

Introduction to Stata for Statistical Analysis: Part 1

STATA BASICS

Four Default Windows in Stata

Command – interactive commands can be typed in and submitted by hitting enter
Results – displays submitted commands and all output
Variables – displays all variables
Review – records of all Stata commands submitted in the current Stata session

Getting Help in Stata

Use Help tab (last tab at the top) – click Contents or Search (can search by keyword)
In Command Window – type **help summarize** and hit Enter to get help on Stata command summarize
In Command Window – type **search log** to get a full list of help file, FAQ's, examples, etc. related to **log**.

Using Directories to Organize Your Work

Stata displays your current working directory in the lower left hand corner of the program window. Here's what you may see depending on the computer you are using:

C:\data

You'd probably want to make a directory to do your work. Working in that directory will allow you to be able to refer to files by just their file name instead of the entire path. To make a directory called **IntroStata** and work from that directory when running Stata:

```
mkdir c:\IntroStata  
cd c:\IntroStata /* change directory to IntroStata */
```

Now to use the Stata data file:

```
use demo /* instead of by the entire path such as use c:\templdemo */
```

Using the Do-file Editor

Stata can be used interactively--that is you submit each line of command and get output without saving any of the commands. This is very useful when you just want to take a quick look at a dataset. But when you need to do substantial data management or analysis, you will want to create a program (called a "do" file in Stata) which contains all of the necessary commands or save all your commands in a separate text file. This allows you to document your methods and replicate your analyses.

In order to do this you will keep a text editor window open while you work. Though you can use any text editor such as Notepad or Wordpad, Stata contains a built in text editor called the "do-file editor." It is accessed from within Stata by clicking on "new do file editor" icon (8th from left on main Stata window). Once you have your commands written you can submit the set of commands in the do file by clicking on "do current file" icon (10th from left in do file editor window). Alternatively you can save the file, say as "file1.do" by clicking on the "save" icon (3rd from left in do file editor window) and then type the following in the command window: **do file1**. Or, what I often do is to keep the text editor open with the set of commands saved, and copy the particular line of command that needs to be run in Command window interactively. Note that even when working from a do file, you can also work interactively with Stata.

In the results window you will see your commands and the output they have produced. If you want to change a command or add to the program simply type into the do-file editor and again submit the entire set of program or submit just a line of command. At the end of your session when you close the do-file editor, make sure to save it, but treat it like any document and keep saving the file intermittently as you add more commands.

Keeping a Log File

Keeping a log file is a way to keep a copy of your output. The log file is a simple text file. It will capture everything that appears in the Stata Results window. The use of a log file can be built into your program by placing near the beginning of your program the command **capture log using filename, text replace** and **capture log close** at the end of your program.

The optional "capture" command tells Stata to suppress any error messages that might arise from the command. The capture command is mainly important when you want your do-file to continue to run even if the command appears to be redundant. So if the log file is already open when we ask Stata to open it we want our program to keep going instead of giving us the "log file already open" error message. The option "replace" is used here because we don't want Stata to give an error message every time we write over the log file. In this case we just want to go ahead and write over the log file each time we run the program, but be careful as you may not want to do this. The option "text" is used if you want this file to be in ascii text form. I prefer this because then I can open it using a text editor such as Notepad.

To temporarily turn on and off the logging, you can type **log on** and **log off**.

Helpful Tips Specific to Stata

Equal Signs: Stata distinguishes between equal signs which indicate equality and those which assign a value. To express equality in Stata double equal signs must be used **==**. To assign value, a single **=** is used.

For example: **replace patgrp10=1 if patgrp10==10**
Not equal to is represented by **~=** or **!=**. Greater than or equal to is **>=**, and less than or equal to is **<=** as expected.

Missing Values: In Stata, missing numeric data is represented with a period ".". In some operations (notably sort and if, although not in statistical calculations such as means or correlations), a missing value in Stata is considered to have the largest possible positive value for that variable. For example if you have missing values in your age variable and you want to create an indicator variable for people over age 65, you have to be careful or all of your missing values will be included as over age 65.

Instead of: **replace over65=1 if age>=65**
You'll want to say: **replace over65=1 if age>=65 & age~.**

Top 50 Commands

Most of what you will need to do to organize files and clean and manage your data can be accomplished with these 50 commands. Additionally Stata includes almost 500 specific statistical commands which may be applied to your specific analysis plan.

Category	Stata Commands
Getting on-line help	search, help
Operating system interface	pwd, cd, sysdir, mkdir, dir, erase, copy, type
Using and saving data from disk	use, save, append, merge, compress
*Inputting data into Stata	input, edit, infile, infix, insheet
The Internet and Updating Stata	update, net, ado, news
Basic data reporting	describe, codebook, list, browse, count, inspect, summarize, table, tabulate
Data manipulation	generate, replace, egen, rename, drop, keep, sort, encode, decode, order, by, reshape
Formatting	format, label
Keeping track of your work	log, notes
Convenience	display

Portions from Stata Netcourse 101 Copyright StataCorp. and "Statistics with Stata 5.0" by Lawrence Hamilton.

