Ecog-314 Project guide – part of lecture #4

Three types of dataset

- **Cross-section data:** data on one or more variables collected ***at the same point in time*** (i.e., **multiple subjects** or individuals at the **same time**).

Structure

A: x= Profit (year 2015, billions $), i = {Apple, Microsoft, GE, IBM, etc. }

xi : x1 (AAPL): 40
x2 (MSFT): 50
x3 (GE): 75
x4 (IBM): 100

B: x = GDP (year 2016, billions $) , I = { US, China, Japan, Germany, etc}

xi : x1(US): 18,558.130
x2 (China): 11,383.030
x3 (Japan): 4,412.600
x4 (Germany): 3,467.780

C:

Rank	Country/Economy	GDP Nominal (billions of $) 2016	2017	2018	2019	2020
1	United States	18,558.130	19,285	20,145	21,016	21,874
2	China	11,383.030	12,263	13,338	14,605	16,144
3	Japan	4,412.600	4,514	4,562	4,676	4,800
4	Germany	3,467.780	3,592	3,697	3,822	3,959
5	United Kingdom	2,760.960	2,885	2,999	3,123	3,256
6	France	2,464.790	2,538	2,609	2,700	2,804
7	India	2,288.720	2,488	2,725	3,007	3,315
8	Italy	1,848.690	1,902	1,943	1,994	2,051
9	Brazil	1,534.780	1,556	1,609	1,677	1,749
10	Canada	1,462.330	1,531	1,596	1,667	1,740
11	Korea	1,321.200	1,379	1,435	1,499	1,566
12	Spain	1,242.360	1,291	1,332	1,380	1,433
13	Australia	1,200.780	1,262	1,330	1,399	1,469
14	Russia	1,132.740	1,268	1,355	1,447	1,531
15	Mexico	1,082.430	1,167	1,228	1,300	1,381

Cross sectional data for 2016

Source: http://statisticstimes.com/economy/countries-by-projected-gdp.php

D:

TABLE 1.1 U.S. EGG PRODUCTION

	State	Y_1	Y_2	X_1	X_2	State	Y_1	Y_2	X_1	X_2
	AL	2,206	2,186	92.7	91.4	MT	172	164	68.0	66.0
Alaska	AK	0.7	0.7	151.0	149.0	NE	1,202	1,400	50.3	48.9
	AZ	73	74	61.0	56.0	NV	2.2	1.8	53.9	52.7
	AR	3,620	3,737	86.3	91.8	NH	43	49	109.0	104.0
California	CA	7,472	7,444	63.4	58.4	NJ	442	491	85.0	83.0
	CO	788	873	77.8	73.0	NM	283	302	74.0	70.0
	CT	1,029	948	106.0	104.0	NY	975	987	68.1	64.0
	DE	168	164	117.0	113.0	NC	3,033	3,045	82.8	78.7
	FL	2,586	2,537	62.0	57.2	ND	51	45	55.2	48.0
	GA	4,302	4,301	80.6	80.8	OH	4,667	4,637	59.1	54.7
	HI	227.5	224.5	85.0	85.5	OK	869	830	101.0	100.0
	ID	187	203	79.1	72.9	OR	652	686	77.0	74.6
	IL	793	809	65.0	70.5	PA	4,976	5,130	61.0	52.0
	IN	5,445	5,290	62.7	60.1	RI	53	50	102.0	99.0
	IA	2,151	2,247	56.5	53.0	SC	1,422	1,420	70.1	65.9
	KS	404	389	54.5	47.8	SD	435	602	48.0	45.8
	KY	412	483	67.7	73.5	TN	277	279	71.0	80.7
	LA	273	254	115.0	115.0	TX	3,317	3,356	76.7	72.6
	ME	1,069	1,070	101.0	97.0	UT	456	486	64.0	59.0
	MD	885	898	76.6	75.4	VT	31	30	106.0	102.0
	MA	235	237	105.0	102.0	VA	943	988	86.3	81.2
	MI	1,406	1,396	58.0	53.8	WA	1,287	1,313	74.1	71.5
	MN	2,499	2,697	57.7	54.0	WV	136	174	104.0	109.0
	MS	1,434	1,468	87.8	86.7	WI	910	873	60.1	54.0
	MO	1,580	1,622	55.4	51.5	WY	1.7	1.7	83.0	83.0

Note: Y_1 = eggs produced in 1990 (millions)
Y_2 = eggs produced in 1991 (millions)
X_1 = price per dozen (cents) in 1990
X_2 = price per dozen (cents) in 1991
Source: World Almanac, 1993, p. 119. The data are from the Economic Research Service, U.S. Department of Agriculture.

