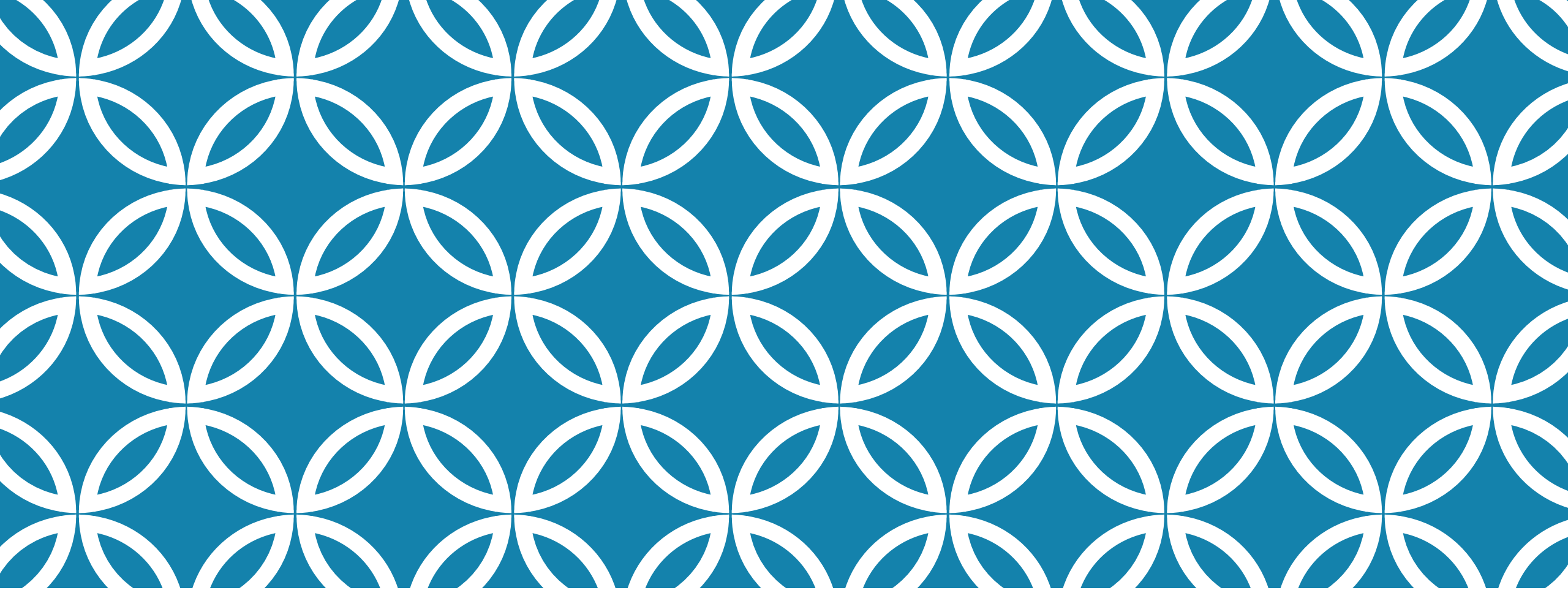


REGRESSION DIAGNOSTICS, UNUSUAL AND INFLUENTIAL DATA

Dr. Syed Hassan Raza

REGRESSION DIAGNOSTICS

Previously, we learned how to do ordinary linear regression with Stata, concluding with methods for examining the distribution of our variables. Without verifying that your data have met the assumptions underlying OLS regression, your results may be misleading. This lecture will explore how you can use Stata to check on how well your data meet the assumptions of OLS regression. In particular, we will consider the following assumptions.

REGRESSION DIAGNOSTICS

- Linearity – the relationships between the predictors and the outcome variable should be linear.

- Normality – the errors should be normally distributed – technically normality is necessary only for hypothesis tests to be valid, estimation of the coefficients only requires that the errors be identically and independently distributed.

- Homogeneity of variance (homoscedasticity) – the error variance should be constant.

REGRESSION DIAGNOSTICS

- Independence – the errors associated with one observation are not correlated with the errors of any other observation.

- Errors in variables – predictor variables are measured without error.

- Model specification – the model should be properly specified (including all relevant variables, and excluding irrelevant variables).

REGRESSION DIAGNOSTICS

Additionally, there are issues that can arise during the analysis that, while strictly speaking are not assumptions of regression, are none the less, of great concern to data analysts.

