Introducción a Stata

Ventajas de utilizar STATA

Stata es un paquete estadístico desarrollado y distribuido por Stata-Corporation. Es un paquete que cuenta con varias ventajas que podrían resumirse en lo siguiente:

- ✓ Contiene muchas de las técnicas estadísticas mas recientes.
- ✓ Se actualiza frecuentemente.
- ✓ Métodos gráficos muy poderosos.
- ✓ Buena interfase con procesadores de texto e impresoras.
- ✓ Requiere de poco espacio en el disco duro.
- ✓ Requiere de poca memoria. Precio accesible.
- ✓ Lenguaje de programación amigable y sencillo.

Aunque Stata también tiene ciertas desventajas : No puede leer directamente de manejadores de bases de datos, para esto necesita de un programa de interfase:

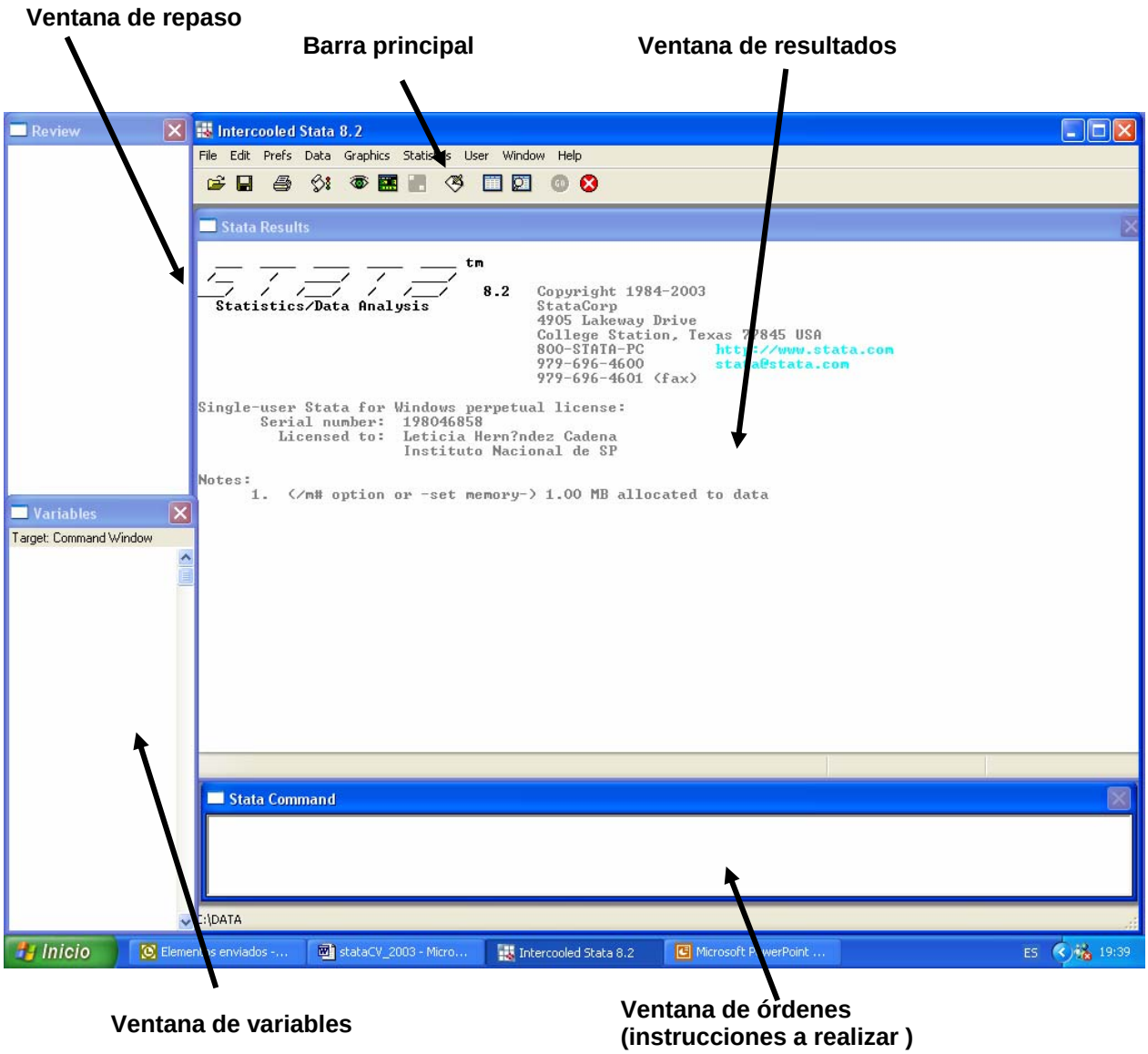
DBMSCOPY o
Stattransfer

El DBMSCOPY y el Stattransfer son programas de traducción de archivos, de todo a todo, DBASE a Foxpro SPSS a STATA, de STATA a SAS, de SAS a Epi Info, etc.

Para iniciar con Stata

Stata se ejecuta pulsando doblemente el icono de Stata en el menú de **Start** en Windows. Al ejecutar Stata aparecerá la siguiente pantalla:

Las barra de herramientas de Stata



Stata para Windows tiene trece botones. Si se le olvida lo que hace cualquier botón, ponga el puntero del ratón sobre el icono y en unos segundos aparecerá la descripción en inglés.

La lista de botones son los siguientes:

1) Open (Abrir)

Abre una base de datos de Stata.

2)Save (Guardar)

Guarda en el disco la base de datos actualmente en memoria.

3) Print Graph/Print Log (Imprimir gráficas/Imprimir el registro)

Imprime una gráfica o el archivo de registro (log)

4) Log open/Stop/Suspend (Abrir/cerrar/ o suspender un archivo de registro) (Log in Windows)

Abre un archivo de registro nuevo o añade a otro.

Cierra o suspende provisionalmente el registro.

5) Start View to Front (Coloca la ventana de registro al frente)

Coloca la ventana de registro sobre la ventana de Stata.

6) Bring Results to Front

Coloca la ventana de resultados al frente

7) Bring Graph to Front (Coloca la ventana de gráficas al frente)

Coloca la ventana de gráficas al frente de las otras ventanas de Stata

8) Do-file Editor (Editor de archivos-do)

Abre el editor de archivos-do, lo coloca al frente de las otras ventanas de Stata

9) Data Editor (Editor de datos)

Abre el editor de datos o lo coloca al frente de las otras ventanas de Stata

10) Data Browser (Visualizador o hojeador de datos)

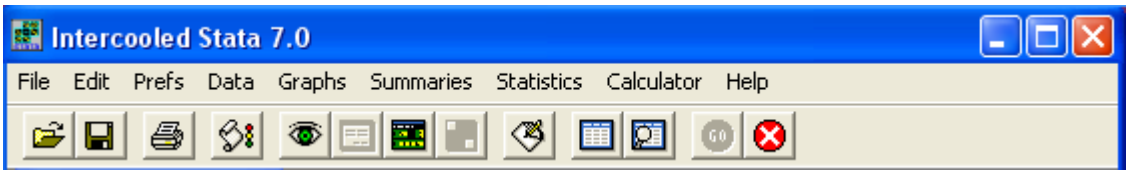
Abre el visualizador de datos o lo coloca al frente de las otras ventanas de Stata

11) Clear -more- Condition (continuar)

Le instruye a Stata que continúe después de parar durante una producción larga

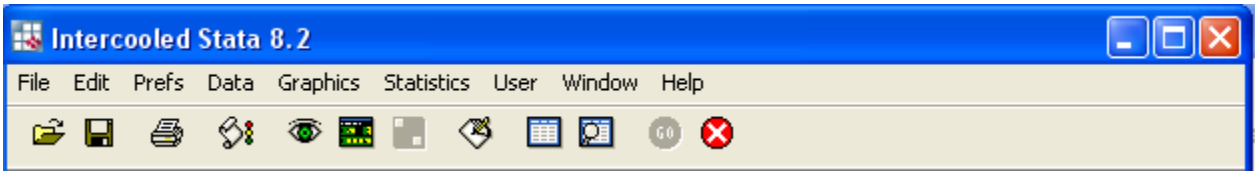
12) Break (interrumpir) Interrumpe lo que esté haciendo Stata.

Stata versión 7 tiene la opción de activar un submenú llamado **quest** el cual se proprociona en la página web de Stata.



Con el ***quest*** se pueden ejecutar algunas órdenes desde los menún a través de uso de ventanas como algunos gráficos, estadísticas de resumen, modelos estadísticos simples y empleo de calculadora.

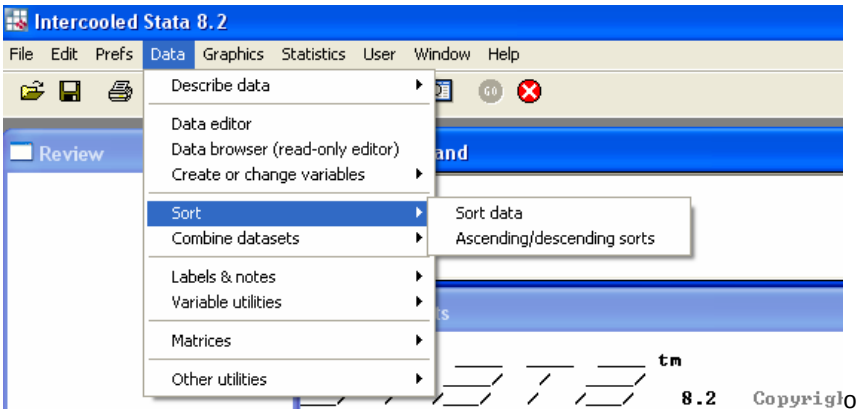
La nueva versión de stata (Stata 8) trae activada esta opción automáticamente y proprociona además el acceso directo desde el menú a opciones que stata 7 no contiene como por ejemplo Data, Graphics Statistics y el User, lo cuales permiten realizar a travéz de ventanas muchas de las órdenes que se hacen vía programación en la ventana de comandos.



Casi todas las órdenes de Stata se han implementado como diálogos y se pueden obtener por medio de menús que se han organizado por temas. Sólo elija una orden de los menús de **Statistics**, **Graphics** o **Data**, complete el diálogo y la orden se emitirá a Stata. Con estos nuevos menús y diálogos de Stata.

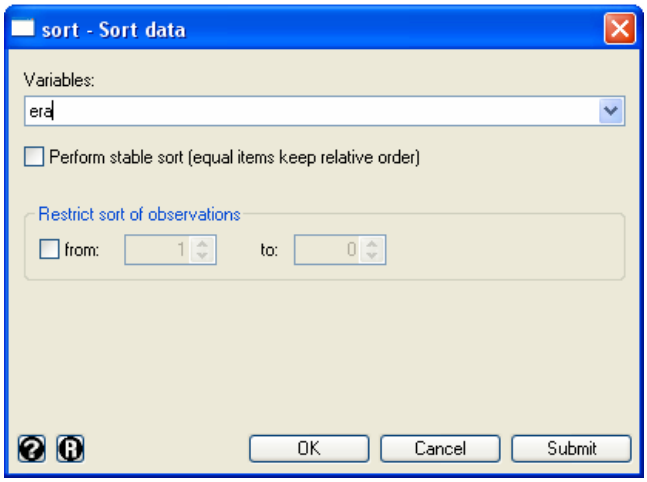
Ejemplo:

La orden para ordenar los datos de menor a myor en base a una columna o varaible es *sort,* si quiero aplicar la orden desde ventanas entonces con el cursos selecciono el menú **Data** en el cual aparecerá una lista de opciones. Con el mismo cursor navego hasta la opción Sort y selección la orden deseada: sort data



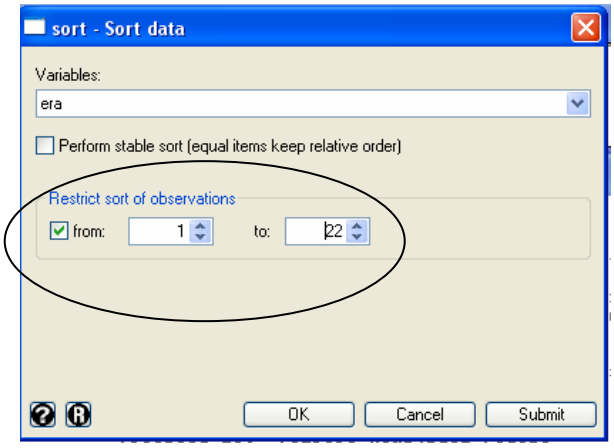
o:

Una vez seleccionada la orden aparecerá una ventana de diálogo en la cual solo tengo que agregar el nombre de la varaible por la cual deseo ordenar mis datos.



y eligo OK si deseo concluir la orden o submit si deseo realizar la orden pero contunuar con la misma ventana de diálogo. En la ventana de resultados aparecerá lo siguiente: **`sort era`** con lo cual se muestra que la orden se ejecutó.

Si es necesaro se pueden elegir las opciones que cada orden puede contener, por ejemplo si deseo puede dar click con el ratón en la opción *from* dentro de las opciones para restringir a un numero de observaciones y elegir de que a que observación deso ordenar:



En la ventana de resultados aparecerá:

```
. sort era in 1/20
```

y los datos se ordenarán de menor a mayor por la variable era solo en los registros del 1 al 20.

Lista de los Menús de Data y Stathistics contenidos en STATA 8.

Data: Contiene instrucciónes útiles para el manejo de bases de datos desde STATA.

Data:

I. Describe data
   1) Describe variables in memory
   2) Describe variables in file
   3) Describe data contents (codebook)
   4) Inspect variables
   5) List data
   6) Compactly list variable names
   7) Summary statistics

II. Data editor
III. Data browser (red-only editor)
IV. Create or change variables
   1) Create new variable
   2) Create new variable (extend)
   3) Other variable creation commands
      - Interaction expansion
      - Create indicator variables
      - Fill in missing values
      - Linearly interpolate/extrapolate values
      - Draw samples from normal distribution
      - Create dataset with specified correlation
      - Orthogonalize variables
      - Orthogonal polynomials
      - Create fractional polynomial powers
      - Linear spline construction
      - Zero-skewness log transform
      - Box-Cox transform
      - Generate numerical ranges
      - Calculate numeric derivatives
      - Calculate numeric integrals
   4) Change contents of variable
   5) Other variable transformation commands

V. Sort
    1) Sort data
    2) Ascending/descending sort

VI. Combine datasets
    1) Merge datasets
    2) Form all pairwise combinations within groups
    3) Append datasets
    4) Form every pairwise combination of two datasets

VII. Labels & notes
    1) Label dataset
    2) Label variable
    3) Define value label
    4) Assign value label to variable
    5) Set or change language for labels
    6) List value labels
    7) Drop value labels
    8) Save value labels as do-file
    9) Produce codebook describing value labels
    10) Add/remove numeric values from values labels
    11) Make dataset from value labels
    12) Add notes to data
    13) List notes
    14) Delete notes

VIII. Variable utilities
    1) Rename variable
    2) Set variable display format
    3) Eliminate variables or observations
    4) Change order of variables in dataset
    5) Alphabetize variables
    6) Relocate variable
    7) Compare two variables
    8) Compare two datasets
    9) Optimize variable storage
    10) Check for unique identifiers
    11) Check for duplicate observations
    12) Count observations satisfying condition

IX. Matrices
    1) Input matrix by hand
    2) Define matrix from expression
    3) Convert variables to matrix
    4) Convert matrix to variables

5) List contents of matrix
6) Rename matrix
7) Drop matrices
8) Eigenvalues & vectors of symmetric matrices
9) Singular value decomposition
10) Eigenvalues of square matrices

X. Other utilities
1) Hand Calculator
2) ICD-9 utilities
- Verifiy variable is valid
- Clean and verify variable
- Generate new variable frop existing
- Display code descriptions
- Search for codes from descriptions
- Display ICD-9 code surce

Graphs

I. asy graphs
1) Scatter plot
2) Connected scatter plot
3) Line graph
4) Area graph
5) Overlaid twoway graphs
6) Bar chart
7) Horizontal bar carth Dot chart
8) Pie chart (by variables)
9) Pie chart (by category)
10) Histrogram
11) Box plot
12) Horizontal box plot
13) Scatterplot matrix
14) Regression fit
15) Function graph

II. Twoway graph (scatterplot, line, etc.)
III. Overllaid twoway graphs
IV. Bar chart
V. Pie chart
VI. Histogram
VII. Box plot

VIII. Horizontal box plot

IX. Scatterplot matrix

X. Distributional graphs

1) Symetry plot
2) Quantiles plot
3) Normal quantile plot
4) Normal probability plot
5) Chi-squared quantile plot
6) Chi-squared probability plot
7) Quantile-quantile plot
8) Ladder of powers histograms
9) Ladder of powers normal quantiles plots
10) Spike plot and rootogram

XI. Smoothing and densities

1) Kernel density estimation
2) Lowess smoothing

XII. Regression diagnostics plots

1) Added-variable
2) Component-plus-residual
3) Augmented component-plus-residual
4) Leverage-versus-squared residual
5) Residual-versus –fitted
6) Residual-versus-predictor

XIII. Cross-sectional time-series line plots

XIV. Survival analysis graphs

1) Line plots
2) Correlogram(ac)
3) Partial correlogram (pac)
4) Periodogram
5) Cumulative spectral distribution
6) Bivariate cross-correlogram
7) Barlett´s white noise test
8) Vector autoregression (VAR) graphs

XV. ROC analysis

1) Nonparametric ROC curve
2) Parametric ROC curve after rocfit
3) Compare ROC Curves
4) Compare ROC curves against a gold standard
5) ROC curve after logistic/logit/probit
6) Sensitivity/specificity plot

XVI. Quality control

1) Cumulative sum(cusum)
2) C chart
3) P chart
4) R chart
5) X-bar chart
6) Vertically aligned X-bar and R chart
7) Standar error bar chart

XVII. More statistical graphs
1) Dendograms for hierartchical cluster analysis
2) Eigenvalues after factor analysis
3) Fractional polynomial regression plot
4) Odds of failure by category
5) Pharmacokinetic measures
6) Pharmacokinetic data summary
7) Means/medians by group
8) Comparative scatterplot

XVIII. Table of graphs

XIX. Manage graphs
1) Rename graph in memory
2) Copy graph in memory
3) Drop graphs
4) Describe graph
5) Make memory graph current
6) Query styles and schemes

XX. Change scheme/size

XXI. Graph preferencies

Statistics: Contiene las ordenes y funciones necesarias para análisis de cualquier nivel.

Statistics:

I. Summaries, tables & tests
3) Summary statistics
- Summary statistics
- Confidence intervals
- Normal CI calculator
- Binomial CI calculator
- Poisson CI calculator
- Correlations & covariances

- Pairwise correlations
- Partial correlations
- Arith./geometric/harmonic means
- Graph means/medians by groups
- Centiles with CIs
- Create varaible of percentiles
- Create variables of quartiles

4) Tables
   - Table of summary statistics (table)
   - Table of summary statistics (tabstat)
   - One/two-way tables
   - Mulple one-way tables
   - Two-way tables with measures of association
   - All possible two-way tabulations
   - Tabel calculator
5) Classical tests of hypotheses
   - One-sample mean comparison test
   - Two-sample mean comparison test
   - One –sample mean comparison calculator
   - Two-sample mean comaprison calculator
   - Binomial probability test
   - Binomial probability test calculator
   - One-sample proportion test
   - Two-sample proportion test
   - Group proportion test
   - One-sample proportions calculator
   - Two-sample proportions calculator
   - One-sample variance camparison test
   - Two-sample variance camparison test
   - Group variance camparison test
   - One-sample variance camparison calculator
   - Two-sample variance camparison calculator
   - Robust equal variance test
   - Sample size & power determination
6) Nonparametric test of hypotheses
   - One sample Kolmogorov-Simirrov test
   - Two sample Kolmogorov-Simirrov test
   - Kruscal-Wallis rank test
   - Wicoxon matched-pairs signed-rank test
   - Test equality of matched pairs
   - Mann-Whitney two-sample ranksum test

- k-sample equality of medians test
- Test for random order
- Trend test across order groups
- Spearman's rank correlation
- Kendall's rank correlation

7) Distributional plots & tests
   - Symetry plot
   - Quantiles plot
   - Normal quartile plot
   - Chi-squared quantile plot
   - Quantile-quantile plot
   - Stem & leaf display
   - Letter-value display
   - Cumulative distribution graph
   - Skewness & Kurtosis normality test
   - Shapiro-Wilk normality test
   - Shapiro-Francia normality test
   - Ladder of powers
   - Ladder of powers histograms
   - Ladder of powers normal quantile plots

II. Linear regression and related

1) Lineal regression
2) Regression diagnostics
   - Added variable plot
   - Component-plus-residual plot
   - Augmented component-plus-residual plus
   - Levarege-versus-squared residual plot
   - Residual versus-fitted plot
   - Residual versus-predictor plot
   - Ramsey RESET omitted varaible test
   - Score test for heteroskedaticity
   - DFBETAs
   - Variance inflation factors
   - Informations matrix test
   - Szroeter´s rank test for homoskedasticity
1) Box Cox regression
2) Errors-in-variables regression
3) Frontier models
4) Truncated regression
5) Constrained linear regression
6) Multiple equations model

- Instrumental varaibles & two stage least square
- Tree stage estimation
- Seemingly unrelated regression

7) Censored regression
   - Tobit regression
   - Censored normal regression
   - Interval regression
8) Fractional polynomial
   - Fractional polynimial regression
   - Multivariate fractional polynomial models
   - Fractional polynomial regression plots
   - Create fractional polynomial powers
9) Others
   - Variance-weighted least square
   - Robust regression
   - Nonlinear least square
   - Linear regression absrobing one cat. Variable

III. Binary outcomes
1) Logistic regression
2) Logistic regression reporting odds ratio
3) Probit regression
4) Probit regression (reporting change in probability)
5) Bivariate probit regression
6) Seemingly unrealted bivariate probit regression
7) GLM for the binomial family
8) Complementary log-log regression
9) Heteroskedastic probit regression
10) Skewedd logit regression
   - Grouped data
   - Logit regression for grouped data
   - Probit regression for grouped data
   - Weighted least-squares logyt regression
   - Weighted least-squares probit regression
11) Post-estimations
   - Goodness-of-fit for logistic/logyt/probit
   - Summary statisctics after logistic/logyt/probit
   - ROC curva after logistic/logyt/probit
   - Sensitivity/specificity plot

IV. Ordinal outcomes
1) Ordered logyt regression
2) Ordered probit regression

V. Count outcomes
   1) Poisson regression
   2) Goodness-of-fit after poisson regression
   3) Negative binomial regression
   4) Generalized negative binomial regression
   5) Zero-inflated poisson regression
   6) Zero-inflated negative regression

VI. Categorical outcomes
   7) Multinomail logistic regression
   8) Conditional logistic regression
   9) Nested logit regression
   10) Setup for nested logit regression
   11) Display nested logyt tree structure
   12) Rank-ordered logistic regression

VII. Selection models
   1) Heckman selection model(ML)
   2) Heckman selection model(two-step)
   3) Probit estimation with selection
   4) Tratment effects model (ML)
   5) Tratment effects model(two-step)

VIII. Generalized linear models (GLM)

   Generalized linear models(glm)

IX. Nonparametric analysis
   1) Kernel density estimation
   2) Lowess smoothing
   3) Robust nonlinear smoother
   4) Quantile regression
   5) Interquantile regression
   6) Simultaneous-quantile regression
   7) Bootstrapped quantile regression

X. Time series
   1) Setup & utilities
      - Declarate dataset to be time series data
      - Fill in missing time values
      - Report time-series aspects of dataset
      - Append obs to time-series dataset
   2) ARIMA models
   3) ARCH/GARCH
      - ARCH & GARCH models
      - Nelson EGARCH model
      - Threshold ARCH model

- GJR form of threshold ARCH model
- Simple asymmetric ARCH model
- Power ARCH model
- Nonlinear ARCH model
- Nonlinear ARCH model a single shift
- Asymmetric power ARCH model
- Nonlinear power ARCH modelx

4) Prais-Winsten regression
5) Regression with newey-West std. errors
6) Smoother/univariate forecasters
   - Single exponencial smoothig
   - Double exponencial smoothing
   - Holt-winters nonseasonal smoothing
   - Holt-winters seasonal smoothing
   - Nonlinear filter
   - Moving average filter
7) Tests
   - Augmented Dickey-Fuller unit-root test
   - Perform DF-GLS test for a unit root
   - Phillips-perron units roots test
   - Barlett´s periodogram-based white noise test
   - Portmanteau with noise test
   - Breusch-Godfrey LM test after regress
   - LM test for ARCH after regress
   - Durbin-Watson d statistic after regress
   - Durbin´s alternative test after regress
8) Graphs
   - Line plots
   - Autocorrelations & partial autocorrealations
   - Correlogram (ac)
   - Partial correlogram(pac)
   - Periodogram
   - Cumulative spectral distribution
   - Cross-correlogram for bivaraite time series

XI. Multivariate time series

1) Setup & utilities
   - Declarate dataset to be time series data
   - Fill in missing time values
   - Report time-series aspects of dataset
   - Append obs to time-series dataset

2) Basic vector autorregresive model
3) Vector autorregresive model (VAR)
4) Structural vector autorregresive model
5) VAR diagnostics and test
   - Granger causality tests
   - LM statistics for residual aucorrelation
   - Test for normally distributed disturbances
   - Lag-order selection statisticsx
   - Check VAR stability condition
   - Wald lag exclusion statistics
6) VAR dinamics forecasts
   - Compute forecasts (required for graph)
   - Graph forecasts
7) IRF & variance decomposition analysis
   - Create IRF result set
   - Impulse-response functions graphs
   - Overlaid graph
   - Impulse-response function talbes
   - Combined tables
8) Manage IRF results and files
   - Add an IRF results set
   - Rename IRF result set
   - Drop IRF result set(s)
   - Describe an IRF file
   - Erase an IRF file
   - Set active IRF file

XII. Cross-sectional time series

1) Setup & utilities
2) Linear models
3) Endogenous covariates
4) Dynamic panel data
5) Contemporaneos correlations
6) Random coefficientes
7) Frontier models
8) Binary outcomes
9) Count outcomes
10) Censored outcomes
11) Generalized estimating ecuations (GEE)
12) Line plots

XIII. Survival anaysis

1) Setup & utilities
2) Summary statistics, tests & tables
3) Regression models

XIV. Observational/Epi. Analysis
1) ROC analysis
2) Tables of epidemiologists
3) Other

XV. Survey data analysis
1) Setup & utilities
2) Distribution-specific models
3) Univariate estimator

XVI. ANOVA/MANOVA
1) Analysis of variance & covariance
2) Test linear hypotheses after anova
3) One-way analysis of variance
4) Large one-way ANOVA, random effects, and reliability
5) MANOVA
6) Multivaraite test after MANOVA
7) Wald test after MANOVA
8) Hotelling´s T-squared generalized means test

XVII. Cluster analysis
1) Kmeans cluster analysis
2) Kmedians cluster analysis
3) Single linkage clustering
4) Average linkage clustering
5) Complete linkage clustering
6) Weighted average linkage clustering
7) Median linkage culstering
8) Centroid linkage clustering
9) Wards linkage clustering
10) Post-clustering
    - Dendograms for hierachical cluster analysis
    - Cluster analysis stopping rules
    - Generate summary variables from cluster analysis
    - Cluster analysis notes
    - Detailed listing of cluster
    - Drop cluster analysis
    - Rename a cluster or cluster variables

XVIII. Other multivariate analysis
1) Multivariate regression
2) Factor analysis

3) Principal component analysis
4) Rotation of factor analysis
5) Scoring after principal component analysis
6) Scoring after factor analysis
7) Graph of eigenvalues after factor analysis
8) Cronbash´s alfa
9) Canonical correlations

XIX. Resampling & simulation

1) Bootstrap estimation
2) Bootstrap statistical from variables
3) Bootstrap statistical from file
4) Jackknife estimation
5) Montecarlo permutation test
6) Bootstrap sampling
7) Draw random sample
8) Draw a sample from a normal distribution
9) Create a dataset with a specified correlation structure

XX. General post-estimation

1) Obtain predictions, residuals, etc, after estimation
2) Nonlinear predictions after estimation
3) Tables of adjusted means & proportions
4) Tests
5) Linear combinations of estimators
6) Nonlinear combinations of estimators
7) Obtain marginal effects or elaticities after estimation
8) Replay marginal effects
9) Manage estimation results
10) Display variance-covariance matrix of estimators

XXI. Other

1) Collect statistics for a command across a by list
2) Stpwise estimation
3) Constrains
4) Quality control

Ayuda

Stata para Windows tiene un sistema de ayuda integrada. El sistema **Help.**

El help cuenta con las siguientes características para la utilización del mismo y del programa STATA.

Puede mantener la ventana de ayuda abierta mientras entra órdenes o instrucciones.
Al seleccionar ayuda **Help** usando la barra principal, podrá hacer una de las siguientes cosas:

- Ver el contenido de ayuda *(table of contents)*
- Buscar información sobre algún tema y obtener ayuda sobre alguna orden de Stata
- Listar las últimas adiciones a Stata, Además instalar la última versión oficial de Stata contenida en un disco flexible (o bajándola de la web si usa Stata para Windows 98/95/NT), programas de Stata escritos por otros usuarios o del boletín técnico (Stata Technical *Bulletin).*

Al seleccionar **(Search ...) usando** el menú de **Help** puede buscar información usando palabras claves y producir una pantalla que contiene:

- Enlaces de hipertexto (palabras pulsables de color claro) las cuales lo conectan con los archivos de ayuda correspondientes.
- Referencias a temas en los manuales de referencia y de gráficas (Reference Manual y Graphics Manual), a la guía del usuario (User's Guide) y al boletín técnico (Stata Technical B*ulletin.)*
- FAQs preguntas frecuentemente hechas sobre el tema en el sitio-web de Stata.

Ejemplo:

- Usando el menú de **Help,** seleccione **Search...**
- escriba **regression y oprima OK**

Verá todas las referencias sobre el tema **regression** en el manual de referencia y la guía del usuario. También verá una lista de todas las órdenes de Stata que tengan algo que ver con **regression.**

- Otras órdenes de Stata como **qreg, cnreg, y cnsreg** aparecerán en verde al colocar el puntero del ratón cerca del enlace de hipertexto, el puntero se cambiará a una mano. Si pulsa mientras la mano está sobre una de las órdenes, por ejemplo **qreg,** irá al archivo de ayuda para **qreg.**

Se pueden buscar temas múltiples usando el **Search** Al añadir temas se reduce los resultados de la búsqueda; por ejemplo:

•Entre regression residuals

Usando el menú Help, al seleccionar **Contents** obtendrá el contenido del sistema de ayuda.

- •Puede seleccionar uno de los enlaces para obtener ayuda sobre la orden
- •ó puede entrar el nombre completo de la orden en la ventana.