⇨ For **each year** the data on the 50 states are cross-sectional data.

R-Code:

```
#US Egg production R

state_egg_production_1990 <- read.table(header = TRUE, text = "
State Number_of_eggs_produced Price
AL 2206 92.7
AK 0.7 151.0
AZ 73 61.0
```

Break

```
WA 1287 74.1
WV 136 104.0
WI 910 60.1
WY 1.7 83.0
")
```

```
> # Verify your data
> dim(state_egg_production_1990)
```

R code

```
[1] 50  3
```

Output

```
> head(state_egg_production_1990)
```

R code

```
  State Number_of_eggs_produced Price
1    AL                  2206.0  92.7
2    AK                     0.7 151.0
3    AZ                    73.0  61.0
4    AR                  3620.0  86.3
5    CA                  7471.0  63.4
6    CO                   788.0  77.8
```

Output

```
> # Descriptive statistics
> summary(state_egg_production_1990)
```

R code

```
     State     Number_of_eggs_produced      Price
 AK     : 1   Min.   :   0.7          Min.   : 48.00
 AL     : 1   1st Qu.: 229.4          1st Qu.: 61.25
 AR     : 1   Median : 818.0          Median : 75.35
 AZ     : 1   Mean   :1355.6          Mean   : 78.29
 CA     : 1   3rd Qu.:1543.5          3rd Qu.: 87.42
 CO     : 1   Max.   :7472.0          Max.   :151.00
 (Other):44
```

Output

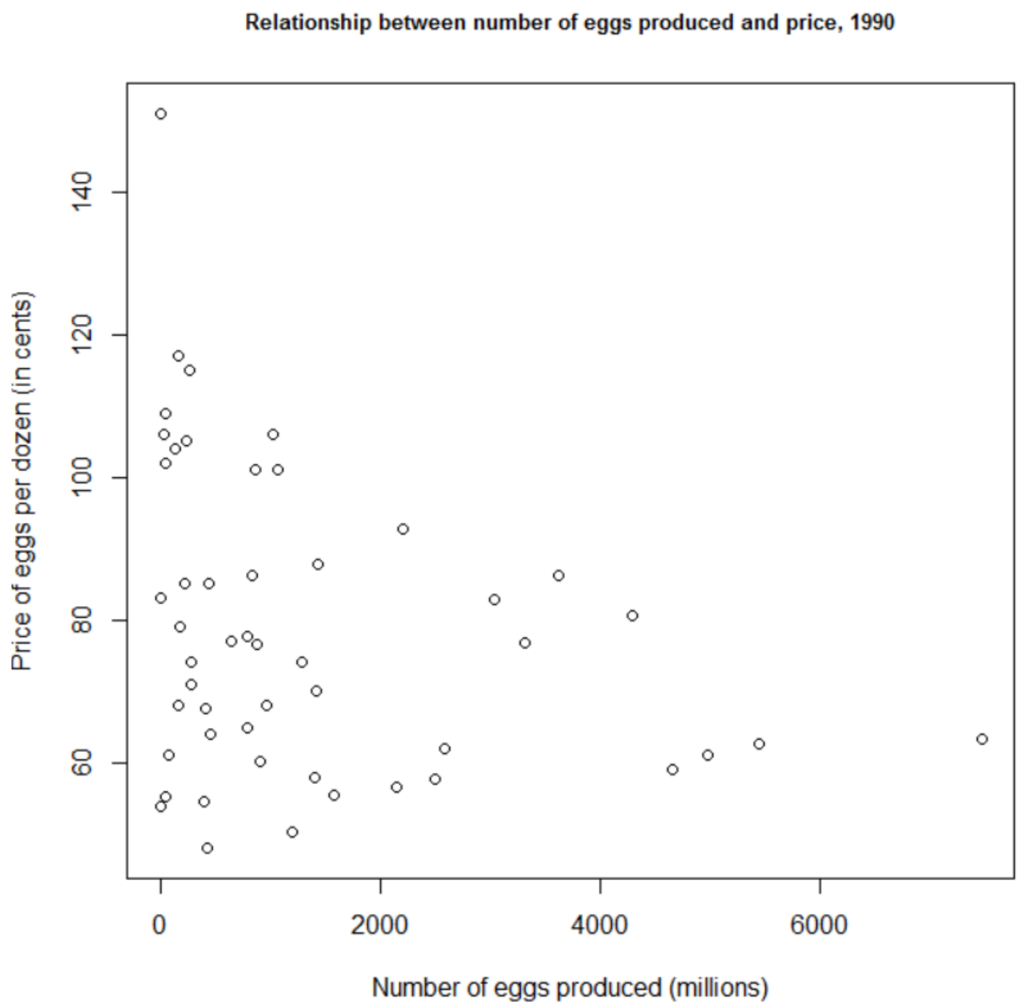
```
#Get documentation on the plot function
?plot

# get your x and y variables
x = state_egg_production_1990$Number_of_eggs_produced
y = state_egg_production_1990$Price

#Open-up a separate plotting window
windows()     #alternatively you can use dev.new()

#Plot the relationship between Quantity of eggs produced and the price
plot(x, y, main="Relationship between number of eggs produced and price, 1990",
     xlab = "Number of eggs produced (millions)",
     ylab = "Price of eggs per dozen (in cents)",
     cex.main=0.8 )
```