- Influence – individual observations that exert undue influence on the coefficients.

- Collinearity – predictors that are highly collinear, i.e., linearly related, can cause problems in estimating the regression coefficients.

REGRESSION DIAGNOSTICS

Many graphical methods and numerical tests have been developed over the years for regression diagnostics. Stata has many of these methods built-in, and others are available that can be downloaded over the internet. In particular, Nicholas J. Cox (University of Durham) has produced a collection of convenience commands which can be downloaded from SSC (**ssc install** *commandname*). These commands include **indexplot**, **rvfplot2**, **rdplot**, **qfrplot** and **ovfplot**. In this chapter, we will explore these methods and show how to verify regression assumptions and detect potential problems using Stata.

UNUSUAL AND INFLUENTIAL DATA

A single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis. If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. There are three ways that an observation can be unusual.

UNUSUAL AND INFLUENTIAL DATA

- **Outliers**: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

- **Leverage**: An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far an observation deviates from the mean of that variable. These leverage points can have an effect on the estimate of regression coefficients.

- **Influence**: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

UNUSUAL AND INFLUENTIAL DATA

How can we identify these three types of observations? Let's look at an example dataset called **crime**. This dataset appears in *Statistical Methods for Social Sciences, Third Edition* by Alan Agresti and Barbara Finlay (Prentice Hall, 1997). The variables are state id (**sid**), state name (**state**), violent crimes per 100,000 people (**crime**), murders per 1,000,000 (**murder**), the percent of the population living in metropolitan areas (**pctmetro**), the percent of the population that is white (**pctwhite**), percent of population with a high school education or above (**pcths**), percent of population living under poverty line (**poverty**), and percent of population that are single parents (**single**).

UNUSUAL AND INFLUENTIAL DATA

- use https://stats.idre.ucla.edu/stat/stata/webbooks/reg/crime
- describe

```
Contains data from crime.dta
  obs:             51                          crime data from agresti &
                                                 finlay - 1997
 vars:             11                          6 Feb 2001 13:52
 size:          2,295 (98.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
sid             float  %9.0g
state           str3   %9s
crime           int    %8.0g                 violent crime rate
murder          float  %9.0g                 murder rate
pctmetro        float  %9.0g                 pct metropolitan
pctwhite        float  %9.0g                 pct white
pcths           float  %9.0g                 pct hs graduates
poverty         float  %9.0g                 pct poverty
single          float  %9.0g                 pct single parent
-------------------------------------------------------------------------------
Sorted by:
```

