Transformaciones de Box y Cox

Box y Cox (1964) propusieron una familia de funciones de potencia para la variable de respuesta con el objetivo de garantizar el cumplimiento de todos los supuestos de un modelo lineal, es decir:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

Estas transformaciones combinan el objetivo de encontrar una relación simple, con homogeneidad de varianzas, mejorando la normalidad.

Las transformaciones originales de Box y Cox están dadas por:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log y & \text{si } \lambda = 0 \end{cases}$$

Mediante la regla de L' Hôpital podemos probar que

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log y$$

En ese mismo trabajo estos autores proponen la familia

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{si } \lambda_1 \neq 0 \\ \log(y+\lambda_2) & \text{si } \lambda_1 = 0 \end{cases}$$

para contemplar el caso de valores de y negativos. En la práctica se elige λ_2 para que $y_i + \lambda_2 > 0$ para todo i. De manera que sólo veremos a λ_1 como parámetro de estas transformaciones.

Esta familia es continua en λ y monótona creciente para cada λ, es decir que el orden original entre las y's es preservado: si $y_1 > y_2$, luego $y_1^{(\lambda)}) > y_2^{(\lambda)}$.

Es claro que no toda distribución puede ser transformada a una normal. Draper y Smith (1969)estudiaron este problema y concluyeron que aún en aquellas distribuciones para las que transformando por potencias no es posible lograr exacta normalidad, los estimadores usuales de λ conducen a distribuciones cuyos primeros 4 momentos corresponderían a simetría.

John y Draper (1980) propusieron la siguiente modificación:

$$y^{(\lambda)} = \begin{cases} sg(y)\ \frac{(|y|+1)^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ sg(y)\ \log(|y|+1) & \text{si } \lambda = 0 \end{cases}$$

que puede funcionar mejor para distribuciones simétricas.

Supongamos que las observaciones transformadas $\mathbf{Y}^{(\lambda)} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Nosotros observamos la matriz de diseño $\mathbf{X}$, el vector de respuestas $\mathbf{Y}$, de manera que los parámetros del modelo son $(\lambda, \boldsymbol{\beta}, \sigma^2)$. Box y Cox (1964) mostraron que λ puede ser estimado por el método de máxima verosimilitud. Sin embargo, si planteáramos las tres ecuaciones de scores, resolverlas simultáneamente podría ser complicado. Por este motivo, se suele resolver la búsqueda de los estimadores de $(\boldsymbol{\beta}, \sigma^2)$ para cada λ fijo y luego se elige el λ más adecuado.

En este caso tendríamos que la densidad de $\mathbf{Y}^{(\lambda)}$ es

$$f(\mathbf{y}^{(\lambda)}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\ e^{-\frac{\left(\mathbf{y}^{(\lambda)}-\mathbf{x}\boldsymbol{\beta}\right)'\left(\mathbf{y}^{(\lambda)}-\mathbf{x}\boldsymbol{\beta}\right)}{2\sigma^2}}$$

donde

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \log y & \text{si } \lambda = 0 \end{cases}$$

¿Cuál sería en este caso $f(\mathbf{y})$? Tendríamos

$$f(\mathbf{y}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \, e^{-\frac{\left(\mathbf{y}^{(\lambda)}-\mathbf{x}\boldsymbol{\beta}\right)'\left(\mathbf{y}^{(\lambda)}-\mathbf{x}\boldsymbol{\beta}\right)}{2\sigma^2}} J(\lambda, \mathbf{y})$$

donde $J(\lambda, \mathbf{y})$ es el jacobiano de la transformación de y a $y^{(\lambda)}$. Por lo tanto:

$$J(\lambda, \mathbf{y}) = \Pi_{i=1}^{n} \frac{\partial y_i^{(\lambda)}}{\partial y_i} = \Pi_{i=1}^{n} y_i^{\lambda-1}$$