STATA FOR STATISTICAL ANALYSIS

How to Load the Data and Look at the Data

For Stata Formatted Data:

```
. use demo /* Loads dataset demo. Stata dataset has an extension of dta, and so the file name is demo.dta. */
```

From Excel Sheet to Stata:

```
. insheet using demo.txt /* demo.txt is a dataset saved in Excel as a tab delimited ascii file */
```

Using Stat Transfer (commercial software: <http://www.stattransfer.com>)

Can Hand Enter the Data

```
. Input id height weight
```

List and Summarize the Data:

```
. list in 1/5 /* lists first 5 observations */  
. sum /* summarize the data; gives variable name, non-missing N, mean, SD, min, max */  
. list id mstatus age /* lists variables of id, mstatus and age for first 10 observations */  
. list wt ht if age<50 /* lists weight and height of those whose are younger than 50 */
```

How to Display and Describe Single Variable (Univariate Analysis)

Nominal Scale Data

Ex: Death status, Blood group, Marital status

Name only, no order, magnitude not important, categories need to be a disjoint and exhaustive list

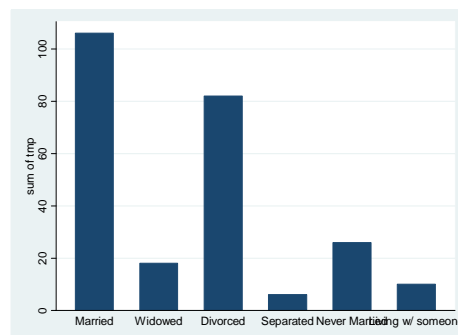
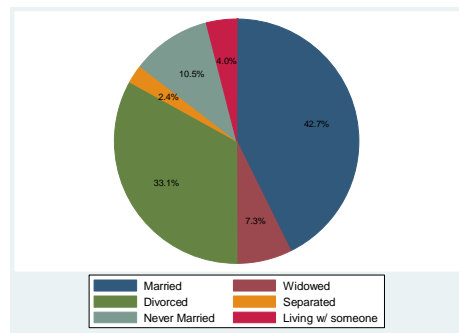
```
. tab mstatus /* tabulation by marital status */
```

marital status	Freq.	Percent	Cum.
Married	106	42.74	42.74
Widowed	18	7.26	50.00
Divorced	82	33.06	83.06
Separated	6	2.42	85.48
Never Married	26	10.48	95.97
Living w/ someone	10	4.03	100.00
Total	248	100.0	

```
. graph pie, label(_all name) over(mstatus) legend(off)
```

```
. graph pie, label(_all percent, format(%3.1f)) over(mstatus)
```

/* left figure */



```
. gen tmp = 1 /* creates a temporary variable giving values of 1 to every observation */
```

```
. graph bar (sum) tmp, over(mstatus) ytitle(Frequency)
```

/* right figure */

Note: Better to show relative percent rather than N.

Ordinal Scale Data

Ex: Diagnostic Test Result (highly unlikely (1); unlikely (2); ambivalent (3); likely (4); highly likely (5))

. tab d5confus

confusion score	Freq.	Percent	Cum.
0	118	57.56	57.56
1	40	19.51	77.07
2	15	7.32	84.39
3	21	10.24	94.63
4	11	5.37	100.00
Total	205	100.00	

. gen confus = (d5confus>1)
. replace confus = . if d5confus==.

Relative frequency of confusion (confusion score 2 or above)

confus	Freq.	Percent	Cum.
0	158	77.07	77.07
1	47	22.93	100.00
Total	205	100.00	

Continuous (Interval Scale) Data

. sum wt /* summarize the data */

Variable	Obs	Mean	Std. Dev.	Min	Max
wt	285	201.407	38.86328	95	348

STEMPLOTS (Stem-and-leaf plots)

. stem wt

```

9* | 5
10* |
11* | 8
12* | 0
13* | 0011124888
14* | 055555557
15* | 000002345556789
16* | 000222223557777888
17* | 000000011223455555678
18* | 0000000000000000000122334555556667788
19* | 0000000024445555555556777899
20* | 00000000012222355555555677889
21* | 00000001223445555566677888
22* | 0000002555555556899
23* | 0000001233555578
24* | 0000245558
25* | 00000000222344555
26* | 00002255689
27* | 005
28* | 00136
29* | 55
30* |
31* |
32* | 0
33* |
34* | 8
  
```

Note: Can split stems if two few stems or round the numbers if too many digits (too many empty stems (no leaves)).

