

Introduction to Stata

17.871

Spring 2015

Before I start...

- You will be greatly helped if you download the datafiles that are associated with the Kohler and Kreuter book.

See pp. xxii and xxiii for how to do this.

The role of statistical packages in research

- Obvious answer
 - Manage data
 - Carry out appropriate statistical tests
 - Assist in displaying data
- Less obvious answer
 - Channel the type of research you are likely to do
 - Limitations as to variables and cases
 - Types of analysis is sometimes guided by choice of package

Analysis -> Packages

- Baby exercises
 - Minitab, spreadsheets
- Time series
 - TSP
- Cross-sectional
 - SPSS, SAS
- Time series & cross-sectional
 - Stata, R

Logic of quant research in this class

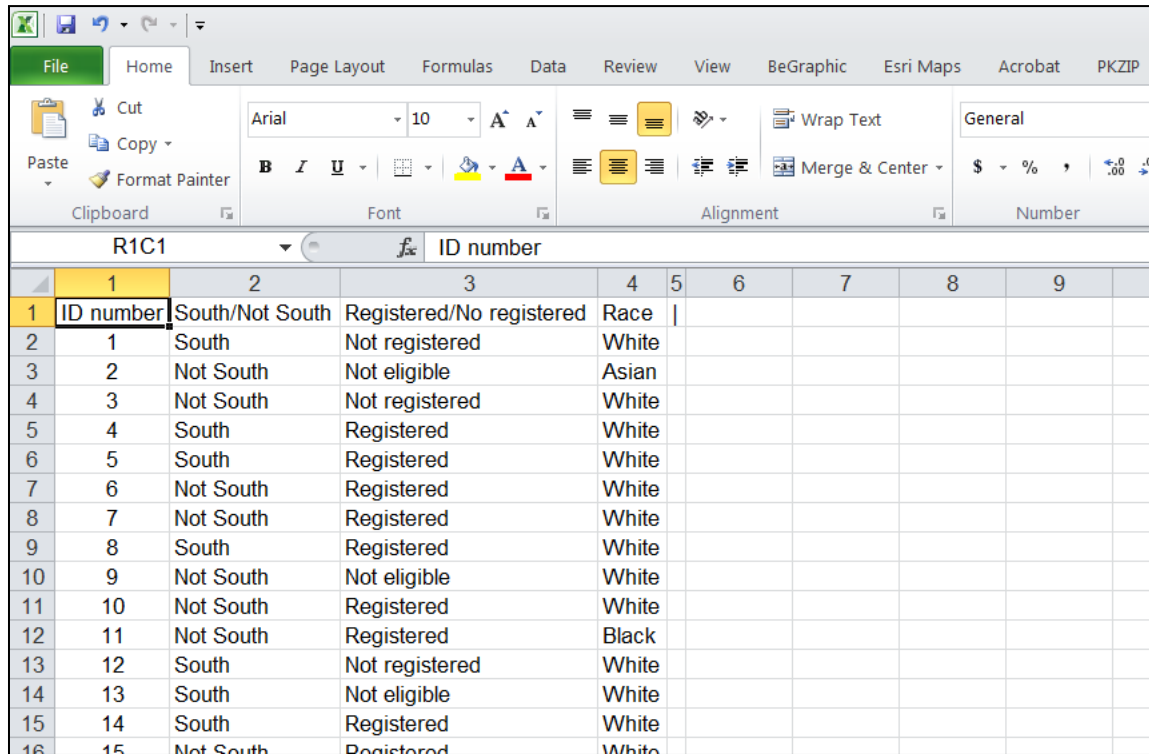
$$y_i = f(x_i, \beta, \varepsilon_i)$$

Logic of data setup:

	V_1	V_2	...	V_j
Obs ₁				
Obs ₂				
...				
Obs _i				

Suppose I wanted to test whether African Americans and whites were registered to vote at the same rates in southern states, compared to all other states.

What you would like to find



The screenshot shows the Microsoft Excel interface with the following data table:

R1C1 ID number										
	1	2	3	4	5	6	7	8	9	
1	ID number	South/Not South	Registered/No registered	Race						
2	1	South	Not registered	White						
3	2	Not South	Not eligible	Asian						
4	3	Not South	Not registered	White						
5	4	South	Registered	White						
6	5	South	Registered	White						
7	6	Not South	Registered	White						
8	7	Not South	Registered	White						
9	8	South	Registered	White						
10	9	Not South	Not eligible	White						
11	10	Not South	Registered	White						
12	11	Not South	Registered	Black						
13	12	South	Not registered	White						
14	13	South	Not eligible	White						
15	14	South	Registered	White						
16	15	Not South	Registered	White						

What you actually find*

HRHHID	GESTCEN	PES1	PES2	PTDTRACE
130609914027385	21	2	2	1
988009445479950	87	-1	-1	4
608038914219250	64	2	2	1
612981028909492	61	1	-1	1
9093130995291	12	1	-1	1
701099064904672	51	1	-1	1
645081921524170	54	2	1	1
506904455991733	91	1	-1	1
756490679500849	57	-2	-2	1
140809958804389	74	1	-1	1
927389210899305	56	1	-1	2
40621280985198	32	2	2	1
108903931992469	42	-1	-1	1
5570150694693	71	1	-1	1
689400611360960	93	-1	-1	1

*From the 2012 Voting and Registration Study of the Current Population Survey, U.S. Census

Dataset: CPS//Voting and Registration/Nov 2012

GESTCEN
Geography-census state code

With the following Ranges:

- 11 ME
- 12 NH
- 13 VT
- 14 MA
- 15 RI
- 16 CT
- 21 NY
- 22 NJ
- 23 PA
- 31 OH
- 32 IN
- 33 IL
- 34 MI
- 35 WI
- 41 MN
- 42 IA
- 43 MO
- 44 ND
- 45 SD
- 46 NE
- 47 KS
- 51 DE
- 52 MD
- 53 DC
- 54 VA
- 55 ...