UNUSUAL AND INFLUENTIAL DATA

- summarize crime murder pctmetro pctwhite pcths poverty single

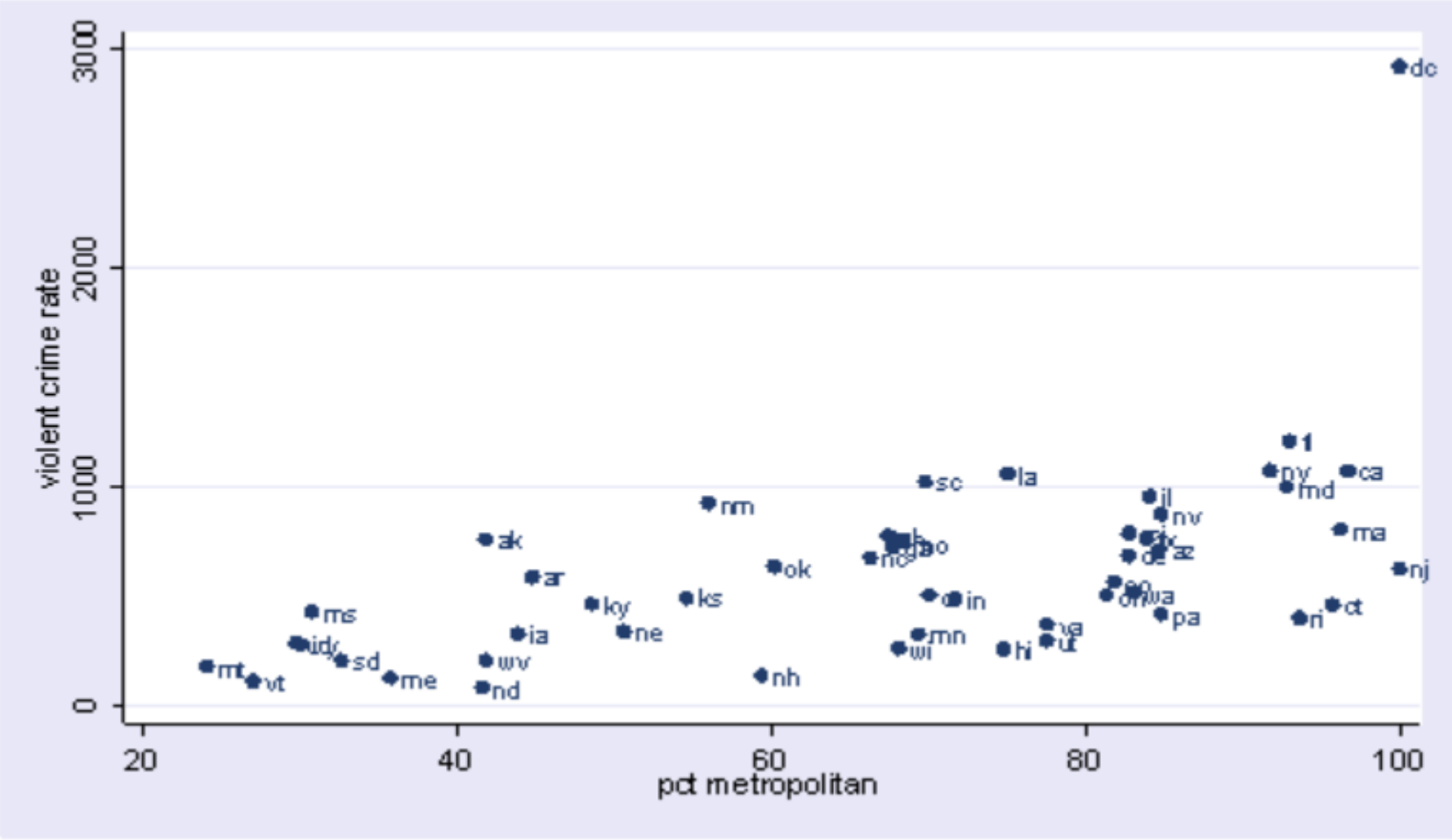
Variable	Obs	Mean	Std. Dev.	Min	Max
crime	51	612.8431	441.1003	82	2922
murder	51	8.727451	10.71758	1.6	78.5
pctmetro	51	67.3902	21.95713	24	100
pctwhite	51	84.11569	13.25839	31.8	98.5
pcths	51	76.22353	5.592087	64.3	86.6
poverty	51	14.25882	4.584242	8	26.4
single	51	11.32549	2.121494	8.4	22.1

UNUSUAL AND INFLUENTIAL DATA

Let's say that we want to predict **crime** by **pctmetro**, **poverty**, and **single**. That is to say, we want to build a linear regression model between the response variable **crime** and the independent variables **pctmetro**, **poverty** and **single**. We will first look at the scatter plots of crime against each of the predictor variables before the regression analysis so we will have some ideas about potential problems. Let's make individual graphs of **crime** with **pctmetro** and **poverty** and **single** so we can get a better view of these scatterplots. We will add the **mlabel(state)** option to label each marker with the state name to identify outlying states.

