

Lecture 5

1. Data checking: correct formats, unusual values
2. Scatterplot matrix
3. Editing data in the data step
4. Simple scatter plot: Plot, Insight, SGplot, Gplot

1

Data Checking

1. *Check that SAS read each variable as the correct type:*
 - Numeric data: NUM
 - Character data: CHAR
 - Date-time data: converted to numeric
2. *Check that SAS read the correct number of observations.*

Proc Contents answers both these questions.

```
Proc Contents data = PH6470.child_iq;
```

2

The CONTENTS Procedure

Data Set Name	PH6470.CHILD_IQ	Observations	434
Member Type	DATA	Variables	6
Engine	V9	Indexes	0
Created	Thu, Sep 10, 2009 11:55:16 AM	Observation Length	48

Filename	C:\Documents and Settings\Administrator\Desktop\SAS Class\child_iq.sas7bdat
Release Created	9.0201M0
Host Created	XP_PRO

[other stuff]

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
1	ID	Num	8	ID
2	child_IQ	Num	8	child IQ
6	male	Num	8	male
3	mom_HS_grad	Num	8	mom HS grad
5	mom_IQ	Num	8	mom IQ
4	mom_age	Num	8	mom age

3

```

Data N;
  input ID gender $ birthdate MMDDYY10. ;
  format birthdate MMDDYY10.;
  cards;
  4833 F 5/16/1978
  4834 F 7/4/1980
  4855 M 12/14/1988
  ;
Proc Contents data=N;

```

The CONTENTS Procedure

Data Set Name	WORK.N	Observations	3
Member Type	DATA	Variables	3

.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format
1	ID	Num	8	
3	birthdate	Num	8	MMDDYY10.
2	gender	Char	8	

4

3. What is the pattern of missing data?

```
Proc Means nmiss n data=pubh.0GTT_hw2;
```

NMISS = count of missing values, N = count of non-missing

The MEANS Procedure

Variable	Miss	N
min000	0	1019
min030	3	1016
min060	2	1017
min090	4	1015
min120	0	1019
id	0	1019

5

4. Find unusual observations—are there outliers or incorrect values?

Use Insight for quick scatterplots, Proc Univariate to identify extreme observations, Proc Freq for list of distinct values

5. Should some variables be transformed?

For positive variables, when $\frac{\text{maximum value}}{\text{minimum value}} > 10$, take logs.

6

Scatterplot matrix can help with both questions. Three ways to do this:

```
Proc SGscatter data=ph6470.child_iq; SG = statistical graphic  
  matrix child_iq mom_iq mom_age / group = male;
```

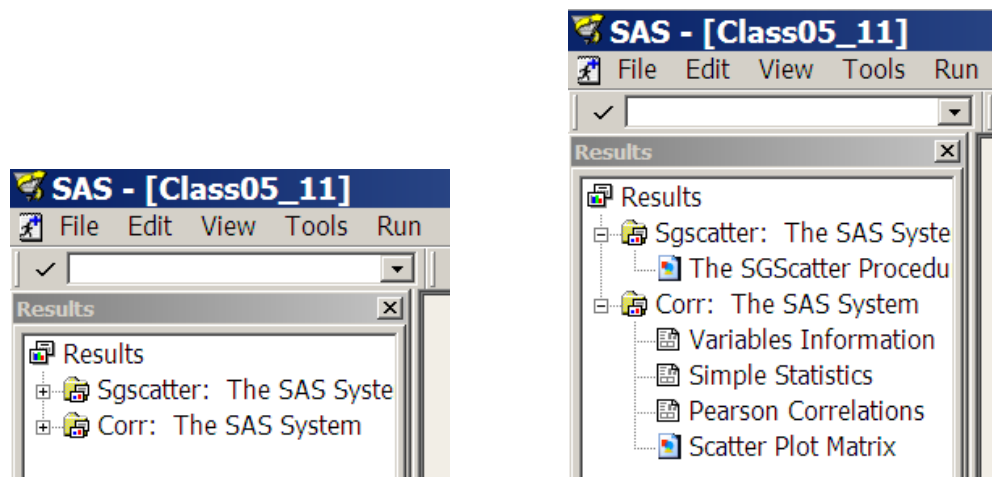
```
Proc Insight data=ph6470.child_iq; interactive, but no grouping variable  
  scatter child_iq mom_iq mom_age * child_iq mom_iq mom_age ;  
Insight will not be included in later versions of SAS
```

ODS graphics on;

```
Proc Corr data=ph6470.child_iq plots=matrix ; no grouping variable  
  var child_iq mom_iq mom_age;  
run; run the procedure before turning off ODS graphics  
ODS graphics off;
```