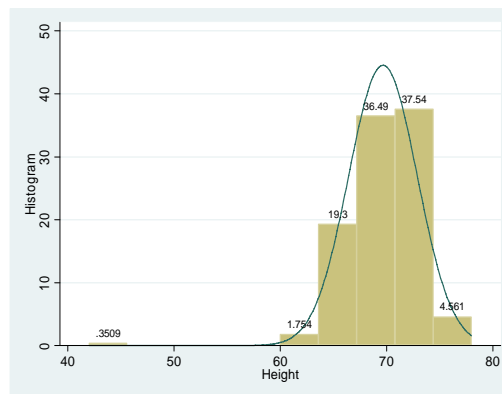
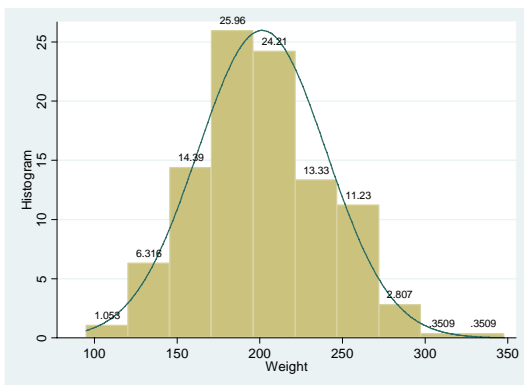
Describing data distribution

Unimodal or bimodal? (*mode = peak*)
 Symmetric or skewed?
 Gaps? Outliers?

HISTOGRAMS

- Breaks the range of values of a variable into intervals
- Displays count or % of the data that fall into each interval
- Should have adjoining bars and equal width intervals

`. histogram wt, bin(10) percent addlabels normal ytitle(Histogram) xtitle(Weight) /* left */`
`. histogram ht, bin(10) percent addlabels normal ytitle(Histogram) xtitle(Height) /* right */`



TABLES: counts, relative frequency, cumulative frequency

`. egen wtcut = cut(wt), at(0, 100, 200, 300, 400, 500) icode`
 /* The cut value should include upper and lower end */
`. tabstat wt, by(wtcut) stat(n min max mean)`

wtcut	N	min	max	mean
0	1	95	95	95
1	142	118	199	171.3239
2	140	200	295	230.7857
3	2	320	348	334
Total	285	95	348	201.407

`. tab wtcut`

Weight Distribution

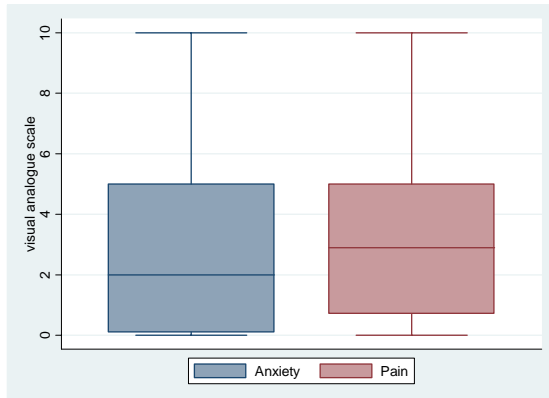
wtcat	Freq.	Percent	Cum.
0	1	0.35	0.35
1	142	49.82	50.18
2	140	49.12	99.30
3	2	0.70	100.00
Total	285	100.00	

`. drop wtcut` /* drops the variable wtcut */
`. egen wtcut = cut(wt), group(5) icode` /* cuts into 5 groups of about equal size */

BOXPLOT

- Graphical representation of min, Q_1 , median, Q_3 , max, and suspected outliers.
- Boxplot is not appropriate when the distribution is bimodal.

`. graph box d5preanx d5preint, ytitle(visual analogue scale)`

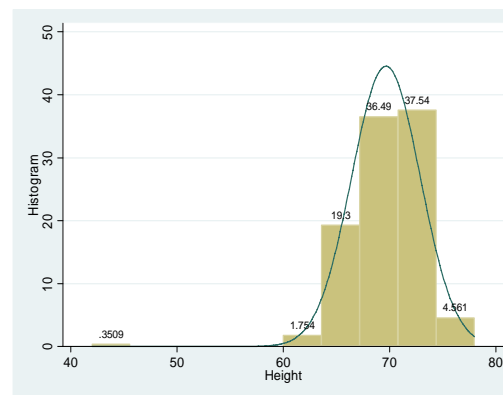
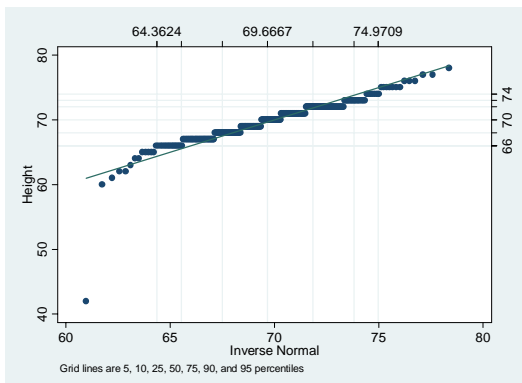


How to judge whether data are approximately normal?

