

Denny, Kevin

Working Paper

Basic Stata graphics for economics students

UCD Centre for Economic Research Working Paper Series, No. WP18/11

Provided in Cooperation with:

UCD School of Economics, University College Dublin (UCD)

Suggested Citation: Denny, Kevin (2018) : Basic Stata graphics for economics students, UCD Centre for Economic Research Working Paper Series, No. WP18/11, University College Dublin, UCD School of Economics, Dublin

This Version is available at:

<https://hdl.handle.net/10419/198564>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

UCD CENTRE FOR ECONOMIC RESEARCH

WORKING PAPER SERIES

2018

Basic Stata Graphics for Economics Students

Kevin Denny, University College, Dublin

WP18/11

July 2018

**UCD SCHOOL OF ECONOMICS
UNIVERSITY COLLEGE DUBLIN
BELFIELD DUBLIN 4**

Basic Stata graphics for economics students

Kevin Denny

School of Economics & Geary Institute for Public Policy

University College Dublin

3rd July 2018¹

This paper provides an introduction to the main types of graph in Stata that economics students might need. It covers univariate discrete and continuous variables, bivariate distributions, some simple time plots and methods of visualising the output from estimating models. It shows a small number of the many options available and includes references to further resources.

Keywords: Stata, graphics, data visualization

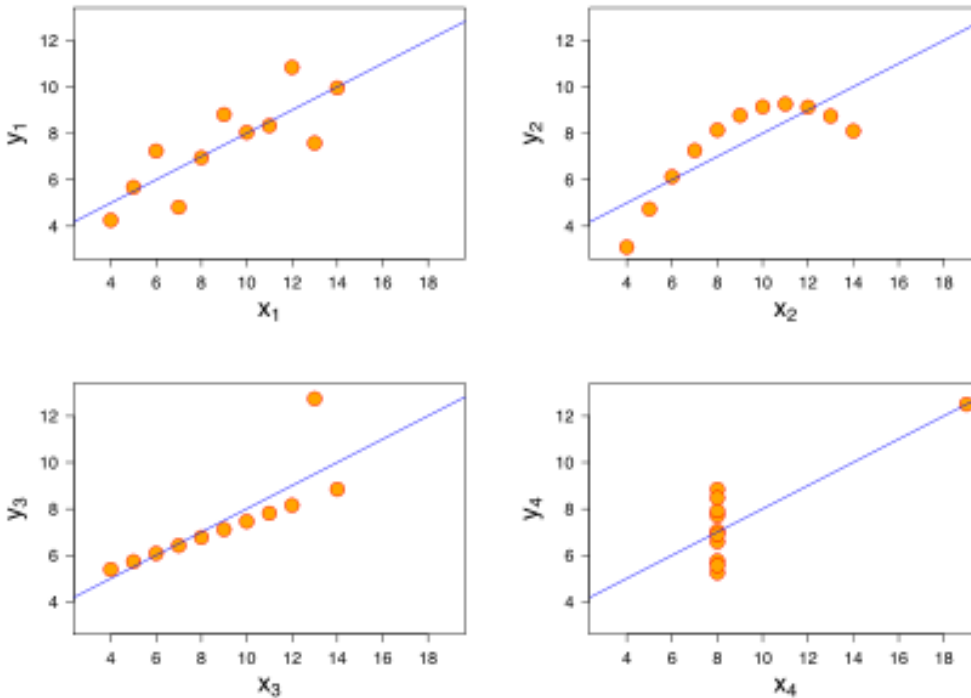
JEL codes: A2, C87, Y10

¹ This guide was written with my students in mind. I am releasing it into the wild as others may find it helpful but mainly as a commitment device to prevent me extending it further. Comments to kevin.denny@ucd.ie

1. Introduction

It has often been said that a picture paints a thousand words. This is certainly true when it comes to data analysis. There are two good reasons to acquire some skills in graphing your data: (1) Graphical methods are a powerful way for a researcher to explore data and (2) graphs can be a very useful way of illustrating your data and results whether it is in a presentation, a project or a thesis.

To motivate the first reason above, consider the following set of graphs:



These four graphs are collectively known as Anscombe's Quartet. You may be surprised to learn that the x variables in these four graphs have the same mean and the same variance. This is also true of y and, moreover, the covariance between x and y is the same & hence the regression line is the same. Clearly they are very different relationships. Without graphing the data, you would probably never know.

In this document I show how to use Stata to generate some of the key graphs that economics students should know about and should consider using in their projects, presentations and theses. There are several good online treatments of Stata graphics (listed at the end). Stata's Youtube channel has videos on graphics which are excellent. The book by Michael Mitchell is a fantastic resource which you could also draw on. Andrew Jones' guide, though designed for health econometrics, is of general interest if you are using Stata. Jesse Shapiro has a great set of slides on preparing a good applied microeconomics presentation but is relevant for any applied economics talk & probably other social sciences. Here I am going to outline the main methods that I think economics students should know at a minimum. Along the way I show a few of the many options available to whet your appetite. The definitive source of information is the Stata Graphics Manual which is a mere 739 pages long. A classic text on data visualization and graphics is Tufte (2001). For a shorter guide targeted at economists see the paper by Schwabish (2014).

All of the datasets I use here are either available online & can be accessed in Stata using the **webuse** command or they are provided with Stata and can be accessed using **sysuse**. To switch from one dataset to another you need to use **clear** first. Stata commands will be in **bold**. A basic knowledge of Stata is required. There are two ways to create graphs in Stata. You can either (a) use a written command which can be done interactively in the command line or written in a do-file or (b) you can use the dialogue boxes/pull-down menus at the top.

A nice feature is that if you use the dialogue box to create a graph Stata will show you the equivalent syntax in the output window so you can learn how to generate the graph. You could copy the syntax into a do-file so you can repeat the exercise. So I tend to use the pull-down menus to experiment until I get the graph looking like I want. Then I copy the syntax that generates it from the output window into my do-file so I can replicate it later.

When Stata produces a graph for you on the screen click on “file” at the top left: you can either save it or you can open the editor to make further changes. Stata’s native format for graphs is .gph. If you want to include it in a Word document or a Powerpoint file for example you need to save it to a format like portable network graph (.png), a tiff file (.tiff), a Windows meta file (.wmf) or a postscript file (.ps). You may need to experiment saving to different formats to get something that works with your document. As .wmf files only work with Windows, if you are a Mac user or you are collaborating with one it might be best to avoid them: .emf is better. If in doubt I recommend saving the graph as .png. Postscript files can end up taking a lot of space if there are a large number of data points in your graph.

A feature I will not discuss here is that you can create two graphs separately and then combine them into one graph. Koffman (2015) has a few slides on this or **help graph combine**.

The Stata graphics editor has numerous options & you can customize the graph in many many ways. It is beyond the scope of this document to describe how. Here I am mostly going to use the graph commands that come with Stata. However there are some good user-written commands for Stata graphics that are freely available online. You can find and download them within Stata using the **findit** command. Here I will draw on four of these: **binscatter**, **coefplot**, **fabplot** and **vioplot**. To download the first of these say, just type:

findit binscatter in the command line or **ssc install binscatter**. Hit return and follow the steps.

As these user created commands are occasionally revised, it is worth using the **adoupdate** command periodically to ensure you have the latest version.

2. Distributions

When you are analysing data it is essential that you carefully explore the data before you get stuck into modelling using it using econometric methods. You really need to get to know your data. There are a few reasons for this. One is that exploring the data will sometimes show up anomalies, for example there might be crazy values like missing values code as -1 or 99. The main reason is to get a sense of what the basic patterns are. This is particularly the case for variables that you create from the raw data. It is very easy to make a mistake – even experienced users do - so if you generate a new variable you want to check does it look sensible.

2.1 Univariate

We will first consider looking at the distribution of a single variable. You should certainly have a good look at your key variables before you do any modelling.

2.1.1 Discrete

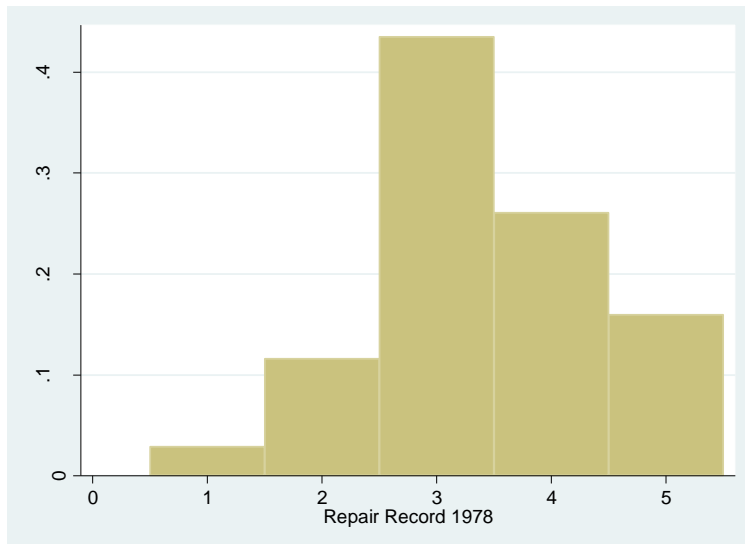
For a discrete variable you should use a histogram:

webuse fullauto

ta rep78 generates a table of this discrete variable. This is fine as far as it goes and you may want to include a table like this in your document particularly if this is your dependent variable. Note that to keep the table nicely aligned as it is in Stata you need to use the Courier font. However it may be hard to get a sense of the distribution simply by looking at the table. If you are preparing a presentation, for example, you want the audience to easily grasp what the data looks like. Let's graph it next.