HRHHID
Household-identifier,scrambled

PES1
CPS PES1 Vote-Vote in the November election

With the following Ranges:

- 9 No response
- 3 Refused
- 2 Don't Know
- 1 Not in Universe
- 1 Yes
- 2 No

PES2
CPS PES2 Vote-Registered to vote in the November election

With the following Ranges:

- 9 No response
- 3 Refused
- 2 Don't Know
- 1 Not in Universe
- 1 Yes
- 2 No

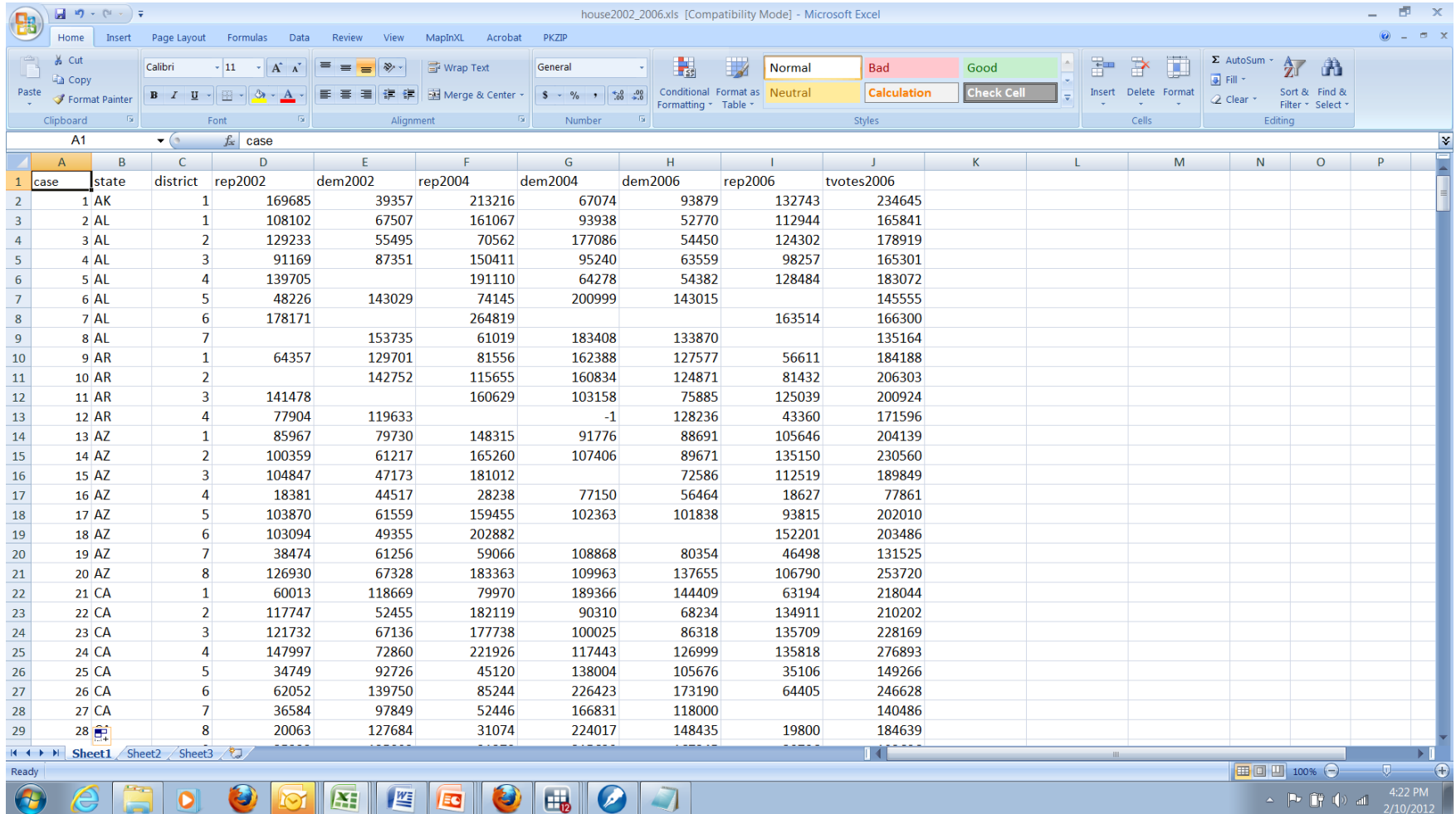
PTDTRACE
Demographics- race of respondent

With the following Ranges:

- 1 White Only
- 2 Black Only
- 3 American Indian, Alaskan Native Only
- 4 Asian Only
- 5 Hawaiian/Pacific Islander Only
- 6 White-Black
- 7 White-AI
- 8 White-Asian
- 9 White-Hawaiian
- 10 Black-AI
- 11 Black-Asian
- 12 Black-HP
- 13 AI-Asian
- 14 Asian-HP
- 15 W-B-AI
- 16 W-B-A
- 17 W-AI-A
- 18 W-A-HP
- 19 W-B-AI-A
- 20 2 or 3 Races
- 21 4 or 5 Races

PWSSWGT
Weight-second stage weight (rake 6 final step weight)