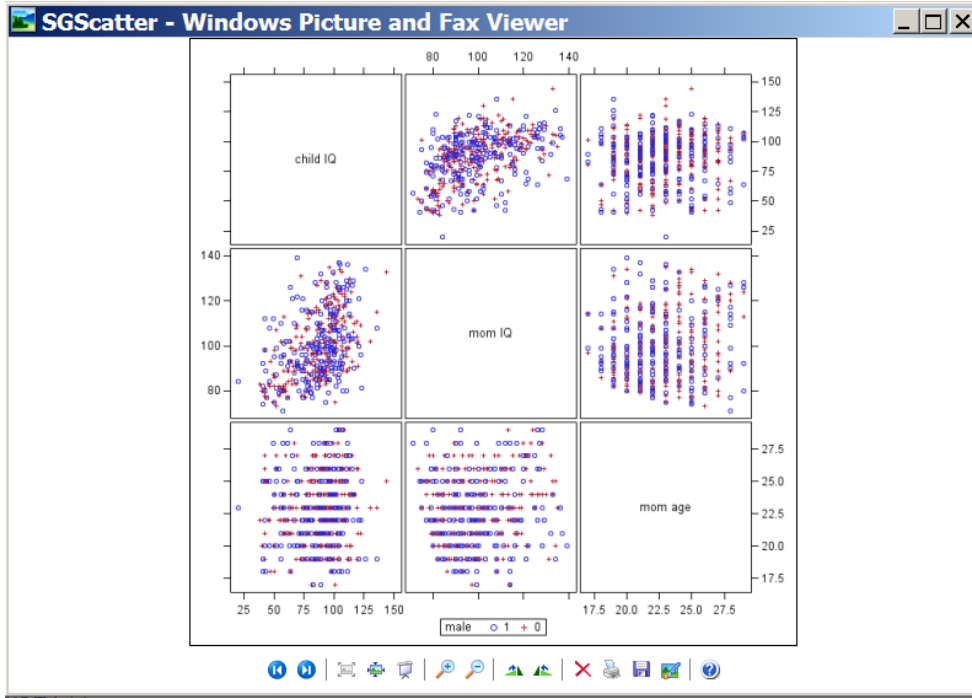
7

To open an SG graphic, go to the Results panel on the left
click the + next to Sgscatter



Then click on the icon next to Sgscatter to see plot icons

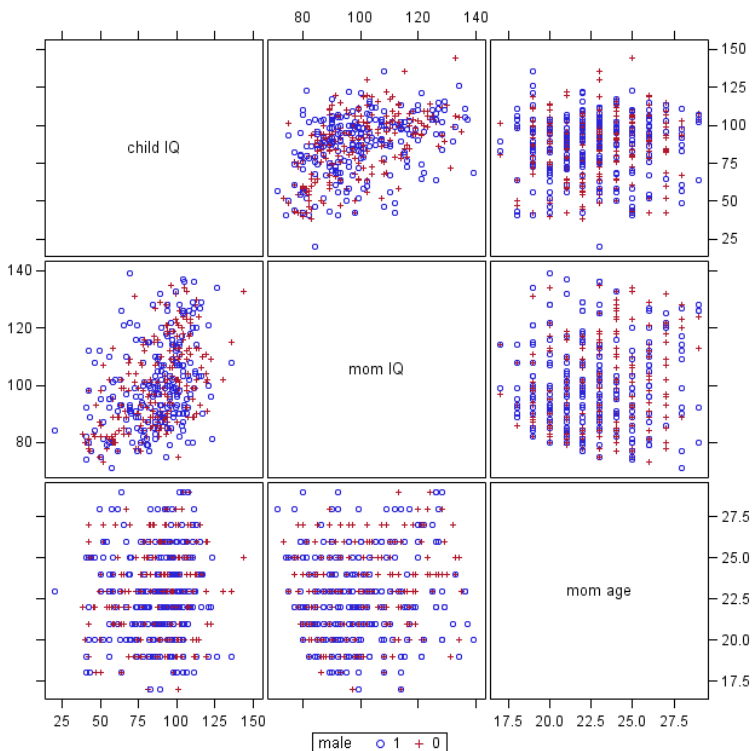
8



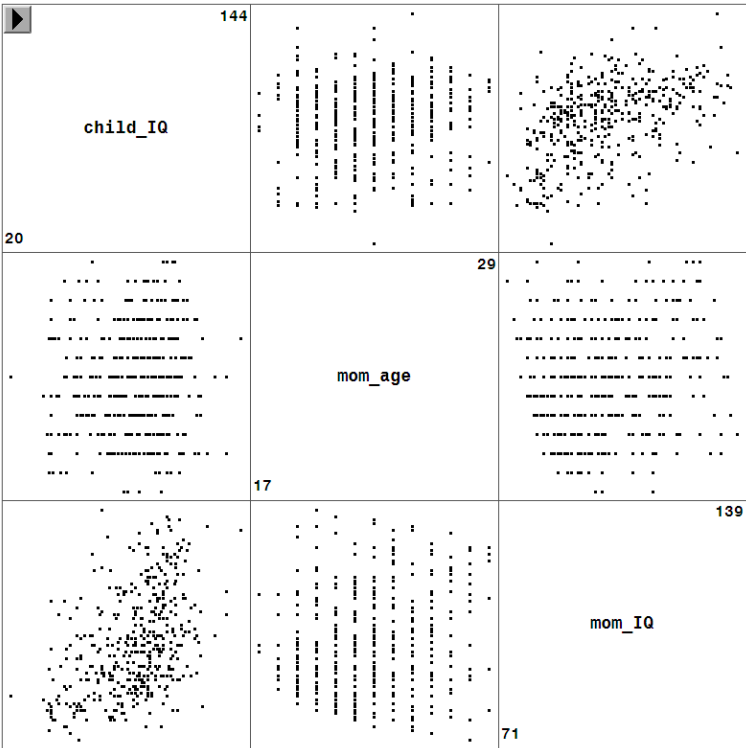
Use the floppy-disk icon to save to desktop as PNG, JPG or GIF.
Then delete the graphic (otherwise saved deep in SAS program folders).