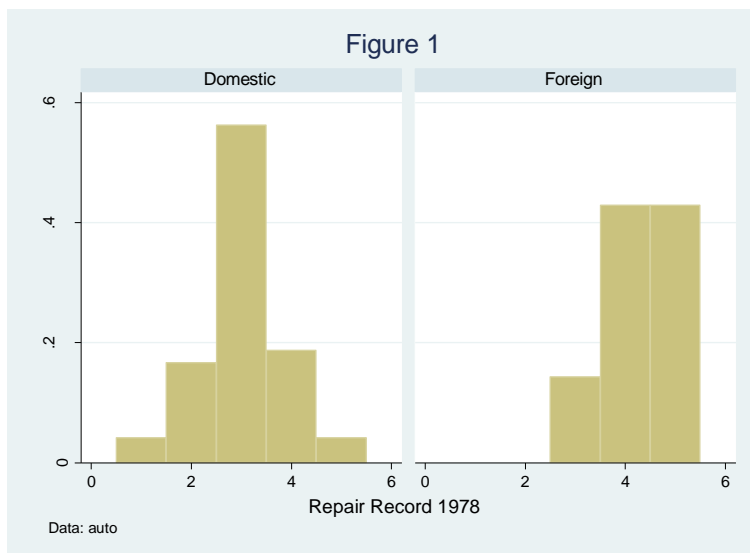
Repair				
Record 1978		Freq.	Percent	Cum.
-----+-----				
Poor		2	2.90	2.90
Fair		8	11.59	14.49
Average		30	43.48	57.97
Good		18	26.09	84.06
Excellent		11	15.94	100.00
-----+-----				
Total		69	100.00	

histogram rep78, discrete fraction creates a histogram where the heights of the bars gives the fractions in each category. If you want to show percentages instead simply replace **fraction** with **percent**. If you want absolute frequencies use **frequency**.



You can generate a graph with separate histogram for different groups beside each other:

histogram rep78, discrete fraction by(, title(Figure 1) note(Data: auto)) by(foreign)

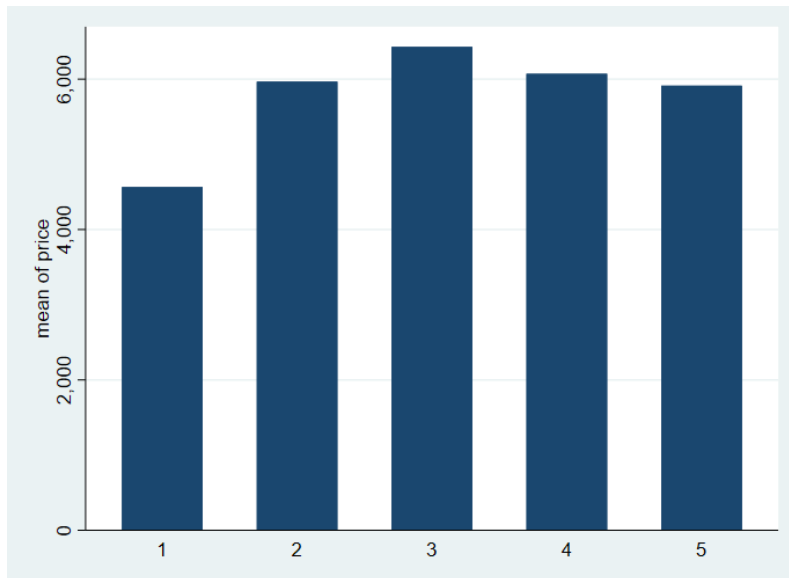


See how I have added a title and a note at the bottom?

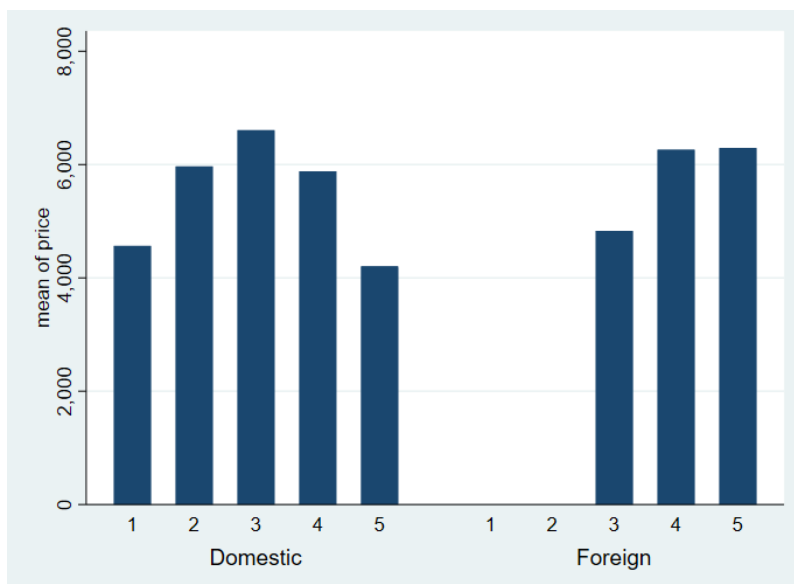
Another way of illustrating shares across categories is the pie chart. Pie charts are not to everyone's taste and some are less than helpful so use them wisely. Try **graph pie, over(rep78)** as an alternative to a histogram. If you would like to see what % each slice contains: **graph pie, over(rep78) plabel(_all percent)**

A bar chart is a useful way of comparing some characteristic of a variable (like the mean) across different categories of a variable. For example

graph bar (mean) price, over(rep78)



graph bar (mean) price, over(rep78) over(foreign) shows the mean across categories of two variables



An alternative way of doing this where the two graphs are separate is:

graph bar (mean) price, over(rep78) by(foreign)

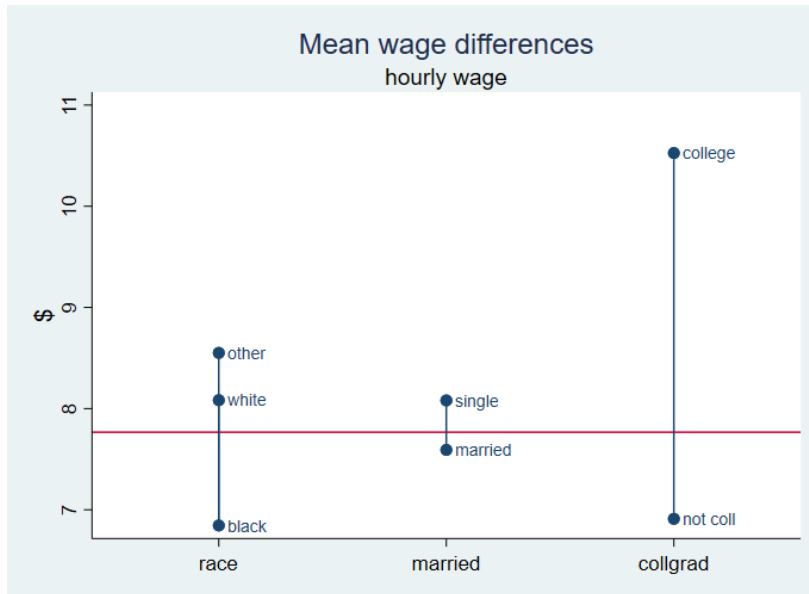
Note that bar charts aren't just for means. You can use them to compare other statistics like medians, standard deviations, maxima etc. When you use the pull-down menu there is a "Statistic" option beside the variable. If you want the bars to be horizontal, replace **bar** with **hbar**. The mean is the default. To generate a bar chart of standard deviations across the categories of rep78 for example try:


```
graph bar (sd) price, over(rep78)
```

Stata has another type of graph which can be used to illustrate differences in means or medians of a continuous variable across categories of one or more variables.

```
sysuse nlsw88, clear
```

```
grmeanby race married collgrad, summarize(wage) ytitle($ ytitle(, size(medlarge)) title(Mean wage differences) subtitle(hourly wage)
```



This example shows in one graph the differences in average earnings between the categories of three variables. A glance at this graph suggests that education seems to be more important than marital status. Note that I have added a title to the Y axis (the “\$”) and changed the size of the title. The command line I used continues onto a second line. If you were using this in a do-file Stata has to know to read it as one line. Entering it like this in your do-file will work:

```
grmeanby race married collgrad, summarize(wage) ytitle($ ytitle(, size(medlarge)) /*
```

```
*/ title(Mean wage differences) subtitle(hourly wage)
```

The `/* */` comments out the end of the line so Stata just reads it as one line. Adding **median** after the “,” (the comma) in the command will show the medians instead e.g.:

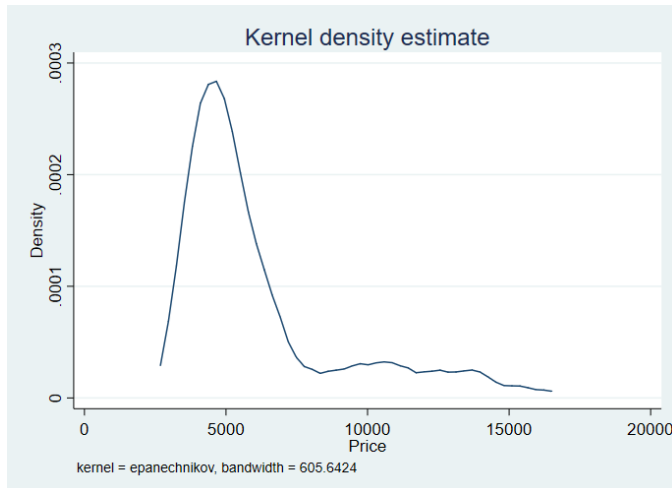
```
grmeanby race married collgrad, summarize(wage) median
```

.