Example, House Elections



house2002_2006.xls [Compatibility Mode] - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	case	state	district	rep2002	dem2002	rep2004	dem2004	dem2006	rep2006	tvotes2006						
2	1 AK		1	169685	39357	213216	67074	93879	132743	234645						
3	2 AL		1	108102	67507	161067	93938	52770	112944	165841						
4	3 AL		2	129233	55495	70562	177086	54450	124302	178919						
5	4 AL		3	91169	87351	150411	95240	63559	98257	165301						
6	5 AL		4	139705		191110	64278	54382	128484	183072						
7	6 AL		5	48226	143029	74145	200999	143015		145555						
8	7 AL		6	178171		264819			163514	166300						
9	8 AL		7		153735	61019	183408	133870		135164						
10	9 AR		1	64357	129701	81556	162388	127577	56611	184188						
11	10 AR		2		142752	115655	160834	124871	81432	206303						
12	11 AR		3	141478		160629	103158	75885	125039	200924						
13	12 AR		4	77904	119633		-1	128236	43360	171596						
14	13 AZ		1	85967	79730	148315	91776	88691	105646	204139						
15	14 AZ		2	100359	61217	165260	107406	89671	135150	230560						
16	15 AZ		3	104847	47173	181012		72586	112519	189849						
17	16 AZ		4	18381	44517	28238	77150	56464	18627	77861						
18	17 AZ		5	103870	61559	159455	102363	101838	93815	202010						
19	18 AZ		6	103094	49355	202882			152201	203486						
20	19 AZ		7	38474	61256	59066	108868	80354	46498	131525						
21	20 AZ		8	126930	67328	183363	109963	137655	106790	253720						
22	21 CA		1	60013	118669	79970	189366	144409	63194	218044						
23	22 CA		2	117747	52455	182119	90310	68234	134911	210202						
24	23 CA		3	121732	67136	177738	100025	86318	135709	228169						
25	24 CA		4	147997	72860	221926	117443	126999	135818	276893						
26	25 CA		5	34749	92726	45120	138004	105676	35106	149266						
27	26 CA		6	62052	139750	85244	226423	173190	64405	246628						
28	27 CA		7	36584	97849	52446	166831	118000		140486						
29	28 CA		8	20063	127684	31074	224017	148435	19800	184639						

Example, World Bank World Development Indicators

57a6e89c-6c95-4b52-82ea-23c3ad6097ca.xlsx - Microsoft Excel

Country Name						
A	B	C	D	E	F	G
Country Name	Country Code	Series Name	Series Code	2010 [YR2010]	2011 [YR2011]	2012 [YR2012]
Afghanistan	AFG	Adjusted net enrollment rate, primary (% of primary school age children)	SE.PRM.TENR	..	..	..
Afghanistan	AFG	Adjusted net enrollment rate, primary, female (% of primary school age children)	SE.PRM.TENR.FE	..	..	..
Afghanistan	AFG	Adjusted net enrollment rate, primary, male (% of primary school age children)	SE.PRM.TENR.MA	..	..	..
Afghanistan	AFG	Children out of school, primary	SE.PRM.UNER	..	..	..
Afghanistan	AFG	Children out of school, primary, female	SE.PRM.UNER.FE	..	..	..
Afghanistan	AFG	Children out of school, primary, male	SE.PRM.UNER.MA	..	..	..
Afghanistan	AFG	Primary education, pupils	SE.PRM.ENRL	5279326	5291624	5767543
Afghanistan	AFG	Primary education, pupils (% female)	SE.PRM.ENRL.FE.ZS	39.34104	40.47591	40.71588
Afghanistan	AFG	Ratio of female to male primary enrollment (%)	SE.ENR.PRIM.FM.ZS	67.801	71.094	71.808
Afghanistan	AFG	Ratio of female to male secondary enrollment (%)	SE.ENR.SECO.FM.ZS	49.142	54.879	55.338
Afghanistan	AFG	Ratio of female to male tertiary enrollment (%)	SE.ENR.TERT.FM.ZS	..	33.227	..
Afghanistan	AFG	Ratio of girls to boys in primary and secondary education (%)	SE.ENR.PRSC.FM.ZS	62.136	65.97	66.593
Afghanistan	AFG	School enrollment, preprimary (% gross)	SE.PRE.ENRR	..	..	..
Afghanistan	AFG	School enrollment, preprimary, female (% gross)	SE.PRE.ENRR.FE	..	..	..
Afghanistan	AFG	School enrollment, preprimary, male (% gross)	SE.PRE.ENRR.MA	..	..	..
Afghanistan	AFG	School enrollment, primary (% gross)	SE.PRM.ENRR	99.39882	97.40102	103.94475
Afghanistan	AFG	School enrollment, primary (% net)	SE.PRM.NENR	..	..	..
Afghanistan	AFG	School enrollment, primary, female (% gross)	SE.PRM.ENRR.FE	79.9847	80.64234	86.57215
Afghanistan	AFG	School enrollment, primary, female (% net)	SE.PRM.NENR.FE	..	..	..
Afghanistan	AFG	School enrollment, primary, male (% gross)	SE.PRM.ENRR.MA	117.96969	113.43012	120.56038
Afghanistan	AFG	School enrollment, primary, male (% net)	SE.PRM.NENR.MA	..	..	..
Afghanistan	AFG	School enrollment, primary, private (% of total primary)	SE.PRM.PRIV.ZS	..	1.67143	2.48133
Afghanistan	AFG	School enrollment, secondary (% gross)	SE.SEC.ENRR	50.1815	51.77018	54.03824
Afghanistan	AFG	School enrollment, secondary (% net)	SE.SEC.NENR	..	..	..
Afghanistan	AFG	School enrollment, secondary, female (% gross)	SE.SEC.ENRR.FE	32.82917	36.45946	38.26146
Afghanistan	AFG	School enrollment, secondary, female (% net)	SE.SEC.NENR.FE	..	..	..
Afghanistan	AFG	School enrollment, secondary, male (% gross)	SE.SEC.ENRR.MA	66.80491	66.43656	69.14169
Afghanistan	AFG	School enrollment, secondary, male (% net)	SE.SEC.NENR.MA	..	..	..
Afghanistan	AFG	School enrollment, secondary, private (% of total secondary)	SE.SEC.PRIV.ZS	..	1.27354	1.94761
Afghanistan	AFG	School enrollment, tertiary (% gross)	SE.TER.ENRR	..	3.74394	..
Afghanistan	AFG	School enrollment, tertiary, female (% gross)	SE.TER.ENRR.FE	..	1.85067	..