Con lo cual, la función de verosimilitud, que coincidiría con $f(\mathbf{y})$, resultaría:

$$f(\mathbf{y}, \lambda, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \, e^{-\frac{\left(\mathbf{y}^{(\lambda)} - \mathbf{x}\boldsymbol{\beta}\right)'\left(\mathbf{y}^{(\lambda)} - \mathbf{x}\boldsymbol{\beta}\right)}{2\sigma^2}} \, \Pi_{i=1}^{n} y_i^{\lambda-1}$$

Para cada λ fijo los estimadores de máxima verosimilitud de $\boldsymbol{\beta}$ y de σ^2 son:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(\lambda) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}^{(\lambda)} \\ \widehat{\sigma^2}(\lambda) &= \mathbf{Y}^{(\lambda)}(\mathbf{I} - \mathbf{P})\mathbf{Y}^{(\lambda)}/n \end{aligned}$$

Si consideramos la log–versosimilitud y reemplazamos por dichos valores resulta:

$$\begin{aligned} \log f(\mathbf{y}, \lambda, \boldsymbol{\beta}, \sigma^2) &= cte - \frac{n}{2} \log \widehat{\sigma^2}(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log y_i \\ &= cte - \frac{n}{2} \log S^2(\lambda) \end{aligned}$$

$S^2(\lambda)$: ¿Por qué puede ser visto como un estimador de la escala: ?

Sea g la media geométrica de las observaciones y_i: $g = (\Pi_{i=1}^{n} y_i)^{1/n}$ y definimos

$$y(\lambda, g) = y^{(\lambda)}/g^{\lambda-1}$$

Si hiciéramos la regresión de $y(\lambda, g) \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, ¿Cuánto daría S^2_λ ?

Veremos que es la cantidad que

$$-\frac{n}{2}\log S^2_\lambda = \frac{n}{2}\log \widehat{\sigma}^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n} \log y_i$$

Por lo tanto, el estimador de λ se obtendrá maximizando

$$-\frac{n}{2}\log S^2(\lambda)$$

A partir de la teoría que conocemos de cociente de verosimilitud, podemos ver que si nos interesa testear la hipótesis

$$H_0 : \lambda = \lambda_0$$

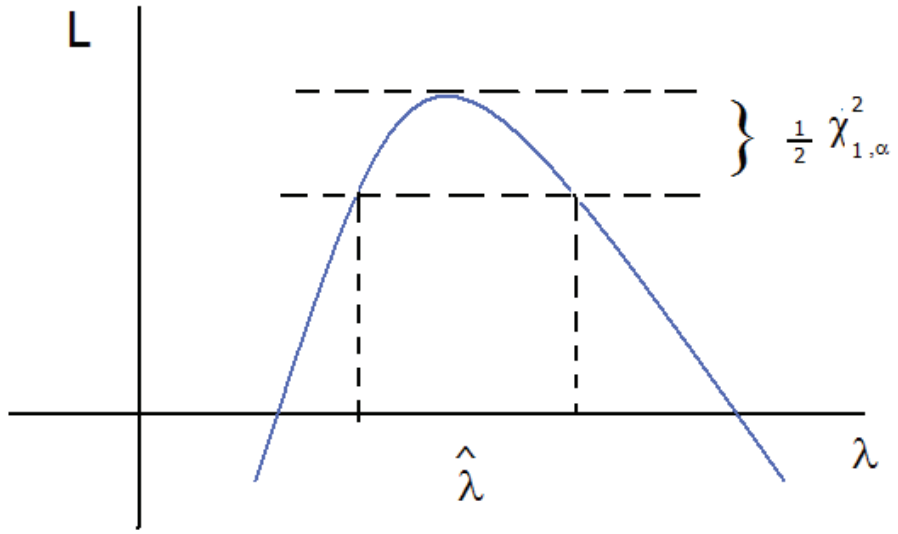
el estadístico:

$$W = 2(-\frac{n}{2}\log S^2(\widehat{\lambda}) + \frac{n}{2}\log S^2(\lambda_0))$$

tiene distribución asintótica χ^2_1. En consecuencia:

$$P(-\frac{n}{2}\log S^2(\widehat{\lambda}) + \frac{n}{2}\log S^2(\lambda_0) \leq \frac{1}{2}\chi^2_{1,\alpha}) \cong 1 - \alpha$$

y podemos deducir un intervalo de confianza para λ.