Ejemplo:

1) Pulse en la ventana **Help**

2) Entre ttest (ttest es una orden de Stata). Al oprimir **Enter** irá al archivo de ayuda para ttest

3) Oprima **Back** para regresar al archivo anterior

4) Oprima **Top** para regresar al contenido o a los resultados del **Search**

Stata cuenta con manuales para su uso, el help es solo una parte específica de los que se desea saber de Stata, es por eso que cuando en un texto en el help aparece la expresión "[R] se refiere a la anotación para la orden de interés en el manual de referencia. [R] es de referencia y la expresión "[G] **graph options",** se refiere a la anotación para **graph options** en el manual de gráficas. [G] es de gráficas.

Las órdenes de ayuda y búsqueda

1. Se puede entrar al sistema de ayuda desde la *ventana de* órdenes.
   Al hacer esto, los resultados aparecen en la ventana de resultados o en la ventana de ayuda.

2. Teclear *search tema* en la ventana de órdenes es igual que seleccionar **Search...** después de seleccionar Help de la barra principal y poner el *tema de interés.* Sin embargo, los resultados aparecen en la ventana de resultados.

3. Teclear *help nombre de la ordene*s igual que seleccionar de la barra principal Help, después **Stata command...,** y entrar *nombrede la orden,* pero los resultados no aparecerán en la ventana de resultados.

4. **Diferencia importante:**
   Con las órdenes *help y search,* no tendrá enlaces de hipertexto en la ventana de resultados.

5. Se puede obtener ayuda con enlaces de hipertexto en la ventana de órdenes.
   En lugar de teclear *help nombre de la orden,* teclee w*help nombre de la orden.* El archivo de ayuda aparecerá en la ventana de ayuda y podrá usar los enlaces de hipertexto. (Teclar *whelp nombre de la orden* es igual que usar la barra principal, seleccionar **Help Stata command...,** y teclear *nombre de la orden.)*

El editor de datos

Para ejecutar el editor de datos:

- Se oprime el botón **Data Editor**
- ó se teclea edit en la ventana de órdenes y se oprime **Enter** ↵

El editor de datos funciona como una hoja de cálculo, cada columna es una variable y cada fila una observación. Dentro del editor puede navegar pulsando la celda deseada o usando las flechas del teclado y también puede copiar datos de otras hojas de cálculo al editor de Stata y viceversa:

1) En el editor de Stata o en la otra hoja de cálculo resalte los datos que desea copiar. Seleccione **Edit** y después **Copy.**
2) Ahora hay que pegar los datos en el editor de Stata o en la otra hoja de cálculo. Esto se hace seleccionando la celda superior en el lado izquierdo del área donde desea copiar los datos.
3) Seleccione **Edit** y después **Paste**

Para modificar o añadir datos

1) Se selecciona la celda, se teclea el valor y se oprime **Enter** *o* ***Tab***

Nota: La diferencia entre Enter y T*ab* es que:

- Enter lo mueve de fila en fila en la misma columna y
- *Tab lo* mueve de columna a columna en la misma fila hasta al final y después a la primera columna de la próxima fila.

Para añadir variables:

1) Se pulsa en la primera celda de la primera columna vacía
2) Se teclea el valor
3) Se oprime **Enter** para bajar a la próxima celda

Para añadir observaciones:

1) Se pulsa en la primera celda de la primera fila vacía
2) Se teclea el valor
3) Se oprime Enter para moverse hacia abajo
4) Después de terminar con la primera observación, se pulsa en la primera celda de la segunda fila
5) Se teclean los valores de la segunda observación y se oprime Tab para moverse a la derecha
6) Al terminar de entrar cada observación, *Tab* automáticamente lo llevará a la primera columna de la próxima fila.

Datos numéricos y alfanuméricos

(Datos compuestos de letras y números) se añaden de la misma manera.

- No necesita usar comillas alrededor de valores alfanuméricos

Valores numéricos que faltan (nulos) son simbolizados con un punto '**.**' y se añaden oprimiendo Enter o *Tab* 0 tecleando '**.**' y oprimiendo Enter o *Tab*

Valores alfanuméricos nulos se dejan simplemente vacíos y se añaden oprimiendo Enter o *Tab*

El editor: nombra las variables var1, var2,

Para cambiar el nombre de una variable:

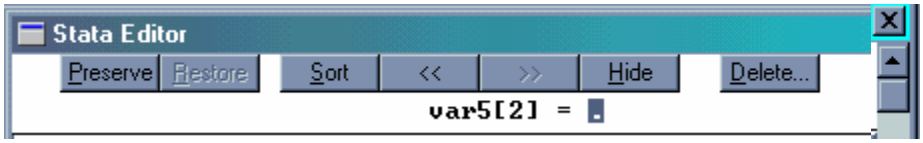
1) Se pulsa doblemente en cualquier lugar en la columna de la variable de interés. Esto abre la ventana de la variable (**Variable Information:)**

2) Teclee el nombre nuevo de la variable en la línea que dice Name

El nombre debe tener de 1 a 8 caracteres. Aunque en STATA ver. 7 y Stata versión 8 pueden ocupar más de 8 caracteres. Una recomendación es utilizar nombres cortos para que puedan ser compatibles con otros programas como Epi-Ino y SPlus.

- Los caracteres pueden ser letras: A - Z, a - z, números: 0 - 9 ó el "-"
- No se pueden usar espacios u otros caracteres Ejemplo: Mi-nombre. El primer carácter debe ser una letra o el "-", pero no se recomienda empezar el nombre con el "-"

Los botones del editor de datos



El editor de datos tiene siete botones:

Preserve (preservar). Se oprime este botón si está satisfecho con los cambios que ha hecho y desea permanecer en el editor para hacer más cambios, puede actualizar el archivo de seguridad antes de seguir.

Restore (restaurar). Al abrir el editor, Stata automáticamente hace una copia de seguridad del archivo de datos.
Si desea cancelar los cambios que haya hecho antes de salir del editor y restaurar la copia de seguridad oprima este botón.

Sort (ordenar, clasificar). **Sort** pone las observaciones en orden ascendente según los valores de la variable resaltada.

<< El botón **<<** mueve la variable resaltada a la primera columna.

>> El botón **>>** mueve la variable resaltada a la última columna.

Hide (esconder). **Hide** esconde la variable resaltada. La variable existe pero el editor no la sigue mostrando.

Delete... (Borrar) Delete... abre otra ventana que le deja: borrar la variable resaltada, borrar la observación resaltada o borrar todas las observaciones en la base de datos que tengan el mismo valor que la variable resaltada.

Todas las órdenes dadas en el editor se registran en la ventana de resultados. Las órdenes son idénticas a las órdenes que se usan en Stata. El guión al frente de la orden indica que el cambio fue hecho en el editor de datos.

➢ Creando una base de datos con el editor

Nota para personas con experiencia usando Stata: El editor de datos hace todo lo que hace la orden **input** y mucho más.

Ilustramos el uso del editor de datos usando los siguientes datos de mortalidad por neumonía e influenza:

País	Año	Numero de casos	Tasa de mortalidad	Porcentaje
Argentina	1994	560	83.11	3.78
Belice	1989	5	113.38	4.63
Brasil	1993	5534	152	12.64
Canadá	1992	26	6.52	1.07
Chile	1994	368	127.72	10.7
Colombia	1991	1367	152.68	10.64
Cuba	1995	87	59.23	6.29
Estados Unidos	1991	607	14.77	1.65
Guatemala	1993	4206	1439.14	33.42
México	1994	7687	264.7	15.42
Perú	1992	3275	525.77	23.2
Puerto Rico	1992	20	29.5	2.4
Venezuela	1993	875	166.86	7

Ref. Infecciones Respiratorias en niños, Yehuda Benguigui. OPS/OMS. 1997. pag27

Las variables son: País, año de última información, total de casos de muerte por neumonía e influenza, tasa de mortalidad por 100,000 nacidos vivos y porcentaje sobre el total de muertes.

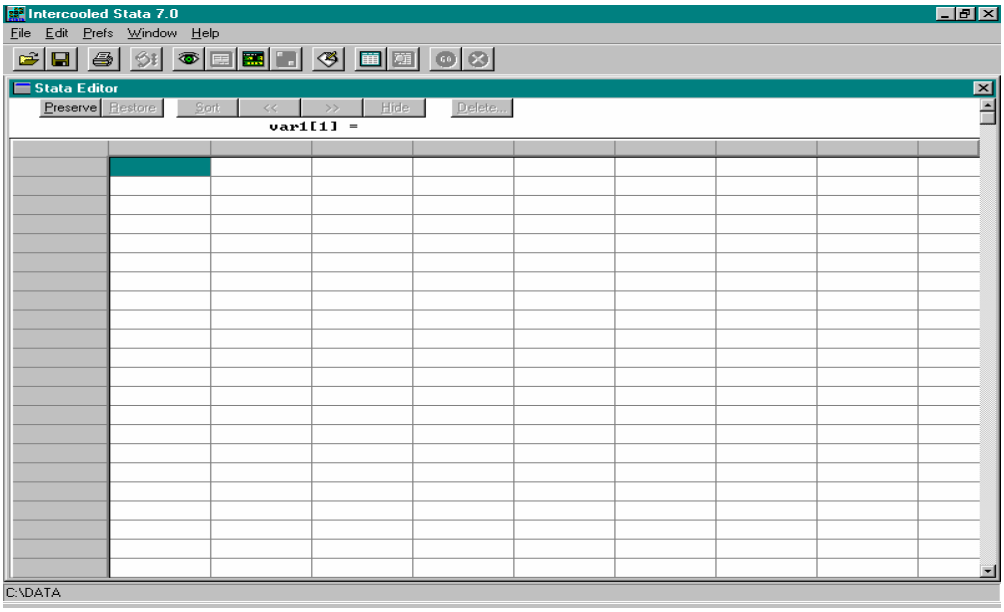
¿Como se genera en stata?

Ahora vamos a crear una base de datos usando el editor de Stata.

1. *Ejecute el editor.*

   Oprima el botón **Data Editor** ó teclee edit en la ventana de órdenes.

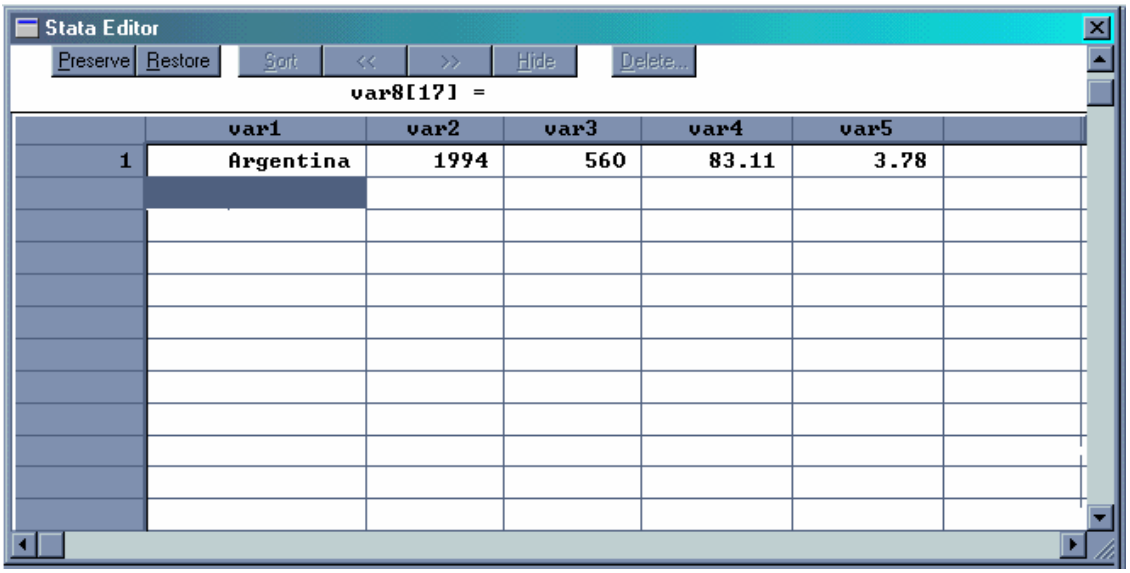
   Usted verá la ventana siguiente:



2. Introducir los datos.
   Los datos se pueden añadir una variable a la vez o una observación a la vez.
   Las columnas corresponden a las variables y las filas a las observaciones.

3. Para añadir una nueva observación, oprima Tab después de teclear cada valor.
   Empezando en la primera celda de la primera fila, teclee el país Argentina y oprima Tab para moverse a la próxima celda a la derecha. No oprima Enter porque eso lo baja a la siguiente observación.

   Ahora entre el año 1994 y oprima Tab. Siga así hasta entrar todos los valores de la primera observación. Ahora pulse la segunda celda en la primera columna y entre los datos de la segunda observación siempre usando la tecla Tab.



4. Después de entrar la primera observación, Stata sabe cuántas variables tiene. Al teclear **Tab** después de entrar el último valor de la segunda observación en adelante, se moverá automáticamente a la primera columna de la próxima observación.

5. Para añadir datos una variable a la vez, oprima **Enter** *después de teclear cada* valor. Pulse la primera celda de la primera columna vacía. Teclee los valores de la variable oprimiendo **Enter** después de cada valor.

Notas que necesita saber para añadir datos

No se necesitan comillas *alrededor de valores* alfanuméricos como en otras órdenes que sí las requieren (**" "**) alrededor de valores alfanuméricos. Puede usar las comillas en el editor pero no es necesario.

Un punto ('.') representa un valor numérico que falta (nulo). O llamado **missing**

Sólo necesita oprimir Tab o Enter para añadir valores alfanuméricos nulos, esto resultará en una variable vacía (sin nada) en esta observación ó teclear ('**.**'). **Enter**

Stata no acepta columnas ni filas vacías en la base *de datos* .
Al añadir una nueva variable o una nueva observación siempre empieza en la primera columna o fila vacía. Si se salta una fila o columna, Stata va a rellenar la columna o fila vacía con valores nulos.

Si ve por *ejemplo, var3 [4]* = en la parte superior del editor :

Esto corresponde a la celda seleccionada. var3 es el nombre predeterminado para la tercera variable, y [4] indica la cuarta Después de entrar la primera observación, Stata sabe cuántas variables tiene. Al teclear Tab después de entrar el último valor de la segunda observación en adelante, se moverá automáticamente a la primera columna de la próxima observación.

Para añadir datos una variable a la vez, oprima Enter *después de teclear cada* valor. Pulse la primera celda de la primera columna vacía. Teclee los valores de la variable oprimiendo Enter después de cada valor.

Observación. Si desea conservar los cambios realizados en su ventana de editor, al cerrar la ventana aparecerá un mensaje preguntando si desea conservar sus cambios, si la opción es **sí** presiones ***preserve.*** Estos datos solo permanecen en la memoria mas no han sido guardados en el disco duro, para tal caso deberá guardar el archivo usando el menú **File** y seleccionando **Save as.** De el nombre deseado.

Será necesario renombrar las variables con nombres que identifiquen mejor a cada una de ellas, esto lo haremos con el comando **rename** que se puede abreviar como ren)

Stata results

```
. ren var1 pais
. ren var2 year
. ren var3 num
. ren var4 tasamort
. ren var5 porcent
```

Esto también se puede hacer con el editor dando doble clic en la columna de la variable que quiero renombrar y en una reemplazar el nombre anterior por el nuevo.

¿Cómo Cambiar y visualizar datos con el editor de datos?

Uso avanzado del editor de datos

Puede seleccionar las variables que van a aparecer en el editor:

Escribiendo en la ventana de órdenes:

Orden	Función
• edit id	Selecciona la variable pais
• edit pais year	Selecciona las variables pais year
Incluir cualquier número de variables, restringir el número de observaciones que aparecen en el editor: Escribiendo en la ventana de órdenes:	
• edit in 1	Sólo usa la primera observación
• edit in 2	Sólo usa la segunda observación
• edit in -2	Sólo usa la penúltima observación
• edit in -1	Sólo usa la última observación
• edit in 1 (Le., l	Sólo usa la última observación
Restringir el editor a una serie de observaciones usando "en" (in):	
• edit in 1/9	Usa de la primera a la novena observación
• edit in 2/-2	Usa de la segunda a la penúltima observación
Restringir el editor a una serie de observaciones que sólo satisfacen una expresión matemática usando el condicional "si" (if):	
• edit if exp	Usa observaciones en las que la expresión exp es cierta
• edit if tasamort>15	Usa observaciones en las cuales tasamort>15
• edit if tasamort==15	Usa observaciones en las cuales tasamort es igual a 20
• edit if num==.	Usa observaciones en las cuales el valor de num falta
Combinar in e if (el orden no importa):	
• edit in 1/9 if tasamort>=1439.14	Usa de la primera a la novena observación, sólo si tasamort mayor o igual que 25
• edit if porcent<15 in 5/-1	Usa de la quinta a la última observación sólo si porcent<15
Puede seleccionar variables y restringir observaciones al mismo tiempo:	
• edit id in 5/-5	Usa sólo la variable id de la quinta a la –5 observación.

Nota: las variables son sin acentos y no se utiliza la ñ. Deben teclearse tal con mayúsculas y/o minúsculas según esté escrito el nombre de la variable

También es posible cambiar los datos dentro del editor escribiendo sólo edit o edit *varnombre(s),* edit if .... etc. ó pulsando **Data Editor** (pero no puede seleccionar variables ni observaciones), al abrir el editor pulse la celda que desea cambiar y entre el nuevo valor de la variable y teclee *Enter o Tab.* Si restringe el editor a las variables y observaciones de interés disminuye la posibilidad de cometer errores. Aunque para hacer cambios globales a los datos, es mejor usar la orden **replace**.

Para borrar variables u observaciones oprima el botón **Delete...** ; pero es preferible que para borrar varias observaciones o variables a la vez, utilice la orden **drop**.

Browser

El editor de datos puede ser usado para visualizar los datos.

Para usar el editor como un visualizador (browse):

- Oprima el botón **Data Browser**
- ó escriba browse en la ventana de órdenes

El visualizador no le deja cambiar los datos. Use el visualizador **(browse), y no edit,** cuando solamente desea examinar los datos, esto permitirá que usted no cometa un error en su base de datos que después no pueda corregir.

En el visualizador también es posible seleccionar variables y observaciones deseadas procediendo igual que con el editor. Ejemplo:

Stata results

```
. browse pais year
. browse in 1/13
. browse if porcent==.
. browse pais year tasamort in 5/-5 if porcent>=15
```

Se da la orden seguida de la lista de variables y opcionalmente seguida de **if** y/o **in.**

El browse puede hacer muchas de las mismas cosas que hace la orden **list.** Pero es más conveniente porque lo deja desplazarse.

Manejo y manipulación de Datos.

Descripción de datos y etiquetas para las bases y/o las variables.

➢ *describe y label*

Función	Instrucción
Cómo describir los datos	
que tiene en memoria:	`describe`
que tiene guardados en el disco:	`describe using c:/archive o "c:/archive"`
Cómo ponerle etiquetas a la base de datos:	`label data "texto"`
Cómo ponerle etiquetas a las variables:	`label var varnombre "texto"`
Cómo ponerle etiquetas a los valores de las variables:	
Definir una etiqueta para los valores:	`label define etiqueta # "texto1" # "texto2"`
Asocie la etiqueta con la variable:	`label values varnombre etiqueta`
Nota: Puede asociar la misma etiqueta para valores con distintas variables.	
Cómo quitar la etiqueta	
de la base de datos:	`label data`
de la variable:	`label var vamombre`

de los valores de las variables:	label values *varnombre*
Cómo borrar una etiqueta para valores:	label drop *etiqueta*
Cómo cambiar una etiqueta para valores:	
Bórrela:	label drop *etiqueta*
Vuelva a definirla:	label define *etiqueta # "texto" # "texto"*
Cómo cambiar una base de datos permanentemente	
vuelva y guarde los datos	Del menú de **File,** seleccione **Save.**
o, teclee:	save *archivo,* replace

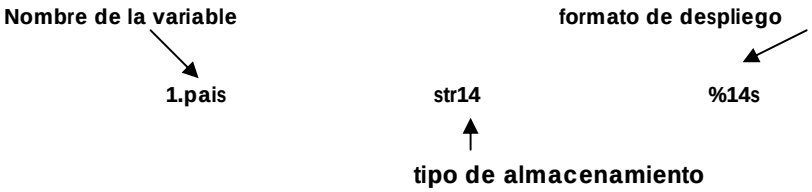
➢ describe

Guardamos la base de datos creada en el editor como el archivo:

save a:/ tasas. dta

Stata results

```
. use a:/tasas.dta

. list
               pais      year       num   tasamort    porcent
  1.      Argentina      1994       560      83.11       3.78
  2.         Belice      1989         5     113.38       4.63
  3.         Brasil      1993      5534        152      12.64
  4.         Canadá      1992        26       6.52       1.07
  5.          Chile      1994       368     127.72       10.7
  6.       Colombia      1991      1367     152.68      10.64
  7.           Cuba      1995        87      59.23       6.29
  8. Estados Unidos      1991       607      14.77       1.65
  9.      Guatemala      1993      4206    1439.14      33.42
 10.         México      1994      7687      264.7      15.42
 11.           Perú      1992      3275     525.77       23.2
 12.    Puerto Rico      1992        20       29.5        2.4
 13.      Venezuela      1993       875     166.86          7

** Usemos la o orden describe para describir estos datos:

. describe
Contains data from A:\Yo.dta
  obs:            13
 vars:             5                          24 Jul 2001 19:49
 size:           390 (100.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
pais            str14  %14s                   pais
year            int    %8.0g                  Anio de última información
num             int    %8.0g                  Total de Casos de muerte por
                                                Neumonia e Influenza
tasamort        float  %9.0g                  Tasa de mortalidad (x100,000
                                                nacidos vivos)
porcent         float  %9.0g                  Porcentajesobre el total de
                                                muertes
-------------------------------------------------------------------------------
```



1. El nombre de la variable es como nos vamos a referir a la columna de datos.
2. Los tipos de almacenamiento se refieren a la amplitud de los datos entrantes en la variable y si los datos son numéricos o alfanuméricos.
3. Los formatos de despliego controlan cómo se representan los valores en la pantalla y en los archivos de registro.

No es necesario cargar el archivo de datos en la memoria de la computadora para describirlo:

.describe using a:/Tasas

Es decir, al teclear la orden describe sin argumentos, Stata describe la base de datos que tiene en memoria, si teclea describe using *archivo,* Stata describe el contenido de la base de datos especificada. (en este caso el archivo llamado *archivo.* dta creado por Stata).

➢ label

Se le pueden poner etiquetas (labels) a una base de datos, a las variables y a los valores de las variables. Como ejemplo, usemos el archivo de **tasas.dta.**

.describe using a:/tasas.dta

Agreguemos a la base de datos **tasas** una nueva variables que tenga el número 1 en los países de Norteamérica, 2 en los países de Centroamérica, 3 en los países de Sudamérica, y 4 en los países del Caribe.

A esta variable ponerle el nombre de **Región**.

1. Describir la base
2. Con **label var** etiquetar la variable.
3. Ponerle también una etiqueta a cada uno de los números identificando la región. Esto es útil para recordar el contenido de las variables. En el caso de cuestionarios muy extensos, lo es más.

➢ Etiquetas para bases de datos y variables

Es decir, **label var** se utiliza para ponerle etiquetas a las variables. Ponga el texto entre comillas, por ejemplo:

.**label var** región "Región de América a la que pertenece:"

Así podremos etiquetar todas las demás variables y además ponerle etiquetas a los valores de las variables.

Stata results

```
. desc

Contains data from A:\Tasas.dta
  obs:            13
 vars:             6                          24 Jul 2001 19:49
 size:           442 (100.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
pais            str14  %14s                   Pais
year            int    %8.0g                  Anio de ultima información
num             int    %8.0g                  Total de Casos de muerte por
                                                Neumonia e Influenza
tasamort        float  %9.0g                  Tasa de mortalidad (x100,000
                                                nacidos vivos)
porcent         float  %9.0g                  Porcentajesobre el total de
                                                muertes
region          float  %9.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved


. label var  region "Región de América a la que pertenece"

.  label  define  region  1  "Norteamerica"  2  "Centroamercia"  3  "Sudamerica"  4
"Caribe"

. label value region region

. tab  region

    Region de |
 America a la |
que pertenece |      Freq.     Percent        Cum.
--------------+-----------------------------------
 Norteamerica |          3       23.08       23.08
Centroamercia |          3       23.08       46.15
   Sudamerica |          5       38.46       84.62
       Caribe |          2       15.38      100.00
--------------+-----------------------------------
        Total |         13      100.00

. tab region, nolabel
  Region de |
```

```
  America a |
     la que |
  pertenece |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          3       23.08       23.08
          2 |          3       23.08       46.15
          3 |          5       38.46       84.62
          4 |          2       15.38      100.00
------------+-----------------------------------
      Total |         13      100.00

. describe

Contains data from A:\Tasas.dta
  obs:            13
 vars:             6                          24 Jul 2001 19:49
 size:           442 (100.0% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
pais            str14  %14s                   Pais
year            int    %8.0g                  Anio de ultima información
No              int    %8.0g                  Total de Casos de muerte por
                                                Neumonia e Influenza
tasamort        float  %9.0g                  Tasa de mortalidad (x100,000
                                                nacidos vivos)
porcent         float  %9.0g                  Porcentajesobre el total de
                                                muertes
region          float  %13.0g      region      Region de America a la que
                                                pertenece
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

Etiquetas para los valores de las variables

Hemos agregado la nueva variable la cual etiquetamos, también etiquetamos los valores de las variables:

Norteamérica=1, Centroamérica=2, Sudamérica=3 y Caribe=4 con label define y label value

Es decir,

1. **label define** para crear una etiqueta de valores. La sintaxis es
   label define *etiqueta # "contenido" # "contenido" ...*

2. **label values** para asociar la etiqueta con la variable. La sintaxis es
   label values *varnombre etiqueta*

3. Para que los cambios sean permanentes hay que volver a guardar el archivo.

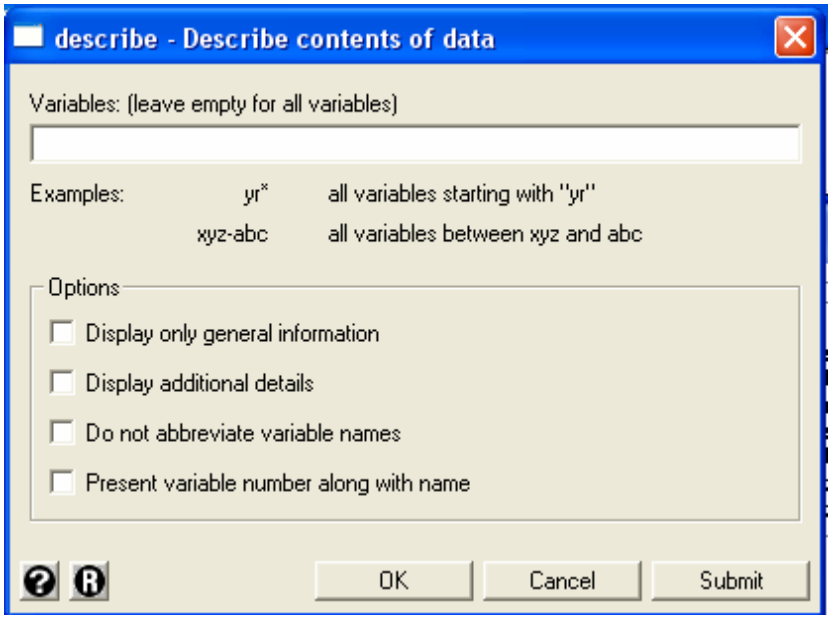
La misma asignación de región que le dimos a la variable **region** se podría utilizar para otra variable que asocie los mismos números con las mismas etiquetas.

Si nos damos cuenta, al teclear ***describe*** en la variable region aparece en la columna de value labels el nombre region.

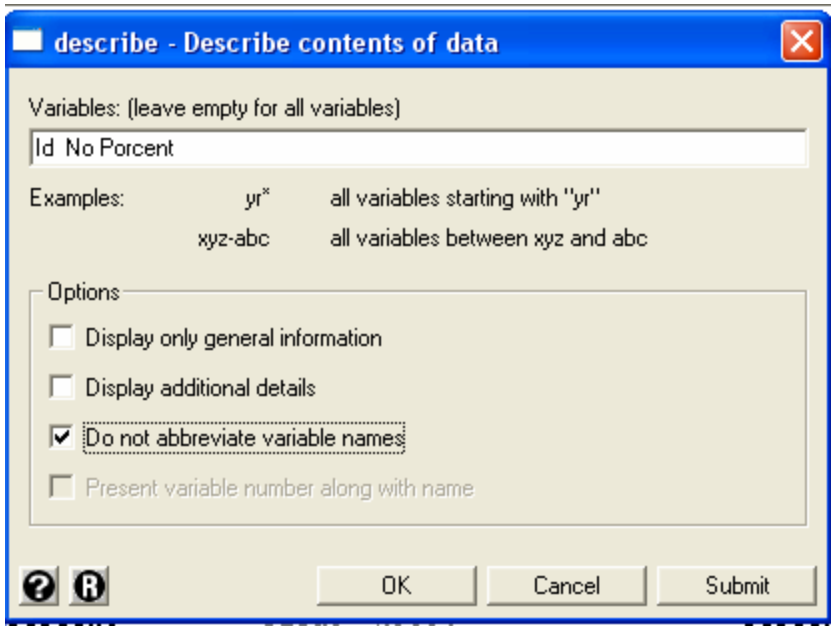
Manejo y manipulación de Datos desde Ventanas de diálogo:

Describe:

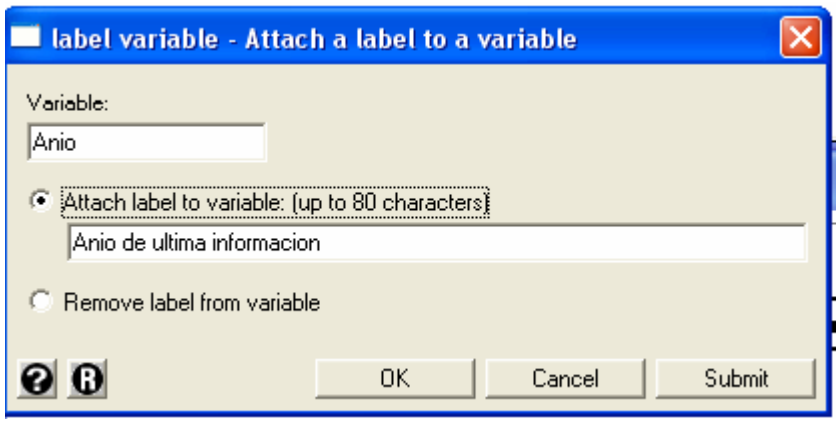
Todas las variables



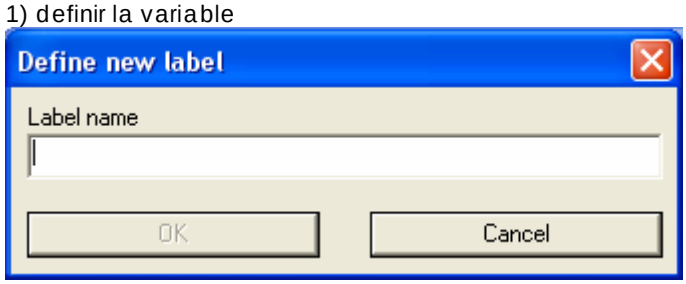
solo una selección de variables



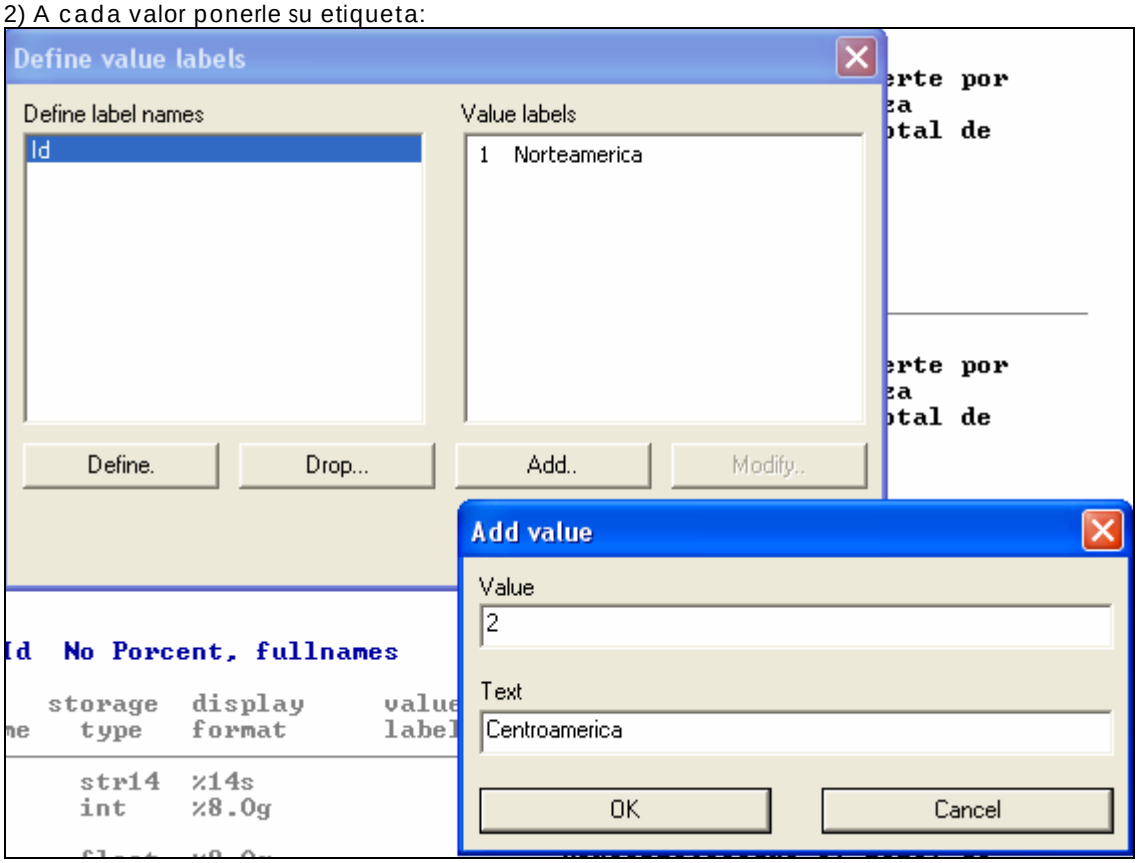
Etiquetas de variables:



Etiquetas de valores
1) definir la variable



2) A cada valor ponerle su etiqueta:



Funciones o métodos abreviados.

• La ventana de repaso (Review window) contiene las órdenes dadas anteriormente. Si usted pulsa una vez cualquier orden previa localizada en la ventana Review, ésta será copiada a la ventana de órdenes. Si pulsa doblemente cualquier orden previa ésta será copiada y ejecutada.
• Si crea un archivo de registro (log) podrá ver todas las órdenes y los resultados anteriores.

En la ventana de variables (Variables window) se ven las variables actualmente en la memoria. Pulse una vez cualquier variable y el nombre será copiado a la ventana de órdenes. (Si pulsa doblemente, la variable se copiará dos veces). Esta ventana tiene una barra de desplazamiento.

El texto en la ventana de órdenes se edita de la misma manera que el texto en las ventanas de Windows. Las teclas para editar texto en la ventana de órdenes son:

Delete o Supr	Borra caracteres a la derecha del cursor
Backspace	Borra caracteres a la izquierda del cursor
Esc	Borra la línea completa
Home o Inicio	*Mueve* el cursor al principio de la línea
End o Fin	Mueve el cursor al final de la línea
Page Down o Re Pág	Se mueve hacia abajo
Page Up o Av Pág	Recupera la orden anterior

Page *Down Se mueve hacia abajo.* Hace lo contrario de Page Up que se mueve el cursor hacia arriba. (Las teclas Page Up y Page *Down* hacen lo mismo que pulsando una vez cualquier orden en la ventana de repaso.)

Listar datos

➢ list (lista)

La orden list y la orden browse son muy parecidas.

Función	Orden
Para listar en la ventana de resultados, teclee:	. list
Si la palabra --more-- (más) aparece en la ventana de resultados, lo cual pasa con listas largas,	
Para ver la próxima línea:	Teclee Enter.
Para ver la próxima pantalla:	Oprima cualquier tecla.
o:	Oprima el botón More.
Para interrumpir completamente una orden de Stata	
y regresar al estado en que se encontraba antes e dar la orden:	Oprima el botón **Break.**
o:	Teclee *Ctrl-Break.*
Para listar una variable sola:	list *varnombre*
Ejemplo:	list pais
list se puede abreviar:	1 pais
También puede abreviar	list pa

el nombre de la variable:	
Parar listar varias variables:	list *vamombres(s)*
Ejemplo:	list pais year
Puede abreviar:	1 pais year
Para listar de la variable *varnombrei* a *vamombrej:*	list *vamombrei-vamombrej*
Ejemplo:	list pais-porcent
Puede abreviar:	1 pais-porcent
Para listar las variables que empiezan con	
la letra p:	list p*
Puede combinar todo lo anterior:	**list** year-tasamort p*
Para listar la tercera observación:	list in 3
la penúltima observación:	list in -2
la última:	list in -1
la primera*:*	list in 1
Para listar de la primera a la tercera observación: *list in 1/3*	
de la 5 ala 17:	list in 5/17
de la 3 a la penúltima:	list in 3/-2
Puede combinar todo lo anterior:	list year-tasamort p* in 3/-3
Para listar observaciones que satisfacen una	
condición, use *if exp* (si la expresión):	list if exp
Ejemplo:	list if year==1992
Puede combinar todo lo anterior:	list year-tasamort p* if year==1992
	list year-tasamort pop* if year==1992 in 3/-3
Todo lo que aparece en la ventana de resultados,	
incluyendo la lista producida por *list,*	
puede ser registrado en un archivo (log)	
Especificar que se dibujen líneas horizontaleles entre las observaciones	list, separator(5)

Notas

1) La orden list sin argumentos produce una lista de todas las observaciones y variables. Puede oprimir el botón **Break** e interrumpir la lista en cualquier momento.
2) Puede producir una lista de un subconjunto de variables especificando los nombres de las variables. Por ejemplo: produce una lista de las variables *pais year num*. Puede abreviar: *list p** produce una lista de las variables que empiezan con la letra p. *list pais-num* produce una lista de todas las variables localizadas entre las variables *pais y num*, dependiendo en órden en que usted las tenga.
3) Puede abreviar list como l (la letra l).
4) Hay que tomar en cuenta que "in" restringe la lista a un rango de observaciones, los números positivos cuentan desde la primera observación hacia abajo mientras que los números negativos cuentan desde la última observación hacia arriba.

listas usando "if"

Stata results

