

## Auxiliar 3: STATA - Semestre Otoño 2015

31 de Marzo, 2015

Canción de hoy: Matador, de Los Fabulosos Cadillacs

Esta auxiliar será una guía individual para aprender a usar STATA de una forma práctica. Para esto, se usará la base de datos **campus.dta** que se puede encontrar en Material Docente de U-Cursos. Esta base de datos posee registros del número de crímenes en los Campus de Estados Unidos.

### Exploración de datos y regresión

1. Abra la base de datos desde su ubicación. Pestaña *File > Open > campus.dta*.
2. Para observar los datos, en la pestaña superior apretar el símbolo . Para visualizar un detalle de cada variable, utilice el comando *codebook (nombre\_variable)*. Para ver una tabla con un resumen de ellas, utilice *summarize (nombre\_variable)*. Para ver una tabla de frecuencias de cada variable, utilice *tab nombre\_variable*. **¿Qué diferencias entrega cada forma de ver los datos? ¿Cuántas observaciones hay y cuántos missing values? ¿Qué diferencia tiene la variable *priv* y qué representa?**
3. Para visualizar el comportamiento de algunas variables, se puede realizar un gráfico Q-Q de normalidad. Para graficar en Q-Q la variable de número de crímenes utilice el comando *qnorm crime*, y para la variable de número de policías, *qnorm police*. **¿Cómo se puede interpretar el gráfico? Ahora compare el mismo gráfico con las variables logarítmicas *lcrime* y *lpolice*. ¿Qué se puede decir?** Para realizar tests sobre normalidad, 2 alternativas son el test de normalidad de Shapiro-Wilk, y el test de Skewness y Kurtosis. Para realizar estos tests sobre todas las variables utilice los comandos *sktest police lpolice crime lcrime* y *swilk police lpolice crime lcrime*. **¿Qué variables se puede afirmar que se comportan de forma normal?** (Queda propuesta la interpretación de los tests a través de los p-valores).
4. Para visualizar previamente algunas relaciones entre las variables, se pueden analizar sus correlaciones de forma gráfica o analítica. Para graficar dos variables con un gráfico de puntos (scatter) utilice *twoway (scatter crime police)*. En vez de utilizar el comando *scatter* utilice *line* para graficar con una línea, **¿Cuál es mejor para visualizar los datos?** Grafique ahora las variables *lcrime* y *lpolice*, **¿Qué puede decir?**
5. Ahora, para ver las correlaciones entre todas las variables, utilice *correlate crime lcrime police lpolice priv*. **¿Qué le dicen esos números? ¿Se condice con lo graficado anteriormente?**<sup>1</sup>

---

<sup>1</sup>Para ver alternativas y complementos al gráfico, utilice *help twoway*. O bien realice el gráfico desde la pestaña *Graphics > Twoway graphs (scatter, line, etc)*, desde allí se pueden variar los títulos, leyendas, escalas, etc.

6. Para hacer una regresión utilice el comando *regress*. La primera variable que se coloque después del comando será la variable dependiente, y luego se colocarán las variables independientes. Es decir, si se quiere regresar el número de crímenes sobre el número de policías, la matrícula y la condición de Campus privado, se utilizará el comando *regress crime police enroll priv*. **¿Qué variables son significativas, cómo se puede rechazar la hipótesis nula mirando el p-valor y los intervalos de confianza? ¿Cómo es la bondad de ajuste del modelo? ¿Qué se puede interpretar de los coeficientes?**
7. Posterior a la visualización de los coeficientes y de la regresión, se pueden realizar tests para obtener más información sobre los datos. Para obtener un test de medias, por ejemplo de las medias de policías diferenciando por campus privado, utilice el comando *sdtest police, by(priv)* esto hará el test con respecto a la variable *priv*. Para evaluar las medias de *crime* bajo el mismo test, utilice *sdtest crime, by(priv)*. **¿Qué puede interpretar?**

## Multicolinealidad

1. Para evaluar la multicolinealidad en los datos se han estudiado algunas maneras. Utilice los mismos comandos de regresión para evaluar el coeficiente de bondad de ajuste  $R^2_i$  de las variables independientes. **¿Hay evidencia de multicolinealidad entre las variables?**

Para evaluar el FIV, se generará un escalar que tome la información de la regresión, por ejemplo su bondad de ajuste. Luego de la regresión, el valor de  $R^2$  puede accederse a través del escalar  $e(r2)$ . Por lo tanto, para elaborar el FIV utilice el comando *scalar FIV = (1 / (1 - e(r2)))*. Luego para visualizar ese valor, utilice *display FIV*. **¿Qué valores de FIV son aceptables y cuáles preocupantes?**