L
$\frac{1}{2}\chi^2_{1,\alpha}$
$\hat{\lambda}$
λ

Ejemplo (Draper y Smith, 1981)

Los siguientes datos corresponden a un estudio más extenso presentado por Draper y Smith (1981) en el que se quiere estudiar la viscosidad en función de dos componentes $FF =$ filler y $PP =$ Oil (aceite).

```
FF PP WW
0  0  26
12  0  38
24  0  50
26  0  76
48  0  108
60  0  157
0  10  17
12  10  26
24  10  37
36  10  53
48  10  83
60  10  124
0  20  13
12  20  20
24  20  27
36  20  37
48  20  57
60  20  87
12  30  15
24  30  22
36  30  27
48  30  41
```

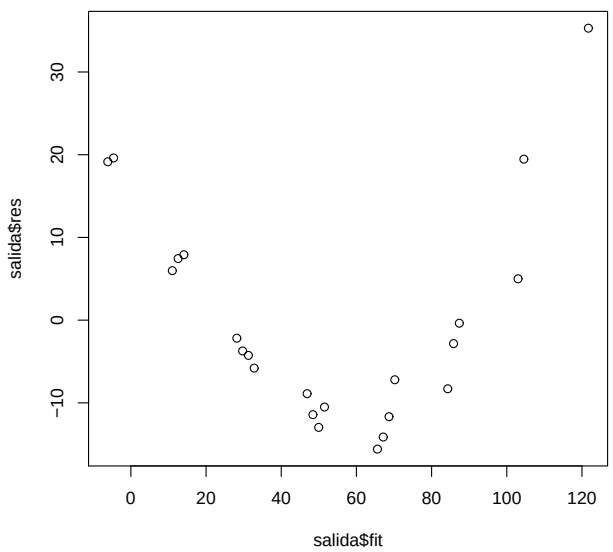
```
60 30 63
```

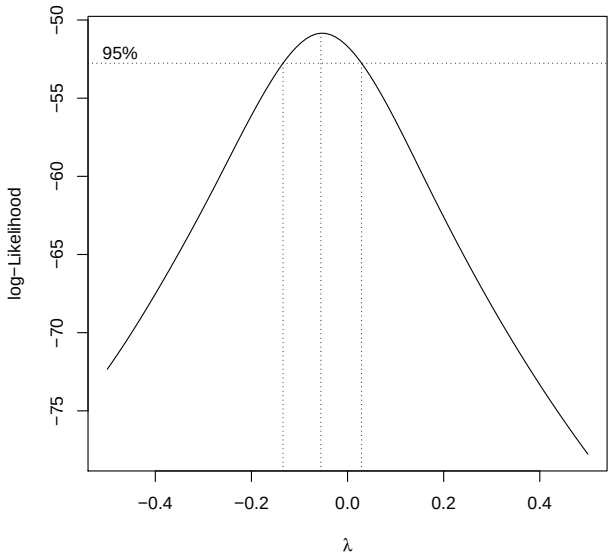
El modelo propuesto es:

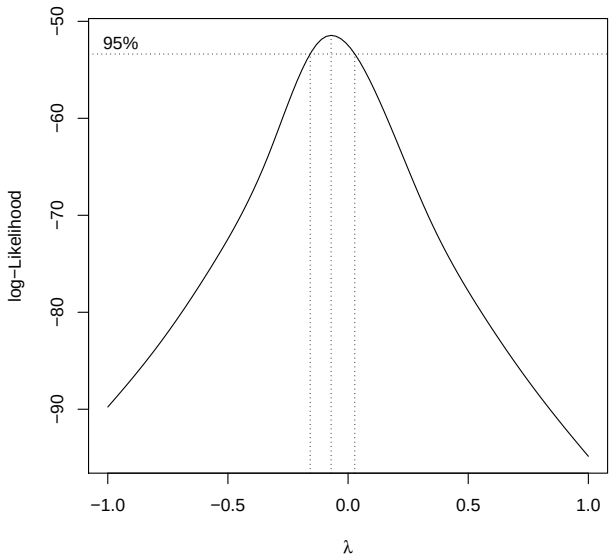
$$WW = \beta_0 + \beta_1 FF + \beta_2 PP + \epsilon$$