2.1.2 Continuous

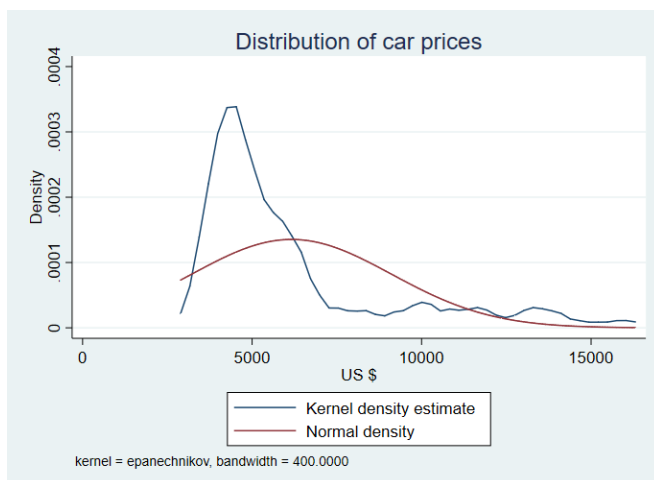
There are a few ways to show the distribution of a continuous variable. You can use a histogram as shown on page 5. My preferred method is to generate something called a kernel density function.

kdensity price



The smoothness of the density is controlled by a bandwidth parameter. Stata calculates a default parameter & reports it. You can see it is 605.6 in the above example. You can over-ride this if necessary using the bandwidth option. For example by reducing the bandwidth to 400 you make it less smooth. Be careful not to over-smooth i.e. setting the bandwidth so high that you remove key features of the data. There is a handy download **akdensity** which allows the bandwidth to adapt optimally to how much data there is at a particular part of the distribution. The “norm” option superimposes a normal distribution which can be useful if you have reasons to believe that the variable should be normal:

kdensity price, bw(400) norm xtitle(US \$) title(Distribution of car prices)

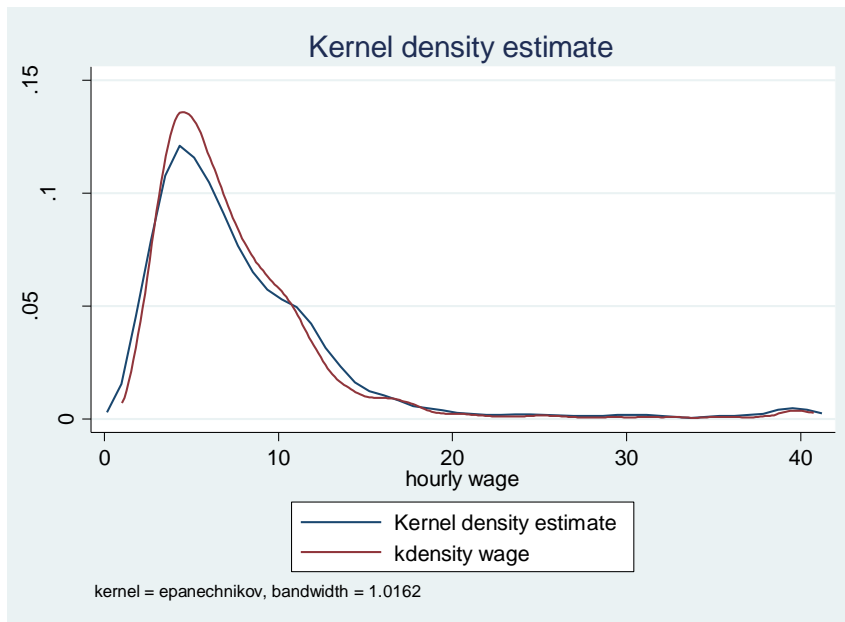


Note how I have added a title and a label for the x axis. Sometimes you may wish to superimpose two densities on top of each other. For example if you are looking at the distribution of earnings, it might be useful to compare the earnings of men and women or married and unmarried people.

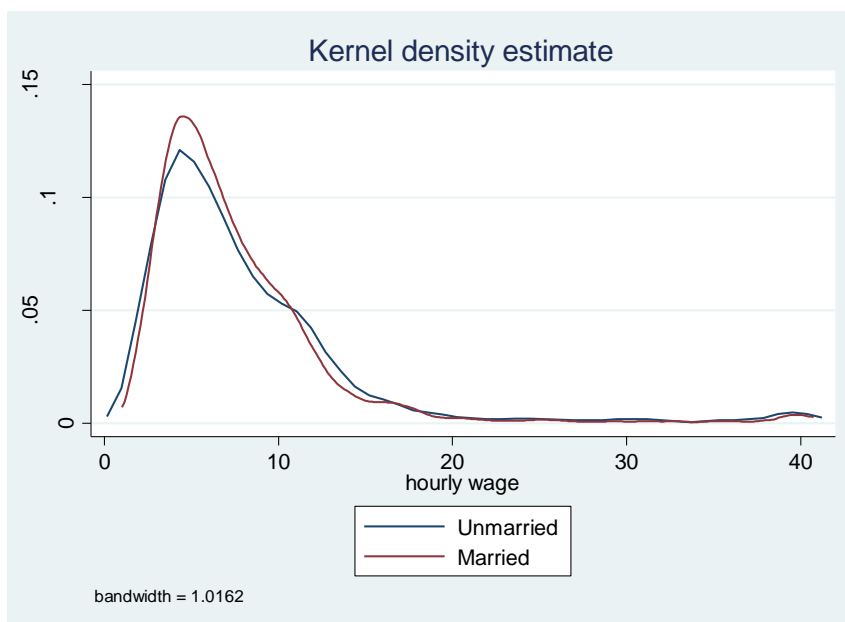
Use the nlsw88 data

`sysuse nlsw88, clear`

`kdensity wage if married==0 , addplot(kdensity wage if married==1)`



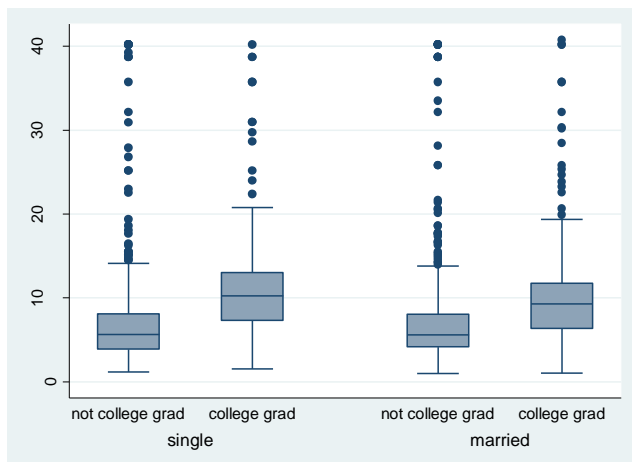
The density shown in blue is for the unmarried (**married==0**) and red is for the married. The legend below the graph is not very helpful unfortunately and you will need to edit it in the graphics editor so you can end up with something like:



An alternative way of comparing the distribution of a continuous variable across categories of some other discrete variable is a boxplot (aka “box and whisker” plot). The middle line in the box shows the median, the bottom and top of the box show the 25th & 75th quartiles, respectively. So the height of the box is the IQR, the inter-quartile range.

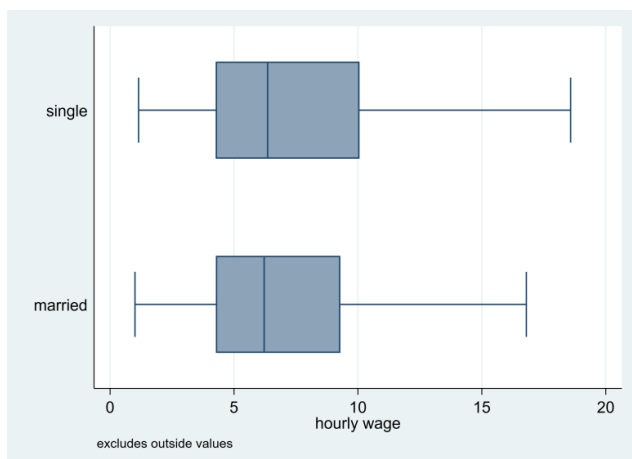
The “whiskers” from the box extend vertically to the upper and lower adjacent values. Their definition is somewhat tricky: think of a value $U = \text{the } 75^{\text{th}} \text{ percentile} + (3/2) * \text{IQR}$ and $L = \text{the } 25^{\text{th}} \text{ percentile} - (3/2) * \text{IQR}$. The upper adjacent value is the value of x which is $\leq U$. The lower adjacent value is defined as the value of x which is $\geq L$.² Essentially, the whiskers pick up the extent to which the distribution is spread out outside of the IQR. Points outside this range are shown as dots. In the example below we show the distribution over categories of two variables, college graduate status and marital status:

graph box wage , over(collgrad) over(married):



The points at the top of each plot show that this variable is right (positively) skewed. These points can sometimes distort the diagram so if you wish to omit them adding **noout** at the end of the line will do. The command below will plot the bars horizontally and removes the outliers:

graph hbox wage, over(married) noout

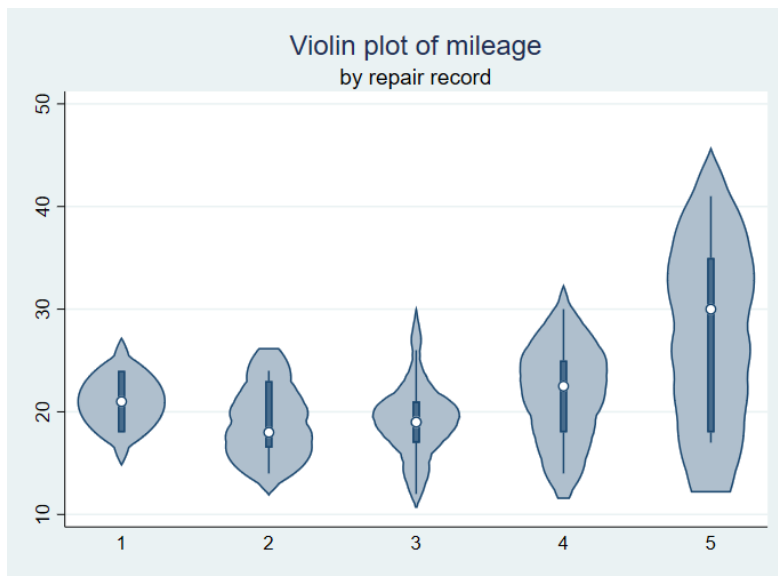


² Note that this is Stata’s implementation of box plots which goes back to the influential work of John Tukey (1977). Other approaches are possible. For example some have the whiskers extend to the 10th & 90th percentiles instead.

