

Exercise No. 7: Managing Data, Simple Correlation & Regression

New Stata Commands		Old Commands Reviewed	
infile	graph twoway scatter	label var	tabulate
compress	graph twoway mspline	label data	
rename		describe, detail	
gsort		clear	
regress		generate	
predict		list	
preserve		sort	

{Be sure to open a log and keep it for your records}

1 Introduction

This exercise continues work on using Stata to manage and analyze data. In particular, we will be discussing various data management techniques, discussion of which will be found in Hamilton, *Statistics with Stata 10*, pp. 44-70 along with methods to characterize variables and relationships between variables.

- ◆ Download the following *three* data sets from the course web site:

ces941_exercise.dta
fatdeath.xls
fatdeath.txt

The data for this exercise come from two sources: (1) a study of the relationship between nutrition and prostate cancer that showed a relationship between fat consumption and the death rate from prostate cancer (among males, obviously!); (2) the 1994 Survey of Consumer Expenditures (CES) conducted on an ongoing basis by the Bureau of Labor Statistics.

2 Importing data from other programs

It's easy to use Stata's data editor to enter small data sets; however, we often need to import data that have been entered into the computer using other programs. We will show how to import data into Stata from two outside sources: (1) an Excel spreadsheet, and (2) a raw text file that might have been created in a word processor or in a text editor like Notepad or WordPad.

2.1. Entering data from a text file (*fatdeath.txt*) *(Hamilton, pp. 44-47)*

Open "fatdeath.txt" in a text editor (Notepad or WordPad) or in Microsoft Word. You'll see data

that look something like this:

```
"El Salvador" 38 0.9
"Philippines" 29 1.3
"Japan" 42 1.6
"Mexico" 57 4.5
"Greece" 96 4.8
"Colombia" 47 5.4
.
.
"Sweden" 132 18.4
```

Each row contains three variables: (1) “country” (with quotation marks to signify a string variable that may have spaces in it), (2) “fatgrm” -- Dietary fat (grams/day), and (3) “dthrate” -- male death rate (per 100,000 per year). Of course, I’ve just arbitrarily assigned names to these three variables; we’ll have to tell Stata what the names are. Print out a copy of the raw data file and exit the editor/word processor.

- ◆ Open Stata and try this command (be sure to include quotes around path\file name):

```
infile country fatgrm dthrate using "your path\fatdeath.txt",
clear
```

Oops! You should have gotten a bunch of error messages saying that the variable “country” could not be read as a number. Good, because it’s *not* a number! We need to specify each string variable as such in order to get around this problem. So, try this:

```
infile str30 country fatgrm dthrate using "your
path\fatdeath.txt", clear
```

Ah yes, that’s better.

- ◆ Now, label the variables as follows:

```
. label var country "Country"
. label var fatgrm "Dietary fat (grams/day)"
. label var dthrate "Death Rate (per 100,000)"
```

- ◆ Let’s label the data set, too:

```
label data "Fat Consumption & Prostate Cancer"
```

- ◆ Finally, in specifying “country” as a string variable, we gave it lots of space (30 spaces, to be exact). We don’t need all that space so let’s *compress* the data set and then describe it:

```
compress
describe,detail
```

- ◆ Note that the *compress* command shrank the maximum size of “country” to 14 spaces *and* it changed “fatgrm” from a floating point variable to an integer in order to save space.
- ◆ Now, save the Stata data set as “fatdeath_1.dta”.
- ◆ Finally, execute the “describe,detail” command to see what your data set looks like. Paste the results into your exercise. They should look like the following:

```
. describe,detail

Contains data from your path\fatdeath_1.dta
  obs:          30 (max=          1,129)      Fat Consumption & Prostate
                                         Cancer
  vars:          3 (max=          99)         25 Aug 2002 07:33
  width:        20 (max=         200)
-----
variable name   storage   display   value   variable label
                type     format    label
-----
country         str14    %14s     Country
fatgrm          int      %9.0g    Dietary fat (grams/day)
dthrate         float    %9.0g    Death Rate (per 100,000)
-----
Sorted by:
```

The “infile” command allows you to read in raw data and describe and format them in such a way that Stata can work its magic. Read the section of Hamilton referenced above to learn more about “infile” and associated commands.

2.2. Entering Data from a Microsoft Excel File


(This section assumes that you have Microsoft Excel on your computer. If you don’t have it available, then note that in your exercise and either (1) skip this subsection, or (2) if possible, open the Excel file in a spreadsheet program that you *do* have and attempt to complete this section.)

Often, you’ll find that data are available in the form of a spreadsheet from a program like Excel, Lotus 123, etc. In that case, it’s often easy to transport the data into Stata simply by cutting and pasting. For example, the file “fatdeath.xls” that you’ve downloaded is an Excel file containing exactly the same data as “fatdeath.txt” that we just used.