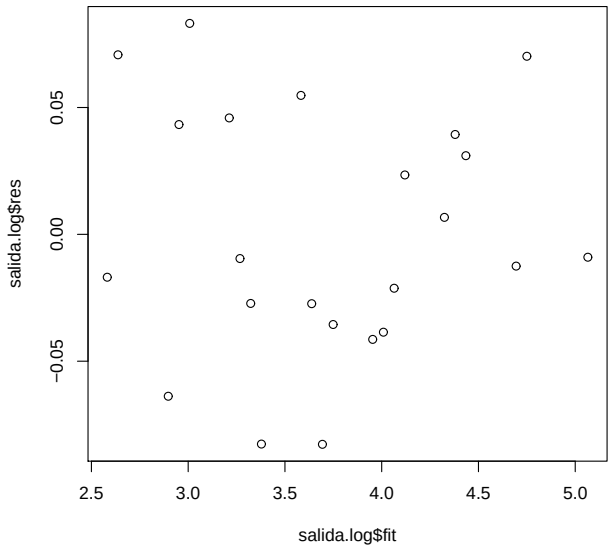
```
Call:
lm(formula = WW ~ FF + PP)

Residuals:
    Min      1Q  Median      3Q     Max
-15.592  -9.695  -3.722   6.713  35.296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1837     6.3322   4.451 0.000245 ***
FF            1.5587     0.1452  10.735 9.48e-10 ***
PP           -1.7166     0.2640  -6.502 2.44e-06 ***

Residual standard error: 13.82 on 20 degrees of freedom
```

```
Multiple R-squared: 0.8793,     Adjusted R-squared: 0.8673
F-statistic: 72.87 on 2 and 20 DF,  p-value: 6.543e-10

library(MASS)

boxcox(WW~FF+PP, data = viscosity,lambda = seq(-1, 1, length = 10))

boxcox(WW~FF+PP, data = viscosity,lambda = seq(-1, 1, length = 10))

salida.log<- lm(logww~FF+PP)
```


salida$res
salida$fit
30
20
10
0
−10
0
20
40
60
80
100
120


95%
log−Likelihood
λ
−50
−55
−60
−65
−70
−75
−0.4
−0.2
0.0
0.2
0.4


95%
log−Likelihood
λ
−50
−60
−70
−80
−90
−1.0
−0.5
0.0
0.5
1.0


salida.log$res
salida.log$fit
0.05
0.00
−0.05
2.5
3.0
3.5
4.0
4.5
5.0

Errores Correlacionados

Consideremos el caso particular en que los errores siguen el siguiente un modelo autorregresivo de orden 1, AR(1), es decir:

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \ ,$$

donde u_t son i.i.d, $E(u_t) = 0$ y $Var(u_t) = \sigma_u^2$. Asumimos que $|\rho| < 1$. Ya hemos probado que

$$\begin{aligned} E(\epsilon_t) &= 0 \\ Var(\epsilon_t) &= \frac{\sigma_u^2}{1-\rho^2} \\ Cov(\epsilon_t, \epsilon_{t-r} &= \rho^r \frac{\sigma_u^2}{1-\rho^2} \end{aligned}$$

Removiendo la autocorrelación mediante una transformación

Supongamos que

$$\begin{aligned} y_t &= \alpha + \beta x_t + \epsilon_t \\ \epsilon_t &= \rho \epsilon_{t-1} + u_t \end{aligned}$$

donde u_t son i.i.d. $u_t \sim N(0, \sigma_u^2)$. Notemos que:

$$\begin{aligned} y_t &= \alpha + \beta x_t + \epsilon_t \\ y_{t-1} &= \alpha + \beta x_{t-1} + \epsilon_{t-1} \end{aligned}$$

por lo tanto:

$$y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(x_t - \rho x_{t-1}) + \epsilon_t - \rho \epsilon_{t-1}$$

con lo cual

$$y_t^* = \alpha^* + \beta^* x_t^* + u_t$$

es decir las nuevas variables satisfacen las condiciones habituales del modelo lineal.