Violin plots are a useful way of combining box plots and densities invented by Hintze & Nelson (1998). First download the **vioplot** package ("**ssc install vioplot**"). Then, using the auto dataset:

```
vioplot mpg , over(rep78) title("Violin plot of mileage") subtitle("by repair record")
```

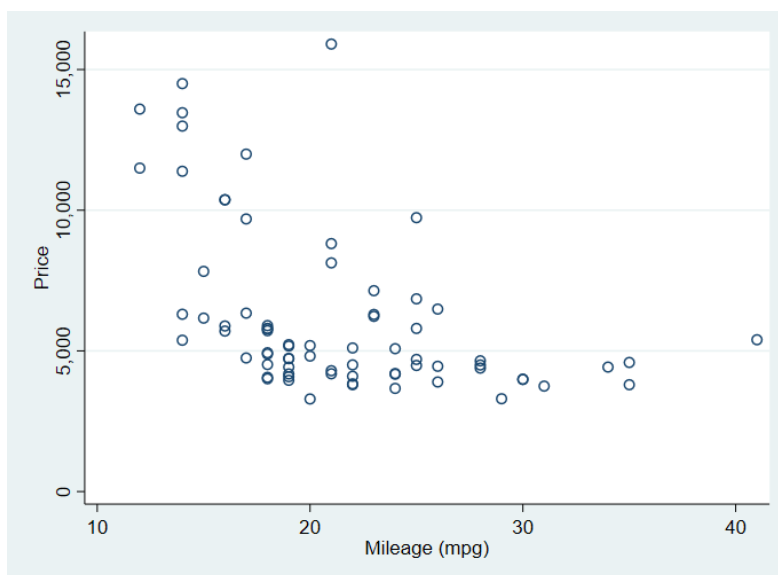
The white dot is a marker for the median, the thick line shows the interquartile range with whiskers extending to the upper & lower adjacent values (as defined above). This is overlaid with a density of the data. Violin plots contain a lot of information though you may need to fiddle with the options to get it looking right.



2.2 Bivariate

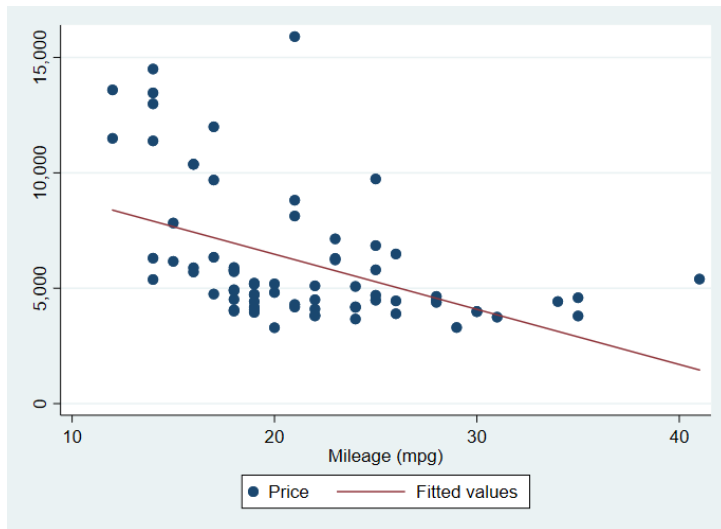
To examine a bivariate distribution, start with a scatterplot

```
twoway (scatter price mpg , msymbol(Oh))
```



I used the **msymbol** option above to change the dots to an “O”. Scatterplots are not always very illuminating and you may want to adjust them. It is simple to fit and plot a linear regression to this data:

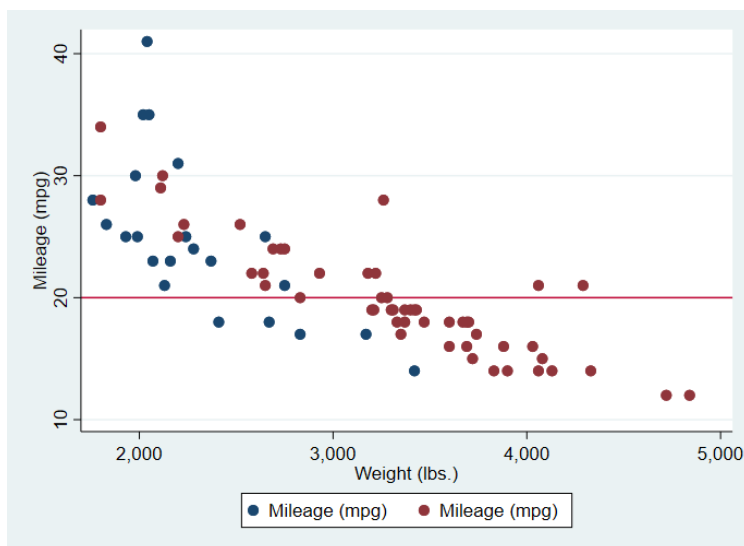
twoway (scatter price mpg) (lfit price mpg)



If you replace **lfit** with **lfitci** it shows the confidence interval around the line. Using **qfit** instead fits a quadratic curve (and hence **qfitci** instead show the confidence intervals).

If you want to show a scatterplot for two different subsets of the data try:

scatter mpg weight if foreign || scatter mpg weight if !foreign , yline(20)



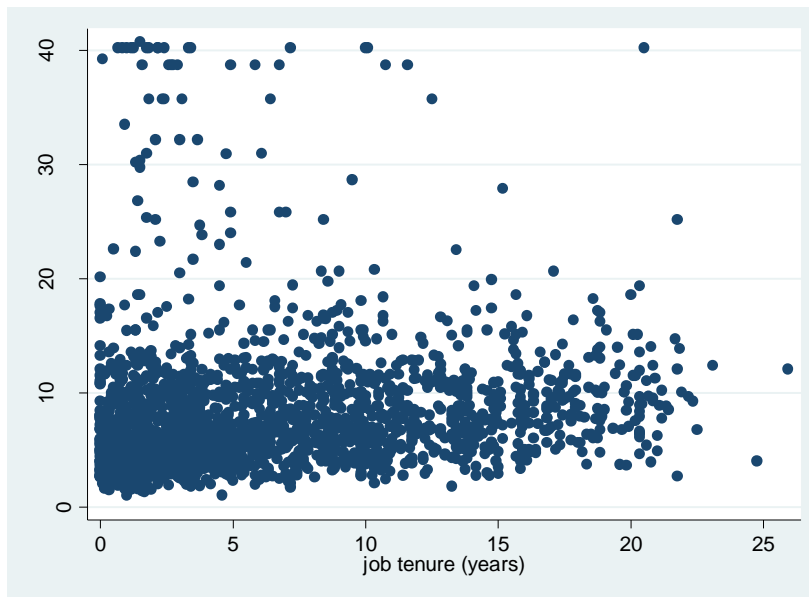
The blue dots refer to the first named subset (foreign cars). I have added a line corresponding to $y=20$ with the **ylines(20)** option. You can have more than one line: using say **xline(3000 4000)** would create vertical lines corresponding to those values of x. You can use this if there is a particular x or y value that is important (e.g. a particular year). This syntax is another way of creating the same basic diagram:

twoway (scatter mpg weight if foreign) (scatter mpg weight if !foreign)

Note the variable `foreign` is either 0 or 1 so **!foreign** means “not foreign” i.e. **foreign==0**. Sometimes a scatter plot has so many datapoints that you end up with a graph that’s not very illuminating. Let’s switch to the `nlsw88` data set to see this:

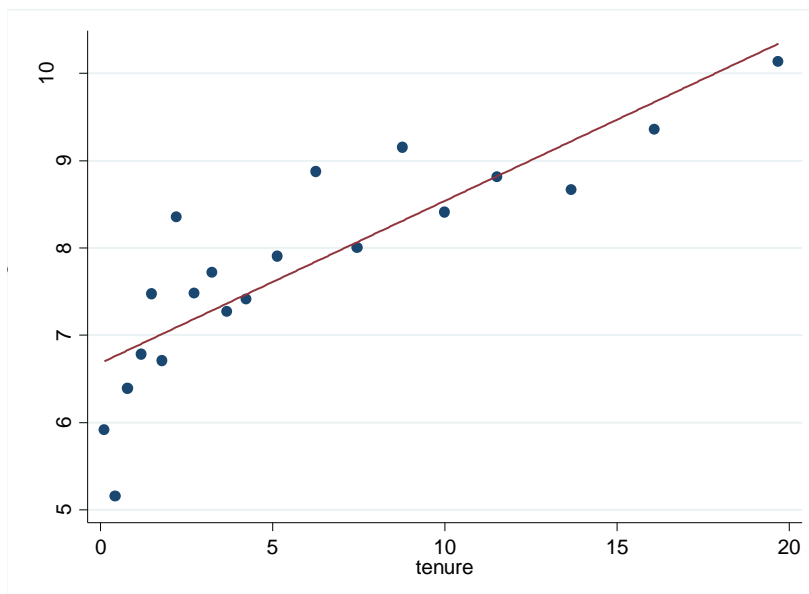
```
sysuse nlsw88, clear
```

```
scatter wage tenure (note this is the same as twoway (scatter wage tenure) )
```