9

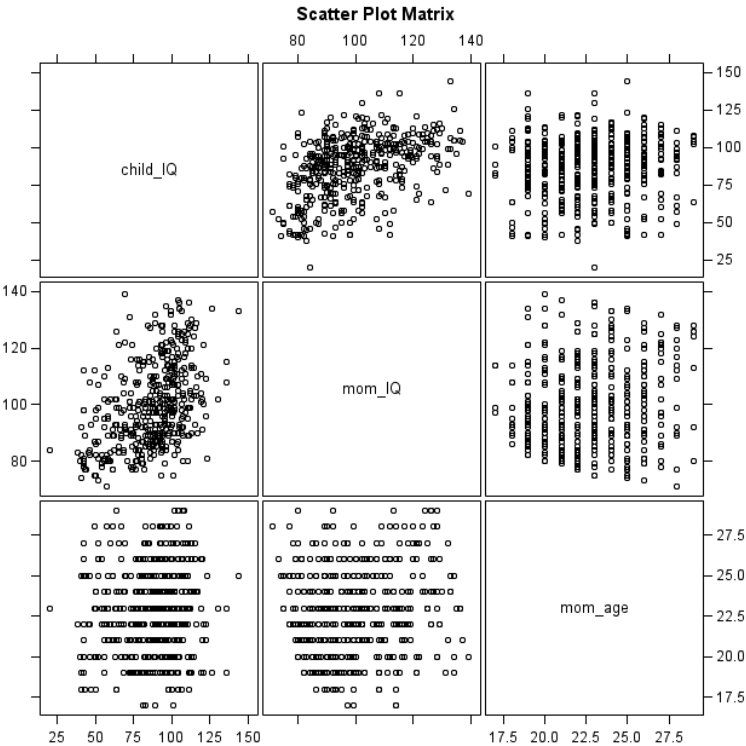
Proc SGscatter shows groups with distinct plotting characters



Proc Insight: click point to identify observation, no grouping; save with ScreenPrint



Proc Corr: not interactive, no grouping; gives correlations



NHANES III.

The third National Health and Nutrition Examination Survey collected data from 33,994 people during 1988–1994, a representative sample of the whole US population. Subset of data for HW 2, adults 20–29.

One of the survey questions was: How many years of school have you had?

Sketch the histogram responses.

13

Two ways to draw histograms:

```
Proc SGplot data=hw2;  
    histogram education_yrs;
```

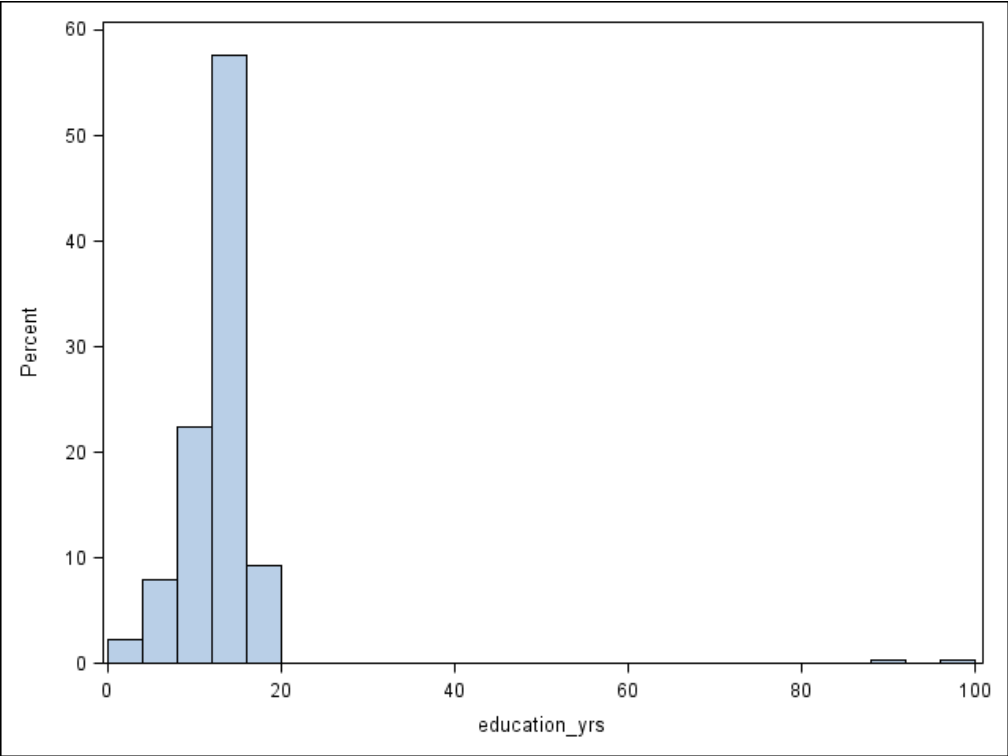
```
Proc Univariate noprint data=hw2;  
    var education_yrs;  
    histogram education_yrs / cfill=graye03; can set endpoints of bins  
    inset q1 median mean q3 / position=NE noframe;
```

Histogram from Univariate opens in SAS window.