¿Cómo estimar a ρ?

El método iterativo de Cochrane–Orcutt propone los siguientes pasos para la estimación en esta situación.

1. Computar los estimadores de mínimos cuadrados ordinarios de α y β.
2. Calcular los residuos e_t y estimar a ρ mediante
$$\widehat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$$
3. Ajustar el modelo (*) usando $\widehat{\rho}$.
4. Examinar los nuevos residuos. Si no están correlacionados terminar computando los estimadores de interés:
$$\widehat{\alpha} = \widehat{\alpha}^*/1 - \widehat{\rho} \quad \widehat{\beta} = \widehat{\beta}^*$$
De lo contrario, repetir el procedimiento usando como estimadores iniciales $\widehat{\alpha}$ y $\widehat{\beta}$.

Método de Prais–Winstein (1954)

Otra posibilidad es el método de Prais–Winstein basado en mínimos cuadrados generalizados. En función de las expresiones vistas para la varianzas y las correlaciones de los errores, tenemos que $\Sigma_e psilon = \sigma^2 \Omega$, donde

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \dots & & & & \\ \dots & & & & \\ \rho^{n-1} & \rho^{n-2} & & \dots & 1 \end{pmatrix}$$

Utilizando el estimador del paso [2.] anterior, $\widehat{\rho}$, podríamos estimar a Ω por $\widehat{\Omega}$

$$\widehat{\Omega} = \begin{pmatrix} 1 & \widehat{\rho} & \widehat{\rho}^2 & \dots & \widehat{\rho}^{n-1} \\ \widehat{\rho} & 1 & \widehat{\rho} & \dots & \widehat{\rho}^{n-2} \\ \dots & & & & \\ \dots & & & & \\ \widehat{\rho}^{n-1} & \widehat{\rho}^{n-2} & & \dots & 1 \end{pmatrix}$$

para luego computar el estimador de mínimos cuadrados generalizados:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{Y}$$

.

Detección de Puntos Influyentes

Residuos

En general, los puntos con residuos standarizados r_i que van más allá del rango [-2, 2] (o [-2.5,2.5], según los autores) se consideran sospechosos.

Leverage

El leverage mide cuán extrema es una observación en el espacio de las covariables $\mathbf{x}'s$.

Se llama **leverage** de una observación a

$$p_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

En la práctica probarán propiedades de p_{ii} que son útiles para interpretar qué miden.

De hecho, si $\mathbf{X} \in \Re^{n\times p}$ contiene una columna de $1's$, sin pérdida de generalidad asumamos que la primera, $\mathbf{X} = [\mathbf{I}, \mathbf{X}_2]$ y la matriz de proyección

$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ satisface:

a) $\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2$ donde $\mathbf{P}_1 = n^{-1}\mathbf{1}\,\mathbf{1}'$ $(\mathbf{1} \in \Re^n, \mathbf{1} = (1, 1, .., 1)')$ y $\mathbf{P}_2 = \widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'$ siendo $\widetilde{X} = (\mathbf{I} - n^{-1}\mathbf{1}\,\mathbf{1}')\mathbf{X}_2$ la matriz con las columnas centradas.

b) $p_{ii} \geq \frac{1}{n}$.

c) $p_{ii} = \frac{1}{n} + \widetilde{p}_{ii}$ donde $\widetilde{p}_{ii} = (P_2)_{ii}$.

Con lo cual, p_{ii} mide la distancia de $\mathbf{x}_i$ a su centro $\overline{\mathbf{x}}$.