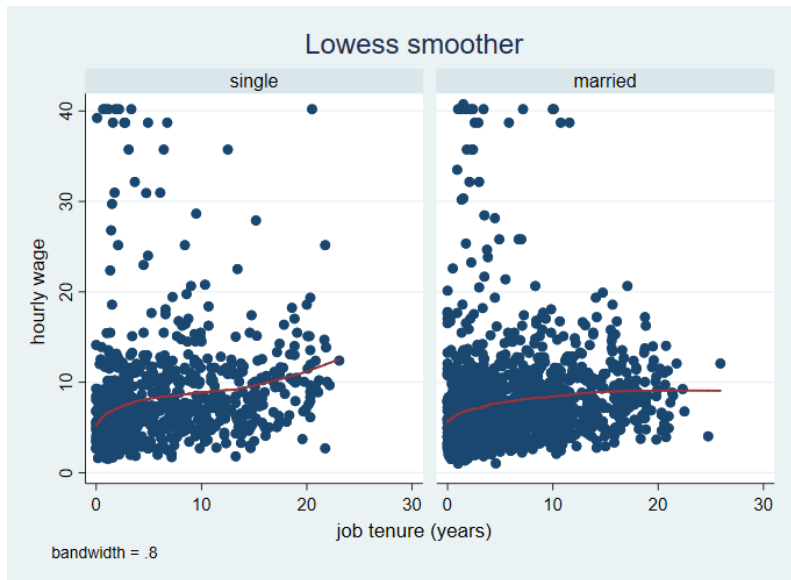
Not terribly clear is it? There is a handy Stata download called **binscatter**. It groups the x-axis variable into equal-sized bins, computes the mean of the x-axis and y-axis variables within each bin, then creates a scatterplot of these data points.

```
binscatter wage tenure
```



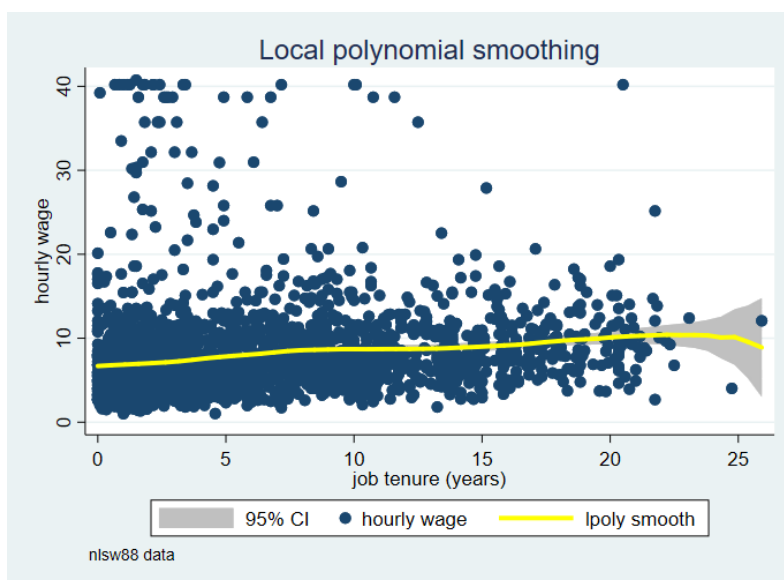
What this brings out is there is a positive relationship between wages and job tenure. **binscatter** has many nice feature that you can explore. It is possible to generate scatterplots of 3 variables i.e. with 3 dimensional graphs using a download **graph3d**. These are trickier to get in a form that is helpful. An alternative way of illustrating the bivariate relationship between two variables is to fit a curve to the data. Stata has several ways of doing this. A popular method is called lowess (for locally weighted scatterplot smoothing).

lowess wage tenure, by(married) lineopts(lwidth(medthick))



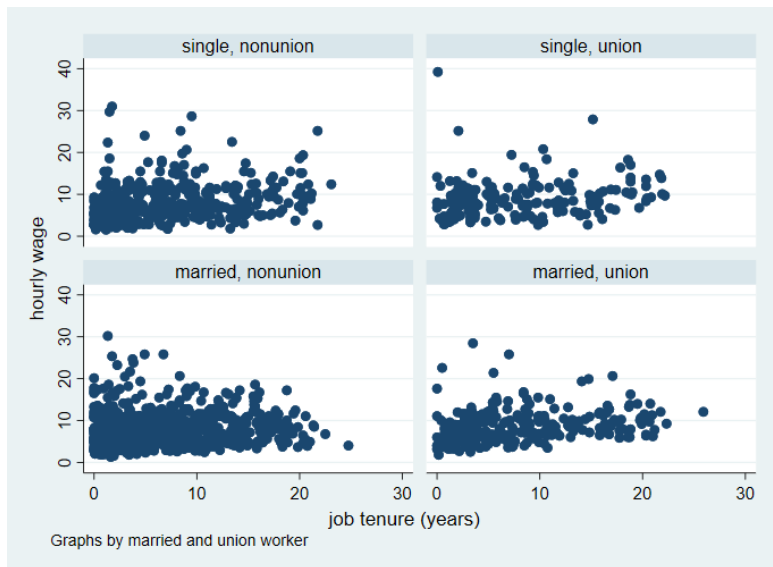
Here I have also made the line thicker than the default. An alternative curve-fitting technique is local polynomial smoothing:

lpoly wage tenure, ci lineopts(lcolor(yellow) lwidth(thick)) title(Local polynomial smoothing) note(nls88 data)



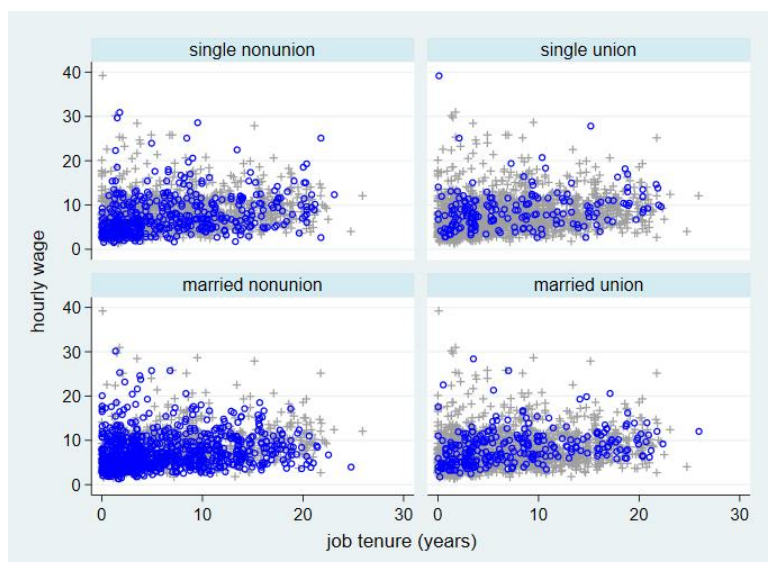
I have made the line even thicker & changed the colour to yellow. The confidence interval is so tight for most of the range of tenure you can't see it here except in the right tail where it spreads out as there is so little data. Which technique you use is partly a matter of taste and what works best with your data.

In the first graph on the previous page I used **"by(married)"** to generate separate graphs for two subsets according to that variable. If I used **twoway scatter wage tenure, by(married union)** I would have four separate graphs for each of the combinations e.g. single & non-union, single & union.. etc:.



A download **fabplot**, created by Nick Cox, shows graphs in which each panel contains all the points but the particular subsets are highlighted. For example:

fabplot scatter wage tenure , by(married union)

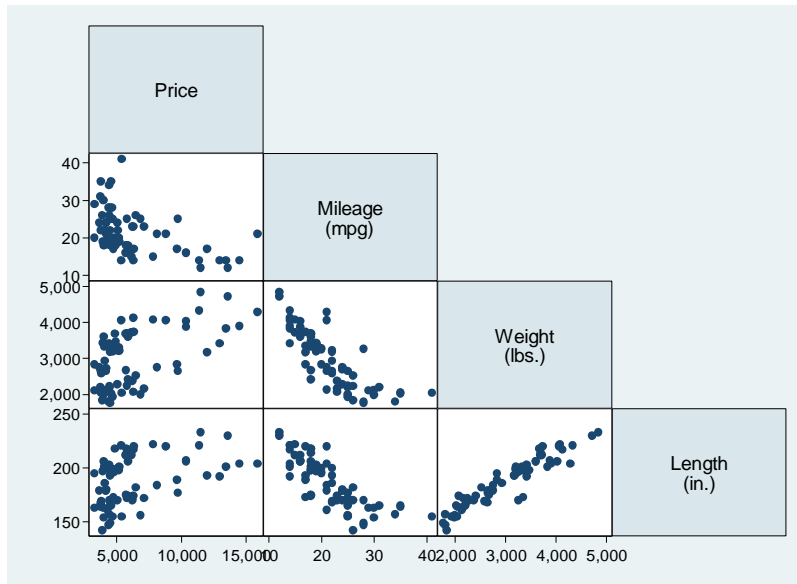


So in the top left graph, the blue dots show you where the single non-union members are in the whole distribution. This may facilitate uncovering where in the joint distribution a particular subset is concentrated.

To generate a matrix of scatterplots for several variables:

webuse auto, clear

graph matrix price mpg weight length, half



Omitting the “**half**” option means that the upper triangle (symmetric to the lower one) is also shown.