```
. list
                pais       year        num    tasamort     porcent          region
  1.       Argentina       1994        560       83.11        3.78      Sudamerica
  2.          Belice       1989          5      113.38        4.63    Centroamercia
  3.          Brasil       1993       5534         152       12.64      Sudamerica
  4.          Canada       1992         26        6.52        1.07     Norteamerica
  5.           Chile       1994        368      127.72        10.7      Sudamerica

  6.        Colombia       1991       1367      152.68       10.64     Norteamerica
  7.            Cuba       1995         87       59.23        6.29           Caribe
  8. Estados Unidos       1991        607       14.77        1.65     Norteamerica
  9.       Guatemala       1993       4206     1439.14       33.42    Centroamercia
 10.          Mexico       1994       7687       264.7       15.42    Centroamercia

 11.            Peru       1992       3275      525.77        23.2      Sudamerica
 12.     Puerto Rico       1992         20        29.5         2.4           Caribe
 13.       Venezuela       1993        875      166.86           7      Sudamerica

. list if  region==2

                pais       year        num    tasamort     porcent          region
  2.          Belice       1989          5      113.38        4.63    Centroamercia
  9.       Guatemala       1993       4206     1439.14       33.42    Centroamercia
 10.          Mexico       1994       7687       264.7       15.42    Centroamercia


. list if region==2 &  tasamort>15

                pais       year        num    tasamort     porcent          region
  2.          Belice       1989          5      113.38        4.63    Centroamercia
  9.       Guatemala       1993       4206     1439.14       33.42    Centroamercia
 10.          Mexico       1994       7687       264.7       15.42    Centroamercia

. list if region==2 &  tasamort>15 & porcent<10

                pais       year        num    tasamort     porcent          region
  2.          Belice       1989          5      113.38        4.63    Centroamercia

. list if region==2 | region==1 & (tasamort>15 & porcent<10)

                pais       year        num    tasamort     porcent          region
  2.          Belice       1989          5      113.38        4.63    Centroamercia
  9.       Guatemala       1993       4206     1439.14       33.42    Centroamercia
 10.          Mexico       1994       7687       264.7       15.42    Centroamercia
```

En muchas de las órdenes de Stata es necesario utilizar condiciones como en el edit, browse, list, generete, etc., los más utilizados son los mencionados en los ejemplos anteriores como el "if" que es el condicional "si". "if *exp"* quiere decir: si la expresión *(exp)* es cierta. Algunas expresiones pueden ser más complicadas como el "'&" que es la conjunción "y", y el " | " es la conjunción "o".

Los operadores lógicos son:

<	menor que
<=	menor que o igual
==	igual
>=	mayor que o igual
>	mayor que
~=	no es igual
&	la conjunción: y
|	la conjunción: o
~	no (la negación lógica)
( )	paréntesis para especificar la orden de las operaciones

La conjunción siempre es evaluada antes de la conjunción |; así que, **a | b&c** resulta en **a | (b&c),** lo cual es cierto si **a** es cierto o si ***b*** y ***c*** son ambas cierto. Para especificar que **a** o ***b*** sea cierto, y que **c** también sea cierto, escriba **(a| *b)&c***.

Crear variables nuevas

➢ generate y replace (crear y reemplazar)

Para crear una variable nueva la cual contiene el resultado de una expresión algebraica	generate *newvar* = *exp*
La orden generate (crear o generar) se puede abreviar:	g *newvar* = *exp*
Para cambiar (reemplazar) el contenido de una variable:	replace *o1dvar* = *exp*

La orden replace no se puede abreviar.

exp es una expresión algebraica que puede ser una combinación de otras variables, operadores y funciones.

Operadores:

Matemáticos		Lógicos		Relacionales (numéricos y alfanuméricos)	
+	adición	~	no ó !	>	mayor que
-	substracción	|	o	<	menor que
*	multiplicación	&	y	>=	> o igual
/	división			<=	< o igual
^	exponente			==	igual
				~= ó !=	no es igual

+ concatenación de valores alfanuméricos

Algunos ejemplos de funciones que se pueden utilizar con el generate son:

Cos( ), exp( ), ln( ), lnfact( ), sqrt( ), chiprob( ), fprob( ), uniform( ), lower( ), real(), rtrim( ), string( ), substr(), upper ( ), date( ), day( ), dow( ), mdy( ), month( ), year( ), e(sample), float( ), max( ), min( ), missing( ), recode( ), sum().

➢ generate (crear, generar)

La sintaxis de la orden generate es

generate nuevavar = exp

donde nuevavar es el nombre de la variable nueva (tiene que ser un nombre nuevo, distinto al nombre de las otras variables en la base de datos) y exp es cualquier expresión válida. La orden generate puede abreviarse como *g, ge, gen*, etc. Dicha expresión puede ser una combinación de variables, operadores y funciones. Las expresiones pueden ser simples o complejas. Cuando se generan valores nulos o missing, Stata informa del número de éstos generados al generar la nueva variable. Si no se presenta el mensaje, quiere decir que no creó ningún valor nulo.

Stata results

```
. gen nacvivos=num/tasamort

. browse

. list  pais num tasamort  nacvivos, table clean
              pais       num   tasamort   nacvivos
  1.      Argentina       560      83.11   6.738058
  2.         Belice         5     113.38   .0440995
  3.         Brasil      5534        152   36.40789
  4.         Canada        26       6.52    3.98773
  5.          Chile       368     127.72   2.881303
  6.       Colombia      1367     152.68   8.953367
  7.           Cuba        87      59.23    1.46885
  8. Estados Unidos       607      14.77   41.09682
  9.      Guatemala      4206    1439.14   2.922579
 10.         Mexico      7687      264.7   29.04042
 11.           Peru      3275     525.77   6.228959
 12.    Puerto Rico        20       29.5   .6779661
 13.      Venezuela       875     166.86   5.243917
```

Al generar una variable hay que especificar que tipo de variable es, siempre y cuando ésta sea alfanumérica.

En ocasiones es posible que aparezca un mensaje de error "type mismatch" (tipo equivocado), esto ocurre porque **generate** por defecto crea variables numéricas en las cuales no se pueden guardar valores alfanuméricos. Para crear una variable alfanumérica se tiene que declarar, inmediatamente antes del nombre, el tipo y dimensión de la variable.

Cuando se usa el operador '+' con variables alfanuméricas, éstas se unen. Por ejemplo: la expresión **"esto" + "eso"** resulta en el valor alfanumérico **"estoeso".**

➢ replace (reemplazar)

Generate se usa principalmente para crear nuevas variables, sin embargo es necesario usar la orden replace para cambiar los valores de las variables que existen.

La orden replace no se puede abreviar. Por razones de seguridad Stata no deja que se abrevien órdenes que cambian datos.

Stata results

```
. replace  nacvivos= nacvivos*100000
(13 real changes made)

. list  pais num tasamort  nacvivos, table clean
              pais       num   tasamort   nacvivos
 1.      Argentina       560      83.11   673805.8
 2.         Belice         5     113.38   4409.949
 3.         Brasil      5534        152    3640790
 4.         Canada        26       6.52     398773
 5.          Chile       368     127.72   288130.3
 6.       Colombia      1367     152.68   895336.8
 7.           Cuba        87      59.23     146885
 8. Estados Unidos       607      14.77    4109682
 9.      Guatemala      4206    1439.14   292257.9
10.         Mexico      7687      264.7    2904042
11.           Peru      3275     525.77   622895.9
12.    Puerto Rico        20       29.5   67796.61
13.      Venezuela       875     166.86   524391.7
```

Borrar variables y observaciones

➢ clear, drop y keep (limpiar, borrar y retener)

Función	Orden
Borrar todos los datos de la memoria de la computadora:	clear
O:	drop_all
Borrar una variable sola:	drop varnombre
Ejemplo	drop pais
Borrar varias variables a la vez:	drop pais year
Borrar la variable varnombre$_i$ a la variable varnombre$_j$:	drop varnombre$_i$-varnombre$_j$
Ejemplo	drop tasamort-nacvivos
Borrar todas las variables que empiezan con **p:**	drop p*
Combinar:	drop tasamort-nacvivos a*
Borrar una determinadan observación en la base:	drop in # (renglón)
Borrar observaciones condicionalmente:	drop if exp
Ejemplo	drop if region==4
O combinando	drop if region==4 in 3/-3
La orden keep es parecida al drop pero tiene que especificar las variables u observaciones que quiere retener	keep if region==4 in 3/-3

Análisis exploratorio de datos

El análisis exploratorio de datos es la primera fase del análisis estadístico. Se puede realizar mediante el cálculo de diferentes estadísticos y mediante la presentación gráfica de la información. Estos procedimientos son de gran utilidad ya que permiten resumir grandes cantidades de información utilizando procedimientos estandarizados muy simples, que son accesible en casi todos los paquetes estadístico comerciales.

Como se mencionó anteriormente, las técnicas de análisis exploratorio de datos se utilizan en las primeras fases del análisis estadístico y sirven para:

a) Evaluar la calidad y consistencia de la información
b) Detectar valores "Fuera de serie "(VFS) o " no plausibles"
c) Investigar la distribución de las variables de interés
d) Investigar adherencia a las suposiciones estadísticas, que se deben cumplir en etapas posteriores del análisis estadístico
e) Resumir información mediante diferentes estadísticos y gráficos
f) Explorar formas de categorizar variables (puntos de corte)

En cualquier investigación es necesario evaluar la calidad y consistencia de la información antes de iniciar cualquier análisis estadístico. Este análisis inicial permite detectar sesgos sistemáticos, que de ignorarse, podrían ser la principal fuente de sesgos. En el campo de la investigación epidemiológica, se recolecta información sobre un gran número de variables, ya sea mediante cuestionario o con instrumentos de medición. En ocasiones se utilizan datos de fuentes secundarias que no están sujetos a controles de calidad estrictos, por lo que es conveniente realizar evaluaciones completas. Por ejemplo, cuando se obtiene información de las estaciones de monitoreo ambiental, se pueden detectar valores negativos o valores muy exagerados. La falla en detectar y corregir estos valores podría condicionar la introducción de errores importantes.

Las evaluaciones iniciales que se realizan dependen de la naturaleza de los datos obtenidos. Frecuentemente, la evaluación que se realiza es la búsqueda de valores no plausibles o valores faltantes en la escala de medición de los valores plausibles.

Existen diferentes criterios de valoración que pueden ayudar a los investigadores a tomar decisiones sobre valores que potencialmente podrían ser considerados como errores o valores aberrantes -outliers-.

En general los valores aberrantes se identifican como valores que se encuentran lejos del total de observaciones y estas se diferencian notablemente de la nube de puntos. Existen diferentes criterios y técnicas estadísticas para el tratamiento de los valores aberrantes. Sin embargo, la acción mas importante es la de identificar plenamente la fuente de error. Es muy importante poder diferenciar si se trata de una observación con plausibilidad biológica -es decir dentro del rango de observaciónes-, o de una observación no plausible, que queda fuera del rango de mediciones posibles. En el primer caso se recomienda dejar el valor observado y explorar su efecto en las etapas subsecuentes del análisis estadístico. En el segundo caso se recomienda excluir el valor, para análisis subsecuentes. En ambos casos es recomendable consultar las fuentes primarias de información para descartar la posibilidad de error.

Mediante las técnicas de análisis exploratorio de datos, es posible estudiar la distribución de la información, detectar asimetrías, rangos observados, así como los valores máximos y mínimos. La información sobre la distribución de las variables es importante, ya que muchas de las técnicas estadísticas utilizadas a menudo, asumen una serie de suposiciones sobre el comportamiento y distribución de la variables en estudio. Así por ejemplo, la regresión lineal simple considera que la variable dependiente debe estar normalmente distribuida. Cuando no se cumplen las suposiciones sobre la distribución, se puede realizar una transformación de la variable, de tal manera que la re-expresión de esta si cumple con los requisitos de normalidad. Finalmente, el análisis exploratorio de datos es importante y permite identificar re-expresiones de las variables para recategorizar o re-expresar en una escala de medición diferente. Por ejemplo en cuartiles o terciles.

Por otra parte los métodos utilizados proporcionan al investigador métodos gráficos, de fácil interpretación, que son muy útiles para la presentación gráfica de la información.

Las técnicas comúnmente utilizadas para variables continuas son:

Técnica	**Instrucción en Stata**
• Estadísticas univariadas	**summarize y summarize,detail** **tab (frecuencias)**
• Diagrama de tallo hoja	**stem**
• Diagrama de letras	**lv**
• Diagrama de caja	**graph box nomvar, medtype(line)**
• Gráfica de simetría	**symplot, qnorm**
• Normalidad	**sktest, swilk**
• Medias	**means**

Gráficos

Stata cuenta con una gran variedad de gráficos, stata (ha modificado las presentaciones de los gráficos de tal manera que puedan ser utiles para publicaciones. Las nuevas gráficas, proveen no solo flexibilidad en su apariencia, sino tambien en su contenido. Las gráficas pueden contener líneas gruesas o claras, regiones de confidencia sombreadas y otros componentes gráficos basados en y calculados de los datos. Estas se implementan en el nuevo idioma de programación orientada a objetos de Stata y eso significa que los usuarios que se sienten motivados pueden agregar esquemas nuevos estilos nuevos, tipos nuevos y características nuevas. Estas adiciones nuevas se pueden obtener y pueden ser instaladas automáticamente por medio del Internet, usando las órdenes net y update que actualmente posee Stata. Las gráficas nuevas de Stata tienen un número casi ilimitado de opciones, y la GUI de Stata provee una interfase fácil de usar para esas opciones a través de sus diálogos. Los diálogos nuevos de las gráficas permiten que cambie fácilmente los títulos, los colores, los símbolos de los marcadores, las líneas cuadriculadas, etc. sin requerir editores externos de gráficas para que la gráfica se vea como usted quiere. Las gráficas pueden ser exportadas también a otros formatos tal como PostScript y PNG (Gráfica Portátil de la Red o Portable Network Graphics).

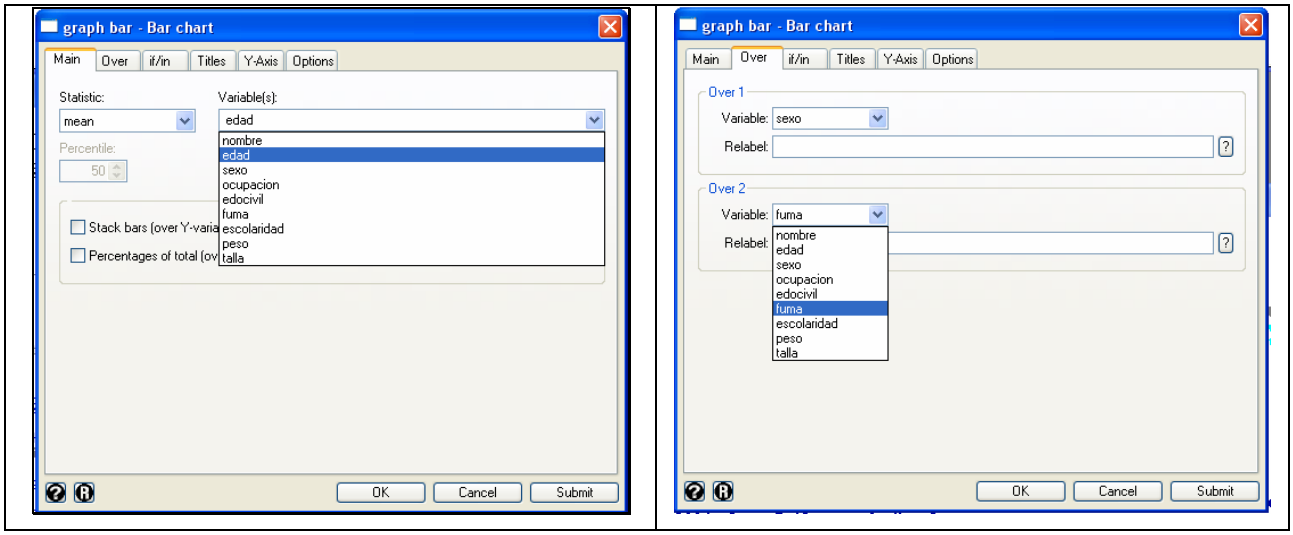
Algunos tipos de g´raficos son:

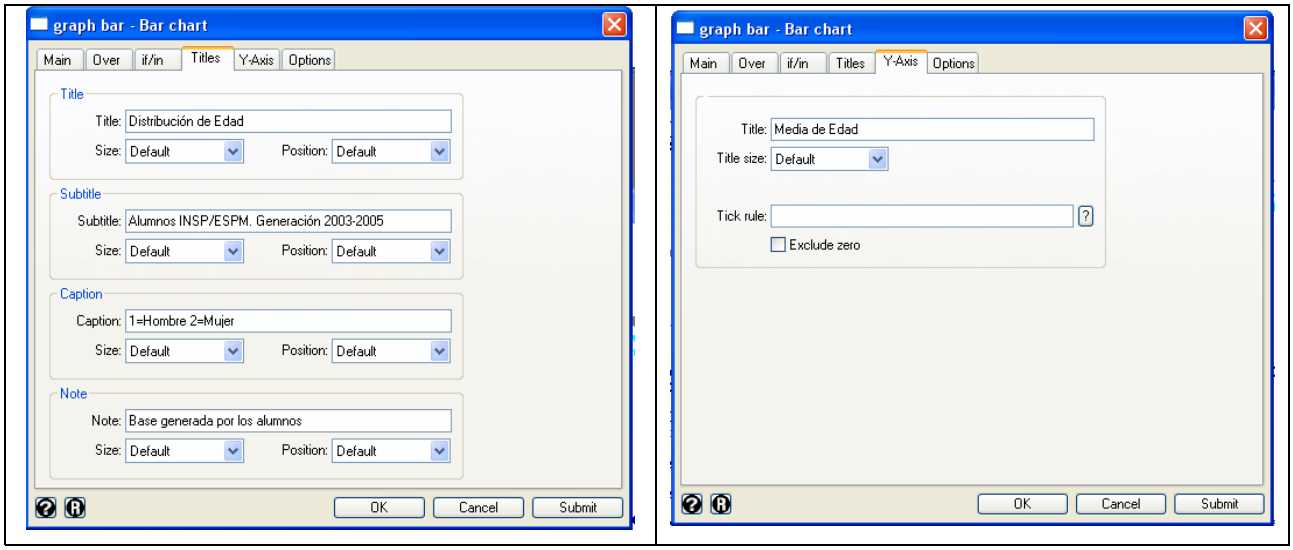
- Histogramas
- Caja
- Tallo y hoja
- Scatterplot
- Estrella
- Pastel

Gráfico de Barras:

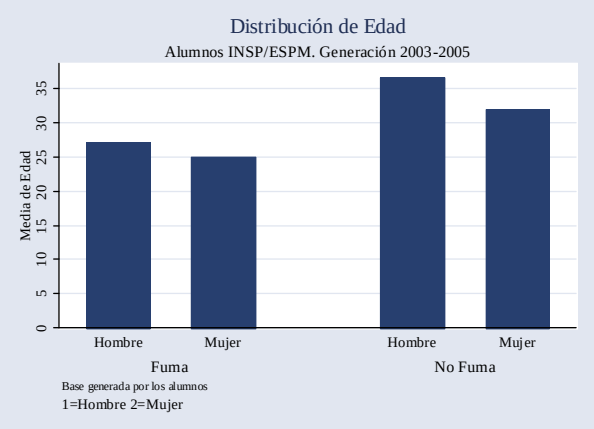
Podemos hacer los gráficos de dos formas como todas las órdenes de STATA, a partir de los menús y ventanas de diálogos, para abreviar las rutas que hay que seguir desde los menus y submenús de gráficos en el caso del siguiente gráfico de barras podemos seguir los pasos siguientes:

Entrar al menu graphs **Graphics** /Easy graphs /Bar chart/





submit o **OK**

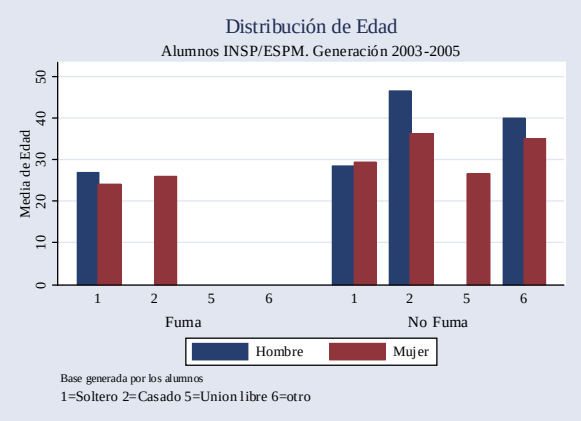


La instrucción o la orden completa para dicho gráfico sería:

.graph bar (mean) edad, over(sexo, relabel(1 "Hombre" 2 "Mujer")) over(fuma, relabel(1 "Fuma" 2 "No Fuma")) title(Distribución de Edad) subtitle(Alumnos INSP/ESPM. Generación 2003-2005) caption(1=Hombre 2=Mujer) note(Base generada por los alumnos) ytitle(Media de Edad) ylabel(#8) scheme(s2color) snack

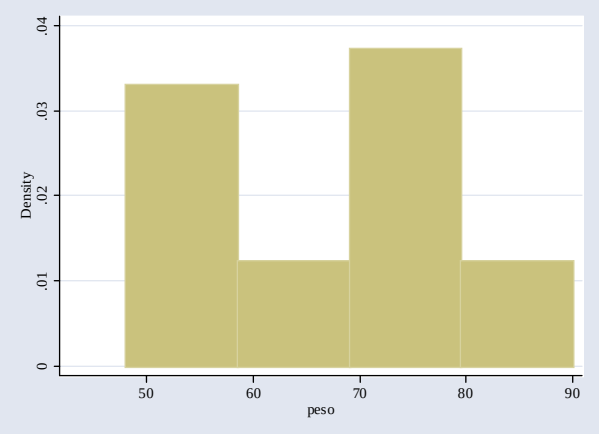
si el gráfico lo queremos hacer separando además por estado civil intercambiando con colores a los hombres y mujeres:

.**graph bar** (mean) edad, **over**(sexo, relabel(1 "Hombre" 2 "Mujer")) **over**(fuma, relabel(1 "Fuma" 2 "No Fuma")) **over**(edocivil, relabel(1 " soltero" 2 " casado" 5 " Union libre")) **title**(Distribución de Edad) **subtitle**(Alumnos INSP/ESPM. Generación 2003-2005) **caption**(1=Hombre 2=Mujer) note(Base generada por los alumnos) **ytitle**(Media de Edad) **ylabel**(#8) **scheme**(s2color)stack

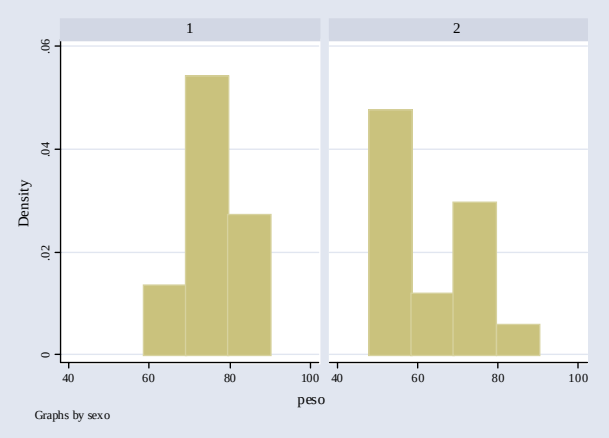


Histogramas

.histogram peso

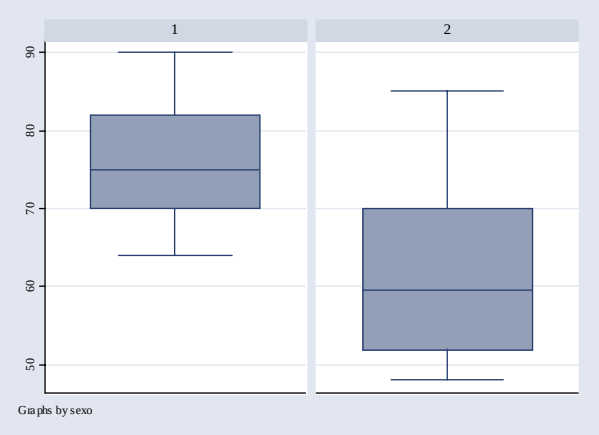


.histogram peso, by(sexo)

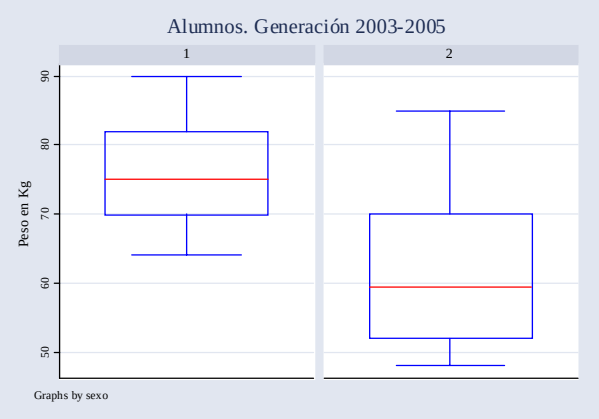


Bax plot (caja)

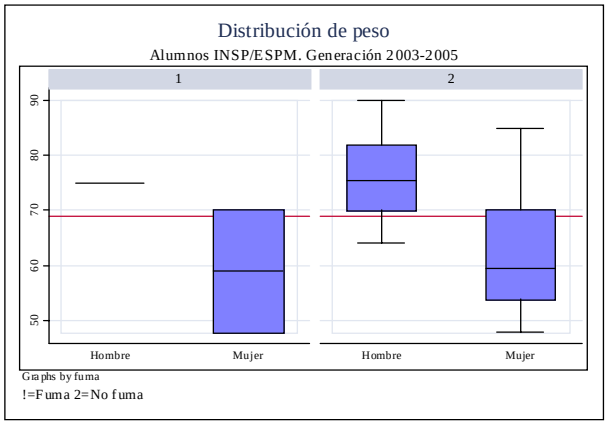
.graph box peso, medtype(line) by(sexo)



.**graph box** peso, **medtype**(cline) **medline**(lcolor(red) **lwidth**(medthick)) **by**(sexo, **title**(Alumnos. Generación 2003-2005)) **box**(1, bfcolor(none) blcolor(blue) blwidth(medthick)) **ytitle**(Peso en Kg)

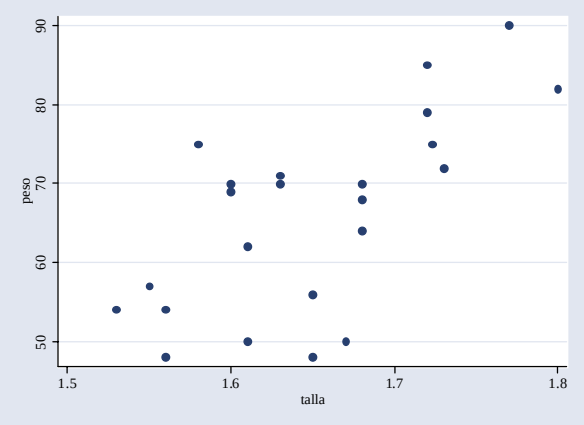


.**graph box** peso, medtype(line) **over**(sexo, relabel(1 "Hombre" 2 "Mujer")) **by**(fuma, graphregion(fcolor(white) **lcolor**(black)) **plotregion**(fcolor(white) lcolor(black)) **title**(Distribución de peso) **subtitle**(Alumnos INSP/ESPM. Generación 2003-2005) **caption**(!=Fuma 2=No fuma)) **box(**1, bfcolor(blue) blcolor(black) blwidth(medthick)) **mark**(1, msymbol(smtriangle)) **yline**(69, lwidth(medthick)) **scheme**(s2color**) plotregion**(fcolor(white) **ifcolor**(white))

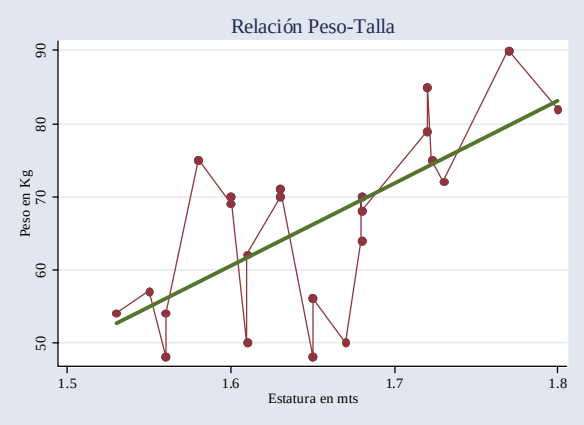


Scatterplot

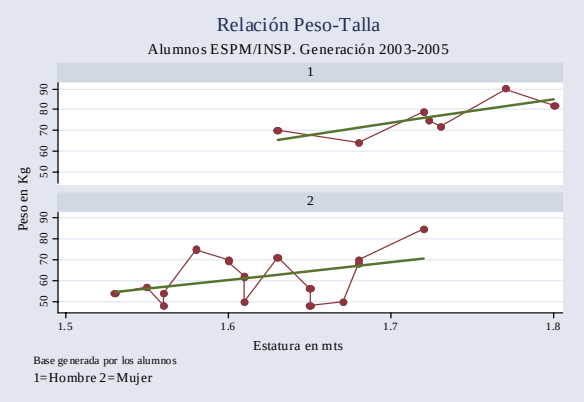
scatter peso talla



.**twoway** (**scatter** peso talla) (**connected** peso talla, sort connect(direct)) (**lfit** peso talla, sort clwidth(thick)), **ytitle**(Peso en Kg) **xtitle**(Estatura en mts) **title**(Relación Peso-Talla) **legend**(off)

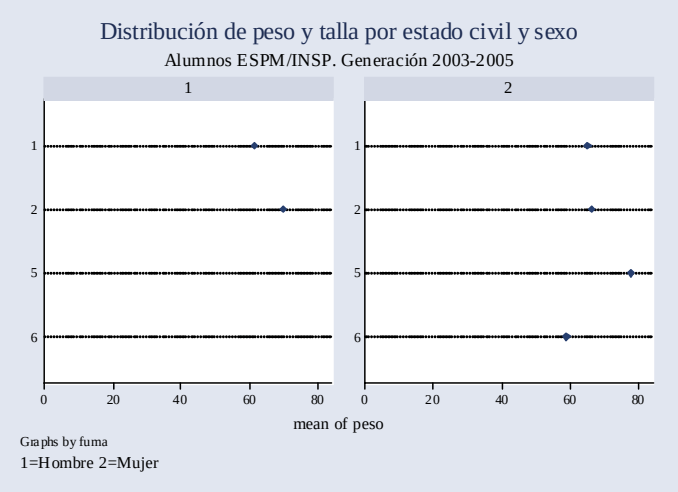


.**twoway** (**scatter** peso talla) (**connected** peso talla, sort connect(direct)) (**lfit** peso talla, sort clwidth(thick)), **by**(sexo, cols(1) **title**(Relación Peso-Talla) **subtitle**(Alumnos ESPM/INSP. Generación 2003-2005) **caption(**1=Hombre 2=Mujer) **note**(Base generada por los alumnos) **legend**(off)) **ytitle**(Peso en Kg) xtitle(Estatura en mts) **legend**(off)



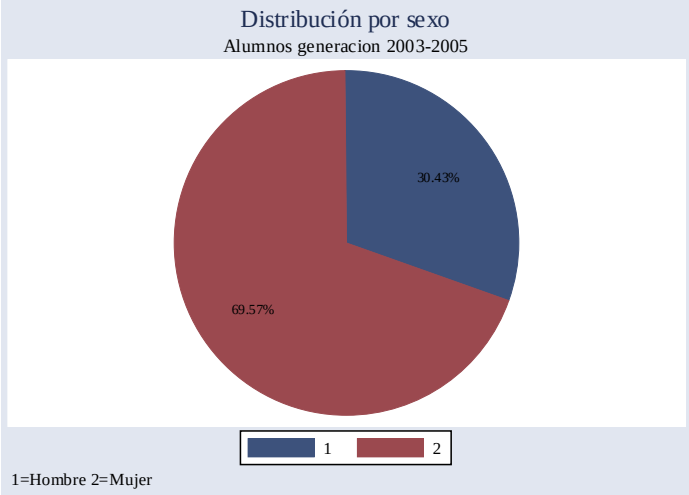
Dot plot

.**graph dot** (mean) peso, **over**(edocivil) **by**(fuma, **title**(Distribución de peso y talla por estado civil y sexo) **subtitle**(Alumnos ESPM/INSP. Generación 2003-2005) **caption**(1=Hombre 2=Mujer)) **mark**(1, **msymbol**(smdiamond) **msize**(medium)) **linetype**(dot)



Pastel

graph pie, **over**(sexo) **title**(Distribución por sexo) **subtitle**(Alumnos generacion 2003-2005) plabel(_alpercent)



Estadísticas Univariadas

Las estadísticas univariadas incluyen la media, la mediana y diferentes percentiles, moda, los valores máximos y mínimos, así como las medidas de dispersión (rango, desviación estándar), comúnmente utilizadas en estadística para resumir información.

Para ilustrar los diferentes estadísticos y gráficos utilizados se emplearán algunas bases de datos obtenidas de investigaciones epidemiológicas realizadas por investigadores del Instituto Nacional de Salud Pública.

Inicialmente se utilizará la información relativa a un estudio realizado en la Ciudad de Tapachula, Chiapas, en el que se midieron parámetros seminales en hombres y se aplicó un cuestionario de exposición a DDT y sus metabolitos. Las mediciones de parámetros seminales se encuentran en diferentes unidades, pero principalemente son porcentajes de funcionalidad.

La base de datos se encuentra en Stata, y se puede acceder a la misma mediante el programa con el comando ***use***

Stata results

