Statistical Methods and Computing, STAT 2010 / STAT 4200
Lab 2

1 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        *
************************ *

* If you are running SAS on the computer you are on,
* download the files ‘‘billion.dat" and ‘‘OECD.dat" from the course web page,
   suppose that it is saved in the ‘temp’ folder
* and use the following code:

* data billion ;
* infile 'c:\temp\billion.dat' ;
* input wlth age region $ ;
* run ;

* If you are running SAS on the Virtual Desktop,  use the following code:

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 in 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 ;
```

See next few pages for selected SAS output

The SAS System

The CORR Procedure

2 Variables:	pcgdp pch

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pcgdp	30	20381	6752	611441	6720	34536
pch	29	1509	760.95177	43758	232.00000	3898

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	pcgdp	pch
pcgdp	1.00000 30	0.87420 <.0001 29
pch	0.87420 <.0001 29	1.00000 29

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: pch

Number of Observations Read	30
Number of Observations Used	29
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12390695	12390695	87.52	<.0001
Error	27	3822638	141579		
Corrected Total	28	16213333			

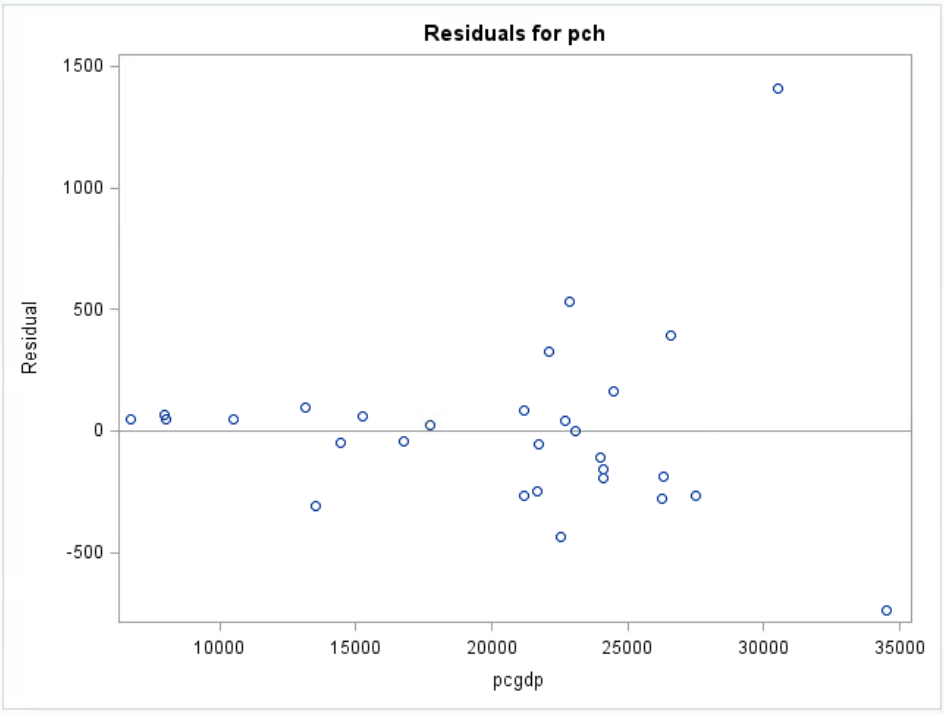
Root MSE	376.27009	R-Square	0.7642
Dependent Mean	1508.89655	Adj R-Sq	0.7555
Coeff Var	24.93677		

Parameter Estimates

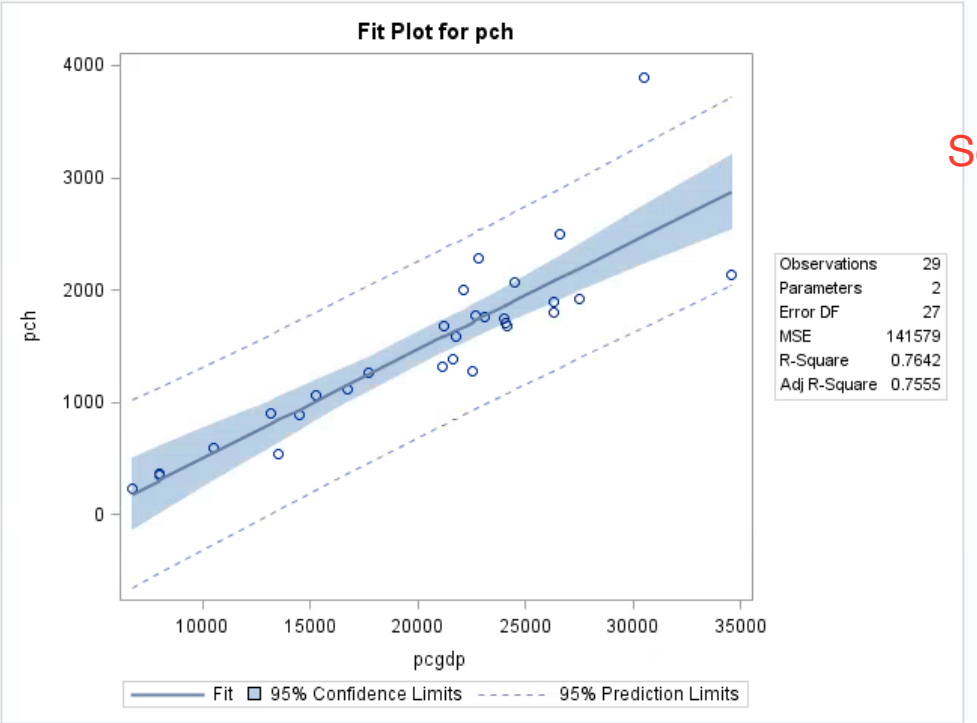
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-465.66368	222.33244	-2.09	0.0457
pcgdp	1	0.09682	0.01035	9.36	<.0001

Estimate of intercept parameter, a

Estimate of slope parameter, b



residual plot



Scatter plot with regression line imposed on it

We haven't learned about confidence bands yet. Ignore them for now

2 **Remember to exit from SAS and log out of your hawk ID**