

Statistical Methods and Computing, 22S:30/105

Instructor: Cowles

Lab 2

Feb. 6, 2015

1 Downloading files and accessing SAS.

We will be using the "billion.dat" dataset again today, as well as the OECD dataset on health care expenses. Read the file "OECD.info" to learn about the OECD dataset. Also look at the OECD.data dataset.

Then call up SAS.

2 Sorting, scatterplots, correlation and regression

In the following SAS code, lines that begin with an asterisk are comments and do not need to be typed.

```
*****
* Setting the number of characters *
* in output lines and pages      *
***** ;

options linesize = 79 pagesize = 60 ;

***** *
* Reading the billionaire *
* dataset into SAS      *
***** *

* Use this version if you are running SAS on the computer you are on ;
data billion ;
infile 'c:\temp\billion.dat' ;
input wlth age region $ ;
run ;

* Use this version if you are running SAS on the Virtual Desktop ;
data billion ;
input wlth age region $ ;
datalines ;
< paste data in here >
;
run ;

*****
* Sorting a dataset *
***** ;

* Note: If we want to produce separate output for different subsets of
* a dataset, we must first sort the dataset by the variable that
* defines those subsets ;

proc sort data = billion ;
by region ;
```

```
run ;

*****
* Producing separate analysis for *
* each region                      *
***** ;

* Note: In addition to a complete univariate analysis within each
* region, this procedure produces side-by-side boxplots of wealth
* by region ;

proc univariate plot data = billion ;
var wlth ;
by region ;
run ;

*****
* Producing a scatterplot *
***** ;

* Note: the following code plots wlth on the y-axis and age on the x-axis;

proc plot data = billion ;
plot wlth * age ;
run ;

*****
* Reading the OECD dataset *
* into SAS                  *
***** ;

* Note: the "13." in the "input" statement tells SAS the number of
* characters in the longest country name. Without this information,
* SAS would truncate the country names to 8 letters each ;

data OECD ;
input country $ 13. pcgdp pch beds los docs infmort ;
datalines ;
<paste data here>
;
run ;

*****
* Better text scatter plots *
***** ;

proc plot data = OECD ;
plot pch * pcgdp = '.' / vpos = 20 hpos = 40;
run ;

*****
* Correlation *
***** ;
```

```

proc corr data = OECD ;
var pcgdp pch ;
run ;

*****
* Regression *
***** ;

proc reg data = OECD ;
model pch = pcgdp ; * model <resp vbl> = <explanatory vbl> ;
id country ;
* identifies observations in list of predicted
values and residuals ;
run ;

```

```

*****
* Predicted values *
* and residuals *
*****

```

\* Note: the "p" option on the "model" statement gets list of predicted values and residuals ;

```

proc reg data = OECD ;
model pch = pcgdp / p ;
id country ;
run ;

```

```

*****
* Scatterplots and *
* Residual plots *
*****

```

\* Note: the "lp" option on the "proc reg" statement makes any plots become text plots that appear in the output window. Without this option, you get prettier plots that are harder to print ;

```

proc reg data = OECD lp ;
model pch = pcgdp / p ;
plot pch * pcgdp / symbol = ' .' hplots = 2 vplots = 2 ;
run ;
plot residual. * predicted. / symbol = ' .' hplots = 2 vplots = 2 ;
run ;

```

### 3 Analyst for regression

Use the following steps to get into "Analyst" from the menu:

- Solutions
  - Analysis
    - \* Analyst

You must specify which dataset you wish to use. Do so by clicking

- File
  - Open by SAS name
    - \* Work library (double click)
      - OECD (double click)

To create a scatterplot, choose "Graphs/Scatterplot." Use the interactive window to specify the explanatory variable on the X axis and the response variable on the Y axis.

To do regression analysis, choose "Statistics/ Regression/Simple." Again, interactively specify the explanatory and response variables. Other choices in the window can be used to request predicted values and specific plots.

### 4 Insight for regression

Insight is another point-and-click facility built into SAS. We will be using its graphical features later on when we study multiple regression. In case you want to try it now, here are some instructions.

From the main pull-down window, select the following sequence of choices:

- Solutions
  - Analysis
    - \* Interactive data analysis

In the window that appears, you must specify which dataset you wish to use. Do so by clicking