- Histogram or Stemplot
- Normal Probability Plot (**pnorm**; a plot of data probability against standardized normal probability) or a Normal Quantile Plot (**qnorm**: normal quantiles against data quantiles (ordered data values)). If data are approximately normal, these plots will be close to a 45-degree straight line.

`. qnorm ht, grid`

Recall earlier histogram of ht

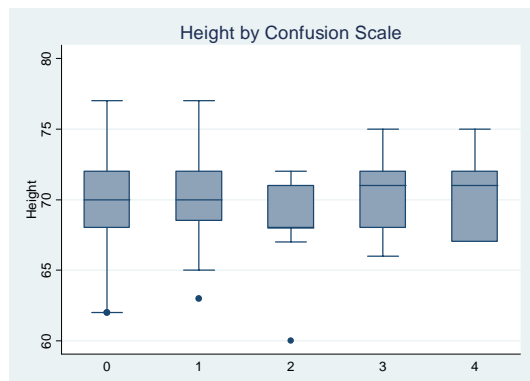


How to Display Bivariate Relationships

RELATIONSHIP BETWEEN A CONTINUOUS AND AN ORDINAL (OR CATEGORICAL) VARIABLE

Example: Is there a relationship between height (continuous) and confusion (ordinal)?

. graph box ht, by(d5confus) title(Height by Confusion Scale)



. tabstat ht, by(d5confus) stat(n mean sd)

Summary for variables: ht
by categories of: d5confus

d5confus	N	mean	sd
0	116	69.98276	2.78812
1	40	69.825	2.706947
2	15	68.6	2.848559
3	21	70.52381	2.561622
4	11	70.45455	2.769969
Total	203	69.93103	2.76025

Note: Similar box plots and table can be made for a relationship between a continuous variable and a nominal (categorical or dichotomous) variable. Ex: height and marital status.

RELATIONSHIP BETWEEN A CATEGORICAL (OR ORDINAL) AND A CATEGORICAL (OR ORDINAL) VARIABLE

Example: Is there a relationship between weight categories (ordinal) and confusion (ordinal)?

Relative % of weight categories across five confusion status

. tab wtcacat d5confus, col

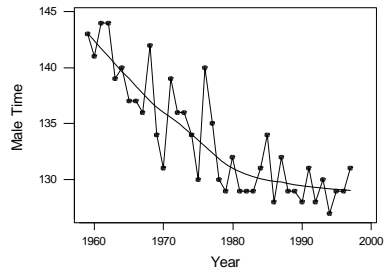
wtcat	d5confus					Total
	0	1	2	3	4	
0	24 20.69	10 25.00	1 6.67	3 14.29	1 9.09	39 19.21
1	28 24.14	6 15.00	4 26.67	4 19.05	2 18.18	44 21.67
2	20 17.24	9 22.50	4 26.67	5 23.81	2 18.18	40 19.70
3	21 18.10	6 15.00	3 20.00	5 23.81	2 18.18	37 18.23
4	23 19.83	9 22.50	3 20.00	4 19.05	4 36.36	43 21.18
Total	116 100.00	40 100.00	15 100.00	21 100.00	11 100.00	203 100.00

RELATIONSHIP BETWEEN A CONTINUOUS VARIABLE AND TIME

TIME PLOT

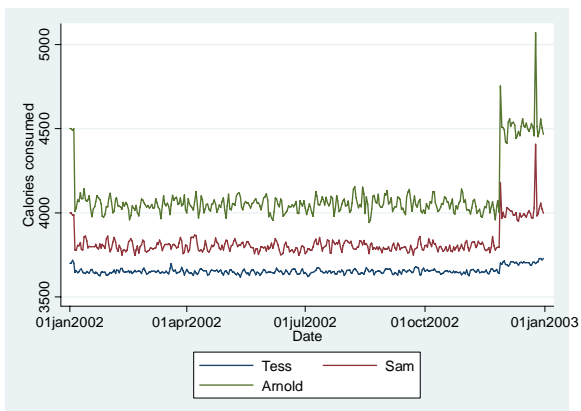
- Plot of measurements against time or order in which measurements taken.
- Can show a systematic change over time

When measurements are taken over time, stemplots or histograms can be misleading if there is a systematic change over time.