- ◆ Open “fatdeath.xls” in Excel (or other spreadsheet program). You’ll see a screen that looks something like this:

Country	fatgrm	dthrate
El Salvador	38	0.9
Philippines	29	1.3
Japan	42	1.6
Mexico	57	4.5
Greece	96	4.8
Colombia	47	5.4
Bulgaria	67	5.5
Yugoslavia	72	5.6
Poland	93	6.4
Panama	58	7.8
Israel	95	8.4
Romania	67	8.8
Venezuela	62	9.0
Czechoslovakia	96	9.1
Italy	86	9.4
Spain	97	10.1
Portugal	73	11.4
Finland	112	11.1
Hungary	100	13.1
United-Kingdom	143	12.4
Germany	134	12.9
Canada	142	13.4
Austria	119	13.9
France	137	14.4
Netherlands	152	14.4
Australia	129	15.1
Denmark	156	15.9
United-States	147	16.3
Norway	133	16.8
Sweden	132	18.4

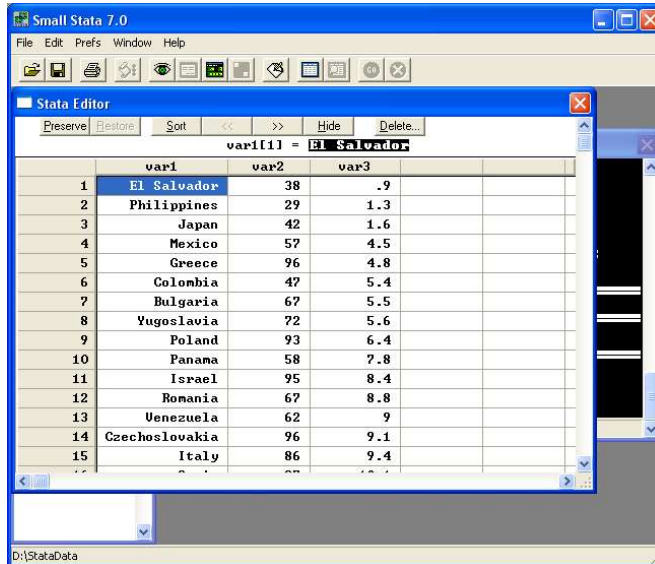
- ◆ Now, start Stata and enter the command “clear” to clear any data out of memory in Stata.

Next, open the data editor by clicking the icon  in the toolbar. Back in Excel, use the mouse to select the rectangular collection of cells containing *the data only, not the variable names in the first row*. You should see something like this:

Country	fatgrm	dthrate
El Salvador	38	0.9
Philippines	29	1.3
Japan	42	1.6
Mexico	57	4.5
Greece	96	4.8
Colombia	47	5.4
Bulgaria	67	5.5
Yugoslavia	72	5.6
Poland	93	6.4
Panama	58	7.8
Israel	95	8.4
Romania	67	8.8
Venezuela	62	9.0
Czechoslovakia	96	9.1
Italy	86	9.4
Spain	97	10.1
Portugal	73	11.4
Finland	112	11.1
Hungary	100	13.1
United-Kingdom	143	12.4
Germany	134	12.9
Canada	142	13.4
Austria	119	13.9
France	137	14.4
Netherlands	152	14.4
Australia	129	15.1
Denmark	156	15.9
United-States	147	16.3
Norway	133	16.8
Sweden	132	18.4

- ◆ Pres CNTL-C to copy the selection to the clipboard, and change to Stata’s data editor,

putting your cursor in the top left-hand cell of Stata's spreadsheet editor. Press CNTL-V to paste the data into Stata's editor. If all has gone well you should now have the data in Stata's editor, looking something like this:



- ◆ Click “Preserve” in the data editor, and close the editor by clicking the “x” in the upper right corner of the editor. Describe the data. The result should look like this:

```
. describe,detail

Contains data
  obs:          30 (max=          1,130)
  vars:          3 (max=           99)
  width:        20 (max=          200)

-----
variable name   storage   display   value   variable label
                type     format    label
-----
var1            str14   %14s
var2            int     %8.0g
var3            float  %9.0g

-----
Sorted by:
  Note:  data set has changed since last saved
```

Notice that Stata appears to have understood which variables are string variables, integers and floating point variables. Now, you need to rename the variables:

```
. rename var1 country
. rename var2 fatgrm
. rename var3 dthrate
```

- ◆ Finally, label the data set and each variable as you did when preparing the raw data set

above. Save the data set as “fatdeath_2.dta” in order to distinguish it from the first version that we created above. “Describe” the data set and paste the result into your exercise. Are the two datasets you’ve created the same?

So, now you can import data into Stata from raw text files and from spreadsheet files (The procedure that I’ve just described should work for *any* spreadsheet, not just Excel).

3 Using Explicit Subscripts with Variables and Sorting (Hamilton, p. 43)

“When Stata has data in memory, it also defines certain *system variables* that describe those data. For example, `_N` represents the total number of observations. `_n` represents the observation number: `_n = 1` for the first observation, `_n = 2` for the second, and so on to the last observation (`_n = _N`.)” (Hamilton p. 43) Often, we may wish to sort the data in different ways for different purposes, but at the same time be able to recover the original ordering when ever we wish.

- ◆ Load “fatdeath_2.dta” into Stata, if it’s not already there. The data do not appear to be sorted in any particular fashion; however, let’s assume that we always want to be able to come back to the present sorting of the data. We can create a “caseID” variable (you can call it anything you like) that shows the ordering of the data set as it currently is:

```
generate caseID = _n
```

For each case, this command creates a new case number based on the case’s position in the data set. Then, no matter how the data are later sorted, we can always regain the original sorting by sorting on “caseID”.