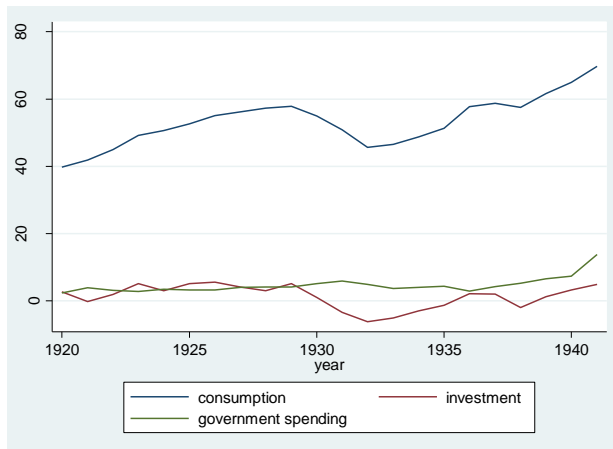
2.3 Time plots

If your data is time series it is best to use the dedicated line plot for time series command.

webuse klein, clear

tsset yr this tells Stata that this variable is the time variable

tsline consump invest govt will graph the three variables over time:



Use **twoway (tsline consump, recast(scatter))** If you do not wish the points to be connected.

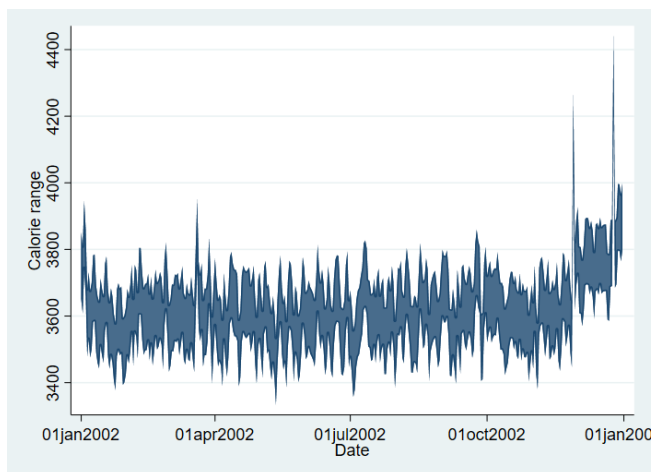
To plot the correlogram i.e. autocorrelations between y_t and y_{t-1} , y_t and y_{t-2} etc:

ac consump, lags(8)

Sometimes you have two variables and you want to illustrate the range between them over time. For example they could be the upper and lower bounds for a given outcome, like a daily price high and low.

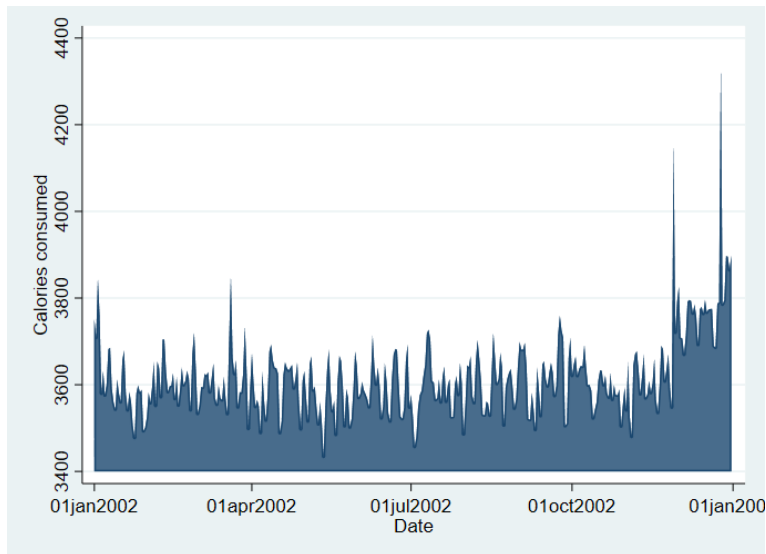
sysuse tsline2

twoway rarea ucalories lcalories day



For a slightly different look replace **rarea** with **rcap** or **rbar**.

twoway area calories day shades the area under the curve

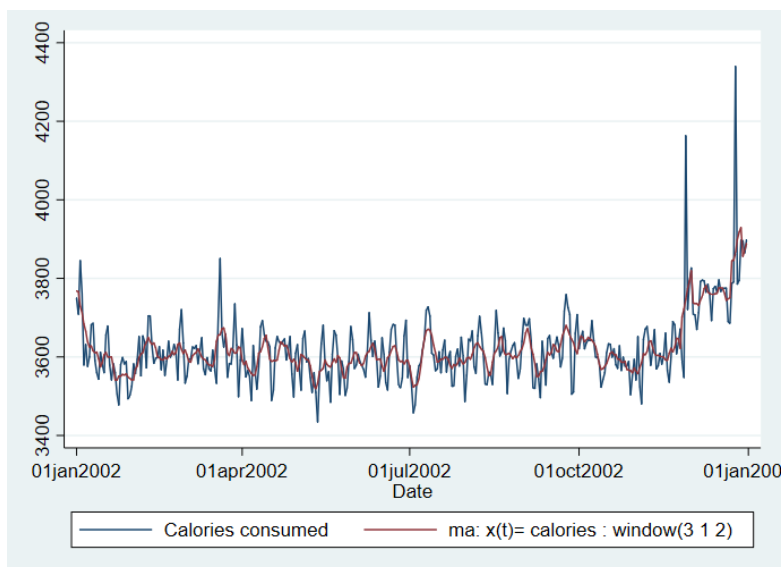


Some time series fluctuate have high variance and it may be preferable to create a smoother version of the data that removes much of the short run volatility. The simplest way of doing this is a moving average. To graph a moving average in Stata, create it using **tssmooth ma** first.

tssmooth ma caloriesm = calories , window(3 1 2)

This generates a 6 year moving average with three lags, the current value and two leads. You can apply weights to the different values if you wish. To show the original and the moving average series:

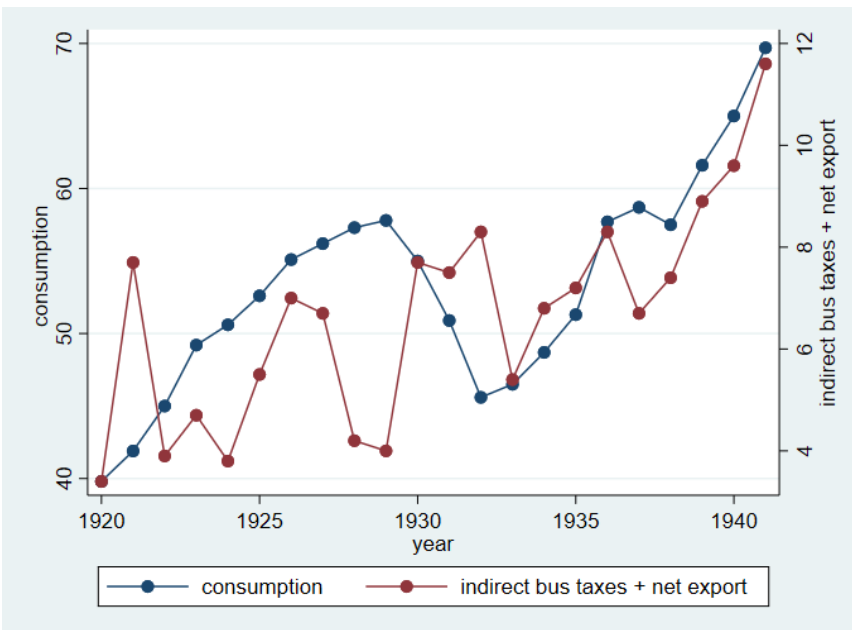
tsline cal*



The widely used Hodrick-Prescott filter for macroeconomic data is available with the **tsfilter hp** command.

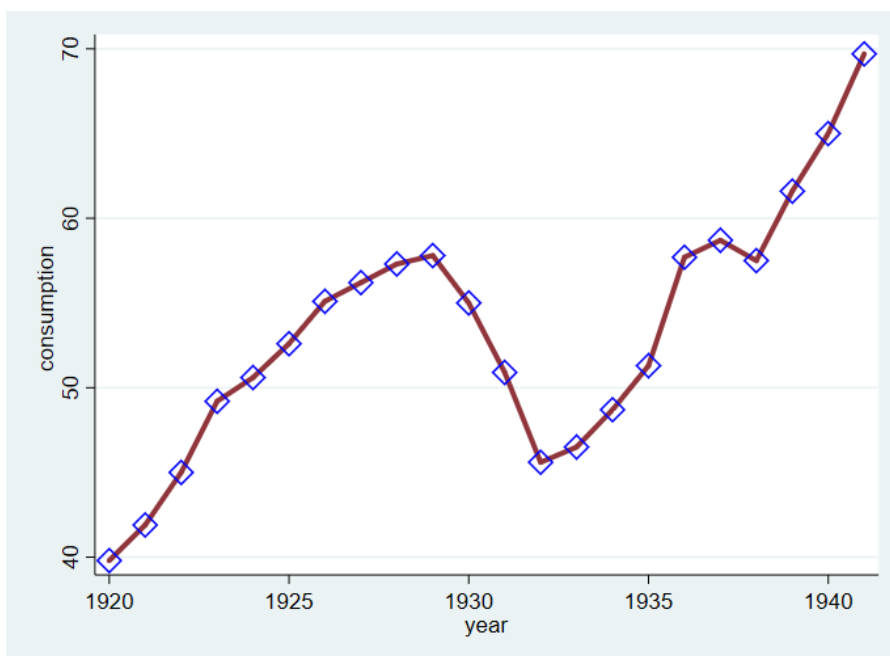
Sometimes you might wish to plot two or more variables which have different dimensions for example GNP and the unemployment rate. In that case you can use a separate y axis for each of the two. Using the **klein** dataset:

twoway (scatter consump yr, c(l) yax1s(1)) (scatter taxnetx yr, c(l) yax1s(2))