Using Stata to Analyze Data in Matrix Form

- Question: Did Ron Paul do better in Iowa in 2012, compared to 2008 in counties with college students?
- Data sources:
 - 2008: Des Moines Register web site, (<http://caucuses.desmoinesregister.com/data/iowa-caucus/caucus-history-gop/>)
 - 2012: Iowa Republican Party, Google Doc (<https://www.google.com/fusiontables/DataSource?dsrcid=2475248>)

2008

Firefox

AACSB Business... standards-busn... Support | IS&T sfprwhmr_pdf.s... MIT - people di... Mccorm-I Admi... Google Docs - ... Largest College... Iowa Caucuses ... Iowa Caucus 20...

caucuses.desmoinesregister.com/data/iowa-caucus/caucus-history-gop/ iowa republican party caucus results 2008 county

LibX MIT Keyword Search MIT's WorldCat Clear Scholar

In the past four Republican presidential caucuses, the winner of the Iowa caucuses has gone on to win the nomination twice and the presidency once. The main role of the caucuses is seen as an important test of strength that winnows the field. From the start of the modern Iowa caucuses in 1972 through 2004, no candidate who finished worse than third had gone on to win a nomination. In 2008, ultimate nominee John McCain broke the pattern, finishing fourth by a whisker. Besides trimming the field, the caucuses have changed the contours of the race in other ways. In 1988, a disappointing third-place Iowa finish persuaded Vice President George H.W. Bush to change his campaign strategy from an assumption of inevitability to a more aggressive campaign, with more frequent public appearances. In 1996 and 2000, the Iowa caucuses confirmed the front-runner status of Bob Dole and George W. Bush respectively. And in 2008, the caucuses propelled a little-known candidate, former Arkansas Gov. Mike Huckabee, into national prominence.

YEAR
1988 1996 2000 2008

Former Arkansas Gov. Mike Huckabee rode a populist message and support from Christian conservatives to victory. His win reinforced the caucuses' More ...

Results Turnout

Iowa
Click to change map VOTES PCTG.

1. Mike Huckabee	40,954	34.4%
2. Mitt Romney	30,021	25.2%
3. Fred Thompson	15,960	13.4%
4. John McCain	15,536	13%
5. Ron Paul	11,841	9.9%
6. Rudy Giuliani	4,099	3.4%

NOMINEE IN RED

COUNTIES WON BY
Huckabee Thompson Romney McCain Paul Giuliani

County by county (Click the map to filter the list. Hold down the shift key to select more than one county.)

All counties	Mike Huckabee	Fred Thompson	Mitt Romney	John McCain	Ron Paul	Rudy Giuliani
Adair	73 votes 34.1%	67 votes 31.3%	40 votes 18.7%	14 votes 6.5%	14 votes 6.5%	—
Adams	76 votes 40.6%	26 votes 13.9%	44 votes 23.5%	21 votes 11.2%	14 votes 7.5%	—
Allamakee	113 votes 26.8%	29 votes 6.9%	117 votes 27.7%	58 votes 13.7%	88 votes 20.9%	—
Appanoose	262 votes 55.2%	62 votes 13.1%	53 votes 11.2%	39 votes 8.2%	41 votes 8.6%	—
Audubon	68 votes 34.9%	26 votes 13.3%	52 votes 26.7%	29 votes 14.9%	13 votes 6.7%	—
Benton	341 votes 34.9%	187 votes 19.2%	204 votes 20.9%	83 votes 8.5%	109 votes 11.2%	—
Black Hawk	1,084 votes 34.8%	453 votes 14.5%	846 votes 27.2%	321 votes 10.3%	312 votes 10%	—
Butte	543 votes	192 votes	279 votes	128 votes	149 votes	—

XFINITY INTERNET
Comcast

Meet the candidates

Bachmann Gingrich Huntsman

Johnson Paul Perry

Roemer Romney Santorum

Barack Obama

About the Democratic caucuses

Barack Obama's record

OTHER CANDIDATES

Campaign trail videos

Michele Bachmann drop...

Mitt Romney on caucus night

Find: google d

Next Previous Highlight all Match case Phrase not found

Windows Explorer Mail Word PowerPoint

4:38 PM 2/10/2012

2012

Firefox | AACSB Business... | standards-busn... | Support | IS&T | sfprwhmr.pdf.s... | MIT - people di... | Mccorm-I Admi... | Google Docs - ... | Largest College... | Iowa GOP Relea... | Iowa Caucus 20... | + | - | [Icons]

google.com | https://www.google.com/fusiontables/DataSource?dsrclid=2475248 | ☆ | [Icons] | google doc iowa republican caucus precinct | [Icons] | Search MIT's WorldCat | Clear | Scholar

+You | Gmail | Calendar | Documents | Photos | Sites | Search | More | Sign in | [Icon]

Iowa Caucus 2012 Precinct Results

Republican Party of Iowa (RPI)

File | View | Edit | Visualize | Merge | Experiment

Showing all rows [options](#) 1 - 100 of [many](#) [Next >](#)