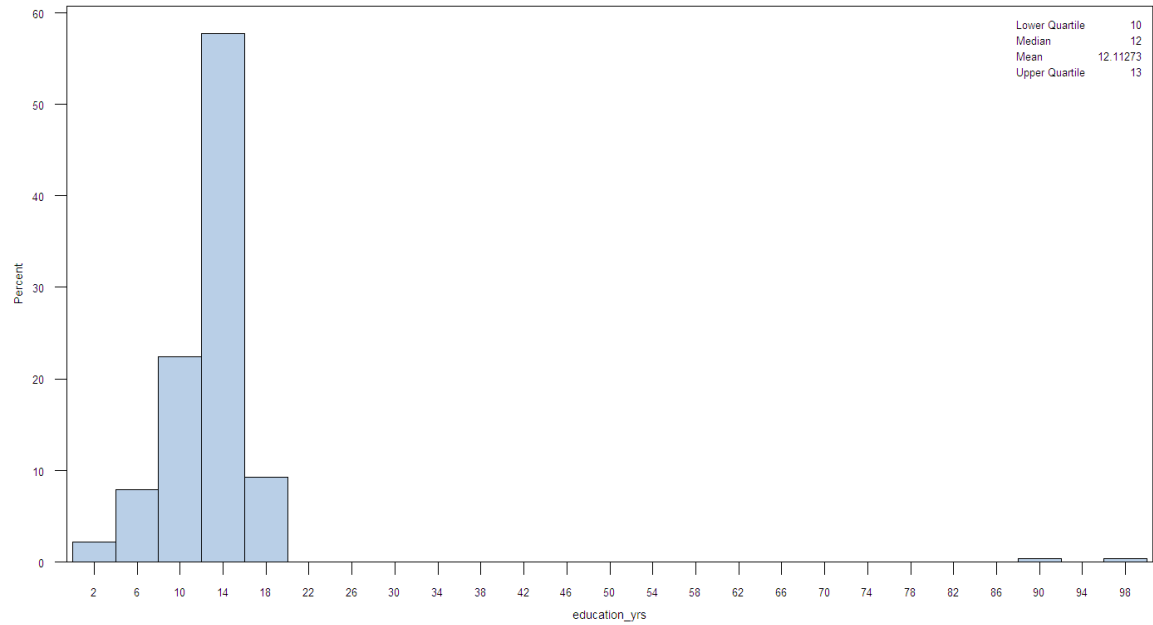
Choose Edit > Export as Image... then save file in desired format.

14

Proc SGplot histogram



Proc Univariate



What are these extremely high values of education_yrs?

Use Proc Freq to list all distinct values.

```
Proc FREQ data=hw2;  
  tables education_yrs;
```

education_yrs	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	29	0.83	29	0.83
1	8	0.23	37	1.06
2	17	0.48	54	1.54
.
11	246	7.01	1139	32.48
12	1268	36.16	2407	68.63
13	297	8.47	2704	77.10
14	290	8.27	2994	85.37
15	169	4.82	3163	90.19
16	219	6.24	3382	96.44
17	102	2.91	3484	99.34
88	11	0.31	3495	99.66
99	12	0.34	3507	100.00

17

The variable education_yrs is really HFA8R in the NHANES III data.

From the documentation for NHANES III adult data:

```
HFA8R    What is the highest grade or year of  
         regular school -- has completed?
```

00 Never attended or kindergarten only

01-17

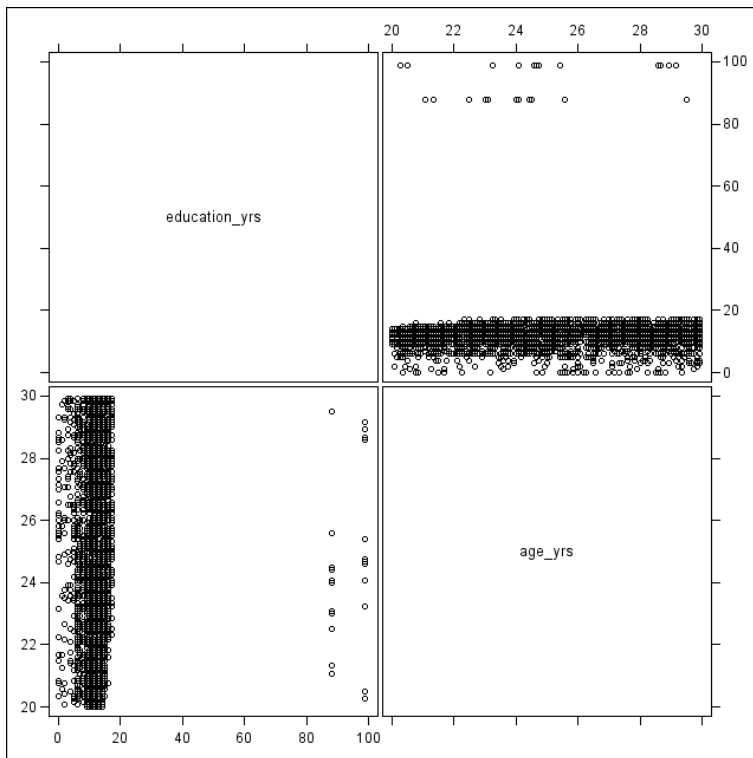
88 Blank but applicable

99 Don't know

88 and 99 are codes for missing data. So exclude values larger than 17 years.

18

We could have found this problem with a scatterplot matrix, too.



19

Editing data in a DATA step

We want to exclude values of `education_yrs` larger than 17 years.

Don't edit a spreadsheet. Make the change in the SAS code, with comments to documents the edit.

Data one;

```
SET hw2;
```

```
IF (education_yrs LE 17) ; * omit obs with missing education;
```

Disadvantage?

20

Better approach—replace 88 and 99 with “missing value code” (period indicates a missing number):

```
Data one;
```

```
  SET hw2;
```

```
  IF (education_yrs > 17) THEN education_yrs = . ;
```

Still better—keep the original variable and make a new corrected variable:

```
Data one;
```

```
  SET hw2;
```

```
  education_yrs_corrected = education_yrs;
```

```
  IF (education_yrs_corrected > 17)
```

```
    THEN education_yrs_corrected = . ;
```

21

What about this?

```
Data one;
```

```
  SET hw20;
```

```
  IF (education_yrs > 17) THEN education_yrs_corrected = . ;
```

What is the value of `education_yrs_corrected`

when `education_yrs = 12`?

First create the variable, then edit it.