- Library: Work
  - Dataset: OECD
    - \* Open

To do regression in Insight, choose

- Analyze
  - Fit

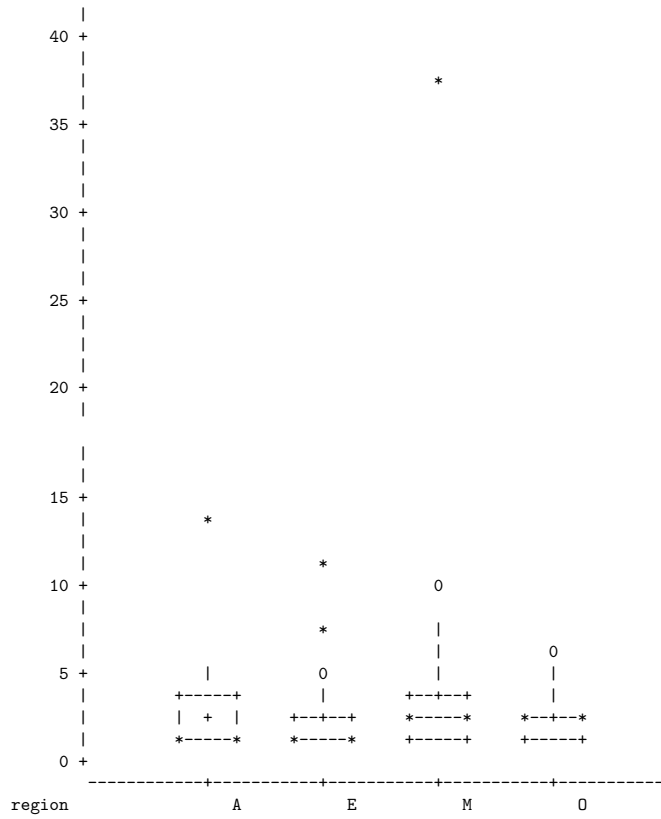
To identify the response variable, use your mouse to click "PCH" and then "Y." Similarly, copy "PCGDP" into the "X" column. Click "OK" and lots of regression output and plots will appear.

To get out of Insight and back into command mode, click in the window showing the data in spreadsheet form. Then pull down the "File" menu and choose "End."

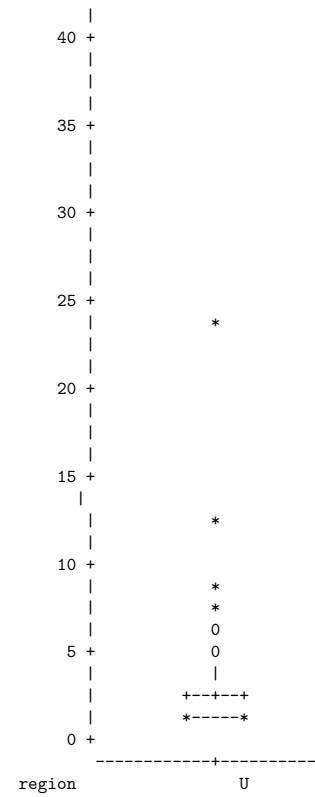
### 5 Remember to exit from SAS and log out of your hawkid

Output

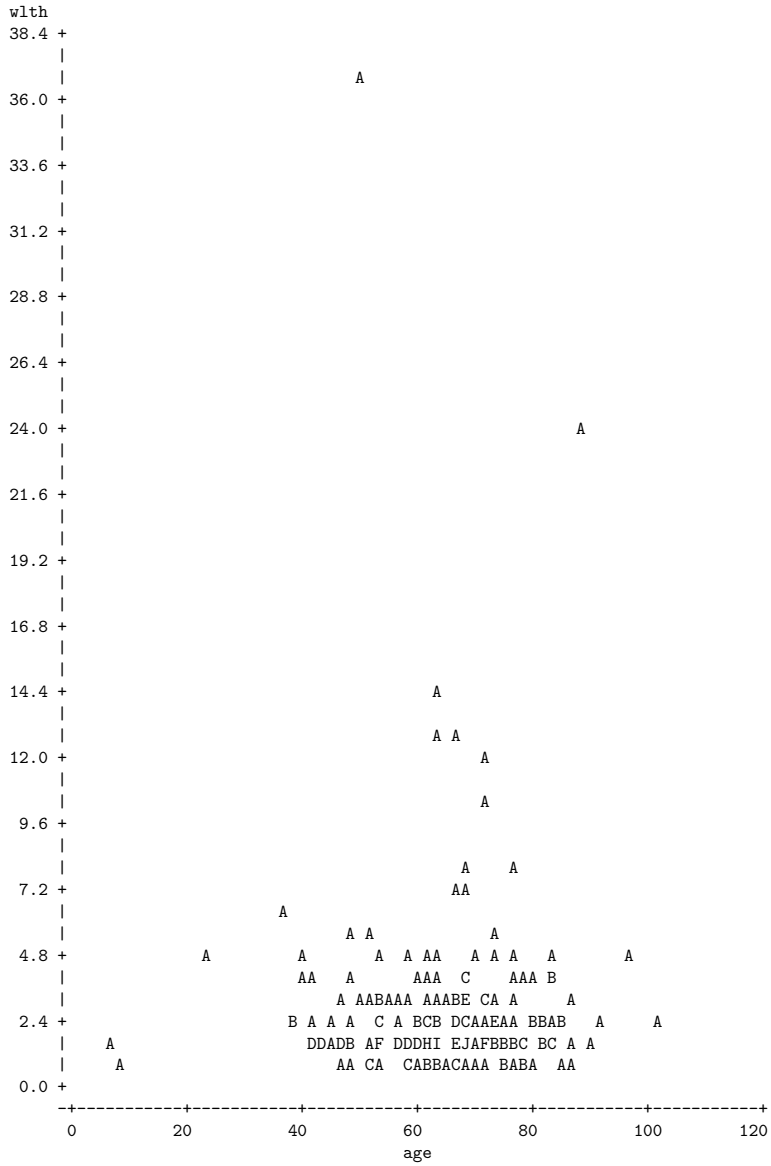
The UNIVARIATE Procedure  
Schematic Plots