ID ▾	ParentID ▾	DistrictName-en ▾	NumVoters ▾	NumBall	NumCou	VoteCount-P	VoteCount-B	VoteCount-Johnson ▾	VoteCount-Gi	VoteCount-S	VoteCou	[Icon]	[Icon]
10001	19	Buchanan - Aurora	3	1	1	0	0		2	0	0	[Icon]	
10002	19	Buchanan - Brandon	20	1	1	6	2		1	10	0	[Icon]	
10003	19	Buchanan - Fairbank	12	1	1	2	1		0	5	0	[Icon]	
10004	19	Buchanan - Hazleton	18	1	1	1	4		5	4	0	[Icon]	
10005	19	Buchanan - Indep 1st Ward/Washington	94	1	1	13	4		15	28	0	[Icon]	
10006	19	Buchanan - Indep 2nd Ward	25	1	1	4	6		2	7	0	[Icon]	
10007	19	Buchanan - Indep 3rd Ward	37	1	1	7	2		12	4	1	[Icon]	
10008	19	Buchanan - Indep 4th Ward	12	1	1	1	0		4	2	0	[Icon]	
10009	19	Buchanan - Indep 5th Ward	21	1	1	4	3		2	1	0	[Icon]	
1001	1	Adair - 1NW ADAIR	46	1	1	7	4		11	10	0	[Icon]	
10010	19	Buchanan - Jesup	36	1	1	7	0		10	13	0	[Icon]	
10011	19	Buchanan - Lamont	15	1	1	2	0		0	9	0	[Icon]	
10012	19	Buchanan - Perry/Westburg	35	1	1	4	5		4	16	0	[Icon]	
10013	19	Buchanan - Quasqueton	51	1	1	7	5		5	17	0	[Icon]	
10014	19	Buchanan - Rowley	45	1	1	5	4		11	11	0	[Icon]	
10015	19	Buchanan - Winthrop	35	1	1	3	4		4	6	0	[Icon]	
1002	1	Adair - 2NE STUART	51	1	1	8	5		3	15	1	[Icon]	
1003	1	Adair - 3SW FONTANELLE	55	1	1	9	6		16	14	0	[Icon]	
1004	1	Adair - 4SE ORIENT	50	1	1	4	6		6	15	0	[Icon]	
1005	1	Adair - 5GF GREENFIELD	67	1	1	14	5		8	12	0	[Icon]	
11001	21	Buena Vista - Albert City Fairfield Coon	73	1	1	9	4		11	35	0	[Icon]	
11002	21	Buena Vista - Alta Nokomis Elk	97	1	1	29	2		11	21	0	[Icon]	
11003	21	Buena Vista - Lakeside Hayes Maple Valley	80	1	1	20	2		20	9	1	[Icon]	
11004	21	Buena Vista - Linn Grove Brooke	26	1	1	0	0		0	14	1	[Icon]	
11005	21	Buena Vista - Marathon Poland	11	1	1	4	0		2	2	0	[Icon]	
11006	21	Buena Vista - Newell Newell Providence	85	1	1	22	4		15	22	0	[Icon]	
11007	21	Buena Vista - Rembrandt Lincoln Scott	32	1	1	8	0		5	10	0	[Icon]	
11008	21	Buena Vista - Sioux Rapids Lee	30	1	1	6	0		7	5	0	[Icon]	
11009	21	Buena Vista - Storm Lake 1	30	1	1	7	1		9	5	0	[Icon]	

Find: google d | Next | Previous | Highlight all | Match case | Phrase not found

[Icons] | 4:33 PM 2/10/2012

Switch over to Stata run-through

Emacs & Stata Exercise

Return from Emacs & Stata Exercise

- Some more Stata commands about manipulating data

Input commands

insheet

- Data is output from a spreadsheet into “csv” or “comma-delimited” format
- Data is a simple $I \times J$ matrix, and all the variables are separated either by a tab or comma
- Stata is now smart enough to figure out that the first line of the file contains the variable names

insheet

Assume the following file was created by outputting a file from Excel in csv format:

HRHHID	GESTCEN	PES1	PES2	PTDTRACE
130609914027385	21	2	2	1
988009445479950	87	-1	-1	4
608038914219250	64	2	2	1
612981028909492	61	1	-1	1
9093130995291	12	1	-1	1
701099064904672	51	1	-1	1
645081921524170	54	2	1	1
506904455991733	91	1	-1	1
756490679500849	57	-2	-2	1
140809958804389	74	1	-1	1
927389210899305	56	1	-1	2
40621280985198	32	2	2	1
108903931992469	42	-1	-1	1
5570150694693	71	1	-1	1
689400611360960	93	-1	-1	1

insheet using *filename*

infile

- Data is not in Stata format, is in an ASCII file, but is *not* separated *only* by a tab or comma (e.g., by a space)

infile

Assume the following file was created using an ASCII text editor (e.g., EMACS), and that spaces separate the variables:

```
130609914027385 21 2 2 1
988009445479950 87 -1 -1 4
608038914219250 64 2 2 1
612981028909492 61 1 -1 1
9093130995291 12 1 -1 1
701099064904672 51 1 -1 1
645081921524170 54 2 1 1
506904455991733 91 1 -1 1
756490679500849 57 -2 -2 1
140809958804389 74 1 -1 1
927389210899305 56 1 -1 2
40621280985198 32 2 2 1
108903931992469 42 -1 -1 1
5570150694693 71 1 -1 1
689400611360960 93 -1 -1 1
```

infile HRHHID GESTCEN PES1 PES2 PTDTRACE using *filename*

Or

infile str HRHHID GESTCEN PES1 PES2 PTDTRACE using *filename*

infix

- Data is in an ASCII file, but you cannot rely on spaces, commas, or other standard “delimiters” to separate variables
- Datasets may have observations on more than one line

infix

Assume the following file was created using an ASCII text editor:

1															2														
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2								

1	3	0	6	0	9	9	1	4	0	2	7	3	8	5	2	1	2	2	1										
9	8	8	0	0	9	4	4	5	4	7	9	9	5	0	8	7	-	1	-	1	4								
6	0	8	0	3	8	9	1	4	2	1	9	2	5	0	6	4	2	2	1										
6	1	2	9	8	1	0	2	8	9	0	9	4	9	2	6	1	1	-	1	1									
9	0	9	3	1	3	0	9	9	5	2	9	1	1	2			1	-	1	1									

Handy label, not in dataset

Dataset

infix HRHHID 1-15 GESTCEN 16-17 PES1 18-19 PES8 20-21

PTDTRACE 22 using *filename*

Or

infile str15 HRHHID 1-15 GESTCEN 16-17 PES1 18-19 PES8 20-21

PTDTRACE 22 using *filename*



House Roll Call votes in the 38th Cong.

[illegible]

	1	2	3	4	5	6	7	8
1234567890123456789012345678901234567890123456789012345678901234567890								
01R338219801000011101ALLEN, J.C.	1121625222666611161166611691166699699999999999999							
02R338219801000011101ALLEN, J.C.	161666611616166966161116616611116199991111611111							
03R338219801000011101ALLEN, J.C.	111161616161166119116611166661116116996916199911							
04R338219801000011101ALLEN, J.C.	666161111619916966999161661666661966666116196111							
05R338219801000011101ALLEN, J.C.	666616161619666696661666611919161666699116669996							
06R338219801000011101ALLEN, J.C.	999999999999999999999999961991111119111111166161661							
07R338219801000011101ALLEN, J.C.	166661611911196166996166169999116111616611611666							
08R338219801000011101ALLEN, J.C.	611161616696696161616666661661691691111199999999							
09R338219801000011101ALLEN, J.C.	99							
10R338219801000011101ALLEN, J.C.	9616666996611699999999999999999999999999999999999999							
11R338219801000011101ALLEN, J.C.	616111166616999996166119991966116166669169611161							
12R338219801000011101ALLEN, J.C.	116116166961111911616611666161616166191666119111							
13R338219801000011101ALLEN, J.C.	969111111961166661991161							
01R338211301000013301ALLEN, W.J.	112162262266661116116991169199999199619961611169							
02R338211301000013301ALLEN, W.J.	191616619996166966161699616619116199999111999911							
03R338211301000013301ALLEN, W.J.	1111616166669661199999999999999999999999999999999999							
04R338211301000013301ALLEN, W.J.	99							
05R338211301000013301ALLEN, W.J.	9666161696166666666616669999991699611999999999966							