22

Scatter plots

Procedures to make a scatterplot:

1. Proc Plot is an old procedure that makes “teletype” graphics in the output file
2. Proc Insight makes better plots, but saving as usable file requires many steps
3. Proc SGplot decent plot, shows groups, easy to save
4. Proc Gplot many options, more complicated

23

Proc Plot is an old procedure that makes “teletype” graphics.

Fixed character width (**hpos**) and fixed interline spacing (**vpos**).

```
Proc Plot data=AAA.child_iq;
```

```
plot child_iq * mom_iq / vpos=25 hpos=60 ;
```

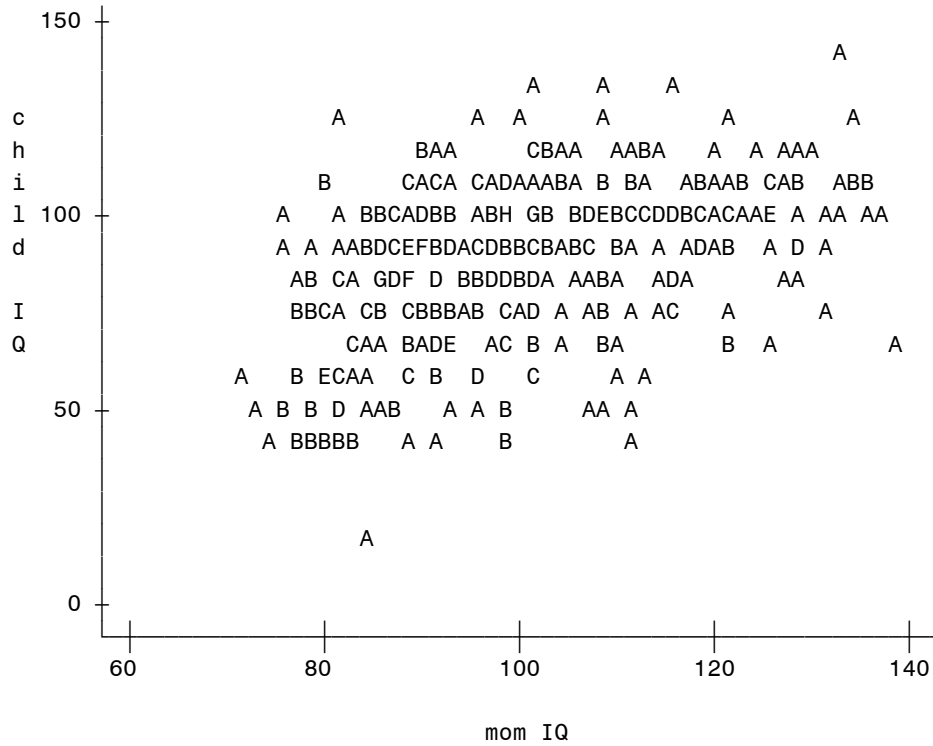
options set vertical and horizontal dimensions in character units

vref = *constant* draws horizontal line (useful for residual plots)

Plots use SAS Monospace font. Paste graphic into MSWord, then convert to PDF.

24

Plot of child_IQ*mom_IQ. Legend: A = 1 obs, B = 2 obs, etc.



25

Scatter plots 2: Proc Insight

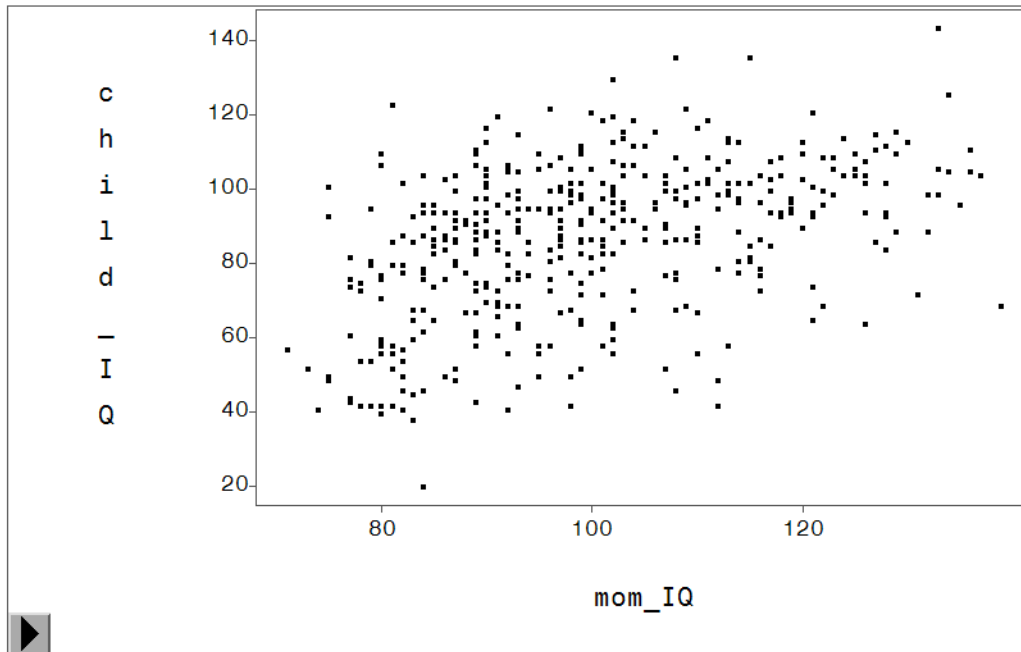
Proc Insight makes graphics interactively, and can be called in a program:

```
Proc Insight data = child_iq;
  scatter child_iq * mom_iq ;
  scatter Y- variable * X-variable
```

Procedures after this close the graphics window, so run Proc Insight separately.