change the font size of the title.



- **Time series data**: A time series is a set of observations on the values that a variable takes at different times. It is collected at **regular time intervals**, such as daily, weekly, monthly quarterly, etc.

Structure and example

Example 1: Sales time series

Period (t)	Year	Quarter	Sales1	Sales2
1	2005	1	10	10
2	2005	2	12	15
3	2005	3	9	18
4	2005	4	13	14
5	2006	1	10	13
6	2006	2	11	17
7	2006	3	15	22
8	2006	4	8	19

Example 2: Financial Accounts time series

```
* $open fof
$open FOF already open as /fame/data/database/fof/fof.db

* repo <dec 1: date 2015q1 to 2016q2> FL152000005.Q as "Assets", FL154190005.Q as "Liabilities", FL152090005.Q as "Net worth"
```

Time	Assets	Liabilities	Net worth
2015q1	99977718.0	14152363.0	85825355.0
2015q2	100696372.0	14293360.0	86403012.0
2015q3	99461570.0	14372802.0	85088769.0
2015q4	101679772.0	14509524.0	87170248.0
2016q1	102512039.0	14524319.0	87987720.0
2016q2	103750285.0	14687565.0	89062720.0

www.federalreserve.gov/releases/z1/Current/z1.pdf

Page: 152 of 198 Automatic Zoom

B.101 Balance Sheet of Households and Nonprofit Organizations (1)