The UNIVARIATE Procedure  
Schematic Plots

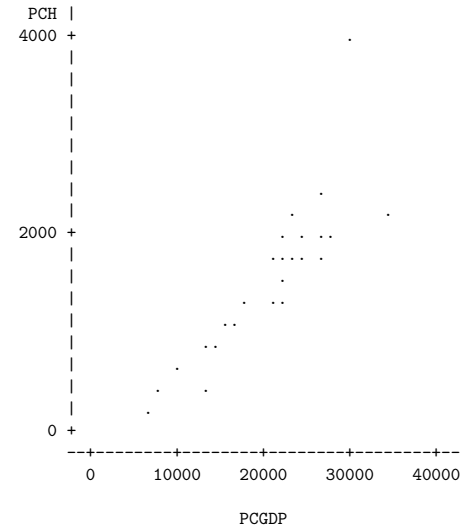


Plot of wth\*age. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 8 obs had missing values.

Plot of PCH\*PCGDP. Symbol used is '.'.



NOTE: 4 obs hidden.

The CORR Procedure

2 Variables: pcgdp pch

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pcgdp	29	20395	6871	591441	6720	34536
pch	29	1509	760.95177	43758	232.00000	3898

Pearson Correlation Coefficients, N = 29  
Prob > |r| under H0: Rho=0

	pcgdp	pch
pcgdp	1.00000	0.87420 <.0001
pch	0.87420 <.0001	1.00000

Model: MODEL1  
 Dependent Variable: PCH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	12390694.827	12390694.827	87.518	0.0001
Error	27	3822637.8631	141579.18012		
C Total	28	16213332.69			
Root MSE	376.27009	R-square	0.7642		
Dep Mean	1508.89655	Adj R-sq	0.7555		
C.V.	24.93677				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	-465.663682	222.33243595	-2.094	0.0457
PCGDP	1	0.096818	0.01034925	9.355	0.0001

Model: MODEL2  
 Dependent Variable: PCH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	12390694.827	12390694.827	87.518	0.0001
Error	27	3822637.8631	141579.18012		
C Total	28	16213332.69			
Root MSE	376.27009	R-square	0.7642		
Dep Mean	1508.89655	Adj R-sq	0.7555		
C.V.	24.93677				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >  T
INTERCEP	1	-465.663682	222.33243595	-2.094	0.0457
PCGDP	1	0.096818	0.01034925	9.355	0.0001

Obs	COUNTRY	Dep Var PCH	Predict Value	Residual
1	Australia	1775.0	1731.0	43.9558
2	Austria	1748.0	1856.5	-108.5
3	Belgium	1708.0	1867.4	-159.4
4	Canada	2065.0	1903.3	161.7
5	CzechRepub	904.0	806.2	97.7631
6	Denmark	1802.0	2078.7	-276.7
7	Finland	1380.0	1631.3	-251.3
8	France	2002.0	1673.1	328.9
9	Germany	2278.0	1745.2	532.8
10	Greece	888.0	934.6	-46.6178
11	Hungary	602.0	553.3	48.7491
12	Iceland	1893.0	2080.3	-187.3
13	Ireland	1276.0	1713.6	-437.6
14	Italy	1584.0	1639.1	-55.0669
15	Japan	1677.0	1868.5	-191.5
16	Korea	537.0	845.3	-308.3
17	Luxembourg	2139.0	2878.0	-739.0
18	Mexico	358.0	308.7	49.3118
19	Netherlands	1766.0	1769.1	-3.0938
20	NewZealand	1270.0	1249.2	20.8199
21	Norway	1928.0	2196.5	-268.5
22	Poland	371.0	307.5	63.4736
23	Portugal	1071.0	1012.4	58.6372
24	Spain	1115.0	1155.1	-40.0728
25	Sweden	1675.0	1588.1	86.8594
26	Switzerland	2499.0	2108.3	390.7
27	Turkey	232.0	185.0	47.0454
28	UnitedKingdom	1317.0	1584.0	-267.0
29	UnitedStates	3898.0	2488.6	1409.4

Model: MODEL1