Example: Daily calorie intake data of three people collected over one year

- ```
. preserve /* preserve current Stata session to temporarily use another dataset */
. sysuse xtline1, clear /* Use xtline1.dta in the system */
. xtset person day /* Set person and day as id and time of panel data */
. xtline calories, overlay /* Time plot */
. restore /* restore the previous Stata session */
```

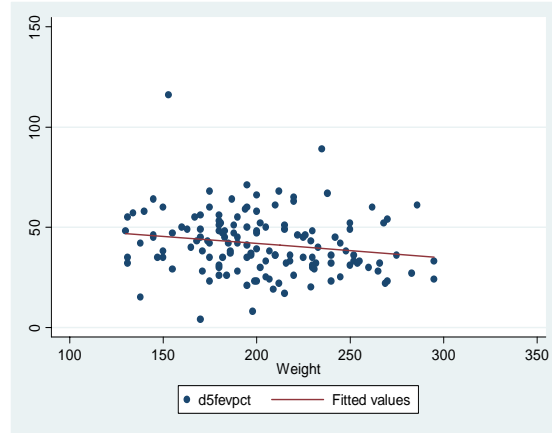
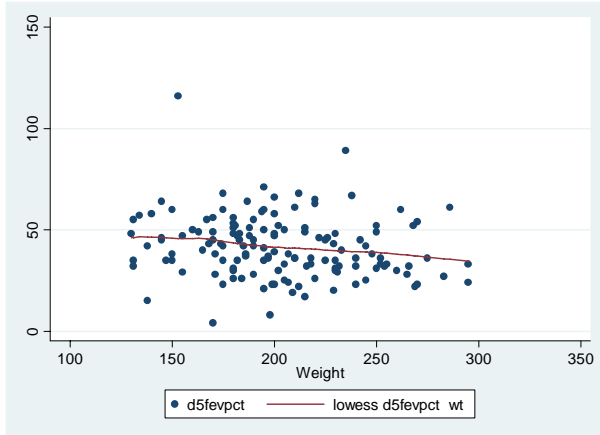


## RELATIONSHIP BETWEEN TWO CONTINUOUS VARIABLES

### SCATTER PLOTS

Example: Is there a relationship between FEV % (lung capacity) and weight?

```
. scatter d5fevpct wt /* simple scatter plot */
. twoway (scatter d5fevpct wt) (lowsess d5fevpct wt) /* left figure: overlaid with lowsess */
```



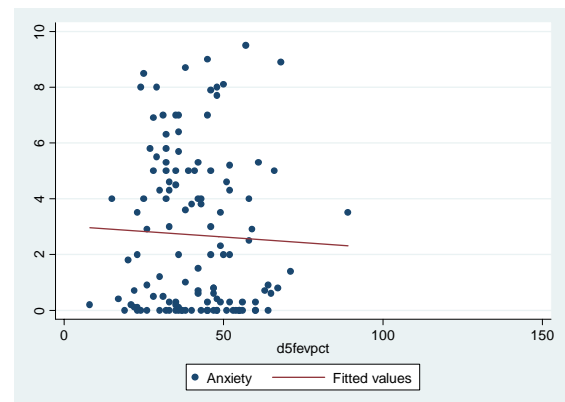
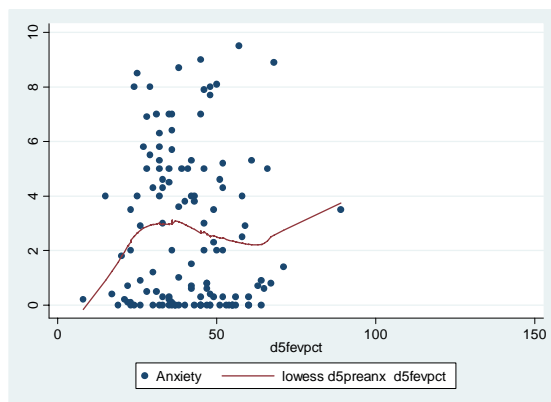
```
. twoway (scatter d5fevpct wt) (lfit d5fevpct wt) /* right figure: overlaid with predicted regression */
```

In scatterplots, look for:

- Overall pattern?
- Consistent **direction** (positive or negative)?
  - **Form** (linear, quadratic)?
  - **Strength of the relationship**?
- Deviation from the pattern? Any outlier?
- Equal variance?
- Clusters?

Example: Is there a relationship between Anxiety and FEV %?

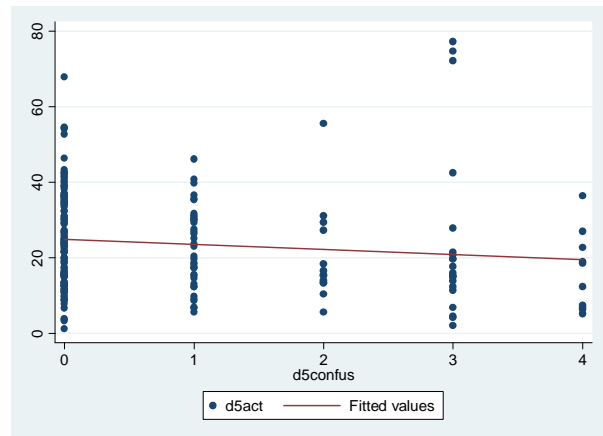
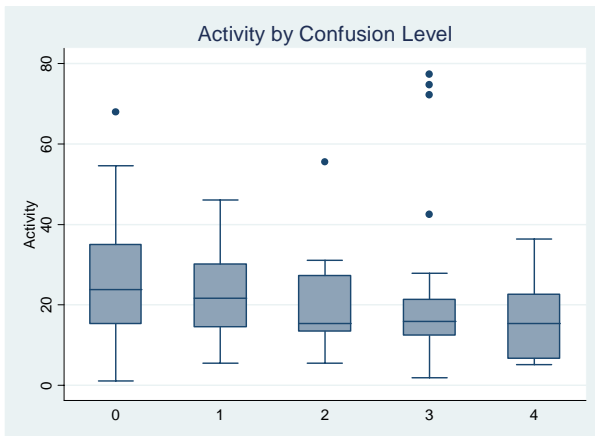
```
. twoway (scatter d5preanx d5fevpct) (lowsess d5preanx d5fevpct) /* left figure */
. twoway (scatter d5preanx d5fevpct) (lfit d5preanx d5fevpct) /* right figure */
```