```
.use a: fertil

.describe

Contains data from a:\fertil.dta
  obs:           144
 vars:            11                          30 Jul 2001 23:46
 size:         6,912 (99.9% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
folio           long   %8.0g
morf            float  %9.0g                 morfología
morfnor         float  %9.0g                 morfología normal (%)
cpdroplm        float  %9.0g                 mean morphology cpdropl
motrapi         float  %9.0g                 motilidad rápida
motprog         float  %9.0g                 motilidad progresiva
motabc          float  %9.0g                 motilidad tipo a+b+c
volumen         float  %9.0g                 volumen
densid          long   %12.0g                densidad del semen
cta_tot         long   %12.0g                cuenta total de espermas
abstin          float  %9.0g                 días de abstinencia
------------------------------------------------------------------------------
Sorted by:
```

➢ **sum**

```
. summarize volumen

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |     144    1.753125   .9404882         .1       4.65}
```

```
. sum  volumen,detail

                          volumen
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .15             .1
 5%           .6             .15
10%           .8             .3       Obs                 144
25%         1.025            .4       Sum of Wgt.         144

50%           1.5                     Mean           1.753125
                       Largest        Std. Dev.      .9404882
75%         2.275           4.25
90%             3            4.4      Variance       .8845181
95%           3.7            4.6      Skewness       .9912472
99%           4.6           4.65      Kurtosis        3.76054
```

Las estadísticas que se obtienen con la instrucción ***summ*** o ***summarize*** son de gran utilidad, ya que permiten evaluar los valores máximos y mínimos, así como los puntos de corte para los percentiles más utilizados. La **"Skewness**" y la **"Kurtosis"** proporcionan información sobre la simetría de la distribución. (para skewness el valor esperado es cero cuando la distribución es perfectamente simétrica y para la Kurtosis el valor esperado es de 3 cuando la distribución es normal).
Los percentiles son estadísticas que indican la posición de diferentes valores en relación al resto de las observaciones y estas se obtienen al ordenar las observaciones de menor a mayor.

En el ejemplo el percentil 50 o la mediana es el valor 1.5, es decir el 50% de las observaciones tienen un volumen igual o menor que 1.5 ml.

Otra manera de presentar los datos es mediante el cálculo de las medias armónica y geométrica.

- **means**

Stata results

```
. means  volumen

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+--------------------------------------------------------
     volumen | Arithmetic     144    1.753125        1.598204   1.908046
             |  Geometric     144    1.501494        1.359277   1.658591
             |   Harmonic     144    1.178716        .9842439   1.468961
-------------------------------------------------------------------------
```

La media armónica se define como:

$$\text{Media harmónica} = \frac{n}{\sum \frac{1}{xi}}$$

La media geométrica se define como:

$$\text{Media geométrica} = e^{\sum \frac{\ln(xi)}{n}}$$

Existen otros estimadores del centro de la distribución que se basan en la exclusión de cierta proporción de los valores **extremos.** Estos estimadores se conocen como ***"trimmed means"*** **o medias recortadas**

La manera de estimar las medias recortadas se puede entender fácilmente comparando la manera de estimar la media y la mediana. Para estimar la media se asume que todas las observaciones tienen un peso específico igual a 1.

De esta manera, es posible definir medias recortadas (MR), una MR (0.0) es equivalente a la media. La mediana se obtiene al eliminar **(1 -(1/(2n))** observaciones; MR (0.05) elimina el 5% de las observaciones. Para eliminar las observaciones es necesario ordenar la variable de mayor a menor y eliminar los valores extremos que corresponden al porcentaje que se requiere eliminar. Al comparar las medias con diferentes proporciones de exclusión de datos, se puede evaluar el efecto de los valores extremos sobre la media.

Stata results

```
. sum volumen

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |     144    1.753125   .9404882         .1       4.65

. sum volumen if volumen>.15 & volumen<4.6

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |     142    1.744366   .9044825        .15        4.6

. sum volumen if volumen>.6 & volumen<3.7

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |     129    1.679845   .7114718         .6       3.65

. sum volumen if volumen>.8 & volumen<3

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |     116    1.612069   .5652864         .8       2.95
```

```
. sum volumen if volumen>1.025 & volumen<2.275

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |      72      1.5875   .3291849       1.05       2.25
```

Comparando estos valores con los de la mediana(1.5), y las medias armónica (1.17) y geométrica (1.50) se puede observar como estos estimadores de la muestra son mas resistentes al efécto de los valores extremos y cómo tienden a disminuir conforme eliminamos algunas observaciones. La media recortada en el 75% es 1.58.

- **Diagrama tallo-hoja**

  stem

En su estructura más simple, se trata de una serie de números. La presentación del tipo de tallo-hoja permite explorar la estructura de los datos, mediante este gráfico se puede evaluar:

- Si la estructura es simétrica
- La dispersión
- Situación especial de algún valor
- Concentración de datos
- Valores faltantes dentro de la serie
- Patrones de dispersión y errores de dígitos

El procedimiento para construir este tipo de gráfico es muy simple y consiste en una presentación de los datos ordenados de mayor a menor. Así por ejemplo, en el caso de los datos de nuestro ejemplo de volumen:

Valores de volumen ordenados de menor a mayor y tabulados para gráfico de tallo hoja en decenas.
Cuando se realizan los diagramas de tallo-hoja a mano, la manera de calcular el número de intervalos y la amplitud de los intervalos es la siguiente: para el número de intervalo es **L=[10xlog(10)n]** y para la amplitud del intervalo se divide L entre la amplitud de valores observados en los datos. Para el caso de los datos de volumen **L=[10xlog(10)144]=21,** se estiman 21 intervalos; como la amplitud de los datos va de 0.1 a 4.65, se estima una amplitud de 5.78. Otro método para estimar el número de intervalos es **raíz de n,** en este caso sería 12.

La instrucción que se utiliza es:

Stem *variable*

```
. stem volumen
Stem-and-leaf plot for volumen (volumen)

volumen rounded to nearest multiple of .01
plot in units of .01

  0** | 10,15
  0** | 30
  0** | 40,45,55,55
  0** | 60,60,60,65,70
  0** | 80,80,80,85,85,85,85,85,90,90,90,95
  1** | 00,00,00,00,00,00,00,00,00,00,05,10,10,10,10,15,15,15
  1** | 20,25,25,25,25,30,30,30,30,30,30,30,35,35,35,35
  1** | 40,40,40,45,45,45,45,45,45,50,50,50,50,50
  1** | 60,60,65,65,70,70,75,75,75
  1** | 80,80,80,80,80,80,80,85,85,85,95,95,95
  2** | 00,00,00,00,05,10,10,15,15,15
  2** | 20,25,30,30,30,35,35
  2** | 40,45,45,45,50,50,50,50,55
  2** | 65,65,70
  2** | 80,90,95
  3** | 00,00,05
  3** | 35
  3** | 40,50
  3** | 60,65,70
  3** | 80,85
  4** | 00
  4** | 25
  4** | 40
  4** | 60,65
```

Don de por ejemplo:

0** | 10,15 *corresponde a los valores mínimos: valor de 0.10 y otro de 0.15*

0** | 30 *corresponde una observación con valor de 0.30*

1** | 00,00,00,00,00,00,00,00,00,00,05,10,10,10,10,15,15,15

4** | 60,65 *corresponden a los valores máximos uno de 4.60 y otro de 4.65*

En la gráfica del ejemplo podemos observar que la distribución de las frecuencias está un poco sesgada hacia uno de los lados, es decir hacia la izquierda, lo cual ocasiona que la distribución de la misma no sea normal.

- **Gráfico de letras (lv)**

Al igual que el gráfico de tallo-hoja, el diagrama de letras se basa principalmente en el ordenamiento de los datos, **de menor** a mayor, y en el cálculo de diferentes estadísticos que evalúan el impacto de los extremos de la distribución, "de las colas", de los datos, asumiendo diferentes puntos de corte. El nombre de diagrama de letras se origina en el hecho de que a cada punto de corte se le ha asignado una letra.

El procedimiento para obtener los estadísticos de diagrama de letras, consiste en ordenar los datos -de menor a mayor- y en extraer información sobre los valores que definen el punto medio (la mediana), los que definen los cuartos, es decir los percentiles 25 y 75; los octavos con los percentiles 12.5 y 87.5, los y

dieciseisavos, los treintadosavos, y así sucesivamente.

				Punto de corte en %	
Fracción de corte	**Símbolo**	**%**	**Fracción**	**Inferior**	**Superior**
Mediana	M	0.5	1/2	50.0	50.0
Cuartiles	F	0.25	1/4	25.0	75.0
Octiles	E	0.125	1/8	12.5	87.5
Dieciseisciles	D	0.0625	1/16	6.25	93.75
Treintaidosciles	C	0.03125	1/32	3.125	96.87
Sesentaicuatrosciles	B	0.01562	1/64	1.56	98.44
Cientoveintiochoavos	A	0.00781	1/128	0.78	99.22

Como ya se mencionó, a cada punto de corte se le ha asignado una letra, esta asignación es arbitraria, es decir no sigue un orden particular, pero es la que se usa convencionalmente en la representación gráfica.

A continuación se examinará el diagrama de letras para una de las variables de estudio:

Stata results

```
. lv morf

 #    139               morfología
             --------------------------------
 M     70   |               1.435            |    spread  pseudosigma
 F     35.5 |      1.37     1.4625      1.555 |  .1849999    .1386847
 E     18   |      1.29       1.46       1.63 |       .34    .1489038
 D      9.5 |     1.245    1.54625     1.8475 |     .6025    .1997806
 C      5   |      1.21    20.2475     39.285 |    38.075    10.39295
 B      3   |       1.2       20.3       39.4 |      38.2    9.218723
 A      2   |       1.2      20.32      39.44 |     38.24    8.466334
 Z      1.5 |       1.2      29.71      58.22 |     57.02    11.91776
        1   |       1.2       39.1         77 |      75.8    14.70818
             |                                |
             |                                |   # below     # above
inner fence |    1.0925               1.8325 |         0          10
outer fence |  .8150002                 2.11 |         0           7
```

La primera línea **# 139 morfología** muestra el número de observaciones y la etiqueta de la variable.

La segunda línea, **M 70 | 1.435**, contiene información sobre la mediana y el número de observaciones que se encuentran por debajo de la mediana. En este caso la mediana es de 1.435 y separa 70 observaciones. En la segunda línea aparecen las estadísticas asociadas con los cuartos, lo que corresponde a la letra F. El 1.37 y 1.555 marcan los valores límite para el cuartil inferior (percentil 25) y el cuartil superior (percentil 75). La cifra de 35.5 indica que, por debajo y por arriba de estos puntos de corte, quedan 165 observaciones (17.25 en cada extremo). El valor 1.4625 indica el punto medio de las observaciones que quedan entre los puntos de corte inferior y superior, en este caso (1.37+1.555)/2.

Si la distribución fuese perfectamente simétrica, se esperaría que los punto medios fueran iguales a la mediana. El ***"spread"*** o dispersión, se obtiene al calcular la diferencia entre el valor del límite superior y el

inferior, en este caso 1.555 -1.37. La ***pseudosigma*** es una estimación de la desviación estándar, -para el cálculo se asume que la variable se distribuye normalmente- utilizando los valores que quedaron en los extremos de cada punto de corte. Si la variable tiene una distribución normal, los valores para los diferentes puntos de corte deben ser similares. En la interpretación de los valores de la pseudosigma se puede inferir lo siguiente: a) si se observan valores decrecientes, se puede concluir que tiene menor dispersión que la distribución normal; b) si se incrementa ello indicaría mayor dispersión; ambos comportamientos indican asimetrías en la distribución.

En la parte inferior del diagrama se presenta información sobre los valores que se encuentran separados de la nube de puntos. Es importante detectar estos valores, ya que dentro del análisis estadístico ameritan atención especial puesto que pueden tener un impacto importante sobre los resultados y conclusiones. Como ya se mencionó, estos valores pueden deberse a errores reales, en cuyo caso deben corregirse o excluirse del análisis, o a valores reales, con cierta plausibilidad, en cuyo caso deben incluirse en el análisis y evaluarse en términos del impacto que tienen sobre los resultados y conclusiones. Una alternativa es excluirlos de análisis final y evaluar la diferencia en los resultados.

Como convención, se definen dos puntos de corte y se cuenta el número de observaciones que quedan dentro de ellos; éstas observaciones merecen atención especial.

La información se presenta en dos categorías que marcan lejanía hacia la nube de puntos. En general, se manejan dos puntos de corte basados en el rango intercuartil. Los puntos de corte se definen como **límite interno,** que identifica los puntos que podrían ser considerados como valores aberrantes o "outliers" y el **limite externo,** que identifica los valores con una alta probabilidad de ser aberrantes. Si las observaciones se originaran de una distribución normal, los valores para el límite interno equivaldrían a -2.698 σ y a +2.698 σ, y para los límites externos a -4.721 σ y a +4.721 σ.

Se utiliza el valor del rango intercuartil dado que es una medida robusta que no se afecta por la presencia de valores extremos, a diferencia de la desviación estándar o la dispersión (rango). Los límites interno y externo se definen de la siguiente manera:

Diferencia intercuartil *DI = C75 - C25*
Limite interno inferior *LIi = C25 - 1.5x DI*
Limite interno superior *LIs = C75 + 1.5 x DI*
Limite externo inferior *LEi = C25 - 3.0 x DI*
Limite externo superior *LEs = C75 + 3.0 x DI*

Para identificar las observaciones se puede realizar un "list", estableciendo los puntos de corte calculados para los valores de los puntos de corte. En el ejemplo anterior:

Stata results

```
. list  folio morf if morf>=1.8325, table clear

        folio       morf
 11.      122      39.22
 26.      109          .
 36.      124          .
 45.      127          .
 52.       53       1.91
 59.       75       39.4
 62.       14          .
 75.      132     39.165
 84.       33      39.34
101.       78      39.44
102.       34     39.285
118.       84          .
125.       24         77
131.       51       1.84
141.       58      1.855
```

si existen otras variables con las cuales podamos comparar estos valores, es decir con los cuales la morfología se pudiera correlacionar, sería adecuado analizarlos y evaluar si esos datos que en la variable de morfología son altos, en la otra variable también son altos.

Es importante tomar nota y evaluar el impacto de estas observaciones en las fases subsecuentes del análisis.

- **Gráfico de caja (boxplots)**

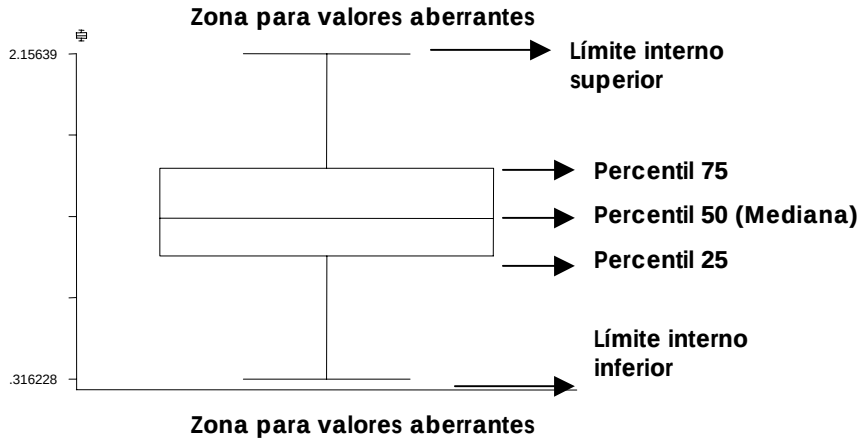
  Graph box variable

Este tipo de gráfico es una representación simple de la información, que indica:

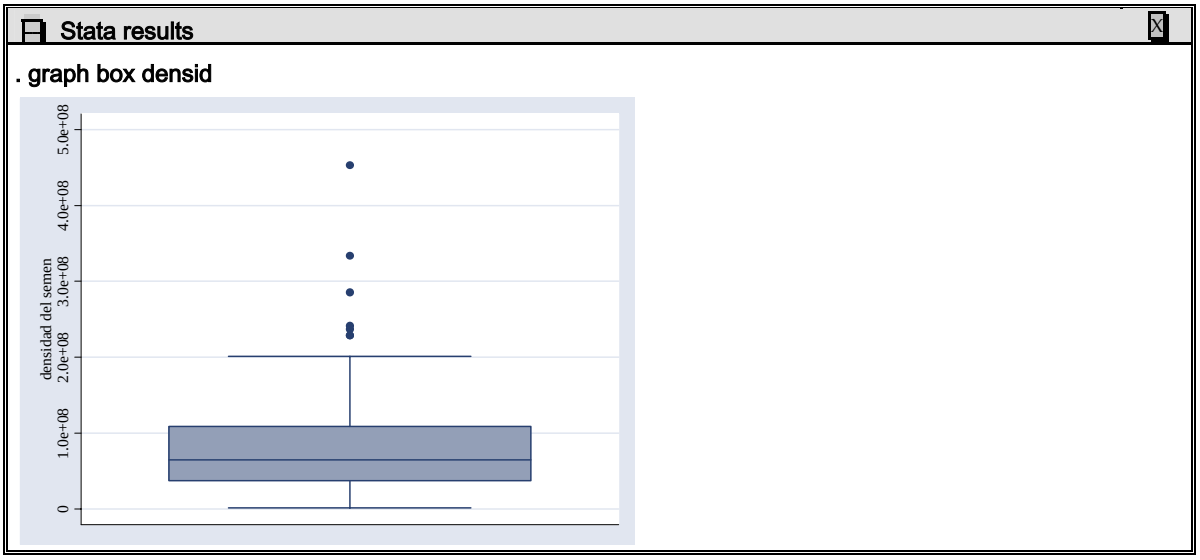
1. la localización del centro de los datos
2. la dispersión
3. la simetría
4. la extensión de los extremos (colas de la distribución)
5. la existencia de valores aberrantes (outliers)

La sencillez de este gráfico lo convierte en un buen instrumento para realizar comparaciones entre diferentes categorías, por ejemplo, entre densidad de la muestra de semen en los hombres del estudio de Tapachula, Chiapas, por días de abstinencia.

Estructura del diagrama de caja:



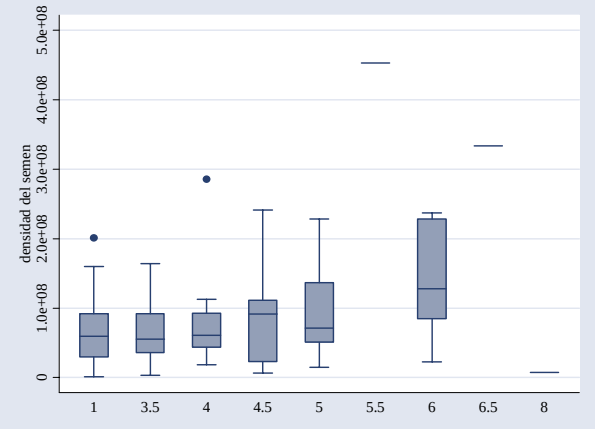
La ventaja del diagrama de caja, basado en los rangos intercuartiles, es que es resistente al impacto de valores extremos. De hecho, podrían presentarse valores extremos en el 25% de las observaciones y no tener un impacto importante sobre los límites de la caja. En relación con los límites para detectar valores aberrantes, éstos se definen de manera arbitraria. Si se aplicaran a una distribución normal, se esperaría que únicamente el 0.7% de las observaciones tomarán valores superiores a estos punto de corte.



Al graficar la información sobre la densidad de las muestras podemos observar asimetría en los datos, con algunos valores aberrantes, esta información concuerda con la información que nos proporciona un diagrama de letras de la misma variable.