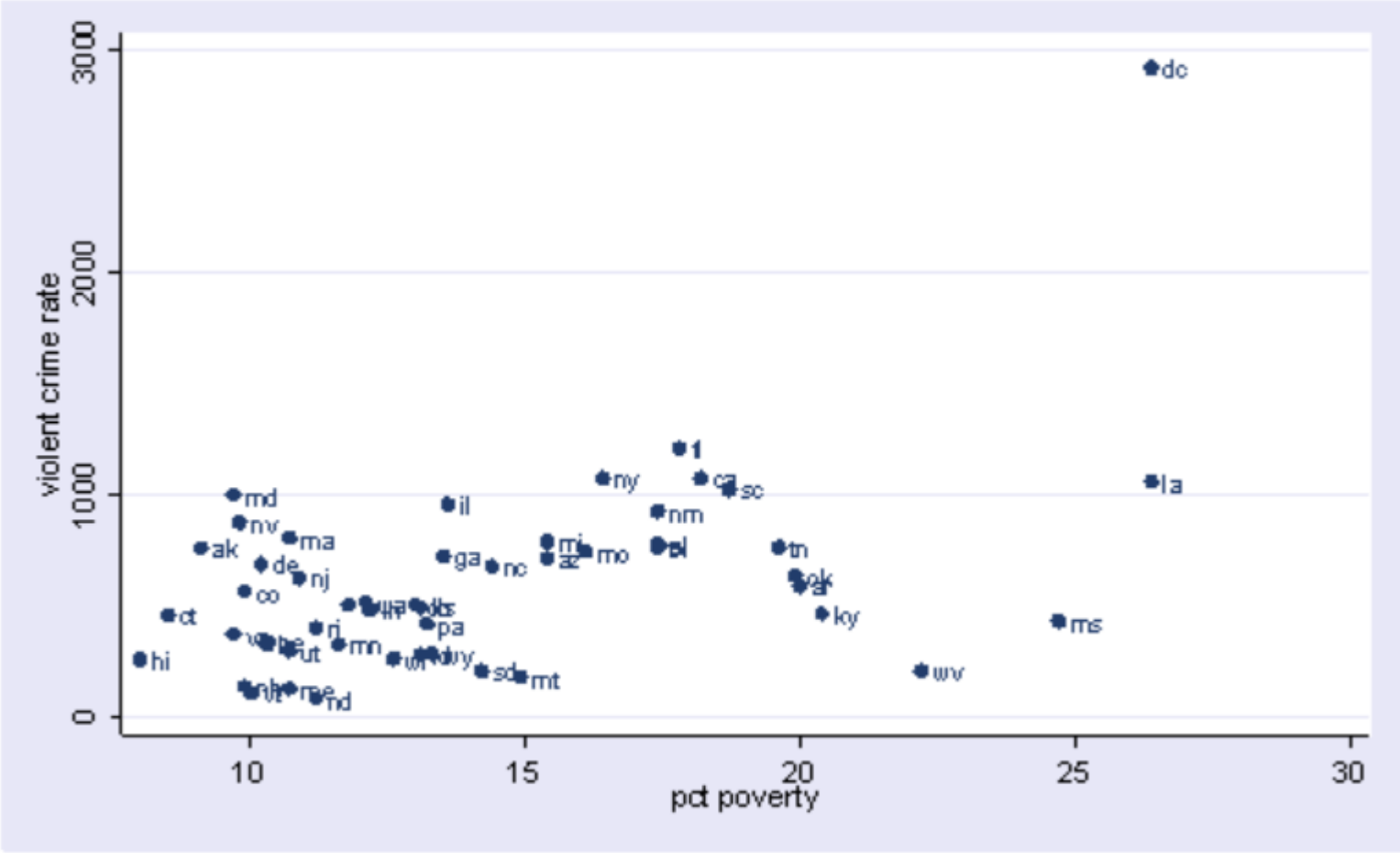
UNUSUAL AND INFLUENTIAL DATA

- scatter crime pctmetro, mlabel(state)