Save or copy/paste from the graphics window that opens.

26



No way to get different plotting symbols by gender.

27

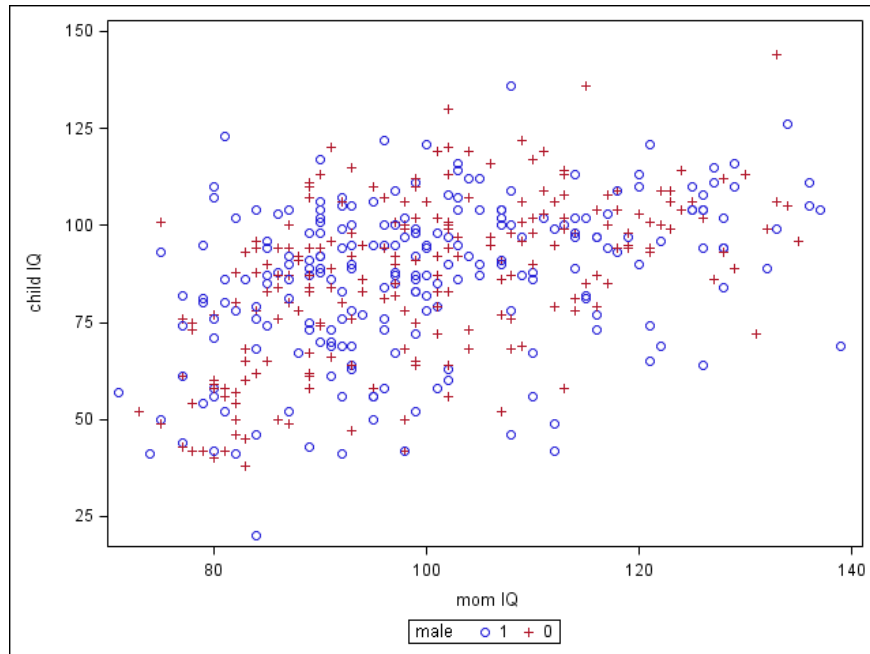
Scatter plots 3: Proc SGplot

SG = *statistical graphics*

SAS Help > SAS/GRAPH > STATISTICAL GRAPHICS PROCEDURES > SGPLOT

```
Proc SGplot data=ph6470.child_iq;
  scatter y = child_iq x= mom_iq / group = male ;
```

28



Default often has poor choice of horizontal and vertical limits.

29

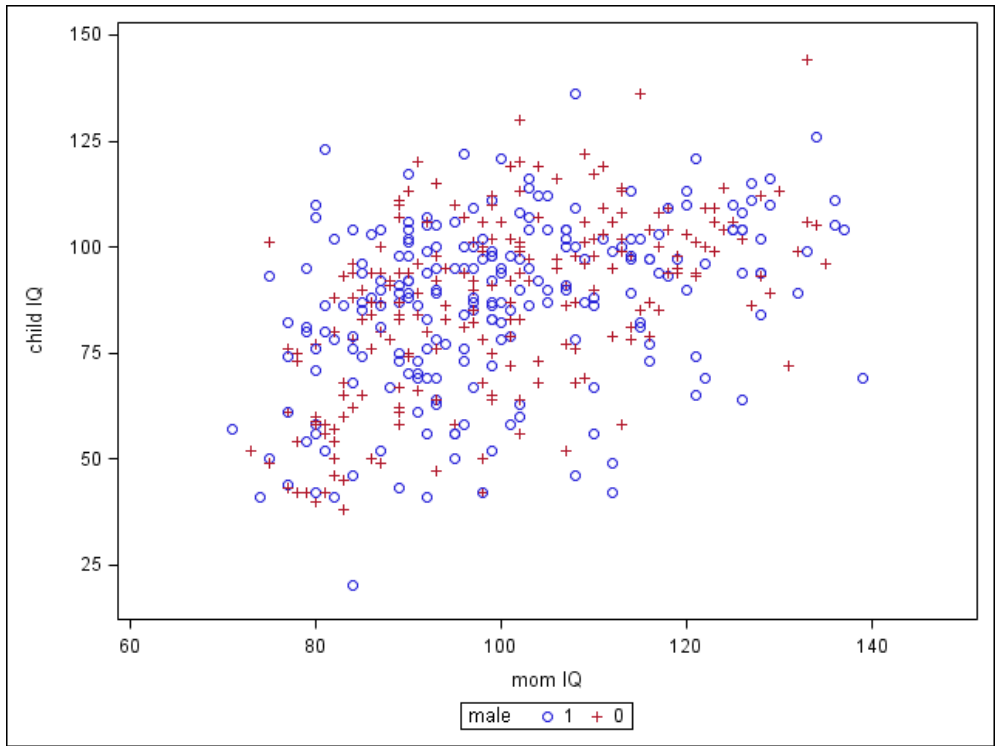
```
Proc SGplot data=ph6470.child_iq;
  scatter y = child_iq x= mom_iq /group = male ;
  xaxis min=60 max=150; set limits of horizontal axis
  yaxis min=15 max=150; set limits of vertical axis
```

```
Proc SGplot data=ph6470.child_iq;
  scatter y = child_iq x= mom_iq /
  group = male MARKERATTRS=(symbol="circlefilled" size=8) ;
  xaxis min=60 max=150; yaxis min=15 max=150;
```