```
. lv  densid
 #     139                densid
              ---------------------------------
 M      70   |               648.25             |    spread  pseudosigma
 F      35.5 |    365.75   724.4375   1083.125 |   717.375    537.7782
 E      18   |     197.5        815     1432.5 |      1235    540.8713
 D       9.5 |       100   967.1875   1834.375 |  1734.375    575.0944
 C       5   |        27     1198.5        2370 |      2343    639.5452
 B       3   |        23       1439        2855 |      2832    683.4404
 A       2   |      20.5       1679      3337.5 |      3317    734.3836
 Z       1.5 |     15.75    1974.75     3933.75 |      3918    818.9017
         1   |        11     2270.5        4530 |      4519    876.8635
             |                                  |
             |                                  |   # below     # above
inner fence | -710.3125              2159.188   |         0           7
outer fence | -1786.375               3235.25   |         0           2
```

. graph box densid,over(abstin)



Es de utilidad poder tener el gráfico de caja para comparar la distribución de los valores observados (en este caso se graficaron los valores observados en densidad por días de abstinencia).

En este gráfico se pueden observar diferencias entre los días de abstinencia y la densidad, al parecer los días de abstinencia son un factor para que aumente la densidad de la muestra de semen, a mayor días de abstinencia mayor es la densidad. Y a mayor días de abstinencia mayor dispersión de los datos. Este patrón podría sugerir la necesidad de una transformación, es decir, de re-expresar los valores observados para lograr una dispersión similar, logrando una mejor representación gráfica y datos mas apropiados para los análisis estadísticos tradicionales, como el de varianza y la regresión lineal. En el análisis de varianza se hace la suposición sobre igualdad de varianzas dentro de los diferentes grupos de comparación.

- **Normalidad y Transformaciónes**

Transformación de variables.

Una de las aplicaciones del análisis exploratorio de datos, es la evaluación de la necesidad de realizar transformaciones. Las principales razones para realizar transformaciones son:

a) Normalizar las distribuciones
b) Ganar interpretabilidad
c) Corregir asimetrías fuertes
d) Categorías con dispersiones diferentes
e) Residuales influyentes (detectados en regresión lineal)

Las transformaciones más frecuentemente usadas son:

$$T_p(x) = ax^p + b \quad \text{cuando } p \neq 0$$
$$T_p(x) = c\log + d \quad \text{cuando } p=0$$

Se trata de transformaciones fuertes y, en general, cambian la forma de los datos; forman parte de un grupo conocido como transformaciones de potencia, que tienen la siguiente forma:

$$Tp(x) = ax^p + b \quad \text{cuando } p \neq 0$$
$$Tp(x) = c\log + d \quad \text{cuando } p=0$$

Se requiere que a, b, c, d y p sean números reales; y que a>0 para p>0 y a<0 para p<0. Con estas condiciones se asegura lo siguiente:

a) Se conserva la secuencia original de orden en los datos
b) Se conservan los valores asociados a las letras, en el diagrama de letras.
c) Son funciones continuas
d) Son funciones sin variaciones bruscas
e) Se utilizan transformaciones simples, que pueden re-expresarse sin dificultad

Las transformaciones llevan la información a escalas que no resultan familiares por lo que, en general, se pierde interpretación. Los problemas surgen principalmente en el área de la interpretación y no tanto en la de análisis. Por las razones anteriores, solo se deben transformar los datos cuando:

a) Existe una dispersión muy amplia en los datos. Si la relación entre el valor menor y el mayor es superior a 20, es probable que la transformación tenga un buen efecto.
b) Se encuentran residuales con valores grandes

c) Existen asimetrías importantes

Entre los usos que se pueden hacer de las transformaciones, está el de lograr "normalidad", es decir, que los datos se distribuyan de acuerdo con la distribución normal. Para evaluar en forma inicial si las observaciones se apegan a esta distribución, se mencionaron anteriormente los resultados que se obtienen del diagrama de letras. En este gráfico, si la distribución se apega a la normalidad, se esperaría que los valores de la pseudosigma fuesen constantes en las estimaciones asociadas a las diferentes letras.

A continuación se presenta el diagrama de letras asociado a los datos de espermatobioscopías en los hombres del estudio de Tapachula, Chiapas para la variable **volumen**.

Stata results

```
. lv volumen

 #    144                 volumen
             --------------------------------
 M     72.5 |                  1.5             |    spread  pseudosigma
 F     36.5 |     1.025       1.65       2.275 |      1.25    .9291277
 E     18.5 |       .85       1.85        2.85 |         2    .8724843
 D      9.5 |        .6     2.1125       3.625 |     3.025    .9911818
 C      5   |       .45      2.225           4 |      3.55    .9607234
 B      3   |        .3       2.35         4.4 |       4.1     .982601
 A      2   |       .15      2.375         4.6 |      4.45    .9793717
 Z      1.5 |      .125      2.375       4.625 |       4.5    .9354966
        1   |        .1      2.375        4.65 |      4.55    .8787322
            |                                  |
            |                                  |   # below     # above
inner fence |      -.85                 4.15 |         0           4
outer fence |    -2.725                6.025 |         0           0
```

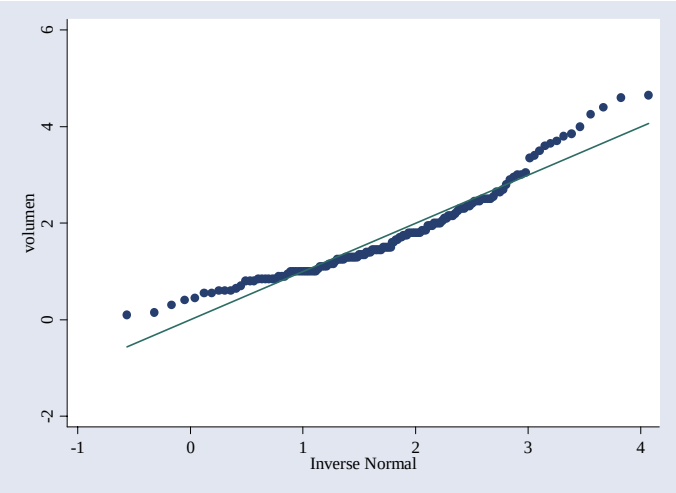
Se puede apreciar que la pseudosigma varía de, lo .9291277 a .8787322 lo que sugiere que no se apega a una distribución normal.

Existen otros métodos para evaluar la normalidad; probablemente el más utilizado es el gráfico de la variable original, en relación a su transformación como una variable normalizada. De este gráfico se puede obtener información sobre la falta de normalidad y se puede construir graficando la variable original (y) versus la variable transformada **(f[(X_i - μ]/σ)**.

- **qnorm nor, title("gráfico de normalidad)**

Stata results

.qnorm volumen



- **symplot**

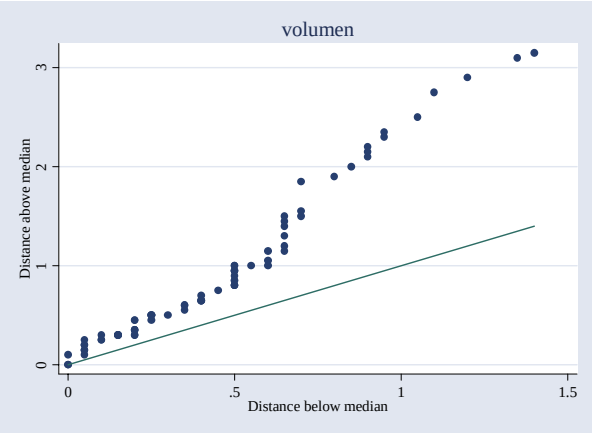
Existen otros gráficos de simetría que pueden ser utilizados. La distancia que tiene cada observación de la mediana se ha utilizado como un indicador de simetría. Si la distribución es simétrica se esperaría que los datos se comportaran de manera similar en ambos extremos de la distribución.

Para realizar este gráfico debemos calcular la diferencia entre la mediana y el valor observado. Como valores esperados podemos graficar el valor observado vs. el mismo valor observado. Si la distribución es simétrica todos los valores deben quedar por debajo del valor esperado.

Posición	volumen observado	mediana observada	diferencia
1.	.1	1.5	1.4
2.	.15	1.5	1.35
3.	.3	1.5	1.2
4.	.4	1.5	1.1
5.	.45	1.5	1.05
140.	4	1.5	2.5
141.	4.25	1.5	2.75
142.	4.4	1.5	2.9
143.	4.6	1.5	3.1
144.	4.65	1.5	3.15

Stata results

```
. symplot volumen
```



Los puntos que se grafcican son:

$$\text{mediana-}y \text{ vs } y_{i(\_N+1-1)}$$

Si la distribución es simpetrica la distancia entre los puntos que se encuentran por debajo de la mediana es igual a la distancia de los puntos que se encuentran por arriba. La líne sólida refleja el valor esperado.

Otra forma de evaluar normalidad de los datos es mediante pruebas estadísticas de ajuste. En este caso se asume que la distribución es normal y se estima la probabilidad de que los valores observados se deriven de una distribución normal. Este procedimiento tiene la desventaja de que el resultado dependerá del tamaño de muestra. Para muestras grandes, diferencias pequeñas son altamente significativas, para muestras pequeñas diferencias importantes pueden pasar desapercibidas.

- **Sktest**

Un comando para realizar esta prueba es el sktest, esta prueba se basa en la kurtosis (curvatura) y la skewness(simetría) de la variable.

Para las variables de las base de fertil, se obtienen los siguientes valores:

Stata results

```
. sktest  morf morfnor motrapi motprog motabc volumen densid cta_tot
                  Skewness/Kurtosis tests for Normality
                                                  ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
        morf |      0.000         0.000               .       0.0000
     morfnor |      0.000         0.000               .       0.0000
     motrapi |      0.000         0.000           34.14       0.0000
     motprog |      0.000         0.015           22.68       0.0000
      motabc |      0.000         0.028           20.85       0.0000
     volumen |      0.000         0.075           17.86       0.0001
      densid |      0.000         0.000           57.76       0.0000
     cta_tot |      0.000         0.000           33.07       0.0000
```

En este caso nosotros rechazamos la hipótesis nula para todas las variables, ninguna de ellas se distribuye normalemente.

- **swilk**

Otro estadístico para determinar la normalidad de los datos es la prueba de Shapiro –Wilk. En Stata la instrucción es swilk.

Del mismo ejemplo anterior aplicando esta prueba tenemos:

Stata results

```
. swilk  morf morfnor motrapi motprog motabc volumen densid cta_tot
                  Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
        morf |    139    0.23472     83.430      9.989  0.00000
     morfnor |    136    0.75502     26.215      7.367  0.00000
     motrapi |    118    0.86906     12.422      5.640  0.00000
     motprog |    139    0.91363      9.416      5.063  0.00000
      motabc |    139    0.91611      9.145      4.997  0.00000
     volumen |    144    0.93266      7.566      4.578  0.00000
      densid |    139    0.83553     17.930      6.518  0.00000
     cta_tot |    139    0.86934     14.244      5.998  0.00000
```

En este caso, se puede observar que para todas las variables se rechaza la hipótesis de que se ajustan a una distribución normal. Tomando en cuenta que el valor esperado para el estadístico V es de 1.0 se puede observar que la variable morf presenta los valores más extremos y que la variable volumen se acerca más a una distribución normal.

- **Ladder**

Otra manera de encontrar la mejor re-expresión de la variable para normalizarla (corregir simetría) es ensayar diferentes transformaciones y evaluar cual se ajusta mejor a la distribución normal. Stata puede hacer transformaciones a diferentes potencias mediante el comando ladder.

Aplicando este comando a una de las variables de nuestra base de datos fértil:

Stata results

```
. ladder volumen
Transformation         formula               chi2(2)       P(chi2)
-----------------------------------------------------------------
cube                   volumen^3                   .        0.000
square                 volumen^2               53.60        0.000
raw                    volumen                 17.86        0.000
square-root            sqrt(volumen)            1.76        0.415
log                    log(volumen)            27.85        0.000
reciprocal root        1/sqrt(volumen)             .        0.000
reciprocal             1/volumen                   .        0.000
reciprocal square      1/(volumen^2)               .        0.000
reciprocal cube        1/(volumen^3)               .        0.000
```

Vemos que la transformación mas adecuada que normaliza la variable *volumen* es la raíz cuadrada.

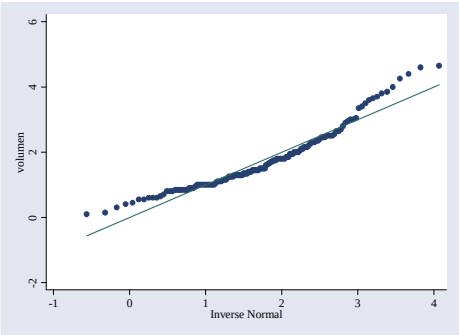
Entonces debemos generar una variable utilizando una función que es raíz cuadrada (sqrt) sugerida por el comando anterior.

Stata results

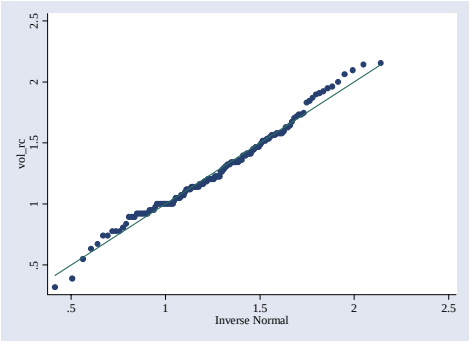
```
. sum volumen
    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     volumen |     144    1.753125   .9404882         .1       4.65

. gen vol_rc=sqrt(volumen)

. label var vol_rc "Transformación raíz cuadrada de volumen"

. sum vol_rc
    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      vol_rc |     144    1.277159    .350489   .3162278   2.156386
```

si graficamos la variable las dos variables por medio de barras de frecuencias tenemos que:

. qnorm volumen



.qnorm vol_rc



Podemos observar cómo la transformación mejora sustancialmente la distribución de la variable.

Algo que podemos concluir de las transformaciones es que:

Se gana simetría.

Se pierde "interpretabilidad"

Si la media > mediana	Desviación positiva
Si la media = mediana	Simétrica
Si la media < mediana	Desviación negativa
Cubo: ^3	Reduce asimetría negativa muy fuerte
Cuadrado ^2	Reduce asimetría negativa leve
Raíz cuadrada	Reduce asimetría positiva leve moderada
Logaritmo	Reduce asimetría positiva

Introducción al Modelamiento estadístico

El modelamiento estadístico generalmente es consecuencia de un proyecto en el cual, con anterioridad, se ha planteado una pregunta de investigación y en la cual se pretende buscar una asociación o bien una predicicón.

Este tiene como objetivos principales: determinar la existencia y la magnitud de la asociaicón entre una variable de respuesta con uno o mas factores (variables de exposición), controlando por variables exógenas (variables de control) y/o determinar que factores (variables predictoras) son las que mejor predicen una respuesta.

La evaluación de la respuesta en los estudios epidemiológicos están muy comunmente relacionados con un proceso de Salud-enfermedad y estos difieren de acuerdo al tipo de diseño empleado:

1. Prevalencia
2. Incidencia (densidad de incidencia)
3. Riesgo (Razón de incidencias)
4. Probabilidad de sobrevida
5. Riesgo instantáneo
6. Razones de momios
7. Razones de prevalencia

La base de toda investigación epidemiológica antes que cualquier método de análisis estadístico, es el diseño de investigación con el cual se recaba la información. Al mismo tiempo que estos determinan el tipo de análisis a realizar y el método estadístico mas apropiado. En los estudios transversales por ejemplo, es común utilizar un análisis de prevalencias aunque también, se pueden obtener Razones de Momios utilizando una regresión logística o razones de prevalencia. Los estudios de Casos y Controles que son los diseños mas comune para evaluar factores de riesgo sobre la probabilidad de presentar o no una enfermedad determinada se utiliza también regresión logística sobre la cual se pueden obtener Razones de Momios.

Por otro lado, en los estudios de cohorte y ensayos clínicos, puede ser posible determinar desde Riesgos de incidencia, razones de riesgos (Riesgos Relativos), tasas de incidencia, análisis estratificado, curvas de sobrevida, utilizando el análisis estadístico apropiado: regresión Poisson, regresión logísitica, Survas de Sobrevida , regresión de Cox, medidas repetidas, etc.

Indroducción al análisis comparativo bivariado y multivariado en STATA

La estadística representan una herramienta muy importante para comprender los fenómenos biológicos, y nos permiten:

1) Comunicar y describir información en forma estandarizada
2) Contestar hipótesis
2) Modelar y cuantificar diferentes relaciones entre parámetros.

Sin embargo, es muy importante recordar que su aplicación se basa en una sobre simplificación de los fenómenos biológicos y una serie de suposiciones, sobre el comportamiento de las variables en las que se ha operacionalizado la medición de los fenómenos biológicos.

Análisis bivariado

El análisis bivariado consta de diferentes pruebas para encontrar la asociación entre dos variables simples, la elección de la prueba estadística va a depender del tipo de variable que se examine, es decir, la escala de medición tanto de la variable dependiente como de la independiente, así como de su distribución.

- **Tab var1 var2, column all exact**

Esta opción del comando tab despliega una tabla de 2 x 2 mostrando además las proporciones por columna para cada una de las categorías. La opción "all exact" es equivalente a especificar "chi2 lrchi2 V gamma taub". S incluyendo prueba exacta de Fisher's. Con la prueba de chi2 podemos evaluar la diferencia de proporciones.

- **Tablas cc para OR**

Esta prueba en STATA se utiliza para evaluar la asociación entre dos variables categóricas (variable que indica caso o no caso y la variable de expuesto o no expuesto), las cuales se pueden graficar en una tabla de 2 x 2. Con ello calcula Razones de Momios y sus intervalos de confianza, además de las fracción atribuible o prevenible entre los expuestos y la fracción atribuible o prevenible poblacional.

Razón de momios instantáneas cci. Puede utilizarse para calcular el OR conociendo el valor de las celdas.

Cc var1 varr2,, by(var3) permite probar diferencias entre los OR calculados entre estratos utilizando medias ponderadas. El estadístico utilizado para dicha prueba es la de Mantel –Hanzel.

- **Sdtest**

Esta prueba se utiliza para comparar las varianzas entre dos grupos o categorías (varible continua y una dicotómica). La hipótesisi nula para este estadístico es probar que las varianzas entre ambas categorías son iguales, mediante una prueba de significancia: Valor P.

- **ttest**

El comando ttest se utiliza para probar la hipótesis nula de que las medias de distribución entre dos grupos son iguales. Al igual que la prueba de diferencia de varianzas, la prueba de diferencia de medias requiere una variable categórica (dicotómica) y una variable continua, dicha variable se espera que tenga una distribución normal entre ambos grupos, que su varianza sea homogénea y que entre las observaciones haya independencia.
Ttest prueba t de student se emplea para muestras pequeñas

$$t= \frac{X-u}{SX}$$

- **ANOVA**

Análisis de varianza, prueba la hipótesis nula de que no hay diferencias entre los grupos contra la hipótesis alterna de que al menos un grupo es diferente. Esta prueba requiere de varios supuestos para su uso:
Las muestras se hayan seleccionado aleatoriamente, que la variable dependiente se distribuya como una variable normal en cada uno de los grupos y que la varianza de la misma sea constante en cada grupo.
La prueba ANOVA es una generalización de la prueba t para comparar dos muestras independientes.

$$\}SST = (k\text{-}1)MST= \sum_{i=1}^{k} n_i(Y_i - \bar{Y})^2$$

$$SST = (n\text{-}k)MSE= \sum_{i=1}^{k} \sum_{j=1}^{nk} (Y_{ij} - \bar{Y}i)^2$$

La prueba de bonferroni se aplica cuando hay diferencias de medias entre los grupos y su objetivo es establecer la diferencia específica entre grupos y el nivel de significancia.

- **Kwallis**

Prueba la hipótesis de que dos o más muestras probienen de una misma población. Se utiliza para pruebas en las cuales la distribución de la población es no paramétrica, es decir no requiere que las poblaciones estudiadas estén normalmente distribuidas. La prueba de Kruskal-Wallis es una generalización de la

prueba de rangos de signos de Wilcoxon para dos muestras (llamada también de Mann-Whitney). Las muestras de tamaño n_j j=1,....,m se combinan en rangos en orden ascendente de magnitud, a cada rango se le asigna su promedio.

➢

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{m} \frac{R^2 j}{nj} - 3(n+1)$$

En la fórmula *n* denota el total del tamaño de la muestra y *Rj* la suma de rangos para cada muestra *jth.*
La distribución de la muestra H es aproximadamente X^2 con m-1 grados de libertad.

- **Correlate x1 x2 x3**

Esta prueba pretende encontrar la correlación entre dos variables. El estimador puntual que utiliza son las medias y determina los coeficientes de correlación entre ellos. La hipótesis nula para esta prueba es que las variables no están correlacionadas.
Corr despliega una matriz de correlación de Pearson usando solamente observaciones con valores no missing sobre todas las variables especificadas. Adicionando la opción covarianza produce una matriz de varianza-covarianza proveniente de la correlación

- **pwcorr x1 x2 x3 y, sig**

Despliega una matriz de correlaciones de Pearson usando parejas y deleción de valores missing y mostrando probabilidades de t test (de Ho:p = 0) sobre cada correlación.

- **Spearman x1 x2**

Correlación de rangos que se calcula como la correlación de Pearson sólo que estimada sobre sobre los rangos y promedios en cada rango, además calcula la significancia de la correlación. Asume que la variable 1 y la variable 2 son independientes.

- **Gráficas de dispersión**

Muestra la tendencia de la correlación entre dos variables continuas.

Modelos de Regresión:

El análisis de regresión lineal es una herramienta más para el análisis estadístico entre las asociaciones de parámetros, la regresión lineal en Stata ofrece un amplio rango de procedimientos, desde elementales a sofisticados, desde los comandos que realizan regresiones ordinarias de mínimos cuadrados simples y múltiples (OLS) hasta las órdenes que calculan valores predichos, residuos, y estadísticas de diagnóstico como datos influyentes y Cooles D.

Ejemplos de Comandos

Orden	**Función**
regress yx	Estima la ecuación de la regresión de mínimos cuadrados entre la variable y (variable dependiente y la variable X (variable independiente
regress yx if var1 == 3 & var2 > 50	Obtiene la regresión estratificando por loa variable 2 cuando esta sea mayor que 50 y si var1==3
predict yhat	Genera una nueva variable la cual arbitrariamente la nombra como yhat igual al valor predicho de la última regresión
predict e, resid	Genera una nueva variable (Nombrada arbitrariamente *e*, igual a los residuos de la regresión mas reciente.
graph y x, || line yhat x **o** **twoway (lfit y x)**	Dibuja un scatterplot (gráfica de puntos) con la línea de regresión usando la variable *y*, yhat, y *x*
scatter e yhat, twoway box yline (0)	Dibuja una gráfica de los residuos contra los valores predichos usando la variable *e* y yhat.
regress y x_1 x_2 x_3	Estima una regresión lineal múltiple con tres predictores x_1 x_2 y x_3.
regress y x_1 x_2 x_3, robust	Calcula estimados robustos de errors estándar (Huber/White).
regress y x_1 x_2 x_3, beta	Estima una regresión múltiple y muestra los coeficientes de la regresión en forma estanadarizada (coeficientes) sobre una tabla de resultados.
correlate x_1 x_2 x_3 y	Despliega una matriz de correlación de Pearson usando solamente observaciones con valores no missing sobre todas las variables especificadas. Adicionando la opción covarianza produce una matriz de varianza-covarianza proveniente de la correlación
pwcorr x_1 x_2 x_3 y, sig	Despliega una matriz de correlaciones de Pearson usando parejas deleción de valores missing y mostrando probabilidades de t test (de Ho:p = 0) sobre cada correlación.
graph matrix x_1 x_2 x_3 y, half	Dibuja una matríz de scatterplot s. Como sus listas de variables son las mismas, este ejemplo produce una matriz de scatterplots teniendo la misma organización como la matriz de correlación producida por el comando pwcorr .

test x1 x2	Estima una prueba F de la hipótesis nula que los coeficientes sobre X_1 y X_2 ambos son igual a cero, sobre el modelo de regresión más reciente.
sw regress yx1 x2 x3, pr(05)	Estima paso a paso un modelo de regresión usando backward (hacia atrás o eliminando) bajo predictores señalados que resultan significativos a un nivel de 0.05. O Forward (hacia delante) parte del modelos más simple utilizando los predictores señalados hasta el mas complicado tomando el mismo criterio de selección de predictores que el backward. El vlor de P, puede ser cambiante.

Por ejemplo, si analizamos el efecto del plomo sobre el peso al nacer:

Hipótesis: Las altas concentraciones de plomo en sangre en las mujeres embarazadas están relacionadas con una disminución del peso al nacer del recién nacido (RN).

Evento de estudio: peso del RN medido en gramos al momento del parto.
Exposición: Concentraciones de plomo en sangre en las mujeres embarazadas antes del parto.
Covariables: Edad gestacional, perímetro cefálico, talla de la madre, lactancia previa, fuma y otras.

En este estudio los investigadores están interesados en modelar el efecto del plomo sobre el peso al nacer por exposición a plomo durante el embarazo. En este caso la operacionalización de la variable independiente – la exposición a plomo- se hizo mediante la medición de plomo en sangre durante el embarazo en diferentes etapas del mismo (cada 3 meses) y 1 mes después del parto. La operacionalización de la variable dependiente – la medición del efecto (Peso al nacer) – se hizo mediante la evaluación del pediatra sobre el RN , dando como resultado la medición del peso en Kilogramos.

En este estudio es necesario entonces resumir y entender la información recolectada en este estudio (estudio de cohorte) mediante un modelo estadístico. Para esto necesitamos una representación sobre una ecuación matemática que nos permita modelar dicho efecto.

El modelo estadístico se debe ajustar a la siguiente ecuación:

$$\textit{Peso del recién naciodo} = \alpha + \text{exposición a plomo} * \text{efecto}$$

donde:

y_i = peso al nacer

α = es la media del peso al nacer

βx = exposición a plomo

Utilizando la base da datos pesorn:

Ejemplo de regresión lineal simple.

1.- Primero deberá seguir los pasos necesarios para conocer la base de datos, explorarla, detectar valores aberrantes u outliers.

2.- proceda a realizar un análisis univariado para conocer el comportamiento de las principales variables, si es necesario transformar la variable dependiente, hágalo.

3.- Ahora puede realizar el análisis bivariado, conozca la relación simple entre la variable dependiente y la independiente, además la relación entre las otras covariables, una por una. Con esto tendrá una idea de que variables pueden estar influyendo en la relación entre el peso al nacer y la exposición a plomo. Asegúrese de que las covariables no estén correlacionadas entre sí, pues podrían llevarlo a resultados erróneos.

Abriendo la base de datos **pesorn.dta**

Antes que nada debo empezar con la limpieza de la base, como conozco cuales son las variables por las cuales debo iniciar el análisis iniciaré con ellas explorándolas.

```
. sum peso_rn talla_rn  pecef_rn edges_rn

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     peso_rn |     274    3.080109   .4750916          1      4.525
    talla_rn |     274    49.85949   2.472442         35         56
    pecef_rn |     274    34.72993   5.817013         28       99.9
    edges_rn |     274    39.01095   5.488906         27         99
```

Observamos que las variables de percef_rn y edges_rn tienen valores de 99.9 y 99 Para un niño recién nacido estos valores no son posibles. Esto indica que tengo aun valores en los cuales las participantes no contestaron y a ellos se les aplicó un 99.

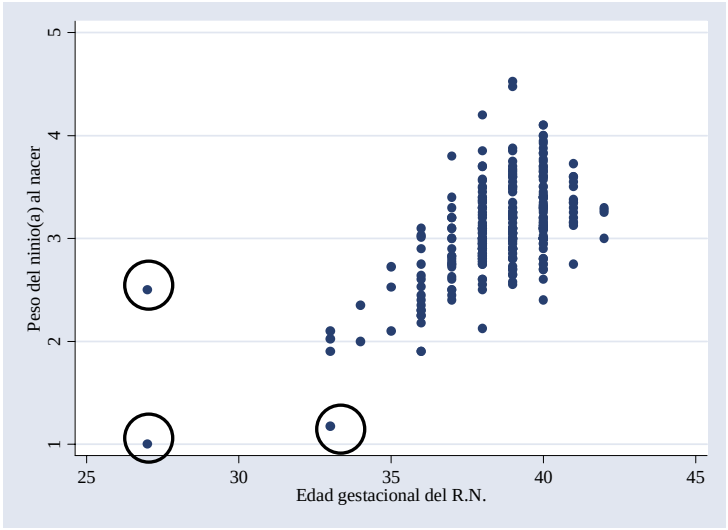
```
. sum  pecef_rn if pecef_rn<99

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pecef_rn |     272    34.25074   1.585148         28         43

. sum   edges_rn if edges_rn<99

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    edges_rn |     272    38.56985   1.896465         27         42
```

Vemos que al no incluir el valor 99 la media de ambas variables disminuye y el numero de observaciones también disminuye.

Podemos realizar algunas gráficas en las que veamos la correlación y evaluemos si existen o no puntos que pueden ser erróneos.

- **Aplicando gráficas de dispersión**

```
. scatter edges_rn peso_rn if edges_rn<99
```

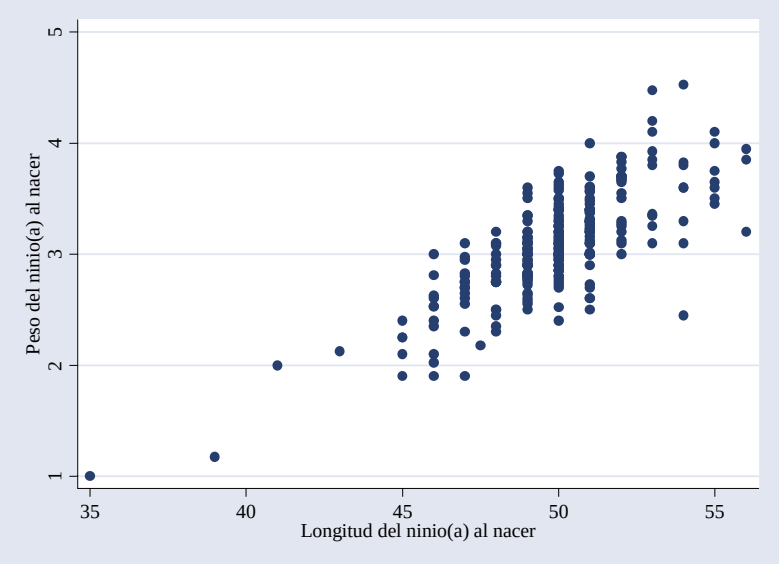


Esperaríamos que la relación fuera lineal que, es decir que todos los puntos quedaran alineados siguiendo una línea recta, los puntos que salen de la nube de puntos son los que debemos explorar.
¿Cómo podemos hacer esto?
Con list o browse.

El primer punto corresponde a una edad gestacional de 127 semanas y tiene un peso de 1 Kg., lo que es realmente bajo, sin embargo para su edad gestacional, lo podríamos creer, a menos que al verificar los cuestionarios estos no fueran los reales.

El siguiente punto que sale de la recta es el que corresponde a una edad gestacional de 33 semanas y peso de 1.175, el siguiente corresponde a una edad gestacional de 27 y un pese de 2.5. Estos últimos dos puntos hay que evaluarlos o tomarlos en cuenta en el análisis.

```
. scatter peso_rn talla_rn
```



Se correlacionan bien,

¿Como podemos evaluar que edad gestacional este bien determinada?: si conociéramos la fecha de ultima regla y la fecha de nacimiento del niño podríamos calcular una edad gestacional nosotros mismos.

Evaluaremos si tenemos puntos outliers:

Por ejemplo:

```
. lv  peso_rn

#    274        Peso del ninio(a) al nacer
                ---------------------------------
 M    137.5 |                  3.1              |    spread  pseudosigma
 F     69   |      2.81       3.08       3.35   |       .54    .4008705
 E     35   |       2.6        3.1        3.6   |         1    .4371645
 D     18   |       2.4      3.085       3.77   |      1.37    .4509953
 C      9.5 |    2.1125    3.00625        3.9   |    1.7875    .4876353
 B      5   |       1.9          3        4.1   |       2.2    .5189138
 A      3   |       1.9       3.05        4.2   |       2.3    .4920915
 Z      2   |     1.175      2.825      4.475   |       3.3    .6579634
 Y      1.5 |    1.0875    2.79375        4.5   |    3.4125    .6484396
        1   |         1     2.7625      4.525   |     3.525    .6288308
            |                                   |
            |                                   |   # below     # above
inner fence |         2                  4.16   |         6           3
outer fence |      1.19                  4.97   |         2           0
```

Los puntos que salen de los límites inferior internos y los límites exteriores externos son los que hay que evaluar.