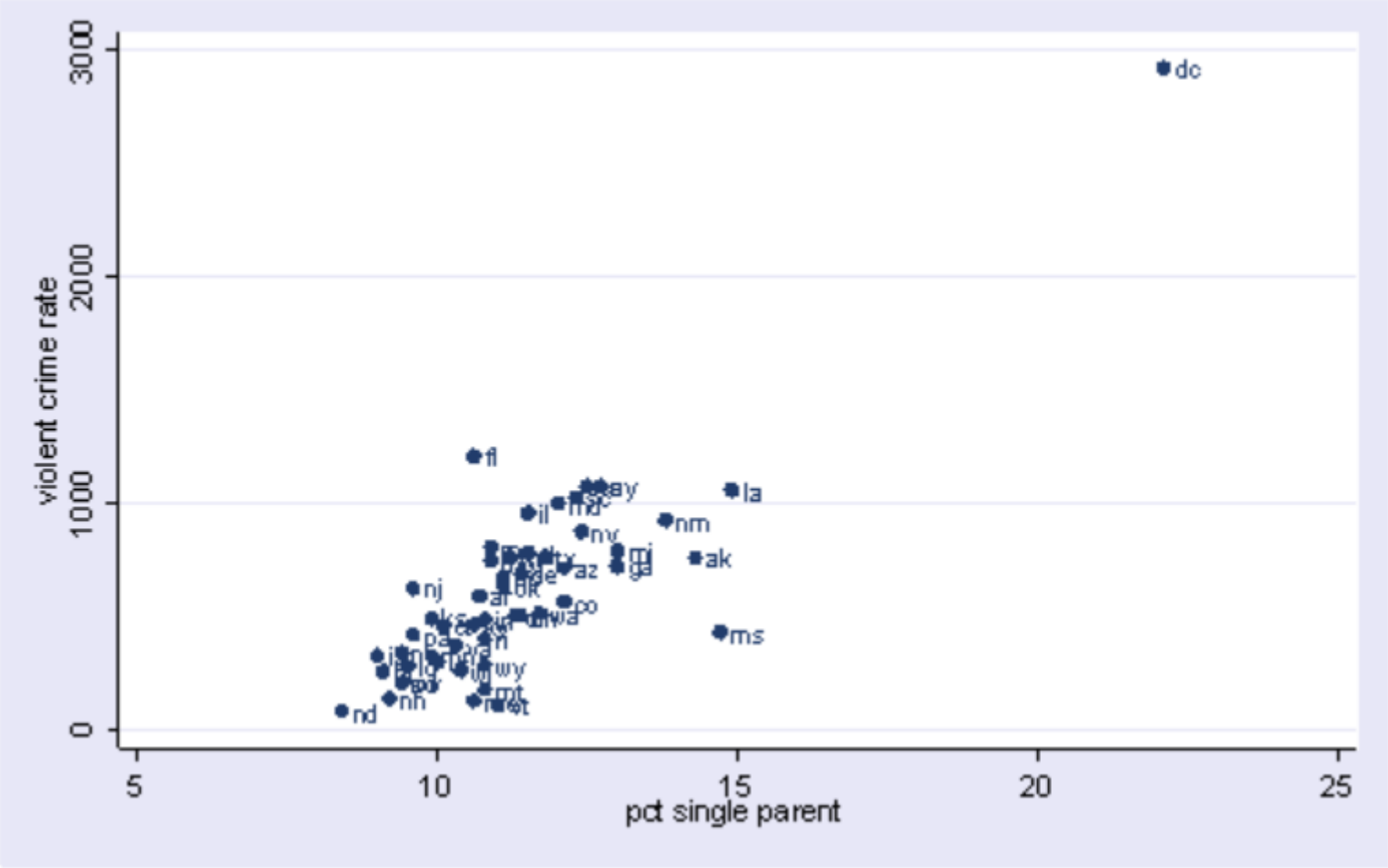
UNUSUAL AND INFLUENTIAL DATA

- scatter crime poverty, mlabel(state)



UNUSUAL AND INFLUENTIAL DATA

- scatter crime single, mlabel(state)



UNUSUAL AND INFLUENTIAL DATA

All the scatter plots suggest that the observation for **state** = dc is a point that requires extra attention since it stands out away from all of the other points. We will keep it in mind when we do our regression analysis.

Now let's try the regression command predicting **crime** from **pctmetro poverty** and **single**. We will go step-by-step to identify all the potentially unusual or influential points afterwards.

UNUSUAL AND INFLUENTIAL DATA

- regress crime pctmetro poverty single

```
  Source |       SS       df       MS              Number of obs =      51
---------+------------------------------           F(  3,    47) =   82.16
   Model |  8170480.21     3  2723493.40            Prob > F      =  0.0000
Residual |  1557994.53    47  33148.8199            R-squared     =  0.8399
---------+------------------------------           Adj R-squared =  0.8296
   Total |  9728474.75    50  194569.495            Root MSE      =  182.07

------------------------------------------------------------------------------
       crime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pctmetro |    7.828935   1.254699     6.240   0.000     5.304806    10.35306
     poverty |    17.68024    6.94093     2.547   0.014     3.716893    31.64359
      single |    132.4081   15.50322     8.541   0.000     101.2196    163.5965
       _cons |   -1666.436    147.852   -11.271   0.000    -1963.876   -1368.996
------------------------------------------------------------------------------
```

UNUSUAL AND INFLUENTIAL DATA

Let's examine the studentized residuals as a first means for identifying outliers. Below we use the **predict** command with the **rstudent** option to generate studentized residuals and we name the residuals **r**. We can choose any name we like as long as it is a legal Stata variable name. Studentized residuals are a type of standardized residual that can be used to identify outliers.