VAR #	0490	SESSION 2	WIDTH = 0001	MD=0	DK 10	COL 80-80	H38
	G-35-1-531A	J 38-2-170			JAN. 31, 1865		
	H382049	Y=119 N=56			ASHLEY, OHIO		
	TO PASS S.J. RES. 16. (P. 531-2)						

SEE NOTE 16

NOTE 016 S.J. RES. 16 IS A RESOLUTION SUBMITTING TO THE LEGISLATURES OF THE SEVERAL STATES, A PROPOSITION TO AMEND THE CONSTITUTION BY ADDING ARTICLE XIII PROHIBITING SLAVERY AND INVOLUNTARY SERVITUDE.

Enter data yourselves

Return again to Stata run-through

merge command

- Used when you want to add data to a pre-existing data set, or you have more than one dataset that has all the variables you need for analysis.
- Most important thing: each dataset must have (at least) one identifier that links observations, and allows merging.
- Second thing: both datasets must be sorted on the common identifier(s)*

*Stata will now take care of this, but if you know your datasets are already sorted on the common identifier, things will go faster.

Example: one-for-one match

Election results, election_results.dta

county	cand1	cand2	cand2
A	10	20	30
B	40	50	60
C	70	80	90
Z	500	40	30

Demographics, demographics.dta

county	income	educ	catholic
A	10000	.2	.3
B	40000	.5	.6
C	70000	.8	.9
Z	5000	.95	.3

merge command results

- [assume both datasets have previously been sorted on county, by typing the command `sort county`]
- use `election_results.dta`
- `merge 1:1 county using demographics.dta`

*The former merge syntax would have you type:

```
merge county using demographics.dta
```

You can still do this, but Stata will give you a warning.

Voila!

county	cand1	cand2	cand2	income	educ	catholic
A	10	20	30	10000	.2	.3
B	40	50	60	40000	.5	.6
C	70	80	90	70000	.8	.9
Z	500	40	30	5000	.95	.3

many-to-one merge

Demographic data, demographic_data.dta

county_code	town	income	education
A	Aville	50000	.3
A	Bobville	60000	.4
B	Candiceville	70000	.5
B	Dogville	80000	.5
C	Catville	100000	.5

County code mapping, county_code_mapping.dta

county_code	county_name
A	Adams
B	Brooks
C	Calhoun

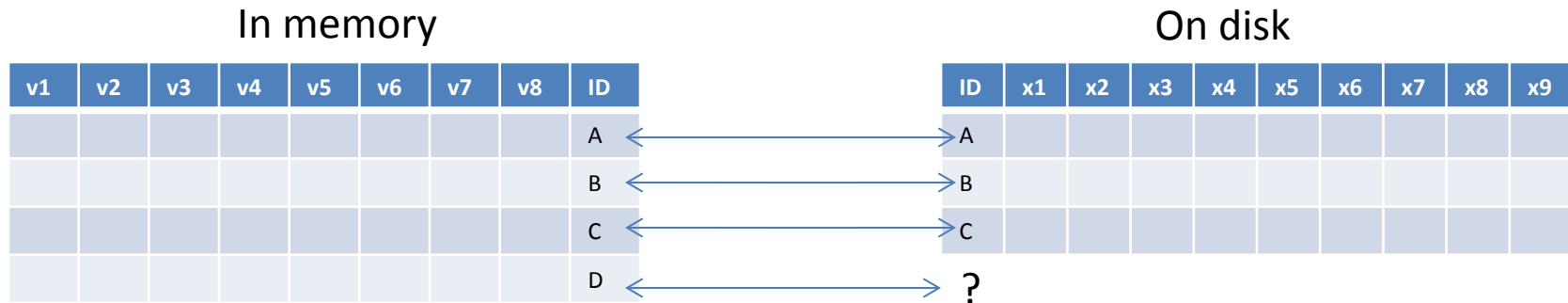
merge command

- [make same sorting assumptions as before]
- `use demographic_data.dta`
- `merge m:1 county_code using
county_code_mapping.dta`

Voila!

county_code	town	income	education	county_name
A	Aville	50000	.3	Adams
A	Bobville	60000	.4	Adams
B	Candiceville	70000	.5	Brooks
B	Dogville	80000	.5	Brooks
C	Catville	100000	.5	Calhoun