- ◆ “list” the data to see that the new variable does, indeed, equal the case number:

```
list
```

- ◆ Paste this listing into your exercise.
- ◆ Sort the data by “dthrate” and then list the data again:

```
sort dthrate
```

```
list
```

- ◆ Are the data still in the same order they were before the sort?
- ◆ Now, let’s sort the data in reverse order from the *original ordering*:

```
sort caseID
```

```
gsort -dthrate
```

list

The command “gsort -” allows a sort in either the ascending *or* descending direction, while “sort” allows only a sort in the ascending direction. Notice that you can sort on multiple variables, which is really handy for complicated sorts of data. You can use “gsort” any place that you can use “sort” and, because it’s a more versatile command, it’s probably good practice to use “gsort” in place of “sort” whenever you can. Here’s Stata’s help file discussion of “gsort”:

```

-----
help for gsort                                     (manual:  [R] gsort)
-----
Ascending and descending sort

       gsort [+|-]varname [[+|-]varname [...]] [, generate(newvar) mfirst]

Description

gsort arranges the observations to be in ascending or descending order of the specified
varnames and so differs from sort in that sort can produce only ascending-order arrangements;
see help sort.

Each varname can be numeric or string.

The observations are placed in ascending order of varname if + or nothing is typed in front of
the name and in descending order if - is typed.

Options

generate(newvar) creates newvar containing 1, 2, 3, ..., for each of the groups denoted by the
ordered varnames. This is useful when you wish to use the ordering with a subsequent by; see
help by.

mfirst specifies that missing values are to be placed first in descending orderings rather
than last.

```

Finally, note that the death rate from prostate cancer is highest in countries that otherwise are thought to have low death rates.

- ◆ Let’s use “gsort” to sort the data by fat consumption (in decreasing order) and create a new case ID that reflects this sort:

```
gsort -fatgrm, generate(caseFAT)
```

- ◆ Finally, let’s sort the data alphabetically by country and then save the file:

```
gsort country
```

```
save "your path\fatdeath_2.dta", replace
```

- ◆ “Describe” the data set and paste the result into your exercise.

4 Using the CES

The CES is conducted quarterly to provide information on, among other things, the “market basket” of goods and services that the typical household in America consumes. This information, in turn, contributes to the computation of the Consumer Price Index. We will be using data on 781 “consumer units” collected during 1994. A consumer unit is an autonomous consuming entity, composed of one or more individuals who share expenditure decisions. Consumer units (CU) may be single individuals living alone or with other persons with whom they do *not* share expenditure decisions, or groups of individuals, related or unrelated by blood, who share expenditure decisions. If you’re living in an apartment with a group of students and pool resources for food and other expenditures, then you’re part of a consumer unit. If you live independently in this group, you’re an independent consumer unit all by yourself. Families, of course, make up a large number of the consumer units in the U.S. BLS identifies 9 types of consumer units:

FAM_TYPE CU type is based on relationship of members to reference person. "Own" children include blood-related sons and daughters, step children and adopted children.

CODED

- 1 Husband and wife (H/W) only
- 2 H/W, own children only, oldest child < 6
- 3 H/W, own children only, oldest child > 5, <= 17
- 4 H/W, own children only, oldest child > 17
- 5 All other H/W CU's
- 6 One parent, male, own children only, at least one child age < 18
- 7 One parent, female, own children only, at least one child age < 18
- 8 Single persons
- 9 Other CU's

The “reference person” is the person who answered the survey for the CU. In families, the reference person is generally the husband or wife, but it doesn’t need to be. We’ll see shortly how many different CU types there are in our data set.

The first thing you need to do is to create yourself a *data set catalog* so you can have a written list of the variables in the data set. To do that,

- ◆ load the data into Stata and enter the command:

```
describe, detail
```

You’ll see that the data contain 46 variables (the master data set from which the current data set is extracted contains close to 500 variables) including expenditure data (usually of storage type “float”), income data, demographic data (e.g., family size, number of males/females, relationships between CU members, age of CU members, etc.), etc. It might be worthwhile to copy and paste the results of this “describe” into a Word document and then print it out so you’ll have a data set catalog for “ces941_exercise.dta.”

- ◆ Let’s see how many cases fall into each of the 9 CU-types listed above:


```
tabulate fam_type
```

- ◆ Paste the results of this command into your exercise and answer the following questions:
 - What percentage of the sample contains at least a husband and wife?
 - In which CUs might college students not in dormitories be found? (List the appropriate family type codes)
 - In which CUs might we find unrelated persons living together?
- ◆ From the data catalog, locate the appropriate variable and use Stata to determine how many of the CUs are located in urban and rural areas respectively. Enter the results of your analysis into your exercise.

5 Creating Dummy Variables (Hamilton pp. 40-42)

A “dummy variable” takes on the value 1 when a statement is true and 0 when it is false. Dummy variables are often used in economic analysis to indicate whether an observation is a member of a group (e.g., male/not male, Caucasian/not Caucasian, etc.).