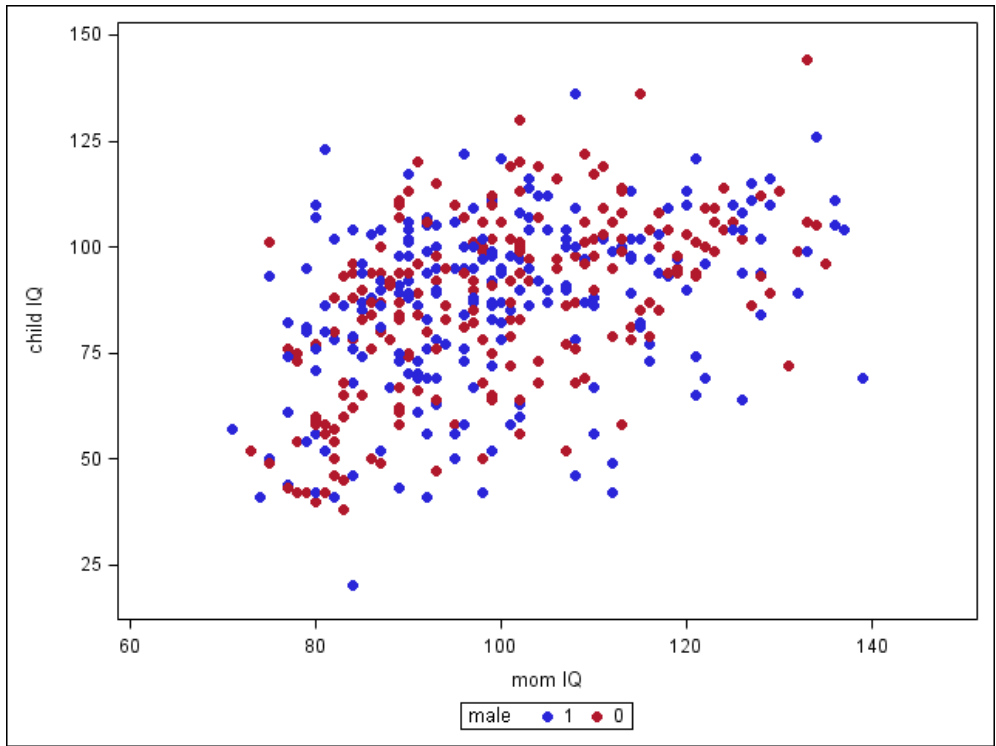
Another option, for residual plots:

```
refline 0 / axis = y; adds horizontal line at y = 0
```

30



31



with “Circlefilled” plotting characters

32

SGplot will add regression line(s) or nonparametric smooth(s) to the plot.

```
Proc SGplot data=ph6470.child_iq;
```

```
  reg y = child_iq x= mom_iq /group = male; linear regressions by group
```

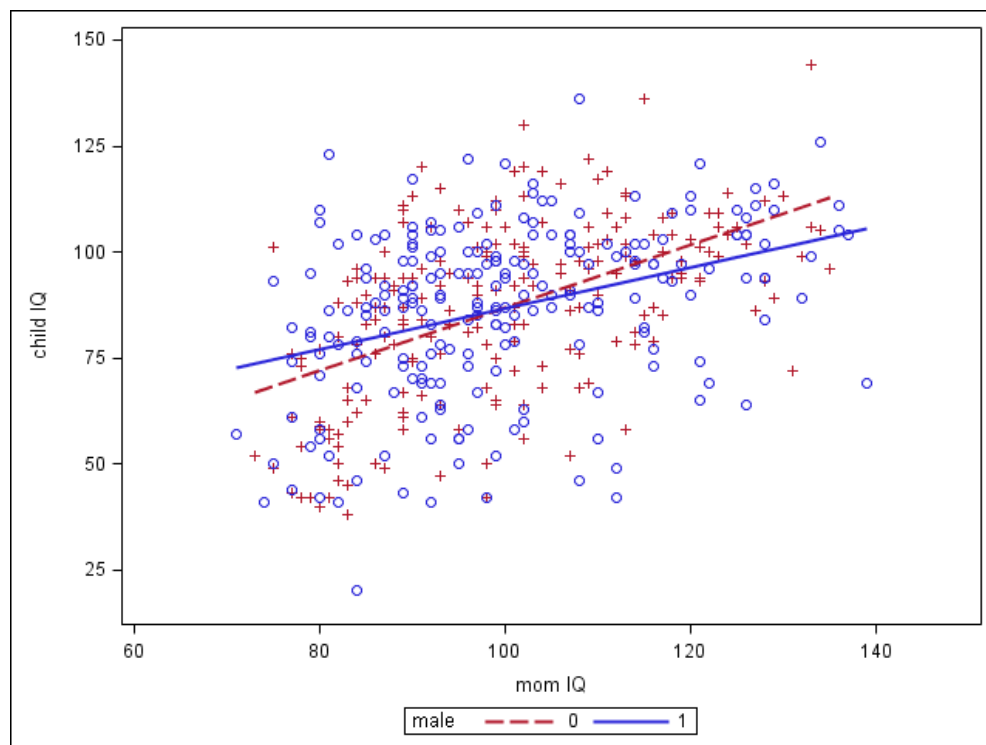
```
  xaxis min=60 max=150; yaxis min=15 max=150;
```

```
Proc SGplot data=ph6470.child_iq;
```