Billions of dollars; amounts outstanding end of period, not seasonally adjusted

			2013	2014	2015	2015 Q1	2015 Q2	2015 Q3	2015 Q4	2016 Q1	2016 Q2	
1	**FL152000005**	**Assets**	**92701.4**	**98111.0**	**101679.8**	**99977.7**	**100696.4**	**99461.6**	**101679.8**	**102512.0**	**103750.3**	**1**
2	FL152010005	Nonfinancial assets	27228.0	28701.2	30462.6	29190.2	29597.3	30008.4	30462.6	30896.7	31419.8	2
3	LM155035005	Real estate	21849.2	23195.1	24756.0	23639.2	23979.5	24350.5	24756.0	25120.8	25594.7	3
4	LM155035015	Households (2,3)	19194.5	20273.2	21532.6	20614.9	20902.2	21195.0	21532.6	21882.0	22290.0	4
5	LM165035005	Nonprofit organizations	2654.7	2921.9	3223.4	3024.3	3077.2	3155.5	3223.4	3238.8	3304.7	5
6	FL165015205	Equipment (nonprofits) (4)	311.6	320.4	331.0	323.9	326.2	328.7	331.0	333.0	334.6	6
7	FL165013765	Intellectual property products (nonprofits) (4)	126.0	132.8	138.9	134.9	136.7	138.1	138.9	140.9	143.2	7
8	FL155111005	Consumer durable goods (4)	4941.2	5052.9	5236.8	5092.2	5154.9	5191.1	5236.8	5301.9	5347.4	8
30	**FL154190005**	**Liabilities**	**13792.6**	**14167.0**	**14509.5**	**14152.4**	**14293.4**	**14372.8**	**14509.5**	**14524.3**	**14687.6**	**30**
31	FL163162003	Debt securities (municipal securities) (10)	235.6	228.8	220.8	228.4	224.7	222.1	220.8	220.9	220.8	31
32	FL154123005	Loans	13274.1	13651.0	13998.8	13635.2	13778.7	13860.9	13998.8	14012.1	14174.7	32
33	FL153165105	Home mortgages (9)	9403.9	9397.1	9479.6	9368.8	9409.3	9456.7	9479.6	9497.4	9553.2	33
34	FL153166000	Consumer credit	3096.2	3318.0	3535.7	3322.8	3397.8	3481.4	3535.7	3539.4	3605.3	34
35	FL153168005	Depository institution loans n.e.c.	90.8	211.9	325.7	232.5	246.3	268.7	325.7	339.3	366.5	35
36	FL153169005	Other loans and advances	480.5	513.7	437.2	499.0	510.1	436.5	437.2	413.5	424.5	36
37	FL163165505	Commercial mortgages (10)	202.7	210.3	220.7	212.2	215.1	217.7	220.7	222.6	225.3	37
38	FL163170003	Trade payables (10)	255.0	258.1	259.4	258.5	258.9	259.2	259.4	259.4	259.4	38
39	FL543077073	Deferred and unpaid life insurance premiums	27.9	29.1	30.6	30.2	31.0	30.6	30.6	31.8	32.7	39
40	**FL152090005**	**Net worth**	**78908.9**	**83944.0**	**87170.2**	**85825.4**	**86403.0**	**85088.8**	**87170.2**	**87987.7**	**89062.7**	**40**

RCode

csv file name

```r
# Time series data data

#help on how to read a csv file
?read.table

#read the time series data file
b_101_table = read.table(file="b101.csv", header = TRUE, sep=",", stringsAsFactors=FALSE)


#Verify your data
dim(b_101_table)
```

```
> dim(b_101_table)
[1] 283  53
```

R output

```r
#show a section of the file
head(b_101_table[, c(1:5)], n=10)
```

```
> head(b_101_table[, c(1:5)], n=10)
      date FL152000005.Q FL152010005.Q LM155035005.Q LM155035015.Q
1  1945:Q4     832955.93     189327.00     134635.00     116049.00
2  1946:Q1            ND            ND            ND            ND
3  1946:Q2            ND            ND            ND            ND
4  1946:Q3            ND            ND            ND            ND
5  1946:Q4     917781.00     220716.00     158074.00     133422.00
6  1947:Q1            ND            ND            ND            ND
7  1947:Q2            ND            ND            ND            ND
8  1947:Q3            ND            ND            ND            ND
9  1947:Q4    1024628.00     280815.00     206381.00     177473.00
10 1948:Q1            ND            ND            ND            ND
> #What do we have
```

R output

```r
#What do we have
class(b_101_table)
```

```
> class(b_101_table)
[1] "data.frame"
```

R output

Pooled data: In pooled, or combined, data are elements of ***both time series and cross-section dat***a. The data in Table 1.1 are an example of pooled data. For each year we have 50 cross-sectional observations and for each state we have two-time series observations on prices and output of eggs, a total of 100 *pooled* (or combined) observations.

