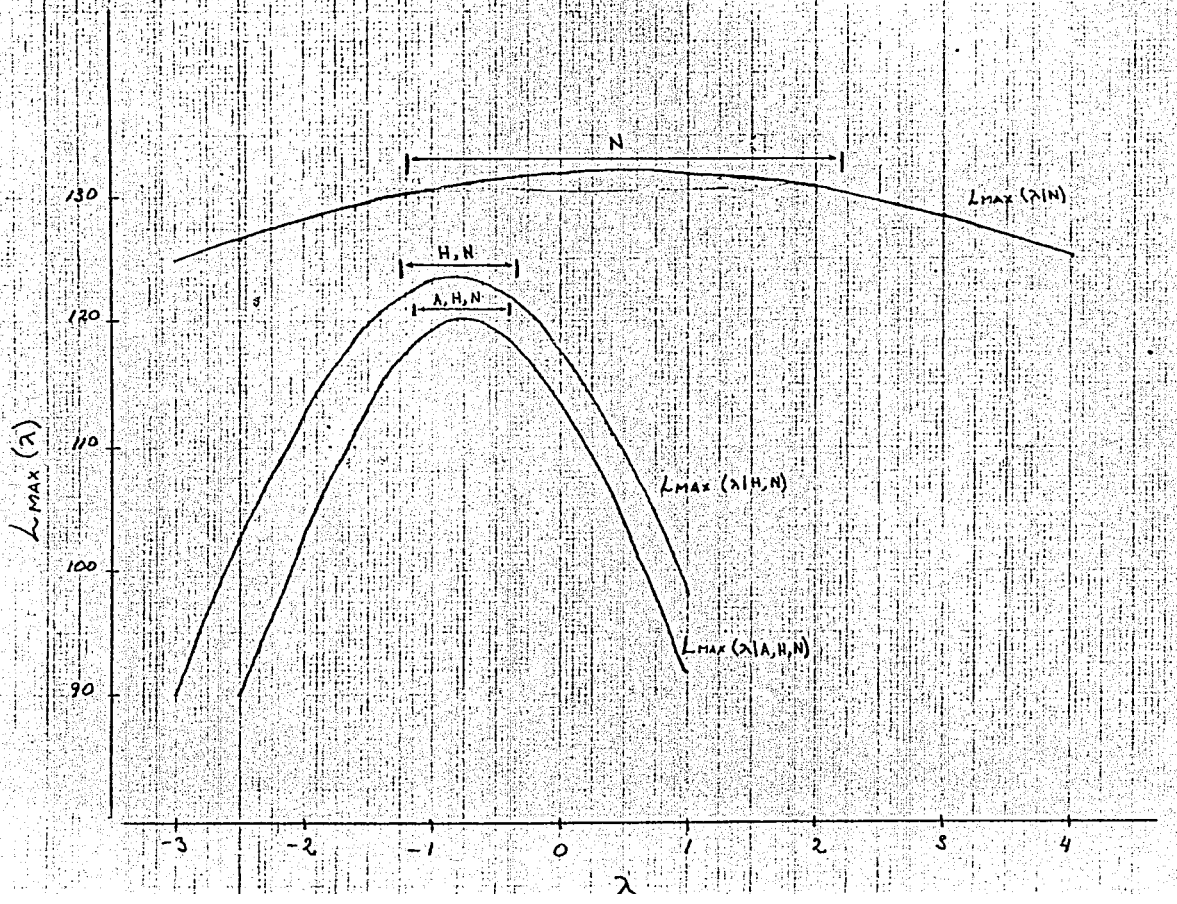
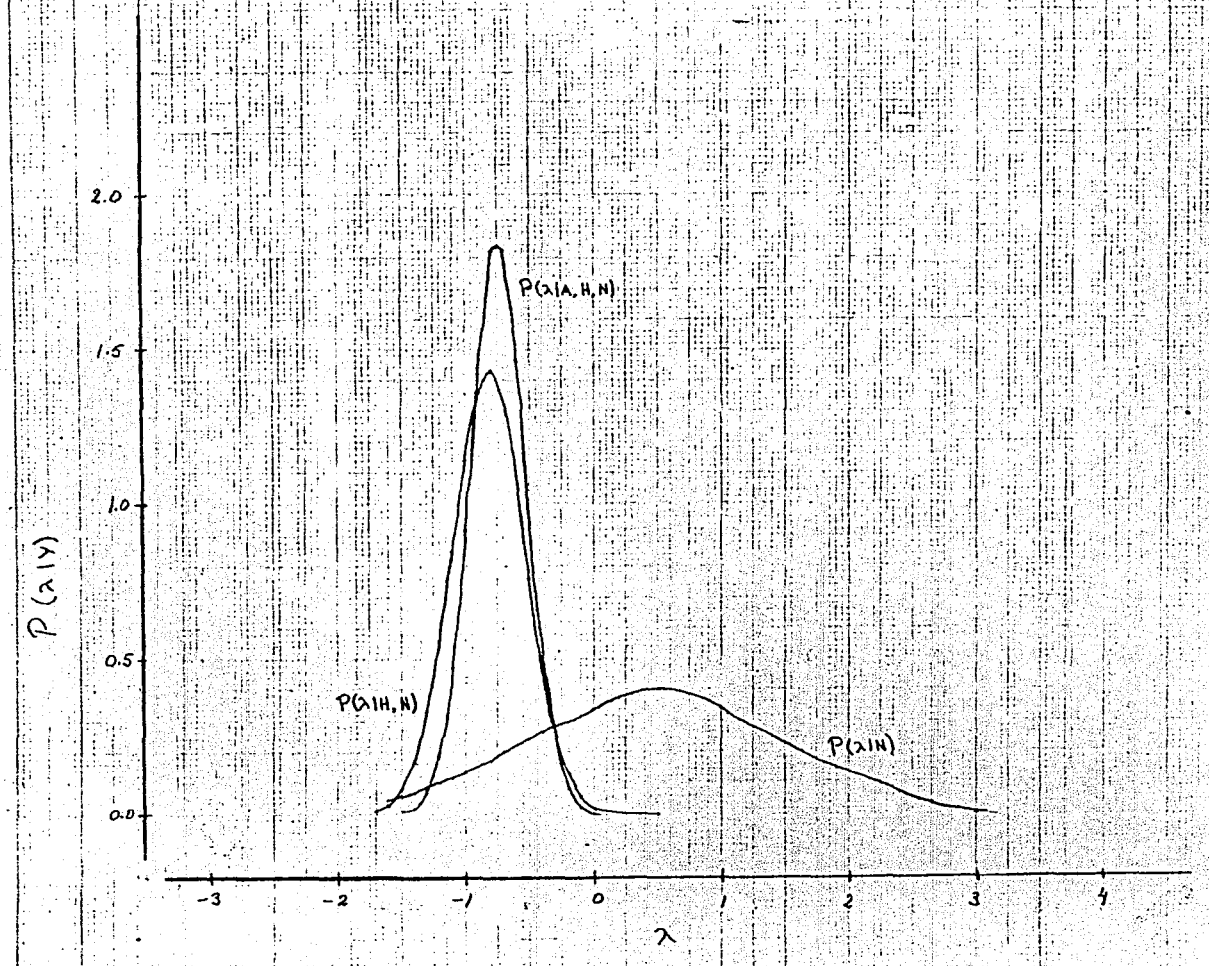