merge command summary I



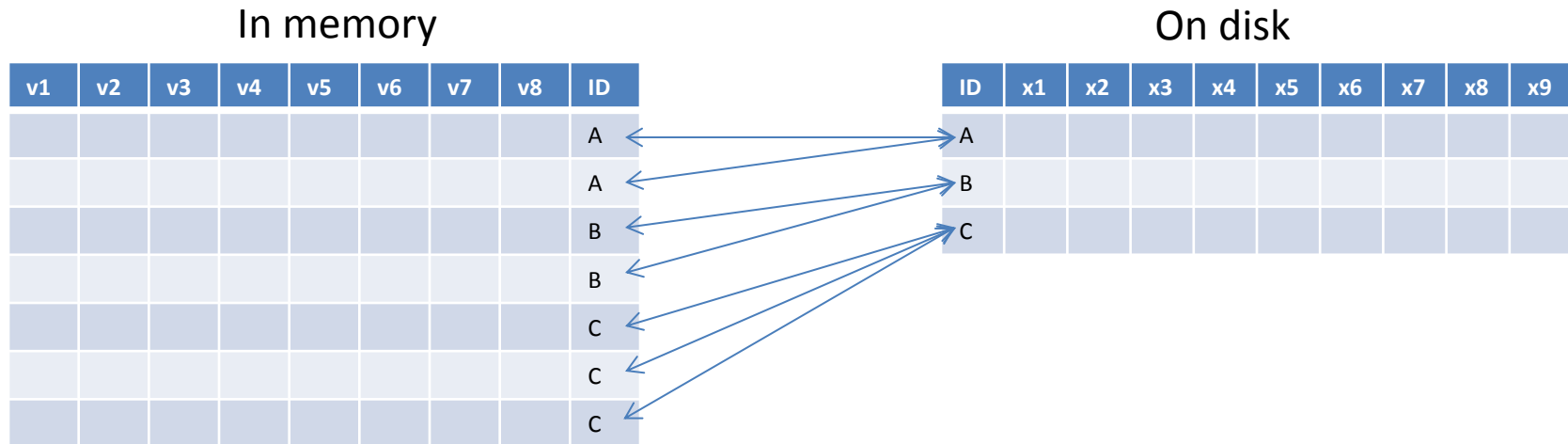
Syntax:

```
merge 1:1 ID using <on disk filename>
```

Examples

- Economic data / political data [ID = country name/code]
- Data from year t / Data from year $t+1$ [ID = unit name/code]
- Data about spouses [ID = household code]
- Presidential popularity / unemployment [ID = month]

merge command summary II



Syntax:

```
merge m:1 ID using <on disk filename>
```

Examples

- Individual economic data / Household economic data [ID = household code]
- County-level data / State-level data [ID = state]
- Student performance data / School resource data [ID = school code]

collapse command

county	DistrictName-en	voters	Paul	Bachmann	Johnson	Gingrich	Santorum	Huntsman	Other	Roemer	Romney	Perry	Cain	
Adair	Adair - 1NW ADAIR		46	7	4		11	10	0	0	0	8	6	0
Adair	Adair - 2NE STUART		51	8	5		3	15	1	0	0	6	13	0
Adair	Adair - 3SW FONTANELLE		55	9	6		16	14	0	0	0	3	7	0
Adair	Adair - 4SE ORIENT		50	4	6		6	15	0	0	0	13	6	0
Adair	Adair - 5GF GREENFIELD		67	14	5		8	12	0	0	0	13	15	0
Adams	Adams – Carbon		28	7	0		5	12	0	0	0	3	1	0
Adams	Adams - Corning 1A		19	7	0		1	6	0	0	0	4	1	0
Adams	Adams - Corning 1B		3	3	0		0	0	0	0	0	0	0	0
Adams	Adams - Corning 2A		9	2	0		2	0	0	0	0	5	0	0
Adams	Adams - Corning 2B		8	5	1		0	1	0	0	0	0	1	0
Adams	Adams - Corning 3A		12	4	0		0	0	0	0	0	6	2	0
Adams	Adams - Corning 3B		19	9	0		1	6	0	0	0	1	2	0
Adams	Adams – Nodaway		10	1	1		5	0	0	0	0	2	1	0
Adams	Adams – Prescott		32	21	3		2	1	0	0	0	3	2	0
Adams	Adams – Quincy		22	7	0		2	8	0	0	0	3	2	0
Adams	Adams - SE Adams		38	8	4		6	13	0	0	0	5	2	0
Allamakee	Allamakee - FV/TL/HF CITY		28	7	0		6	9	0	0	0	6	0	0
Allamakee	Allamakee - LF/CN/LS/LS CITY		64	20	2		21	7	0	0	0	4	10	0
Allamakee	Allamakee - PC/LT/WV CITY		42	20	0		7	9	0	0	0	5	1	0
Allamakee	Allamakee - PO/FK		20	4	1		5	3	0	0	0	6	1	0
Allamakee	Allamakee - PV CITY		35	7	1		2	3	0	0	0	21	1	0
Allamakee	Allamakee - UC/IA/NA CITY		31	4	1		3	4	0	0	0	16	3	0
Allamakee	Allamakee - UP/MK/FC/JF/LL		122	53	2		18	14	0	0	0	28	7	0
Allamakee	Allamakee - WK 1 CITY		33	8	1		6	12	0	0	0	5	1	0

collapse command

county	DistrictName-en	voters	Paul	Bachmann	Johnson	Gingrich	Santorum	Huntsman	Other	Roemer	Romney	Perry	Cain
Adair	Adair - 1NW ADAIR	46	7	4		11	10	0	0	0	8	6	0
	Adair - 2NE STUART	51	8	5		3	15	1	0	0	6	13	0
	Adair - 3SW FONTANELLE	55	9	6		16	14	0	0	0	3	7	0
	Adair - 4SE ORIENT	50	4	6		6	15	0	0	0	13	6	0
	Adair - 5GF GREENFIELD	67	14	5		8	12	0	0	0	13	15	0
Adams	Adams – Carbon	28	7	0		5	12	0	0	0	3	1	0
	Adams - Corning 1A	19	7	0		1	6	0	0	0	4	1	0
	Adams - Corning 1B	3	3	0		0	0	0	0	0	0	0	0
	Adams - Corning 2A	9	2	0		2	0	0	0	0	5	0	0
	Adams - Corning 2B	8	5	1		0	1	0	0	0	0	1	0
	Adams - Corning 3A	12	4	0		0	0	0	0	0	6	2	0
	Adams - Corning 3B	19	9	0		1	6	0	0	0	1	2	0
	Adams – Nodaway	10	1	1		5	0	0	0	0	2	1	0
	Adams – Prescott	32	21	3		2	1	0	0	0	3	2	0
	Adams – Quincy	22	7	0		2	8	0	0	0	3	2	0
	Adams - SE Adams	38	8	4		6	13	0	0	0	5	2	0
Allamakee	Allamakee - FV/TL/HF CITY	28	7	0		6	9	0	0	0	6	0	0
	Allamakee - LF/CN/LS/LS CITY	64	20	2		21	7	0	0	0	4	10	0
	Allamakee - PC/LT/WV CITY	42	20	0		7	9	0	0	0	5	1	0
	Allamakee - PO/FK	20	4	1		5	3	0	0	0	6	1	0
	Allamakee - PV CITY	35	7	1		2	3	0	0	0	21	1	0
	Allamakee - UC/IA/NA CITY	31	4	1		3	4	0	0	0	16	3	0
	Allamakee - UP/MK/FC/JF/LL	122	53	2		18	14	0	0	0	28	7	0
	Allamakee - WK 1 CITY	33	8	1		6	12	0	0	0	5	1	0