TABLE 1.1 U.S. EGG PRODUCTION

State	Y_1	Y_2	X_1	X_2	State	Y_1	Y_2	X_1	X_2
AL	2,206	2,186	92.7	91.4	MT	172	164	68.0	66.0
Alaska AK	0.7	0.7	151.0	149.0	NE	1,202	1,400	50.3	48.9
AZ	73	74	61.0	56.0	NV	2.2	1.8	53.9	52.7
AR	3,620	3,737	86.3	91.8	NH	43	49	109.0	104.0
California CA	7,472	7,444	63.4	58.4	NJ	442	491	85.0	83.0
CO	788	873	77.8	73.0	NM	283	302	74.0	70.0
CT	1,029	948	106.0	104.0	NY	975	987	68.1	64.0
DE	168	164	117.0	113.0	NC	3,033	3,045	82.8	78.7
FL	2,586	2,537	62.0	57.2	ND	51	45	55.2	48.0
GA	4,302	4,301	80.6	80.8	OH	4,667	4,637	59.1	54.7
HI	227.5	224.5	85.0	85.5	OK	869	830	101.0	100.0
ID	187	203	79.1	72.9	OR	652	686	77.0	74.6
IL	793	809	65.0	70.5	PA	4,976	5,130	61.0	52.0
IN	5,445	5,290	62.7	60.1	RI	53	50	102.0	99.0
IA	2,151	2,247	56.5	53.0	SC	1,422	1,420	70.1	65.9
KS	404	389	54.5	47.8	SD	435	602	48.0	45.8
KY	412	483	67.7	73.5	TN	277	279	71.0	80.7
LA	273	254	115.0	115.0	TX	3,317	3,356	76.7	72.6
ME	1,069	1,070	101.0	97.0	UT	456	486	64.0	59.0
MD	885	898	76.6	75.4	VT	31	30	106.0	102.0
MA	235	237	105.0	102.0	VA	943	988	86.3	81.2
MI	1,406	1,396	58.0	53.8	WA	1,287	1,313	74.1	71.5
MN	2,499	2,697	57.7	54.0	WV	136	174	104.0	109.0
MS	1,434	1,468	87.8	86.7	WI	910	873	60.1	54.0
MO	1,580	1,622	55.4	51.5	WY	1.7	1.7	83.0	83.0

Note: Y_1 = eggs produced in 1990 (millions)
Y_2 = eggs produced in 1991 (millions)
X_1 = price per dozen (cents) in 1990
X_2 = price per dozen (cents) in 1991

Source: World Almanac, 1993, p. 119. The data are from the Economic Research Service, U.S. Department of Agriculture.

- **Panel, Longitudinal, or Micropanel Data**: This is a ***special type*** of pooled data in which the ***same cross-sectional unit*** *(say, a family or a firm)* is surveyed over time.

Another definition: In pooled, or combined, data are elements of ***both time series and cross-section*** *dat*a. panel (longitudinal)

- multiple subjects (individuals)
- at different times, you have the same subject at different times, and you have many subjects at the same time; think of it as a table where rows are time points, and columns are subjects.

For example, part of a longitudinal dataset could contain specific students and their standardized test scores in six successive years.

Student Name	**Grade 1 (2001) Raw Score**	**Grade 2 (2002) Raw Score**	**Grade 3 (2003) Raw Score**	**Grade 4 (2004) Raw Score**	**Grade 5 (2005) Raw Score**	**Grade 6 (2006) Raw Score**
Mike	339	350	361	366	381	390
Jasmine	332	343	350	351	351	355
Thomas	360	380	400	420	430	438

The primary advantage of longitudinal databases is that they can measure *change*. So we can estimate, for example, the effect of various factors on *improvement* in student achievement. We can also estimate the overall effectiveness of individual teachers by examining the performance of successive classes of students they teach, as well as examine the extent to which teacher effectiveness changes with experience or the composition of their class.

Source: http://www.caldercenter.org/what-are-longitudinal-data

Motivations for Multilevel Models

Consider the following the data from Agresti (1996). Researchers were interested in possible bias in death penalty cases based on the defendant's race. Here is a simple table examining defendant's race and whether they were given the death penalty:

Defendant's race	Death Penalty		Percent Yes
	Yes	No	
White	53	430	11.0
Black	15	176	7.9

Based on these data, we would conclude – if anything – that there is a slight bias against White defendants. However, there was also data on the victim's race as well.