- ◆ Stata makes it relatively easy to create a dummy variable, so let’s create them for the rural/urban variable you just tabulated:

```
tabulate bls_urb, generate(bls_urb)
```

If you look at your variable list, you’ll see that the generate command has created two new “bls_urb” variables: “bls_urb1” and “bls_urb2”.

- ◆ Tabulate these two variables to see what they contain:

```
tabulate bls_urb1
```

You get the following table:

bls_urb==		Freq.	Percent	Cum.
1.0000				
0		89	11.40	11.40
1		692	88.60	100.00
Total		781	100.00	

where bls_urb1 = 1 if the CU comes from an urban area and zero, if not.

- ◆ Try the same command with “bls_urb2.” How do these two variables differ?

- ◆ Try the following command to produce a two-way table:

```
tabulate bls_urb1 bls_urb2
```

- ◆ Paste the resulting table into your exercise and explain what's happening in this two-way table. Would you need to use both "bls_urb1" and "bls_urb2" to have complete information about the rural/urban location of a CU? Why/Why not?
- ◆ Save your data set *with a new name* (i.e., use "Save As")

6 Managing Memory (Hamilton, pp. 69-70)

The CES data set that we're using for this exercise is much smaller (fewer cases and many fewer variables) than the original data set, because I had to make sure that it fit into Small Stata when students were still using that crippled program. Users of Intercooled Stata have considerable control over the amount of memory they use to analyze really large data sets. Read pages 69-70 in Hamilton's *Statistics with Stata 10* to learn how to adjust memory capacity in Intercooled Stata.

7 Descriptive Methods in Regression & Correlation

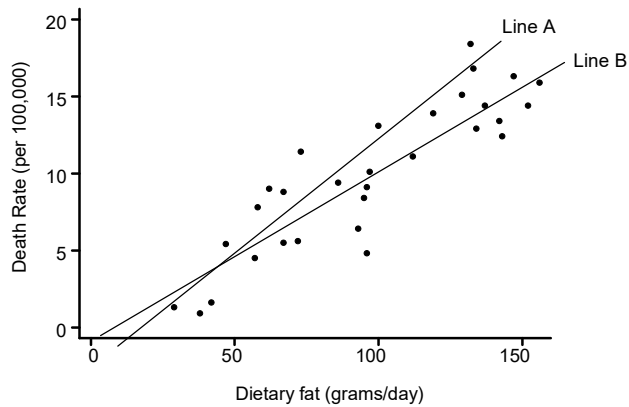
Empirical analysis in economics deals with relationships among variables. Often these relationships are quite complex; sometimes, they're rather simple. This exercise will continue our semester-long discussion of the analysis of relationships among variables. You've already been exposed to bivariate graphical analysis and we will now extend that to a more quantitative methodology.

- ◆ Load "fatdeath_1.dta" into Stata and "describe" the data set in order to make sure that every thing is as it should be.
- ◆ Do a two-way scatter plot of the data, making sure that the "independent" variable "fatgrm" is on the horizontal axis, using the following command:

```
graph twoway scatter dthrate fatgrm, ti("Crude Death Rate vs.  
Dietary Fat by Country")
```

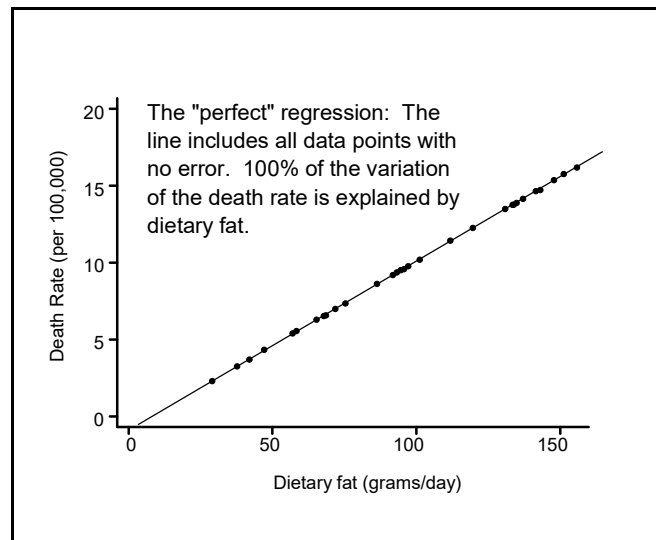
- ◆ Paste the graph into your exercise and, after you've printed it out, try to draw a line through the scatter that best represents the relationship, if any, that you see between the variables.

One way to develop a more precise quantitative understanding of the relationship between these two variables, it to try to fit a line through the scatter plot that best represents the relationship. The following graph suggests some possible lines:



I've drawn both these lines as straight lines because the scatter suggests that the relationship is linear. Lines A and B differ in their **slopes** and in their **intercepts**. They each have an equation of the form: $Y = a + BX$, where a is the vertical intercept and B is the slope of the line. Choosing the "best" line means choosing the intercept and slope values that are "optimal" in some sense.