collapse (sum) voters-Cain,by(county)

county	voters	Paul	Bachmann	Johnson	Gingrich	Santorum	Huntsman	Other	Roemer	Romney	Perry	Cain
Adair	269	42	26	0	44	66	1	0	0	43	47	0
Adams	200	74	9	0	24	47	0	0	0	32	14	0
Allamakee	518	157	18	0	82	77	0	0	0	155	28	0
Appanoose	537	77	25	0	71	174	1	0	12	87	90	0
Audubon	223	41	17	0	32	54	0	0	0	48	31	0
Benton	1042	202	66	0	121	290	5	1	0	184	168	4
Black Hawk	3642	870	262	0	596	783	29	0	4	835	259	1
Boone	1344	276	104	0	160	400	4	0	0	230	170	0
Bremer	933	194	57	0	98	215	14	2	0	246	105	0
Buchanan	459	66	40	0	77	133	1	2	0	78	62	0
Buena Vista	716	169	26	0	128	154	3	0	0	124	110	2
Butler	552	99	41	0	71	157	4	0	0	92	87	0
Calhoun	435	75	31	0	54	131	2	2	0	69	71	0
Carroll	716	133	32	0	145	168	2	0	1	146	85	1
Cass	674	116	32	0	147	170	2	0	0	141	66	0
Cedar	711	188	34	0	84	167	4	1	0	165	67	0
Cerro Gordo	1571	304	100	0	235	345	5	1	0	408	170	2
Cherokee	537	95	20	0	78	155	0	0	0	126	63	0
Chickasaw	443	142	14	0	53	72	3	0	0	85	74	0
Clarke	367	98	42	0	46	51	1	2	0	65	62	0
Clay	733	150	40	0	137	165	4	2	0	149	75	0
Clayton	625	205	28	0	72	122	1	0	0	116	81	0
Clinton	1384	295	62	0	149	354	9	0	0	437	73	5
Crawford	437	72	22	0	84	101	0	0	0	93	64	0

Do **Merge** and Collapse Exercises I

- Merge together the following two datasets in /mit/17.871/Examples
 - majors2014.dta
 - majors2004.dta
- list out the data

Syntax:

```
merge 1:1 <common var> using <on disk filename>
```

Do Merge and **Collapse** Exercises II

- Merge together the following two datasets in /mit/17.871/Examples
 - majors2014.dta
 - school_course.dta
- Collapse across the “school” to find the total number of majors in each school in 2014

Syntax:

```
collapse (sum) <varname>, by(<unitname>)
```

Collapse summary

```
collapse clist [if] [in] [weight] [, options]
```

Collapse summary

```
collapse clist [if] [in] [weight] [, options]
```

`[(stat)] varlist...`

`[(stat)] target_var=varlist...`

```
collapse (sum) majors2014, by(school)
```

```
collapse (sum) majors2014 (max) max2014=majors2014, by(school)
```

Collapse summary

```
collapse clist [if] [in] [weight] [, options]
```



```
by(varlist)
```



```
collapse (sum) majors2014, by(school)
```

Do-files (Ch. 2 of K&K)

- Do-files are the Stata scripting language to automate analysis.
- Here is how the first five lines of the Iowa exercise would look in a do-file:

```
#delimit;  
insheet using iowa_example_csv.dat;  
list;  
generate paulpct08=paul08/tvotes08;  
generate paulpct12=paul12/tvotes12;
```

Log-files (pp. 33-34 of K&K)

- Log files capture both commands and results (but not graphs) in a file for later use.
- Different extensions produce different files
 - .smcl : preserves colors and fonts
 - .log: plain text
- Please save log files using the .log extension:
 - `log using filename.log`
 - `log close`

```
/*This file does the simple Iowa example,  
saving a log file in smcl format*/
```

```
#delimit;
```

```
set more off;
```

```
clear;
```

```
log using iowa_example.smcl,replace;
```

```
insheet using iowa_example_csv.dat;  
list;
```

```
generate paulpct08=paul08/tvotes08;
```

```
generate paulpct12=paul12/tvotes12;
```

```
set more on;
```

```
capture noisily log close;
```

```
/*This file does the simple Iowa example,  
saving a log file in log format*/
```

```
#delimit;
```

```
set more off;
```

```
clear;
```

```
log using iowa_example.log,replace;
```

```
insheet using iowa_example_csv.dat;  
list;
```

```
generate paulpct08=paul08/tvotes08;
```

```
generate paulpct12=paul12/tvotes12;
```

```
set more on;
```

```
capture noisily log close;
```

Three tricks for building do-files

- Select commands from command window
- Use `cmdlog` command to capture commands for inclusion in do files
- You can cut-and-paste from results window (really useful for capturing graphing commands)

Some final points about Stata, from my observing people using it

- Do not use variable or file names with embedded spaces
- Remember that Stata is CaSe-sENsitive, for commands and variable names, but NOT file names
 - a useful command: `rename *, lower`
- Avoid pointing and clicking in Stata (graphing may be an exception).
- Unix reminder on next slide

- Remember that the whole `/afs/athena.mit.edu/...` business can usually be reduced to:
- `/mit/...`
- So, for instance
- **`/afs/athena.mit.edu/c/s/cstewart`** reduces to
- **`/mit/cstewart`**

On to the do-file/log-file exercise

- Write a do file that
 - merges the three files: majors2004.dta, majors2014.dta, and school_course.dta
 - collapses across schools to find the total number of majors in both 2004 and 2014
 - Captures everything in a log file called majors0414.log