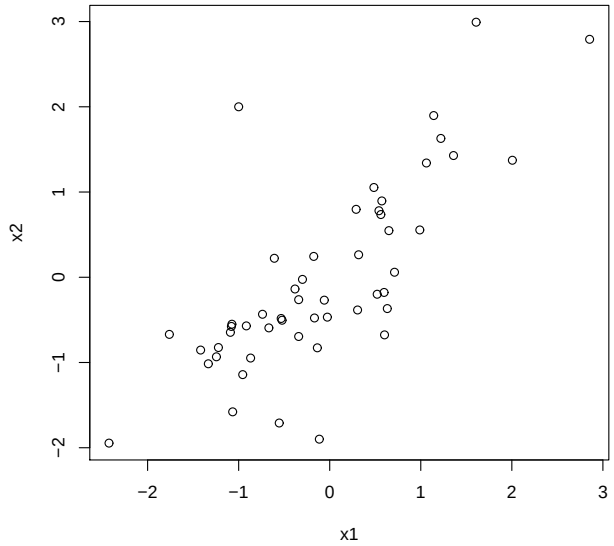
Sabemos que

$$\sum_{i=1}^{n} p_{ii} = p \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} p_{ii} = \frac{p}{n}$$

y por esta razón se sugiere considerar como punto de corte $2\frac{p}{n}$ (algunos autores sugieren $3\frac{p}{n}$)

Por lo tanto, se estudiarán especialmente aquellos puntos tales que $p_{ii} > 2\frac{p}{n}$.

Además se sugiere considerar los siguientes gráficos:



- i vs. p_{ii}
- tallo y hoja (o histograma) de p_{ii}
- boxplots de p_{ii}

Distancia de Cook

Las conclusiones de los métodos de diagnóstico podrían depender de la presencia de puntos influyentes.

Al excluir un punto influyente del análisis, las conclusiones a partir del conjunto restante podrían cambiar considerablemente.

En principio, desearíamos que pequeñas perturbaciones introdujeran pequeños cambios.

Supongamos que $\widehat{\boldsymbol{\beta}}$ es el estimador de mínimos cuadrados obtenidos a partir de toda la muestra $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, mientras que $\widehat{\boldsymbol{\beta}}_{(i)}$ es el estimador de mínimos cuadrados obtenido al excluir la i–ésima observación, $(\mathbf{x}_i, y_i)$, de la muestra.

Se define la Curva de Influencia Muestral (SIC) como:

$$SIC = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{1/n} = n(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})$$

Como SIC es un vector, podríamos considerar su norma o su norma respecto de una matriz simétrica definida positiva $\mathbf{M}$ y eventualmente un factor de escala:

$$\begin{aligned} D_i(M,c) &= \frac{n^{-2} SIC' \, \mathbf{M} \, SIC}{c} \\ &= \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})' \mathbf{M} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{c} \end{aligned}$$

Si eligiéramos $\mathbf{M} = \mathbf{X}'\mathbf{X}$ y $c = p\widehat{\sigma}^2 = ps^2$ obtendríamos algo conocido:

$$\frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})' (\mathbf{X}'\mathbf{X}) (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{p\widehat{\sigma}^2}$$

De hecho el elipsoide de confianza lo obtenemos como:

$$\frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})' (\mathbf{X}'\mathbf{X}) (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{p\widehat{\sigma}^2} \leq F_{p,n-p,1-\alpha}$$

La distancia de Cook (1977) es:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{ps^2}$$

Notemos que

$$D_i = \frac{(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})'(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})}{ps^2}$$

$$= \frac{\sum_{j=1}^{n} (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{ps^2}$$

donde $\widehat{\mathbf{Y}}_{(i)}$ denota al vector de valores predichos obtenido a partir de $\widehat{\boldsymbol{\beta}}_{(i)}$.

En la práctica se mostrará que

$$D_i = \frac{1}{p} \frac{p_{ii}}{1 - p_{ii}} r_i^2$$

donde p_{ii} es el elemento i de la diagonal de la matriz de proyección $\mathbf{P}$ y r_i es el $i-$ésimo residuo standarizado. En esta expresión se ve que esta distancia

conjuga tanto el efecto sobre los residuos como el leverage de las observaciones, por lo tanto D_i implica residuo o leverage grandes.

