Stata Intermedio

Sesión 1.1

Juan Carlos Abanto Orihuela
j.abanto@giddea.com

Grupo IDDEA Consulting

Abril - 2018

Parte I

Valores Atípicos

Agenda

- Tratamiento de Valores Atípicos

Agenda

- Tratamiento de Valores Atípicos
- Puntos Leverage

- Tratamiento de Valores Atípicos
- Puntos Leverage
- Puntos Outliers

- Tratamiento de Valores Atípicos
- Puntos Leverage
- Puntos Outliers
- Puntos de Influencia

- Algunas observaciones podrian tener influencia inusual en determinados parámetros estimados y predicciones del modelo. Estas observaciones de influencia podrían ser detectadas usando una de varias medidas cuando observamos un amplio residuo.

- Algunas observaciones podrian tener influencia inusual en determinados parámetros estimados y predicciones del modelo. Estas observaciones de influencia podrían ser detectadas usando una de varias medidas cuando observamos un amplio residuo.
- La medida de leverage de la i-esima observación, denotado por h_i es igual al i-esimo elemento de la diagonal de la matriz $H = X(X'X)^{-1}X'$. Si h_i es amplio, entonces y_i tiene una gran influencia sobre las predicciones del MCO $\hat{y}_i$ porque $\hat{y} = Hy$.

- Algunas observaciones podrian tener influencia inusual en determinados parámetros estimados y predicciones del modelo. Estas observaciones de influencia podrían ser detectadas usando una de varias medidas cuando observamos un amplio residuo.
- La medida de leverage de la i-esima observación, denotado por h_i es igual al i-esimo elemento de la diagonal de la matriz $H = X(X'X)^{-1}X'$. Si h_i es amplio, entonces y_i tiene una gran influencia sobre las predicciones del MCO $\hat{y}_i$ porque $\hat{y} = Hy$.
- El punto umbral que marca a los puntos leverage es $2(k/n)$

- En algunos casos, la existencia de amplios residuos tendrán un efecto determinante sobre el cálculo de la desviación estandar del parámetro por lo que finalmente afectaria a la inferencia del parámetro.

- En algunos casos, la existencia de amplios residuos tendrán un efecto determinante sobre el cálculo de la desviación estandar del parámetro por lo que finalmente afectaria a la inferencia del parámetro.
- Una forma de detectar a los outliers es mediante el cálculo de los residuos studentizados, $r_i = \frac{\epsilon_i}{\sigma\sqrt{1-h_i}}$

- En algunos casos, la existencia de amplios residuos tendrán un efecto determinante sobre el cálculo de la desviación estandar del parámetro por lo que finalmente afectaria a la inferencia del parámetro.
- Una forma de detectar a los outliers es mediante el cálculo de los residuos studentizados, $r_i = \frac{\epsilon_i}{\sigma\sqrt{1-h_i}}$
- Aquellos valores que resulten mayores a 2 implicarán la existencia de outliers.

Cook's D

- Resulta deseable considerar tanto la ubicación del punto en el espacio X y la respuesta de la variable como medida de influencia.
- Cook (1977, 1979) sugiere usar una medida de la distancia cuadrática entre la estimaciòn mínimo cuadrática basada sobre una estimación de n-puntos y las estimación de borrar el i-esimo punto.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{kMSE}$$

$$D_i = \frac{e_i^2}{kMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{(1+k)s^2}$$

- Los puntos de influencia son tomados para aquellas observaciones donde $D_i > 4/n$.

DFITS

- Welsch y Kuh (1977) introducen dos medidas de detección de los puntos de influencia

$$DFITS_i = r_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

- Donde r_i es el residuo Studentizado.
- Se sugiere que los valores de los DFITS que superen los valores de $2\sqrt{k/n}$ serán clasificados como puntos de influencia.

Distancia Welsch

- Otra medida útil para detectar los puntos de influencia es basarse en las medidas de distancia de Welsch, cuyo calculo puede ser planteado de la siguiente manera.

$$W_i = DFITS_i \sqrt{\frac{n-1}{1-h_{ii}}}$$

- Se sugiere que los valores de los Welsch que superen los valores de $3\sqrt{k}$ serán clasificados como puntos de influencia.

COVRATIO

- Belsley, Kuh y Welsch (1980) midieron los puntos de influencia de la i-esima observación considerandoel efecto de la matriz de varianza y covarianza de las estimaciones. La medida es el ratio de los determinantes de la matriz de covarianzas con y sin la i-esima observación.

$$COVRATIO_i = \frac{1}{1-h_{ii}}\left(\frac{n-k-\hat{\epsilon}_i^2}{n-k-1}\right)^k$$

- Donde $\hat{\epsilon}_i$ es el residuo estandarizado.
- Para las observaciones que no son de influencia, el valor del COVRATIO es proximo a 1. Mayores valores del residuo o de los valores del leverage causaran desviaciones por lo que se sugiere que el punto de corte sea considerado a aquellos valores en los que $|covratio_i - 1| \geq 3k/n$

DFBETA

- En estas medidas podemos identificar el parámetro asociado al punto de influencia.

$$DFBETAS_{j,i} = \frac{(\hat{\beta}_j - \hat{\beta}_{j(i)})}{s^2_{(i)} C_{jj}}$$

- Donde C_{jj} es el j-esimo elemento de la diagonal de $(X'X)^{-1}$.
- Para el punto de corte se considera aquellos valores por encima de $2/\sqrt{(n)}$ como influenciados.
- Una alternativa al calculo del DFBETA podría ser:

$$DFBETA_j = \frac{r_j \mu_j}{\sqrt{U^2(1 - h_{jj})}}$$

- Donde μ_j es el residuo obtenido de la regresión de x_i sobre los restantes X, y $U^2 = \sum_j \mu_j^2$.