Another view:

Let's look at what the data look like when we disaggregate the data by victim's race:

Victim	Defendant	Death Penalty		Percent Yes
White		Yes	No	
	White	53	414	11.3
	Black	11	37	22.9
Black		Yes	No	
	White	0	16	0.0
	Black	4	139	2.8

Now things look quite a bit different.

Once we take into account the victim's race, a greater percentage of Black defendant's are given the death penalty – regardless of victim's race!

Source: https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf

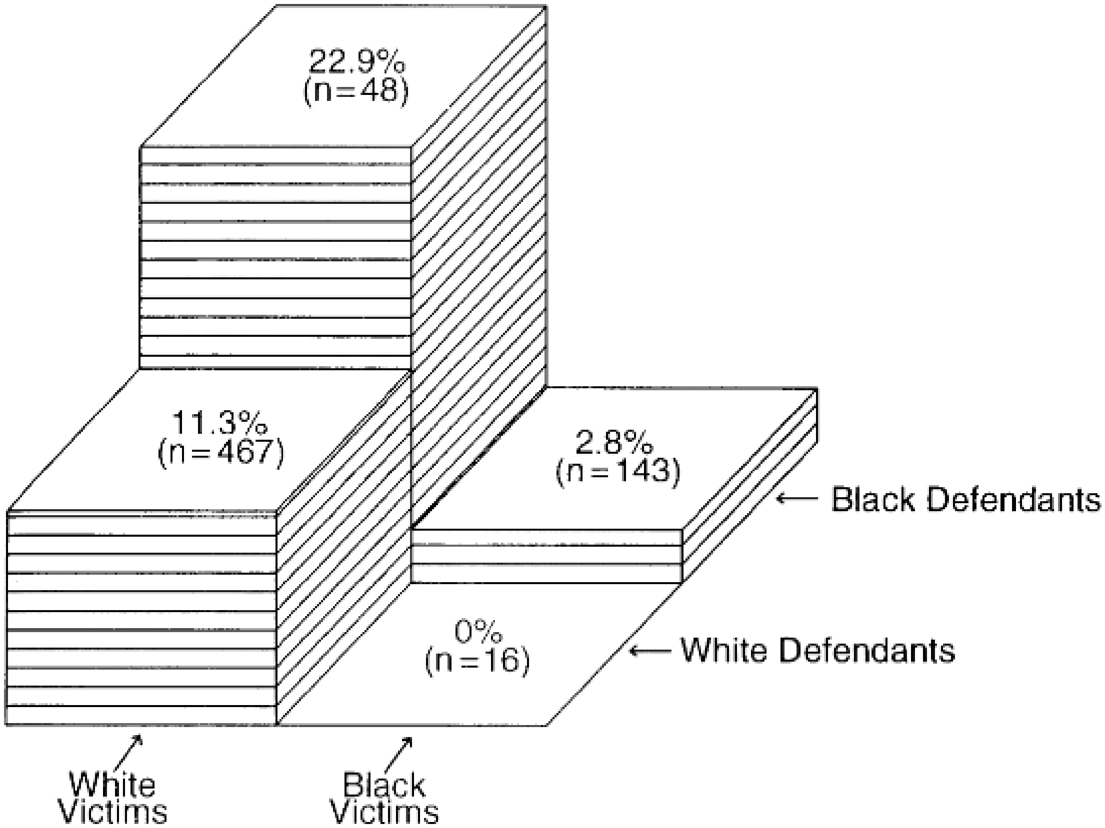


Figure 2.3. Percentage receiving death penalty, by defendant's race and victims' race.

Working with Panel dataset

Example1:

Each year, beginning at age 14, 82 teenagers completed a 4-item questionnaire assessing their alcohol consumption during the previous year. Using a 8-point scale (0 = "not al all, 8 = "every day") teenagers described the frequency with which they
(1) drank beer or wine,
(2) drank hard liquor,
(3) had five or more drinks in a row, and
(4) got drunk.

Two potential predictors of alcohol use are whether the teenager is a child of an alcoholic parent; and alcohol use among the teenager's peers. The teenager used a 6-point scale to estimate the proportion of their friends who drank alcohol occasionally (item 1) or regularly (item 2). This was obtained during the first wave of data collection.

Source: Currant, P. et al. (1997). Reported in Singer, J., & Willet (2003). Applied Longitudinal Data Analysis. p. 76-77.