```
. list folio talla_rn peso_rn edges_rn pecef_rn if  peso_rn>=4.16 & peso_rn<.

        folio   talla_rn    peso_rn  edges_rn   pecef_rn
152.      217         53      4.475        39         37
254.      334         53        4.2        38         37
283.      363         54      4.525        39         38


. list folio talla_rn peso_rn edges_rn pecef_rn if peso_rn<=2

        folio   talla_rn    peso_rn  edges_rn   pecef_rn
 43.       65         46        1.9        36         33
212.      287         45        1.9        33         30
223.      301         41          2        34         32
228.      306         47        1.9        36         32
241.      319         39      1.175        33         28
360.      444         35          1        27         36
```

Todas nuestras variables son continuas. Los valores que aquí parecen ser aberrantes debemos evaluarlos según nuestro criterio si no revisar que en el cuestionario correspondan y si no verificarlos con la participante.

Debemos también evaluar si la variable de plomo en sangre presenta o no discrepancias.

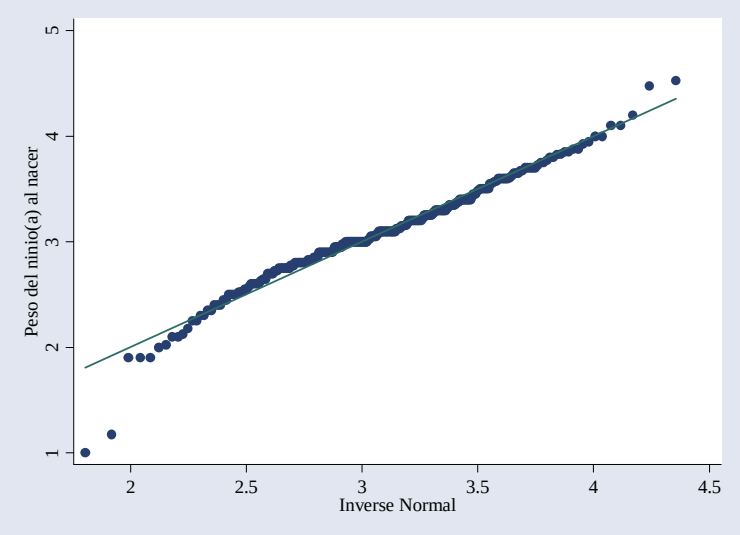
```
. sum  pb_3 pb_6 pb_8

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    pb_3 |     184    .2081887   .1934007     .03542     1.0869
    pb_6 |     183    .1668055   .2579198      .0274     3.0782
    pb_8 |     181    .1790993   .3123257      .0296     2.6329


. lv  pb_3 pb_6 pb_8

 #    167           pb_en plasma et.3
             ---------------------------------
 M     84   |                .134             |    spread  pseudosigma
 F     42.5 |    .09922     .16616      .2331 |    .13388    .1001744
 E     21.5 |     .0753    .214775     .35425 |    .27895    .1220124
 D     11   |      .065      .3097      .5544 |     .4894    .1605564
 C      6   |     .0557      .3772      .6987 |      .643    .1759903
 B      3.5 |     .0512     .43915      .8271 |     .7759    .1868296
 A      2   |     .0466     .48365      .9207 |     .8741    .1877487
 Z      1.5 |     .0465   .5000509   .9536018 |  .9071018     .184468
        1   |     .0464   .5164518   .9865037 |  .9401037    .1780867
             |                                |
             |                                |   # below     # above
inner fence |    -.1016                .43392 |         0          15
outer fence |   -.30242                .63474 |         0           7


 #    167           pb_en sangre et.6
             ---------------------------------
 M     84   |                .1085            |    spread  pseudosigma
 F     42.5 |     .0778      .1206      .1634 |     .0856    .0640494
 E     21.5 |    .06245      .1424     .22235 |     .1599    .0699401
 D     11   |     .0519     .18165      .3114 |     .2595    .0851336
 C      6   |     .0504     .23325      .4161 |     .3657    .1000928
 B      3.5 |    .04375     .30175     .55975 |      .516    .1242481
 A      2   |     .0364   .3520058   .6676117 |  .6312117    .1355785
 Z      1.5 |     .0319   .9524029   1.872906 |  1.841006    .3743865
        1   |     .0274     1.5528     3.0782 |    3.0508    .5779225
             |                                |
             |                                |   # below     # above
inner fence |    -.0506                .2918 |         0          11
outer fence |     -.179                .4202 |         0           5
```

```
 #    167           pb_en plasma et. 8
              ---------------------------------
 M     84   |                .1155             |     spread  pseudosigma
 F     42.5 |    .07915      .1263     .17345  |      .0943    .0705591
 E     21.5 |    .06365   .1453194   .2269887  |   .1633387    .0714442
 D     11   |     .0498      .1882      .3266  |      .2768    .0908092
 C      6   |     .0452      .2269      .4086  |      .3634    .0994633
 B      3.5 |    .04085     .63665    1.23245  |    1.1916    .2869264
 A      2   |     .0356    1.24485     2.4541  |    2.4185    .5194718
 Z      1.5 |     .0326    1.28805     2.5435  |    2.5109     .510616
        1   |     .0296    1.33125     2.6329  |    2.6033    .4931512
            |                                  |
            |                                  |   # below     # above
inner fence |    -.0623                .3149  |         0          11
outer fence |   -.20375                .45635 |         0           4
```

Existen valores que parecen outliers, los observamos y algunos de ellos corresponden en etapa al otro valor extremo en la etapa anterior y/o posterior.

Con lo anterior evaluamos normalidad de las variables y además detectamos valores alejados de la nube de puntos. Si no se realiza alguna corrección en los mismos porque se consideren plausibles, podemos evaluar si la distribución se asemeja a una distribución normal:

.qnorm peso_rn




```
. sktest peso_rn

                    Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
     peso_rn |      0.004         0.000           20.15       0.0000
```

La variable aunque gráficamente muestra apego a la línea normal en la prueba estadística rechazamos la hipótesis de que peso_rn tiene una distribución normal.

. ladder peso_rn

Transformation	formula	chi2(2)	P(chi2)
cube	peso_rn^3	41.99	0.000
square	peso_rn^2	12.94	0.002
raw	peso_rn	20.15	0.000
square-root	sqrt(peso_rn)	52.27	0.000
log	log(peso_rn)	.	0.000
reciprocal root	1/sqrt(peso_rn)	.	0.000
reciprocal	1/peso_rn	.	0.000
reciprocal square	1/(peso_rn^2)	.	0.000
reciprocal cube	1/(peso_rn^3)	.	0.000

¿Qué pasa aquí? Tendríamos que excluir los valores extremos?

Tenemos que decidir que variables podrían ser predictoras del peso al nacer y cuales potenciales confusoras para poderlas incluir en el modelo final, para esto debemos de realizar el análisis bivariado.

Sabemos que edad gestacional peso y talla deben tener una correlación ya que pensemos que a mayor edad gestacional el niño será mas grande y viceversa. Para esto realizaremos una prueba de correlación entre ellas.

La correlación es altamente significativa.

- **Aplicando pwcorr**

```
. pwcorr peso_rn talla_rn pecef_rn edges_rn pb_3 pb_6 pb_8,sig
```

	peso_rn	talla_rn	pecef_rn	edges_rn	pb_3	pb_6	pb_8
peso_rn	1.0000						
talla_rn	0.7701 0.0000	1.0000					
pecef_rn	0.0965 0.1110	0.1052 0.0823	1.0000				
edges_rn	0.5953 0.0000	0.5219 0.0000	0.1413 0.0198	1.0000			
pb_3	-0.1334 0.0734	-0.0898 0.2293	-0.0428 0.5672	-0.1210 0.1065	1.0000		
pb_6	-0.1535 0.0385	-0.0962 0.1964	-0.0184 0.8056	-0.0397 0.5968	0.4652 0.0000	1.0000	
pb_8	-0.0105 0.8885	0.0453 0.5461	-0.0167 0.8240	0.0345 0.6471	0.1752 0.0219	0.5263 0.0000	1.0000

Como habíamos visto, pwcorr despliega una matriz de correlaciones de Pearson usando parejas y eliminando los valores missing. Muestra probabilidades de t test (de Ho:p = 0) sobre cada correlación. Las correlaciones pueden tomar valores de 0 a 1 tanto en forma positiva como negativa, en nuestro caso vemos que plomo en sangre (pb_) en todas las etapas se correlaciona en forma negativa con el peso al

nacer. Sin embargo la correlación peso_rn - pb_8 no es significativa. Perímetro cefálico tampoco muestra una correlación significativa con el peso al nacer.

Podemos también evaluar otras variables que podrían ser confusoras:

```
. sum peso_m3 emba cipa_m6 edad_m n_hijos hijos_bp hijos_pm hijos_m abortos presis3
presis3

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    peso_m3  |     264    61.20758    10.4896         42        105
     talla_m |     458    155.7707   6.209847        140        192
        emba |     462    1.761905    1.13685          0          6
    cipa_m6  |     270     34.9363   3.256052       23.6       47.4
      edad_m |     463    27.15119   5.294655         14         43
     n_hijos |     456    .8135965   .9128577          0          8
    hijos_bp |     407    .0614251   .2870968          0          3
    hijos_pm |     406    .0763547   .2838579          0          2
     hijos_m |     407      .02457   .1550012          0          1
     abortos |     418    .2822967    .601136          0          4
     sexo_rn |     274    1.478102   .5004343          1          2
    presis3  |     256     110.293   10.58727         70        132
    predia3  |     256    70.03516   8.561813         40         90
```

y así para todas las etapas..

- **Aplicando pcorr**

```
. pcorr peso_rn peso_m3 emba cipa_m6 edad_m n_hijos hijos_bp hijos_pm hijos_m abortos
presis3 predia3 talla_m
(obs=200)

Partial correlation of peso_rn with

    Variable |    Corr.     Sig.
-------------+------------------
    peso_m3  |  -0.0389    0.596
        emba |   0.1793    0.014
    cipa_m6  |   0.1411    0.053
      edad_m |   0.1402    0.055
     n_hijos |  -0.1129    0.123
    hijos_bp |  -0.0255    0.728
    hijos_pm |  -0.0644    0.380
     hijos_m |   0.0371    0.613
     abortos |  -0.1861    0.011
    presis3  |   0.113    0.126
    predia3  |  -0.0973    0.184
     talla_m |   0.0528    0.472
```

pcorr permite realizar una prueba de correlaciones parciales únicamente entre la variable dependiente contra las variables independientes. No despliega la matriz de correlación de todas las variables. Únicamente las variables emba y abortos resultan en correlación significativa, aunque cipa_m6 y edad_m quedan en el valor límite.

Para analizar la variable como sexo del RN podemos aplicar una prueba t.

- **Aplicando ttest**

```
. ttest peso_rn, by(sexo_rn)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |     143    3.098566    .0422772    .5055622    3.014992    3.182141
       2 |     131    3.059962    .0384852    .4404829    2.983824      3.1361
---------+--------------------------------------------------------------------
combined |     274    3.080109    .0287013    .4750916    3.023605    3.136614
---------+--------------------------------------------------------------------
    diff |            .0386046    .0575157               -.074628    .1518371
------------------------------------------------------------------------------
Degrees of freedom: 272

                     Ho: mean(1) - mean(2) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   0.6712                 t =   0.6712                  t =   0.6712
   P < t =   0.7487             P > |t| =   0.5027              P > t =   0.2513
```

El comando ttest se utiliza para probar la hipótesis nula de que las medias de distribución entre dos grupos son iguales. En este caso nosotros no rechazamos la hipótesis nula, es decir, no existen diferencias en las medias de peso al nacer en los niños con respecto a las niñas, ya que el valor p de significancia es 0.5027 (p>0.05) . También podemos apreciar que las medias entre niños y niñas son 3.098 y 3.05 respectivamente.

O utilizar una prueba no paramétrica en el caso de que no conociéramos la distribución de la variable talla_rn de acuerdo al sexo del recién nacido.

- **Aplicando Kwallis**

```
. kwallis  talla_rn,by( sexo_rn)

Test: Equality of populations (Kruskal-Wallis test)

 sexo_rn          _Obs   _RankSum
       1           143   20109.50
       2           131   17565.50

chi-squared =     0.465 with 1 d.f.
probability =     0.4951

chi-squared with ties =     0.480 with 1 d.f.
probability =     0.4883
```

Al igual que la prueba *t* a través de la prueba de kwallis comprobamos que en no hay diferencias en cuanto a la media del peso del recién nacido por sexo (p=0.4883).

El análisis bivariado también puede hacerse probando por medio de modelos lineales simples, por ejemplo:

```
.  reg peso_rn pb_3

      Source |       SS       df       MS              Number of obs =     181
-------------+------------------------------           F(  1,   179) =    3.24
       Model |  .548885381     1  .548885381           Prob > F      =  0.0734
    Residual |  30.2945436   179   .16924326           R-squared     =  0.0178
-------------+------------------------------           Adj R-squared =  0.0123
       Total |  30.8434289   180  .171352383           Root MSE      =  .41139

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       pb_3  |  -.2876684   .1597375    -1.80   0.073    -.6028793    .0275426
       _cons |   3.185634     .04503    70.74   0.000     3.096776    3.274492
------------------------------------------------------------------------------

.  reg peso_rn pb_6

      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  1,   180) =    4.35
       Model |  .739101466     1  .739101466           Prob > F      =  0.0385
    Residual |    30.61112   180  .170061778           R-squared     =  0.0236
-------------+------------------------------           Adj R-squared =  0.0182
       Total |  31.3502215   181  .173205644           Root MSE      =  .41239

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       pb_6  |  -.2470947   .1185263    -2.08   0.039    -.4809744   -.0132149
       _cons |   3.156136   .0364194    86.66   0.000     3.084272       3.228
------------------------------------------------------------------------------
```

como sabemos el comando ***regress o reg*** estima la ecuación de la regresión de mínimos cuadrados entre la variable y (variable dependiente y la variable X (variable independiente), por lo tanto mediante este podemos ajustar la siguiente ecuación, tomado pb_6 como la principal variable independiente.

$$y_i = \alpha + \beta x$$

Peso al nacer = $\alpha +$ plomo en sangre et 6 * efecto

Peso al nacer= 3.156 – 0.2471plomo en sangre et.6

3.156 es la media esperada del peso al nacer cuando x=0

0.2471 representa el coeficiente β, es decir la medida del efecto, la unidad de cambio.

Podríamos interpretar que por cada μg/dl de plomo que aumenta en plasma de la madre, disminuye en 0.25 kg el peso al nacer, asumiendo que no existen otros confusores.

El valor p asociado al coeficiente indica que la asociación observada es diferente a la magnitud de asociación que se podría observar simplemente por el azar.

Esto se puede hacer con las demás covariables.

Es útil probar una reexpresión de la variable independiente (variable continua) en forma de categorías que me ayuden a evaluar si los grupos mas altos podrían predecir mejor la disminución del peso al nacer.

Dado que no existen datos en la literatura de cómo podríamos agrupar las concentraciones de plomo en sangre, nosotros agruparemos la variable en cuartiles. Mediante esta categorización dividiremos la variable en cuatro grupos que contengan el 25 % de las observaciones cada uno:

```
. sum pb_6,d
                    pb_en plasma et.6
-------------------------------------------------------------
      Percentiles      Smallest
 1%        .0364          .0274
 5%        .0517          .0364
10%        .0598          .0425       Obs                 183
25%        .0807           .045       Sum of Wgt.         183

50%        .1098                      Mean           .1668055
                       Largest       Std. Dev.      .2579198
75%        .1697          .5963
90%        .2789       .6676117      Variance       .0665226
95%        .4161       1.357228      Skewness       8.596641
99%     1.357228         3.0782      Kurtosis       92.44176
```

en un Segundo paso generaremos las variables indicadoras. Para este ejemplo se requiere de 4 variables indicadoras (x1, x2, x3, x4) que indican la presencia o la ausencia en un grupo en particular .

```
. gen qpb6=pb_6
(281 missing values generated)

. recode qpb_6 min/0.0807=1 0.0810/.1098=2 0.1099/.1697=3 .1698/max=4
(183 changes made)
```

```
. tab  qpb_6
    qpb_6   |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         46       25.14       25.14
          2 |         46       25.14       50.27
          3 |         47       25.68       75.96
          4 |         44       24.04      100.00
------------+-----------------------------------
      Total |        183      100.00
```

Una variable indicadora significa que contiene 1 cuando pertenece a ese grupo y = cuando no pertenece.

Podemos realizar una prueba ANOVA de una sola vía para ver si existe alguna diferencia de peso al nacer por categoría de plomo en sangre.

- **Aplicando ANOVA (oneway)**

```
. oneway peso_rn qpbpl6, tab bonferroni
           |   Summary of Peso del ninio(a) al
           |                nacer
     qpb_6 |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |   3.1696739   .32804708          46
         2 |   3.1919565   .41835056          46
```

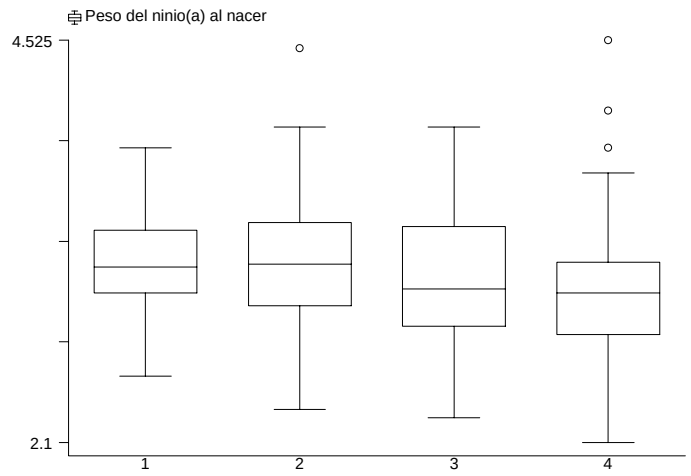
```
         3 |   3.0646739   .42685725          46
         4 |   3.0294318    .4721022          44
-----------+------------------------------------
     Total |   3.1148626   .41617982         182

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      .848596645      3   .282865548      1.65     0.1794
 Within groups      30.5016249    178   .171357443
------------------------------------------------------------------------
    Total           31.3502215    181   .173205644

Bartlett's test for equal variances:  chi2(3) =   5.8611  Prob>chi2 = 0.119

                     Comparison of Peso del ninio(a) al nacer by qpb_6
                                   (Bonferroni)
Row Mean-|
Col Mean |          1          2          3
---------+---------------------------------
       2 |    .022283
         |      1.000
         |
       3 |      -.105   -.127283
         |      1.000      0.852
         |
       4 |   -.140242   -.162525   -.035242
         |      0.659      0.386      1.000
```

Dado que esta prueba nos dice si hay o no diferencia entre los grupos con respecto a la varianza de cada uno de ellos, nosotros necesitamos valores grandes de F para rechazar la hipótesis nula de que los grupos son iguales. En este caso no rechazamos la hipótesis nula.

El tab nos muestra como está la media de los pesos de los niños al nacer por cada una de las categorías. Si tomamos como referencia el primer cuartil para comparar los demás grupos las diferencias entre los cuartiles serían:

Q1-Q1=0 Q1-Q2=-0.0223 Q1-Q3=0.105 Q1-Q4=0.1402

¿Cómo representaríamos gráficamente estas diferencias de medias?



¿Y cómo podríamos expresar en esto en un modelo de regresión lineal?

- **Aplicando regresión lineal simple**

```
. tab qpb_6,gen(qpb6)

. reg peso_rn  qpb6_2 qpb6_3  qpb6_4

      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  3,   178) =    1.65
       Model |  .848596645     3  .282865548           Prob > F      =  0.1794
    Residual |  30.5016249   178  .171357443           R-squared     =  0.0271
-------------+------------------------------           Adj R-squared =  0.0107
       Total |  31.3502215   181  .173205644           Root MSE      =  .41395

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      qpb6_2 |   .0222826   .0863153     0.26   0.797    -.1480503    .1926155
      qpb6_3 |      -.105   .0863153    -1.22   0.225    -.2753329    .0653329
      qpb6_4 |  -.1402421   .0872906    -1.61   0.110    -.3124997    .0320155
       _cons |   3.169674   .0610341    51.93   0.000     3.04923    3.290117
------------------------------------------------------------------------------
```

En el modelo anterior dejamos de referencia la primera categoría, cuando las otras tres variables tomen el valor de cero, entonces la constante corresponde a la media estimada para el primer cuartil. Vemos que los intervalos de confianza se entrecruzan entre cada categoría, los valores de p no son sinificativos. Podemos realizar una prueba para evaluar si existe diferencia entre los tres grupos:

```
. lincom  qpbpl6_2- qpbpl6_3
 ( 1)  qpbpl6_2 – qpbpl6_3 = 0.0

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1272826   .0863153     1.47   0.142    -.0430503    .2976155
------------------------------------------------------------------------------
```

No hay diferencias.

Nota: hacer la prueba para las demás categorías.

Podemos seguir evaluando:

```
. reg peso_rn  qpbpl6_3  qpbpl6_4
```

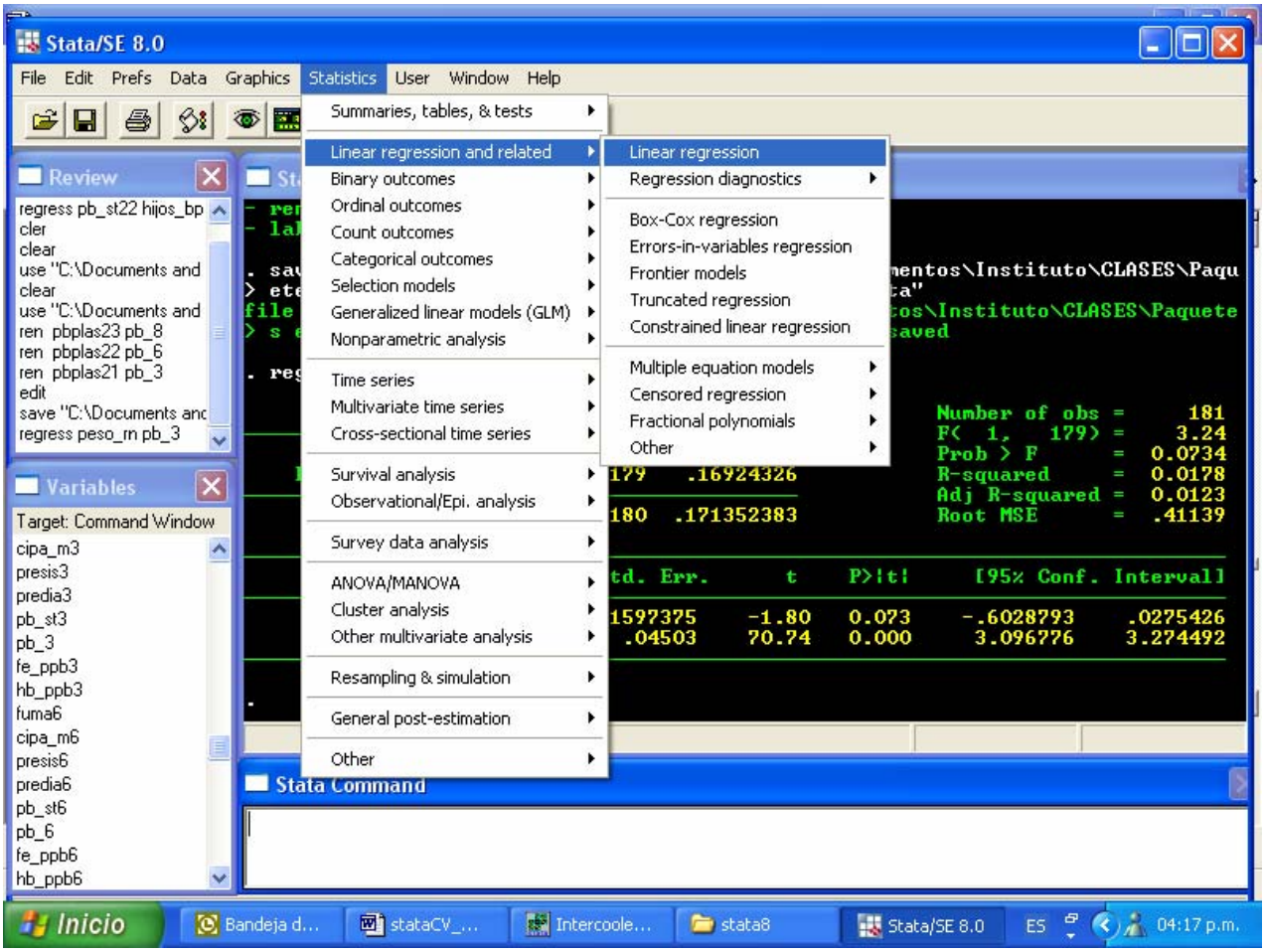
```
      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  2,   179) =    2.46
       Model |    .8371768     2    .4185884           Prob > F      =  0.0887
    Residual |  30.5130447   179  .170463937           R-squared     =  0.0267
-------------+------------------------------           Adj R-squared =  0.0158
       Total |  31.3502215   181  .173205644           Root MSE      =  .41287

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      qpb6_3 |  -.1161413   .0745561    -1.56   0.121    -.2632632    .0309806
      qpb6_4 |  -.1513834   .0756773    -2.00   0.047    -.3007178    -.002049
       _cons |   3.180815    .043045    73.90   0.000     3.095874    3.265756
------------------------------------------------------------------------------


. reg peso_rn  qpbpl6_4

      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  1,   180) =    2.46
       Model |  .423520174     1  .423520174           Prob > F      =  0.1182
    Residual |  30.9267013   180  .171815007           R-squared     =  0.0135
-------------+------------------------------           Adj R-squared =  0.0080
       Total |  31.3502215   181  .173205644           Root MSE      =  .41451

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      qpb6_4 |  -.1126696    .071763    -1.57   0.118    -.2542745    .0289353
       _cons |   3.142101   .0352851    89.05   0.000     3.072476    3.211727
------------------------------------------------------------------------------


. reg peso_rn qpbpl6_1  qpbpl6_2

      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  2,   179) =    2.41
       Model |  .820665332     2  .410332666           Prob > F      =  0.0931
    Residual |  30.5295562   179   .17055618           R-squared     =  0.0262
-------------+------------------------------           Adj R-squared =  0.0153
       Total |  31.3502215   181  .173205644           Root MSE      =  .41298

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      qpb6_1 |   .1222295   .0748519     1.63   0.104    -.0254763    .2699352
      qpb6_2 |   .1445121   .0748519     1.93   0.055    -.0031936    .2922178
       _cons |   3.047444   .0435324    70.00   0.000     2.961542    3.133347
------------------------------------------------------------------------------


. reg peso_rn qpbpl6_1  qpbpl6_2 qpbpl6_3

      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  3,   178) =    1.65
       Model |  .848596645     3  .282865548           Prob > F      =  0.1794
    Residual |  30.5016249   178  .171357443           R-squared     =  0.0271
-------------+------------------------------           Adj R-squared =  0.0107
       Total |  31.3502215   181  .173205644           Root MSE      =  .41395

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      qpb6_1 |   .1402421   .0872906     1.61   0.110    -.0320155    .3124997
      qpb6_2 |   .1625247   .0872906     1.86   0.064    -.0097329    .3347823
      qpb6_3 |   .0352421   .0872906     0.40   0.687    -.1370155    .2074997
       _cons |   3.029432   .0624058    48.54   0.000     2.906281    3.152582
------------------------------------------------------------------------------

. reg peso_rn qpbpl6_1
```

```
      Source |       SS       df       MS              Number of obs =     182
-------------+------------------------------           F(  1,   180) =    1.07
       Model |  .184939669     1  .184939669           Prob > F      =  0.3028
    Residual |  31.1652818   180  .173140455           R-squared     =  0.0059
-------------+------------------------------           Adj R-squared =  0.0004
       Total |  31.3502215   181  .173205644           Root MSE      =   .4161