Se suele comparar a D_i con la distribución $F_{p,n-p}$ y se presta especial atención a aquellos puntos que están por encima del percentil 50 %.

Otras medidas

DFFIT

Una medida bastante natural y cercana a la distancia de Cook es la del cambio en la predicción al eliminar la observación i.

Recordemos que

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\, \mathbf{x}_i \, \frac{e_i}{1-p_{ii}}$$

$$S^2_{(i)} = \frac{(n-p)s^2 - e_i^2(1-p_{ii})}{n-p-1}$$

Por lo tanto el cambio en la predicción resulta:

$$\begin{aligned} DFFIT_i &= \widehat{Y}_i - \widehat{Y}_{i(i)} = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{(i)} \\ &= \frac{p_{ii}}{1-p_{ii}}\, e_i \end{aligned}$$

Como $\Sigma_{\widehat{\mathbf{Y}}} = \sigma^2\mathbf{P}$, una versión standarizada es:

$$DFFIT_i = \frac{\sqrt{p_{ii}}}{S_{(i)}(1-p_{ii})}\, e_i$$

Usando las cotas vistas para los residuos y los leverage, se sugiere como puntos de corte $|DFFIT| > 2\sqrt{\frac{p}{n-p}}$ o si n es mucho mayor que p $|DFFIT| > 2\sqrt{\frac{p}{n}}$.

$DFBETA_j$

Esta medida considera el cambio en cada coordenada de $\widehat{\boldsymbol{\beta}}$ al eliminar la observación i.

Como vimos

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{x}_i \frac{e_i}{1-p_{ii}}$$

Llamemos

$$(a_{0i}, \ldots, a_{p-1i}) = (\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{x}_i$$

entonces para $i = 1, .., n$ y $j = 0, .., p-1$

$$DFBETA_j = \widehat{\beta}_j - \widehat{\beta}_{j(i)} = \frac{a_{ji} e_i}{1-p_{ii}}$$

Ver archivo Complemento

Colinealidad

la calidad de los estimadores, medida a través de su precisión, puede ser muy afectada si las covariables están muy relacionadas entre sí.

Esta situación típicamente puede deberse a:

- Las covariables cumplen una restricción (ejemplo % de cemento)
- Se crean variables a partir de otras existentes y se introduce dependencia
- En los sistemas bológicos o físicos o químicos las variables naturalmente pueden tener dependencia.
- Dependencia inadecuada por un muestreo inadecuado.

De todas formas, no siempre puede identificarse el origen de la colinealidad, aunque es importante detectarla y tratar de entender su naturaleza.

Sabemos caraterizar la singularidad: existe $\mathbf{c}$, $\|\mathbf{c}\| = 1$ tal que

$$\mathbf{Xc} = 0 \quad (\|\mathbf{Xc}\|^2 = 0)$$

Podríamos decir que la *casi–singularidad* corresponde a: existe $\mathbf{c}$, $\|\mathbf{c}\| = 1$ tal que

$$\|\mathbf{Xc}\|^2 = \delta <<$$

Veamos que efecto tiene esta *casi–singularidad*. Por Cauchy–Schwartz tenemos que

$$1 = \mathbf{c}'\mathbf{c} = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{1/2}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{c} \leq \sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})\mathbf{c}}\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}} = \sqrt{\delta}\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}$$

Por lo tanto:

$$1 \leq \delta\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$$

En consecuencia:

$$Var(\mathbf{c}'\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} \geq \sigma^2/\delta >>$$

Como $\mathbf{Xc}$ puede ser afectado por las unidades de $\mathbf{X}$ vamos a escalar las colum-

nas de $\mathbf{X}$ de manera que tengan norma 1:

$$\mathbf{X} = [\mathbf{x}^{[1]} \dots \mathbf{x}^{[p]}] \longrightarrow \mathbf{X}_s = [\mathbf{x}^{[1]}/\|\mathbf{x}^{[1]}\| \dots \mathbf{x}^{[p]}/\|\mathbf{x}^{[p]}\|]$$