- predict r, rstudent

UNUSUAL AND INFLUENTIAL DATA

Let's sort the data on the residuals and show the 10 largest and 10 smallest residuals along with the state id and state name. Note that in the second **list** command the **-10/l** the last value is the letter "l", NOT the number one.

- sort r
- list sid state r in 1/10

	sid	state	r
1.	25	ms	-3.570789
2.	18	la	-1.838577
3.	39	ri	-1.685598
4.	47	wa	-1.303919
5.	35	oh	-1.14833
6.	48	wi	-1.12934
7.	6	co	-1.044952
8.	22	mi	-1.022727
9.	4	az	-.8699151
10.	44	ut	-.8520518

UNUSUAL AND INFLUENTIAL DATA

- list sid state r in -10/l

	sid	state	r
42.	24	mo	.8211724
43.	20	md	1.01299
44.	29	ne	1.028869
45.	40	sc	1.030343
46.	16	ks	1.076718
47.	14	il	1.151702
48.	13	id	1.293477
49.	12	ia	1.589644
50.	9	fl	2.619523
51.	51	dc	3.765847

UNUSUAL AND INFLUENTIAL DATA

- We should pay attention to studentized residuals that exceed +2 or -2, and get even more concerned about residuals that exceed +2.5 or -2.5 and even yet more concerned about residuals that exceed +3 or -3. These results show that DC and MS are the most worrisome observations followed by FL.

- Another way to get this kind of output is with a command called **hilo**. You can download **hilo** from within Stata by typing **search hilo** (see How can I used the search command to search for programs and get additional help? for more information about using **search**).

- Once installed, you can type the following and get output similar to that above by typing just one command

UNUSUAL AND INFLUENTIAL DATA

- hilo r state

➢10 smallest and largest observations on r

r	state
-3.570789	ms
-1.838577	la
-1.685598	ri
-1.303919	wa
-1.14833	oh
-1.12934	wi
-1.044952	co
-1.022727	mi
-.8699151	az
-.8520518	ut

r	state
8211724	mo
1.01299	md
1.028869	ne
1.030343	sc
1.076718	ks
1.151702	il
1.293477	id
1.589644	ia
2.619523	fl
3.765847	dc

UNUSUAL AND INFLUENTIAL DATA

Let's show all of the variables in our regression where the studentized residual exceeds +2 or -2, i.e., where the absolute value of the residual exceeds 2. We see the data for the three potential outliers we identified, namely Florida, Mississippi and Washington D.C. Looking carefully at these three observations, we couldn't find any data entry error, though we may want to do another regression analysis with the extreme point such as DC deleted. We will return to this issue later.

UNUSUAL AND INFLUENTIAL DATA

- list r crime pctmetro poverty single if abs(r) > 2

	r	crime	pctmetro	poverty	single
1.	-3.570789	434	30.7	24.7	14.7
50.	2.619523	1206	93	17.8	10.6
51.	3.765847	2922	100	26.4	22.1

UNUSUAL AND INFLUENTIAL DATA

Now let’s look at the leverage’s to identify observations that will have potential great influence on regression coefficient estimates.

- predict lev, leverage

UNUSUAL AND INFLUENTIAL DATA

We use the show(5) high options on the hilo command to show just the 5 largest observations (the high option can be abbreviated as h). We see that DC has the largest leverage.

- hilo lev state, show(5) high

```
5 largest observations on lev

      lev      state
.1652769         la
.1802005         wv
 .191012         ms
.2606759         ak
 .536383         dc
```

UNUSUAL AND INFLUENTIAL DATA

Generally, a point with leverage greater than **(2k+2)/n** should be carefully examined. Here **k** is the number of predictors and **n** is the number of observations. In our example, we can do the following.

- display (2*3+2)/51

➢.15686275

UNUSUAL AND INFLUENTIAL DATA

- list crime pctmetro poverty single state lev if lev >.156

	crime	pctmetro	poverty	single	state	lev
5.	208	41.8	22.2	9.4	wv	.1802005
48.	761	41.8	9.1	14.3	ak	.2606759
49.	434	30.7	24.7	14.7	ms	.191012
50.	1062	75	26.4	14.9	la	.1652769
51.	2922	100	26.4	22.1	dc	.536383

As we have seen, DC is an observation that both has a large residual and large leverage. Such points are potentially the most influential.