Otra alternativa es evaluar según el tamaño de la matriz  $X'X$ . Tomando los valores propios de la matriz y seleccionando los valores máximos y mínimos entre éstos ( $\lambda_{min}, \lambda_{max}$ ) un indicador de multicolinealidad es propuesto como  $\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ . Para evaluar este indicador:

- I. Genere una matriz con los datos, la matriz  $X'X$ . Utilice el comando *mkmat* para convertir las variables en vectores: *mkmat crime police enroll*.
- II. Luego, junte los datos en una única matriz utilizando el comando *matrix X = (crime, police, enroll)*. Esto generará la matriz X, para verla utilice el comando *matrix list X*.
- III. Ahora, para obtener una matriz de multiplicación, utilice el comando *matrix XX = X'\* X*<sup>2</sup>. Con esta nueva matriz, se puede obtener la inversa mediante el comando *matrix INV = inv(XX)*. Es una alternativa ver en los valores de la matriz inversa cómo inflará la varianza del estimador.
- IV. Para obtener los valores propios de la matriz, utilice el comando *matrix eigenvalues \_reales\_ \_complejos\_ = XX* para guardar en la variable *\_reales\_* los valores propios reales (un vector de  $k$  valores) y en la variable *\_complejos\_* valores propios complejos. Para ver dichos vectores utilice nuevamente el comando *matrix list \_reales\_*.

---

<sup>2</sup>Notar que  $X'$  representa la traspuesta de X, y el símbolo \* es la multiplicación. Notar que XX es el nombre de la matriz creada

- v. Guarde en escalares dichos valores para ver el indicador. Por ejemplo, si el valor mayor  $\lambda_{max}$  está en la primera columna de la matriz en la primera fila, utilice el comando `scalar lambda_max = reales[1,1]`. Realice lo mismo con el  $\lambda_{min}$
  - vi. Ahora con dichos valores escalares, genere otro escalar para el valor de  $\gamma$ . **¿Qué valores de  $\gamma$  son aceptables y cuáles preocupantes?**
2. Realice el mismo procedimiento ( $R^2_i$ , FIV,  $\gamma$ ) con las variables logarítmicas (variable dependiente e independientes) y responda **¿Qué modelo posee más colinealidad?**.

## Teorema Frisch-Waugh-Lovell en STATA

1. Para la programación y elaboración de problemas más complejos en STATA, se utilizan los archivos `.do`. Para abrirlos, busque en la pestaña superior el icono , y comience con una planilla en blanco.
2. Para visualizar el teorema de Frisch-Waugh-Lovell en STATA, primero se tomará en cuenta el modelo:  $Y = X\beta_1 + Z\beta_2 + U$  como se vio en la auxiliar 2. En la base de datos esto corresponderá a  $lcrime = lpolice\beta_1 + lenroll\beta_2 + U$ . Entonces,
  - I. Realice la parcialización del efecto de Z en Y (quitar ese efecto), guardando los residuos de la regresión de *lcrime* sobre *lenroll*. Para hacer la regresión utilice `regress lcrime lenroll`.
  - II. Luego guarde los residuos en una variable cualquiera con el comando `predict u_yz, resid` donde *u\_yz* es el nuevo nombre de la variable guardada con los errores de la regresión realizada.
  - III. Ahora, guarde los residuos de la regresión de X sobre Z, es decir utilice el comando `regress lpolice lenroll` y guarde los residuos con `predict u_xz, resid`.
  - IV. Con ambas variables de error realice la última regresión con el comando `regress u_yz u_xz`. **¿Da el mismo beta que el modelo original? ¿Para cuando es útil utilizar este teorema?**

## Abreviaciones y comandos relevantes

1. STATA entiende abreviaciones cuando son únicas. Por ejemplo, los comandos *regress*, *display*, *matrix*, *summarize*, *correlate* se pueden abreviar en *reg*, *dis*, *mat*, *summ*, *corr*. De la misma forma, cuando existen variables con el mismo comienzo, por ejemplo *X\_1*, *X\_2*, *X\_3*, etc se pueden listar todas a través de un asterisco: *X\_\**.
2. Para entender algún comando, buscar la sintaxis, ejemplos u opciones internas, utilizar el comando *help* (*comando*) para abrir la ayuda de STATA.