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      qpb6_1 |   .0733504   .0709719     1.03   0.303    -.0666936    .2133944
       _cons |   3.096324   .0356804    86.78   0.000     3.025918    3.166729
------------------------------------------------------------------------------
```

¿cómo haríamos esto en stata 8?

Los comandos son los mismos.

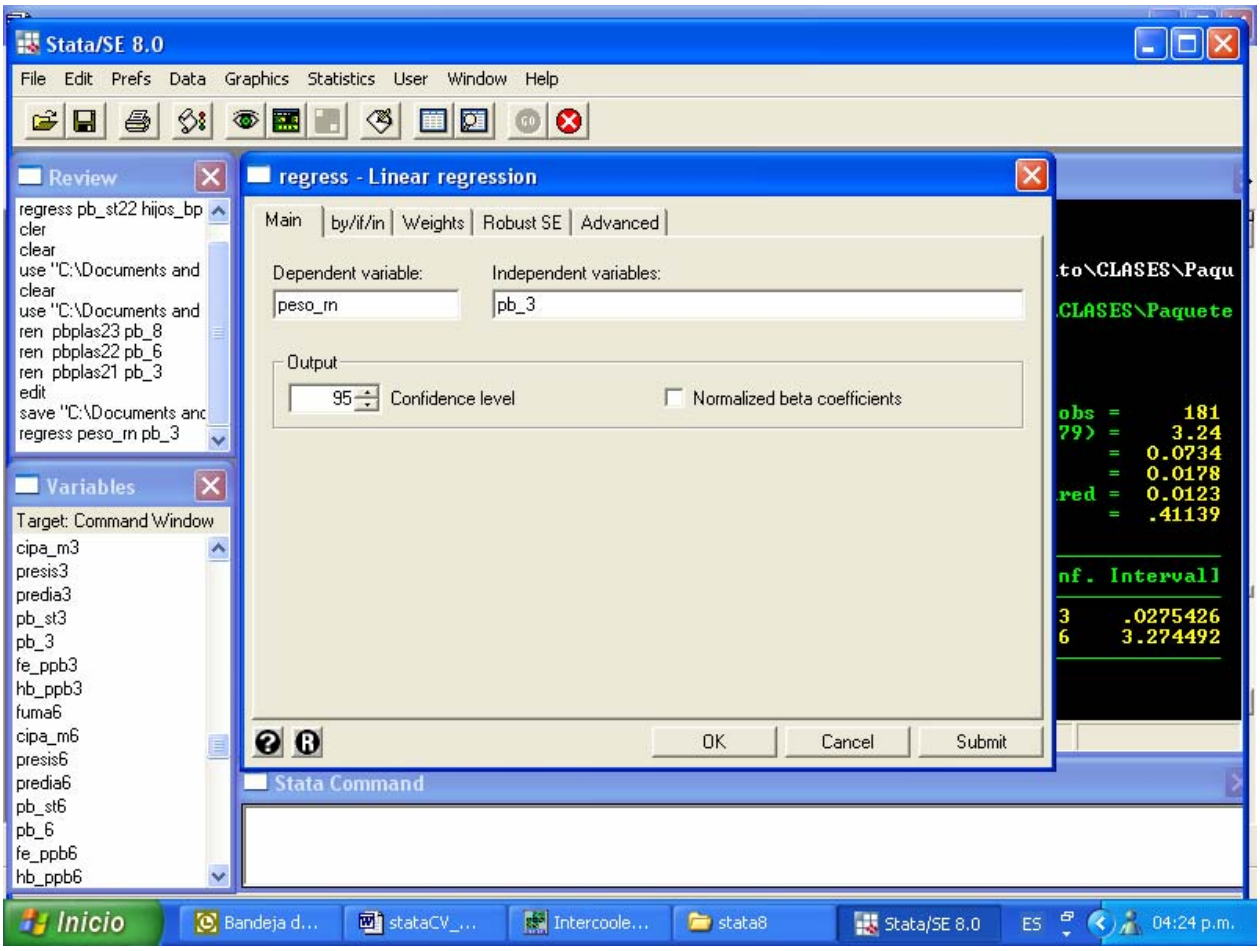
Si lo queremos hacer a partir de las ventanas:



En el menu de opciones seleccionamos [statisctics] luego nos vamos a la opción[linear regression and relateded] y ahí presionamos [linear regression], en donde nos presentará una ventana en la cual nos pide introducir los datos de las variables sobre las cuales queremos realizar la regresión.

En dicha ventana debemos introducir el nombre de la varaibles dependiente y el nombre de la (las) variable(s) independientes.

Existen otras opciones que se pueden cambiar como por ejemplo el nivel de confianza. Además incluir algunas otras como es dar peso por alguna variable, hacer un análisis estratificado, etc.


Stata/SE 8.0
File Edit Prefs Data Graphics Statistics User Window Help
Review
regress pb_st22 hijos_bp
cler
clear
use "C:\Documents and
clear
use "C:\Documents and
ren pbplas23 pb_8
ren pbplas22 pb_6
ren pbplas21 pb_3
edit
save "C:\Documents and
regress peso_rn pb_3
Variables
Target: Command Window
cipa_m3
presis3
predia3
pb_st3
pb_3
fe_ppb3
hb_ppb3
fuma6
cipa_m6
presis6
predia6
pb_st6
pb_6
fe_ppb6
hb_ppb6
regress - Linear regression
Main by/if/in Weights Robust SE Advanced
Dependent variable:
peso_rn
Independent variables:
pb_3
Output
95 Confidence level
Normalized beta coefficients
OK Cancel Submit
Stata Command
.to\CLASES\Paqu
.CLASES\Paquete
obs = 181
79) = 3.24
= 0.0734
= 0.0178
red = 0.0123
= .41139
nf. Interval]
3 .0275426
6 3.274492
Inicio
Bandeja d...
stataCV_...
Intercoole...
stata8
Stata/SE 8.0
ES
04:24 p.m.

- **Regresión lineal múltiple**

Tomando de referencia el artículo de Cossio Et al. Es necesario evaluar un modelo que incluya potenciales confusores de la relación anterior.
En este caso la ecuación anterior cambia por la siguiente:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots\ldots + \varepsilon_{ij}$$

Con este modelo se muestra la importancia de los dos niveles de acción necesarios para utilizar los métodos estadísticos ya que hay que evaluar la hipótesis tanto desde el punto de vista estadístico como desde el punto de vista conceptual.

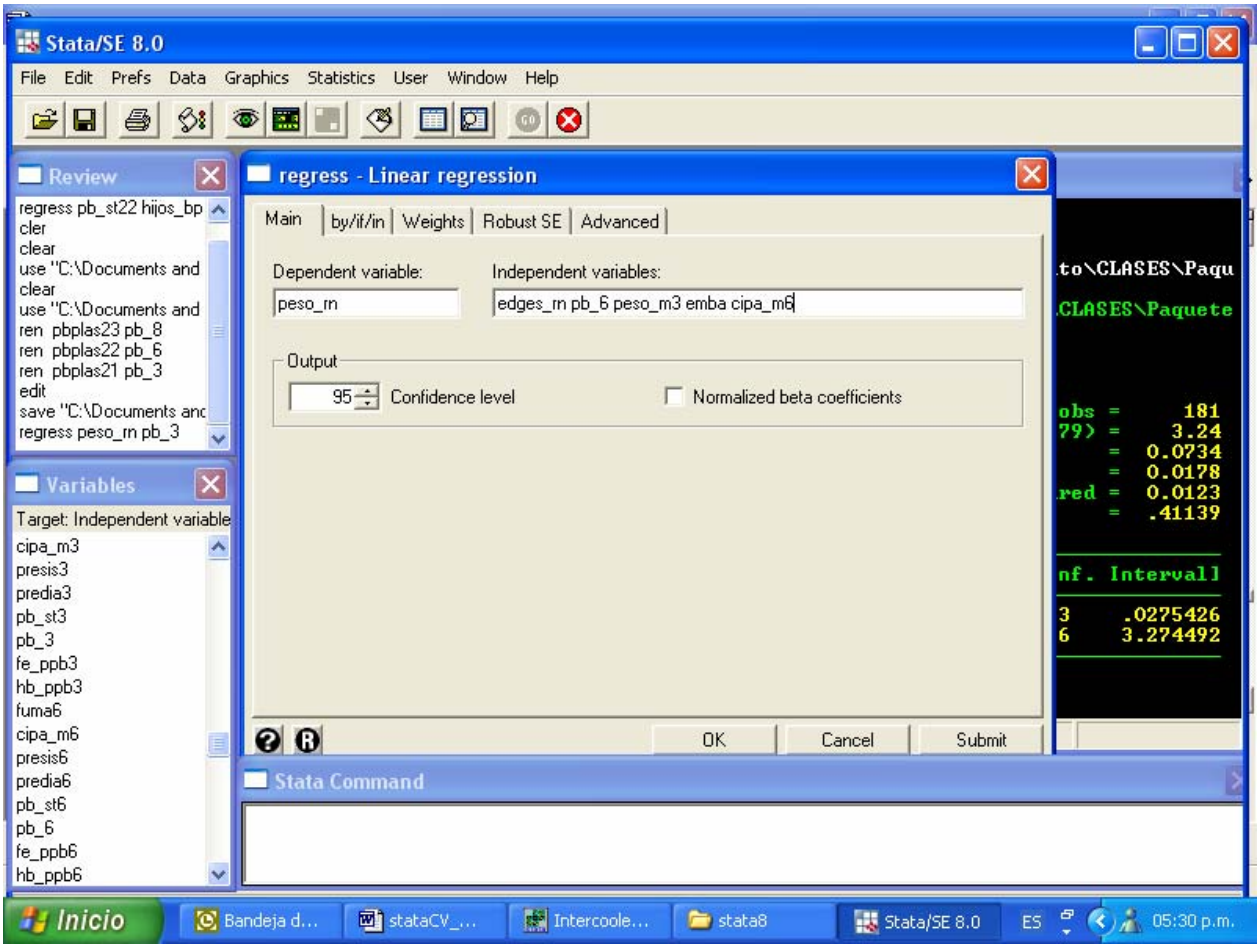
Aplicando Stata, nosotros tenemos que traducir esa ecuación en aplicación de comandos.

Continuación del ejercicio de Peso al nacer y plomo en sangre...

4.- Ahora sí, realice el modelo con las variables que mejor predicen la relación lineal. Tome en cuenta los criterios correspondientes.

```
. reg peso_rn edges_rn pb_6 peso_m3 emba cipa_m6

      Source |       SS       df       MS              Number of obs =     170
-------------+------------------------------           F(  5,   164) =   14.57
       Model |  8.89872814     5  1.77974563           Prob > F      =  0.0000
    Residual |  20.0330005   164  .122152442           R-squared     =  0.3076
-------------+------------------------------           Adj R-squared =  0.2865
       Total |  28.9317286   169   .17119366           Root MSE      =   .3495

------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pb_6 |  -.2064495   .1011997    -2.04   0.043    -.4062718   -.0066273
    edges_rn |   .1286125   .0188388     6.83   0.000     .0914146    .1658104
     peso_m3 |   -.003908   .0046138    -0.85   0.398    -.0130182    .0052022
        emba |   .0620872   .0251905     2.46   0.015     .0123477    .1118267
     cipa_m6 |   .0389931   .0158391     2.46   0.015     .0077182     .070268
       _cons |  -3.108531   .7824896    -3.97   0.000    -4.653583   -1.563478
------------------------------------------------------------------------------
```

En Stata 8 lo podríamos hacer de la siguiente manera:



¿cómo interpretamos estos resultados?

Peso al nacer= -3.108531 -0.2064495pb_6+ 0.1286125edges_rn -.003908peso_m3 + 0.0620872emba + 0.0389931cipa_m6

Podríamos interpretar que por cada μg/dl de plomo que aumenta en plasma de la madre, disminuye en 0.2065 kg el peso al nacer, asumiendo que el resto de las covariables permanecen constantes.

En el caso de las variables indicadoras, ¿cómo sería la interpretación?

Cuando la variable indicadora es 1, ej. fumar durante el embarazo se espera una reducción en el peso al nacer de ***x*** kgs. Cuando la varible indicadora toma el valor de cero -las mujeres no fumaron durante el embarazo- el valor esperado es el de la media.

El valor p asociado a los coeficientes, indica que la asociación observada es diferente a la magnitud de asociación que se podría observar simplemente por el azar.

- **Coeficiente de determinación R2**

En nuestro modelo tenemos una R^2 de 0.3076, esto es que nuestro modelo explicar el 30.76 5 de la variabilidad del peso al nacer, el resto queda explicado por variables desconocidas. La raíz cuadrada positiva de R2 es el coeficiente de correlación múltiple de ***y*** con el conjunto de regresores incluidos en el modelo. En el ejémplo r es 0.5546.

5.- Evalúe el modelo. ¿cumple con los supuestos de la regresión lineal?

- **verificar los supuestos :**

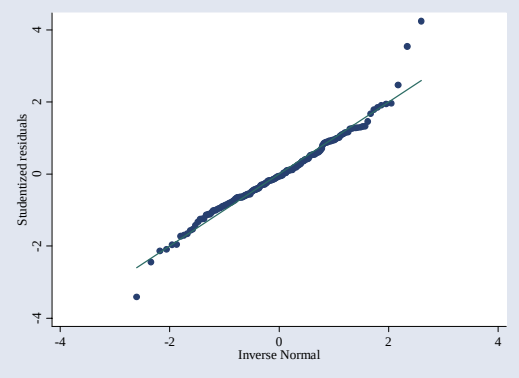
- **predict residuos, r**

para nuestro modelo:

. predict residuo,rstu

(294 missing values generated)

- **qnorm residuos**



Los residuos son discrepancias entre el valor estimado con el modelo y el valor observado. Los residuos pueden verse como la variabilidad que no puede explicarse mediante el modelo de regresión. También se pueden interpretar como el valor de error. Es por eso que observamos los residuos para saber si se cumplen o no las suposiciones básicas del modelo. En este caso vemos que existen residuos demasiado grandes que aun no ajustan a la línea normal, esos residuos podemos evaluarlos.

sum residuos (ojo, estos residuos son estudentizados)

```
. sum residuo

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     residuo |     170    .0033081   1.016795  -2.356718   4.656434
```

list if abs(residuos)>2.5 & abs(residuo)<.

```
. list folio peso_rn pb_6 if abs(residuo)>2.5 & abs(residuo)<.

        folio    peso_rn   pb_6
 92.      217      4.475      .1007
158.      363      4.525      .1727
```

count if abs(residuos)>1.645

```
. count if abs(residuo)>1.645 & residuo<.
   16
```

display

```
. display 16/170
       .0941176516
```

count if abs(residuos)>1.96

```
. count if abs(residuo)>1.96 & residuo<.
    8
```

swilk

```
. swilk residuo

                  Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
     residuo |    170    0.94693      6.876      4.400  0.00001
```

Estas pruebas de Shapiro Wilk da información sobre el grado de concordancia entre la gráfica normal y la distribución esperada sobre la línea recta.

La **W** representa los valores de las pruebas Shapiro wilk y la **V** el valor de la prueba. El valor esperado de V para distribuciones normales es de 1. No debo rechazar la hipótesis nula para normalidad.

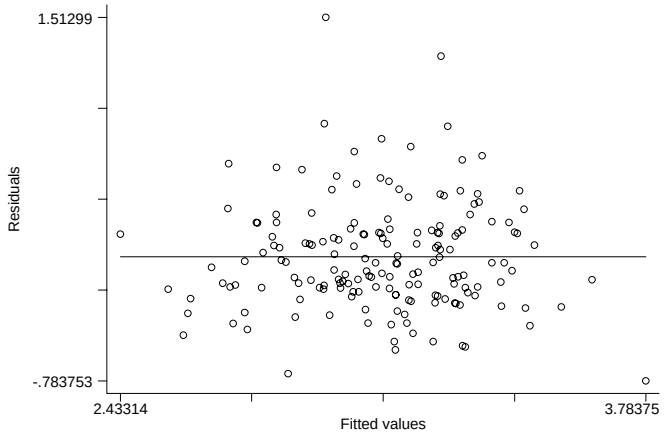
Dado que los valores observados en la variable independiente y los residuos no son independientes, no se recomienda realizar gráficos diagnósticos utilizando estas variables.

Lo esperado es los gráficos de *ei* contra y_i *estimada* es que no exista relación entre los residuos y el valor esperado. Cualquier patrón de dependencia indica problema.

Para el modelo rechazo la hipótesis nula de normalidad.

- **rvfplot, ylab( ) xlab( )**

. rvfplot, ylabel xlabel



Gráfica de los residuos comunes contra el valor estimado de la variable respuesta, para evaluar media cero y varianza constante.

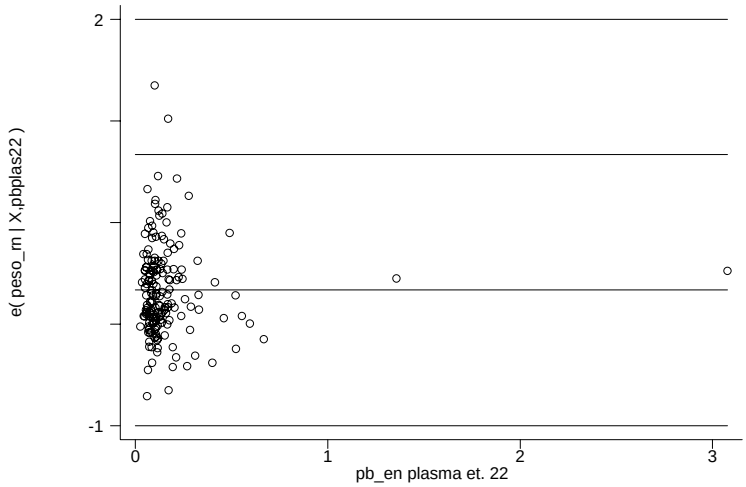
- **hettest**

```
. hettest
Cook-Weisberg test for heteroskedasticity using fitted values of peso_rn
     Ho: Constant variance
         chi2(1)      =      0.33
         Prob > chi2  =      0.5635
```

hettest es una prueba de heterocedasticidad. No debería de encontrarse algún patrón de comportamiento, en el ejémplo el valor p es .5635 con lo cual no rechazo la hipótesis nula de varianzas constantes. En este sentido el modelo propuesto es bueno puesto que no existe algun patrón de comportamiento en los valores esperados del peso al nacer. Las varianzas son constantes.

- **rvpplot plomo yline xlab**

gráfica de los residuos comunes contra cada una de las variables independientes.



¿Qué se observa ene sta gráfica?

- ✓ **predict hat, hat** Predice los puntos influyentes

Una medida de la distancia de cada punto al centroide de puntos se conoce como "Hat Matriz" y los valores que puede tomar van desde:

$$\frac{1}{n} \leq h_{ij} \leq \frac{1}{c}$$

```
. predict sombrero, hat
(294 missing values generated)

. count if sombrero>2*6/170 & sombrero<.
    9

. list folio peso_rn pb_6 peso_m3 emba cipa_m6  if sombrero>2*6/170 & sombrero<.

        folio    peso_rn       pb_6      peso_m3      emba   cipa_m6
 86.      393        3.5      .0834       68.2         6         39
 94.      229       2.55      .1335         60         1       29.5
100.      256        3.6      .1153         98         3         42
147.      152       2.35      .4607         52         3         36
151.      237      2.995   1.357228         54         3       33.5
171.      167       3.05      .5232        105         4       47.4
178.        7      2.525      .5539        100         1       45.3
180.      396       2.75      .2042         77         5       35.7
182.      139      2.575     3.0782         48         1         33
```

El valor mínimo se obtiene si todos los elementos de x_i son iguales a la media de la variable y si los datos caen en el centroide de la distribución. El valor máximo se presenta en observaciones alejadas del centroide. Si se tiene el valor más alto, de 1, entonces el punto es tan influyente que forza la dirección de la recta hasta pasar por el punto.

count if hat>2*p/n. Se considera que las observaciones que toman valores dos veces por arriba del valor espérado, pueden ser de gran peso para los parámetros estimados.

- ➢ **distancia de cook**

predict cook, cooksd

La distancia de Cook nos permite detectar posibles valores aberrantes: la media de cook cuantifica el impacto de la observación o del punto sobre el modelo; cuantifica que tanto cambia el modelo, es decir, los coeficientes de regresión, al excluir cada uno de los puntos.

Se espera que los resultados de la regresión no dependan de una sola observación o de un punto de la regresión.

Distancia de Cook: $D_i = \frac{r_i^{\,2}}{p'}\left(\frac{h_{ij}}{1-h_{ij}}\right)$

Donde r_i^2 es el residual estandarizado, h_{ij} la diagonal de la matriz sombrero (hat) y p' el número de parámetros en el modelo.

La distancia de Cook combina una medida de influencia y de falta de ajuste y se distribuye como una F con p+1 y n-p-1 grados de libertad.

```
. predict cook, cooksd
(294 missing values generated)

. sum cook

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        cook |     170    .0071119   .0236071   2.40e-07   .2515637

. count if cook>1 & cook<.
    0
```

Los puntos que toman valor por arriba de uno ameritan averiguarlos. Si existen puntos arriba de 2 entonces si hay problemas.

- **dfbeta** puntos influyentes en β

  DFBETAS= $\dfrac{b_k - b_{k(i)}}{S_{e(i)} \Big/ \sqrt{RSS_k}}$ Si DFBETAS>0 sobre estima las b´s. O si DFBETAS<0 sub estima las b´s.

  DBETAS>2/ $\sqrt{n}$

  Cumpliendo con normalidad y corrigiendo por el tamaño de muestra.

Este diagnóstico nos ayuda a evaluar el impacto sobre el vector de β's . No todas los outliers o valores aberrantes influyen en los datos estimadores.

Nos indica el impacto que ejercería sobre las betas el eliminar las observaciones en cuestión y expresa la magnitud de cambio en unidades de desviación estándar.

```
. dfbeta

(294 missing values generated)
            DFedges_rn:  DFbeta(edges_rn)
(294 missing values generated)
            DFpb_6:  DFbeta(pb_6)
(294 missing values generated)
            DFpeso_m3:  DFbeta(peso_m3)
(294 missing values generated)
              DFemba:  DFbeta(emba)
(294 missing values generated)
```

DFcipa_m6: DFbeta(cipa_m6)

```
. sum  DFedges_rn DFpb_6 DFpeso_m3 DFemba DFcipa_m6

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  DFedges_rn |     170     .000399   .0734837  -.3696917   .2911262
      DFpb_6 |     170    .0048854   .0983734  -.1666424   1.208837
   DFpeso_m3 |     170   -.0001463   .0732985  -.1845765    .297517
     DFemba  |     170   -.0003174   .0999066  -.4132033   .8465961
  DFcipa_m6  |     170   -.0005006   .0759841  -.5158963    .194616
```

count if abs(df*)>2/sqrt(n)

```
. for var  DFedges_rn- DFcipa_m6: count if abs(X)>2/sqrt(170) & X<.

->  count if abs(DFedges_rn)>2/sqrt(170) & DFedges_rn<.
  10
->  count if abs(DFpb_6)>2/sqrt(170) & DFpb_6<.
  3
->  count if abs(DFpeso_m3)>2/sqrt(170) & DFpeso_m3<.
  10
->  count if abs(DFemba)>2/sqrt(170) & DFemba<.
  9
->  count if abs(DFcipa_m6)>2/sqrt(170) & DFcipa_m6<.
```

- **dffits**

dffits >2*sqrt(p/n)

$$\text{DFFITS} > 2 * \sqrt{p/n}$$

```
.precit dfits, dfits

. list folio peso_rn pb_6 peso_m3 emba cipa_m6 if abs(dfit)>2*sqrt(6/170) & dfit<.

       folio    peso_rn         pb_6   peso_m3      emba   cipa_m6
  1.     170       3.85       .0637         65         4      34.5
```

```
 37.       11          3      .0621         63         4       38.5
 92.      217      4.475      .1007         54         1       31.4
 95.       77        3.8      .1191         80         2         39
158.      363      4.525      .1727         54         5         33
171.      167       3.05      .5232        105         4       47.4
182.      139      2.575     3.0782         48         1         33
```

Informan de acerca de cómo cambia el valor predicho al excluir la x_i observación. Su interpretación es muy similar a la distancia de Cook.

Hay que explorar los puntos antes de exluirlos:

```
. reg peso_rn pb_6 edges_rn peso_m3 emba cipa_m6 if abs(dfit)<2*sqrt(6/170)

      Source |       SS       df       MS              Number of obs =     163
-------------+------------------------------           F(  5,   157) =   22.57
       Model |  9.88283419     5  1.97656684           Prob > F      =  0.0000
    Residual |  13.7504885   157  .087582729           R-squared     =  0.4182
-------------+------------------------------           Adj R-squared =  0.3996
       Total |  23.6333227   162  .145884708           Root MSE      =  .29594
------------------------------------------------------------------------------
     peso_rn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pb_6 |  -.2563846   .1631968    -1.57   0.118    -.5787292    .0659599
    edges_rn |   .1377506   .0162653     8.47   0.000     .1056235    .1698777
     peso_m3 |  -.0061606   .0039894    -1.54   0.125    -.0140405    .0017193
        emba |   .0605842   .0226563     2.67   0.008     .0158336    .1053347
     cipa_m6 |   .0558272   .0136893     4.08   0.000     .0287883    .0828661
       _cons |  -3.927776   .6883575    -5.71   0.000    -5.287412    -2.56814
------------------------------------------------------------------------------
```

¿Qué observamos? Al parecer uno de los puntos influyentes era en pb_6 el valor de 3.07 ya que cambia considerablemente el valor del coeficiente del mismo. Podríamos solo evalura sin ese valor.

- **vif**

multicolinealidad. Vector de Inflación de la varianza.

```
. vif
    Variable |       VIF       1/VIF
-------------+----------------------
     peso_m3 |      3.17    0.315057
     cipa_m6 |      3.13    0.319620
        emba |      1.03    0.968075
    edges_rn |      1.03    0.970605
        pb_6 |      1.01    0.994793
-------------+----------------------