FIGURA 3.4

FUNCIÓN $P(\lambda|Y)$ BAJO DIFERENTES MODELOS:
ADITIVIDAD, HOMOSCEDASTICIDAD, NORMALIDAD.



Finalmente, se analiza a partir de la curva $M(\lambda;Z)$ en la figura 3.5 que la restricción de homoscedasticidad es, en efecto, compatible con valores de λ en el rango que incluye el valor $\lambda = -1$; y la curva $F(\lambda;Z)$ indica que en el mismo rango, los datos son consistentes con aditividad. Se ve además que $M(\lambda;Z)$ alcanza su mínimo cercano a $\lambda = -1$

4. LAS PRUEBAS DE ANDREWS Y ATKINSON

Podríamos ahora aplicar las pruebas T_A y T_b propuestas por Andrews y Atkinson respectivamente, y comparar con el análisis y resultados anteriores.

Usando la estadística T_A (2.1) se encuentra una región de confianza para λ de (-1.18, 0.40) a un nivel de significancia del 5%, este intervalo contiene el valor $\lambda = -1$. Pero se encuentra un valor óptimo para λ de -0.5.

Como habíamos dicho en el capítulo II Andrews investiga qué tan sensible a los valores erráticos son los métodos de verosimilitud y el "exacto" y encuentra que los estimadores maximoverosímiles para el método de verosimilitud se reduce mucho dentro del 75% de los límites de confianza; véase tabla 3.4

Es fácil ver que el estimador maximoverosímil se vió altamente afectado, mientras que el estimador F-mínimo no se vió afectado. De los intervalos de confianza, el mas afectado fué el obtenido por el método para muestras grandes. Razones por las cuales se había concluido, en el capítulo II sección 1, que el método de verosimilitud es más sensible a fallas de normalidad.