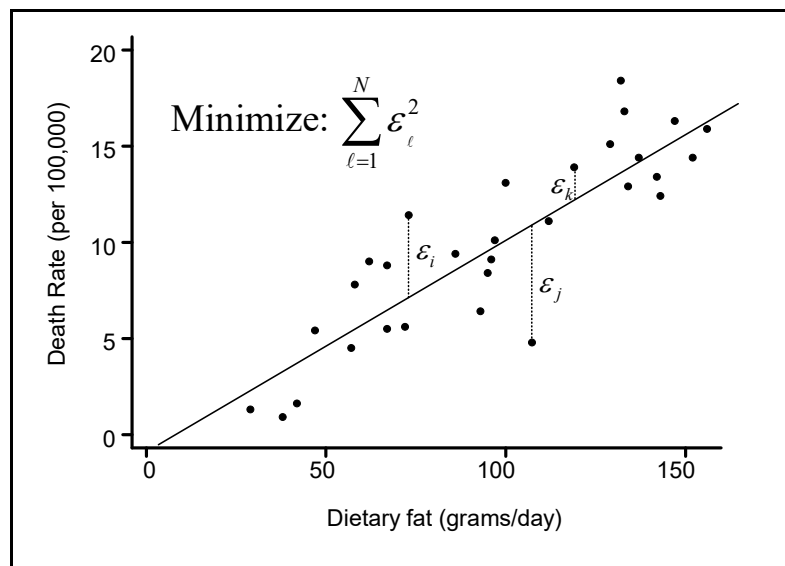
How do we define "optimal"? If all the scatter points lay along a single straight line, then we would say that the line optimally shows the relationship between dietary fat and death rate from prostate cancer:



In this case we say that the intercept and slope terms produce a line that "explains" 100% of the variation in death rates. It also suggests that getting dietary fat consumption very low will almost eliminate deaths from prostate cancer, a dubious proposition perhaps.

Unfortunately, regression lines like this almost never show up in the real world. In fact, if they do, it suggests that there's something wrong with the data. More likely is the scatter in the previous graph in which no single line can explain *all* the variation in death rates. So, perhaps a way to draw a line through the data would be to choose the line that *minimizes the unexplained variation in death rates that remains once the line has been drawn*. In the graph below, Each dotted vertical line represents the error from the regression line to the actual observation. Each observation has an implicit dotted line drawn vertically from it to the regression line. We want to find a line that minimizes the sum of these *squared errors*. Why *squared errors*? If we simply added up all the numerical errors around a regression line we would find that the sum of all these errors adds to zero! That is, the positive errors would just be canceled out by the negative errors. So, we square each error and find the line the minimizes the sum, over all N observations, of the *squared error terms*.

$$\text{Minimize: } \sum_{\ell=1}^N \varepsilon_{\ell}^2$$



This method, called the method of **least squares regression**, is ubiquitous in data analysis, and economists in particular make heavy use of it. Stata, naturally, has very extensive regression capabilities; so, let's see what kind of a regression line Stata computes for the dietary data:

- ◆ With the “fatdeath_1.dta” data set loaded into Stata, execute the following regression command:

```
regress dthrate fatgrm
```

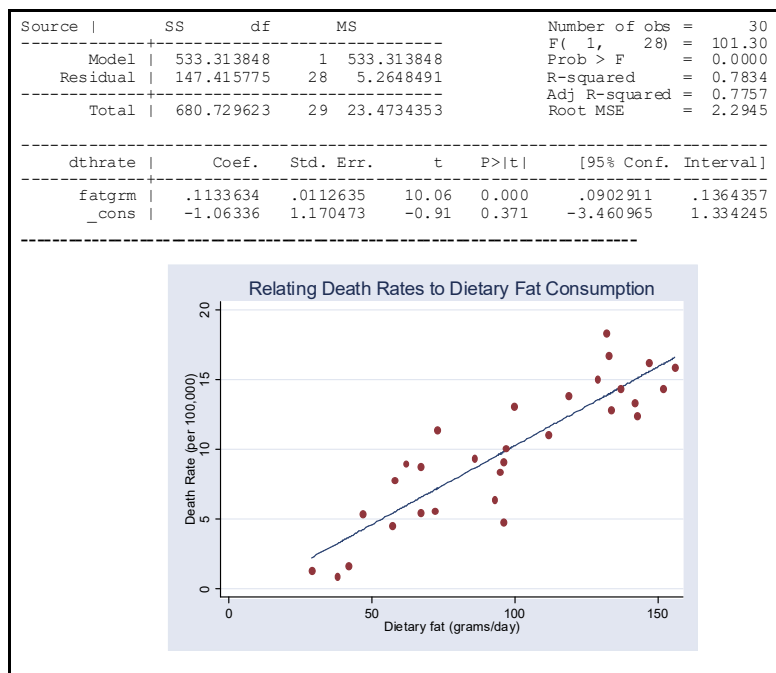
- ◆ Paste the output from this command into your exercise.

We'll return to the regression results shortly; however, now we're going to graph the regression line that this analysis tells us is the optimal straight line through these data. (An easier way to do this would be to use Stata's graphical line fitting capability. Try the command "**graph twoway lfit dthrate fatgrm || scatter dthrate fatgrm**" and see what you get. We'll do the more complicated version because it also works with multiple regression.)

