Statistical Methods and Computing, STAT 2010 / STAT 4200
Lab 3

1 Correlation

Let $(x_i, y_i),\ i = 1, \cdots, n$ be pairs of observations of (x, y). Then the (sample) correlation between x and y is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right),$$

where $\overline{x}$, $\overline{y}$ denote the (sample) mean of x and y respectively and s_x, s_y denote the (sample) standard deviations of x and y respectively.

- (Sample) correlation is a measure of *linear* relationship between two variables.
- r is always between -1 and 1
- $r > 0$ indicates positive linear association, $r < 0$ indicates negative linear association and $r = 0$ indicates no linear association
- $r = 0$ doesn't mean no relationship and it is entirely possible for two variables to have some relationship (such as quadratic relationship) while having $r = 0$
- (Sample) correlation is unit free

2 Correlation and regression in SAS

```
  ****************************
  * 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 ;

filename oec url "http://homepage.divms.uiowa.edu/~kcowles/Datasets/OECD.dat";
data OECD;
infile oec;
input country $ 13. pcgdp pch beds los docs infmort ;
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

r=0.87420

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			

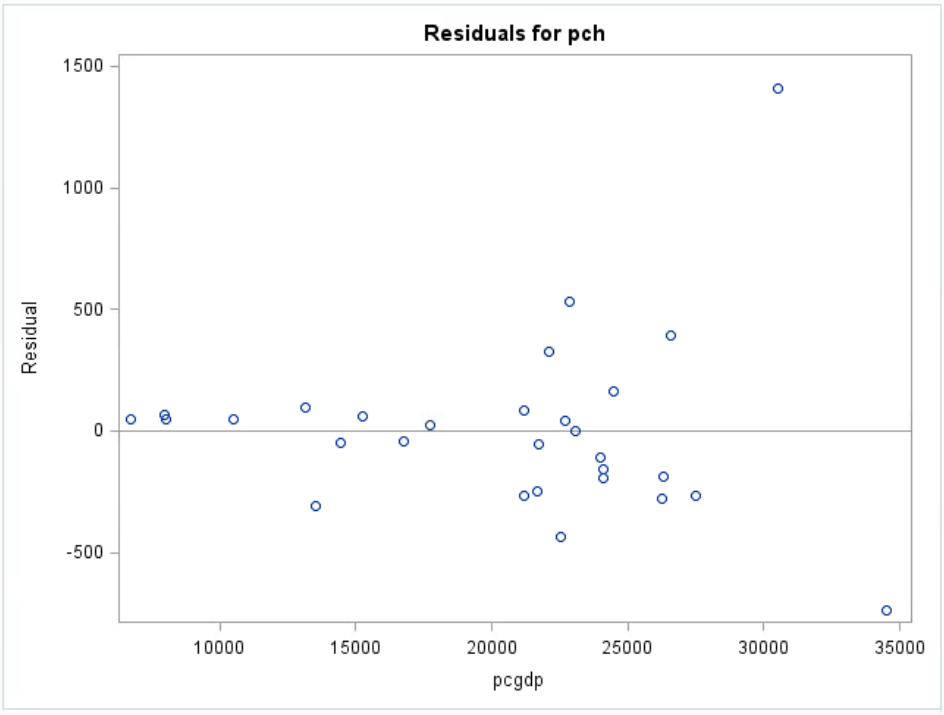
Root MSE	376.27009	R-Square	0.7642
Dependent Mean	1508.89655	Adj R-Sq	0.7555
Coeff Var	24.93677		

R-Square = 0.7642

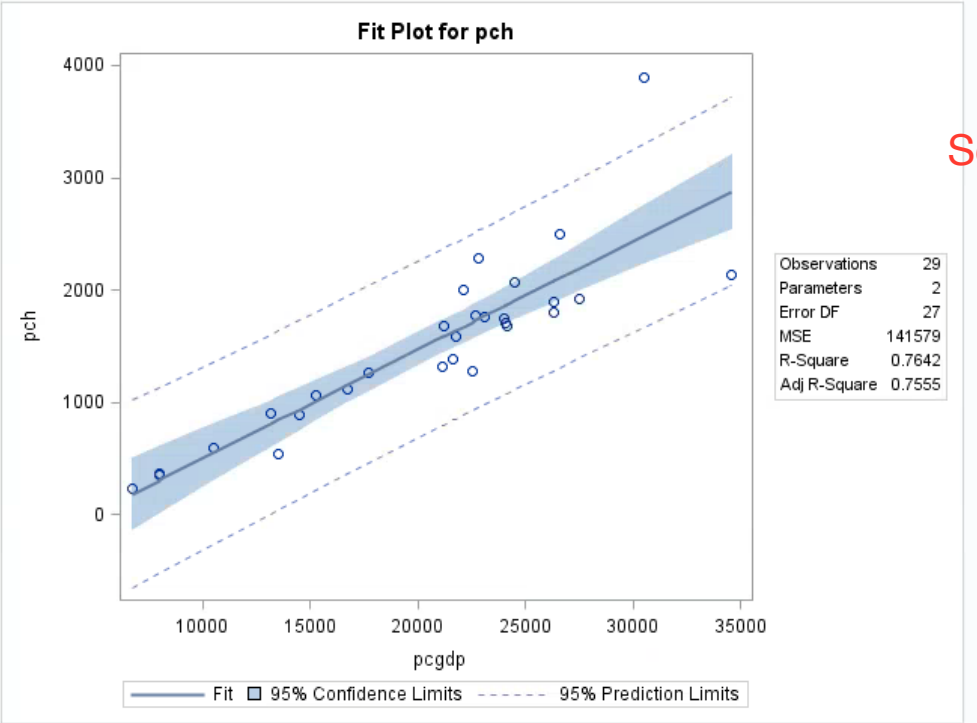
Estimate of intercept parameter, a

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 slope parameter, b



residual plot



Scatter plot with regression line imposed on it