La figura 3.6 presenta la potencia de la prueba para diferentes valores de λ ; para probar $\lambda = -1$. En ella la pendiente de la curva indica la potencia de la prueba sin tomar en cuenta el tamaño.

FIGURA 3.5

CRITERIO DE BARTLETT $M(\lambda;z)$ PARA HOMOSCEDASTICIDAD COMO UNA FUNCIÓN DE λ

RAZON DE VARIANZA $F(\lambda;z)$ PARA INTERACCIÓN CONTRA ERROR COMO UNA FUNCIÓN DE λ

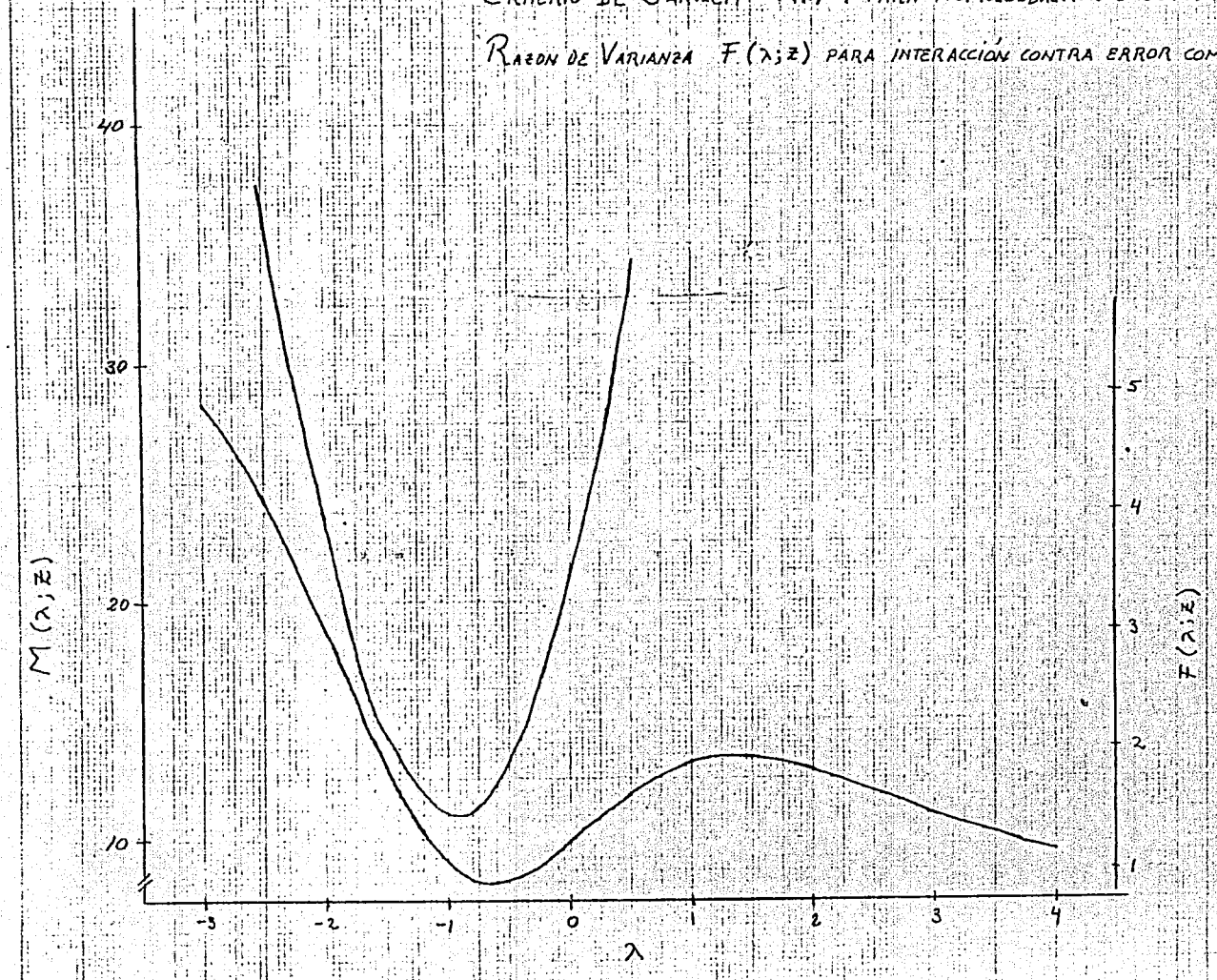
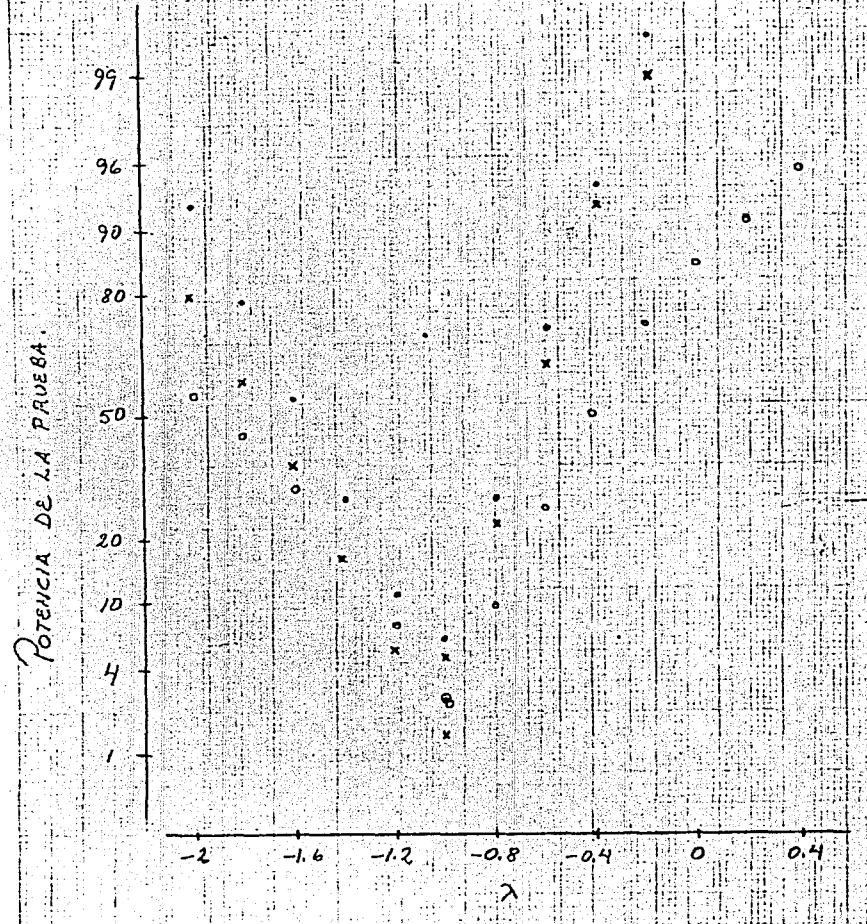