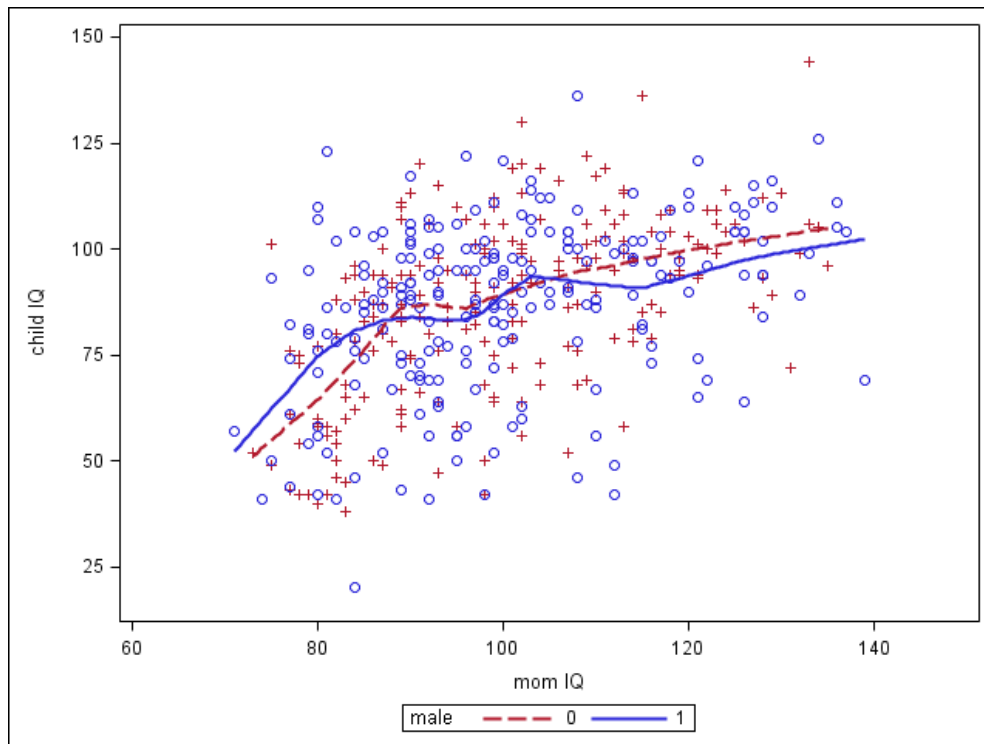
```
  loess y = child_iq x= mom_iq /group = male ; smooth by group
```

```
  xaxis min=60 max=150; yaxis min=15 max=150;
```

33



34



35

Scatter plots 4: Proc Gplot

GPlot offers more options and control than SGplot.

SAS Help > SAS/GRAPH > Procedures and Statements > All Procedures > GPLOT

Options *formatting and output statements* ;

```
Proc Gplot data=ph6470.child_iq;
```

```
plot child_iq * mom_iq = male ; Y-variable * X-variable = Group variable
```

36

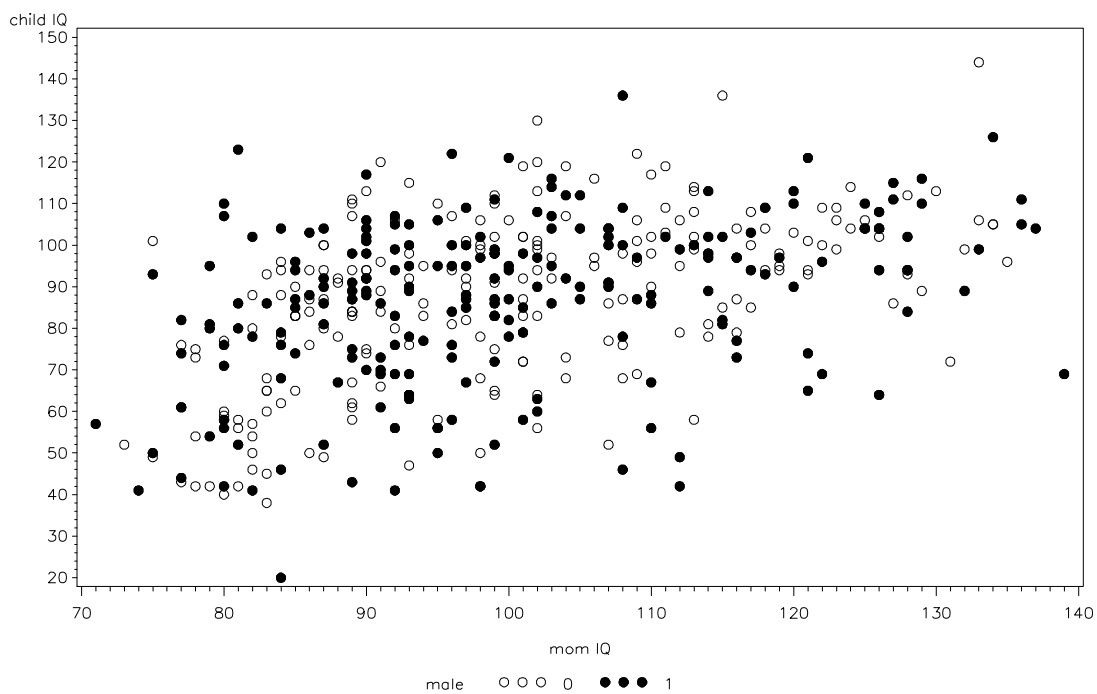
```
Goptions reset=all device=GIF gsfname=graphout gsfmode= replace
noborder vsize=5in figure height in inches
ftext=simplex lfactor=2.5 ; thicker lines
filename graphout "C: ...(path to desktop)... Class05.gif ";
```

```
symbol1 value=circle height=1 color=black; open circles
symbol2 value=dot height=1 color=black; filled circles
```

```
Proc Gplot data=ph6470.child_iq;
plot child_iq * mom_iq = male ;
```

One *symbolN* statement for each group (value of the grouping variable)

37



38