Notemos que si $\mathbf{D}^{-1} = diag(\|\mathbf{x}^{[1]}\|, \dots, \|\mathbf{x}^{[p]}\|)$, entonces

$$\mathbf{X}_s = \mathbf{X}\mathbf{D}^{-1}$$

y por lo tanto:

$$(\mathbf{X}_s'\mathbf{X}_s)^{-1} = \mathbf{D}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}$$

En este sentido podríamos considerar el modelo equivalente

$$\mathbf{Y} = \mathbf{X}_s\boldsymbol{\beta}_s + \epsilon$$

donde $\boldsymbol{\beta}_s = \mathbf{D}\boldsymbol{\beta}$.

Tenemos que:

$$\widehat{\boldsymbol{\beta}}_s = \mathbf{D}\widehat{\boldsymbol{\beta}} \quad \text{y} \quad \Sigma_{\widehat{\boldsymbol{\beta}}_s} = \mathbf{D}\Sigma_{\widehat{\boldsymbol{\beta}}}\mathbf{D}$$

Una consecuencia de escalar es que se remueve la casi–singularidad debida a que una columna de $\mathbf{X}$ tiene longitud pequeña.

Para $\mathbf{d} = \mathbf{D}\mathbf{c}$:

$$\mathbf{c}'(\mathbf{X}'\mathbf{X})\mathbf{c} = \mathbf{c}'\mathbf{D}\mathbf{D}^{-1}(\mathbf{X}'\mathbf{X})\mathbf{D}^{-1}\mathbf{D}\mathbf{c} = \mathbf{d}'(\mathbf{X}_s'\mathbf{X}_s)\mathbf{d} \geq \lambda_{min}\|\mathbf{d}\|^2$$

siendo d_{min} es el mínimo autovalor de $(\mathbf{X}_s'\mathbf{X}_s)$

Luego, si hay multicolinealidad $\mathbf{c}'(\mathbf{X}'\mathbf{X})\mathbf{c}$ puede ser pequeño (aún con $\|\mathbf{d}\|^2$ no tan pequeño) y por lo tanto λ_{min} será pequeño.

Detección de Colinealidad

Autovalores y Número de Condición

Como hemos visto los autovalores pueden darnos indicios de colinealidad.

Sean $\lambda_1, \ldots, \lambda_p$ los autovalores de $(\mathbf{X}_s'\mathbf{X}_s)$ y lllamemos

$$\lambda_{max} = \text{máx}\, \lambda_i \qquad \lambda_{min} = \text{mín}\, \lambda_i$$

Definimos:

$$\text{índice de condición :} \quad \delta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$$

Un número de condición grande indica una matriz pobremente condicionada.

Belsey, Kuh y Welsch (1980) sugieren que índices $\delta_j > 30$ o 100 indicarían colinealidad de moderada o severa

Factor de Inflación de la Varianza

Podemos medir la relación entre una variable x_j y las demás mediante el coeficiente de correlación múltiple R_j^2.

Se define el Factor de Inflación de la Varianza como

$$VIF_j = \frac{1}{1 - R_j^2}$$

Si $R_j^2 \simeq 1$ entonces $VIF_j >>$ y si x_j es ortogonal a todas las demás $VIF_j = 1$.

Se puede demostrar que si $\mathcal{R}$ es la matriz de correlación de las x_j entonces:

$$(\mathcal{R}^{-1})_{jj} = VIF_j$$

Theil (1971) y Berek (1977) probaron que

$$V(\widehat{\boldsymbol{\beta}}_j) = \frac{\sigma^2}{\tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j} VIF_j$$

donde $\tilde{\mathbf{x}}_j$ es la $j-$ésima columna centrada y escalada.

Se suele tomar como punto de corte $VIF_j > 10$ como indicador de colinealidad.

Ver archivo Complemento