FIGURA 3.6

POTENCIA DE LAS ESTADISTICAS T_A (o); T_L (x) Y T_D (·), PARA PROBAR $\lambda = 1$ A UN NIVEL DE SIGNIFICANCIA DEL 5 %



Atkinson tan solo aplica su prueba a este ejemplo y hace comparaciones con las estadísticas anteriores.

La región de confianza obtenida a partir de T_b es (-1.09,0.41) al mismo nivel de significancia. Haciendo una comparación de los tres intervalos de confianza obtenidos, el mayor es el obtenido por el método de significancia (de Andrews)

Haciendo la comparación a partir de la potencia de las estadísticas para probar $\lambda=-1$, se concluye que las pruebas que se derivan de la función de verosimilitud, son uniformemente más potentes que la prueba exacta (recuérdese capítulo II sección 3) Una de las razones se como se vió en el capítulo II, que la baja potencia de la estadística T_A se debe a la omisión del jacobiano; y la otra razón es que T_A está basada en los valores esperados por celda, por tanto no se obtiene información acerca de la transformación; de aquí que la tabla 3.4 muestre que T_A es más robusta que T_L, bajo la presencia de valores erráticos.

5. OTROS CRITERIOS PARA SELECCIONAR TRANSFORMACIONES

Aplicar el criterio de Hinkley, para seleccionar el parámetro de transformación, sería de gran utilidad, pues daría un valor a priori de λ que nos acercaría bastante al verdadero valor de λ y ahorraría mucho trabajo, además que el valor seleccionado bajo este criterio, nos acercaría a la simetría de los datos; punto esencial para la normalidad.