To get a taste for some of the many options you can use in crafting an image, consider this:

twoway (scatter consump yr, c(l) msymbol(Dh) mcolor(blue) msize(large) clwidth(thick) clcolor(maroon))



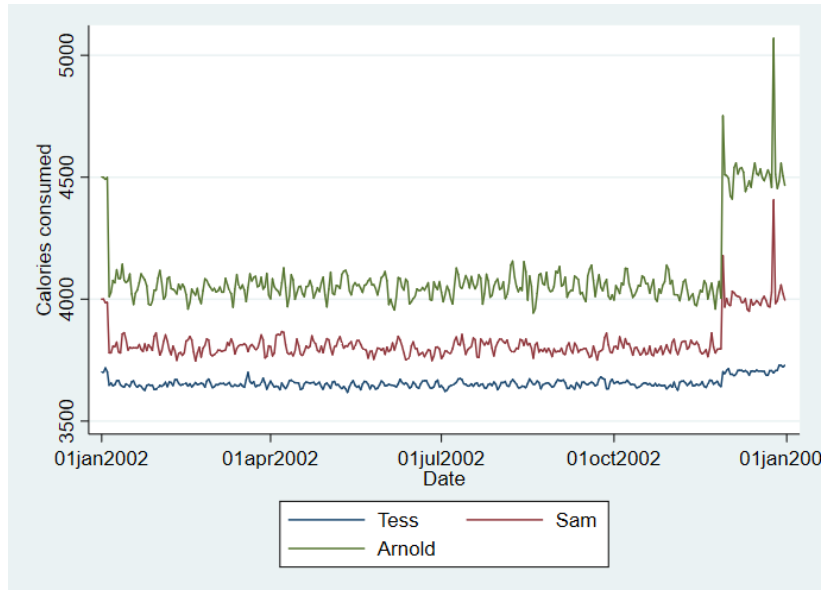
2.4 Panel data

If you are using panel/longitudinal data it is possible to create line plots of a series for the individuals:

sysuse xtline1

xtset person day

xtline calories, overlay



Dropping the **overlay** option generates a separate graph for each individual. If “N” is large i.e. there are many individuals, then this form of graph may not be very useful.

3. Graphs after estimation

After you have estimated your models there are several reasons to use graphs. One is that, post-estimation, it may be wise to examine various characteristics of the residuals. A second is that sometimes a plot of regression coefficients or marginal effects is an easier way of showing the results.

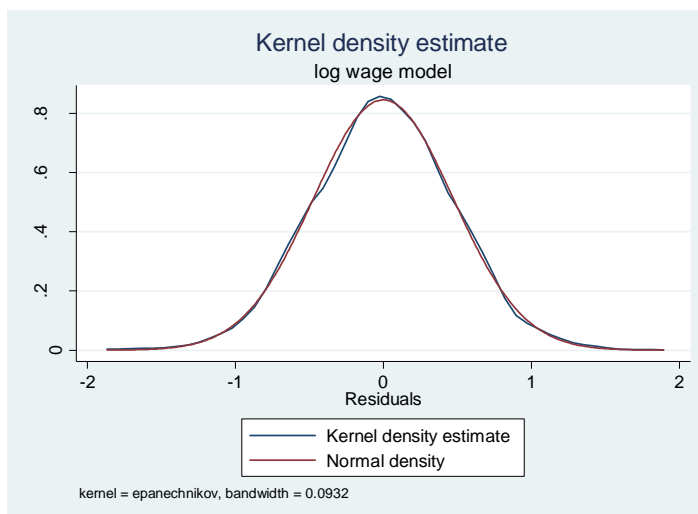
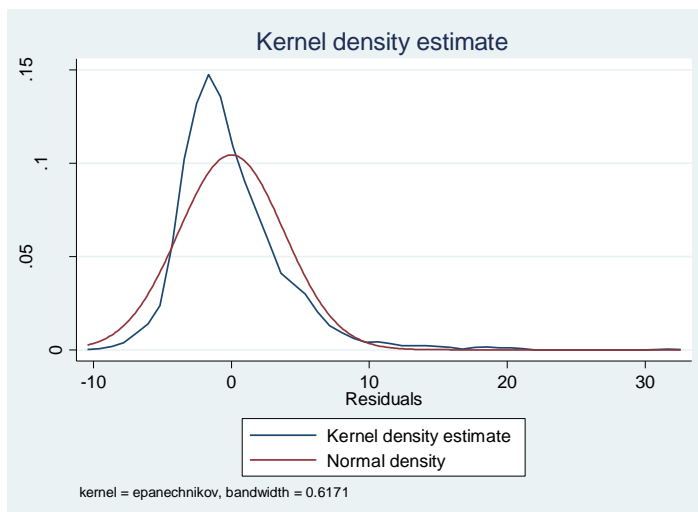
```
sysuse nlsw88, clear
```

```
reg wage age married collgrad union i.race
```

```
predict reshat , residuals
```

```
kdensity reshat , norm
```

This plots the residuals from the model and superimposes a normal distribution. You can see the residuals don't look normal. It is well known that the distribution of earnings is close to being log-normal i.e. the log is normally distributed. If you change the dependent variable to the log of wages & re-estimate the model you will find the residuals are remarkably close to a normal distribution (2nd graph below)



If your data is time-series then you should be interested in whether the residuals are autocorrelated. A graphical way of doing this is examine the correlogram which plots the degree of autocorrelation for different lags i.e. the autocorrelation between the residuals in period t and $t-1$, t and $t-2$ etc.

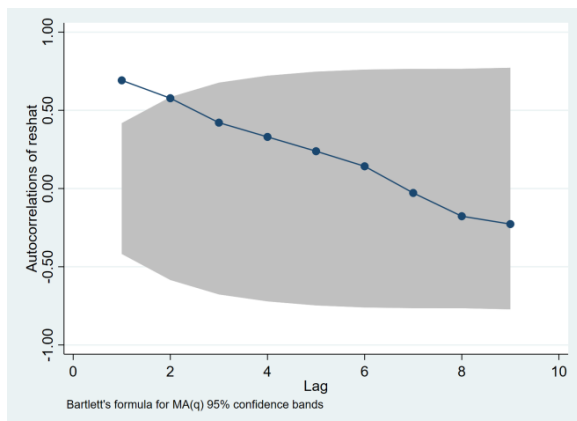
webuse klein, clear

tsset yr

reg consump totinc

predict reshat , residuals

ac reshat, lags(9) recast(connect)

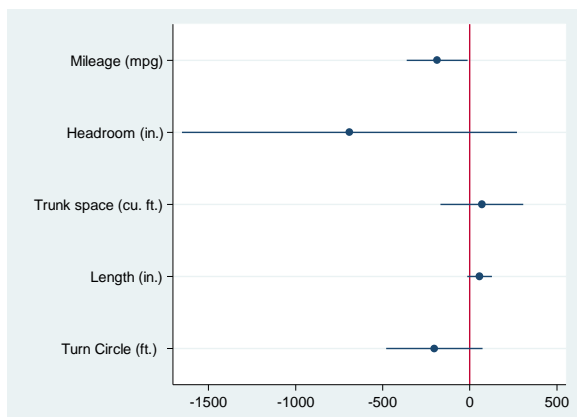


So the autocorrelation between residuals in periods t and $t-1$ is about .65. The lags fade away as we would expect so there is little correlation between residuals in period t and $t-5$, say. The **corrgram** command displays a table of the autocorrelations and has a crude graph of them. To plot regression coefficients there is a user written command **coefplot** written by Jann (2013). By default it shows 95% confidence intervals but you can change that.

sysuse auto

regress price mpg headroom trunk length turn

coefplot, drop(_cons) xline(0)



In the example above the coefficient on each variable is the marginal effect of that variable. That is because the price variable is assumed to be a linear function of the x variables. If any of the x variables enter non-linearly then the marginal effect of the variable will be different at different values. Say for example if price depends on mpg (miles per gallon) and its square:

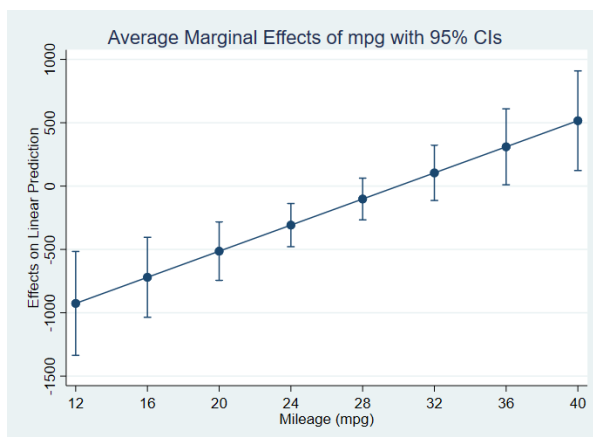
$$price = \beta_0 + \beta_1 MPG + \beta_2 MPG^2 + \dots$$

$$\Rightarrow \frac{\partial price}{\partial MPG} = \beta_1 + 2\beta_2 MPG$$

Stata's **margins** command can be used to evaluate the marginal effect at different values of mpg. First run the model, say: **regress price c.mpg##c.mpg headroom trunk length turn**

margins, dydx(mpg) will give you the average marginal effect with a standard error but doesn't tell you how it varies with mpg. Note that if you had created the square of mpg as a separate variable (say "mpgsq") & then included in the model just like another variable then **margins** would not provide the correct marginal effect of mpg as Stata, in using **margins**, would not know what mpgsq is. That's why it is necessary to use the **c.mpg##c.mpg** formulation. Note that **c.mpg##c.mpg** will generate a quadratic function but without the linear component i.e. the square only.

margins, dydx(mpg) at (mpg=(12(4)41)) evaluates the marginal effect at different values of mpg starting at 12 and increasing by increments of 4 to 41. The results are presented in a table but it is better at this stage to use the **marginsplot** command to get a nice graph of the marginal effect of mpg as it varies with mpg. Since the model was quadratic in mpg, the marginal effect is linear (see the 2nd equation above).