**RECALL: RELATIONSHIP BETWEEN A CONTINUOUS AND AN ORDINAL VARIABLE**

Example: Is there a relationship between activity level and how confused one is?

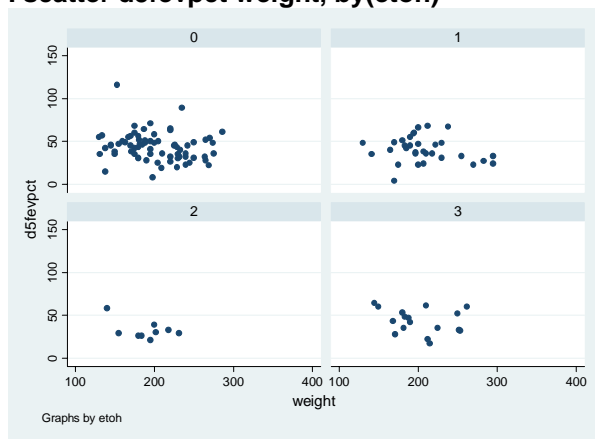
```
. graph box d5act, over(d5conf) title(Activity by Confusion Level) ytitle(Activity) /* left figure */
. twoway (scatter d5act d5conf) (lfit d5act d5conf) /* right figure */
```



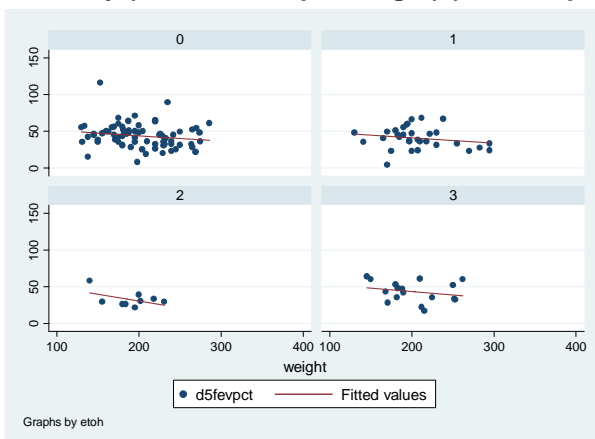
**RELATIONSHIP ACROSS THREE VARIABLES: AN ORDINAL (OR CATEGORICAL) VARIABLE IN A SCATTER PLOT**

Example: Does the relationship between FEV % and weight change depending on different alcohol abuse status?

```
. scatter d5fevpct weight, by(etoh)
```



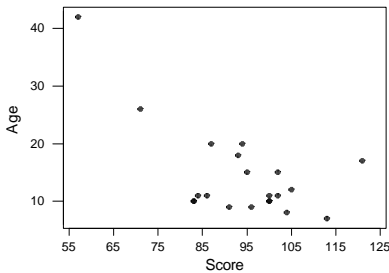
```
. twoway (scatter d5fevpct weight) (lfit d5fevpct weight), by(etoh)
```



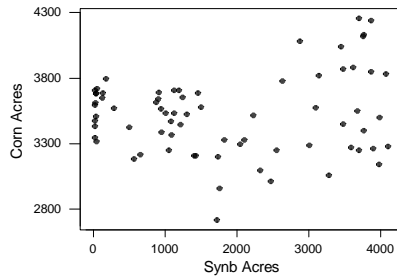
## CORRELATION COEFFICIENT ( $r$ )

- measures the **direction** and **strength** of the **linear** relationship between  $x$  &  $y$ .
- $-1 \leq r \leq 1$ .
- $r > 0$  means a positive linear relationship ( $r = 1$  means a perfect positive linear relationship.)
- $r < 0$  means a negative linear relationship ( $r = -1$  means ....)
- $r = 0$  means no *linear* relationship (does **not** mean that there is no relationship)
- has no unit
- makes no distinction between  $x$  &  $y$
- is not affected by changes in unit of  $x$  or  $y$
- and is sensitive to outlying values