--

The dataset

http://www.ats.ucla.edu/stat/r/examples/alda/data/alcohol1_pp.txt

sid	coa	sex	age14	alcuse	peer	cpeer	ccoa
1	1	0	0	1.73205	1.26491	0.24691	0.549
1	1	0	1	2.00000	1.26491	0.24691	0.549
1	1	0	2	2.00000	1.26491	0.24691	0.549
2	1	1	0	0.00000	0.89443	-0.12357	0.549
2	1	1	1	0.00000	0.89443	-0.12357	0.549
2	1	1	2	1.00000	0.89443	-0.12357	0.549
3	1	1	0	1.00000	0.89443	-0.12357	0.549
3	1	1	1	2.00000	0.89443	-0.12357	0.549
3	1	1	2	3.31662	0.89443	-0.12357	0.549
4	1	1	0	0.00000	1.78885	0.77085	0.549
4	1	1	1	2.00000	1.78885	0.77085	0.549
4	1	1	2	1.73205	1.78885	0.77085	0.549
5	1	0	0	0.00000	0.89443	-0.12357	0.549
5	1	0	1	0.00000	0.89443	-0.12357	0.549
5	1	0	2	0.00000	0.89443	-0.12357	0.549

Data

Column 1: Teenager ID
Column 2: Whether the teenager is a child of a alcohlic parent
Column 3: Sex (male = 1, female = 0)
Column 4: Number of year since age 14
Column 5: Alcohol use of the teenager (sqrt-root of mean of 6 items)
Column 6: Alcohol use among the teenager's peers (sqrt-root of mean of 2 items)
Column 7: Alcoholic parenet variable centered
Column 8: Peer variable centered

Another presentation of the dataset

www.ats.ucla.edu/stat/r/examples/alda/data/alcohol1_pp.txt

```
id,age,coa,male,age_14, alcuse,peer,cpeer,ccoa
1,14,1,0,0,1.7320507764816284,1.2649110555648804,.2469110555648036,.5489999999999999
1,15,1,0,1,2,1.2649110555648804,.2469110555648036,.5489999999999999
1,16,1,0,2,2,1.2649110555648804,.2469110555648036,.5489999999999999
2,14,1,1,0,0,.8944271802902222,-.1235728197097785,.5489999999999999
2,15,1,1,1,0,.8944271802902222,-.1235728197097785,.5489999999999999
2,16,1,1,2,1,.8944271802902222,-.1235728197097785,.5489999999999999
3,14,1,1,0,1,.8944271802902222,-.1235728197097785,.5489999999999999
```

Another example with R code

http://www.ats.ucla.edu/stat/r/examples/alda/data/tolerance1.txt

www.ats.ucla.edu/stat/r/examples/alda/data/tolerance1.txt

```
id,tol11,tol12,tol13,tol14,tol15,male,exposure
9,2.23,1.79,1.9000000000000001,2.12,2.66,0,1.54
45,1.12,1.45,1.45,1.45,1.99,1,1.16
268,1.45,1.34,1.99,1.79,1.34,1,.9
314,1.22,1.22,1.55,1.12,1.12,0,.81
442,1.45,1.99,1.45,1.67,1.9000000000000001,0,1.1300000000000001
514,1.34,1.67,2.23,2.12,2.44,1,.9
569,1.79,1.9000000000000001,1.9000000000000001,1.99,1.99,0,1.99
624,1.12,1.12,1.22,1.12,1.22,1,.98
723,1.22,1.34,1.12,1,1.12,0,.81
918,1,1,1.22,1.99,1.22,0,1.21
949,1.99,1.55,1.12,1.45,1.55,1,.93
978,1.22,1.34,2.12,3.46,3.3200000000000003,1,1.59
1105,1.34,1.9000000000000001,1.99,1.9000000000000001,2.12,1,1.3800000000000001
1542,1.22,1.22,1.99,1.79,2.12,0,1.44
1552,1,1.12,2.23,1.55,1.55,0,1.04
1653,1.11,1.11,1.34,1.55,2.12,0,1.25
```