- ◆ Execute the following commands to produce a *predicted value of dthrate for each value of "fatgrm" in the data set.*¹

```
predict dthratehat
label var dthratehat "Death Rate (Predicted)"
graph twoway mspline dthratehat fatgrm, bands(50) || scatter
dthrate fatgrm || , legend(off) ytitle("Death Rate (per
100,000)") ti("Relating Death Rates to Dietary Fat Consumption")
```

Here are the regression results, along with the graph of the regression line and the observations:

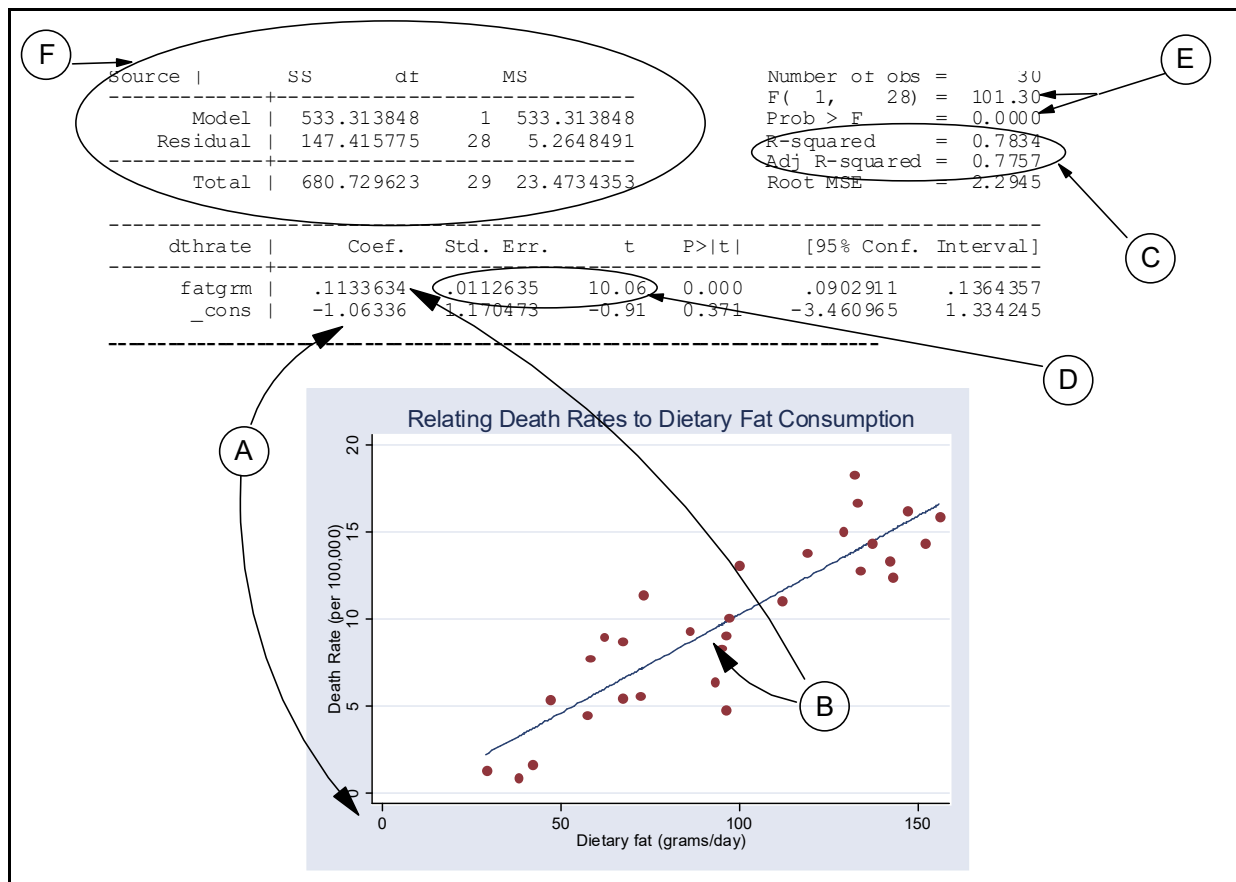


Let's interpret this "graph" command: We have both the actual and predicted values of the death rate on the vertical axis, and fat consumption on the horizontal axis. The command overlays the regression line and a scatter plot of the actual death rate data. The overlays are separated by two vertical lines, ||, typed using the far right key on the second row of keys. "graph twoway mspline" calculates cross medians and then uses the cross medians as knots to fit a cubic spline. The resulting spline is graphed as a line plot." In English this means that the separate values of

¹ See, Hamilton (Chapter "Linear Regression Analysis" Section "Predicted Values and Residuals")

dthrate are connected with a smooth line so that the resulting connected points can be plotted as a straight line (in this particular example). This is just a way of drawing straight lines between points as smoothly as possible. Next, we do a twoway scatter on the actual death rates and fat consumption. Finally, the last part of the command (after the second “||”) does commands for the entire graph -- turning the legend off, putting a title on the Y-axis and adding a title for the graph as a whole.

These commands produced a predicted value of “*dthrate*” for each value of “*fatgrm*” in the data set, based on the regression results in the table above. Let’s look a little more carefully at the regression results and the graph of the predicted regression line and the observations.



The circled letters in the graph above relate to the following paragraphs. You’re getting just a brief introduction here; we will go into much more detail later in the course.