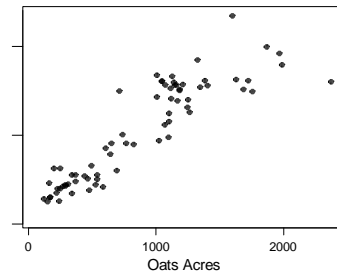
### Examples of $r$



$r = -0.640$



$r = 0.143$



$r = 0.896$

### . pwcorr pfev1pc d5fevpct d5act wt

|          | pfev1pc | d5fevpct | d5act   | wt     |
|----------|---------|----------|---------|--------|
| pfev1pc  | 1.0000  |          |         |        |
| d5fevpct | 0.3921  | 1.0000   |         |        |
| d5act    | 0.1628  | 0.0654   | 1.0000  |        |
| wt       | -0.0556 | -0.1786  | -0.0643 | 1.0000 |

## REGRESSION

**A Simple Regression Line:** A straight line describing how a response variable Y changes as an **explanatory** variable X changes.

$$y = a + bx \quad (\text{Intercept} = a; \text{Slope} = b).$$

**Residuals** = Vertical Deviation for the  $i^{\text{th}}$  observation = observed  $y_i$  – predicted  $y_i$   
 $= y_i - \hat{y}_i = y_i - (a + bx_i) = \text{deviance}_i$

- $R^2$  measures how well the regression line explains the response variable.
- In simple regression,  $R^2$  is the square of the correlation.

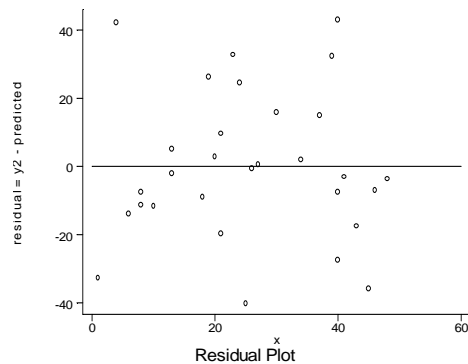
Residual represents “left-over” variation in the response after fitting a regression line.  
 Mean of the least squares residuals is always zero.

**In regression, it is important to determine which the response variable is.**

Assumptions of linear regression analysis:

- Linearity
- Independence
- Equal Variance
- Normality

**Residual Plot:** residuals against x.

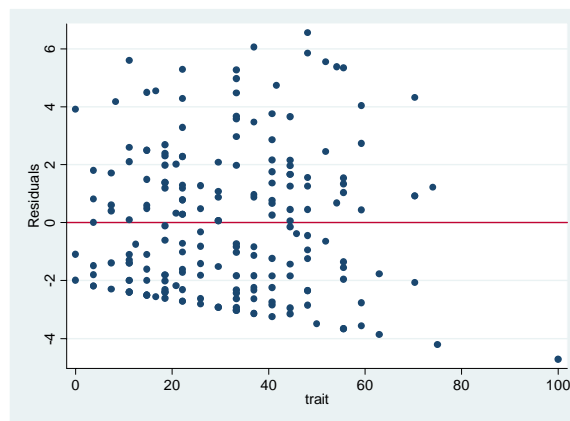
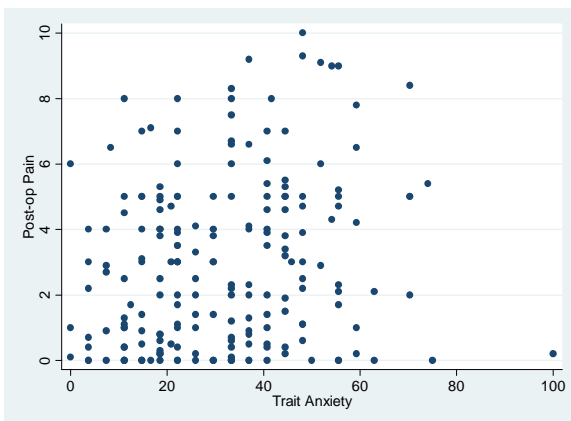


Good = No pattern  
 Bad = Unequal Variance (Heteroscedasticity), Nonlinearity

Example: Relationship between post-op pain (**d5preint**) and trait anxiety (**trait**)

Graphs to see if the relationship is linear and if other assumptions met.