Aplicaremos entonces 2.14, usando como escala muestral, la desviación estandar y el rango intercuartil; la tabla 3.5 nos dará una idea de los resultados,

Puede observarse que el valor de λ para el cual $\trianglelefteq$ es más cercano a cero es -1; además de que $\trianglelefteq$ usando $f_{0.75}-f_{0.25}$ da para $\lambda=-1$ el valor más pequeño, se había visto que era más robusto calcu-

TABLA 3.5

Medidas de Asimetría para Datos Transformados y Escalas

	Parámetro de Transformación λ				
	-1	0	0.5	1	2
Media Muestral	-1.62	-0.86	-0.66	-0.52	-0.35
Mediana Muestral	-1.50	-0.92	-0.74	-0.6	-0.42
Desv. Estandar σ	1.18	0.49	0.34	0.25	0.16
Rango Interquartil	1.73	0.73	0.48	0.33	0.15
Valor de $\leq$ usando σ	-0.10	0.12	0.24	0.32	0.41
Valor de $\leq$ usando RI	-0.07	0.08	0.17	0.25	0.44

lar $\leq$ usando el rango intercuartil como medida de escala, que usando la desviación estándar. Además puede observarse que el valor de λ aquí obtenido, es el mismo que el propuesto por Box & Cox, y el obtenido a partir de la estadística T_3.

CONCLUSIONES

Como ha podido verse a lo largo de todo este trabajo, se llega por lo general a conclusiones similares, aplicando diversos métodos para probar transformaciones. Estos métodos tienen sus ventajas y desventajas, y en algunos aspectos unos mejores que otros. Por ejemplo una gran desventaja del método para muestras grandes, es que los límites de confianza y las pruebas basados en este método, tienen solo validéz asintótica (como el mismo método lo especifica) y el número de parámetros debe ser pequeño comparado con el de observaciones. Sin embargo las otras pruebas teóricamente podrán ser mas robustas y mas potentes, pero no son aplicables ya sea por dificultad en su cálculo o por ser poco prácticas. Otro ejemplo es, el hecho que Andrews hace notar de que el método de máxima verosimilitud puede ser sensible a los valores erráticos; sin embargo, todos los métodos, razonablemente eficientes, dependen de una manera crítica, de las observaciones extremas.

En suma, la elección del método a usar dependerá, en gran parte, de las ponderaciones que se den a la normalidad y a la simplicidad en el modelo; de la "interpretabilidad" de una transformación particular; del tamaño de la muestra y, a través de ésta, la relevancia de las propiedades asintóticas de la función de verosimilitud, y finalmente de los posibles tratamientos de los valores erráticos o extremos y de otras desviaciones o fallas de normalidad.

BIBLIOGRAFIA

Andrews, D.F.- "A note on the selection of data transformations" Biometrika (1971), 58, 2, p.249

Ansley, C.F., Spivey, W.A., Wrobleski, W.J.- "An analysis of - transformations for time series".-Working Paper No. 117 Univ. of Michigan (1975)

Atkinson, A.C.- "Testing transformations to normality".- Journal of the Royal Stat. Soc., B, 31 (1969) p.472

Box, G.E.P, Cox, D.R.- "An analysis of transformations".- Journal of the Roy. Stat. Soc., Series B, 26 (1964) p. 211

Box, G.E.P., Tiao, G.C.- "Bayesian Inference in Statistical Analisis".- Addison Wesley (1973) Capítulo 10, p.522

Box, G.E.P., Tidwell, P.W.- "Transformacion of the independent variables".- Technometrics, Vol 4, No. 4 (1962), p.531

Draper, N.R., Cox, D.R.- "On distributions and their transformations to normality".- Jour. of the Roy. Stat. Soc. Series B, 31 (1969) p. 472

Hinkley, D.V.- "On power transformations to symmetry".- Biometrika (1975), 62,1, pag 101

Hinkley, D.V.- "On quick choice of power transformations".- Appl. Statist. (1977), 26, No.1 p. 67

Schlesselman, J.- "Power families: A note on the Box and Cos transformation".- Journal of the Roy. Stat. Soc. Series B, 33 (1970) p. 307

Tintner, G., Kadekodi, G.- "Note on the transformation of variables in simultaneous equations systems".- Univ. of Southern California, pag 163.

Wood, J.T.- "An extension of the analysis of transformations of Box and Cox".- Appl. Statist, (1974) 23, No 3, p.278

Zarembka, P.- "Transformation of variables in econometrics".- Univ. of New York at Bufalo. (1972) p. 81

Zellner.- "An introduction to Bayesian inference in Econometrics" ,capítulo 6, sección 6.1, p.162