- A. The coefficient labeled “*_cons*” in the regression results shows the estimated value of the death rate when fat consumption per day is zero. The regression suggests that the death rate is -1.06336, which suggests that the death rate from prostate cancer is *negative* when fat consumption is zero. This is highly unlikely, and suggests the difficulty of making predictions outside the range of actual data. In fact, we must eat *some* fat in order to survive, and no society in our data set consumes less than about 29 grams per day, on

average.

- B. The second estimated coefficient is the slope coefficient. Stata estimates the value of that coefficient to be 0.1133634 in these data. The meaning of this coefficient is that, on average, for every increase of one gram of fat consumption per day, the death rate for males from prostate cancer rises 0.1134... per 100,000 men. Or, if average fat consumption were to rise by 10 grams per day, the expected death rate from prostate cancer would rise by 1.134... per 100,000 men. If there were *no* relation between fat consumption and prostate cancer mortality we would expect this coefficient to be very near zero. But, while this coefficient looks small, its absolute size depends very much upon the units in which the death rate is measured. We will see in paragraph D below that this slope is quite significantly different from zero.
- C. The purpose of the regression is to draw a line that “explains” as much of the variation in the death rate as possible. The R^2 and the Adjusted- R^2 both measure the *proportion of total variation in the dependent variable, “dthrate,”* that is accounted for by its relationship with fat consumption. If all observations lie on the estimated regression line, as in the previous graph, then $R^2 = 1.0$, and 100% of the variation is explained by the association with fat consumption. If fat consumption and mortality from prostate cancer are completely unrelated, then $R^2 = 0.0$, and none of the variation in death rates is explained by fat consumption. The R^2 is calculated from information in the Analysis of Variance table (see paragraph F, below). $R^2 = \text{Variance Explained by the Regression Model} / \text{Total Variance of the Dependent Variable} = 533.313848 / 680.72962 = 0.7834$, in the current example. The current regression appears to do a good job in accounting for the variation in prostate cancer mortality across countries, explaining 78% of the variation in “dthrate” ($R^2 = 0.7834$). R^2 is very commonly used, often incorrectly, to measure the strength of the regression’s explanatory power.
- D. Stata presents measures of the degree of confidence that we have that the estimated coefficients are *significantly* different from zero. It’s far too early in the course to get specific about the “Std. Err.” and “t” coefficients circled here. Suffice it to say that the *smaller* is the estimated standard error of the coefficient the more confident we are that the regression coefficient “Coef.” (see paragraphs A and B) is significantly different from zero. Likewise, the *larger* the *absolute* value of the *t-statistic*, the more confident we are that the regression coefficient “Coef.” is different from zero. *t-statistics* larger than 1.96 in absolute terms are thought to present strong evidence that the regression coefficient is significantly different than zero. So, for the meantime, look for small standard errors and large *t-statistics* for indications that coefficients are significantly different from zero, and that the regression actually has uncovered a relationship between the two variables.
- E. Finally, Stata presents a probabilistic measure of the strength of the regression in the form of the “F-statistic.” (Calculated by the quotient $533.313848 / 5.2648491$ in the Analysis of Variance table discussed below). This statistic is derived from a random variable in the F-distribution and we can compare its value with tabulated values of the F-distribution to see the probability that the explanatory power of the regression is better than zero. The larger the F-statistic (101.30 in this example) and the smaller the associated “Prob > F”

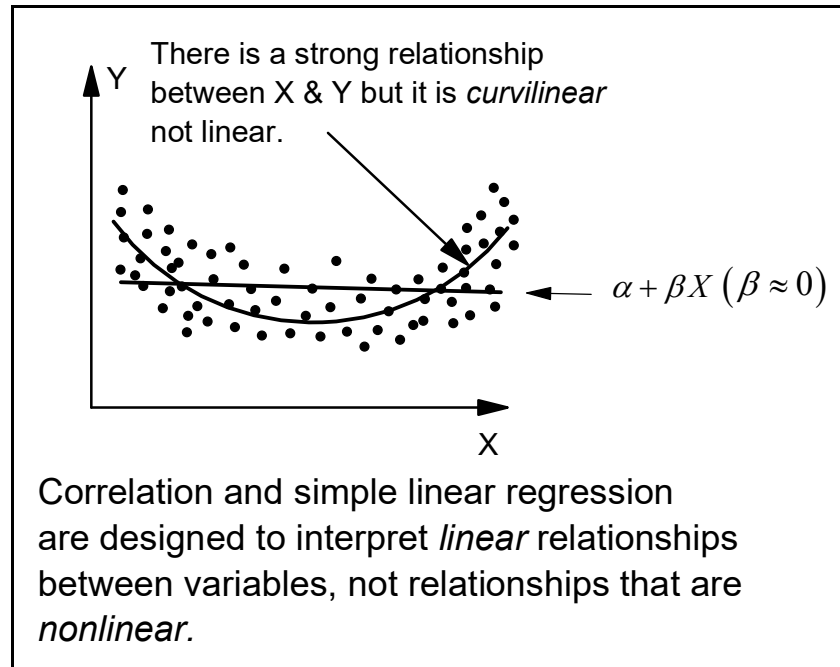
(0.0000 in this example) the higher the probability that there is a significant relationship between the two variables.