**. scatter d5preint trait, ytitle(Post-op Pain) xtitle(Trait Anxiety) /\* below left figure \*/**



**. reg d5preint trait**  
**. rvpplot trait, yline(0)**

**/\* simple regression: Y=d5preint, X<sub>1</sub>= trait \*/**  
**/\* residual vs. trait; above right figure \*/**

**Other problems** to watch out for in examining the fit of a regression line:

Outliers: Large standardized residuals?  $((y_i - \hat{y}_i) / s_{res})$

Influential observations: Individual data points that substantially change the regression line. Find regression line both with and without the suspect point, and if the two lines differ by more than a little, the point is influential. They are often outliers in the x direction. **DFBETA**

**Confounder** is a variable that is correlated with both the predictor variable and the response variable. In the above example, chronic pain (chrnpain) may be a potential confounder in the relationship between post-op pain and trait anxiety. Leaving the confounder out of the regression analysis yields **biased** estimates of the effects of the predictor variable.

```
. graph box d5preint, over(chrnpain) /* figure 1: relationship between post-op pain and chronic pain? */
. graph box trait, over(chrnpain) /* figure 2: relationship between trait anxiety and chronic pain? */
/* chrnpain may confound the relationship between post-op pain and trait anxiety */
. reg d5preint trait chrnpain /* multiple regression: X2 = chrnpain */
. rvfplot /* figure 3: residual vs. fitted */
```

Figure 1

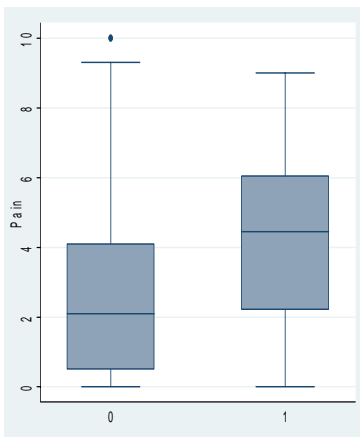


Figure 2

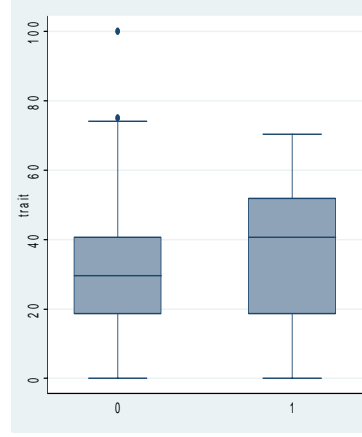
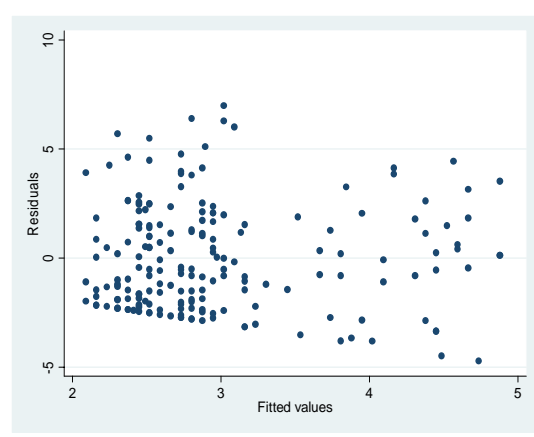


Figure 3



**Interaction (Effect Modification)**

Example: Different effects of trait anxiety on post-op pain depending on whether or not the person has chronic pain or not.

Interaction can be modeled in a regression model by including an interaction term which is a product of the two explanatory variables.

**Diagnostics**

- Do histogram, box plot, stem-and-leaf plot of predictors to find outliers or extreme skewness
- Always construct the scatter plot with the fitted regression line.
- If you have multiple predictors, make scatter plots of each of the predictors against the response variable (matrix).

## Exercises

1. List values for variables **idn**, **gender**, **school**, **income**, and **d5uintp** for patients who are older than 80 years.
2. Rename the variable called **d5uintp** as **pain\_intensity**. If you don't know how to rename, get help in how to rename variables.
3. Find out what type of variable **grp** might be. For instance, what is the range of this variable? If it is a categorical variable, then obtain proportions of sample in each category. If it is a continuous variable, then obtain mean and standard deviation.
4. Get a frequency table for **grp** by **income**.
5. Generate a variable called body mass index (BMI) using height (**ht**) and weight (**wt**) given in the dataset. BMI equals a person's weight in kilograms divided by height in meters squared. ( $BMI = \text{kg}/\text{m}^2$ ). You first need to find out what the unit of the height and weight is in the dataset given. Remember that these are adult patients. In Stata, taking a product is done using **\***, division using **/** and square of **x** is expressed as **x^2**.
6. Get summary statistics (mean and standard deviation) of BMI. Do a histogram of BMI. Does the shape of BMI histogram appear Normal? Is there any outlying value? If there is any outlying value, list the height, weight and BMI of those values and see if they make sense.
7. What proportion has BMI > 30?
8. Create age variable as a five level categorical variable. Before you do this, make a stemplot of age. Then choose cutoffs of <50, [50-60), [60-70), [70-80), >=80. Call this **agecat**. Looking at your stemplot, do you think these cutoffs are appropriate?
9. Create a side by side box plot of BMI by the five level age categories. Does the center of BMI distribution seem to depend on the age category?
10. Describe any relationship you find between weight and age using scatter plot. Do you think the association is weak or strong?