R code to read data

```
> tolerance <- read.csv("http://www.ats.ucla.edu/stat/r/examples/alda/data/tolerance1.txt")
>
> dim(tolerance)
[1] 16  8
>
> head(tolerance, n=10)
    id tol11 tol12 tol13 tol14 tol15 male exposure
1    9  2.23  1.79  1.90  2.12  2.66    0     1.54
2   45  1.12  1.45  1.45  1.45  1.99    1     1.16
3  268  1.45  1.34  1.99  1.79  1.34    1     0.90
4  314  1.22  1.22  1.55  1.12  1.12    0     0.81
5  442  1.45  1.99  1.45  1.67  1.90    0     1.13
6  514  1.34  1.67  2.23  2.12  2.44    1     0.90
7  569  1.79  1.90  1.90  1.99  1.99    0     1.99
8  624  1.12  1.12  1.22  1.12  1.22    1     0.98
9  723  1.22  1.34  1.12  1.00  1.12    0     0.81
10 918  1.00  1.00  1.22  1.99  1.22    0     1.21
>
> summary(tolerance)
       id             tol11           tol12           tol13           tol14           tol15            male            exposure
 Min.   :   9.0   Min.   :1.000   Min.   :1.000   Min.   :1.120   Min.   :1.000   Min.   :1.120   Min.   :0.0000   Min.   :0.8100
 1st Qu.: 410.0   1st Qu.:1.120   1st Qu.:1.195   1st Qu.:1.310   1st Qu.:1.450   1st Qu.:1.310   1st Qu.:0.0000   1st Qu.:0.9225
 Median : 673.5   Median :1.220   Median :1.340   Median :1.725   Median :1.730   Median :1.945   Median :0.0000   Median :1.1450
 Mean   : 762.8   Mean   :1.364   Mean   :1.441   Mean   :1.676   Mean   :1.754   Mean   :1.861   Mean   :0.4375   Mean   :1.1912
 3rd Qu.:1009.8   3rd Qu.:1.450   3rd Qu.:1.700   3rd Qu.:1.990   3rd Qu.:1.990   3rd Qu.:2.120   3rd Qu.:1.0000   3rd Qu.:1.3950
 Max.   :1653.0   Max.   :2.230   Max.   :1.990   Max.   :2.230   Max.   :3.460   Max.   :3.320   Max.   :1.0000   Max.   :1.9900
```

Read in the data

check your data

R output

Create a summary

R output

See worked example on the Wine dataset from UCI data repository:

https://github.com/wampeh1/ECOG_314/blob/master/project1/lecture3_project_guide_multivariate_data_analysis_example.rmd

https://github.com/wampeh1/ECOG_314/blob/master/project1/pdf/lecture3_project_guide_multivariate_data_analysis_example.pdf

Data file:

http://archive.ics.uci.edu/ml/datasets/Wine

archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

```
1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
1,14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450
1,14.39,1.87,2.45,14.6,96,2.5,2.52,.3,1.98,5.25,1.02,3.58,1290
1,14.06,2.15,2.61,17.6,121,2.6,2.51,.31,1.25,5.05,1.06,3.58,1295
1,14.83,1.64,2.17,14,97,2.8,2.98,.29,1.98,5.2,1.08,2.85,1045
1,13.86,1.35,2.27,16,98,2.98,3.15,.22,1.85,7.22,1.01,3.55,1045
1,14.1,2.16,2.3,18,105,2.95,3.32,.22,2.38,5.75,1.25,3.17,1510
1,14.12,1.48,2.32,16.8,95,2.2,2.43,.26,1.57,5,1.17,2.82,1280
1,13.75,1.73,2.41,16,89,2.6,2.76,.29,1.81,5.6,1.15,2.9,1320
1,14.75,1.73,2.39,11.4,91,3.1,3.69,.43,2.81,5.4,1.25,2.73,1150
1,14.38,1.87,2.38,12,102,3.3,3.64,.29,2.96,7.5,1.2,3,1547
```

Data Set Information:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a list of what the 30 or so variables were, but a.) I lost it, and b.), I would not know which 13 variables are included in the set.

The attributes are (dontated by Riccardo Leardi, riclea '@' anchem.unige.it)
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280/OD315 of diluted wines
13)Proline