- F. The Analysis of Variance table circled here decomposes the total variation of the dependent variable, (“dthrate” in this case) into two parts, the model sum of squares and the residual sum of squares. That is the total variation of the variable “dthrate” is comprised of that which is accounted for by the regression line and the residual variation of the data from the regression line. If all the observations lay on the regression line, residual variation would be zero and the model sum of squares would equal the total sum of squares (and R^2 would equal one). The Analysis of Variance table shows the calculations necessary to produce the R^2 statistic and the F-statistic discussed above.

Again, let me reiterate that we will be studying these statistics much more completely and rigorously, later in the course; for now, it is sufficient for you to know that certain values of the statistics discussed above mean that the regression shows a statistically significant relationship between these two variables ... even if we don’t actually know what “statistically significant” means yet!

This regression analysis gives a very strong suggestion that there is a linear relationship between fat consumption and prostate cancer mortality. Of course, we need to know more about the scientific relationship between these two variables. It may well be that an unknown variable, highly correlated with fat consumption, is the real cause of elevated mortality risk; however, since we’re economists, not biologists or biostatisticians, we’ll let that problem ride.

Finally, let me reiterate that linear regression is useful for studying *linear* relationships between variables. Many variables have curvilinear relationships that may not be detected when one tries to fit a straight line to a scatter of data. The graph below shows a situation in which fitting a straight line to a scatter of data might give the mistaken impression that there is no relationship between variables X and Y when, in fact, there is a strong, but curvilinear, relationship. Using graphical analysis to detect nonlinearities in the data is always a good practice; otherwise, you might miss, or mis-specify, some important relationships.



8 Regression Analysis

Previously, you saved the “ces941_exercise.dta” file under a *new name* after you’d added new dummy variables for urban/rural residence.

- ◆ Load that revised file back into Stata so we can do some regression analysis on variables in that data set.

8.1. Predicting after-tax income

Let’s see how well we can predict after-tax CU income by using pre-tax CU income.

- ◆ Do a regression using “fincata2” (Total family income after taxes) as the dependent variable and “fincbta2” (Total family income before taxes) as the independent variable. Paste the results into your exercise and answer the following questions about this regression:
 - What proportion of the total variation in after tax income is “explained” by before tax income?
 - What is the value of the estimated slope coefficient? What does the size of this coefficient’s estimated *t-statistic* suggest about the strength of the relationship between pre- and after-tax income?
 - Look at the size of the estimated intercept coefficient: do its size and standard error indicate that the intercept is significantly different than zero?

- If pre-tax income in the average household were to rise by \$1.00, what would be the change in after-tax income? Could you explain any difference in the changes?
- ◆ Use Stata to create a scatter diagram showing the individual observations and the estimated regression line through them. Paste the resulting graph into your exercise.
 - Does the estimated *linear* regression line look like a good fit to the scatter data.? Would a curve through the data produce a better fit?
 - Speculate on why the scatter of points seems to spread as before tax income rises.

8.2. Predicting total family expenditure using after tax income

We expect that total family expenditure on current consumption (“totexp” in the data set) might be dependent upon after tax income. Let’s see if a regression analysis supports that.

- ◆ Perform a regression in which you attempt to “explain” total expenditures using after tax income. Then produce a predicted total expenditure variable and graph the regression line along with the raw total expenditure and after-tax income variables, as above. Paste the regression results and the graph into your exercise. Answer the following questions:
 - From the regression results, would you say that after-tax income is as successful in predicting (explaining) total expenditures on current consumption as before-tax income is in explaining after tax income? (Justify your answer by appealing to the regression results)
 - Does a straight line fairly represent the relationship between the two variables? Why/why not?
 - Are the slope and intercept coefficients significantly different from zero?
 - If after tax income rises by \$1.00, how much (if any) does consumption expenditure change? What happens to the money that is *not* spent?
 - Look at the scatter of points where after-tax income equals zero. How would you explain the alignment of those points? What does this imply about expenditures of families that exhibit no after-tax income?

8.3. Predicting personal care expenditures with age of head of household

Our data set contains CUs with heads of household of widely varying age.

- ◆ In Stata, produce a stem-leaf diagram for the variable “agehead”. Paste the result into your exercise. (Don’t forget to convert the diagram to “courier” font to retain alignment.)
 - What is the age of the oldest and youngest head of household?
 - Does the distribution of ages look symmetrical or asymmetrical?
- ◆ What are the mean and median ages for heads of household in this sample?

- ◆ “Personal Care Expenditures” (“persca”) are expenditures on such items as makeup, beauty salons, hair cuts, spas, etc. Do you think that these expenditures should vary depending upon the age of the head of household? Run a regression in which you try to explain personal care expenditures using age of head of household as the independent variable. Compute a predicted personal care variable and graph the regression line and the scatter of raw data. Paste the results and the graph into your exercise.
 - What is the value of the slope coefficient? Does the value and significance level of the slope coefficient indicate that there is a non-zero relationship between these two variables? (explain)
 - What is the intercept coefficient’s value? Is it significantly different from zero?
 - Looking just at the graph and regression results, what do you think the value of mean expenditure on personal care items is for this sample?
 - Summarize the “persca” variable and see what median/mean personal care expenditures are. What element(s) of the regression results/graph do they resemble?