With **coefplot** and **margins** there are numerous options to customize the output. I have just shown the basics. **margins** can also be used with interactions between variables. For example

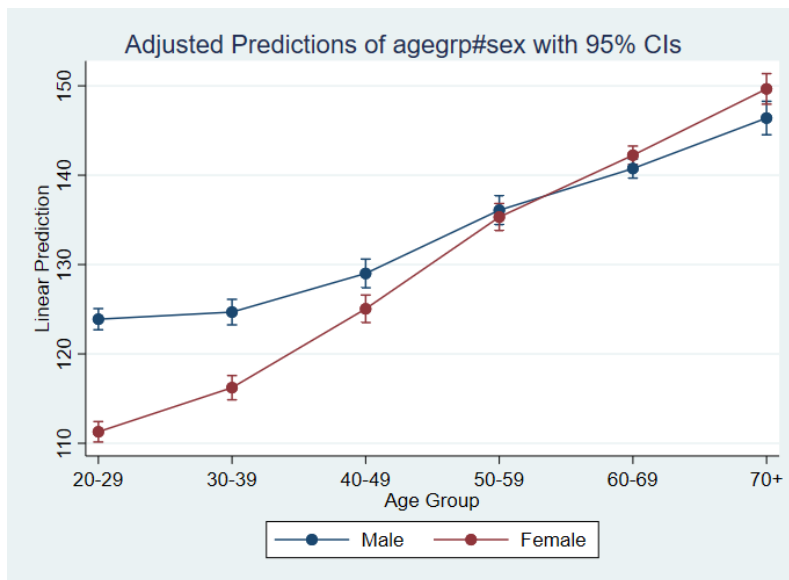
```
sysuse nhanes2
```

```
reg bpsystol agegrp##sex
```

```
margins agegrp#sex
```

<output omitted>

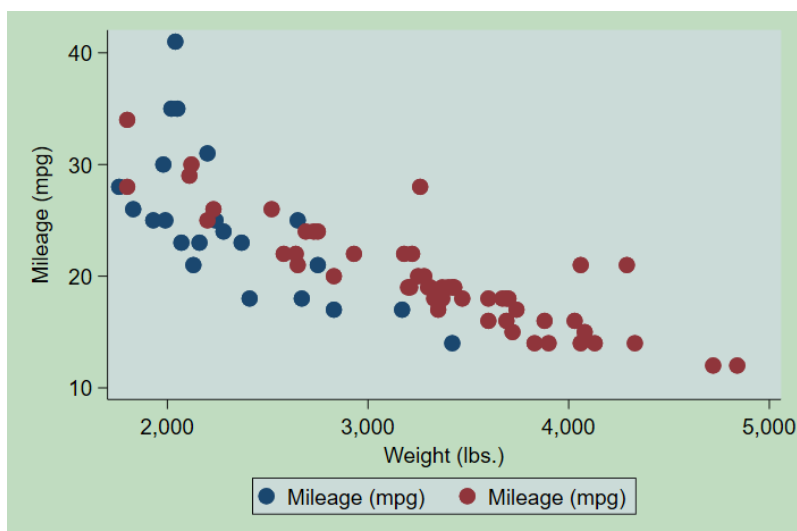
This regresses a continuous variable on a categorical age variable interacted with a dummy variable and then calculates the marginal effects of the interactions. Typing **marginsplot** produces a graph with the predicted value of the outcome for each category by sex with 95% confidence intervals. Older people have higher (systolic) blood pressure and the age gradient is steeper for females.



4. Schemes

While you can tweak the look of graphs in many ways, one approach is to use different styles of graphs (called schemes) that Stata has created. Taking the scatterplot we had on page 12, you could try:

scatter mpg weight if foreign || scatter mpg weight if !foreign , scheme(vg_teal)



Other schemes include **s2color**, **s1mono**, **s2mono**. Using **scheme(economist)** replicates the look of the graphs in **The Economist** magazine. To see the different schemes available, when you open the dialogue box for graphs, open the tab for “overall”. The “scheme” option is at the top left. Alternatively, entering **graph query, schemes** in the Stata command line will provide a full list.

5. The definitive pie chart:



ooOoo

6. References & resources

Hintze, Jerry & Ray Nelson (1998). **Violin Plots: A Box Plot-Density Trace Synergism.** *The American Statistician* 52(2):181-84.

Jann, Ben (2013). **coefplot: Stata module to plot regression coefficients and other results.**

<http://ideas.repec.org/c/boc/bocode/s457686.html>

Jones, Andrew (2017) **Data visualization and health econometrics** <http://eprints.whiterose.ac.uk/120147/>

Koffman, Dawn (2015) **Introduction to Stata 14 graphics**

https://opr.princeton.edu/workshops/Downloads/2015Sep_Stata14GraphicsKoffman.pdf

Mitchell, Michael N (2012) **A visual guide to Stata graphics** 3rd ed. Stata Press

Schwabish, Jonathan (2014) **An economist's guide to visualizing data.** *Journal of Economic Perspectives*, 28(1) 209-234. <https://www.aeaweb.org/articles/pdf/doi/10.1257/jep.28.1.209>

Shapiro, Jesse (nd) **How to give an applied micro talk**

https://www.brown.edu/Research/Shapiro/pdfs/applied_micro_slides.pdf

Tufte, Edward (2001) **The visual display of quantitative information** 2nd ed. Graphics Press.

Tukey, John (1977) **Exploratory Data Analysis**, Addison-Wesley.

Van Kerm, Philippe (2012). **Kernel-smoothed cumulative distribution function estimation with akdensity.** *Stata Journal* 12: 543-548.

Introduction to Stata Graphics: <https://www.ssc.wisc.edu/sscc/pubs/4-24.htm>

Introduction to graphs in Stata <https://stats.idre.ucla.edu/stata/modules/graph8/intro/introduction-to-graphs-in-stata/>

Stata graphics tutorial <http://data.princeton.edu/stata/Graphics.html>

UCD CENTRE FOR ECONOMIC RESEARCH – RECENT WORKING PAPERS

- [WP17/17](#) Andrew E Clark, Orla Doyle, and Elena Stancanelli: 'The Impact of Terrorism on Well-being: Evidence from the Boston Marathon Bombing' September 2017
- [WP17/18](#) Kate Hynes, Jie Ma and Cheng Yuan: 'Transport Infrastructure Investments and Competition for FDI' September 2017
- [WP17/19](#) Kate Hynes, Eric Evans Osei Opoku and Isabel KM Yan: 'Reaching Up and Reaching Out: The Impact of Competition on Firms' Productivity and Export Decisions' September 2017
- [WP17/20](#) Tamanna Adhikari, Michael Breen and Robert Gillanders: 'Are New States More Corrupt? Expert Opinions vs. Firms' Experiences' October 2017
- [WP17/21](#) Michael Spagat, Neil Johnson and Stijn van Weezel: 'David Versus Goliath: Fundamental Patterns and Predictions in Modern Wars and Terrorist Campaigns' October 2017
- [WP17/22](#) David Madden: 'Mind the Gap: Revisiting the Concentration Index for Overweight' October 2017
- [WP17/23](#) Judith M Delaney and Paul Devereux: 'More Education, Less Volatility? The Effect of Education on Earnings Volatility over the Life Cycle' October 2017
- [WP17/24](#) Clemens C Struck: 'On the Interaction of Growth, Trade and International Macroeconomics' November 2017
- [WP17/25](#) Stijn van Weezel: 'The Effect of Civil War Violence on Aid Allocations in Uganda' November 2017
- [WP17/26](#) Lisa Ryan, Karen Turner and Nina Campbell: 'Energy Efficiency and Economy-wide Rebound: Realising a Net Gain to Society?' November 2017
- [WP17/27](#) Oana Peia: 'Banking Crises and Investments in Innovation' December 2017
- [WP17/28](#) Stijn van Weezel: 'Mostly Harmless? A Subnational Analysis of the Aid-Conflict Nexus' December 2017
- [WP17/29](#) Clemens C Struck: 'Labor Market Frictions, Investment and Capital Flows' December 2017
- [WP18/01](#) Catalina Martínez and Sarah Parlane: 'On the Firms' Decision to Hire Academic Scientists' January 2018
- [WP18/02](#) David Madden: 'Changes in BMI in a Cohort of Irish Children: Some Decompositions and Counterfactuals' January 2018
- [WP18/03](#) Guido Alfani and Cormac Ó Gráda: 'Famine and Disease in Economic History: A Summary Introduction' February 2018
- [WP18/04](#) Cormac Ó Gráda: 'Notes on Guilds on the Eve of the French Revolution' February 2018
- [WP18/05](#) Martina Lawless and Zuzanna Studnicka: 'Old Firms and New Products: Does Experience Increase Survival?' February 2018
- [WP18/06](#) John Cullinan, Kevin Denny and Darragh Flannery: 'A Distributional Analysis of Upper Secondary School Performance' April 2018
- [WP18/07](#) Ronald B Davies and Rodolphe Desbordes: 'Export Processing Zones and the Composition of Greenfield FDI' April 2018
- [WP18/08](#) Costanza Biavaschi, Michał Burzynski, Benjamin Elsner, Joël Machado: 'Taking the Skill Bias out of Global Migration' May 2018
- [WP18/09](#) Florian Buhlmann, Benjamin Elsner and Andreas Peichl: 'Tax Refunds and Income Manipulation - Evidence from the EITC' June 2018
- [WP18/10](#) Morgan Kelly and Cormac Ó Gráda: 'Gravity and Migration before Railways: Evidence from Parisian Prostitutes and Revolutionaries' June 2018