  Mean VIF |     1.87
```

Un valor de 10 en la media del factor de infación de la varianza representa multicolinealidad.

Ejercicio práctico:

Regresión lineal:

Con el fin de controlar algunas enfermedades ocasionadas por vectores como es el caso de la Malaria, en México se utilizan algunos compuestos organoclorados y organofosforados para controlar al vector. Como resultado de actividades intensivas de este tipo se ha logrado reducir la el numero de casos de malaria a nivel nacional. El DDT (Dicloro Difenil Tricloroetano) se usó en épocas pasadas y se sigue usando en menor cantidad como spray dentro de las casas en áreas endémicas. El DDT puede metabolizarse en el organismo a p'p-DDE y p'p-DDT, sobre los cuales en algunos estudios se ha reportado que pueden tener efectos estrogénicos y androgénicos en los humanos. Con el propósito de describir las concentraciones de DDT en hombres residentes de un área endémica de paludismo no expuestos ocupacionalmente a DDT, se realizó un estudio transversal en Chiapas México en donde se evaluaron diferentes metabolitos del DDT en plasma y se midieron algunos factores potenciales asociados al incremento de dichos Biomarcadores.
Referencia: Non-Occupational Determinants of Plasma DDT and P, P'-DDE in men from Chiapas, Mexico

En base al artículo de referencia y a la base que se le proporciona (ddt.dta), realice el siguiente ejercicio:

1. Antes de iniciar con el análisis estadístico:

    a) Explore las variables para detectar valores Outliers, si encuentra valores outliers deberá decidir si hay que eliminar o reemplazar los datos por valo res perdidos.

    b)Mediante un gráfico de barras evalúe la distribución de las variables: stature weight opdde ppdde opddt ppddt ppdde_li ppddt_li

    c)mediante grafico evalue las frecuencias de las siguientes variables: age adress time_res born_pla pest_inf actual_o frecupe_ ddt smoke

2. Realice las estadisticas de resumen que considere necesarias y suficientes para describir las varaibles antes descritas.

3. Evalue la correlación entre las variables los principales metabolitos de ddt.

4.- Proponga un modelo que explique los niveles de DDT en sangre.

    a) Evalúe si el modelo cumple los supuestos de

        Normalidad

        Linealidad

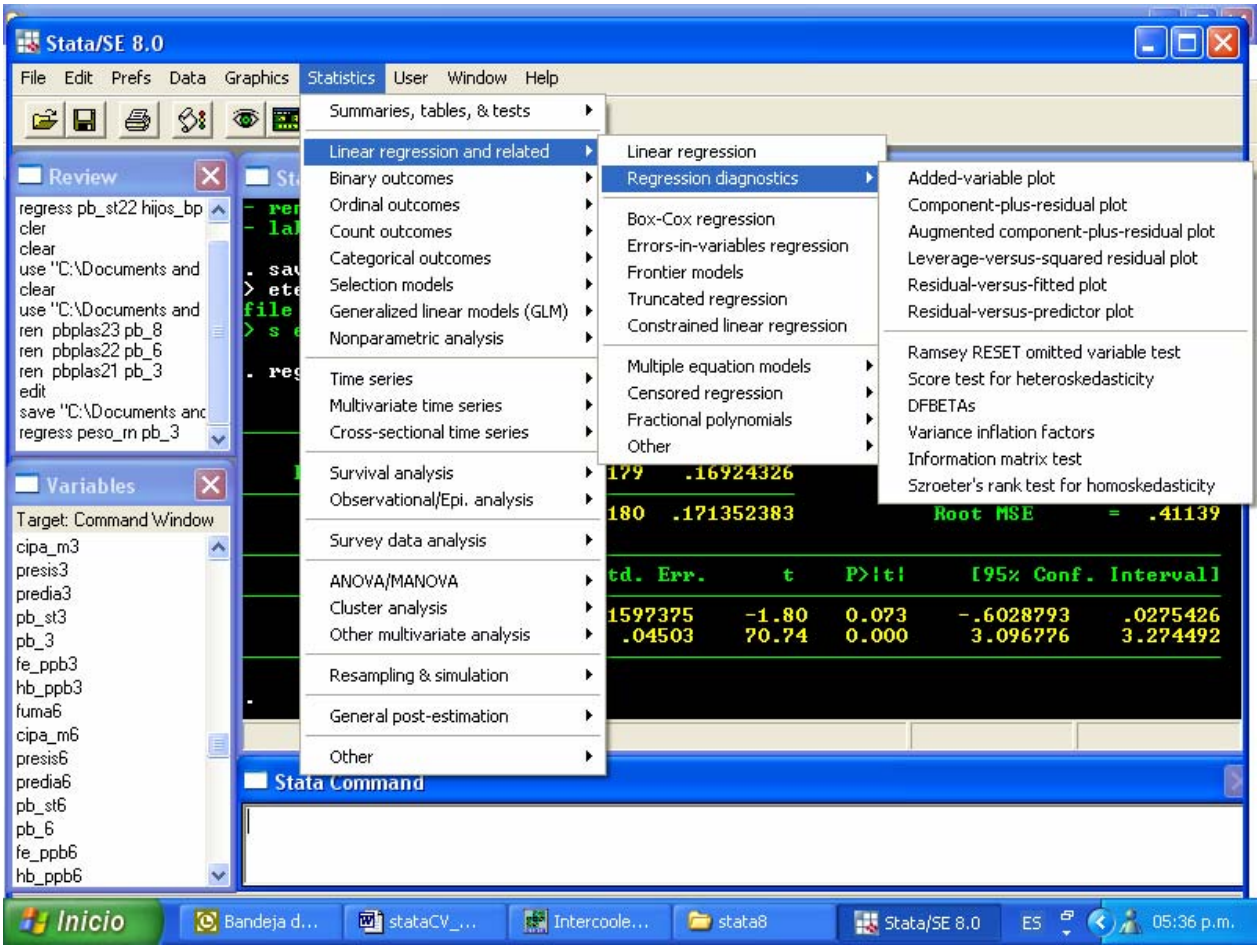
        Homocedasticidad

    b) Mediante técnicas diagnósticas determine si es un buen modelo.

5. Interprete los coeficientes de regresión, R^2 del modelo.

6.- ¿Considera que el modelo al que llegó puede cambiar en base a algunas varaibles no medidas?

- **Diagnóstico utilizando Stata 8.**



A través del menú seleccionamos nuevamente [statistics] dentro del submenu para [Linear regression] seleccionamos [Regressions diagnostics]. Podemos ver una lista de opciones de diagnósico desde gráficos para análisis de residuos hasta opciones para evaluación de puntos influyentes.

Modelos con Regresión logística

Stata también ofrece muchas técnicas para modelar variables dependientes categóricas, variables ordinales y variables censuradas.

En la regresión logística se estima la regresión de una variable dependiente contra las variables independientes, donde la variable dependiente es dicotómica, es decir puede tomar valores de 0 y 1, ya que sigue una probabilidad Bernouli. La regresión logística utilizando en Stata el comando **logistic** se estima Razones de Momios y para ver los coeficientes habría que utilizar la función **logit.**

Un modelo logit o logísitco se estructura de la siguiente manera:

$ln(p/(1-p)= \beta_0 + \beta_1 X$ *En el caso de un modelo simple*

$logit\ p= ln(p/(1-p)= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_p X_p = \underline{X i \beta}$

De este modo:

$$\frac{p}{1-p} = \exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots\ldots\ldots+\beta_p X_p}$$

$$P = \frac{1}{1+\exp^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots\ldots\ldots+\beta_p X_p)}}$$

En el modelo logísitico ***y*** es la variable dicotómica que puede tomar valores de 0 o 1, donde 1 es caso y 0 no caso, 0 tiene una probabilidad p de ocurrir y 0 una probabilidad de 1-p.

La función de riesgo puede tomar valores desde **$-\infty$** a **$+\infty$**

Donde $\underline{xi}$ representa el vector de las variables independientes o factores de riesgo, Ej: *x1*= tabaco, *x2*= alcohol, *x3*= hipertensión, … y β representa el vector de parámetros.

En cuanto a los comantdos (sintaxis) a continuación se presenta un lista parcial de comandos relevantes para utilizarse en regresión logística:

logistic y x1 x2 x3	Estima una regresión logística de {0, 1} variable y sobre los predictores x1, x2 y x3.
Lrtest, s(0) --→ lrtest **est store A----→ lrtest A**	Compara el modelo saturado contra el modelo propuesto a través de las máximas verosimilitudes de ambos modelos.
Lfit,	Presenta una prueba de chi2 de Pearson de máxima verosimilitud del modelo logistico estimado.
lstat	Presenta varias estadísticas de resumen incluyendo una tabla de clasificación.

lstat,lroc y lsens	Se utilizan para evaluar el modelo. El punto de análisis es la clasificación
lroc	Grafica la curva receiver operating characteristic (ROC) Calcula el área bajo la curva.
lsens	Grafíca ambos la sensibilidad y especificidad vs el punto de corte de probabilidades.
lpredict phat	Genera una nueva variable (arbitrariamente nombrada pht) igual a las probabilidades predichas de que y=1 basada sobre el modelo logistico mas reciente.
lpredict dX2, dx2	Genera una nueva variable nombrada dX2(arbitrariamente), la medidia diagnóstica "oportunidad en chi-cuadrada de Pearson," del análisis logístico mas reciente.
mlogit y x1 x2 x3, base (3) rrr nolog	Estima una regresión logística **multinomial** de variables *y* de múltiples categorías sobre las variables *x*. Usa y=3 como la categoría basal de comparación; dando riesgos relativos provenientes de los coeficicientes de regresión.
predict P2, outcome (2)	Genera una nueva variable (arbitrariamente nombrada P2) la cual representa la probabilidad de que y sea igual a 2, basada sobre el análisis mlogit mas reciente.
glm success x1 x2 x3, family (binomial) eform	Estima una regresión logística a partir de un modelo lineal generalizado. Eform se agrega para obtener resultados en forma de OR.

lpredict *newvar*	Predice la probabilidad de que y = 1.
lpredict *newvar,* dbeta	ΔB estadístico de puntos influyentes en B, análogo a Cook´s *D.*
lpredict newvar, deviance	Residuos de Devianza para jth patrón *de x*, dj.
lpredict *newvar, d*x2	Cambio en X^2 Pearson, escrito como ΔX^2 *o* ΔX^2 *P.*
lpredict *newvar,* ddeviance	Cambio en la devianza X^2, *escrito como* ΔD *o* $\Delta X^2 D$. .
lpredict newvar, hat	Influencia de la jth patróbn de x , hj
lpredict newvar, number	Asigna número al patrón de x, j = 1,2,3…j
lpredict newvar, resid	Residuos de Pearson para jth patrón x, rj.
lpredict *newvar, r*standard	Residuos estandarizados de Pearson.

Nota los estadísticos obtenidos de the **dbeta, dx2, ddeviance y hat** no miden la influencia de observaciones individuales como su contraparte en la regresión ordinal. Esto es, logit mide la influencia estadística "patrones de covarianza", es decir la consecuencia de borrar todas las observaciones con estas combinaciones particulares de valores de *x.*

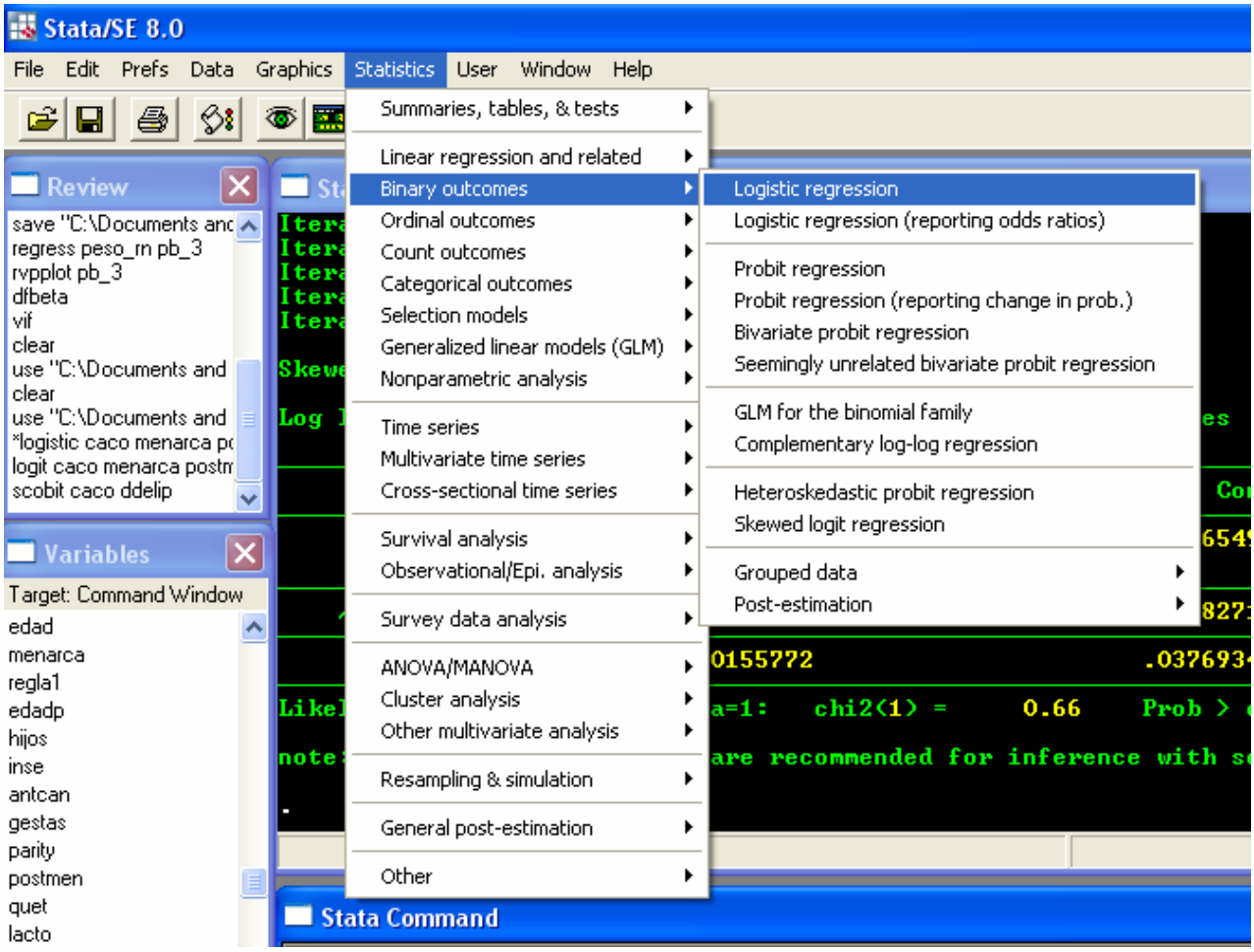
Sesion en Stata

Construcción de un Modelo de Regresión Logística:

En un estudio realizado en la ciudad de México se analizó la relación entre las concentraciones de metabolitos del DDT y el riesgo de cáncer de mama. El análisis siguiente parte de los datos obtenidos en dicho estudio:

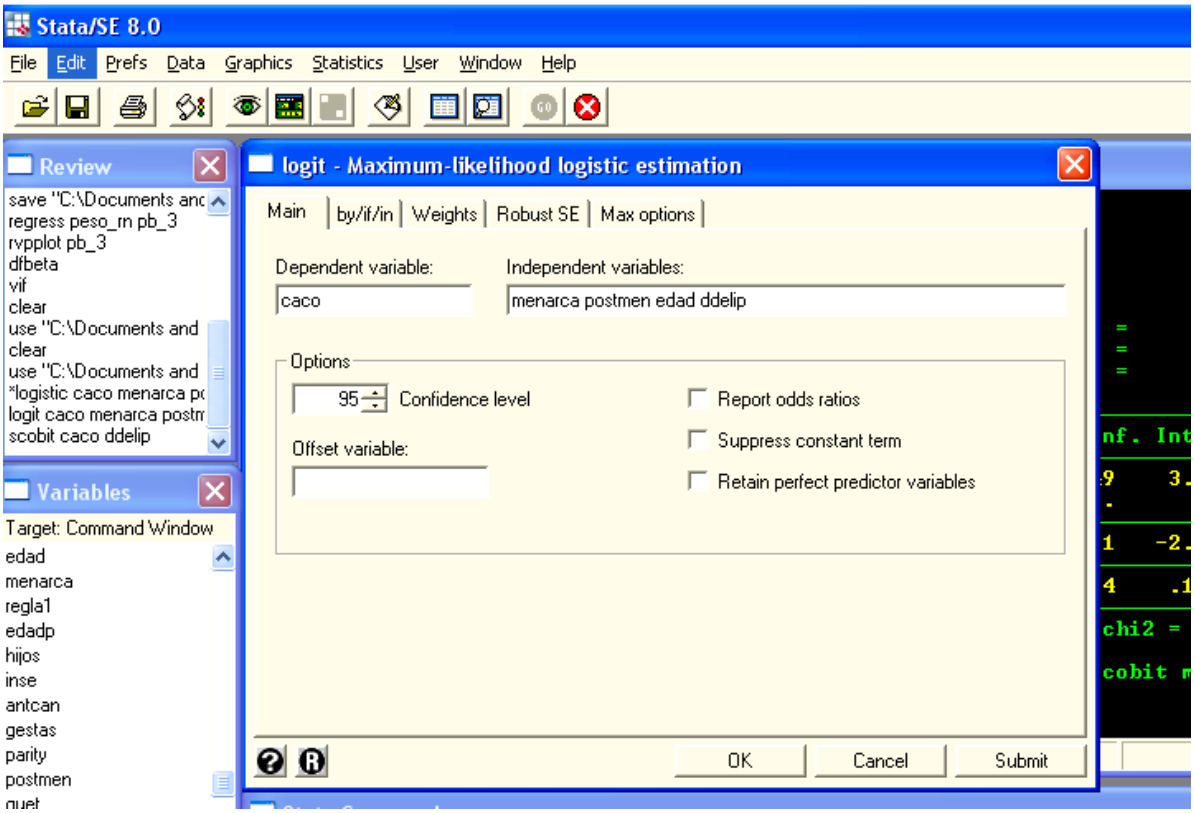
```
. logistic caco menarca postmen edad  ddelip if  ddelip<14

Logit estimates                                   Number of obs   =        242
                                                  LR chi2(4)      =      26.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -154.28118                       Pseudo R2       =     0.0795

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ddelip |   1.200524    .107166     2.05   0.041     1.007831     1.43006
     menarca |   .7108641   .0730371    -3.32   0.001     .5812066    .8694459
     postmen |   .2940498   .1403307    -2.56   0.010      .115398    .7492789
        edad |    1.05168   .0192144     2.76   0.006     1.014686    1.090021
------------------------------------------------------------------------------
```

Si la regresión la hacemos en Stata 8 y deseamos hacerla a través de los menús:

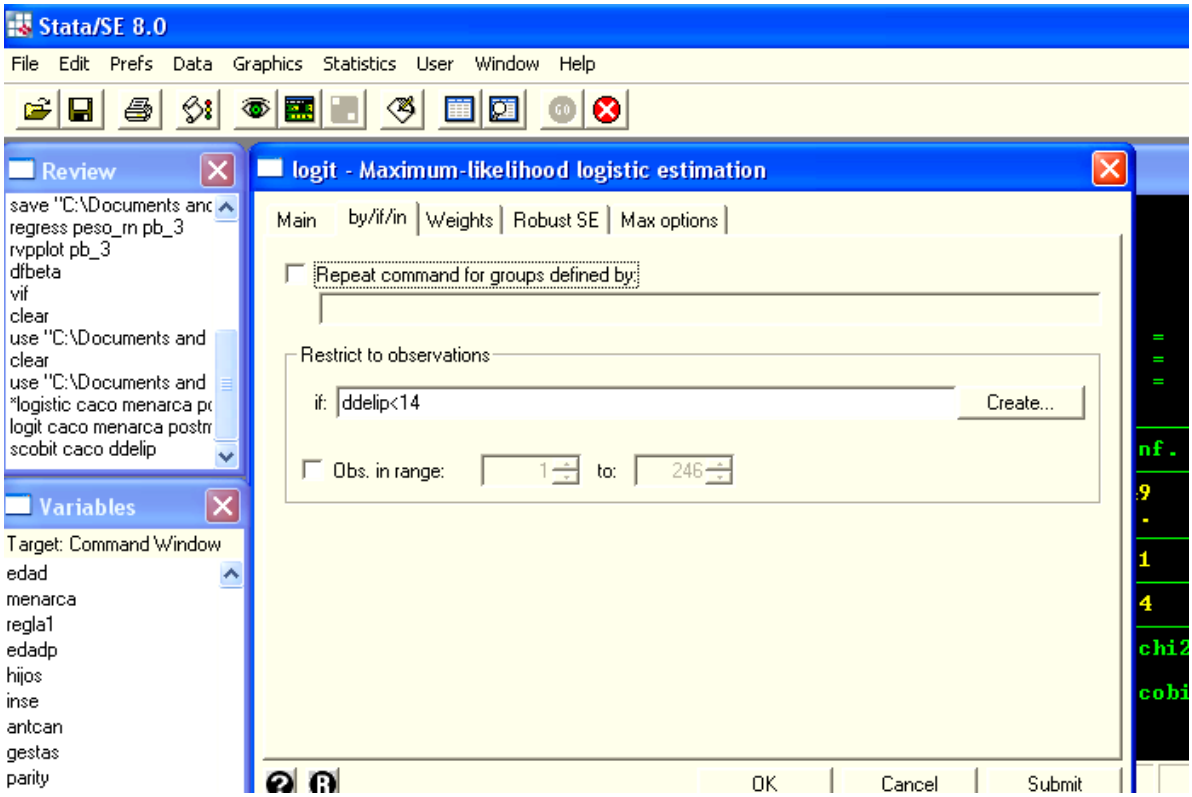


Habrá que seleccionar el submenú para análisis de datos binarios [binary outcomes] ahí encontraremos la opción entre otros para análisis de regresión logística [Logistic regression].

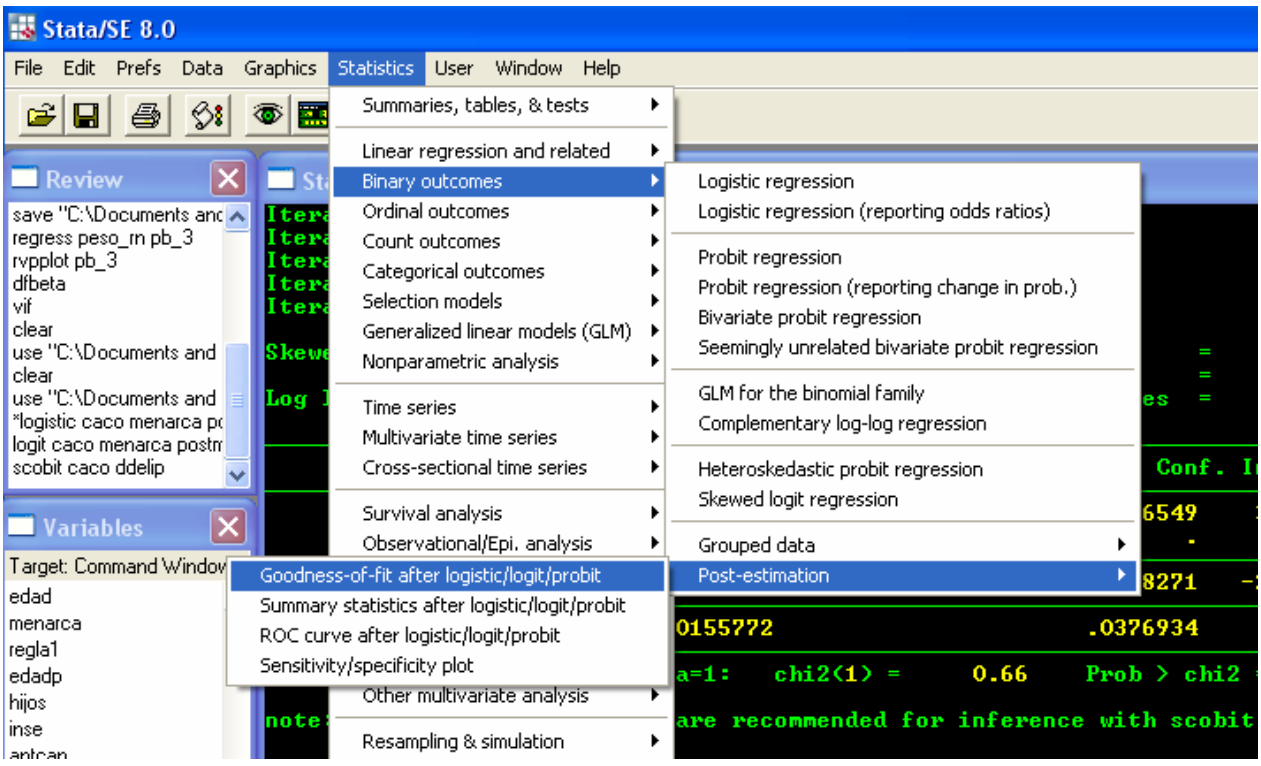
Una vez que entramos en la ventana del submenu, introducimos la variable dependiente e independientes.



Si queremos condicionar por el valor que proponíamos anterioremente, es decir hacer la regresión sólo para cuando la variable ddelip sea menor de 14 entonces en el submenú [by/if/in]:



El diagnóstico del modelo también se puede realizar en stata 8 a través de ventanas (iconos):



➢ **lrtest**

. lrtest,s(0)

Guarda información a cerca del modelo realizado mas recientemente y estima una prueba de razón de verosimilitudes entre pares de máxima verosimilitud de modelos estimados. La opción saving especifica a Stata que guarde con un nombre el resumen de las estadísticas asociadas con el modelo estimado mas recientemente. Generalmente el modelo mas grande se guarda como lrtest,saving(0).
Lrtest, using(0) se emplea entonces en el siguiente modelo con el cual queremos comparar las estadísticas guardadas del modelo anterior. Si no especificamos using(0), Stata por default utiliza el modelo grabado como 0.

Suponiendo que L_0 y L_1 son los valores de log-verosimilitud asociados con el modelo saturado y el modelo propuesto respectivamente. Entonces :

$$X^2 = -2(L_0 \text{ y } L_1)$$

con L_0 y L_1 grados de freedmon, donde d_0 y d_1 son los grados de libertad de freedmon del modelo asociados con el modelo saturado y el modelo propuesto.

La prueba de hipótesis para este estadístico es que las log-verosimilitudes del modelo saturaco y el modelo propuesto son iguales.

```
. logistic caco menarca postmen edad quet ddelip if  ddelip<14

Logit estimates                                   Number of obs   =        242
                                                  LR chi2(5)      =      30.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -152.49782                       Pseudo R2       =     0.0902

------------------------------------------------------------------------------
        caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ddelip |   1.210784   .1089659     2.13   0.034      1.01499    1.444347
     menarca |   .7163815   .0743977    -3.21   0.001     .5844472    .8780988
     postmen |    .264507   .1274727    -2.76   0.006      .102854    .6802255
        edad |   1.052668   .0192988     2.80   0.005     1.015515    1.091181
        quet |   1.059535   .0328435     1.87   0.062     .9970796    1.125903
------------------------------------------------------------------------------
. lrtest,using(0)
Logistic:  likelihood-ratio test                      chi2(-1)     =      -3.57
                                                      Prob > chi2 =          .

. vce

             |  menarca  postmen     edad     quet   ddelip    _cons
-------------+------------------------------------------------------
     menarca |  .010785
     postmen |  .004745  .232252
        edad | -.000292 -.006687  .000336
        quet |  .000059 -.002098  .000029  .000961
```

```
      ddelip |  -.00003 -.006953 -.000317  .000187  .008099
       _cons | -.023116  .276953 -.012356  -.02759 -.007718  1.34723


. logistic caco menarca postmen edad quet ddelip if  ddelip<14

Logit estimates                                   Number of obs   =        242
                                                  LR chi2(5)      =      30.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -152.49782                       Pseudo R2       =     0.0902

------------------------------------------------------------------------------
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
     menarca |   .7163815   .0743977    -3.21   0.001     .5844472    .8780988
     postmen |    .264507   .1274727    -2.76   0.006      .102854    .6802255
        edad |   1.052668   .0192988     2.80   0.005     1.015515    1.091181
       quet |   1.059535   .0328435     1.87   0.062     .9970796    1.125903
      ddelip |   1.210784   .1089659     2.13   0.034      1.01499    1.444347
------------------------------------------------------------------------------

. lrtest,s(0)

. logistic caco menarca postmen edad  ddelip if  ddelip<14

Logit estimates                                   Number of obs   =        242
                                                  LR chi2(4)      =      26.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -154.28118                       Pseudo R2       =     0.0795

------------------------------------------------------------------------------
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
     menarca |   .7108641   .0730371    -3.32   0.001     .5812066    .8694459
     postmen |   .2940498   .1403307    -2.56   0.010      .115398    .7492789
        edad |    1.05168   .0192144     2.76   0.006     1.014686    1.090021
      ddelip |   1.200524    .107166     2.05   0.041     1.007831     1.43006
------------------------------------------------------------------------------

. lrtest,using(0)
Logistic:  likelihood-ratio test                  chi2(1)     =       3.57
                                                  Prob > chi2 =     0.0589

. **no rechazamos la hipótesis nula*
```

- **vce**

vce calcula la matriz de varianza –covarianza de los estimadores (VCE) después de la estimación del modelo VCE puede ser utilizado después de cualquier comando de estimación.

. vce

```
             |  menarca  postmen     edad   ddelip    _cons
-------------+---------------------------------------------
     menarca |  .010556
     postmen |  .004561  .227753
        edad | -.000286 -.006659  .000334
      ddelip | -.000089 -.006585 -.000303  .007968
       _cons | -.020776  .219783 -.011526 -.002832  .553021
```

Este estadístico nos muestra el patrón de varianza covarianza

- **Diagnóstico del modelo de regresión logística:**

Evaluación global del ajuste del modelo.

Después de realizar el modelo y de estar relativamente conformes con él, entonces vamos a evaluar la calidad del mismo.

Estrategia:
Evaluación global del modelo.
Revisión de gráficas diagnósticas.
Revisión de residuos

En regresión logística, la validez de la **X^2 de Pearson** depende del número de "patrones de las covariables".

Si J: Número de valores distintos observados del vector $\underline{x}$ y p: número de parámetros en el modelo, entonces

$$X^2 \text{ de Pearson} \sim X^2_{(J-p)}$$

Pero si $J \approx n$, lo que sucede frecuentemente cuando se tienen covariables continuas, entonces los *p-values* obtenidos son poco confiables, por lo que se propone una alternativa:

Prueba de Hosmer y Lemeshow:

Generar grupos basados en las probabilidades estimadas por el modelo, concretamente en sus percentiles.

Proponen una estadística equivalente a la X^2 de Pearson pero que se distribuye como

$$X^2_{(g-2)}$$

donde g es el número de grupos generados. Comúnmente g=10.

Ejemplo de comandos:

- **lfit**

. lfit

```
Logistic model for caco, goodness-of-fit test

       number of observations =       242
 number of covariate patterns =       242
           Pearson chi2(237) =       239.64
                 Prob > chi2 =         0.4398
```

Prueba de Hosmer y Lemeshow x^2 (g-2). Presenta una prueba de chi^2 de Pearson de máxima verosimilitud del modelo logístico estimado: frecuencias observadas vs esperadas de y=1, usando celdas definidas por el comportamiento de la(s) covariable(s) (variables x). Cuando el patrón de x es grande, se pueden agrupar entonces de acuerdo a probabilides estimadas. lfit, group(10) puede estimar la prueba con 10, aproximadamente igual al tamaño del grupo.

```
. lfit,group(10)

Logistic model for caco, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

       number of observations =       242
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         6.96
                  Prob > chi2 =         0.5408
```

También se propone, como técnica diagnóstica, construir la tabla de clasificación de la variable dependiente vs un predictor dicotómico las cuales se utilizan cuando el estudio sobre el cual estimamos la ecuación logit es un estudio de seguimiento o longitudinal y en los cuales podemos estimar B_0. Algunas de estas pruebas son.

- **lstat**

  Presenta varias estadísticas de resumen incluyendo una tabla de clasificación, sensibilidad y especificadad para el modelo estimado por logistic, logit o probit.

```
. lstat
Logistic model for caco

             -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        71            42  |        113
     -     |        46            83  |        129
-----------+--------------------------+-----------
   Total   |       117           125  |        242

Classified + if predicted Pr(D) >= .5
True D defined as caco ~= 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   60.68%
Specificity                     Pr( -|~D)   66.40%
Positive predictive value       Pr( D| +)   62.83%
Negative predictive value       Pr(~D| -)   64.34%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   33.60%
False - rate for true D         Pr( -| D)   39.32%
False + rate for classified +   Pr(~D| +)   37.17%
False - rate for classified -   Pr( D| -)   35.66%
-------------------------------------------------
Correctly classified                        63.64%
-------------------------------------------------

Cambiando el punto de corte:
. lstat, cutoff(0.7)
```

Los símbolos en la tabla de clasificación tienen las siguientes mediciones:

D ocurrencia del evento de interés (esto es Y=1). En este ejemplo, D indica que ocurre: la enfermedad (caso de cáncer de mama)

~D No ocurrencia del evento (es decir y=0). En este ejemplo, ~ D corresponde a la ausencia de la enfermedad x (en los controles)

+ La probabilidad predicha por el modelo logístico es mayor o igual al punto de corte. Debido a que nosotros utilizamos por default el 0.5 + esto indica que el modelo predice una probabilidad de 0.5 o mas extrema tener la enfermedad x.

- La probabilidad predicha es menor que la del punto de corte. Aquí, el – indica que el modelo predice una probabilidad media menor de 0.5 de tener la enfermedad x (la probabilidad es baja).

Por default lstat emplea una probabilidad de 0.5 como punto de corte (sin embargo se puede cambiar esta al adicionar la opción cutoff( ).

- **lroc**

curva ROC

Grafica la curva receiver operating characteristic **(ROC**). Calcula el área bajo la curva. Esta es una gráfica de la sensibilidad contra (1-especificidad), es decir, grafica el número de casos positivos correctamente clasificados (predichos por el modelo) contra el número de no casos que fueron clasificados incorrectamente como casos, así como la clasificación del entrecruzamiento c. Esta herramienta gráfica es muy útil cuando el objetivo del análisis fue la clasificación.

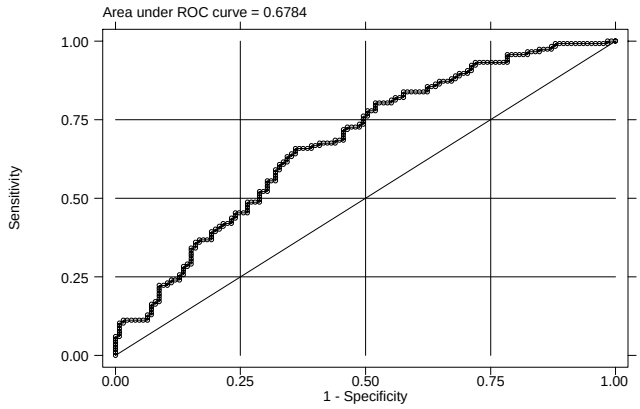
El área bajo la curva se usa como medida del valor predictivo.

Ejemplo de comandos:

. lroc

Logistic model for caco

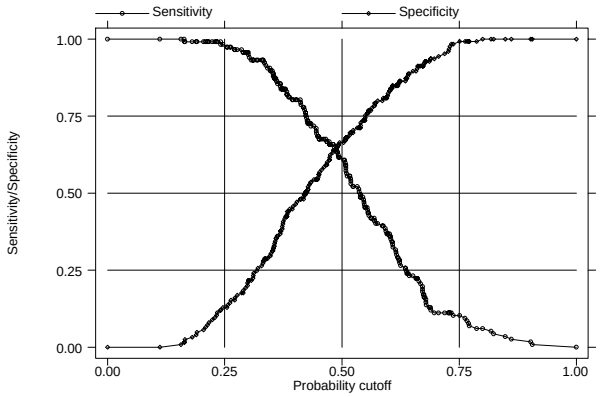
number of observations = 242
area under ROC curve = 0.6784



El área bajo la curva es el área sobre lo mas bajo de esta gráfica, y es determinada por integración de la curva . Los vértices de la curva son determinados por ordenación de los datos de acuerdo al índice predicho y la integral es calculada utilizando la regla trapezoide.

Un modelo sin poder predictivo tendría una curva con inclinación de 45° y el área bajo la curva sería 0.5. El modelo con mayor poder de predicción formaría un arco y el área bajo la curva sería 1.

➢ **lsens**
Lsens también grafica sensibilidad y especifidad.



. lsens

La gráfica muestra en el eje *y* la sensibilidad y la especificidad contra la probabilidad de entrecruzamiento c en el eje x.. Esta equivale a los datos de lstat si cambiáramos los datos del punto de corte del 0 al 1.

Para nuestro modelo la sensibilidad y la especificidad son demasiado bajos, esto querría decir que mi modelo no esta estimando correctamente los casos, sin embargo el diagnóstico con estas pruebas, son preferentemente útiles en el caso de estudios de clasificación como es en el caso de estudios de tamizaje.

Es importante mencionar que en cuanto a diseño:

Aunque el modelo logístico puede aplicarse a un estudio de casos y controles y uno transversal, es importante reconocer algunas **limitaciones:**

- En un **estudio de seguimiento**, el modelo logístico puede usarse para predecir el **riesgo de un individuo** de padecer la enfermedad, dados valores específicos de las variables independientes.

- En un estudio de seguimiento, el parámetro de regresión $_0$ puede estimarse de manera válida porque se conoce la fracción de muestro.

- La estimación adecuada de B_0 permite estimar el riesgo individual de contraer la enfermedad.

- En un estudio de **casos y controles o un estudio transversal**, sólo se pueden obtener estimaciones del **cociente de momios.**

- En un estudio de casos y controles o un estudio transversal, el parámetro B_0 no puede estimarse de manera válida sin que se conozca la fracción de muestreo.

- Sin la estimación adecuada de B_0 no podemos obtener un buen estimador del riesgo.

Cuando las variables por las que se ajusta se consideran fijas pero no se especifican en su totalidad:

- Se puede usar la regresión logística para obtener directamente un estimador del OR pero no podemos estimar el riesgo relativo.

_ Se puede estimar el RR indirectamente ya que el OR iguala al RR si la enfermedad es rara:

Ejercicio práctico.
Regresión logística

1) Haga un análisis exploratorio y bivariado de la información que se le presenta.
2) Mediante regresión logística estime el mejor modelo que prediga el OR de enfermar entre los expuestos a Asma. Compare sus resultados con los del artículo de referencia al respecto. Obtenga intervalos de confianza del 95 % para la Razón de Odds.
3) Justifique, si es el caso, la inclusión en el modelo de las variables de control.
1) Aplique los comandos necesarios para realizar el diagnóstico del modelo propuesto.

Anexos:

Artículos de referencia que usará para los Ejercicios y talleres.

Secciones de del Manual de STATA 8.0

Bases a utilizar:

1) Factores predictores de los niveles de DDT en sangre en pobplación masculina en Chiapas.
2) Factores de riesgo para Asma en niños escolares de la cd de México.